Hamilton Institute

Maynooth University
National University of Ireland Maynooth

Bayesian Statistical Machine Learning Models for Predicting Multivariate Data with Non-Ignorable Partial Missingness

A thesis submitted in fulfillment of the requirements for the Ph.D. degree in Statistics by

Yong Chen Goh

Under the supervision of

Prof. Andrew C. Parnell
Dr. Keefe Murphy

at the

Hamilton Institute
Maynooth University
Maynooth, Co. Kildare, Ireland

February 2025

"But in the end, it's only a passing thing, this shadow. Even darkness must pass. A new day will come. And when the sun shines it will shine out the clearer."
Samwise Gamgee

To Obi, Mocha, and Haydn.

Declaration

I hereby declare that this thesis and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: YONG CHEN GOH

Date: 27/02/2025

Sponsor

This work was sponsored by the Science Foundation Ireland Insight Centre for Data Analytics under grant number SFI/12/RC/2289 P2.

Collaborations

Andrew C. Parnell: As my supervisor, Prof. Parnell (UCD) 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.

Wuu Kuang Soh: Dr. Soh (National Botanic Gardens of Ireland) contributed some text concering the interpretation of the results for the *global Amax* data in Chapter 4.

Publications

The material in Chapter 3 and Chapter 4 has been submitted to the peer-reviewed, open-access journal *Bayesian Analysis*.

Submitted articles (under review):

- Yong Chen Goh, Wuu Kuang Soh, Andrew C. Parnell, and Keefe Murphy (2025+). Joint models for handling non-ignorable missing data using Bayesian additive regression trees. Under review with *Bayesian Analysis*.
  arXiv pre-print: `https://arxiv.org/abs/2412.14946`.

Acknowledgements

This PhD thesis would not have been possible without the support and guidance of many outstanding people.

First and foremost, I would like to express my deepest gratitude to my supervisors, Andrew and Keefe. Their exceptional expertise, unwavering support, and encouragement have been invaluable throughout this journey. They struck the perfect balance between guidance and independence, offering both motivation and empathy when it was needed most. I am profoundly grateful for their patience, insight, and dedication. It has been a privilege to learn from such exceptional and dedicated individuals.

To my non-biological family here in Ireland: words cannot fully express how much your support and friendship have meant to me. Through the most challenging moments, you brought strength, laughter, and companionship, making even the toughest times more bearable. I am forever grateful to have had you by my side throughout this journey. I wouldn't be the person I am today without you. Eleni, you have been with me since I first started this PhD, helping me navigate every disaster along the way. I honestly don't know what I would have done without you. Blake and Akash, they say you should never live with your friends, but I wouldn't change a single day of living with you both. From deep chats to Tesco trips, game nights, and random dance sessions, those moments will always stay with me. Orla, you brought nothing but joy into my life, no misery business in sight. Fergal and Jommy, thank you for all the laughs and chats. They never failed to brighten my day.

To my family, thank you for your support throughout this PhD journey. A special thank you to my sister, Leah, for the phone calls and visits, and to Celine, my best friend and sister by extension, for always picking up the phone, despite the 11-hour time difference. I also thank Niloufar, whose generosity and warmth made even the toughest days a little lighter.

I would also like to thank Kate and Rosemary. Their kindness and dedication made every step of the process so much smoother, easing many of the challenges along the way. Their warmth and generosity never went unnoticed, and seeing them in the Hamilton always brightened my day. Additionally, I'd like to thank Gráinne and Janice for all their

help within the Maths & Stats department.

I also extend my gratitude to those from Andrew's research group—Alan, Alessandra, Amin, André, Anthony, Dáire, Danilo, Emma, Estevão, Maeve, Mateus, and Nathan—from whom I learned so much during our group meetings.

I am also grateful to Ahmed, Anna, Aoife, Beatrice, Conor, Cormac, Dara, Darshana, Jack, Kevin, Paddy, and Shauna, along with many other friends and colleagues from the Hamilton Institute, for their support and camaraderie. Beyond the PhD, I'd also like to extend my thanks to Amaya, Antoine, Dimitra, James, Jason, Murilo, Neisha, Osvaldo, Saurabh, Shane, and Shuchi. Whether through shared music taste, great conversations, or other much-needed distractions, you gave me a life beyond academia, and for that I am truly grateful.

Abstract

Missing data is a pervasive challenge in statistical modelling, particularly in multivariate response settings where partial missingness leads to complex, overlapping missingness patterns. Standard methods often rely on strong and unrealistic ignorability assumptions, such as missing completely at random (MCAR) or missing at random (MAR), typically employing complete-case analysis or imputation, leading to inefficiencies and biases. This thesis introduces three novel Bayesian joint models, integrating the selection model framework with Bayesian additive regression trees (BART) to provide a flexible, non-parametric solution for handling non-ignorable partial missingness in multivariate data.

The motivation for these models arises from limitations of standard missing data techniques, as exemplified by the *global Amax* dataset which exhibits substantial, overlapping missingness in the response variables. Original methods applied to this dataset implicitly assume ignorability, leading to biased inferences and loss of information. To address this, our novel models jointly estimate both the response and missingness processes, enabling the recovery of non-ignorable missing not at random (MNAR) mechanisms, in addition to MCAR and MAR. These models also extend to settings with partially observed covariates with ignorable missingness.

By leveraging BART's ability to flexibly model complex, non-linear relationships, we adopt a multivariate BART framework to capture dependencies across responses while maintaining predictive flexibility. For the missingness mechanism, we explore both parametric and non-parametric Bayesian approaches. The probit regression model allows for the incorporation of prior information on the missingness mechanism, offering greater interpretability when domain knowledge is available. In contrast, the probit extension of BART allows for automatic variable selection and flexibly models complex interactions. Additionally, we adopt a seemingly unrelated framework to model dependencies across responses while allowing dynamic response-covariate relationships. These methods are evaluated through extensive simulations and applied to the *global Amax* dataset, demonstrating strong performance in identifying non-ignorable missingness structures and recovering unobserved values.

Contents

List of Figures

List of Tables

1 Introduction

Missing data presents a significant challenge in statistical modelling, particularly in multivariate response settings where missingness patterns can be complex due to overlap, and response variables may be interdependent. Section 1.1 of this chapter introduces the motivation behind this research, highlighting the limitations of existing missing data methods in the presence of non-ignorable missingness and the need for more flexible approaches. Section 1.2 provides an overview of the thesis structure, summarising the aims and content of each chapter.

1.1 Motivation

Missing data is a pervasive challenge in statistical modelling, often compromising the reliability and robustness of inferential and predictive analyses. This issue is particularly pronounced in multivariate response settings, where missingness in one response may depend on the values of other responses, which themselves may be partially observed. Moreover, missingness patterns can vary across observations, making traditional approaches—such as complete-case analysis or standard imputation techniques—potentially inadequate. Addressing missing data appropriately is essential for drawing valid conclusions, yet many analyses either fail to account for missingness explicitly or rely on strong, often unverified assumptions about the missing data mechanism.

The *global Amax* dataset, originally analysed in Maire et al. [2015], presents a compelling example of these challenges. The study examined how 20 soil and 26 climate variables influence five key leaf photosynthetic traits: specific leaf area (SLA), leaf photosynthetic rate per area ($Aarea$), leaf nitrogen content ($Narea$), leaf phosphorus content ($Parea$), and stomatal conductance (Gs). Despite having fully observed covariates, the dataset exhibits substantial missingness in the response variables, with only 217 out of 2368 observations being fully complete. The missing data patterns vary considerably across traits, with $Aarea$ being nearly complete, while Gs and $Parea$ exhibit high missingness rates of 57.1% and 77.5%, respectively.

While Maire et al. [2015] conducted several analyses, including separate regressions, redundancy analysis, and path analysis, the study provided little to no discussion on how missing data was handled. The only explicit mention was the exclusion of *Parea* from the redundancy analysis due to its small sample size, implying a reliance on complete-case analysis. This approach implicitly assumes the data are missing completely at random (MCAR), where missingness is independent of both observed and unobserved values. However, MCAR is a highly restrictive assumption which is rarely realistic in real-world datasets [Pigott, 2001, Baraldi and Enders, 2010, Van Buuren, 2018]. Even if data are MCAR, complete-case analysis still leads to a substantial loss of information and biased estimates, especially when the proportion of missingness is large [Pigott, 2001, Eekhout et al., 2012, Van Buuren, 2018].

A more flexible assumption is missing at random (MAR), where missingness depends on the observed data but not the unobserved values themselves. Most modern missing data techniques, including multiple imputation and likelihood-based approaches, rely on the MAR assumption [Little, 1993, Van Buuren, 2018]. While such approaches leverage information from observed values to make informed imputations and inferences, they generally do not extend to MNAR scenarios where missingness depends on unobserved values. When data are MNAR, ignoring the missingness process leads to severe biases and invalid inferences [Little, 1995, Tierney et al., 2015]. Unlike MCAR and MAR, MNAR requires explicit modelling of the missingness mechanism, as standard techniques do not fully account for the missingness-induced structure [Kaciroti and Raghunathan, 2014, Van Buuren, 2018, Linero and Yang, 2018, Little and Rubin, 2019, Linero, 2024].

Despite the existence of various missing data methods, including complete-case analysis, single imputation, and multiple imputation approaches, most have only been proven to be valid for MCAR and MAR settings [Eekhout et al., 2012]. While some methods can accommodate MNAR data, they typically require prior knowledge that the data follow an MNAR mechanism, which is rarely known in practice [Baraldi and Enders, 2010, Tierney et al., 2015]. Furthermore, sensitivity analyses are typically necessary to assess the validity of missing data mechanism assumptions, adding an extra step to the analysis process [Kenward and Molenberghs, 1999, Little, 2008, Kaciroti and Raghunathan, 2014, Linero and Yang, 2018].

The aim of this thesis is to develop novel joint modelling approaches for analysing data with multivariate missing responses based on the selection model framework [Heckman, 1976] without ignoring the missingness or imposing strict assumptions on the underlying missing data mechanism. The original framework of Heckman [1976] was developed as a 'two-step' procedure to correct for sample selection bias in univariate response data, where the missingness process is first modelled to estimate the probability of missingness, and these estimates are then incorporated as covariates into the response model. How-

ever, a fully joint modelling approach that simultaneously estimates both the response and missingness process has been shown to be preferable and more efficient in likelihood-based inferences [Puhani, 2000, Bushway et al., 2007, Galimard et al., 2018]. Our models operate in the Bayesian paradigm extend this framework by introducing a multivariate response structure, enabling the handling of data with missingness in multiple responses, such as in the *global Amax* dataset, while addressing key limitations of existing methods. By adopting a fully joint Bayesian modelling approach, our methods ensure that missing data is handled without requiring strong or unverifiable assumptions about its underlying structure. This allows for direct accommodation of all missing data mechanisms, eliminating the need for explicitly specifying the missingness mechanism or conducting separate sensitivity analyses. Additionally, missing responses are imputed simultaneously within the joint model, rather than requiring imputation as a separate pre-processing step.

To ensure flexibility and robustness, our models leverage Bayesian additive regression trees (BART), a non-parametric Bayesian approach developed by Chipman et al. [2010] which is well-suited for capturing non-linear relationships and complex interactions between covariates. Unlike traditional parametric models, BART does not require explicit specification of interactions, as it automatically identifies and models them [Tan and Roy, 2019]. Moreover, BART is highly robust to hyperparameter selection, meaning that little to no tuning is required, and default settings are generally sufficient for strong predictive performance [Chipman et al., 2010, Ročková and Saha, 2019].

To model multiple responses, we utilise the multivariate version of BART [Hahn et al., 2020, Um et al., 2023, McJames et al., 2024], which extends the standard BART framework to simultaneously model multiple correlated responses. While standard BART is designed for univariate outcomes and typically requires separate models for each response, multivariate BART enables information sharing across responses, allowing dependencies between outcomes to be captured naturally. In addition to multivariate BART, we also employ seemingly unrelated BART (suBART) [Chakraborty, 2016, Esser et al., 2024], which consists of separate sets of univariate BART models for each response while explicitly modelling correlations between responses. Unlike multivariate BART, which imposes shared predictor-response relationships, suBART allows each response to maintain distinct associations with covariates, providing greater flexibility in accommodating complex covariate-response structures. This additional flexibility ensures that response-specific patterns are preserved while still leveraging cross-response dependencies, making it particularly useful when responses exhibit heterogeneous relationships with predictors.

In terms of the missingness model, we consider three approaches. First, we propose the parametric probit regression model for multivariate outcomes [Chib and Greenberg, 1998] which allows users to incorporate prior knowledge about the missingness mechanism when such information is available. Second, we propose a non-parametric alternative us-

ing multivariate probit BART, which performs automatic variable selection and requires minimal prior specification. This makes it particularly useful in scenarios where little is known about the missingness mechanism, allowing the model to learn relationships directly from the data. Finally, another non-parametric approach using seemingly unrelated BART supports response-wise modelling, enabling missingness mechanisms to be modelled separately for each response, as opposed to being shared across all responses as in multivariate BART.

By integrating these features, our methods offer a robust, flexible, and principled approach to handling missing data in multivariate response settings. Unlike standard methods that assume a fixed missingness mechanism or rely on separate imputation procedures, our models provide a unified framework for inference and prediction, ensuring that missing data is addressed appropriately and efficiently.

1.2 Thesis outline

The remainder of this thesis is structured as follows: Chapter 2 provides a background on the standard BART model, including its mathematical formulation, prior settings, and posterior computations. Additionally, we discuss probit BART, which extends BART to binary response data using a latent variable framework. We then introduce key concepts in missing data analysis, presenting missing data mechanisms both conceptually and within a mathematical framework. This is followed by a review of common methods for handling missing data and the distinction between ignorable and non-ignorable missingness in likelihood-based approaches. In particular, we describe the formulation of the selection model for MNAR data, which serves as the foundation for the methods developed in this thesis. Finally, we provide background information on the real dataset, *global Amax*, used throughout this thesis. This includes details from Maire et al. [2015] on data collection, compilation, and analysis, as well as their key findings. We also describe the missingness patterns observed in the response variables and discuss the challenges they pose for analysis.

Chapter 3 presents two novel joint models, 'missBART1' and 'missBART2', which integrate Bayesian additive regression trees and the selection model framework to handle multivariate data with missing responses, such as those found in *global Amax*. Both models jointly model the data using multivariate BART, which captures complex, non-linear relationships among variables while introducing flexibility in the data model. The missingness mechanism is then modelled using either multivariate probit regression (missBART1) or multivariate probit BART (missBART2), allowing for a data-driven approach without imposing strong assumptions about the missing data process. By incorporating a fully Bayesian framework, they allow for the simultaneous recovery of MCAR, MAR,

and MNAR data, while also imputing missing values within the MCMC setup. While the models are designed to handle data with missingness in the response which are potentially non-ignorable, they are also capable of accounting for ignorable missingness in the covariates. The chapter begins with an overview of the models, followed by an outline of the selection model framework and how it is adapted for missBART1 and missBART2. We then introduce the Bayesian probit regression model for both univariate and multivariate settings, which serves as the foundation for modelling missingness in missBART1. Next, we formulate multivariate BART, which is used as the data model in both missBART1 and missBART2, detailing its prior specifications and model settings. We also describe the probit counterpart of multivariate BART for binary responses, which is used as the missingness modelling approach in missBART2. Finally, we outline both joint models, providing a comprehensive description of their posterior sampling procedures and inference strategies for handling multivariate data with missing responses.

Chapter 4 evaluates the performance of missBART1 and missBART2 through a series of simulation studies, covering various missing data scenarios, and through application to the *global Amax* data. The simulation experiments include univariate response data with non-linear MNAR missingness, bivariate data with MAR or MNAR missingness under both linear and non-linear missingness mechanisms, and multivariate data with MNAR missingness in the responses and ignorable missingness in the covariates. The proposed models are compared against alternative BART-based missing data methods, including complete-case analysis with multivariate or univariate BART as well as imputation followed by model fitting with multivariate or univariate BART. Model performance is assessed using various error and calibration metrics, along with an evaluation of imputation accuracy. The challenges in assessing model performance in real-world missing data scenarios, where the true values of missing observations remain unknown, are highlighted here. Finally, we apply missBART1 and missBART2 to the *global Amax* dataset, gaining insights into the underlying missingness mechanisms through posterior inferences and variable importance analyses. The results are also compared with those from Maire et al. [2015], highlighting key differences and improvements.

Chapter 5 extends the joint modelling framework by introducing missSUBART, a novel approach that combines the flexibility of seemingly unrelated regression with BART, while accounting for non-ignorable missingness in the responses. The model allows each response to be modelled with its own set of predictors while simultaneously capturing correlations between responses. Unlike missBART1 and missBART2, which impose a shared tree structure across responses, missSUBART offers greater flexibility in scenarios where correlated responses exhibit distinct predictive relationships by allowing separate tree structures for each response while still modelling dependencies. We describe the seemingly unrelated BART structure, detailing its prior specifications and probit extension.

Following this, we formally introduce the missSUBART formulation, along with its posterior computation details. The performance of missSUBART is then evaluated against missBART1, missBART2, and seemingly unrelated BART using both complete cases and imputed data. This evaluation is first conducted in a new simulated scenario, where data and missingness mechanisms exhibit distinct outcome-predictor dependencies. We then extend the comparison to two simulation settings from Chapter 4, further assessing the robustness and adaptability of missSUBART across different missing data structures. Next, missSUBART is applied to the *global Amax* dataset, uncovering additional insights into both the data structure and the underlying missingness mechanisms. While missSUBART introduces added flexibility, it also presents certain limitations, which are discussed alongside potential future improvements to enhance its applicability.

Finally, in Chapter 6, we provide discussions and conclusions on the work carried out in this thesis, offering a comprehensive reflection. The chapter begins by summarising the key contributions, including the development of novel joint modelling approaches for handling multivariate missing data using BART. It then discusses the theoretical and practical implications of these methods, highlighting their advantages over traditional missing data techniques. The chapter also addresses limitations and discusses possible extensions for future work.

2
Background

This chapter gives a background on Bayesian additive regression trees (BART), missing data, and the *global Amax* dataset. In Section 2.1, we introduce BART, which is a flexible, non-parametric Bayesian regression method for capturing complex relationships without strict functional assumptions. BART's sum-of-trees model, Bayesian framework, and backfitting MCMC algorithm are outlined, along with its extension to probit regression for binary outcomes. Next, Section 2.2 discusses missing data, covering the MCAR, MAR, and MNAR frameworks as well as various handling methods. In Section 2.3, the *global Amax* dataset is introduced, highlighting its plant trait, soil, and climate variables. The dataset exhibits high missingness, particularly in the response variables, raising concerns about prior analyses relying on complete-case methods. This motivates the need for advanced missing data techniques, setting the stage for the next chapter. Finally, Section 2.4 summarises the key takeaways from this chapter, laying the groundwork for the development of our joint models that handle multivariate responses while accounting for complex missingness structures.

2.1 Bayesian Additive Regression Trees

Bayesian additive regression trees (BART) is a flexible, non-parametric Bayesian regression approach introduced by Chipman et al. [2010] to model complex relationships using an ensemble of trees. Inspired by ensemble learning methods and boosting algorithms [Freund and Schapire, 1997, Friedman, 2001], BART approximates functions through a sum-of-trees model, where each individual tree contributes a small portion to the overall fit. This additive structure allows BART to capture intricate non-linear relationships and higher-order interactions without requiring explicit specification from researchers [Sparapani et al., 2016, Tan and Roy, 2019]. Unlike traditional parametric regression models, which impose rigid functional forms, BART adapts to the data, making it particularly effective in non-linear settings. Additionally, BART operates within the Bayesian paradigm, providing full posterior distributions that allow for natural uncertainty quantification and

more interpretable inferences than many frequentist and deterministic machine learning methods. In its standard formulation, BART is designed for univariate response variables, where a single outcome is modelled as a function of predictor variables. We extend this approach to multivariate response settings in Chapter 3.

BART builds upon Bayesian CART [Chipman et al., 1998, Denison et al., 1998], a Bayesian extension of classification and regression trees [Breiman et al., 1984], where regularisation is achieved through a prior rather than pruning [Ročková and Saha, 2019]. While Bayesian CART relies on a single tree, BART extends this approach by employing an ensemble of weak learners, with each tree contributing only a small portion to the overall fit. This structure prevents any single tree from dominating the model, enhancing robustness and reducing overfitting.

Each tree in BART represents a simple stepwise function. When combined, they approximate smooth, non-linear functions with minimal user intervention. This ability to automatically detect interactions and complex patterns makes BART especially useful when the underlying relationships between predictors and outcomes are unknown or difficult to model explicitly. The model is invariant to monotone transformations of predictor variables, eliminating the need for extensive data preprocessing [Chipman et al., 2010]. Furthermore, BART is robust to hyperparameter selection, requiring minimal tuning compared to many machine learning algorithms [Ročková and Saha, 2019]. BART is implemented through a Bayesian backfitting Markov chain Monte Carlo (MCMC) algorithm [Hastie and Tibshirani, 2000], efficiently estimating the posterior distribution of the regression function while accounting for prediction uncertainty. The methodology is widely accessible through R-package implementations, such as `BART` [Sparapani et al., 2021b] and `dbarts` [Dorie, 2024], which facilitate its application across various fields.

Since its introduction, BART has demonstrated remarkable success across diverse applications. Empirical studies highlight its effectiveness in various domains: Chipman et al. [2010] introduced BART and showed its superiority over traditional methods in handling complex, non-linear regression problems; Hill [2011] applied BART to causal inference, demonstrating robust treatment effect estimation compared to propensity score matching and other approaches; Bleich et al. [2014] used BART for gene regulation analysis, where it outperformed lasso regression in identifying key predictors; Sparapani et al. [2016] extended BART to survival analysis, outperforming the Cox proportional hazards model by flexibly modelling hazard functions; Linero and Yang [2018] introduced a version of BART with sparsity-inducing soft decision trees, where decision rules within the trees are treated as probabilistic; Prado et al. [2021] developed local linear predictors at BART's terminal nodes, reducing the number of trees needed without sacrificing performance; and Murray [2021] applied BART to count and categorical data, where it outperformed generalised linear models in cases where parametric assumptions were violated.

Beyond its empirical success, BART's theoretical foundation has been rigorously studied. Ročková and Van der Pas [2020] established posterior convergence guarantees, reinforcing BART's reliability in predictive modelling. Additional theoretical advancements have also been made by Linero and Yang [2018], and Ročková and Saha [2019], who explored various aspects of BART, including Bayesian regularisation, theoretical properties of tree ensembles, and posterior contraction rates. In Tan and Roy [2019], a comprehensive tutorial on BART's theoretical underpinnings and practical implementation is provided. This work also introduced a generalised BART framework, which discussed several extensions, including semi-parametric models, methods for handling clustered data with repeated measurements, and statistical matching problems, further expanding BART's scope.

2.1.1 BART Model Setup and Priors

From Chipman et al. [2010], given a fully observed dataset with n observations, suppose the relationship between a univariate outcome $\mathbf{Y}$ and covariates $\mathbf{X}$ can be represented by

$$Y_i = \sum_{k=1}^{K} g(\mathbf{X}_i; \mathcal{T}_k, \boldsymbol{Q}_k) + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \tau^{-1}\right), \tag{2.1}$$

where $i = 1, \ldots, n$, K is the total number of decision trees used in the ensemble, $\mathcal{T}_k$ is the k^{th} decision tree, $\boldsymbol{Q}_k$ is the set of terminal node parameters in tree k, τ is the residual precision parameter, and $g(\cdot)$ is the function which assigns the parameters from $\boldsymbol{Q}_k$ to $\mathbf{X}_i$. In other words, $g(\cdot)$ represents the contribution of a single tree towards the overall sum-of-trees model. We note that in the original work from Chipman et al. [2010], a residual standard deviation σ is specified instead of the residual precision τ.

The posterior distribution of the BART model, given the data $\mathbf{Y}$, takes the form

$$\begin{aligned} p\left((\mathcal{T}_1, \boldsymbol{Q}_1), \ldots, (\mathcal{T}_K, \boldsymbol{Q}_K), \tau \mid \mathbf{Y}\right) \propto & \prod_{i=1}^{n} p(Y_i \mid (\mathcal{T}_1, \boldsymbol{Q}_1), \ldots, (\mathcal{T}_K, \boldsymbol{Q}_K), \tau) \\ & \times \left[\prod_{k=1}^{K} p(\mathcal{T}_k) p(\boldsymbol{Q}_k \mid \mathcal{T}_k)\right] \times p(\tau). \end{aligned}$$

The prior for each tree $\mathcal{T}_k$, $p(\mathcal{T}_k)$, is fundamental to ensuring regularisation and preventing overfitting. It directly influences both tree depth and the probability of variable selection at each split. Following the specifications of Chipman et al. [2010] and the refinements from Tan and Roy [2019], the tree prior $p(\mathcal{T}_k)$ consists of three key components: the probability of a node being non-terminal, the selection of a covariate for splitting, and the choice of a split value given the selected covariate.

First, the probability that a node at depth d will split is given by

$$p_{\text{SPLIT}}(d) = \alpha(1+d)^{-\beta},$$

where $\alpha \in [0,1]$ governs the overall likelihood of node splits, and $\beta > 0$ controls tree depth. Larger values of β discourage deep trees by penalising excessive splitting, thereby promoting regularisation and improving model convergence. By default, Chipman et al. [2010] use a default of $\alpha = 0.95$ and $\beta = 2$. The resulting tree structure for tree $\mathcal{T}_k$ has probability

$$\prod_{\ell'} \alpha(1+d_{\ell'})^{-\beta} \prod_{\ell} \left[1 - \alpha(1+d_{\ell})^{-\beta}\right],$$

where ℓ is the terminal node index and ℓ' is the non-terminal node index within tree k.

Secondly, for each node selected for splitting, a covariate must be chosen to define the split. By default, this selection follows a discrete uniform prior, where each covariate is chosen with equal probability from the set of available covariates, ensuring that no single variable is inherently favoured over others. However, in some cases, adaptive priors can be introduced to promote more informative predictors [Bleich et al., 2014, Linero, 2018]. Finally, once a covariate has been selected, a split point must be determined to partition the data. The split value is typically chosen uniformly from the set of unique observed values for the selected covariate.

Within each tree $\mathcal{T}_k$, each terminal node contains a parameter $\mu_{k\ell} \in \boldsymbol{Q}_k$. A normal prior $\mu_{k\ell} \mid \mathcal{T}_k \sim \mathcal{N}(\mu_\mu, \tau_\mu^{-1})$ is assigned to $\mu_{k\ell}$ and the hyperparameters μ_μ and τ_μ are calibrated based on the data. More specifically, Chipman et al. [2010] describe the strategy of eliciting this prior by first scaling and then shifting the data $\mathbf{Y}$ such that each response falls in the range $[-0.5, 0.5]$. It is then reasonable to set $\mu_\mu = 0$. The prior precision τ_μ is calibrated based on some prior probability ρ_μ that $\mathbb{E}(Y_i \mid \mathbf{X_i})$ falls inside this rescaled interval and is thus obtained through solving

$$\sqrt{\frac{K}{\tau_\mu}}\Phi^{-1}(\rho_\mu) = 0.5, \quad \rho_\mu \in [0,1]\,. \tag{2.2}$$

The default setting used by Chipman et al. [2010] is $\rho_\mu = 0.95$.

For the residual precision parameter τ, a conjugate gamma prior with the shape and rate parameterisation is assigned, i.e. $\tau \sim \text{Ga}(\frac{\nu}{2}, \frac{\nu\lambda}{2})$. A value for ν is first chosen based on the shape of the prior curve. Chipman et al. [2010] explored a range of values and selected a default value of 3. Next, a rough data-based under-estimate $\hat{\tau}$ is obtained, either via the sample precision or the estimated residual precision from a least squares linear regression model, with the latter being our default. With the assumption that the BART model estimates a residual precision τ at least as large as the rough estimate $\hat{\tau}$

with prior probability ρ_τ, the hyperparameter λ can be obtained from

$$P\left(\tau > \hat{\tau}\right) = \rho_\tau, \quad \rho_\tau \in [0,1]\,, \tag{2.3}$$

after selecting appropriate values for ν and ρ_τ. The default setting for ρ_τ from Chipman et al. [2010] is 0.9.

2.1.2 Posterior Computation

One of the important features implemented in BART for posterior sampling and inference is the Bayesian backfitting MCMC algorithm from Hastie and Tibshirani [2000]. The full conditional distribution of $(\mathcal{T}_k, \boldsymbol{Q}_k)$ takes the form

$$(\mathcal{T}_k, \boldsymbol{Q}_k) \mid \mathcal{T}_{(-k)}, \boldsymbol{Q}_{(-k)}, \tau, \mathbf{Y},$$

where $\mathcal{T}_{(-k)}$ and $\boldsymbol{Q}_{(-k)}$ are the sets of $K-1$ trees and terminal node parameters obtained by excluding the k^{th} tree. By computing the partial residuals

$$\mathbf{r}_k \equiv \mathbf{Y} - \sum_{t \neq k} g\left(\mathbf{X}; \mathcal{T}_t, \boldsymbol{Q}_t\right), \tag{2.4}$$

it is only necessary to make draws of $(\mathcal{T}_k, \boldsymbol{Q}_k)$ from $(\mathcal{T}_k, \boldsymbol{Q}_k) \mid \mathbf{r}_k, \tau$ followed by a draw of τ from $\tau \mid (\mathcal{T}_1, \boldsymbol{Q}_1), \ldots, (\mathcal{T}_K, \boldsymbol{Q}_K), \mathbf{Y}$. This implies that the trees can be fit iteratively, updating the structure of the k^{th} tree independently of all other $K-1$ trees.

To update each tree structure $\mathcal{T}_k$, a Metropolis-within-Gibbs sampler [Hastings, 1970, Geman and Geman, 1984] is employed. Given the current tree $\mathcal{T}_{t-1,k}$, a new tree $\mathcal{T}_{t,k}$ is proposed by applying one of the following moves:

- GROW: Splitting a terminal node into two child nodes.
- PRUNE: Removing a split, merging two child nodes back into a terminal node.
- CHANGE: Modifying an internal split rule while maintaining the tree structure.
- SWAP: Exchanging the positions of two split rules within the tree.

The proposed tree $\mathcal{T}_{t,k}$ is accepted with probability given by the Metropolis-Hastings acceptance ratio

$$\frac{p\left(\mathcal{T}_{t,k} \mid \mathbf{r}_k, \tau\right) q\left(\mathcal{T}_{t,k} \to \mathcal{T}_{t-1,k}\right)}{p\left(\mathcal{T}_{t-1,k} \mid \mathbf{r}_k, \tau\right) q\left(\mathcal{T}_{t-1,k} \to \mathcal{T}_{t,k}\right)}$$

where $p\left(\mathcal{T}_k \mid \mathbf{r}_k, \tau\right) = p(\mathcal{T}_k) \int p(\mathbf{r}_k \,|\, \mathcal{T}_k, \boldsymbol{Q}_k, \tau) p(\boldsymbol{Q}_k \,|\, \mathcal{T}_k) d\boldsymbol{Q}_k$. Further details on this and the computation of the transition probabilities can be found in Kapelner and Bleich [2016].

Next, independent draws of $\mu_{k\ell}$ can be obtained, where the posterior distributions follow a normal form due to the conjugate prior. Similarly, posterior draws of τ are sampled from a gamma distribution, also arising from conjugacy.

2.1.3 Probit BART

The original work by Chipman et al. [2010] also introduced the BART model for binary outcomes by incorporating the data augmentation scheme for Bayesian probit regression from Albert and Chib [1993], which we discuss in more detail in Chapter 3. The key idea is to introduce a latent continuous variable $\mathbf{Y}^*$, which is modelled as

$$Y_i^* = \sum_{k=1}^{K} g(\mathbf{X}_i; \mathcal{T}_k, \boldsymbol{Q}_k) + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1).$$

The observed binary outcome Y_i is then obtained via

$$Y_i = \begin{cases} 0 & \text{if } Y_i^\star \leq 0 \\ 1 & \text{if } Y_i^\star > 0. \end{cases}$$

This formulation mirrors the classical probit regression model, where the probability that $Y_i = 1$ follows a standard normal cumulative distribution function

$$Pr(Y_i = 1 \mid X_i, \boldsymbol{\mathcal{T}}, \boldsymbol{Q}) = \Phi\left(\sum_{k=1}^{K} g(\mathbf{X}_i; \mathcal{T}_k, \boldsymbol{Q}_k)\right)$$

where $(\boldsymbol{\mathcal{T}}, \boldsymbol{Q})$ is the collection of K trees and all corresponding terminal node parameters.

While the prior settings for $p(\mathcal{T}_k)$ remain unchanged, a few minor modifications allow posterior computation to closely resemble that of the continuous outcome version. In the standard BART regression model, Chipman et al. [2010] scaled the continuous response to the range $[-0.5, 0.5]$ and adjusted the prior mean and precision on $p(\mu_{k\ell} \mid \mathcal{T}_k)$ accordingly (Equation (2.2)). However, in the probit BART setting, the latent variables $\mathbf{Y}^*$ are assumed to mostly fall within the range $[-3, 3]$. Thus, while μ_μ remains fixed at 0, τ_μ can be calibrated using the same approach as in Equation (2.2), replacing 0.5 with 3.

Unlike in the standard continuous response BART model, the sampling of τ is no longer required as it is assumed to be known and fixed at 1. Next, the partial residuals for the k^{th} tree are calculated as $\mathbf{Y}^* - \sum_{t \neq k} g(\mathbf{X}; \mathcal{T}_t, \boldsymbol{Q}_t)$. Each latent variable is sampled from a truncated normal distribution:

$$Y_i^* \mid Y_i, \boldsymbol{\mathcal{T}}, \boldsymbol{Q} \sim \mathcal{TN}\left(\sum_{k=1}^{K} g(\mathbf{X}_i; \mathcal{T}_k, \boldsymbol{Q}_k), 1, \gamma_i\right),$$

where the truncation points γ_i are determined by Y_i such that $\gamma_i = [0, \infty)$ if $Y_i = 1$ and $\gamma_i = (-\infty, 0]$ if $Y_i = 0$.

2.2 Missing Data

Missing data is a fundamental challenge in statistical analysis and predictive modelling, occurring across diverse disciplines, including clinical research, economics, social sciences, and environmental studies. It can arise for various reasons, such as non-response in surveys, dropout in longitudinal studies, instrument failures in experimental research, or data entry errors in administrative records.

Traditionally, many statistical methods assume that missing data occurs completely at random or follows a predictable pattern that can be inferred from observed data. However, these assumptions are often unrealistic, and failing to account for the missingness mechanism itself can introduce bias and lead to misleading conclusions.

To formally define missing data mechanisms and the challenges they present, particularly in a regression context, we introduce the following notation. Let $\tilde{\mathbf{Y}}$ denote the partially observed response variable(s), with missing values occurring in some entries. The missing data indicator is represented by $\mathbf{M}$ and equals 1 if the response is observed and 0 if the response is missing. Additionally, let $\mathbf{X}$ be a set covariates used to predict $\tilde{\mathbf{Y}}$. We assume here that $\mathbf{X}$ is fully observed. Finally, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ represent model parameters for $\tilde{\mathbf{Y}}$ and $\mathbf{M}$, respectively.

2.2.1 Missing Data Mechanisms

The concept of missing data mechanisms was originally introduced by Rubin [1976], providing a formal framework for understanding the relationship between missing data patterns and measured variables. This framework is essential in determining the validity of statistical analyses and guiding the selection of appropriate missing data handling techniques. Incorrect assumptions about the missingness mechanism can lead to biased estimates and misleading inferences [Van Buuren, 2018].

Rubin's classification distinguishes between three fundamental types of missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data are said to be missing completely at random (MCAR) if the probability of missingness is entirely independent of both observed and unobserved data. This can be expressed as

$$P(\mathbf{M} \mid \mathbf{X}, \tilde{\mathbf{Y}}, \boldsymbol{\psi}) = P(\mathbf{M} \mid \boldsymbol{\psi}).$$

Under MCAR, missingness occurs purely by chance, and the observed data can be con-

sidered a random subsample of the full dataset [Baraldi and Enders, 2010]. A common example of MCAR is when missingness results from random technical failures, for instance, if a weighing scale runs out of battery, causing measurements to be missing [Van Buuren, 2018]. If data are MCAR, the probability of data being missing is uniform across all cases, signifying that missingness is entirely independent of the observed data. While this assumption is conceptually straightforward, it often proves to be overly simplistic and unrealistic in real-world scenarios.

Under the missing at random (MAR) mechanism, the probability of missingness depends only on observed variables but not on the missing values themselves:

$$P(\mathbf{M} \mid \mathbf{X}, \tilde{\mathbf{Y}}) = P(\mathbf{M} \mid \mathbf{X}, \boldsymbol{\psi}).$$

This assumption allows missingness to be explained by fully observed covariates, making MAR less restrictive than MCAR and more realistic in many practical settings. An example of MAR is found in survey research. Consider a study where participants are asked their age and how often they smoke cigarettes. As teenagers may be more likely to withhold their smoking habits due to fear of repercussions, missingness in the smoking variable is explained by the observed age variable. Since missingness is not directly related to the smoking variable itself, this scenario satisfies the MAR assumption [Marcelino et al., 2022]. MAR is a more realistic representation of many practical situations and serves as the underlying assumption for the majority of methods designed to address missing data. Acknowledging the potential relationship between missingness and the observed data, MAR allows for more nuanced strategies in handling and imputing missing values.

Finally, data are missing not at random (MNAR) when the probability of missingness depends on the missing values themselves, even after conditioning on observed data. In other words, missingness is directly related to unobserved information, and ignoring the missingness mechanism can lead to severe bias [Little and Rubin, 2019]. A classic example of MNAR occurs in medical research, where individuals may decline to take an HIV test due to concerns about potential stigma, particularly among those who suspect they might test positive [Marra et al., 2017].

2.2.2 Existing Methods for Handling Missing Data

A variety of methods exist for handling missing data, with the choice of approach largely depending on the assumed missingness mechanism. Some key resources that provide comprehensive overviews of these methods include Baraldi and Enders [2010], Van Buuren [2018], Little and Rubin [2019], and Marcelino et al. [2022].

Before selecting an appropriate missing data method, it is often useful to identify and explore missing data patterns. R packages such as `naniar` [Tierney and Cook, 2023]

and `VIM` [Kowarik and Templ, 2016] offer comprehensive visualisation tools, including histograms, heatmaps, and shadow matrices (which create an auxiliary binary representation of the dataset indicating observed and missing values). `mice` [Van Buuren and Groothuis-Oudshoorn, 2011] and `mi` [Gelman and Hill, 2011] provide functions for summarising missingness distributions and assessing imputation diagnostics, while `ggmice` [Oberman et al., 2022] enhances `mice` by generating visualisations tailored to its workflows.

The MCAR assumption, though rare in practice, allows for relatively simple methods such as complete-case analysis, where only observations with no missing values are used in the analysis. While straightforward, this approach often leads to a loss of efficiency due to reduced sample size, a problem that is particularly pronounced in multivariate settings, where differing missingness patterns across variables can result in the removal of many observations even when some variables remain observed. An alternative is available-case analysis, which utilises all available data for each variable, although this can result in inconsistencies if different sample sizes are used across analyses. Another approach under the MCAR assumption is mean imputation, where missing values are replaced with the mean of the observed values for that variable, though this method can distort variance and underestimate uncertainty. More sophisticated approaches include hot-deck imputation, in which missing values are replaced by observed values from similar units in the dataset, maintaining some level of variability in the imputed data.

When the MCAR assumption is violated but MAR is reasonable, the expectation-maximisation [EM; Dempster et al., 1977] algorithm is a well-known classical approach for handling MAR data, using an iterative process to compute maximum-likelihood estimates. More recently, imputation strategies such as model-based imputation, regression imputation, multiple imputation [Van Buuren, 2018], and multivariate imputation by chained equations [MICE; Azur et al., 2011, Van Buuren and Groothuis-Oudshoorn, 2011, Little and Rubin, 2019] are commonly used, leveraging available data to make educated guesses about the missing values. A variety of single and multiple imputation methods are available in R. For example, the `mice` package implements MICE, `simputation` [van der Loo, 2022] provides a flexible interface for various model-based and regression-based imputation techniques, and `VIM` includes several classical imputation methods. Another notable method is `missForest` [Stekhoven and Bühlmann, 2012], which is a non-parametric missing value imputation technique using random forests [Breiman, 2001] to impute datasets containing variables of mixed type.

As well as imputation strategies, advanced modelling techniques offer alternative approaches for handling missing data without explicit imputation. CART models [Breiman et al., 1984] address missing values through surrogate splits, where a correlated variable is used as a substitute when a splitting variable is missing. In Tierney et al. [2015],

boosted regression trees [Elith et al., 2008] and CART models were used to explore missingness patterns, identifying influential variables in the missingness model to understand the missingness structure.

Additionally, methods such as `BARTm` [Kapelner and Bleich, 2015] and `XGBoost` [Chen and Guestrin, 2016] can accommodate missing values—albeit only in the covariates, and only with univariate responses—without requiring pre-processing steps to fill in missing data. During model training, `BARTm` utilises available cases while incorporating missing covariates directly into the tree-splitting rules. Similarly, `XGBoost` leverages available data for model training and treats missing covariates as a distinct category during split decisions within its boosting framework.

While methods designed for MAR have been effective in handling MCAR data, there is little evidence to suggest that they can adequately address MNAR mechanisms. Although multiple imputation can still be applied in MNAR settings (see Galimard et al. [2016, 2018]), its effectiveness depends on correctly specifying the missingness mechanism, though this is rarely undertaken in practice [Tierney et al., 2015]. When the missing data mechanism is misrepresented, standard imputation approaches may introduce bias rather than mitigate it. Handling MNAR data requires careful consideration and often involves more complex modelling techniques, as the probability of missingness depends directly on unobserved values. Unlike MCAR and MAR, where missingness is either random or conditionally independent of missing values given observed data, MNAR introduces a systematic relationship between missingness and the unobserved information. Consequently, failing to account for the missingness mechanism can result in biased estimates and misleading conclusions.

When obtaining additional data to reduce the extent of MNAR missingness is infeasible, model-based approaches that explicitly account for the missingness process become necessary. Two of the most widely used frameworks for addressing MNAR are the pattern-mixture model [Glynn et al., 1986, Little, 1993] and the selection model [Heckman, 1976], as outlined in Little and Rubin [2019]. These methods differ in how they factorise the joint distribution of the partially observed responses and their missingness indicators, denoted as $p(\tilde{\mathbf{Y}}, \mathbf{M} \mid \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\psi})$. A comprehensive review of methods for handling MNAR-type data is provided in Tang and Ju [2018].

In the pattern-mixture model, the joint distribution is factorised into

$$p(\tilde{\mathbf{Y}}, \mathbf{M} \mid \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\psi}) = p(\tilde{\mathbf{Y}} \mid \mathbf{M}, \mathbf{X}, \boldsymbol{\theta}) p(\mathbf{M} \mid \mathbf{X}, \boldsymbol{\psi}).$$

The data are first stratified based on their missingness patterns, and then the response is modelled based on the specific missingness patterns. This method is advantageous for investigating differences in the response distributions across distinct missing data pat-

terns, and may be insightful for exploring the sensitivity of inferences to different missing mechanism assumptions [Michiels et al., 1999]. However, pattern-mixture models often encounter under-identifiability issues, necessitating the imposition of model restrictions or the incorporation of prior information [Little, 1993, Thijs et al., 2002, Little, 2008].

In contrast, the selection model factorises the joint distribution into

$$p(\tilde{\mathbf{Y}}, \mathbf{M} \mid \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\psi}) = p(\tilde{\mathbf{Y}} \mid \mathbf{X}, \boldsymbol{\theta}) p(\mathbf{M} \mid \mathbf{X}, \tilde{\mathbf{Y}}, \boldsymbol{\psi}),$$

formulating the response model separately from the missingness mechanism, followed by modelling the missingness probabilities given the response. This approach directly estimates parameters related to the full population of interest and reflects the natural order of events where the response occurs before missingness is introduced [Little, 2008].

Originally introduced by Heckman [1976], the selection model framework has been widely used in various applications. A notable example is the 'Heckit' model, which specifies a parametric linear regression for modelling a univariate response while the missingness indicator is modelled using probit regression. This can be formulated as

$$\begin{aligned} \tilde{Y}_i \mid \mathbf{X}, \boldsymbol{\theta} &= \mathbf{b}^\top \mathbf{X}_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \tau^{-1}) \\ Pr(M_i = 1 \mid \mathbf{X}, \tilde{Y}_i, \boldsymbol{\psi}) &= \Phi\left(\delta_0 + \boldsymbol{\delta}_X^\top \mathbf{X}_i + \delta_Y \tilde{Y}_i\right), \end{aligned} \tag{2.5}$$

where $\mathbf{b}$ represents the vector of parameters in the linear data model, τ is the residual precision, and $(\delta_0, \boldsymbol{\delta}_X, \delta_Y)$ are probit regression coefficient parameters for the intercept, the covariates and the response, respectively.

Equation (2.5) represents the probability that the i^{th} value of the response is observed, i.e. $M_i = 1$. When $\boldsymbol{\delta}_X = \mathbf{0}$ and $\delta_Y = 0$, the missingness probability only depends on δ_0 and not on the observed or missing data, characterising an MCAR scenario. When $\delta_Y = 0$, the probability of the response being missing does not depend on its own value but may still be related to the observed covariates, thus representing a MAR scenario. However, if $\delta_Y \neq 0$, the missingness of the i^{th} value of the response depends on its own (potentially missing) value, leading to an MNAR mechanism.

The selection model is not limited to univariate responses, nor is it limited to these parametric forms. In principle, any appropriate model can be employed for both the response and missingness mechanism. Next, while the Heckit model was originally implemented using a 'two-step' process, a joint modelling approach which simultaneously estimates the response and missingness has been shown to be preferable in many cases [Puhani, 2000, Bushway et al., 2007, Galimard et al., 2018].

Furthermore, as demonstrated earlier, the selection model explicitly incorporates the missingness mechanism as a function of both the observed and missing data. In likelihood-

based inference, when the missing data mechanism is MCAR or MAR, the missingness process is ignorable, meaning that the missingness model does not need to be explicitly modelled since it is independent of the partially observed responses. Conversely, when the data are MNAR, the missingness mechanism is non-ignorable, as inferences about the missing values depend directly on the missingness model and failing to account for this dependence leads to biased results.

However, while assuming ignorable missingness simplifies inference, recent work by Linero [2024] highlights challenges in fully ignoring the missingness process, particularly in high-dimensional and non-parametric Bayesian models. This reinforces the need for joint modelling techniques that account for selection bias while avoiding restrictive assumptions about ignorability.

2.3 *global Amax* Data

The *global Amax* dataset, originating from Maire et al. [2015], is a comprehensive collection of data that links plant traits, soil properties, and climate variables from multiple sources to analyse their effects on leaf photosynthetic traits and rates. It builds upon the 'Glopnet' dataset from Wright et al. [2004] and expands it by integrating additional datasets to provide a global perspective on how environmental factors shape photosynthetic capacity across diverse ecosystems.

Following pre-processing, the final dataset analysed by Maire et al. [2015] consists of 46 fully observed covariates and five response variables, which exhibit high levels of missingness with varying missingness patterns. Notably, fewer than 10% of observations are complete cases, making missing data a significant challenge. Despite this, the original study did not provide much details on how missing data were handled, apart from briefly mentioning the exclusion of a response variable, suggesting a reliance on complete-case analysis.

As discussed previously, even under the MCAR assumption, complete-case analysis leads to substantial information loss, particularly in this case, where over 90% of the data may be discarded due to overlapping missingness patterns. Moreover, the response variables are likely to be correlated, and non-linear relationships may exist, thus highlighting the need for robust joint modelling approaches that can account for multivariate dependencies, capture non-linearity, and handle missing data flexibly without imposing strong, unverifiable assumptions about the missingness mechanism. Here, we describe the compositions and characteristics of the *global Amax* dataset based on information from Maire et al. [2015], summarise key findings from the original analysis, and finally examine the missingness patterns in the response variables, highlighting the challenges posed by their complexity.

2.3.1 Dataset Composition and Characteristics

The dataset consists of 2400 species-site combinations, covering 288 sampled sites and 1509 species from 165 plant families. It includes a wide range of plant types, such as trees, shrubs, herbs, grasses, ferns, and vines, with varying physiological and phenological traits. The dataset compiles observations from numerous studies, harmonising measurement protocols to ensure comparability. However, given the scale of integration, some standardisation challenges remain, particularly in reconciling differences in sampling methods across sources.

Three main components define the dataset: trait data, soil data, and climate data. The trait data primarily come from Glopnet, supplemented by additional georeferenced observations. The key measured traits include photosynthetic capacity (*Aarea*), stomatal conductance (*Gs*), leaf nitrogen and phosphorus content (*Narea* and *Parea*), and specific leaf area (*SLA*). These traits provide a mechanistic link between photosynthetic efficiency, nutrient investment, and environmental adaptation in plants. The soil data include 20 soil variables related to soil structure, texture, ion exchange capacity, and macronutrient content. These were extracted from SoilGrids [ISRIC, 2013], the Harmonized World Soil Database [FAO et al., 2012], and the ISRIC-WISE dataset [Batjes, 2012], covering parameters such as organic matter content, pH, cation exchange capacity, nitrogen content, and available phosphorus. The climate data comprise 26 variables, including temperature, precipitation, sunshine duration, relative humidity, and aridity, primarily derived from the Climatic Research Unit (CRU) dataset [New et al., 2002]. Additionally, various bioclimatic indices such as global radiation, total annual incident radiation, and equilibrium evapotranspiration were calculated. Aridity was quantified using the soil moisture index, which represents the ratio between precipitation and potential evapotranspiration.

Despite its extensive coverage, the dataset has some limitations. Certain ecosystems, such as high-latitude boreal forests, extreme deserts, and alpine regions, are underrepresented, potentially limiting its applicability in those regions. Furthermore, while soil data were primarily derived from interpolated global datasets, rather than direct measurements at every site, certain soil variables (e.g., phosphorus availability) required conversion factors to harmonise extraction methods, potentially introducing minor errors. Similarly, climate data for many sites were interpolated from CRU gridded datasets, which may not fully capture fine-scale microclimatic variation. These limitations underscore the need for expanded field measurements and enhanced spatial resolution in future research.

2.3.2 Previous Analysis and Findings

In Maire et al. [2015], the final *global Amax* dataset was analysed by treating variables in the leaf trait dataset, i.e. *SLA*, *Aarea*, *Narea*, *Parea*, and *Gs*, as response variables, while

soil and climate properties served as model predictors. The responses and covariates are listed along with their descriptions and measurement units in Tables 2.1 and 2.2 respectively. A log-transformation was applied to the responses to account for right-skewness. The analysis aimed to disentangle the relative contributions of soil and climate factors to plant photosynthetic traits.

Responses	**Description**	**Unit**
SLA	Specific leaf area	cm^2 g^{-1}
Aarea	Light-saturated photosynthetic carbon assimilation per unit leaf area	$\mu mol\ m^{-2}\ s^{-1}$
Narea	Leaf nitrogen content per unit leaf area	gN m^{-2}
Parea	Leaf phosphorus content per unit leaf area	gP m^{-2}
Gs	Stomatal conductance to water vapour	mmol m^{-2} s^{-1}

Table 2.1: List of response variables included in the *global Amax* dataset along with descriptions and measurement units.

Covariates	**Description**	**Unit**
ALU	Exchangeable aluminium percentage	% of ECEC
AWHC	Available water holding capacity (-33 to -1500 kPa; USDA standard)	mm m^{-1}
BULK	Bulk density	kg dm^{-3}
CARB	Calcium carbonate content	g kg^{-1}
CECC	Cation exchange capacity of clay size fraction, corrected from contribution of organic matter	$cmol^+$ kg^{-1}
CECS	Cation exchange capacity	cmolc kg^{-1}
CLAY	Clay content	%wt
CN	CN ratio	gC gN^{-1}
Corg	Organic carbon content	gC kg^{-1}
DEPTH	Depth to the parent rock	cm
GRAVEL	Gravel content	%wt
MIF	Moisture index (MIF = PPTmean / PETF)	mm mm^{-1}
MIQ	Moisture index (MIQ = PPTmean / PETQ)	mm mm^{-1}
Ntot	Total nitrogen content	gN kg^{-1}
PAR	Cumulative photosynthetically active radiation with daily temperature above 0 °C (PAR0) or 5°C (PAR5)	W m^{-2}

Pavail	Available soil phosphate content	mgP_2O_5 kg^{-1}
PETF	Potential evapotranspiration (Penman Monteith equation)	mm $month^{-1}$
PETQ	Equilibrium evapotranspiration (Prentice equation)	mm $month^{-1}$
pH	Soil pH measured in H_2O solution	0-14
PPTcv	Coefficient of variation of monthly precipitation	mm
PPTmax	Maximum monthly precipitation	mm
PPTmean	Mean annual precipitation	mm
PPTmin	Minimum monthly precipitation	mm
PPTseason	Seasonality of precipitation	0-1
RAD	Global radiation	W m^{-2}
RH	Relative humidity	%
SALT	Salinity measured by the electrical conductivity of the soil	dS m^{-1}
SAND	Sand content	%wt
SBA	Base saturation as percentage of CECS	%
SILT	Silt content	%wt
SODIUM	Sodicity measured by the exchangeable sodium percentage	% of ECEC
SUNmax	Maximum monthly fractional sunshine duration	%
SUNmean	Mean annual fractional sunshine duration	%
SUNmin	Mean monthly fractional sunshine duration	%
SUNrange	Range of monthly fractional sunshine duration	%
TBA	Total exchangeable bases	cmol kg^{-1}
TMPgs	Cumulative daily temperature above 0°C or 5°C	°C
TMPiso	Isothermality	-
TMPmax	Maximal monthly temperature	°C
TMPmean	Mean annual temperature	°C
TMPmin	Minimal monthly temperature	°C
TMPnb	Number of days with daily temperature above 0°C or 5°C	#
TMPrange	Mean diurnal temperature range	°C

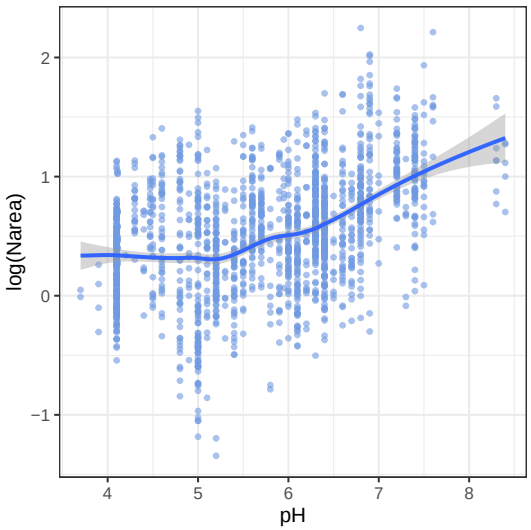
Table 2.2: List of covariates in *global Amax* along with descriptions and measurement units.

The original analysis began with defining key dimensions of soil fertility and quantifying their relationships with leaf traits using mixed regression models. Both quadratic and linear models were fitted, with site and species included as random intercepts to account for the hierarchical structure of the data. Next, stepwise multiple mixed regression models were applied to identify the most influential soil and climate variables for each trait. From an initial set of 26 climate and 20 soil variables, up to four key predictors were selected based on the Akaike information criterion to optimise model simplicity while maintaining predictive accuracy.

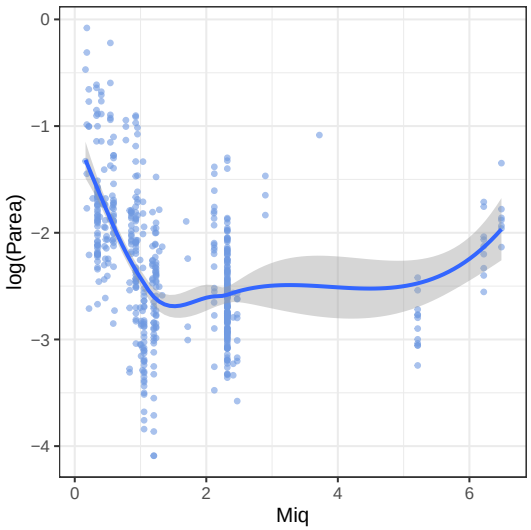
To distinguish unique and joint effects of soil and climate, variation partitioning and Venn diagrams were used. This method decomposed trait variation into components uniquely explained by soil, uniquely explained by climate, or jointly influenced by both. The unique effect of each factor was quantified by comparing adjusted r^2 values between full and partial models. Additionally, redundancy analysis was performed to assess how well the matrix of leaf traits could be explained by soil and climate variables. However, due to its smaller sample size, *Parea* was excluded from this step. Finally, path analysis was conducted to disentangle the direct and indirect effects of soil, climate, and leaf traits on photosynthetic capacity. This technique modelled causal relationships among *Aarea*, *Gs*, *Narea*, *SLA*, and key environmental drivers, helping to clarify the underlying mechanisms governing trait–environment interactions.

The analysis yielded several key findings. Apart from *SLA*, soil properties were found to have more significant influence on leaf traits as compared to climate properties. Notably, *Aarea*, *Narea*, and *Parea* increased with *pH* and decreased with increasing *Miq*, whereas *Gs* declined with increasing *Pavail*. The study revealed that joint effects of soil and climate were dominant for *Narea* and *Parea*, while soil alone was the primary driver of *Aarea* and *Gs*. Three key environmental variables emerged as the most influential for photosynthetic rates: *pH*, *Pavail*, and *Miq*. These insights contribute significantly to understanding how plant photosynthesis varies globally in response to soil fertility and climate constraints.

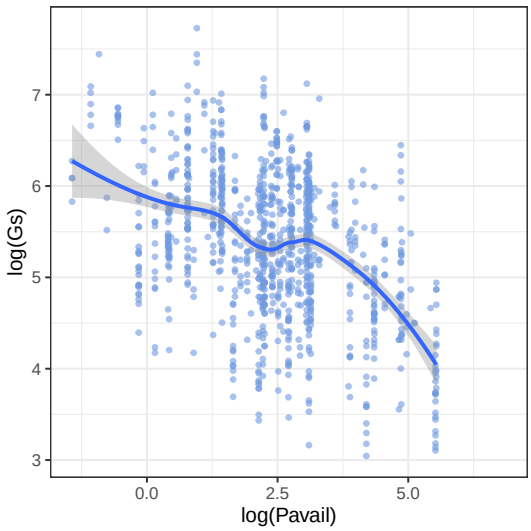
The scatterplots and LOESS curves in Figure 2.1 show the key relationships identified in Maire et al. [2015] between the covariates and some of the log-transformed observed responses, specifically *pH* and log(*Narea*) in Figure 2.1a, *Miq* and log(*Parea*) in Figure 2.1b, and the log-transformed *Pavail* and log(*Narea*) in Figure 2.1c. Note that the log transformation of *Pavail* is applied here for visualisation purposes only, but is not necessary in our analyses. While *Narea* appears to increase with *pH* and *Gs* decreases as *Pavail* increases, the LOESS curves suggest a degree of non-linearity in these relationships. Additionally, despite the limited number of observed *Parea* values, a non-linear trend emerges as *Miq* increases in Figure 2.1b. These observations suggest that the analysis of the *global Amax* dataset may benefit from models capable of capturing non-linear relationships.



(a) Scatterplot and LOESS curve of the observed log(*Narea*) values against *pH* showing a potential non-linear trend.



(b) Scatterplot and LOESS curve of the observed log(*Parea*) values against *Miq* showing a potential non-linear trend.



(c) Scatterplot and LOESS curve of the observed log(*Gs*) values against log(*Pavail*), showing a non-linear decreasing trend.

Figure 2.1: Scatterplots and LOESS curves showing the relationships between observed values of some log-transformed response variables and key variables identified in Maire et al. [2015]. Potential non-linear trends can be observed, especially in the relationship between *Parea* and *Miq*.

2.3.3 Missing Data in *global Amax*

While the covariates in *global Amax* were completely observed, the response variables exhibit substantial levels of missingness, with only 217 complete cases among 2368 observations ($\approx 9.16\%$ completeness). The missing data patterns vary considerably across traits, as illustrated in Figure 2.2. Among the five response variables, *Aarea* is mostly

complete, with only 12 missing cases ($\approx 0.5\%$), while *SLA* and *Narea* show moderate missingness levels, with 433 ($\approx 18.29\%$) and 652 ($\approx 27.53\%$) missing cases, respectively. In contrast, *Gs* and *Parea* exhibit the highest missingness, with 1353 ($\approx 57.14\%$) and 1836 ($\approx 77.5\%$) missing cases, respectively.

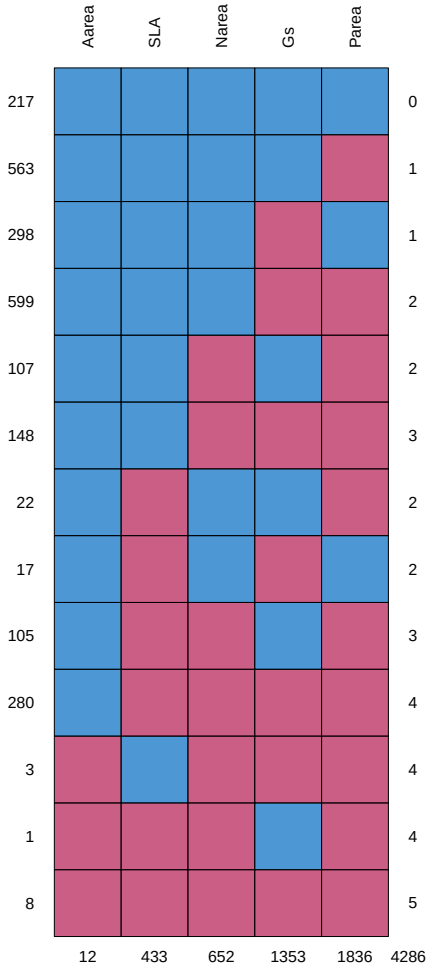


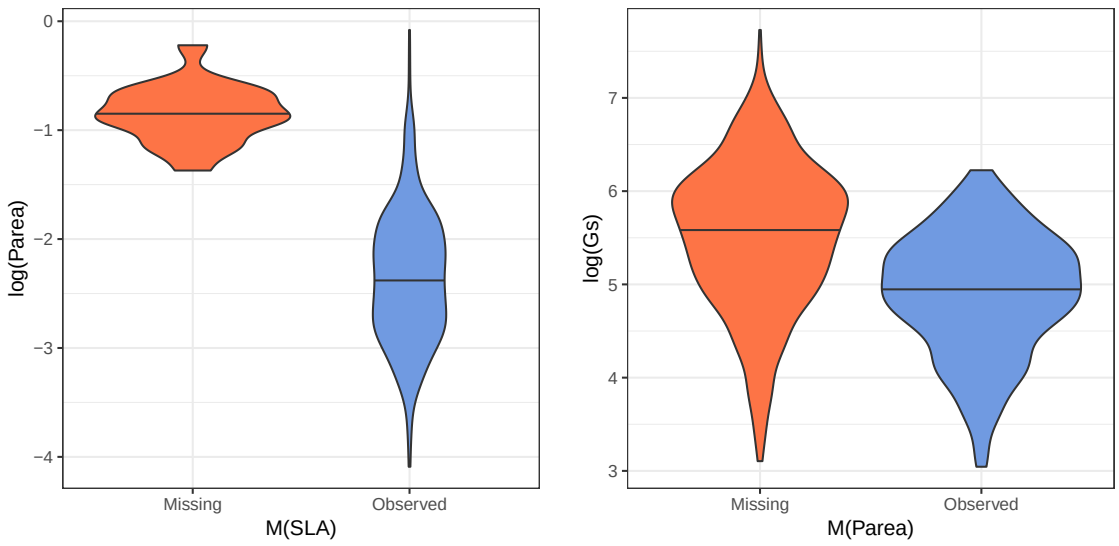
Figure 2.2: Missingness patterns for the *global Amax* response variables with blue boxes indicating 'observed' and red boxes indicating 'missing'. Missingness is present in the 5 response variables, while the 46 covariates are fully observed. Column labels at the top show the response variables. Row labels on the left show the number of cases exhibiting the unique missingness patterns. Row labels on the right indicate the total number of missing variables within that pattern. Column labels on the bottom depict the number of cases where each variable is missing.

Despite the significant proportion of missing responses, Maire et al. [2015] provided limited details on how missingness was handled in their analyses. The redundancy analysis performed in their study briefly noted the exclusion of *Parea* due to its small sample size. This suggests that their approach primarily relied on complete-case analysis. However, even with *Parea* excluded, the number of complete cases would still amount to 780 observations, which is only 33% of the total dataset.

The high degree of missingness in *Parea* as well as *Gs* is likely a consequence of the compilation process of the *global Amax* dataset. Since *Aarea* is a key photosynthetic trait, it is often prioritised in the data collection process, leading to its relatively low missingness. Traits such as *SLA* and *Narea* are frequently quantified in conjunction with *Aarea*, resulting in their relatively low missingness. In contrast, *Parea* and *Gs* may be less frequently measured due to their more complex or less standardised measurement protocols, contributing to their higher missingness levels.

Since missingness in the responses is likely influenced by data collection priorities and measurement complexities, the MCAR assumption is unrealistic. While assuming MAR may be reasonable, it requires that missingness in each response variable is solely explained by the observed covariates and that missingness in one response is independent of the values of other responses. This assumption is appealing, as it allows for the use of numerous MAR-based methods, but if the true missingness mechanism is MNAR, such methods can lead to biased estimates and misleading conclusions.

Figure 2.3 presents violin plots illustrating how the missingness of one response variable relates to the observed values of another. These relationships highlight the impacts of partial missingness, whereby some responses are observed when others are not, and behave differently depending on the missingness of other responses. In Figure 2.3b, when *SLA* is missing, *Parea*—the response with the highest missingness—tends to be observed at higher values. Similarly, in Figure 2.3a, *Gs* tends to be lower when *Parea* is observed, whereas higher values of *Gs* are more frequently observed when *Parea* is missing.



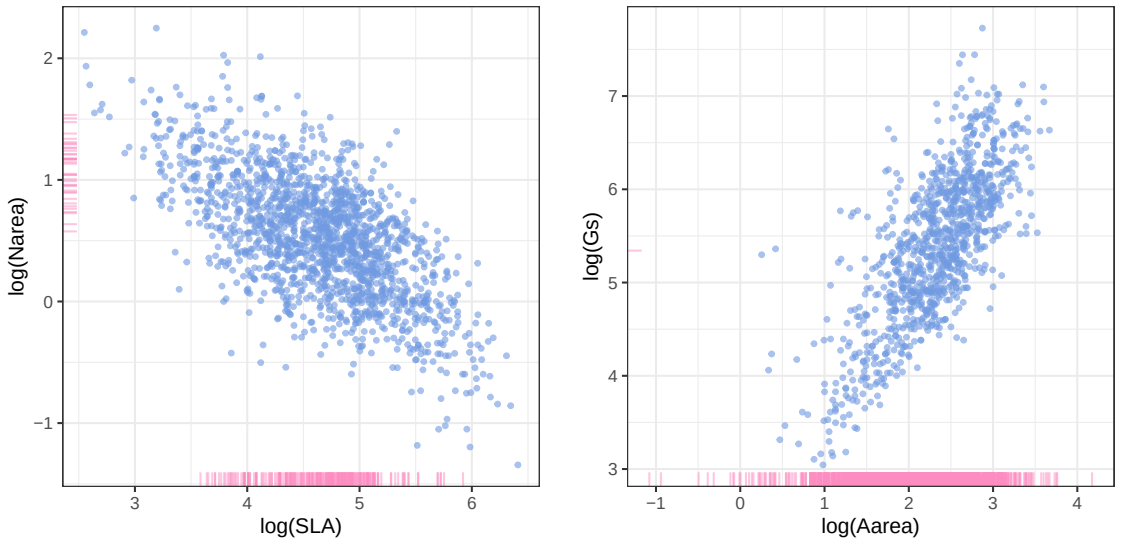
(a) When *SLA* is missing, *Parea*—the response with the highest missingness—tends to be observed at higher values.

(b) *Gs* tends to be lower when *Parea* is observed, whereas higher values of *Gs* are more frequently observed when *Parea* is missing.

Figure 2.3: Violin plots showing the observed values of responses against the missingness others.

This is further illustrated in Figure 2.4, which presents scatterplots of the overlapping observed values for the log-transformed *SLA*–*Narea* and *Aarea*–*Gs* pairs. Rug plots are included, depicting values that are present but excluded from the scatterplot due to missingness in the opposing response variable. In Figure 2.4a, roughly 70.8% of the observed cases between *SLA* and *Narea* overlap. When *SLA* is missing, higher values of *Narea* are still present but are not included in the scatterplot. In Figure 2.4b, only a single *Gs* value was removed from the scatterplot due to missing *Aarea*. However, because *Gs* has high levels of missingness, the resulting scatterplot contains significantly fewer observed values of *Aarea*. Consequently, a substantial portion of *Aarea* values are missing across the entire range, including lower values that fall below those displayed in the scatterplot (i.e., $\log(Aarea) < 0$).

Furthermore, Figure 2.4 also shows strong correlations between response pairs. *SLA* and *Narea* appear to be negatively correlated, while *Aarea* and *Gs* exhibit a strong positive correlation, suggesting that modelling responses independently could overlook important dependencies, potentially leading to a loss of information. A joint modelling approach that simultaneously accounts for multivariate responses and missingness can better capture these correlations while remaining flexible to different missingness mechanisms. By jointly modelling both the responses and the missingness process, such models can improve inference, enhance imputation accuracy, and potentially uncover additional relationships, providing insights that would be lost if missingness were ignored.



(a) Scatterplot of *Narea* against *SLA*. 1677 points (≈ 70.8%) are shown here. These responses appear to be negatively correlated.

(b) Scatterplot of *Gs* against *Aarea*. 1014 points (≈ 42.8%) are shown here. These responses appear to have strong positive correlation.

Figure 2.4: Scatterplots of observed pairs of log-transformed responses showing strong positive and negative correlations. Rug plots depict values that are available but excluded from the scatterplot due to missingness in the opposing response variable.

2.4 Discussion

This chapter provided a foundational background on Bayesian additive regression trees (BART), missing data mechanisms, and the *global Amax* dataset, all of which are central to the development of the methods introduced in later chapters. We first reviewed and characterised BART as a flexible, non-parametric Bayesian approach for regression and binary classification, highlighting its ability to model complex relationships without requiring explicit functional form specifications. We also discussed key prior settings, the Bayesian backfitting MCMC algorithm, and extensions such as probit BART for binary outcomes.

Following this, we explored fundamental concepts of missing data, categorising them into MCAR, MAR, and MNAR mechanisms, and examining their impact on statistical inference. A review of existing missing data handling techniques was presented, noting their strengths and limitations, particularly in cases where missingness may be non-ignorable (MNAR). Particular attention was paid to the selection model framework, on which our methods are based. Finally, the *global Amax* dataset was introduced as a motivating example, illustrating the challenges posed by high levels of missingness in the response variables and underscoring the need for advanced modelling techniques beyond standard complete-case analysis or imputation methods.

By integrating BART with the selection model framework for non-ignorable missing data, we aim to develop joint models capable of handling multivariate responses, capturing complex missingness structures, and improving inference without making strict assumptions about the missingness mechanism. The next chapters build upon this foundation by introducing our proposed joint models, detailing their formulations, and exploring their implementations.

3

Joint Models for Handling Non-Ignorable Missing Data using Bayesian Additive Regression Trees

3.1 Introduction

In this chapter, we focus on a novel selection model framework — outlined later in Equation (3.1) — to address the challenges posed by MNAR missing data in the context of predictive data analysis with multivariate outcomes. More specifically, we present two multivariate response predictive models arising from the selection model factorisation to analyse the *global Amax* data without restrictive assumptions on the missing data mechanism. In both approaches, which we refer to as 'missBART1' and 'missBART2', we specify a multivariate Bayesian additive regression trees [BART; Chipman et al., 2010, Um et al., 2023, McJames et al., 2024] model for the partially-observed responses. The key distinction between the two models lies in their approaches to modelling the missing data mechanisms.

In missBART1, we propose a multivariate probit regression model for the missingness. Probit regression was originally introduced in Bliss [1934]; a full Bayesian model was later described in Albert and Chib [1993] and the multivariate extension was developed in Chib and Greenberg [1998]. One benefit of the probit regression model is its parametric nature, which enables the characterisation of different missing data mechanisms based on interpretable model parameters. Through prior specifications within the probit regression model, we introduce additional flexibility by enabling the incorporation of prior beliefs regarding the underlying missing data mechanism, as well as allowing for efficient Gibbs sampling of the missing responses within the Bayesian framework.

In the probit regression model, the underlying latent structure is inherently linear, making it less suitable in cases where missingness depends on other variables in a non-linear fashion. To address this limitation, missBART2 adopts an alternative approach to modelling missingness via the specification of a multivariate probit BART model. This fully non-parametric joint model leverages BART's ability to capture complex, non-linear

relationships within both the data and missingness sub-models. Although missBART2 lacks the interpretable coefficients provided by the probit regression model in missBART1 and also requires a Metropolis-Hastings step to sample missing responses, BART's variable selection feature mitigates the need for making prior assumptions about the missing data mechanism. This is particularly advantageous when limited information on the missingness mechanism is available.

Figure 3.1 shows a schematic diagram of both joint models. While both models are designed to handle missingness in the responses, they can also accommodate missing covariates, with the constraint that covariates are missing under the ignorability assumption. Due to the parametric nature of the probit regression function, missBART1 requires prior imputation on the covariates. For missBART2, the `BARTm` approach for handling missing covariates can be incorporated, thereby obviating the need for covariate imputation.

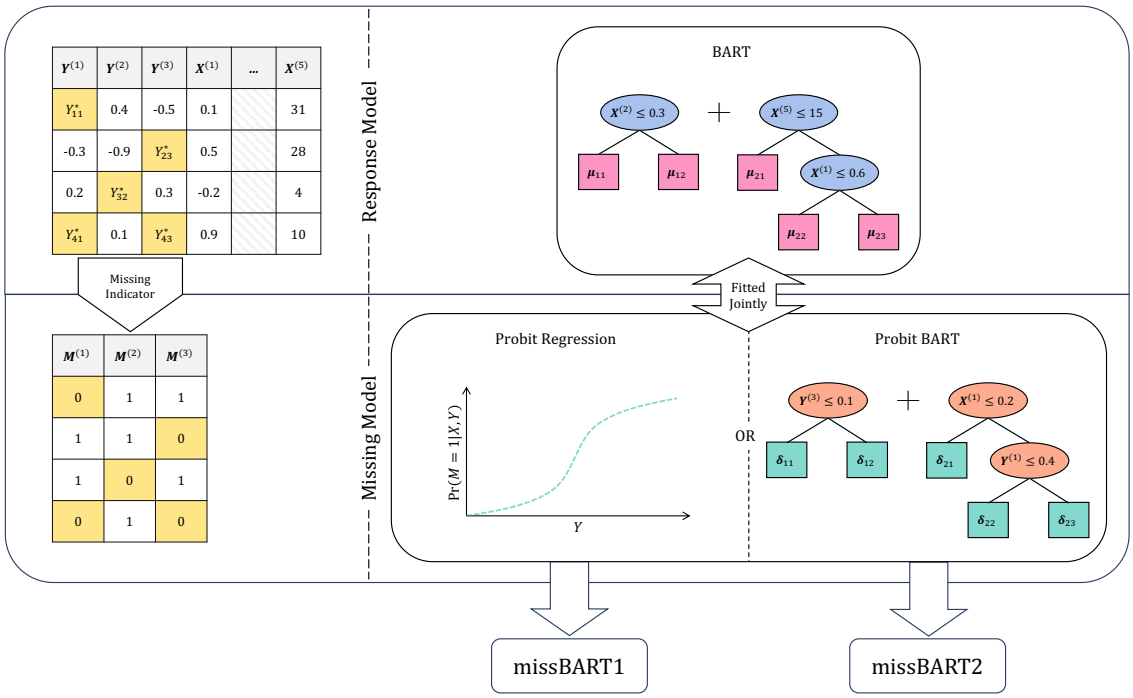


Figure 3.1: Schematic diagram of missBART1 and missBART2 using a toy dataset with three response variables $\left(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(3)}\right)$ and five covariates $\left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(5)}\right)$. The responses have missing entries, denoted by Y_{ij}^{*} with row index i and column index j. Each $\mathbf{M}^{(j)}$ is the resulting missing data indicator for $\mathbf{Y}^{(j)}$, where $M_{ij} = 0$ if Y_{ij} is missing and *vice versa*. Both joint models fit a BART model to the responses, using $\mathbf{X}$ in the splitting rules of the trees. In missBART1, a probit regression model is jointly fitted to model the missing data indicators $\mathbf{M}$, where $\mathbf{X}$ and $\mathbf{Y}$ are used as the missing model covariates. In missBART2, a probit BART model is fitted to $\mathbf{M}$, with $\mathbf{X}$ and $\mathbf{Y}$ used in the splitting rules of the trees. While the goal is to account for unobserved responses, both models are also capable of handling covariate missingness, either via prior covariate imputation (missBART1) or incorporating missingness into the splitting rules of the trees (missBART2).

The remainder of this chapter is structured as follows: Section 3.2 outlines the selection model in the context of handling multivariate missing response data. Section 3.3 describes the probit regression model from a univariate and multivariate standpoint within the Bayesian framework. Section 3.4 outlines the multivariate BART model and its probit counterpart, giving mathematical formulations and prior specifications. Section 3.5 explains our two novel models 'missBART1' and 'missBART2' in detail, formulating the full conditional distributions for posterior sampling along with their sampling algorithms. Following this, in Chapter 4, we present the results from several simulation studies, results from applying 'missBART1' and 'missBART2' to the *global Amax* data, along with a discussion on future work.

3.2 Selection Models for Non-Ignorable Missing Data

We now give a brief outline of the selection model used for modelling partially-observed data such as the *global Amax* data without restrictive ignorability assumptions. We restrict the missingness to the multivariate response variables and assume that all covariates are fully observed, as is the case with the *global Amax* data. However, scenarios with ignorable covariate missingness are discussed in Sections 3.5 and 4.2.

Given a dataset with partially-observed responses $\mathbf{Y}$ and a fully observed set of covariates $\mathbf{X}$, let $\mathbf{M}_i$ ($i = 1, \ldots, n$ observations) be the p-dimensional vector of missing data indicators such that for each $j = 1, \ldots, p$,

$$M_{ij} = \begin{cases} 0 & \text{if } Y_{ij} \text{ is missing} \\ 1 & \text{if } Y_{ij} \text{ is observed.} \end{cases}$$

Under the selection model, the joint distribution is factorised as

$$p\left(\mathbf{Y}, \mathbf{M} \mid \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\psi}\right) = \underbrace{p\left(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}\right)}_{\text{data model}} \times \underbrace{p\left(\mathbf{M} \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\psi}\right)}_{\text{missingness model}}, \tag{3.1}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are sets of parameters in the data and missingness distributions, respectively. This selection model framework first came to prominence in Heckman [1976].

A particularly common specification of a selection model is the so-called 'Heckit' model, wherein a univariate response is modelled via parametric linear regression and the missingness indicator is modelled via probit regression. However, in principle, any model could be used to model the conditional distributions of the data and missingness indicators, so long as both distributions are modelled jointly to account for non-ignorable missingness. As the *global Amax* data consists of five continuous response variables, the multivariate BART model adapted from Um et al. [2023] is considered for the underlying

data model in both missBART1 and missBART2. Further details on the multivariate BART model are provided in Section 3.4.

The second part of the selection model factorisation from Equation (3.1) represents the conditional distribution of the missingness mechanism. If the assumption of MCAR holds, the conditional distribution of $\mathbf{M}$ depends only on $\boldsymbol{\psi}$ and is fully independent of the data. Similarly for MAR, only $\boldsymbol{\psi}$ and the fully observed set of covariates $\mathbf{X}$ are required. These two scenarios negate the need to specify explicit missingness models for likelihood inferences [Little, 2008]. In the context of predictive analysis, it is often sufficient to fit the predictive model on the available cases and impute missing responses from their predictive distribution. However, if the missing data mechanism is MNAR, a full specification of the missingness model is required. In practice, it is often challenging to make definitive assumptions about the underlying missingness mechanisms in real-world data. The joint model approach used in missBART1 and missBART2 allows for the recovery of all three types of missingness mechanism, offering flexibility and robustness even when the missingness mechanism is unknown, as is often the case.

As mentioned previously, we consider two binary models for the conditional distribution of $\mathbf{M}$. The first model, implemented in 'missBART1', is the probit regression model. In Albert and Chib [1993], the univariate probit regression model is described from a fully Bayesian viewpoint and allows for posterior inference via Gibbs sampling. Chib and Greenberg [1998] extended this to the multivariate framework by constraining the covariance matrix to a correlation matrix to ensure identifiability. However, computing posteriors for correlation matrices poses significant challenges due to the lack of conjugate priors. To address this, Talhouk et al. [2012] introduced a parameter-expanded data augmentation strategy, building upon the framework proposed by Liu and Wu [1999], which facilitates posterior inference for all parameters in the multivariate probit model within the Gibbs sampling framework. Thus, we incorporate this technique into 'missBART1' for modelling the multivariate missing data mechanism.

The univariate BART model was also modified in Chipman et al. [2010] to handle binary classification tasks, leveraging principles from the probit regression model described in Albert and Chib [1993]. Combining concepts from the multivariate probit regression model and the multivariate BART model facilitates the extension of the probit BART model to multiple dimensions. In comparison to the probit regression model, integrating the multivariate probit BART approach into our second joint model, 'missBART2', consequently enables us to harness the advantages of the BART model, particularly its ability to capture complex interactions between covariates as well as recover any non-linear patterns of missingness in the data.

Before describing our joint models in detail in Section 3.5, we first review the univariate and multivariate Bayesian probit regression models in Section 3.3, followed by a recap

of the univariate BART model, details on the multivariate BART model, and the probit equivalents of both in Section 3.4.

3.3 Bayesian Probit Regression

From Albert and Chib [1993], M_i is binary with probability

$$\Pr\left(M_i = 1 \mid \mathbf{Z}_i, \mathbf{b}\right) = \Phi\left(\mathbf{Z}_i^\top \mathbf{b}\right), \tag{3.2}$$

where $\mathbf{Z}_i$ is an r-dimensional vector containing $r-1$ model predictors and the intercept, $\mathbf{b}$ is an r-dimensional vector of model parameters, and $\Phi(\cdot)$ is the standard normal cumulative distribution function. To enable posterior sampling entirely within the Gibbs framework, a data augmentation scheme is adopted where n latent variables $M_1^\star, \ldots, M_n^\star$ are introduced such that

$$\begin{aligned} M_i^\star &= \mathbf{Z}_i^\top \mathbf{b} + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0,1\right), \\ M_i &= \begin{cases} 0 & \text{if } M_i^\star \leq 0 \\ 1 & \text{if } M_i^\star > 0. \end{cases} \end{aligned} \tag{3.3}$$

The joint posterior density of $\mathbf{M}^\star = (M_1^\star, \ldots, M_n^\star)$ and $\mathbf{b}$ is then

$$p\left(\mathbf{M}^\star, \mathbf{b} \mid \mathbf{M}\right) \propto \pi\left(\mathbf{b}\right) \prod_{i=1}^{n} \left\{ \phi\left(M_i^\star; \mathbf{Z}_i^\top \mathbf{b}, 1\right) \left[\mathbb{1}(M_i^\star > 0)\, \mathbb{1}(M_i = 1) + \mathbb{1}(M_i^\star \leq 0)\, \mathbb{1}(M_i = 0)\right] \right\},$$

where $\phi(x; \mu, \sigma)$ is the p.d.f of $x \sim \mathcal{N}(\mu, \sigma^2)$. By assigning a conjugate prior to $\mathbf{b}$, posterior samples of $\mathbf{b}$ can be drawn from a multivariate normal distribution, while posterior samples of $\mathbf{M}^\star$ are obtained from a truncated normal distribution.

The multivariate probit regression model from Chib and Greenberg [1998] generalises the univariate probit model from Albert and Chib [1993] and assumes a correlated structure between the multivariate binary outcomes. Given a p-dimensional set of binary outcomes $\mathbf{M}_i = (M_{i1}, \ldots, M_{ip})$, the probability of observing some combination of $\mathbf{M}_i = \mathbf{m}_i$, conditional on a set of r covariates $\mathbf{Z}_i$ and model parameters $(\mathbf{B}, \mathbf{R})$, is given by

$$\Pr\left(\mathbf{M}_i = \mathbf{m}_i \mid \mathbf{Z}_i, \mathbf{B}, \mathbf{R}\right) = \int_{A_{ip}} \cdots \int_{A_{i1}} \phi_p\left(\mathbf{u}; \mathbf{0}, \mathbf{R}\right) \mathrm{d}\mathbf{u}, \tag{3.4}$$

where $\mathbf{B}$ is an $r \times p$ matrix of regression coefficients and $\phi_p(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a p-variate normal density with mean vector $\boldsymbol{\mu}$ and $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. As specified in Chib and Greenberg [1998], $\mathbf{R}$ is restricted to the properties of a correlation matrix and A_{ij} is constrained via

$$A_{ij} = \begin{cases} \left(-\infty, \mathbf{B}^\top \mathbf{Z}_i\right) & \text{if } M_{ij} = 0 \\ \left[\mathbf{B}^\top \mathbf{Z}_i, \infty\right) & \text{if } M_{ij} = 1. \end{cases}$$

Similar to the univariate probit model, a data augmentation scheme can be applied by introducing the p-dimensional latent variables $\mathbf{M}_1^\star, \ldots, \mathbf{M}_n^\star$ such that

$$\mathbf{M}_i^\star = \mathbf{B}^\top \mathbf{Z}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} \mathcal{N}_p\left(\mathbf{0}, \mathbf{R}\right), \tag{3.5}$$

$$M_{ij} = \begin{cases} 0 & \text{if } M_{ij}^\star \leq 0 \\ 1 & \text{if } M_{ij}^\star > 0. \end{cases}$$

Due to the challenging nature of specifying a prior distribution for correlation matrices, we use the parameter-expansion strategy from Talhouk et al. [2012] to sample $\mathbf{B}$ and $\mathbf{R}$.

First, a marginally uniform prior from Barnard et al. [2000] is specified for $\mathbf{R}$, given by

$$p\left(\mathbf{R}\right) \propto |\mathbf{R}|^{\frac{p(p-1)}{2}} \left(\prod_{j=1}^{p} |R_{\{jj\}}|\right)^{-\frac{p+1}{2}}$$

where $R_{\{jj\}}$ represents the j^{th} principal sub-matrix of $\mathbf{R}$. A conjugate prior is assigned to $\mathbf{B}$ such that

$$\mathbf{B} \mid \mathbf{R} \sim \mathcal{MN}_{r\times p}\left(\mathbf{0}, \boldsymbol{\Psi}, \mathbf{R}\right), \tag{3.6}$$

where $\mathcal{MN}_{r\times p}\left(\mathbf{0}, \boldsymbol{\Psi}, \mathbf{R}\right)$ is a matrix-normal distribution with an $r \times p$ mean matrix of zeros, $r \times r$ positive-definite scale matrix of dispersion hyperparameters $\boldsymbol{\Psi}$, and $p \times p$ positive-definite correlation matrix $\mathbf{R}$.

Following this, $\mathbf{M}_i^*$ can be sampled from a multivariate truncated normal distribution from Damien and Walker [2001] with mean vector $\mathbf{B}^\top \mathbf{Z}_i$, covariance matrix $\mathbf{R}$, and truncation points $(-\infty, 0]$ if $M_{ij} = 0$ and $[0, \infty)$ if $M_{ij} = 1$. Next, expansion parameters $\mathbf{D} = \text{diag}(d_1, \ldots, d_p), d_j > 0$, are introduced, followed by a transformation of the probit regression latent variables such that $\mathbf{W} = \mathbf{M}^*\mathbf{D}$. Each $d_j \mid \mathbf{R}$ is assumed to be i.i.d., and can be sampled via $d_j^2 \mid \mathbf{R} \sim \mathcal{IG}\left(\frac{p+1}{2}, \frac{R_{jj}^{-1}}{2}\right)$, where $\mathcal{IG}$ denotes the inverse gamma distribution. This specification leads to $\boldsymbol{\Sigma} = \mathbf{DRD} \sim \mathcal{IW}(2, \mathbf{I}_p)$, where $\mathcal{IW}$ denotes the inverse Wishart distribution from Dawid and Lauritzen [1993]. Defining further $\boldsymbol{\Xi} = \mathbf{BD}$, we can sample $\boldsymbol{\Sigma} \mid \mathbf{W} \sim \mathcal{IW}(2 + n, \mathbf{W}^\top\mathbf{W} + \mathbf{I}_p - \boldsymbol{\Gamma}^\top\boldsymbol{\Lambda}\boldsymbol{\Gamma})$ and $\boldsymbol{\Xi} \mid \mathbf{W}, \boldsymbol{\Sigma} \sim \mathcal{N}_{r\times p}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Lambda}^{-1} = \mathbf{X}^\top\mathbf{X} + \boldsymbol{\Psi}^{-1}$ and $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}\mathbf{X}^\top\mathbf{W}$. Finally, compute $\mathbf{R} = \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}$ and $\mathbf{B} = \boldsymbol{\Xi}\mathbf{D}^{-1}$ where $\mathbf{D} = \text{diag}\left(\Sigma_{11}^{1/2}, \ldots, \Sigma_{pp}^{1/2}\right)$.

3.4 Multivariate BART

BART is a Bayesian sum-of-trees regression model that has earned substantial recognition since its development due to its flexibility and robustness while making accurate probabilistic predictions. Since its development, BART has been extended in various ways to handle multivariate responses. Um et al. [2023] developed a multivariate version of BART to handle multivariate skewed responses, McJames et al. [2024] extended the Bayesian causal forests [BCF; Hahn et al., 2020] model to analyse multivariate response data for causal inference, and Esser et al. [2024] proposed a method of incorporating seemingly unrelated regression [SUR; Zellner, 1962] into the multivariate BART framework.

In the multivariate framework, where there are $p > 1$ outcome variables, the multivariate BART model can be formulated as

$$\mathbf{Y}_i = \sum_{k=1}^{K} g\left(\mathbf{X}_i; \mathcal{T}_k, \boldsymbol{Q}_k\right) + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} \mathcal{N}_p\left(\mathbf{0}, \boldsymbol{\Omega}^{-1}\right), \tag{3.7}$$

where $\mathcal{N}_p$ denotes the p-variate normal distribution, $\boldsymbol{\Omega}$ represents the $p \times p$ residual precision matrix, and $\boldsymbol{Q}_k$ now contains the p-dimensional node-specific vectors $\boldsymbol{\mu}_{k\ell} = (\mu_{1k\ell}, \ldots, \mu_{pk\ell})$.

We calibrate the priors of the multivariate BART model by extending the calibration techniques adopted in the univariate framework to the multivariate setting, using the same prior settings as specified in Chipman et al. [2010]. The prior for $\boldsymbol{\mu}_{k\ell}$ is assigned a p-variate normal distribution, $\boldsymbol{\mu}_{k\ell} \sim \mathcal{N}_p(0, \boldsymbol{\Omega}_\mu^{-1})$ with $\boldsymbol{\Omega}_\mu = \tau_\mu \mathbf{I}_p$, where $\mathbf{I}_p$ denotes the p-dimensional identity matrix and τ_μ is chosen as in the univariate setting. Although the prior on $\boldsymbol{\Omega}_\mu$ assumes no covariance between the components of $\boldsymbol{\mu}_{k\ell}$, the posterior distribution of $\boldsymbol{\Omega}_\mu$ is expected to be non-diagonal when there is information to be shared across response variables. A conjugate Wishart prior is assigned to $\boldsymbol{\Omega}$ such that $\boldsymbol{\Omega} \sim \mathcal{W}_p(\nu, \mathbf{V})$. Given that the Wishart distribution is a multivariate extension of the gamma distribution, we calibrate the Wishart hyperparameters ν and $\mathbf{V}$ by first choosing a value for ν and a vector of p probabilities $\boldsymbol{\rho}_\tau = (\rho_{\tau 1}, \ldots, \rho_{\tau p})$ where $\rho_{\tau j} \in [0, 1] \, \forall \, j = 1, \ldots, p$ as in the univariate BART case. The rough data-based estimates $\hat{\tau}_1, \ldots, \hat{\tau}_p$ are then obtained for each univariate column of the outcome $\mathbf{Y}$. The scale matrix of the Wishart prior is then given by $\mathbf{V} = (\nu \boldsymbol{\lambda})^{-1} \mathbf{I}_p$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$, and each λ_j is calculated as in the univariate case. As before, posterior draws of $(\mathcal{T}_k, \boldsymbol{Q}_k)$ can be obtained by computing the multivariate form of the partial residuals

$$\mathbf{r}_k \equiv \mathbf{Y} - \sum_{t \neq k} g\left(\mathbf{X}; \mathcal{T}_t, \boldsymbol{Q}_t\right). \tag{3.8}$$

Following the same data augmentation scheme adopted in the probit regression model from Albert and Chib [1993] and described in Section 3.3, the univariate BART model in Chipman et al. [2010] also extends to binary classification settings. Assuming that the univariate outcome variable $\mathbf{Y}$ is binary, n latent variables $Y_1^\star, \ldots, Y_n^\star$ are introduced where

$$Y_i^\star = \sum_{k=1}^{K} g\left(\mathbf{X}_i; \mathcal{T}_k, \boldsymbol{Q}_k\right) + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0, 1\right), \tag{3.9}$$

$$Y_i = \begin{cases} 1, & \text{if } Y_i^\star > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The prior calibration for $\mathcal{T}_k$ and μ_k are similar to that specified previously in Section 2.1.1, with the exception that $\mathbb{E}(Y^\star \mid \mathbf{X})$ is expected to fall inside the range $[-3, 3]$. This corresponds to assigning a prior probability of 0.95 to the event $\{\Pr(Y_i = 1 \mid \mathbf{X}_i) \in [\Phi(-3), \Phi(3)] = [0.0013, 0.9987]\}$, which is reasonable for many applications since extremely small or large probabilities are uncommon.

Extending the probit BART model to the multivariate framework follows a similar approach. The vector of latent variables $\mathbf{Y}_i^\star$ now takes the form

$$\mathbf{Y}_i^\star = \sum_{k=1}^{K} g\left(\mathbf{X}_i; T_k, \boldsymbol{Q}_k\right) + \boldsymbol{\epsilon_i}, \quad \boldsymbol{\epsilon_i} \overset{i.i.d.}{\sim} \mathcal{N}_p\left(\mathbf{0}, \mathbf{R}\right). \tag{3.10}$$

As with the multivariate probit regression model, $\mathbf{R}$ is constrained to be a correlation matrix.

3.5 Joint Models for Multivariate MNAR Missing Data

We now describe our two joint models, missBART1 and missBART2, developed under the selection model framework from Equation (3.1) to handle data with missing responses under the non-ignorable assumption. In the presence of missing values, we define the partially-observed responses $\tilde{\mathbf{Y}}$ as an $n \times p$ matrix such that

$$\tilde{Y}_{ij} = \begin{cases} Y_{ij}^{obs} & \text{if } M_{ij} = 1 \\ Y_{ij}^{mis} & \text{otherwise.} \end{cases} \tag{3.11}$$

Here, $\mathbf{Y}^{obs} = \{Y_{ij} \colon M_{ij} = 1\}$ refers to the set of observed responses, which are fixed and known, and $\mathbf{Y}^{mis} = \{Y_{ij} \colon M_{ij} = 0\}$ refers to the set of missing responses which are estimated as part of the model updates. Of key importance in these models is that $\mathbf{X}$ is used as the covariates in the BART data model, and the set $(\mathbf{X}, \tilde{\mathbf{Y}})$ is used as the predictors in the missingness model.

While both models assign the multivariate BART function from Equation (3.7) to the data model $\tilde{\mathbf{Y}} \mid \mathbf{X}, \boldsymbol{\theta}$, missBART1 assigns a multivariate probit regression model from Section 3.3 to the missingness model $\mathbf{M} \mid \mathbf{X}, \tilde{\mathbf{Y}}, \boldsymbol{\psi}$. By contrast missBART2 assigns the multivariate probit BART model from Section 3.4 as the missingness model instead. Details of both models are given in the framework of multivariate responses below, though it is straightforward to reduce both models to their univariate equivalents.

3.5.1 missBART1

In this model, the joint distribution of $\tilde{\mathbf{Y}}$ and $\mathbf{M}$ is obtained by combining Equation (3.4) and Equation (3.7) while setting $\mathbf{Z_i} = (1, \mathbf{X}_i, \tilde{\mathbf{Y}}_i)^\top$ in Equation (3.4) such that the complete data likelihood is:

$$p\left(\tilde{\mathbf{Y}}, \mathbf{M} \mid \mathbf{X}, \boldsymbol{\mathcal{T}}, \boldsymbol{Q}, \boldsymbol{\Omega}, \mathbf{B}, \mathbf{R}\right) = \underbrace{p\left(\tilde{\mathbf{Y}} \mid \mathbf{X}, \boldsymbol{\mathcal{T}}, \boldsymbol{Q}, \boldsymbol{\Omega}\right)}_{\text{BART regression model}} \times \underbrace{p\left(\mathbf{M} \mid \mathbf{Z}, \mathbf{B}, \mathbf{R}\right)}_{\substack{\text{probit regression} \\ \text{missingness model}}}. \tag{3.12}$$

This specification, which includes $\mathbf{X}$ and $\tilde{\mathbf{Y}}$ as predictors in the missingness model as well as an intercept term, accounts for the three different types of missing data mechanism. To illustrate this, assume for simplicity that we have a partially-observed univariate response and only one covariate, which is fully observed. Then, from Equation (3.2), the probability of Y_i being observed is equal to $\Pr(M_i = 1 | X_i, Y_i, \mathbf{b}) = \Phi(b_0 + b_1 X_i + b_2 Y_i)$. If $b_1 = b_2 = 0$, the probability of observing Y_i is constant for all $i = 1, \ldots, n$ and only depends on the intercept b_0, representing an MCAR mechanism. If $b_1 \neq 0$ while $b_2 = 0$, the MAR mechanism is present since the detection probability depends on the value of the observed X_i. If $b_2 \neq 0$, the probability of Y_i being observed depends on the value of Y_i itself, implying an MNAR missingness mechanism, regardless of the values of b_0 and b_1. In the multivariate setting, the matrix of coefficients $\mathbf{B}$ has dimensions $r \times p$ where $r = 1 + p + q$, i.e. the total number of columns in $\mathbf{Z}$. Define $\mathbf{B}^{(j)} = (\mathbf{B}_{1j}, \ldots, \mathbf{B}_{rj})$ as the j^{th} column of $\mathbf{B}$. Each element of $\mathbf{B}^{(j)}$ represents the degree to which each predictor influences the missingness of response j.

Under the data augmentation scheme from Equation (3.5), the joint posterior distribution of the model takes the form

$$p\left(\boldsymbol{\mathcal{T}}, \boldsymbol{Q}, \boldsymbol{\Omega}, \mathbf{M}^\star, \mathbf{B}, \mathbf{R}, \boldsymbol{\Psi}, \mathbf{Y}^{mis} \mid \mathbf{X}, \mathbf{Y}^{obs}, \mathbf{M}\right). \tag{3.13}$$

By assigning the multivariate BART priors specified in previous sections, posterior sampling for $\boldsymbol{\mathcal{T}}$ is carried out via Metropolis-Hastings, while $\boldsymbol{Q}$ and $\boldsymbol{\Omega}$ are sampled through Gibbs updates, where posterior derivations follow closely from the univariate model. Ad-

ditionally, $\mathbf{M}^{\star}, \mathbf{B}$ and $\mathbf{R}$ are also sampled within Gibbs through the incorporation of work by Talhouk et al. [2012], previously outlined in Section 3.3. This allows for posterior inferences to be made on $\mathbf{B}$, where posterior intervals of $\mathbf{B}^{(j)}$ — to the extent that they include or exclude 0 — can give insights into the underlying missing data mechanism of the partially-observed data.

While the multivariate probit regression model from Talhouk et al. [2012] assumes that $\boldsymbol{\Psi}$ in Equation (3.6) is known, we propose a modification which allows the incorporation of prior beliefs about the missing data mechanism. We note that small values on the diagonal of $\boldsymbol{\Psi}$ correspond to probit regression coefficients that are close to zero, and thus have little effect on the missingness. Since our covariates for the missingness model consist of $\mathbf{X}$ and $\tilde{\mathbf{Y}}$, we can tailor the prior distribution to favour MAR (coefficients associated with $\tilde{\mathbf{Y}}$ close to zero) or MNAR structures (coefficients associated with $\tilde{\mathbf{Y}}$ are not zero).

To structure this prior distribution, we first set $\boldsymbol{\Psi}^{-1} = \text{diag}(\tau_{B_0}, \tau_{B_X}\mathbf{1}_q, \tau_{B_Y}\mathbf{1}_p)$, where $\mathbf{1}_q$ and $\mathbf{1}_p$ are q and p-dimensional vectors of ones, respectively. Next, we assign separate gamma priors to τ_{B_0}, τ_{B_X}, and τ_{B_Y} with shape and rate parameters (α_0, β_0), (α_X, β_X) and (α_Y, β_Y), respectively. Following this, the full conditional distribution is

$$\begin{aligned} p(\tau_{B_0}, \tau_{B_X}, \tau_{B_Y} \mid \cdot) \propto |\boldsymbol{\Psi}|^{-\frac{p}{2}} \exp\left\{-\frac{1}{2}\,\text{tr}(\mathbf{R}^{-1}\mathbf{B}^{\top}\boldsymbol{\Psi}^{-1}\mathbf{B})\right\} \times \tau_{B_0}^{\alpha_0-1} \exp\left\{-\beta_0\tau_{B_0}\right\} \\ \times \tau_{B_X}^{\alpha_X-1} \exp\left\{-\beta_X\tau_{B_X}\right\} \times \tau_{B_Y}^{\alpha_Y-1} \exp\left\{-\beta_Y\tau_{B_Y}\right\}. \end{aligned}$$

By using the trace property $\text{tr}(\mathbf{R}^{-1}\mathbf{B}^{\top}\boldsymbol{\Psi}^{-1}\mathbf{B}) = \text{tr}(\boldsymbol{\Psi}^{-1}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\top})$ and letting $\mathbf{A} = \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\top}$, we obtain

$$\begin{aligned} \text{tr}(\boldsymbol{\Psi}^{-1}\mathbf{A}) &= \text{tr}\left((\tau_{B_0}\mathbf{A}_1, \tau_{B_X}\mathbf{A}_2, \ldots, \tau_{B_X}\mathbf{A}_{1+q}, \tau_{B_Y}\mathbf{A}_{2+q}, \ldots, \tau_{B_Y}\mathbf{A}_r)^{\top}\right) \\ &= \tau_{B_0}A_{11} + \tau_{B_X}A_{22} + \ldots + \tau_{B_X}A_{(1+q)(1+q)} + \tau_{B_Y}A_{(2+q)(2+q)} + \ldots + \tau_{B_Y}A_{rr}, \end{aligned}$$

where $\mathbf{A}_i, i = 1, \ldots, r$, is the i^{th} row of the matrix $\mathbf{A}$, and A_{ii} is the i^{th} diagonal entry of $\mathbf{A}$. From here, we have

$$\begin{aligned} p\left(\tau_{B_0}, \tau_{B_X}, \tau_{B_Y} \mid \cdot\right) \propto \left(\tau_{B_0} \times \tau_{B_X}^{q} \times \tau_{B_Y}^{p}\right)^{\frac{p}{2}} \exp\left\{-\frac{1}{2}\left(\tau_{B_0}A_{11} + \tau_{B_X}\sum_{i=2}^{1+q}A_{ii} + \tau_{B_Y}\sum_{i=2+q}^{r}A_{ii}\right)\right\} \\ \times \tau_{B_0}^{\alpha_0-1} \exp\left\{-\beta_0\tau_{B_0}\right\} \times \tau_{B_X}^{\alpha_X-1} \exp\left\{-\beta_X\tau_{B_X}\right\} \times \tau_{B_Y}^{\alpha_Y-1} \exp\left\{-\beta_Y\tau_{B_Y}\right\}, \end{aligned}$$

from which we can easily see that

$$\begin{aligned} p\left(\tau_{B_0} \mid \cdot\right) &\propto \tau_{B_0}^{\frac{p}{2}} \exp\left\{-\frac{1}{2}\tau_{B_0}A_{11}\right\} \tau_{B_0}^{\alpha_0-1} \exp\left\{-\beta_0\tau_{B_0}\right\} \\ &\propto \tau_{B_0}^{\frac{p}{2}+\alpha_0-1} \exp\left\{-\left(\frac{1}{2}A_{11} + \beta_0\right)\tau_{B_0}\right\} \end{aligned}$$

$$
\begin{aligned}
p\left(\tau_{B_X} \mid \cdot\right) &\propto \tau_{B_X}^{\frac{pq}{2}} \exp\left\{-\frac{1}{2}\tau_{B_X}\sum_{i=2}^{1+q} A_{ii}\right\} \tau_{B_X}^{\alpha_X-1}\exp\left\{-\beta_X\tau_{B_X}\right\} \\
&\propto \tau_{B_X}^{\frac{pq}{2}+\alpha_X-1}\exp\left\{-\left(\frac{1}{2}\sum_{i=2}^{1+q}A_{ii}+\beta_X\right)\tau_{B_X}\right\} \\
p\left(\tau_{B_Y} \mid \cdot\right) &\propto \tau_{B_Y}^{\frac{p^2}{2}} \exp\left\{-\frac{1}{2}\tau_{B_Y}\sum_{i=2+q}^{r} A_{ii}\right\} \tau_{B_Y}^{\alpha_Y-1}\exp\left\{-\beta_Y\tau_{B_Y}\right\} \\
&\propto \tau_{B_Y}^{\frac{p^2}{2}+\alpha_Y-1}\exp\left\{-\left(\frac{1}{2}\sum_{i=2+q}^{r}A_{ii}+\beta_Y\right)\tau_{B_Y}\right\}.
\end{aligned}
$$

Finally, the posterior distributions are

$$\tau_{B_0} \mid \cdot \sim Ga\left(\frac{p}{2}+\alpha_0, \frac{A_{11}}{2}+\beta_0\right) \tag{3.14}$$

$$\tau_{B_X} \mid \cdot \sim Ga\left(\frac{pq}{2}+\alpha_X, \frac{\sum_{i=2}^{1+q}A_{ii}}{2}+\beta_X\right) \tag{3.15}$$

$$\tau_{B_Y} \mid \cdot \sim Ga\left(\frac{p^2}{2}+\alpha_Y, \frac{\sum_{i=2+q}^{r}A_{ii}}{2}+\beta_Y\right). \tag{3.16}$$

By carefully tuning these prior parameters, we can articulate any prior beliefs regarding the missingness mechanisms. As a default, we set $(\alpha_0, \beta_0) = (2, 1)$, $(\alpha_X, \beta_X) = (1+q, 1)$ and $(\alpha_Y, \beta_Y) = (1+p+q, 1)$. Scaling the prior mean of τ_{B_Y} with the number of covariates and responses ensures that it remains larger than the prior mean of τ_{B_X}, increasing the likelihood that the model is *a priori* MAR. This also helps keep the values in $\mathbf{B}$ small as the number of covariates and responses increases, and *vice versa*.

Finally, to sample $\mathbf{Y}^{mis}$, we make draws from

$$p\left(\mathbf{Y}^{mis} \mid \mathbf{X}, \mathbf{Y}^{obs}, \boldsymbol{\mathcal{T}}, \boldsymbol{Q}, \boldsymbol{\Omega}, \mathbf{M}^{\star}, \mathbf{B}, \mathbf{R}\right) \propto p\left(\mathbf{Y}^{mis} \mid \mathbf{X}, \mathbf{Y}^{obs}, \boldsymbol{\mathcal{T}}, \boldsymbol{Q}, \boldsymbol{\Omega}\right) p\left(\mathbf{M}^{\star} \mid \mathbf{X}, \tilde{\mathbf{Y}}, \mathbf{B}, \mathbf{R}\right).$$

The sampling distribution takes the form

$$\mathbf{Y}_i^{mis} \mid \mathbf{X}, \mathbf{Y}_i^{obs}, \boldsymbol{\mathcal{T}}, \boldsymbol{Q}, \boldsymbol{\Omega}, \mathbf{M}_i^{\star}, \mathbf{B}, \mathbf{R} \sim \mathcal{N}_{p_i}\left(\left[\boldsymbol{\mu}_Y\right]_{\mathcal{M}_i}, \left[\boldsymbol{\Sigma}_Y\right]_{\mathcal{M}_i,\mathcal{M}_i}\right), \tag{3.17}$$

where $\mathcal{M}_i = \{j \mid M_{ij} = 0\}$ is the set of column indices where $\tilde{\mathbf{Y}}_i$ is missing, p_i is equal to the number of elements in $\mathcal{M}_i$,

$$\boldsymbol{\Sigma}_Y = \left(\boldsymbol{\Omega} + \mathbf{B}_Y\mathbf{R}^{-1}\mathbf{B}_Y^{\top}\right)^{-1}, \quad \boldsymbol{\mu}_Y = \boldsymbol{\Sigma}_Y\left[\boldsymbol{\Omega}\hat{\tilde{\mathbf{Y}}}_i + \mathbf{B}_Y\mathbf{R}^{-1}\left(\mathbf{M}_i^{\star} - \mathbf{B}_{(Y)}^{\top}\mathbf{Z}_{(Y)i}\right)\right],$$

$\left[\boldsymbol{\mu}_Y\right]_{\mathcal{M}_i}$ is the p_i-dimensional subset of $\boldsymbol{\mu}_Y$ obtained by extracting the $\mathcal{M}_i$ elements from $\boldsymbol{\mu}_Y$, and $\left[\boldsymbol{\Sigma}_Y\right]_{\mathcal{M}_i,\mathcal{M}_i}$ is the $p_i \times p_i$ submatrix of $\boldsymbol{\Sigma}_Y$ obtained in a similar fashion. Additionally, we have that

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{(Y)} \\ \mathbf{B}_Y \end{bmatrix} = \begin{bmatrix} \mathbf{B}_0 \\ \mathbf{B}_X \\ \mathbf{B}_Y \end{bmatrix} = \begin{bmatrix} b_{11} & \dots & b_{1p} \\ b_{21} & \dots & b_{2p} \\ \vdots & \dots & \vdots \\ b_{(1+q)1} & \dots & b_{(1+q)p} \\ b_{(r-p+1)1} & \dots & b_{(r-p+1)p} \\ \vdots & \dots & \vdots \\ b_{r1} & \dots & b_{rp} \end{bmatrix},$$

where $\mathbf{B}_Y$ is the $p \times p$ submatrix of $\mathbf{B}$ obtained by removing the first $1+q$ rows of $\mathbf{B}$, and $\mathbf{B}_{(Y)} = \{\mathbf{B}_0, \mathbf{B}_X\}$ is the complementary $(1+q) \times p$ submatrix with probit regression parameters associated with the intercept and covariates $\mathbf{X}$. $\mathbf{Z}_{(Y)i}$ denotes the vector of probit predictors which contain only the intercept term and $\mathbf{X}_i$, i.e. $\mathbf{Z}_{(Y)i} = (1, X_{i1}, \dots, X_{iq})^\top$. With efficient sampling of $\mathbf{Y}^{mis}$ within Gibbs, imputations for missing responses can be obtained via computing the posterior mean of $\mathbf{Y}^{mis}$, along with uncertainty quantification.

The steps for posterior sampling in missBART1 are outlined below:

(1) For all K trees, propose a new tree via a grow, prune, change, or swap move[a] and accept or reject using a Metropolis-Hastings step. For notational simplicity, we drop the tree index k and thus the tree posterior takes the form

$$p\left(\mathcal{T} \mid \mathbf{r}, \boldsymbol{\Omega}\right) \propto \pi\left(\mathcal{T}\right) \prod_{\ell} p\left(\mathbf{r}_\ell \mid \mathcal{T}, \boldsymbol{\Omega}\right),$$

where $\pi(\mathcal{T})$ is the tree prior from Chipman et al. [2010] and $\mathbf{r}_\ell$ denotes the partial residuals from Equation (3.8) assigned to terminal node ℓ in tree k. Note that

$$\begin{aligned} p\left(\mathbf{r}_\ell \mid \mathcal{T}, \boldsymbol{\Omega}\right) &= \int p\left(\mathbf{r}_\ell \mid \mathcal{T}, \boldsymbol{\mu}_\ell, \boldsymbol{\Omega}\right) \pi\left(\boldsymbol{\mu}_\ell\right) \mathrm{d}\boldsymbol{\mu}_\ell \\ &= (2\pi)^{-\frac{n_\ell p}{2}} \tau_\mu^{\frac{p}{2}} |\boldsymbol{\Omega}|^{\frac{n_\ell}{2}} |\boldsymbol{\Sigma}_\mu|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\mu}_0^\top \left(\tau_\mu \mathbf{I}_p\right) \boldsymbol{\mu}_0 - \boldsymbol{\mu}_r^\top \boldsymbol{\Sigma}_r^{-1} \boldsymbol{\mu}_r + \sum_{i=1}^{n_\ell} \mathbf{r}_{\ell i}^\top \boldsymbol{\Omega} \mathbf{r}_{\ell i} \right] \right\}, \end{aligned}$$

where $\boldsymbol{\mu}_r = \boldsymbol{\Sigma}_r \left[\boldsymbol{\Omega}(\sum_{i=1}^{n_\ell} \mathbf{r}_{\ell i}) + (\tau_\mu \mathbf{I}_p)\boldsymbol{\mu}_0\right]$, $\boldsymbol{\Sigma}_r^{-1} = n_\ell \boldsymbol{\Omega} + \tau_\mu \mathbf{I}_p$, and n_ℓ denotes the total number of observations which fall under terminal node ℓ.

(2) For each terminal node ℓ in tree k, once again dropping the tree index k, make a draw of $\boldsymbol{\mu}_\ell \in \boldsymbol{Q}$ from

$$\boldsymbol{\mu}_\ell \mid \mathbf{r}_\ell, \mathcal{T}, \boldsymbol{\Omega} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_\mu, \boldsymbol{\Sigma}_\mu\right),$$

where $\boldsymbol{\mu}_\mu = \boldsymbol{\Sigma}_\mu \boldsymbol{\Omega} \sum_{i=1}^{n_\ell} \mathbf{r}_{\ell i}$ and $\boldsymbol{\Sigma}_\mu = \left(n_\ell \boldsymbol{\Omega} + \tau_\mu \mathbf{I}_p\right)^{-1}$.

[a] See Kapelner and Bleich [2016] for details on these tree-proposal moves.

(3) After carrying out steps 1 and 2 for all K trees, sample $\boldsymbol{\Omega}$ from

$$\boldsymbol{\Omega} \mid \mathcal{T}, \boldsymbol{Q}, \tilde{\mathbf{Y}} \sim \mathcal{W}_p\left(n+\nu, \mathbf{V}_\Omega\right),$$

where $\mathbf{V}_\Omega^{-1} = \sum_{i=1}^{n}\left(\tilde{\mathbf{Y}}_i - \widehat{\tilde{\mathbf{Y}}}_i\right)\left(\tilde{\mathbf{Y}}_i - \widehat{\tilde{\mathbf{Y}}}_i\right)^\top + \mathbf{V}^{-1}$ and $\widehat{\tilde{\mathbf{Y}}}_i = \sum_{k=1}^{K} g\left(\mathbf{X}_i; \mathcal{T}_k, \boldsymbol{Q}_k\right)$.

(4) The posterior distribution of $\mathbf{M}_i^\star$ follows a multivariate truncated normal distribution [Damien and Walker, 2001]

$$\mathbf{M}_i^\star \mid \mathbf{X}_i, \tilde{\mathbf{Y}}_i, \mathbf{B}, \mathbf{M}_i \sim \mathcal{TN}_p(\mathbf{B}^\top \mathbf{Z}_i, \mathbf{R}, \boldsymbol{\gamma}_i),$$

where $\mathbf{Z}_i = (1, \mathbf{X}_i, \tilde{\mathbf{Y}}_i)^\top$ and $\boldsymbol{\gamma}_i$ denotes the p-dimensional vector of truncation points such that $\gamma_{ij} = [0, \infty)$ if $M_{ij} = 1$ and $\gamma_{ij} = (-\infty, 0]$ if $M_{ij} = 0$.

(5) Sample $\boldsymbol{\Psi}^{-1} = \text{diag}(\tau_{B_0}, \tau_{B_X}\mathbf{1}_q, \tau_{B_Y}\mathbf{1}_p)$ using Equations (3.14), (3.15), and (3.16).

(6) Sample ($\mathbf{B}$, $\mathbf{R}$) using the methods from Talhouk et al. [2012], outlined in Section 3.3.

(7) Sample $\mathbf{Y}^{mis}$ from Equation (3.17).

3.5.2 missBART2

Due to its inherent linear structure, the probit regression model may prove inadequate when the true underlying relationship in the missing data model is non-linear. Additionally, the missing data model specification in missBART1 excludes any interaction terms between the model predictors. Consequently, to account for potential non-linearity and predictor interactions, we propose a more flexible, fully non-parametric selection model by applying the multivariate probit BART model outlined in Section 3.4 to the missingness mechanism model. Thus, the model effectively amounts to replacing the probit regression missingness model for $p(\mathbf{M} \mid \mathbf{Z}, \mathbf{B}, \mathbf{R})$ in Equation (3.12) with a multivariate probit BART model such that the complete data likelihood for missBART2 is

$$p\left(\tilde{\mathbf{Y}}, \mathbf{M} \mid \mathbf{X}, \boldsymbol{\mathcal{T}}^y, \boldsymbol{Q}^y, \boldsymbol{\mathcal{T}}^m, \boldsymbol{Q}^m, \boldsymbol{\Omega}, \mathbf{R}\right) = \underbrace{p\left(\tilde{\mathbf{Y}} \mid \mathbf{X}, \boldsymbol{\mathcal{T}}^y, \boldsymbol{Q}^y, \boldsymbol{\Omega}\right)}_{\text{BART regression model}} \times \underbrace{p\left(\mathbf{M} \mid \mathbf{Z}, \boldsymbol{\mathcal{T}}^m, \boldsymbol{Q}^m, \mathbf{R}\right)}_{\substack{\text{probit BART} \\ \text{missingness model}}}. \tag{3.18}$$

Overall, the model uses two distinct sets of trees with corresponding terminal node parameters, denoted by $(\boldsymbol{\mathcal{T}}^y, \boldsymbol{Q}^y) = \left((\mathcal{T}_1^y, \boldsymbol{Q}_1^y), \ldots, (\mathcal{T}_{K_y}^y, \boldsymbol{Q}_{K_y}^y)\right)$ for the regression on $\tilde{\mathbf{Y}}$ and $(\boldsymbol{\mathcal{T}}^m, \boldsymbol{Q}^m) = \left((\mathcal{T}_1^m, \boldsymbol{Q}_1^m), \ldots, (\mathcal{T}_{K_m}^m, \boldsymbol{Q}_{K_m}^m)\right)$ for $\mathbf{M}$, where K_y and K_m are the total number of trees in each respective BART model.

In Esser et al. [2024], a parameter-expanded data augmentation technique from Zhang [2020] is adopted for estimating $\mathbf{R}$. Given the flexibility of tree structures in the missing data model to capture the underlying signal, we find it sufficient to fix $\mathbf{R} = \mathbf{I}_p$. This choice also reduces the computational burden of estimating the correlation structure via additional data augmentation techniques. Furthermore, our simulation studies show that even when the missing data model is simulated using a non-diagonal correlation matrix, missBART2 still performs well despite this simplification.

By including both $\mathbf{X}$ and $\tilde{\mathbf{Y}}$ as predictors available to be used in the splitting rules of $\boldsymbol{\mathcal{T}}^m$, we can account for different missingness mechanisms. While missBART1 has the flexibility of incorporating prior knowledge of the missing data mechanisms via the prior calibration of $\mathbf{B}$, the probit BART model in missBART2 can perform automatic variable selection without needing to pre-specify the functional form of the probit model. Although missBART2 does not have the luxury of being able to examine interpretable probit regression parameters, identifying which variables in $\mathbf{X}$ and $\tilde{\mathbf{Y}}$ are most commonly used to construct splitting rules in the missingness model's set of trees can nonetheless provide an indication of whether the missingness mechanism depends only on observed quantities (i.e., MAR) or unobserved quantities (i.e., MNAR). Simulation studies from Chipman et al. [2010] on BART variable selection show that, as the number of trees increases, the frequency with which each variable is used in the tree-splitting rules becomes more uniform. Conversely, if there are fewer trees, important and influential variables are more likely to be included in the splitting rules of the trees. However, too few trees can hinder convergence and compromise overall model fit. Thus, care should be taken when defining the number of trees to be used in the missingness model, particularly when the recovery of different missing data mechanisms is of interest.

An additional advantage comes from incorporating the `BARTm` methodology introduced in Kapelner and Bleich [2015] to missBART2 such that it can effectively handle missing values in $\mathbf{X}$. This approach seamlessly integrates covariate missingness into the tree-splitting rules, enabling splits based on the available $\mathbf{X}$ values as well as their missingness status. In contrast, the parametric constraints of the probit regression model necessitate prior covariate imputation in missBART1 when faced with missing values in $\mathbf{X}$.

Following the latent variable transformation of $\mathbf{M}$ to $\mathbf{M}^\star$ as in Equation 3.9, the joint posterior distribution takes the form

$$p\left(\boldsymbol{\mathcal{T}}^y, \boldsymbol{Q}^y, \boldsymbol{\mathcal{T}}^m, \boldsymbol{Q}^m, \boldsymbol{\Omega}, \mathbf{M}^\star, \mathbf{Y}^{mis} \mid \mathbf{X}, \mathbf{Y}^{obs}, \mathbf{M}\right). \tag{3.19}$$

The sampling algorithm for missBART2 is similar to that outlined in Section 3.5.1 with a few exceptions. First, posterior sampling for both sets of trees $\boldsymbol{\mathcal{T}}^y$ and $\boldsymbol{\mathcal{T}}^m$ are carried out via Metropolis-Hastings steps. Next, $\boldsymbol{Q}^y, \boldsymbol{\Omega}, \mathbf{M}^\star$, and $\boldsymbol{Q}^m$ are updated via Gibbs steps.

It is notable that, unlike step (6) of the algorithm for missBART1 in Section 3.5.1, it is not necessary to sample ($\mathbf{B}$, $\mathbf{R}$) under missBART2, since the matrix of probit regression coefficients $\mathbf{B}$ is irrelevant for missBART2 and we fix $\mathbf{R} = \mathbf{I}_p$. Thus, the parameter-expansion techniques of Talhouk et al. [2012] or Zhang [2020] are not required.

Defining the regression model parameters $\boldsymbol{\theta} = \{\boldsymbol{\mathcal{T}}^y, \boldsymbol{Q}^y, \boldsymbol{\Omega}\}$ and missing model parameters $\boldsymbol{\psi} = \{\boldsymbol{\mathcal{T}}^m, \boldsymbol{Q}^m\}$, the full conditional distribution of $\mathbf{Y}^{mis}$ is

$$p\left(\mathbf{Y}^{mis} \mid \mathbf{Y}^{obs}, \mathbf{M}^\star, \boldsymbol{\theta}, \boldsymbol{\psi}\right) \propto p\left(\mathbf{Y}^{obs}, \mathbf{Y}^{mis} \mid \boldsymbol{\theta}\right) p\left(\mathbf{M}^\star \mid \mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \boldsymbol{\psi}\right),$$

where no known distributional form is available, necessitating the implementation of a Metropolis-Hastings step. For each missing entry in iteration t, we propose a new value $Y^{mis}_{t,i,j}$ from a random-walk proposal distribution $\mathcal{N}(Y^{mis}_{t-1,i,j}, \sigma^2_Y)$. Given that the observed data are scaled to the range $[-0.5, 0.5]$, we set $\sigma_Y = 0.5/p$ to ensure that proposed values of $\mathbf{Y}^{mis}$ do not go too far outside this range. This tuning choice has demonstrated good empirical performance in simulations.

Overall, the sampling algorithm for missBART2 is given by

(1) Repeat Steps 1 and 2 from Section 3.5.1 for all K_y trees.

(2) Repeat Steps 1 and 2 from Section 3.5.1 for all K_m trees.

(3) Repeat Step 3 from Section 3.5.1 for sampling $\boldsymbol{\Omega}$.

(4) The posterior distribution of $\mathbf{M}^\star_i$ again follows a multivariate truncated normal distribution such that

$$\mathbf{M}^\star_i \mid \mathbf{X}_i, \mathbf{Y}_i, \boldsymbol{\mathcal{T}}^m, \boldsymbol{Q}^m, \mathbf{M}_i \sim \mathcal{TN}_p(\hat{\mathbf{M}}^\star_i, \mathbf{I}_p, \boldsymbol{\gamma}_i),$$

where $\hat{\mathbf{M}}^\star_i = \sum_{k_m=1}^{K_m} g(\mathbf{X}_i, \mathbf{Y}_i; T^m_{k_m}, \boldsymbol{Q}^m_{k_m})$ and $\boldsymbol{\gamma}_i$ is as defined in Step 4 of Section 3.5.1.

(5) For every missing entry, first propose a new value $Y^{mis}_{t,i,j}$ from $\mathcal{N}(Y^{mis}_{t-1,i,j}, \sigma^2_Y)$. Next, calculate the acceptance probability

$$\omega(\mathbf{Y}^{mis}_{t,i}, \mathbf{Y}^{mis}_{t-1,i}) = \frac{p\left(\mathbf{Y}^{mis}_{t,i} \mid \mathbf{Y}^{obs}_i, \boldsymbol{\theta}\right) p\left(\mathbf{M}^\star_i \mid \mathbf{Y}^{obs}_i, \mathbf{Y}^{mis}_{t,i}, \boldsymbol{\psi}\right) q(\mathbf{Y}^{mis}_{t,i} \to \mathbf{Y}^{mis}_{t-1,i})}{p\left(\mathbf{Y}^{mis}_{t-1,i} \mid \mathbf{Y}^{obs}_i, \boldsymbol{\theta}\right) p\left(\mathbf{M}^\star_i \mid \mathbf{Y}^{obs}_i, \mathbf{Y}^{mis}_{t-1,i}, \boldsymbol{\psi}\right) q(\mathbf{Y}^{mis}_{t-1,i} \to \mathbf{Y}^{mis}_{t,i})}$$

and accept or reject the proposed $\mathbf{Y}^{mis}_t$ with probability $\min\left(1, \omega(\mathbf{Y}^{mis}_{t,i}, \mathbf{Y}^{mis}_{t-1,i})\right)$.

Finally, we note that an `R` implementation of both missBART1 and missBART2 is available on GitHub[b].

[b] https://github.com/yongchengoh/missBART.

4

missBART1 and missBART2: Simulation Studies and Application to Real Data

4.1 Introduction

We apply 'missBART1' and 'missBART2' to simulated data and the *global Amax* data. In Section 4.2, results from several simulation studies are discussed, along with comparisons made between other methods for handling missing data to showcase the benefits and importance of appropriately handling MNAR missing data via joint models. Examples with missingness in the covariates are also included. In Section 4.3, we describe the *global Amax* data in detail and show results obtained from applying 'missBART1' and 'missBART2' to the data. The chapter ends in Section 4.4 with concluding statements and discussions on limitations and future work.

Evaluating predictive performance in the presence of MNAR responses requires careful consideration, as models that ignore missingness may perform well on observed values but struggle with making accurate imputations for the unobserved. In real data applications, the "true" missing values remain unknown, presenting a significant challenge for assessing the accuracy of imputations. Nonetheless, we address this in our simulation studies in Section 4.2 when comparing our joint models to other approaches, such as complete case analysis and imputation followed by model fitting.

Using simulated data with fully generated responses and induced missingness, the out-of-sample root mean squared error (RMSE) is calculated in two ways: (1) across different simulated detection probability thresholds and (2) separately for the observed responses, the imputed missing responses, and the combined dataset comprising both observed and imputed responses. The first approach illustrates each model's performance on subsets of data with detection probabilities below specific thresholds, where models ignoring missingness perform worse on data which are more likely to be missing. The second method reiterates this by separating the RMSEs for observed, missing, and combined responses, offering a comprehensive assessment of model performance with respect to missingness.

To provide a more comprehensive evaluation, we also compute the continuous ranked probability scores [CRPS; Gneiting and Raftery, 2007], which assesses both calibration and accuracy of the predicted distributions. The CRPS complements the RMSE by evaluating the entire predictive distribution rather than just point estimates, offering deeper insights into how well models capture uncertainty in the presence of missing data. Additionally, for the combined dataset, we calculate the Frobenius norm, which serves as an overall measure of model performance.

In the real data application presented in Section 4.3, we apply 'missBART1' and 'missBART2' to the *global Amax* dataset. Unlike in simulation studies, where the true values of missing responses are known, real datasets do not provide a direct means of evaluating imputation accuracy. Consequently, standard performance metrics such as RMSE and CRPS cannot be used to assess the quality of the imputations. Instead, we evaluate model performance using a combination of visual diagnostics and posterior inference, while also comparing our findings with those of Maire et al. [2015].

4.2 Simulation Study

We present the results from our simulation studies aimed at demonstrating the advantages of missBART1 and missBART2 for performing predictive tasks in the presence of non-ignorable missingness[a]. First, we illustrate two univariate response cases featuring non-linear patterns of MNAR missingness where the detection probabilities of the response vary non-linearly with the values of the responses themselves. We refer to these two examples as the 'u-shape' and 'n-shape' missingness. For both scenarios, a complete dataset is first simulated from the Friedman function [Friedman, 1991]; this is followed by inducing missingness from a single tree structure. We present the posterior imputations for the missing cases and predictions made on the observed cases obtained by both models. This example reveals insights into the challenges encountered by missBART1, which assumes a monotonic missingness pattern, compared to missBART2, which can capture non-linear patterns.

Through 4-fold cross-validation, we make comparisons between the performances of our joint models with two alternative methods: applying a standard BART model to complete cases ('BART_cc') and applying a standard BART model on the dataset imputed by the `missForest` package in `R` [Stekhoven and Bühlmann, 2012] ('BART_imp'). Particularly when data are MNAR, the complete cases model may perform significantly better on the observed data as it is both trained and evaluated solely on these cases, while performing poorly on the unobserved due to the failure to account for missingness. In contrast, the joint models or the model trained on imputed data — assuming accurate

[a] All simulated data can be accessed via the missBART GitHub repository.

imputations were made — should excel on cases which are less likely to be observed, in other words those with lower detection probabilities. To highlight this, we compute the out-of-sample RMSEs for subsets of the data based on increasing detection probability thresholds and make comparisons between the different methods. We note that this evaluation technique is not applicable when dealing with real data as the "true" missing values are unobtainable.

Next, we demonstrate the performance of our joint models in the bivariate setting. Specifically, we simulate bivariate data from a multivariate BART model, followed by four different scenarios where the missingness is either MAR or MNAR and arises from a multivariate probit regression model or a multivariate probit BART model. For each scenario, we compare the multivariate versions of missBART1 and missBART2 with 6 other methods: multivariate BART on complete cases ('mvBART_cc'), multivariate BART on the `missForest` imputed dataset ('mvBART_imp'), univariate BART on the complete cases of each response variable ('uniBART_cc'), univariate BART on each `missForest` imputed response ('uniBART_imp'), as well as univariate missBART1 and missBART2 on each partially missing response variable ('uni_missBART1', 'uni_missBART2').

After carrying out 4-fold cross-validation, we compute the RMSE and CRPS of each univariate response for the missing, observed, and combined responses. Additionally, the Frobenius norms on the full dataset are also calculated and shown for all methods, providing a single scalar measure that captures the overall predictive performance of the model. To ensure that the joint models can accurately detect the true underlying missingness mechanisms, we provide posterior intervals for the probit regression parameters $\mathbf{B}$ from missBART1 and calculate the variable importance of the missingness trees in missBART2.

Finally, we conclude the section with three multivariate examples where the covariates are either completely observed, missing under the MAR mechanism with a simple missingness pattern, or MAR with more complex missingness. Five responses are simulated from the multivariate version of the Friedman function, and MNAR missingness is induced in the response using the multivariate probit regression model. The missingness in $\mathbf{X}$ is introduced using the `ampute` function [Schouten et al., 2018] implemented in the `mice` package. Given the high dimensionality of the response variables and the presence of missingness in both the response and covariates — resulting in complex and varied patterns of missingness across all variables — we use 8-fold cross-validation to compare the nine methods. Apart from missBART1 and uni_missBART1 which require covariate imputation prior to model fitting, all other models used in the previous examples include the implementation of the `BARTm` strategy for handling partially-observed covariates in the tree-splitting rules. Additionally, "missBART2_impX" fits a missBART2 model with imputed covariates. For consistency, covariate imputations, where necessary, are made using `missForest`.

We run all models for 5000 burn-in and 5000 post-burn-in iterations. The number of trees in the data model is fixed at 100. In the non-linear and bivariate examples, 20 probit BART trees are used in the missBART2 missing model. In the multivariate examples, however, more response variables and covariates are present in the missingness model, as well as induced missingness in the covariates. Given the more challenging simulation setup, we increase the number of missingness trees to 50 to allow for better model convergence without sacrificing the variable selection capabilities of the probit BART trees.

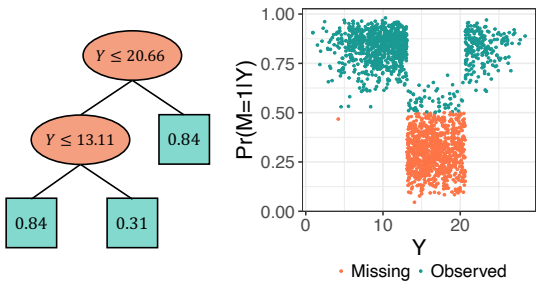
4.2.1 Univariate Examples: Non-Linear Missing Data

In both non-linear missing data examples, we first generate a complete dataset with $n = 2000$ i.i.d. samples using the Friedman function [Friedman, 1991]

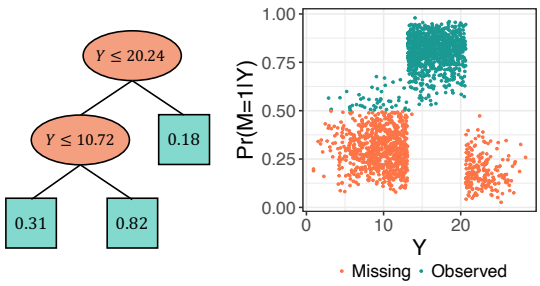
$$Y_i = 10 \sin(\pi X_{i1} X_{i2}) + 20(X_{i3} - 0.5)^2 + 10X_{i4} + 5X_{i5} + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0,1), \tag{4.1}$$

where each $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(5)}$ is simulated independently from a continuous Unif$(0,1)$ distribution. Missingness is then induced through a single tree.

Figure 4.1 shows the missing tree structures of the u-shape and n-shape missing scenarios and their resulting simulated detection probabilities, $\Pr(M_i = 1 \mid Y_i)$, against the true $\mathbf{Y}$ values. All observations with detection probabilities below 0.5 are designated as missing and *vice versa*. In the first example, roughly 55.1% of the data remained observed, with missing values of $\mathbf{Y}$ occurring more frequently around the mid-range of the data. In the second scenario, 50.2% of the data were observed, and missing values predominantly occurred at the extreme ends of the data range.



(a) u-shape: Most of the data between $[13.11, 20.66]$ are missing.

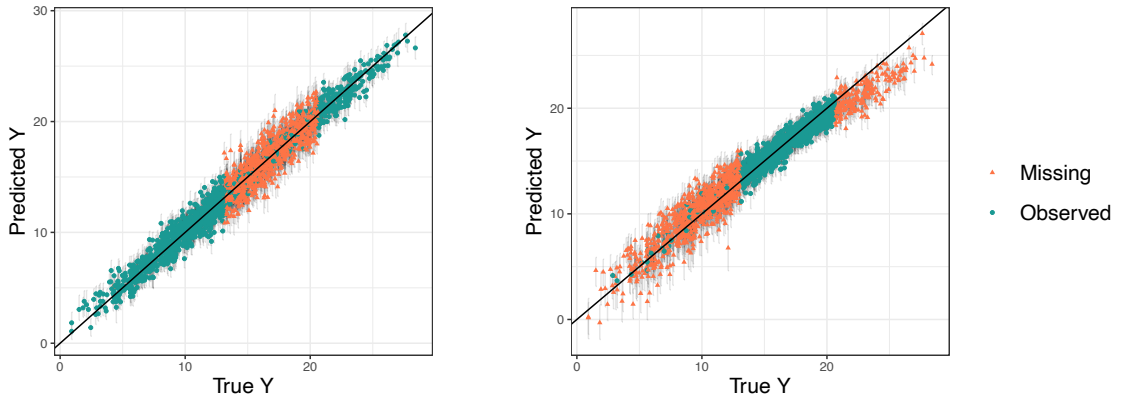


(b) n-shape: Most of the data outside the range $[10.72, 20.24]$ are missing.

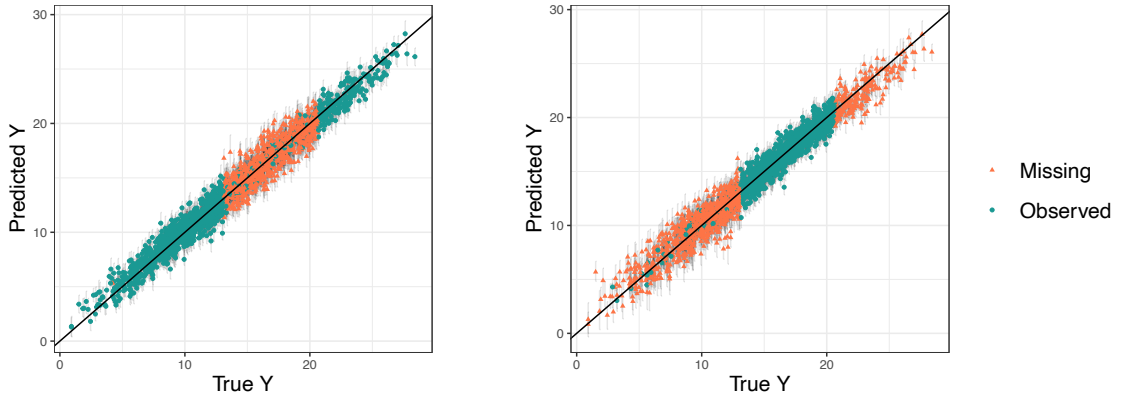
Figure 4.1: Missingness trees with detection probabilities in the terminal nodes (left) and plots of detection probabilities against true $\mathbf{Y}$ values (right) for the u-shape (top) and n-shape (bottom) missing examples.

Figure 4.2 shows the out-of-sample predictions for the observed cases and posterior imputations for the missing cases against their true simulated values obtained from both

models, along with their 95% prediction and posterior intervals. In the case of u-shape missingness, both models perform well in fitting the observed values as well as making accurate posterior imputations for the missing values. However, in the n-shape missingness scenario, while missBART2 maintains accuracy, missBART1 struggles to capture the tail ends of the missing data, particularly in the upper tail where no values are observed. This underlines the limitations of missBART1 in capturing complex non-linear structures while also highlighting the strong capabilities of missBART2 in addressing these challenges.



(a) Out-of-sample predictions and imputations from missBART1 for the u-shape missingness pattern (left) and n-shape pattern (right). The model struggles to capture the upper end of the missing data in the n-shape scenario.

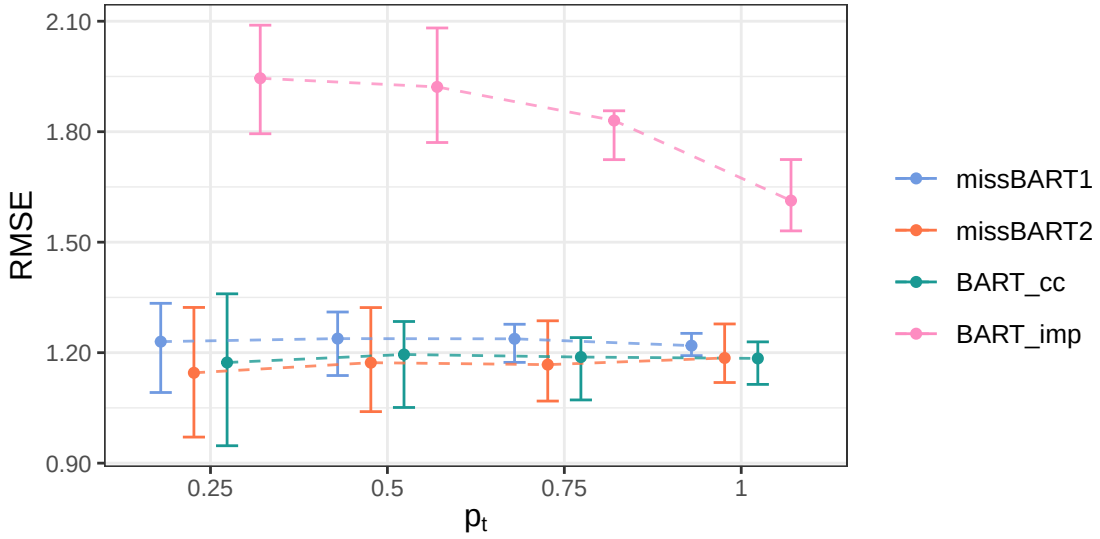


(b) Out-of-sample predictions and imputations from missBART2 for the u-shape missingness pattern (left) and n-shape pattern (right). The model performs well in both scenarios, closely matching the true simulated values.

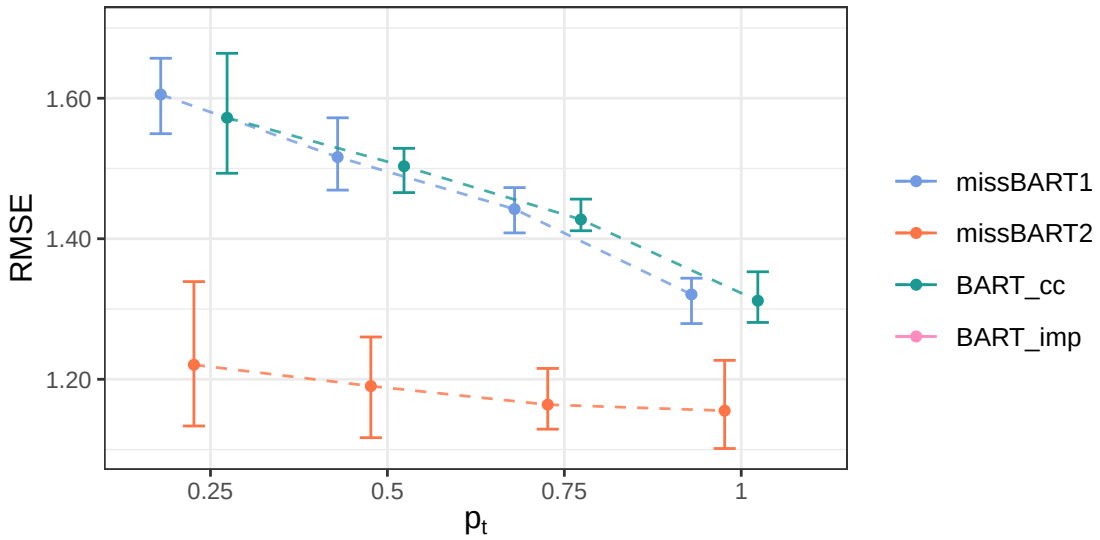
Figure 4.2: Out-of-sample predictions and posterior imputations from missBART1 (a) and missBART2 (b) in the u-shape and n-shape non-linear missing data examples. The left subfigures show performance on the u-shape missingness pattern, while the right subfigures show performance on the n-shape pattern. In general, missBART2 performs well in both cases, while missBART1 struggles with predicting the upper range of the missing data in the n-shape scenario. Vertical lines depict the 95% prediction and posterior intervals.

Next, we compare the performances of four different approaches for handling missing data in the context of predictive data analysis, namely missBART1, missBART2, BART_cc and BART_imp. For each scenario, a 4-fold cross-validation approach was em-

ployed by randomly partitioning the dataset into train and test sets such that $n_{train} = 1500$ and $n_{test} = 500$ for each fold. We calculate the out-of-sample RMSEs for observations based on varying levels of detection probability thresholds, p_t, i.e. $\Pr(M_i = 1 \mid Y_i) \leq p_t$, where $p_t = \{0.25, 0.5, 0.75, 1\}$. The results are shown in Figure 4.3.



(a) Out-of-sample RMSE for u-shaped missingness across varying detection probabilities p_t. The performance of missBART1, missBART2, BART_cc, and BART_imp are shown, with BART_imp demonstrating poor performance across all levels of p_t.



(b) Out-of-sample RMSE for n-shaped missingness across varying detection probabilities p_t. missBART2 outperforms others, while BART_imp is excluded due to poor performance.

Figure 4.3: Out-of-sample RMSE values from 4-fold cross-validation for data with detection probabilities below thresholds p_t. Results for BART_imp are excluded from (b) due to poor performance.

From Figure 4.3a, we note that BART_imp, which imputes the data using `missForest` prior to fitting a BART model, consistently underperforms across all detection probability thresholds in both scenarios. In Figure 4.3b, the RMSE values from BART_imp are omitted due to poor performance (between 2.39 and 3.53). This is most likely explained by the inaccurate initial imputations produced by `missForest`, resulting in the BART model being trained on incorrect values, compromising the overall model fit.

In the u-shape scenario, missBART1, missBART2, and BART_cc perform similarly across all thresholds, implying little difference in the out-of-sample prediction and imputation accuracy, regardless of the detection probabilities. In contrast, missBART2 dominates the other models in the n-shape scenario, while missBART1 has slightly worse performance compared to BART_cc. We observe that, as p_t increases, the RMSE decreases, indicating that the models perform better on data which are more likely to be observed. This is unsurprising, as most missing values lie further away from the observed values, making it exceptionally challenging for inappropriate models to capture the data's true extremes.

4.2.2 Bivariate Examples: Missingness Under MAR and MNAR

We now demonstrate the performance of missBART1 and missBART2 on simulated bivariate data under MAR and MNAR scenarios. We simulate $n = 2000$ i.i.d. bivariate observations from a multivariate BART model with 5 covariates. Using the same complete dataset, we create four different missing scenarios by varying the missing data mechanism — MAR or MNAR — and the underlying missing data model — multivariate probit regression or multivariate probit BART. In MAR scenarios, only $\mathbf{X}$ is used to simulate missingness. In contrast, for MNAR scenarios, the missingness probabilities only depend on $\mathbf{Y}$. A summary of the simulation details is shown in Table 4.1 below.

		MAR 1	**MAR 2**	**MNAR 1**	**MNAR 2**
Data Model	n	2000 $(n_{train} = 1500, n_{test} = 500)$			
	p	2			
	q	5			
	# Data trees	8			
Missingness Model	Model	Probit regression	Probit BART	Probit regression	Probit BART
	Model covariates	$\mathbf{X}$ only	$\mathbf{X}$ only	$\mathbf{Y}$ only	$\mathbf{Y}$ only
	# Missing trees	—	3	—	5
	Observed proportions of $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}$	(57.10%, 78.05%)	(86.50%, 56.10%)	(66.00%, 50.45%)	(58.05%, 75.15%)

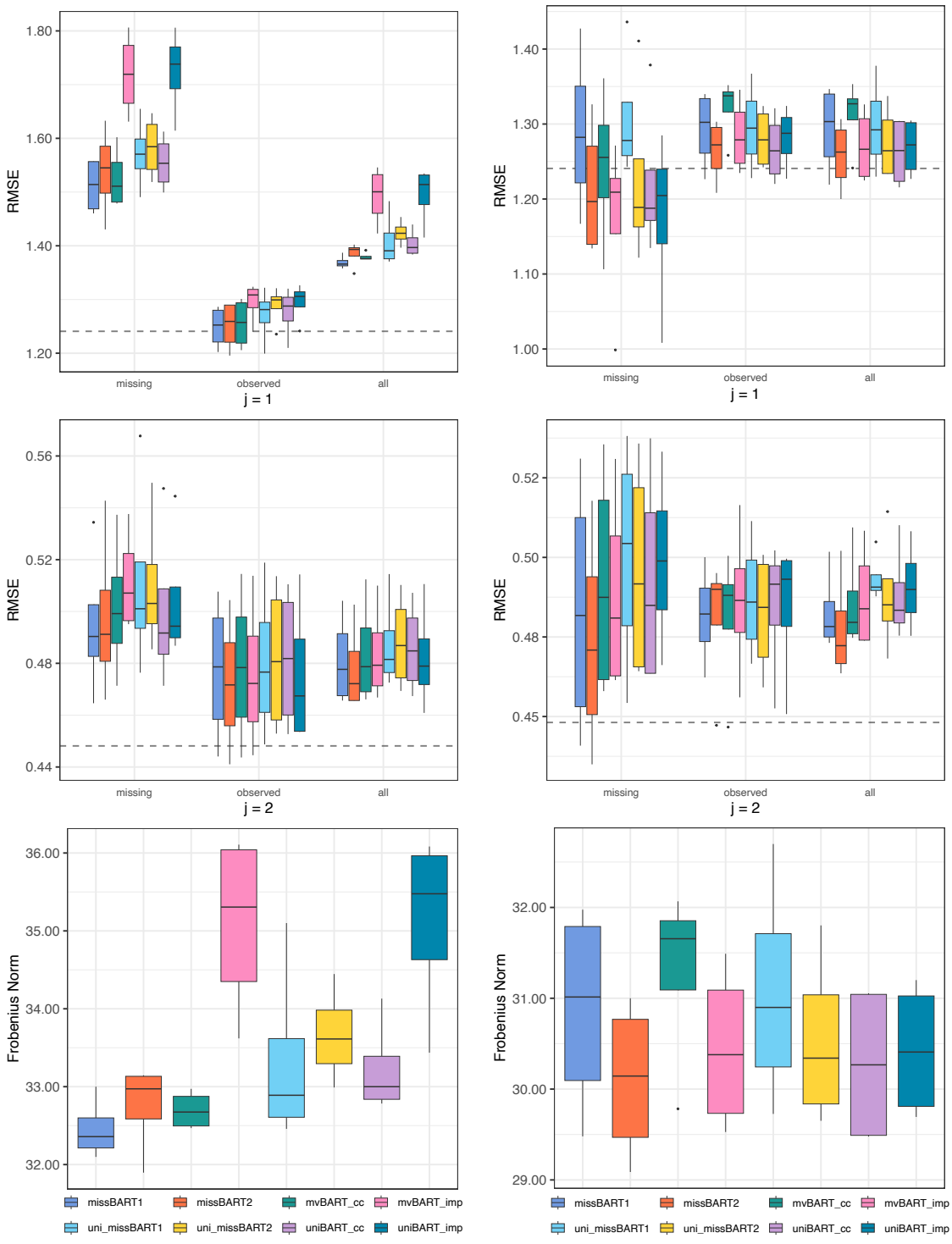
Table 4.1: Simulation recipes for the bivariate simulation studies. The complete dataset is consistent across all 4 scenarios, while the missingness model follows a multivariate probit regression or probit BART model. The final row shows the proportions of observed cases per response variable $(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ under each scenario.

We make comparisons between the eight models through 4-fold cross-validation. For mvBART_imp and uniBART_imp, missing $\mathbf{Y}$ imputations were first obtained by passing the whole dataset $(\mathbf{X}, \mathbf{Y})$ to the `missForest` function before splitting $\mathbf{Y}$ into train/test sets. The default settings of `missForest` were used.

To evaluate imputation accuracy, prediction accuracy, and overall performance, we separately compute the RMSE values of the missing responses, the observed responses, and the combined dataset. Under MAR missingness, we expect the joint models to perform at least as well as the complete-case or `missForest`-imputed models for both missing and observed responses as the missing data mechanism can be ignored and only the regression model is required. In the MNAR case, we anticipate a greater disparity between these models, particularly for the missing responses, as complete-case and `missForest`-imputed models fail to account for any relationship between the responses and their corresponding missingness status. In addition to the RMSE, we also calculate the Frobenius norms of the multivariate dataset as an overall performance evaluation metric.

The results for the MAR and MNAR examples are presented in Figures 4.4 and 4.5 respectively. In each case, the first two panels show the out-of-sample RMSEs for the missing, observed and combined responses for $\mathbf{Y}^{(j)}$ $(j = 1, 2)$, while the last panel shows the Frobenius norms obtained across the multivariate responses. In Figure 4.5b, some results of mvBART_cc, mvBART_imp, and uniBART_imp were omitted due to overly high RMSEs.

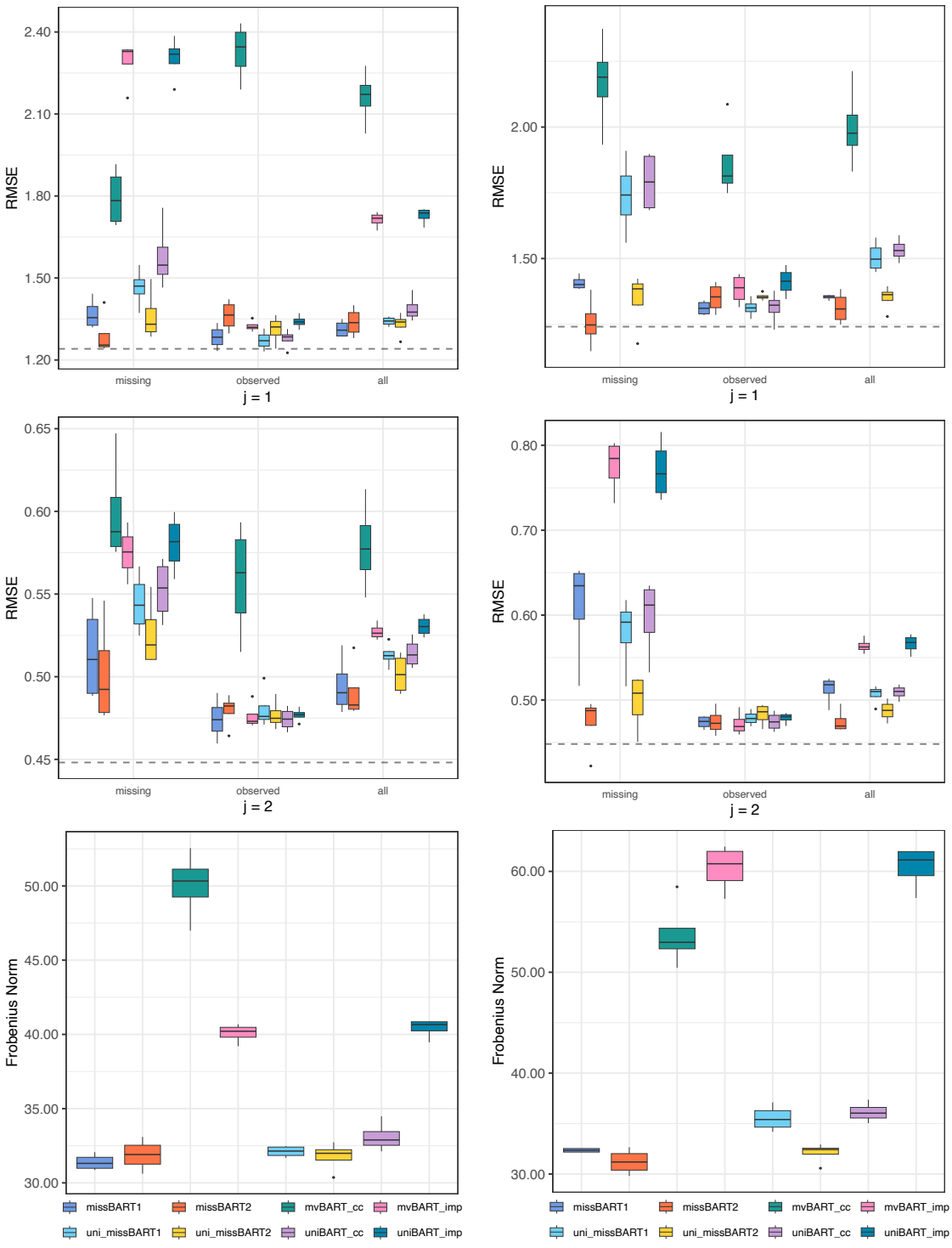
To assess the calibration and accuracy of the predicted distributions, we also compute and compare the CRPS values of the 8 models. Similar to the RMSE, we report the CRPS values for the j^{th} response, separately for the missing responses, the observed responses, and the combined dataset. Under MAR missingness, we expect the joint models to perform at least as well as the complete-case or `missForest`-imputed models for both missing and observed responses. In the MNAR case, we expect the joint models to perform better, particularly for the missing responses, as complete-case and `missForest`-imputed models fail to account for any relationship between the responses and their corresponding missingness status. The results, shown in Figures 4.6 and 4.7, align with these expectations.



(a) MAR 1

(b) MAR 2

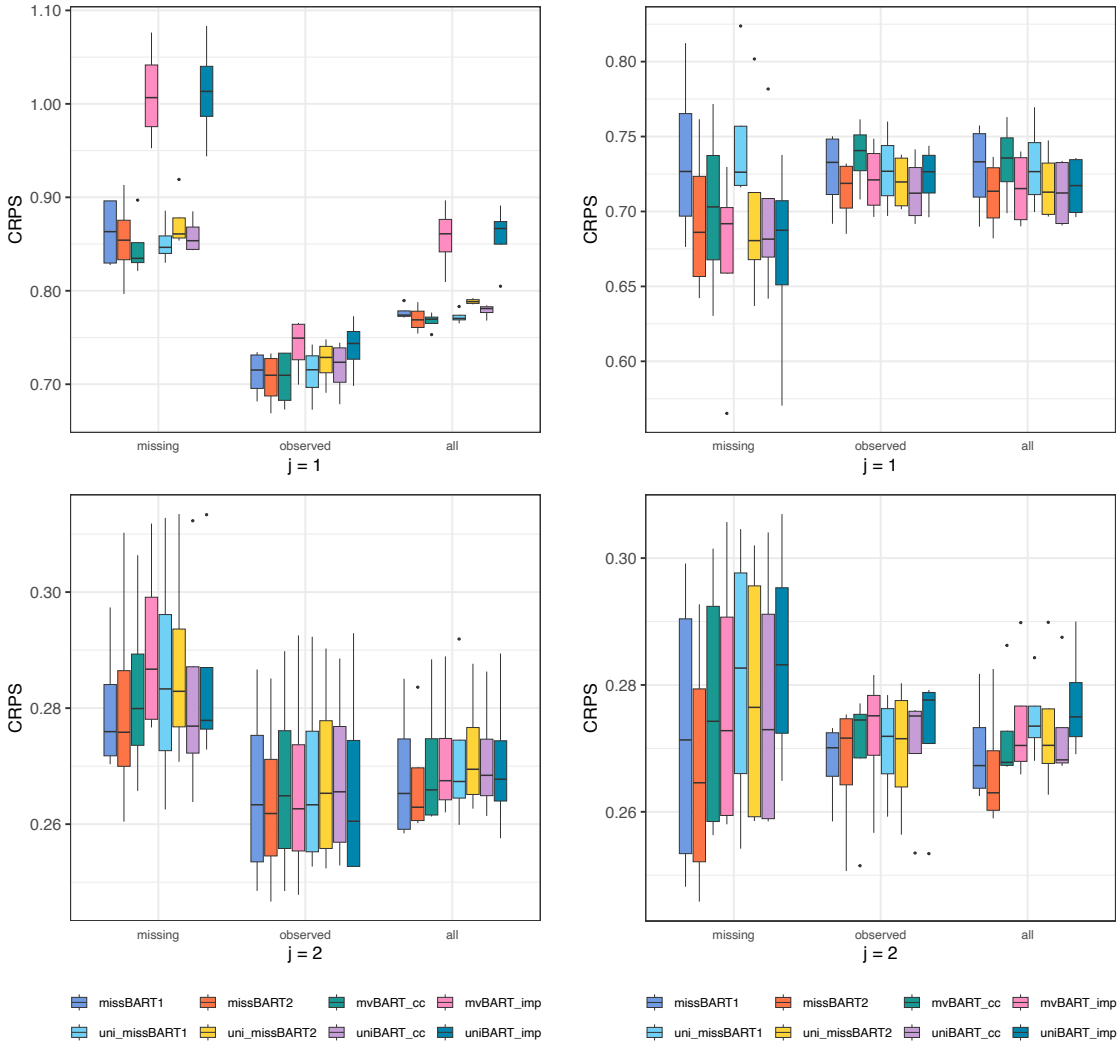
Figure 4.4: RMSEs (top 2 panels) and Frobenius norms (bottom panels) comparing 8 different models in MAR 1 (a) and MAR 2 (b) scenarios. RMSEs are calculated for missing, observed, and all data points, while Frobenius norms provide a single measure of overall model performance. Dashed lines in the top 2 panels represent the true simulated residual standard deviations.



(a) MNAR 1

(b) MNAR 2

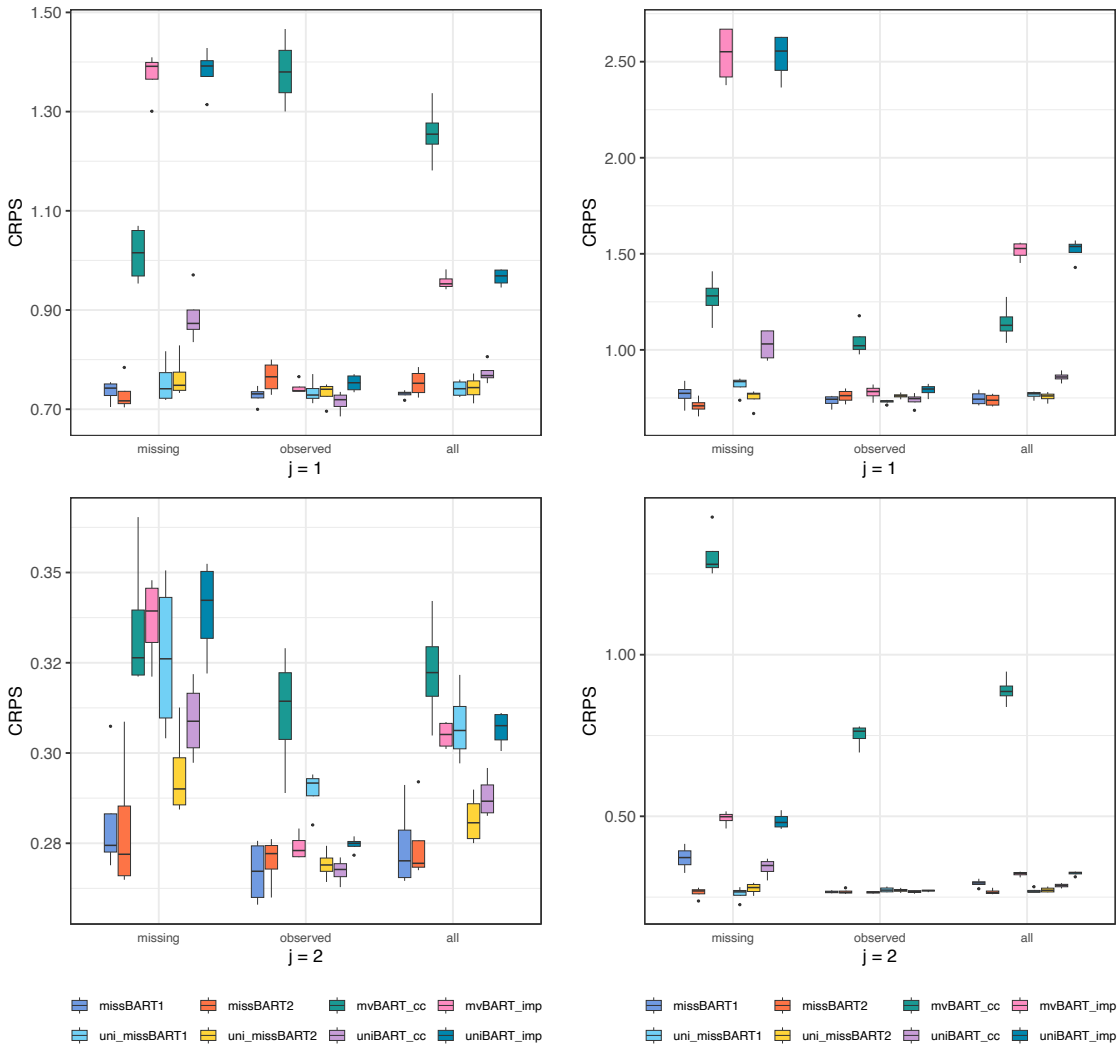
Figure 4.5: RMSEs (top 2 panels) and Frobenius norms (bottom panels) comparing 8 different models in MNAR 1 (a) and MNAR 2 (b) scenarios. RMSEs are calculated for missing, observed, and all data points, while Frobenius norms provide a single measure of overall model performance. Some results of mvBART_cc, mvBART_imp, and uniBART_imp were omitted due to their poor performance in terms of RMSE.



(a) MAR 1

(b) MAR 2

Figure 4.6: CRPS for 8 different models in MAR 1 (a) and MAR 2 (b) scenarios. CRPS values are calculated for missing, observed, and all data points.



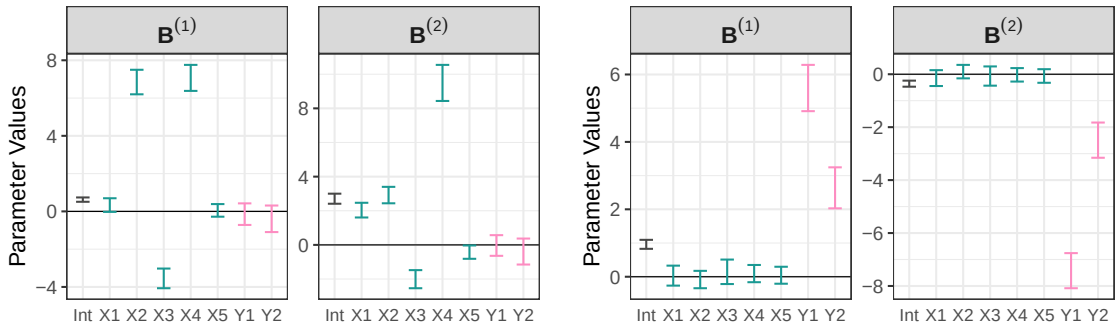
(a) MNAR 1

(b) MNAR 2

Figure 4.7: CRPS for 8 different models in MNAR 1 (a) and MNAR 2 (b) scenarios. CRPS values are calculated for missing, observed, and all data points.

The results obtained from this simulation study were mostly as anticipated. Similar to the non-linear examples, the `missForest` imputed models demonstrate poor performances across MAR 1, MNAR 1, and MNAR 2, which can again be attributed to the inaccurate imputations obtained prior to model fitting, leading to erroneous results. Overall, missBART2 demonstrated superior performances in both the MAR and MNAR scenarios, particularly in making accurate posterior imputations for the missing responses.

In terms of recovering the underlying missing mechanism, we look at the posterior intervals of $\mathbf{B}$ from missBART1. Figure 4.8 shows the 95% posterior intervals obtained from missBART1 for MAR 1 and MNAR 1. Each panel shows the intervals for coefficients of the intercept, five covariates, and two responses within the probit regression model. We observe that the error bars representing coefficients in $\mathbf{B}_Y$ overlap with 0 when the data are MAR, correctly indicating the absence of a relationship between missingness and the response variables. In contrast, the intervals do not overlap with 0 in the MNAR scenario, successfully capturing the non-ignorable missing data mechanism.

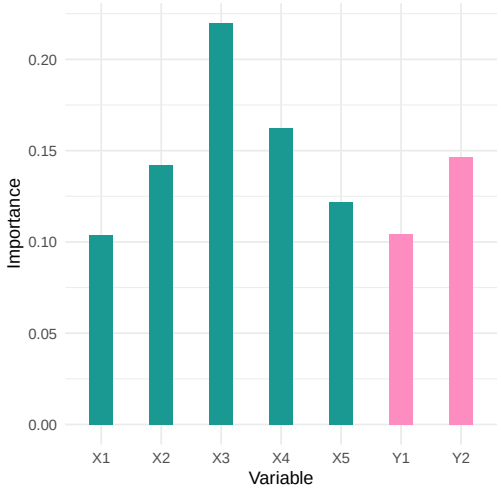


(a) MAR 1: The posterior intervals of $\mathbf{B}_Y$ overlap with 0, capturing the true MAR mechanism.

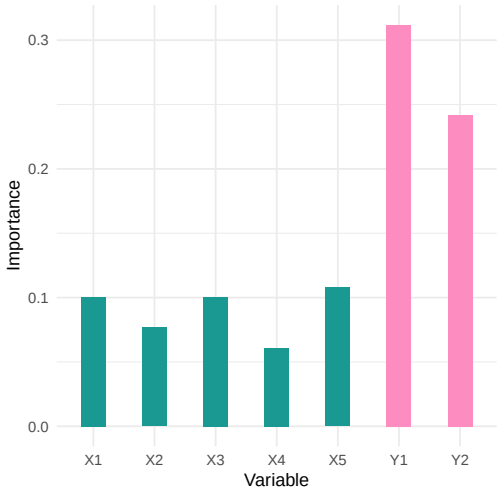
(b) MNAR 1: The posterior intervals of $\mathbf{B}_Y$ do not contain 0, capturing the true MNAR mechanism.

Figure 4.8: 95% posterior intervals of $\mathbf{B}$ from missBART1 for MAR 1 and MNAR 1. In both scenarios, missBART1 accurately captured the true underlying missingness mechanism. Posterior intervals of $\mathbf{B}_Y$ overlapped with 0 when the true missing mechanism was MAR and did not contain 0 in the MNAR case.

To assess the recovery of the missing data mechanism from missBART2, we investigate the variable importance of the missingness trees, calculated as the average number of uses of each variable in the splitting rules of the missingness trees over 5000 post-burn-in iterations. Variable importance for MAR 2 is shown in Figure 4.9a and in Figure 4.9b for MNAR 2. For MNAR 2, most splits are made on $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$, while only a small proportion of splits are made on the covariates. In MAR 2, $\mathbf{X}^{(3)}$ is most frequently used, followed by $\mathbf{X}^{(4)}$.



(a) MAR 2: $\mathbf{X}^{(3)}$ was used most often in the tree splitting rules, followed by $\mathbf{X}^{(4)}$.



(b) MNAR 2: The majority of splits are made on $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$.

Figure 4.9: Variable importance from missingness trees in missBART2 for MAR 2 (a) and MNAR 2 (b). In MAR 2, the covariates $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$ were frequently used in the splitting rules of the missingness trees. In MNAR 2, there is a clear distinction between the importance of response variables and covariates.

4.2.3 Multivariate Examples: MNAR Response, MAR Covariates

While there are no missing covariates in the *global Amax* dataset, we present three simulated scenarios — MNAR_amp0, MNAR_amp1, and MNAR_amp2 — with MNAR missingness in the multivariate responses and different missingness in the covariates to demonstrate the performance of missBART1 and missBART2 in the presence of ignorable missing covariates and non-ignorable missing outcomes.

Similar to the bivariate examples, we use the same dataset for all three examples, this time keeping the missingness in $\mathbf{Y}$ consistent across all examples while varying the missingness in $\mathbf{X}$. We first create a dataset of $n = 2000$ i.i.d. observations with $p = 5$ responses from the multivariate version of the Friedman function:

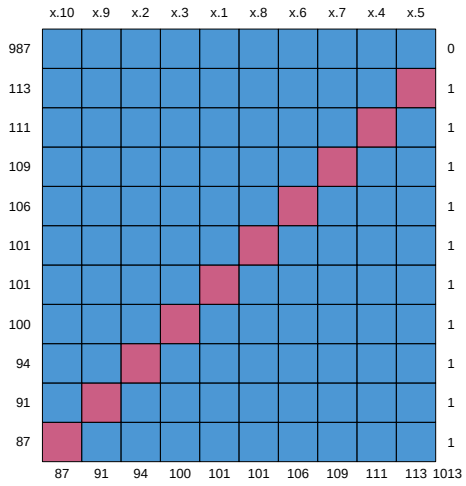
$$\mathbf{Y}_i = \boldsymbol{\xi}_1 \sin(\pi X_{i1} X_{i2}) + \boldsymbol{\xi}_2 (X_{i3} - 0.5)^2 + \boldsymbol{\xi}_3 X_{i4} + \boldsymbol{\xi}_4 X_{i5} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon) \tag{4.2}$$

where $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_4$ are p-dimensional vectors of coefficients simulated from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\xi)$. While the Friedman function only requires five covariates, we include five extra non-informative ones. Each covariate is indepndently drawn from a continuous Unif$(0, 1)$ distribution.

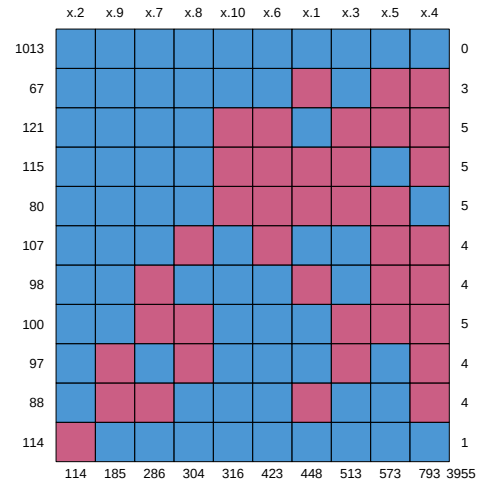
In the previous sections, missingness either depended strictly on $\mathbf{X}$ for the MAR cases or $\mathbf{Y}$ for MNAR. Here, the missing model follows a multivariate probit regression model with non-zero coefficients for $(\mathbf{X}, \mathbf{Y})$, i.e. missingness in $\mathbf{Y}$ depends on both the observed and unobserved variables. However, to ensure that the data are MNAR, we enforce

a stronger relationship between the missingness probabilities and $\mathbf{Y}$ while keeping the coefficients of $\mathbf{X}$ close but not equal to 0.

As a baseline, the complete set of covariates is used in the first example, MNAR_amp0, so missingness is only present in the responses. For MNAR_amp1 and MNAR_amp2, MAR missingness is introduced into $\mathbf{X}$ using the function `ampute` [Schouten et al., 2018] from the `mice` package, which allows the specification of the underlying missing data mechanism, the overall proportion of missingness in $\mathbf{X}$, as well as the missing data pattern [Van Buuren, 2018]. For both examples, we use the default setting of `"MAR"` for the missing data mechanism in the covariates and 0.5 for the associated proportion of missingness. By default, the missing data pattern is diagonal such that only one variable is missing in each row $\mathbf{X}_i$. We use this setting in MNAR_amp1, keeping the missingness proportion across each covariate relatively consistent (between 4.35% and 5.65% missing). In MNAR_amp2, we allow for more complicated patterns of covariate missingness with varying proportions of missingness (between 5.70% and 39.65%). The missingness patterns for MNAR_amp1 and MNAR_amp2 are shown in Figure 4.10a and Figure 4.10b below.



(a) Missingness patterns in $\mathbf{X}$ for MNAR_amp1. The MAR missingness is diagonal such that only one variable is missing in each row of the covariates.



(b) Missingness patterns in $\mathbf{X}$ for MNAR_amp2. The patterns are more complicated and missingness proportions vary between 5.70% and 39.65%.

Figure 4.10: Missingness patterns in the MAR covariates, $\mathbf{X}$, for MNAR_amp1 and MNAR_amp2. Only one covariate is missing for each $\mathbf{X}_i$ in MNAR_amp1, while MNAR_amp2 has more varied and complicated missingness patterns.

We use 8-fold cross-validation to make comparisons between the models described in Section 4.2.2 and an additional model missBART2_impX, the missBART2 model with prior covariate imputation. For MNAR_amp0, since the covariates are fully observed and no covariate imputation is necessary, missBART2_impX is equivalent to missBART2 and thus excluded from the study. Note that, while missBART1, missBART2_impX, and uni_missBART1 are trained and tested on the `missForest` imputed covariates, all other

models include the `BARTm` method for handling missing covariates within the BART trees. The out-of-sample Frobenius norms of the three examples are shown in Figure 4.11 below.

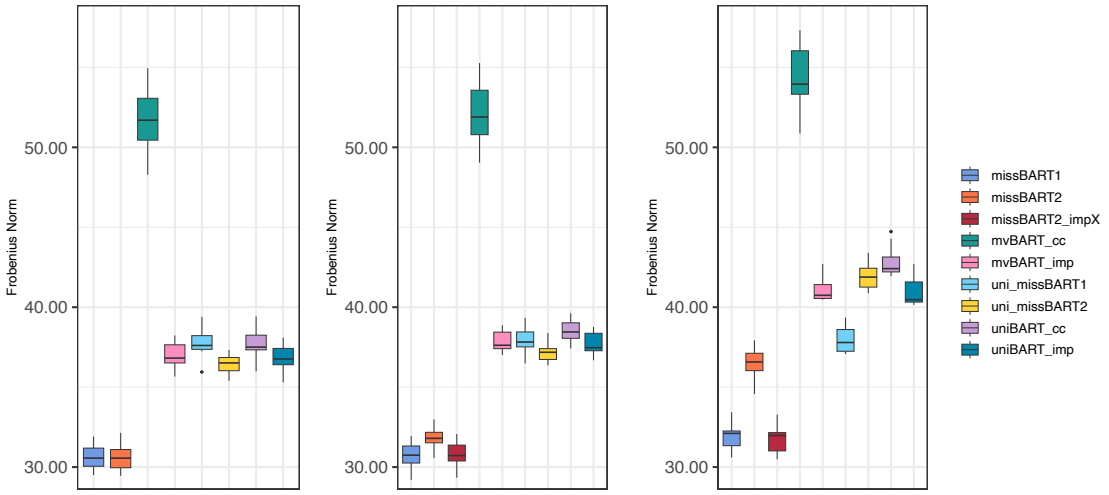


Figure 4.11: Out-of-sample Frobenius norms for MNAR_amp0 (left), MNAR_amp1 (middle), and MNAR_amp2 (right).

The out-of-sample RMSEs for MNAR_amp0, MNAR_amp1, and MNAR_amp2 are shown in Figure 4.12, separately for the missing, observed, and combined univariate responses. The joint models once again demonstrate robust performance when dealing with multivariate non-ignorable response data, both in the presence and absence of missing covariates. When the covariates are fully observed, missBART1 and missBART2 yield comparable results, outperforming all other methods. However, the introduction of missingness in the covariates generally leads to a decline in model performance.

In the MNAR_amp1 scenario with diagonal missing covariates, both missBART1 and missBART2 show increased Frobenius norms compared to the fully observed case, with missBART1 slightly outperforming missBART2. However, when covariate imputation is applied prior to model fitting, missBART2_impX emerges with the strongest overall performance. In the more complex MNAR_amp2 scenario, the disparity in performance between missBART1, missBART2, and missBART2_impX becomes more distinct, yet missBART2_impX maintains the best overall performance.

While missBART2 with covariate imputations demonstrated the strongest overall performance, it is important to note that, as shown in Kapelner and Bleich [2015], different missing covariate scenarios can lead to varying comparative performances between `BARTm` and BART or random forests with `missForest` covariate imputations. Additionally, missBART2_impX is essentially a two-step process that requires covariate imputations followed by model fitting, whereas missBART2 with the incorporation of the `BARTm` technique offers a more straightforward approach to handling missing covariates.

Figure 4.12: RMSEs for multivariate simulated scenarios with missingness in the covariates.

In the simulation studies, convergence of the MCMC sampler was assessed through visual inspection of trace plots. However, as the number of parameters in both missBART1 and missBART2 may scale with the number of responses, covariates, and the amount of missingness, it is not always feasible to monitor all sampled quantities. Instead, we focused on certain key parameters, such as the diagonal entries of the precision matrix $\mathbf{\Omega}$ and selected entries of the missing responses $\mathbf{Y}^{\text{mis}}$. Across all simulation settings, we used a default choice of 5000 burn-in iterations followed by 5000 post-burn-in samples, with 100 regression trees and either 20 or 50 missingness trees, depending on the number of missing responses and the complexity of the missingness structure. These settings yielded satisfactory convergence in the investigated scenarios. However, convergence may vary depending on several factors, such as the number of covariates, responses, missingness proportions, and overall sample size. Therefore, we recommend that practitioners assess convergence using standard MCMC diagnostics and consider adjusting the number of iterations, the number of regression and/or missingness trees, as well as other tuning parameters based on the complexity of the application.

4.3 Application: *global Amax*

The *global Amax* data comprises a rich set of environmental covariates, including 20 soil and 26 climate variables, alongside 5 continuous responses reflecting various leaf photosynthetic traits and rates: light-saturated photosynthetic rate (*Aarea*), stomatal conductance (*Gs*), leaf nitrogen (*Narea*), leaf phosphorus (*Parea*), and specific leaf area (*SLA*).

Originating from Maire et al. [2015], the authors employed various linear and parametric methods such as mixed-effects regression models, variation partitioning, redundancy analysis, and path analysis to quantify the influence of climate and soil properties on each individual leaf photosynthetic trait. The study identified key environmental variables, particularly soil pH (*pH*), moisture index (*Miq*), and available soil phosphate content (*Pavail*), as major influences on photosynthetic traits.

However, the methods employed by Maire et al. [2015] lack the flexibility of non-parametric BART models in capturing non-linear relationships, and do not address the partially overlapping nature of missingness in the multivariate response. In fact, there was minimal discussion of how missingness was addressed, suggesting an implicit reliance on the ignorability assumption. For further details on the methods carried out in this previous study, see Appendix S6 of Maire et al. [2015].

4.3.1 Results

We now apply missBART1 and missBART2 to the *global Amax* data. Prior to model fitting, a log transformation is applied to each response variable to account for right-skewness. As in Section 4.2.3, we use 5000 burn-in and 5000 post-burn-in iterations, 100 regression trees, and 50 missingness trees.

Figures 4.13 and 4.14 show the model predictions for the observed data against their true log-transformed values for missBART1 and missBART2, respectively, along with vertical error bars representing their 95% prediction intervals. Rug plots on the y-axis display the posterior means of the missing value imputations, providing insight into both models' imputations for the missing data.

Both models demonstrate strong predictive performances on the observed data, indicating their robustness in making accurate predictions for observed cases, while simultaneously imputing the missing values. Notably, despite similar fits on the observed data, the imputations differ between the models. While most imputations from missBART2 remain closely aligned with the range of observed values, missBART1 shows more drastic deviations, especially for *SLA*, *Parea*, and *Gs*.

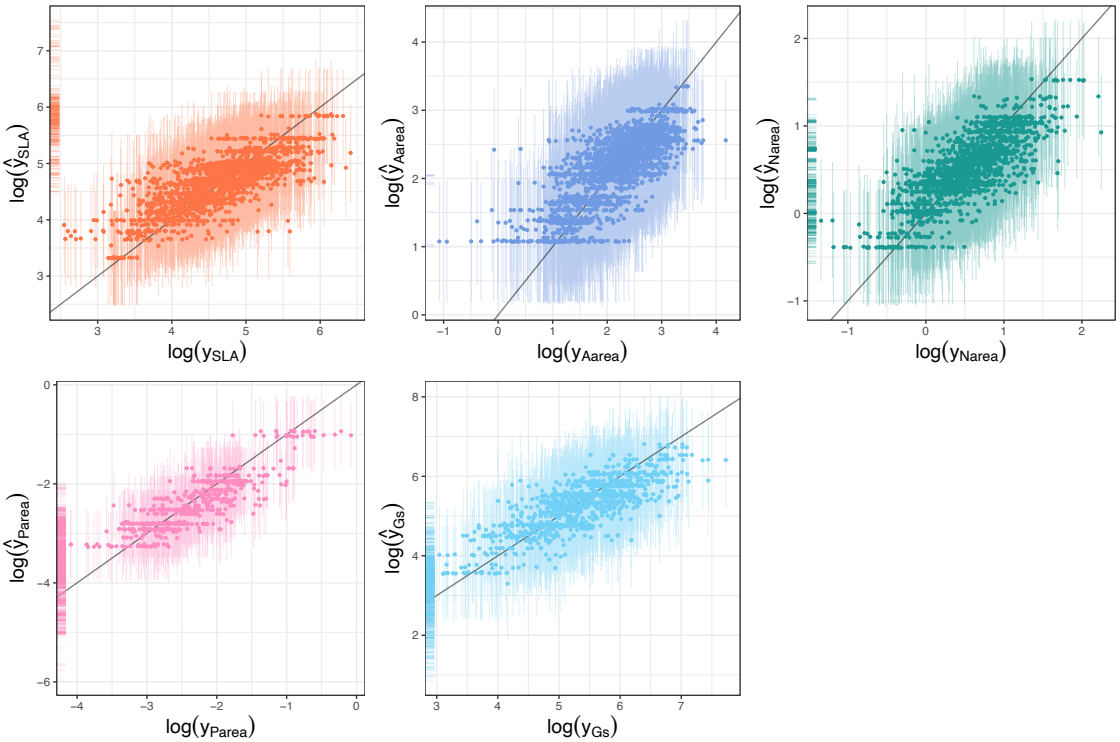


Figure 4.13: missBART1 predictions for the observed data against their true log-transformed values. Vertical error bars represent the 95% prediction intervals for the observed data. Rug plots on the $y-$axes show the posterior means of the missing data imputations. Notably, the imputations for *SLA*, *Parea*, and *Gs* deviate from the range of the observed data.

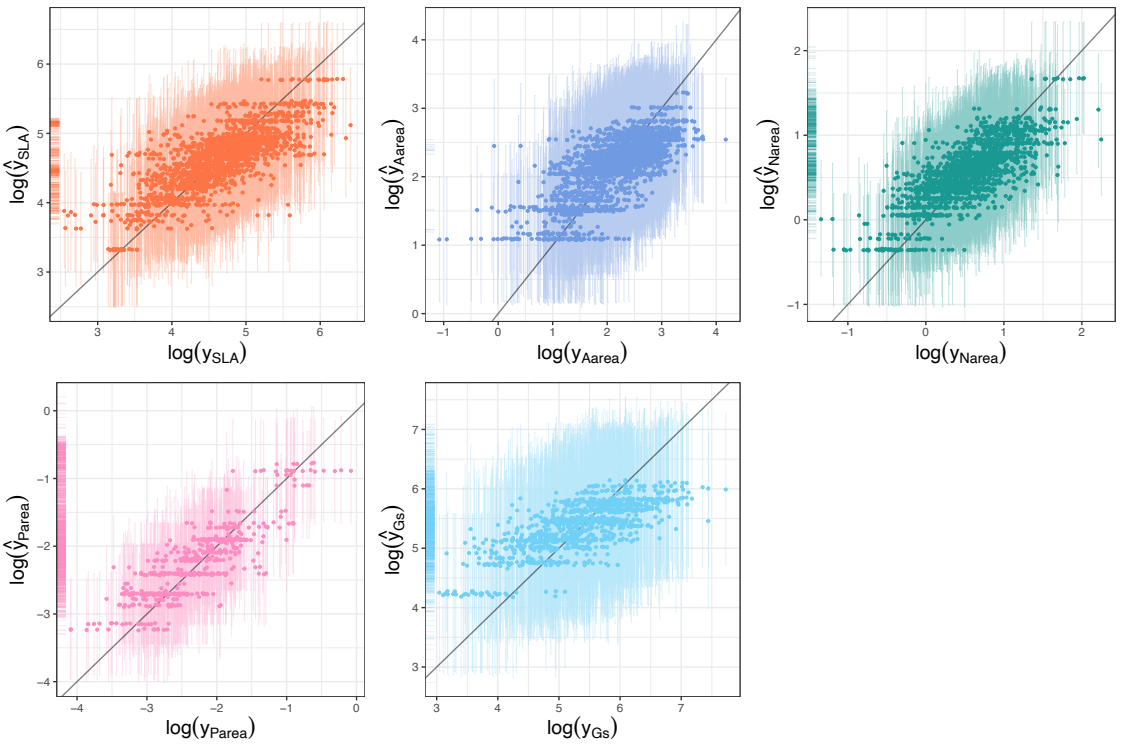


Figure 4.14: missBART2 predictions for the observed data against their true log-transformed values. Vertical error bars represent the 95% prediction intervals for the observed data. Rug plots on the $y-$axes show the posterior means of the missing data imputations. The imputations mostly lie within the range of the observed data, aside from a few values for *Parea* and *Gs*.

Figure 4.15 shows the 95% posterior intervals of $\mathbf{B}_Y$ obtained from missBART1. In each panel, the error bars indicate the degree to which the missingness of each response variable is influenced by the values of the 5 response variables. The coefficients for *Narea* remain consistently non-zero (do not overlap with zero) across multiple panels, apart from that for Gs, indicating that *Narea* strongly influences the missingness probabilities for other responses. Aside from *Aarea*, the missingness of each response variable is often explained by its own values. Specifically, higher values of *SLA* are more likely to be missing, while higher values of *Narea*, *Parea*, and *Gs* are more likely to be observed. This is consistent with the imputations shown in Figure 4.13. Despite *Parea* exhibiting the highest level of missingness, only the error bars for *Narea* and *Parea* do not overlap with 0 in the missingness for *Parea*, indicating that its missingness is primarily influenced by itself and *Narea*. Overall, most of the intervals for $\mathbf{B}_Y$ do not overlap with 0, showing a strong presence of MNAR missingness in the *global Amax* data.

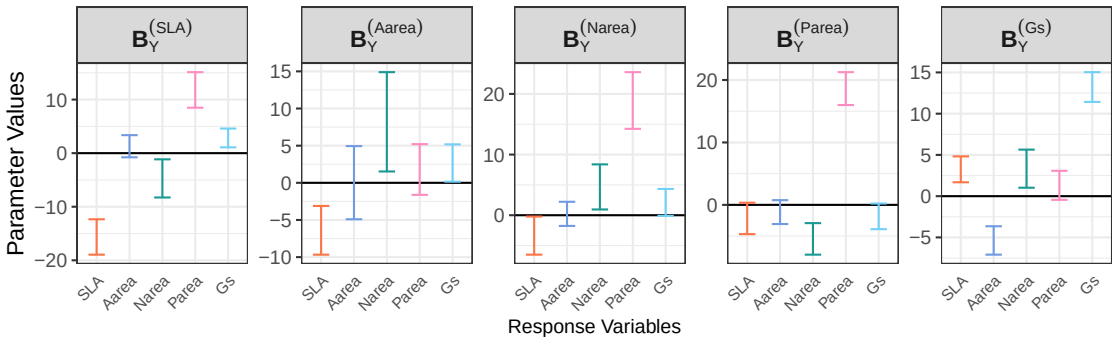
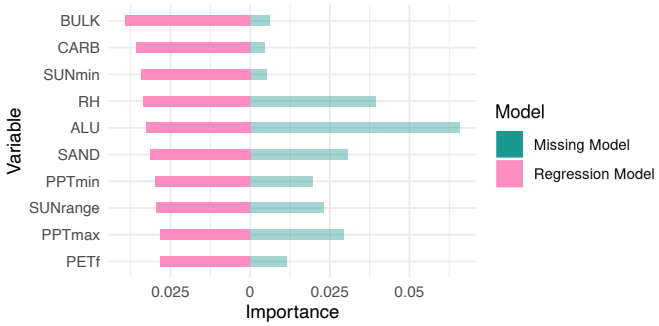
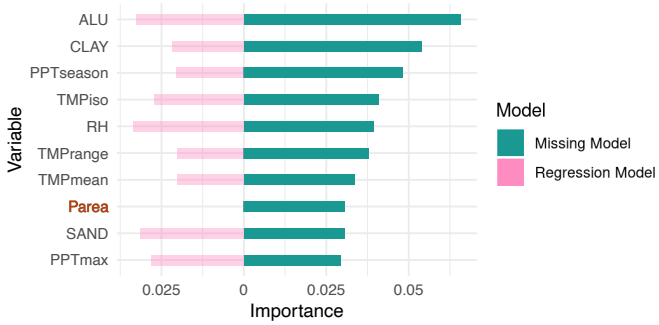


Figure 4.15: 95% posterior intervals of $\mathbf{B}_Y$, the probit regression coefficients associated with the response variables from missBART1. Most intervals do not overlap with 0, indicating a strong presence of MNAR missingness.

From missBART2, we obtain the variable importance from both the regression and missingness models. This is shown in Figure 4.16 below. In Figure 4.16a, the 10 most important variables from the regression trees are displayed, along with the importance measures from both the regression and missingness trees. While bulk density (*BULK*), calcium carbonate content (*CARB*), and mean monthly fractional sunshine duration (*SUNmin*) were the variables most frequently used in the regression trees for predicting the responses, they showed little importance in explaining the missingness of the data. In contrast, Figure 4.16b shows that the most influential variables for the missingness model are exchangeable aluminium percentage (*ALU*), clay content (*CLAY*), and the seasonality of precipitation (*PPTseason*). Of the 5 responses variables, *Parea* is the only variable in the top 10 most important variables of the missingness trees, while the others are among the 20 least important variables. We also note that *Parea* has no importance in the regression trees, as response variables only contribute to the missingness model by construction. The covariates *ALU*, relative humidity (*RH*), sand content (*SAND*), and maximum monthly precipitation (*PPTmax*) are among the top 10 most influential variables for predicting the responses as well as explaining the missingness.



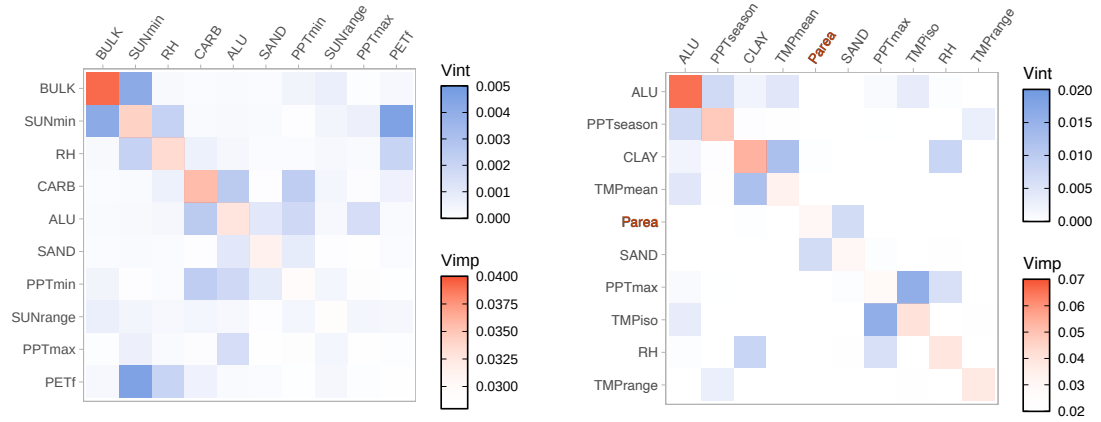
(a) Top 10 variables from the regression trees.



(b) Top 10 variables from the missingness trees.

Figure 4.16: Comparisons of the top 10 important variables in the regression and missingness models from missBART2. On the left, the top 10 variables from the regression model are shown alongside their corresponding importance in the missingness model. On the right, the top 10 variables from the missingness model are displayed and compared with their respective importance in the regression model. Only *RH*, *SAND*, and *PPTmax* are common to both sets. This highlights differences in how variables influence both the missingness mechanism and the regression model's predictions. In panel (b), of the 5 responses, *Parea* is the only one among the top 10 important variables, while the other 4 fall outside the top 40. *Parea* has no importance in the regression trees as it is a response variable and can only contribute to the missingness model by construction.

In addition to variable importance, variable interactions in the regression and missingness trees can be measured by counting the number of times two variables appear consecutively along the same branch in the trees. Using visualisation techniques adapted from Inglis et al. [2022], the average interaction effects of the top 10 variables from the regression and missingness trees are shown using heat maps in Figure 4.17 below, with variable importance shown on the diagonals and variable interactions on the off-diagonals. In Figure 4.17a, *BULK* and potential evapotranspiration (*PETf*) show strong interactions with *SUNmin* implying that *BULK* and *PETf* are often used in conjunction with *SUNmin* for predicting the responses. In Figure 4.17b, *PPTmax* and isothermality (*TMPiso*) display the strongest interaction, followed by *CLAY* and mean annual temperature (*TMPmean*).



(a) Top 10 important variables and their interactions from regression trees.

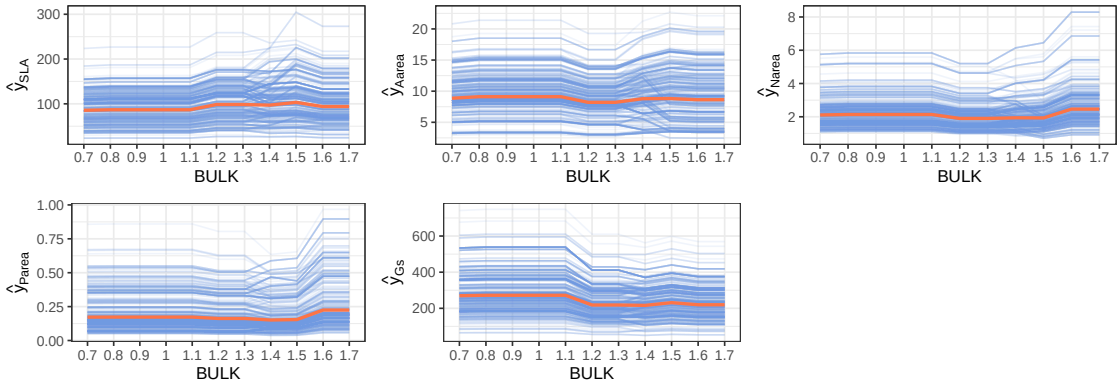
(b) Top 10 important variables and their interactions from missingness trees. The variable *Parea* is highlighted as it is a response variable in the data model and plays the role of a predictor in the missingness model.

Figure 4.17: Heat maps showing the top 10 important variables and their interactions from the regression trees (a) and missingness trees (b) of missBART2.

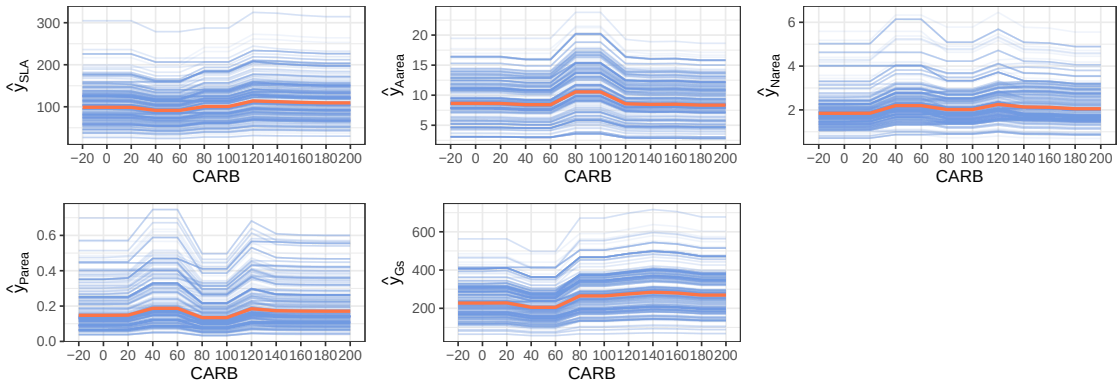
When comparing interactions between variables for the regression and missingness trees, we note that there are more non-zero interaction terms between variables in the regression trees than in the missingness trees. However, the magnitude of interactions is larger in the missingness trees, and similarly for the importance measures. This is largely due to the number of trees used in each model. As mentioned in Section 3.5.2, variables occur more uniformly in splitting rules when more trees are used. In comparison, fewer trees aid the identification of more influential variables. In this case, the regression trees use twice as many trees as the missingness trees, leading to more uniform use of variables in the regression model. In contrast, the smaller number of trees in the missingness model allows the most critical variables to dominate the splitting rules, resulting in fewer but stronger interactions and higher importance scores for those variables. This distinction underscores the need for careful consideration of the number of trees in both sets of trees.

By using partial dependence plots [PDP; Friedman, 2001] and individual conditional expectation [ICE; Goldstein et al., 2015] curves, we can further investigate the marginal effect each variable has on the responses. In Figure 4.18, the PDP (shown in orange) and ICE curves (shown in blue) illustrate how the predictions of each response variable change as the two most important variables in the regression trees of missBART2, *BULK* and *CARB*, increase. As shown in Figure 4.18a, the 5 responses remain relatively stable across increasing values of *BULK*, with *Narea* and *Parea* increasing slightly as *BULK* increases from 1.5 to 1.7, while *Gs* displays a slight decreasing trend as *BULK* increases.

In Figure 4.18b, the responses show a non-linear relationship with *CARB*. *Aarea* shows a spike when *CARB* is between 60 and 120. Both *Narea* and *Parea* exhibit a rise between 20 and 65, followed by a dip between 65 and 120. For *Gs*, there is a dip between 20 and 60, after which a slight increase occurs as *CARB* continues to increase.



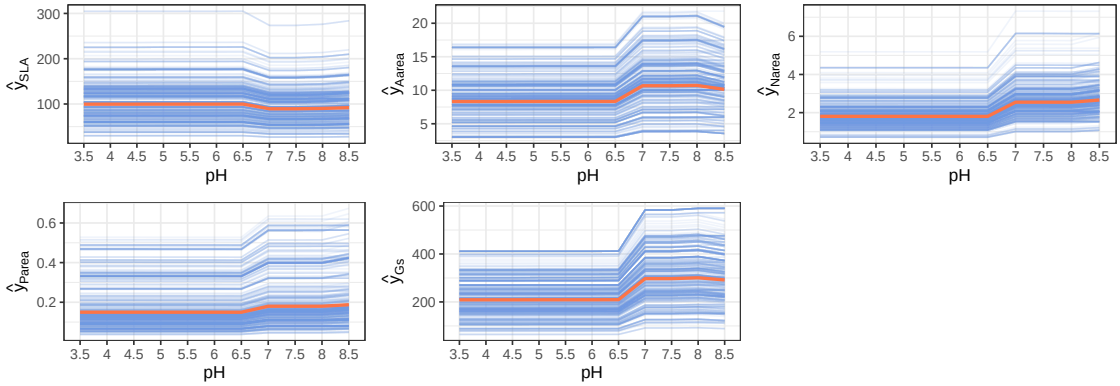
(a) PDP + ICE curves for responses across different levels of *BULK*.



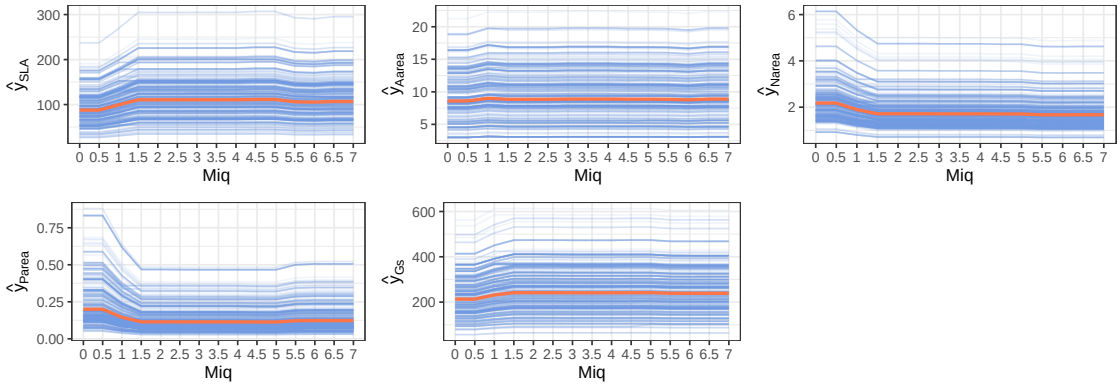
(b) PDP + ICE curves for responses across different levels of *CARB*.

Figure 4.18: PDP + ICE plots from the regression trees of missBART2 for responses across *BULK* and *CARB* levels. The PDP curves (in orange) show the overall average effect of the covariates on the response variables, while the ICE curves (in blue) illustrate the individual variability. The top panel (a) corresponds to *BULK*, with the bottom panel (b) corresponding to *CARB*. The plots reveal little variation in the responses for *BULK* and a non-linear relationship between the responses and *CARB*.

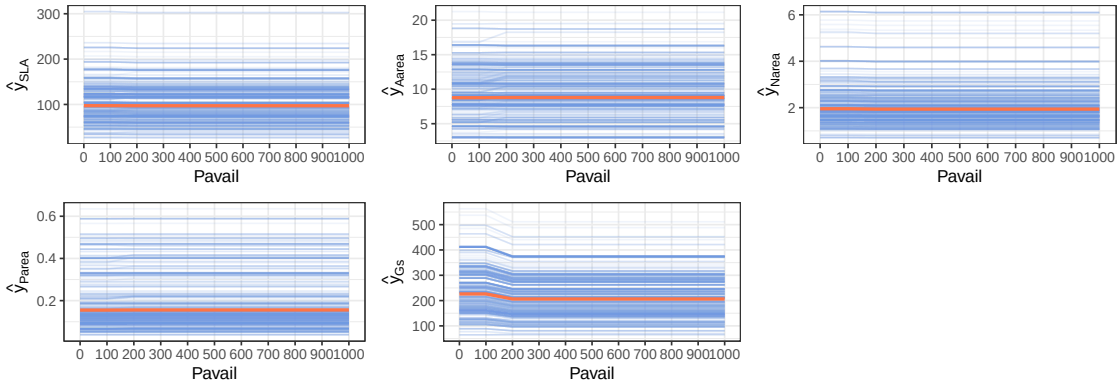
The results from Maire et al. [2015] found relationships between some of the responses and covariates *pH*, *Miq*, and *Pavail*. Specifically, *Aarea*, *Narea*, and *Parea* increased as *pH* increased and *Miq* decreased, while *SLA* decreased. Additionally, *Parea* increased while *Gs* decreased with increasing *Pavail*. Figure 4.19 shows the PDP and ICE curves for the covariates *pH*, *Miq*, and *Pavail*.



(a) PDP + ICE curves for responses across different levels of *pH*.



(b) PDP + ICE curves for responses across different levels of *Miq*.



(c) PDP + ICE curves for responses across different levels of *Pavail*.

Figure 4.19: PDP + ICE plots from the regression trees of missBART2 for different responses across *pH*, *Miq*, and *Pavail* levels. These variables were deemed as influential variables in the Maire et al. [2015]. The PDP curves (in orange) show the overall average effect of the covariates on the response variables, while the ICE curves (in blue) illustrate the individual variability.

In Figure 4.19a, *Aarea*, *Narea*, *Parea*, and *Gs* all show an increase when *pH* is above 6.5, while *SLA* sees a decline. However, there is a small decrease in *Aarea* as *pH* increases from 8 to 8.5, likewise for *Gs*. These results align with established theory, as *pH* has a strong influence on plant availability of soil nutrients, directly impacting photosynthetic

traits. More specifically, nutrient availability is generally highest at moderate *pH* levels and decreases significantly in highly alkaline soils (e.g., *pH* above 8) and highly acidic soils (e.g., *pH* below 6) [Westerband et al., 2023]. As *Miq* increases in Figure 4.19b, *SLA* and *Gs* increase while *Narea* and *Parea* decrease. *Aarea* seems unaffected. From Figure 4.19c, apart from a slight decrease in *Gs* when *Pavail* increases from 0 to 200, we see virtually no changes in all other responses.

As for the missingness, Figure 4.20 shows the effect of log(*Parea*) on the detection probabilities of each response variable. While the detection probabilities of most responses decrease as log(*Parea*) exceeds a value of -2, the detection probabilities of *Gs* show a slight increase on average. This implies that *SLA*, *Aarea*, *Narea*, and *Parea* are more likely to be missing when log(*Parea*) is greater than -2, while *Gs* shows the opposite.

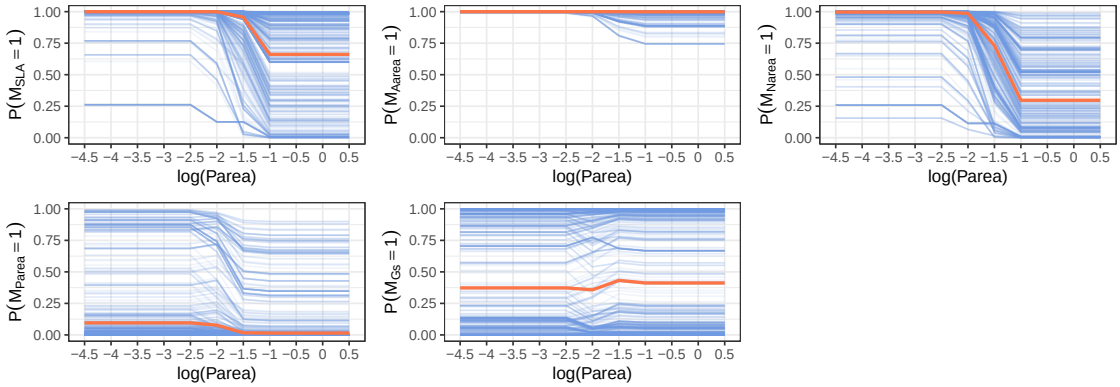


Figure 4.20: PDP + ICE plots from the missingness model of missBART2 for the detection probabilities of different responses across varying levels of log-transformed *Parea*. All variables other than *Gs* are more likely to be missing when log(*Parea*) is greater than -2.

In Figure 4.21, bivariate PDPs across values of *SAND* and log(*Parea*) are shown for the 5 responses. When log(*Parea*) is below -2, the detection probabilities of *SLA* are close to 1 and stay constant with changing *SAND* values. However, when log(*Parea*) is greater than -2, the detection probabilities decrease, especially when *SAND* is between 25 and 75. For detection probabilities of *Narea* and *Parea*, we see a general decrease when *SAND* is below 25 and log(*Parea*) is above -2. Additionally, when log(*Parea*) is greater than -2, the detection probabilities for *Narea* are higher when *SAND* is between 50 and 75. Finally, the detection probabilities of *Gs* are lower when *SAND* is below 75.

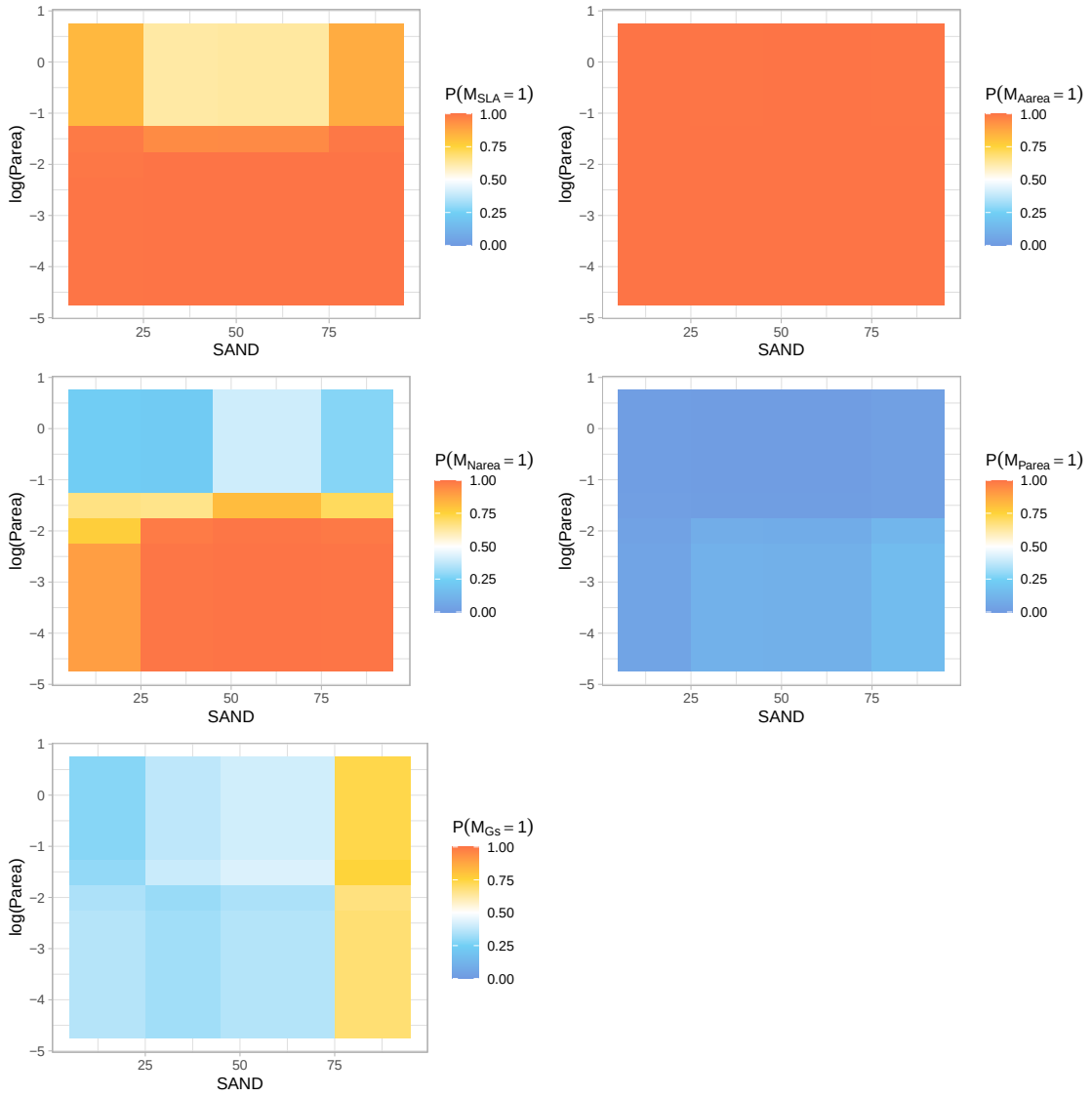


Figure 4.21: Bivariate PDPs for the detection probabilities of different responses across values of *SAND* and log(*Parea*).

4.4 Discussion & Conclusion

Motivated by the *global Amax* data with multivariate missing responses and completely observed covariates, we propose two novel models, missBART1 and missBART2, to address the limitations of existing missing data methods which predominantly focus on MCAR or MAR assumptions. Our models, which can handle MCAR, MAR, and MNAR scenarios for univariate and multivariate response datasets, offer a more flexible approach for the predictive modelling of data where missingness may be a concern. Both models operate within the selection model framework, differing primarily in the specification of the

missing data model, where missBART1 uses a probit regression model and missBART2 uses a probit BART model. While both models were designed to handle non-ignorable missing response data, they have also been adapted to handle ignorable missingness in the covariates. In missBART1, prior covariate imputation is necessary before model fitting, while missBART2 directly incorporates covariate missingness within the splitting rules of the decision trees.

From simulation studies, both models demonstrate strong performance in making accurate predictions on univariate and multivariate response data under various missingness scenarios, as well as the ability to recover the true underlying missingness mechanisms. The results highlight the advantages of our models compared to other methods such as complete case analysis and `missForest`-imputation followed by model fitting. Moreover, our models have the capability to recover both MAR and MNAR mechanisms, showcasing their robustness and flexibility. We highlight the complexities of assessing predictive performances when data are partially observed by reporting out-of-sample RMSEs based on different levels of detection probabilities or missingness statuses. Especially when data are MNAR, solely evaluating model performances based on RMSEs of the observed data can lead to erroneous conclusions, as the missing data imputations of models which ignore the missing data mechanism may be highly inaccurate.

When covariates are partially observed, missBART1 and missBART2 perform well with imputed covariates. While our investigation focused primarily on missing responses rather than covariates and thus limited studies were carried out for varying levels and patterns of missing covariates, we speculate that complicated covariate missingness could result in inaccurate covariate imputations, potentially degrading the performance of missBART1. Without the need for prior covariate imputation, missBART2 is a robust alternative for handling datasets with missingness in both the responses and covariates. This feature makes missBART2 particularly advantageous in scenarios where the two-step process of covariate imputation followed by model fitting may be inefficient or cumbersome.

While interpreting the missingness model, probit regression parameters in missBART1, denoted by $\mathbf{B}$, offer insights into the underlying missing mechanisms of the responses. By assigning priors to the precision of $\mathbf{B}$, we introduce more flexibility for incorporating any prior beliefs about the missing data mechanisms. The default settings scale the prior means of precisions according to the number of covariates and responses, which lessens the emphasis on identifying missingness as MNAR unless strongly supported by the data. This approach aligns with typical assumptions of missingness as MAR, while still allowing for MNAR recovery if indicated. However, careful tuning of these hyperparameters is essential to avoid potential model misspecification.

In contrast, the missingness model within missBART2 does not offer users the option to include any prior information on the missingness mechanism, but instead relies on the

variable selection feature of BART to identify important variables which influence the missingness. By evaluating the variable importance of the missingness trees, primarily whether the responses were important or not, we can gain insights into the underlying missing data mechanisms. However, the efficacy of this is impacted by the number of trees used in the missing model, K_m. As found by Chipman et al. [2010], reducing the number of trees in BART enhances its ability to identify the most influential covariates affecting the response, as it necessitates fewer splitting rules, resulting in a more selective usage of variables within the model. This encourages the inclusion of only the most influential variables, improving the identification of key factors driving the underlying missingness mechanisms. However, the MCMC sampler risks getting trapped in a local mode when too few trees are used, resulting in slower convergence and poorer overall model fit [Chipman et al., 2010, Bleich et al., 2014].

According to Bleich et al. [2014], while higher inclusion proportions suggest variable importance, the raw values cannot be directly interpreted as they do not reflect posterior probabilities. The authors posed a crucial question: *What threshold must the variable inclusion proportion meet to classify a predictor as important?* To address this, they introduced a permutation-based method that establishes optimal thresholds for variable inclusion proportions, allowing for more robust identification of important variables. Thus, in the interest of missing mechanism evaluation and recovery, future work on missBART2 may focus on examining importance thresholds derived from the missingness trees to gain a better understanding of the key variables that influence the missingness mechanisms. Alternatively, a Dirichlet prior may also be applied to the splitting rules of the missingness trees, as introduced in Linero [2018]. This sparsity-inducing method encourages the model to favour using only a subset of the predictors, which could facilitate the recovery of the true underlying missing mechanism.

Our analysis of the *global Amax* data diverged from the approach used by Maire et al. [2015] in several key ways. While Maire et al. [2015] employed separate regression models for each photosynthetic trait, potentially overlooking the effects of missingness in the data, our models utilised a non-linear and non-parametric multivariate BART framework, explicitly accounting for the missingness structure in the responses. This joint modelling approach allows for a more comprehensive understanding of the relationships between covariates and responses while addressing the limitations of missing data. In their study, Maire et al. [2015] identified three covariates — *pH*, *Miq*, and *Pavail* — as significantly influencing photosynthetic traits. While our PDP and ICE plots from Figure 4.19 generally support these findings for *pH* and *Miq*, our models suggest little to no relationship between *Pavail* and the responses. Furthermore, the univariate analyses in Maire et al. [2015] found that most responses were strongly influenced by *pH* and *Miq*, whereas *Pavail* primarily affected *Parea* and *Gs* — the two responses with the

highest proportions of missingness. This raises two important considerations. First, explicitly modelling the missingness structure allows our analysis to uncover more nuanced relationships between the covariates and responses, highlighting potential biases that may arise when missing data mechanisms are ignored. Second, future modification to our models could involve integrating the "seemingly unrelated BART" model from Esser et al. [2024], allowing each response variable to be associated with different sets of BART trees while also accounting for dependencies between the responses, incorporating further flexibility and interpretability of the models in multivariate response settings.

Finally, the *global Amax* dataset used in Maire et al. [2015] consists of continuous data only. However, from the DRYAD Digital Repository where *global Amax* was obtained, several other categorical covariates, such as *Country* and *Genus*, are also available. Including these covariates may improve the predictive performances of our models, but would require the adaptation of **flexBART**, which is an extension of BART from Deshpande [2022] to handle categorical covariates.

5

A Joint Seemingly Unrelated BART Model for Non-Ignorable Missing Data

In this Chapter, we propose a novel joint model, "missSUBART", that handles multivariate response data with potentially non-ignorable missingness, allowing each response to be associated with a distinct set of predictors while simultaneously capturing the correlation structure between the responses.

5.1 Introduction

In Chapter 3, two joint models, missBART1 and missBART2, were introduced within the selection model framework to handle multivariate response data with missing values without imposing strict assumptions on the missing data mechanisms. Both models demonstrated strong predictive performance and effective recovery of missingness mechanisms under various simulated scenarios (Chapter 4), with missBART2 frequently outperforming missBART1 and other competing methods. However, a key limitation of missBART2 is its enforcement of a shared tree structure across all responses, such that all p responses share the same partitioning of the predictor space. In the analysis of the *global Amax* dataset, Maire et al. [2015] highlighted that different predictors influence each response to varying degrees, suggesting the need for response-specific regression functions. Thus, in a tree-based modelling context, distinct responses may require separate tree structures.

A model that allows for separate functions representing the response estimates while still capturing correlated error structures is the seemingly unrelated regression (SUR) model, originally proposed by Zellner [1962]. Given a p-dimensional vector of fully observed responses $\mathbf{Y}_i$ and q covariates $\mathbf{X}_i$, the SUR model can be represented as

$$\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{ip} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1^\top \mathbf{X}_i \\ \vdots \\ \mathbf{b}_p^\top \mathbf{X}_i \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ip} \end{pmatrix},$$

where $(\epsilon_{i1}, \ldots, \epsilon_{ip})^\top \overset{i.i.d.}{\sim} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ is the vector of residual standard deviations with a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. The SUR model estimates a set of p separate linear regression equations, where the conditional expectation of each response $j = 1, \ldots, p$ has its own vector of coefficients $\mathbf{b}_j$. Unlike independent regressions, SUR jointly estimates a non-diagonal covariance matrix $\boldsymbol{\Sigma}$ for the error terms, capturing potential correlations between the residuals associated with each response. However, the strong linearity assumptions of the original SUR model limit its applicability in non-linear settings. To address this, Gallant [1975] introduced a non-linear generalisation of SUR. A comprehensive review of SUR models can be found in Srivastava and Giles [1987], with more recent developments summarised in Fiebig [2003].

Among several extensions made to the original SUR model is that from Chakraborty [2016], who proposed BART–SUR to address issues arising from the strong linear assumptions imposed by traditional SUR. The BART–SUR model was developed as a non-linear, non-parametric SUR model that replaces the linear functions in SUR with Bayesian additive regression trees (BART). This approach allows each response variable to have its own set of trees and predictors for splitting rules while also assigning priors to the number of trees for adaptive tuning, improving model efficiency. However, BART–SUR was introduced only in the context of multivariate continuous responses.

To address broader applications, Esser et al. [2024] introduced seemingly unrelated BART (suBART), a variant of BART–SUR that omits adaptive tree tuning while extending the model to the probit framework for estimating cost-effectiveness in health economics. Further details of the suBART model are included in Section 5.2. A key innovation in suBART is its carefully calibrated prior on the residual covariance matrix, which is based on the hierarchical inverse-Wishart prior of Huang and Wand [2013] rather than the non-informative inverse-Wishart prior used in Chakraborty [2016]. This calibration regularises the residual variance parameters to prevent overfitting while ensuring proper uncertainty quantification, particularly for their application in causal effect estimation.

By maintaining independent tree structures for each response while capturing error correlations, suBART represents a flexible, non-parametric alternative to traditional SUR models, making it particularly well-suited for complex multivariate regression problems. By overcoming the limitation of a shared tree structure while retaining a shared error structure, suBART also represents a flexible alternative to multivariate BART. Through multiple simulation studies on datasets with fully observed responses and covariates, Esser et al. [2024] demonstrated that suBART outperforms competing methods in terms of both predictive accuracy and the accurate estimation of correlation structures. It outperformed alternative methods, including standard BART applied separately to each univariate response, the multivariate BART model such as the one in Chapter 3, and a Bayesian linear seemingly unrelated regression model for multivariate responses.

Here, we propose another novel selection model by replacing the data and missingness models from the selection model with the seemingly unrelated BART models for continuous and binary responses, jointly modelling the response and missingness indicators as before. The proposed model, 'missSUBART', retains the advantages of missBART2, including its non-linear treatment of both responses and missing indicators, its flexibility in handling missing data without restrictive assumptions, and its ability to recover MNAR missingness. However, missSUBART introduces additional flexibility by allowing each response and missingness mechanism to have distinct predictor associations via separate sets of trees. In addition to responses having individualised tree structures—enabling more direct comparisons with results from Maire et al. [2015]—the missingness model also benefits by accommodating heterogeneous missingness patterns across responses. This means that while missBART2 may suggest an overall missingness mechanism shared across all responses, missSUBART enables a more nuanced structure where the missingness mechanisms may vary between each response.

5.2 Seemingly Unrelated BART

In this section, we give an overview of the seemingly unrelated BART model following notation, prior specifications, and model setup of suBART from Esser et al. [2024]. The suBART model is a flexible, non-parametric approach developed to model multiple correlated outcomes in settings where traditional linear methods, such as seemingly unrelated regression (SUR), fall short. Unlike the multivariate BART model introduced in Chapter 3, which assumes a shared tree structure across all responses, suBART allows each outcome to have independent tree ensembles while still modelling residual correlations through a structured error covariance. This enables greater flexibility in capturing response-specific predictor effects and non-linear dependencies, making it particularly useful in complex multivariate settings.

However, despite its increased flexibility and robustness, suBART comes with a key computational drawback. Since each response is assigned an independent set of trees, the total number of trees scales with the number of responses, significantly increasing computational complexity. Moreover, this structure introduces a substantially larger number of parameters, requiring more updates per iteration, further slowing down computation. As a result, while suBART offers greater modelling flexibility, its computational demands can be prohibitive, particularly in high-dimensional settings.

Given a fully observed set of p responses, $\mathbf{Y}_i$, and q model covariates, $\mathbf{X}_i$, the suBART model can be formulated as

$$\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{ip} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{K} g\left(\mathbf{X}_i; \mathcal{T}_k^{y_1}, Q_k^{y_1}\right) \\ \vdots \\ \sum_{k=1}^{K} g\left(\mathbf{X}_i; \mathcal{T}_k^{y_p}, Q_k^{y_p}\right) \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ip} \end{pmatrix}, \tag{5.1}$$

where $\boldsymbol{\mathcal{T}}^{y_j} = \left(\mathcal{T}_1^{y_j}, \ldots, \mathcal{T}_K^{y_j}\right)$, $j = 1, \ldots, p$, is the set of K regression trees associated with the j^{th} response variable, $\mathbf{Y}^{(j)}$, $\boldsymbol{Q}^{y_j} = \left(Q_1^{y_j}, \ldots, Q_K^{y_j}\right)$ is the set of all terminal node parameters associated with $\mathbf{Y}^{(j)}$, each $Q_k^{y_j} = \left(\mu_{k1}^{y_j}, \ldots, \mu_{k\ell_k^{(j)}}^{y_j}\right)$ contains the vector of $\ell_k^{(j)}$ univariate terminal node parameters associated with tree $\mathcal{T}_k^{y_j}$, and $(\epsilon_{i1}, \ldots, \epsilon_{ip})^\top \overset{i.i.d.}{\sim} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ is the vector of residual standard deviations with a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. A key distinction of the suBART model from the multivariate BART model is that each response is associated with a unique set of univariate regression trees. While the multivariate BART model contained $p-$dimensional vectors of parameters in each terminal node, the suBART model requires univariate terminal node parameters for all j ensembles of trees.

The model setup for suBART from Esser et al. [2024] closely follows the structure of standard BART from Chipman et al. [2010]. First, the prior for each tree $\mathcal{T}_k^{y_j}$ follows the specifications from Chipman et al. [2010], using default hyperparameter settings to favour shallow trees and avoid overfitting. Upon scaling and shifting $\mathbf{Y}^{(j)}$ such that they fall within the range $[-0.5, 0.5]$, each terminal node parameter in $\left(\mu_{k1}^{y_j}, \ldots, \mu_{k\ell_k^{(j)}}^{y_j}\right)$ is assigned a $\mathcal{N}\left(0, 1/\tau_\mu^{(j)}\right)$ prior. Following this, each $\tau_\mu^{(j)}$ can be calibrated using Equation (2.2).

Although the multivariate BART model detailed in Section 3.4 specifies a residual precision matrix $\boldsymbol{\Omega}$ and assigns to it a Wishart prior, Esser et al. [2024] specify $\boldsymbol{\Sigma}$ as the covariance matrix of the residual errors and use an approach from Huang and Wand [2013] to assign a prior for $\boldsymbol{\Sigma}$ such that

$$\boldsymbol{\Sigma} \mid \mathbf{a} \sim \text{Inv-Wishart}_p\left(\nu + p - 1, \mathbf{S}_0\right),$$

where $\mathbf{a} = (a_1, \ldots, a_p)$, $\mathbf{S}_0 = 2\nu \times \text{diag}(\mathbf{a})^{-1}$, $a_j \sim \text{Inv-Gamma}\left(0.5, 1/A_j^2\right)$, ν is a fixed hyperparameter, and $A_j > 0$ is a fixed scale hyperparameter, calibrated using a data-based approach. This calibration technique distinguishes suBART from BART-SUR in Chakraborty [2016], which employs a non-informative inverse-Wishart prior for $\boldsymbol{\Sigma}$.

More specifically, Esser et al. [2024] choose A_j to ensure that the residual standard deviation of each response, denoted by σ_j and equal to the square root of the j^{th} diagonal element of $\boldsymbol{\Sigma}$, does not exceed the rough data-based overestimate $\hat{\sigma}_j$. Similar to Chipman et al. [2010], $\hat{\sigma}_j$ can be obtained by calculating the sample standard deviations or computing the estimated residual standard deviations from least squares linear regression models. The latter approach is chosen as the default in Esser et al. [2024], which we also

adopt in our models. Following the prior specifications of $\boldsymbol{\Sigma}$ and $\mathbf{a}$, the induced prior on σ_j follows a half-t distribution, Half-$t(\nu, A_j)$. By assuming *a priori* that $P(\sigma_j < \hat{\sigma}_j) = \rho_\sigma$, where $\rho_\sigma \in [0,1]$, A_j can be obtained through solving

$$\rho_\sigma = \frac{2\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\frac{\nu}{2})A_j\sqrt{\nu\pi}} \int_0^{\hat{\sigma}_j} \left(1 + \frac{u^2}{\nu A_j^2}\right)^{-0.5(\nu+1)} \mathrm{d}u.$$

By default, $\nu = 2$ and $\rho_\sigma = 0.95$. For more details, see Esser et al. [2024].

A probit version of the suBART model is also outlined in Esser et al. [2024] for handling multivariate binary outcomes. Similar to the multivariate probit BART model, probit suBART uses the data augmentation scheme from Chib and Greenberg [1998] to handle multivariate binary outcomes. Assuming instead that $\mathbf{Y}$ is binary and defining $\mathbf{Y}^*$ as the corresponding latent variables, the probit suBART model takes the form

$$\begin{pmatrix} Y_{i1}^* \\ \vdots \\ Y_{ip}^* \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^K g\left(\mathbf{X}_i; \mathcal{T}_k^{y_1^*}, Q_k^{y_1^*}\right) \\ \vdots \\ \sum_{k=1}^K g\left(\mathbf{X}_i; \mathcal{T}_k^{y_p^*}, Q_k^{y_p^*}\right) \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ip} \end{pmatrix}, \tag{5.2}$$

and $Y_{ij} = 0$ if $Y_{ij}^* \leq 0$ and vice versa. Here, $(\epsilon_{i1}, \ldots, \epsilon_{ip})^\top \overset{i.i.d.}{\sim} \mathcal{N}_p(\mathbf{0}, \mathbf{R})$ where $\mathbf{R}$ is a correlation matrix.

5.2.1 Posterior Sampling of suBART

In the continuous response version of suBART, the joint posterior distribution is

$$p\left(\boldsymbol{\mathcal{T}}^y, \boldsymbol{Q}^y, \boldsymbol{\Sigma}, \mathbf{a} \mid \mathbf{X}, \mathbf{Y}\right) \propto \prod_{k=1}^{K_y} \prod_{j=1}^{p} \left\{ p(\mathcal{T}_k^{y_j}) p(Q_k^{y_j} \mid \mathcal{T}_k^{y_j}) \right\} \times p(\boldsymbol{\Sigma} \mid \mathbf{a}) \times \prod_{j=1}^{p} p(a_j).$$

In order to derive posterior sampling distributions, the conditional distribution for each response variable, $Y_{ij} \,|\, \mathbf{X}, \mathbf{Y}^{(-j)}, \boldsymbol{\mathcal{T}}^{y_j}, \boldsymbol{Q}^{y_j}, \boldsymbol{\Sigma}$, is required [Bierens, 2004, Esser et al., 2024]. This is given as

$$\mathcal{N}\left(\hat{Y}_{ij} + \boldsymbol{\Sigma}_{j(-j)}\boldsymbol{\Sigma}_{(-j)(-j)}^{-1}\left(\mathbf{Y}_i^{(-j)} - \hat{\mathbf{Y}}_i^{(-j)}\right), \Sigma_{jj} - \boldsymbol{\Sigma}_{j(-j)}\boldsymbol{\Sigma}_{(-j)(-j)}^{-1}\boldsymbol{\Sigma}_{(-j)j}\right),$$

where $\hat{Y}_{ij} = \sum_{k=1}^{K_y} g(\mathbf{X}_i; \boldsymbol{\mathcal{T}}^{y_j}, \boldsymbol{Q}^{y_j})$ are the fitted values, $\mathbf{Y}_i^{(-j)}$ is obtained by omitting the j^{th} column in $(Y_{i1}, \ldots, Y_{ip})$ and likewise for $\hat{\mathbf{Y}}_i^{(-j)}$. Additionally, Σ_{jj} is the j^{th} diagonal entry of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{j(-j)}$ is the vector obtained by omitting the j^{th} column of the j^{th} row in $\boldsymbol{\Sigma}$ and likewise for $\boldsymbol{\Sigma}_{(-j)j}$. Finally, $\boldsymbol{\Sigma}_{(-j)(-j)}^{-1}$ is the inverse of the matrix obtained after removing the j^{th} row and column of $\boldsymbol{\Sigma}$.

Next, we also require the partial residuals of each response variable in tree k:

$$\mathbf{r}_k^{(j)} = \mathbf{Y}^{(j)} - \sum_{t \neq k} g\left(\mathbf{X}; \mathcal{T}_t^{y_j}, Q_t^{y_j}\right),$$

which enables the back-fitting algorithm [Hastie and Tibshirani, 2000] as in univariate BART. This enables K_y sequential draws of $\left(\mathcal{T}_k^{y_j}, Q_k^{y_j}\right)$ from

$$\left(\mathcal{T}_k^{y_j}, Q_k^{y_j}\right) \mid \mathbf{r}_k^{(j)}, \boldsymbol{\Sigma}, \mathbf{a}$$

where $\mathcal{T}_k^{y_j}$ can be drawn independently for all j and k, as in univariate BART (see Section 2.1.2 and Kapelner and Bleich [2016]). Next, draws of terminal node parameters $\left(\mu_{k1}^{y_j}, \ldots, \mu_{k\ell_k^{(j)}}^{y_j}\right)$ can be obtained from

$$\mathcal{N}\left(\mathcal{G}_j \left(\sum_{i=1}^{n_{k\ell}^{(j)}} r_{ik}^{(j)} - \sum_{i=1}^{n_{k\ell}^{(j)}} \mathcal{U}_i^{(j)}\right), \mathcal{V}_j \mathcal{G}_j\right), \tag{5.3}$$

where we have

$$\begin{aligned}
\mathcal{G}_j &= \left(\tau_\mu^{(j)} \mathcal{V}_j + n_{k\ell}^{(j)}\right)^{-1} \\
\mathcal{V}_j &= \Sigma_{jj} - \boldsymbol{\Sigma}_{j(-j)} \boldsymbol{\Sigma}_{(-j)(-j)}^{-1} \boldsymbol{\Sigma}_{(-j)j} \\
\mathcal{U}_i^{(j)} &= \boldsymbol{\Sigma}_{j(-j)} \boldsymbol{\Sigma}_{(-j)(-j)}^{-1} \left(\mathbf{Y}_i^{(-j)} - \hat{\mathbf{Y}}_i^{(-j)}\right).
\end{aligned}$$

Next, we make j draws of a_j from

$$a_j \mid \boldsymbol{\Sigma} \sim \text{Inv-Gamma}\left(\frac{\nu + n}{2}, A_j^{-2} + \nu \boldsymbol{\Omega}_{jj}\right), \tag{5.4}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Omega}_{jj}$ is the j^{th} diagonal entry of $\boldsymbol{\Omega}$. Finally, we draw $\boldsymbol{\Sigma}$ from

$$\boldsymbol{\Sigma} \mid \mathbf{Y}, \mathbf{a} \sim \text{Inv-Gamma}\left(\nu + p + n - 1, \mathbf{S}_0 + \sum_{i=1}^{n} \left(\mathbf{Y}_i - \hat{\mathbf{Y}}_i\right) \left(\mathbf{Y}_i - \hat{\mathbf{Y}}_i\right)^\top\right). \tag{5.5}$$

The sampling procedure for the probit suBART model follows closely with the univariate probit suBART model and the sampling methods for continuous response suBART, and will not be discussed here. For details, see Esser et al. [2024].

5.3 missSUBART for Multivariate MNAR Missing Data

Using the selection model framework from Section 3.2,

$$p\left(\mathbf{Y}, \mathbf{M} \mid \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\psi}\right) = \underbrace{p\left(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}\right)}_{\text{data model}} \times \underbrace{p\left(\mathbf{M} \mid \mathbf{X}, \mathbf{Y}, \boldsymbol{\psi}\right)}_{\text{missingness model}},$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are the data and missingness model parameters, respectively, we introduce missSUBART, a seemingly unrelated joint model for handling multivariate non-ignorable missing data, where missingness occurs in the response variables. This model allows each response and missing indicator to follow distinct tree structures while simultaneously accounting for correlated dependencies among responses. First, we assign a suBART model as described in Section 5.2 to the data model.

For the missingness model, while it is possible to assign a probit suBART model, assuming correlated missing data indicators and estimating the correlation matrix $\mathbf{R}$, this poses some challenges. In Esser et al. [2024], a parameter-expanded data augmentation strategy from Zhang [2020] was adopted to estimate the correlation matrix $\mathbf{R}$ in the probit suBART model, incorporating additional latent variables and a parameter-expanded Metropolis-Hastings algorithm. However, without careful tuning of the proposal distribution and a sufficiently large number of iterations, this approach can suffer from low acceptance rates and high autocorrelation, leading to substantial computational overhead. Given these computational challenges, along with the inherent flexibility of tree structures to capture true underlying signals, we consider it sufficient to assume that the missingness indicators are uncorrelated, essentially fixing $\mathbf{R} = \mathbf{I}_p$. This specification effectively reduces the missingness model to p independent univariate probit BART models. We note that this is different from the missingness model in missBART2 where missBART2 had uncorrelated error terms through fixing $\mathbf{R} = \mathbf{I}_p$ but retains dependence through shared terminal node parameter vectors within the multivariate probit trees.

Recall from Equation (3.11) that $\tilde{Y}_{ij} = Y_{ij}^{obs}$ if $M_{ij} = 1$ and $\tilde{Y}_{ij} = Y_{ij}^{mis}$ otherwise. Additionally, $\mathbf{Y}^{obs} = \{Y_{ij} \colon M_{ij} = 1\}$ is the set of observed responses and $\mathbf{Y}^{mis} = \{Y_{ij} \colon M_{ij} = 0\}$ is the set of responses which are missing. Denoting $\mathbf{Z}$ as the set of missingness model predictors which includes but is not limited to the observed and missing responses, the complete data likelihood for missSUBART is

$$p\left(\tilde{\mathbf{Y}}, \mathbf{M} \mid \mathbf{X}, \boldsymbol{\mathcal{T}}^y, \boldsymbol{Q}^y, \boldsymbol{\Sigma}, \mathbf{a}, \boldsymbol{\mathcal{T}}^m, \boldsymbol{Q}^m\right) = \underbrace{p\left(\tilde{\mathbf{Y}} \mid \mathbf{X}, \boldsymbol{\mathcal{T}}^y, \boldsymbol{Q}^y, \boldsymbol{\Sigma}, \mathbf{a}\right)}_{\text{suBART regression model}} \underbrace{p\left(\mathbf{M} \mid \mathbf{Z}, \boldsymbol{\mathcal{T}}^m, \boldsymbol{Q}^m\right)}_{\substack{p \text{ univariate probit} \\ \text{BART missingness models}}}, \tag{5.6}$$

where $\boldsymbol{\mathcal{T}}^y$ and $\boldsymbol{\mathcal{T}}^m$ contain p sets of regression and missingness trees respectively for each response variable, such that

$$\boldsymbol{\mathcal{T}}^{y} = \left(\left(\mathcal{T}_1^{y_1}, \ldots, \mathcal{T}_{K_y}^{y_1}\right), \ldots, \left(\mathcal{T}_1^{y_p}, \ldots, \mathcal{T}_{K_y}^{y_p}\right)\right)$$
$$\boldsymbol{\mathcal{T}}^{m} = \left(\left(\mathcal{T}_1^{m_1}, \ldots, \mathcal{T}_{K_m}^{m_1}\right), \ldots, \left(\mathcal{T}_1^{m_p}, \ldots, \mathcal{T}_{K_m}^{m_p}\right)\right),$$

with K_y and K_m being the respective numbers of trees for each variable in the regression and missingness model. While the number of trees assigned to each outcome can vary, we assume that K_y remains constant across all responses and K_m is similarly fixed for the missingness indicators.

In missBART2, incorporating responses as predictors in the missingness model enables the estimation of relationships between a response's missingness and the values of all missing responses, facilitating the recovery of potential MNAR mechanisms. Specifically, the missingness model predictors in missBART2, denoted as $\mathbf{Z}_i$, consist of both the regression covariates and response values, $\left(\mathbf{X}_i, \tilde{\mathbf{Y}}_i\right)^\top$, which are used in the splitting rules of $\boldsymbol{\mathcal{T}}^m$ and are shared across all p missingness indicators. This results in the formulation $\mathbf{Z}_i = \left(\mathbf{X}_i, \tilde{\mathbf{Y}}_i\right)^\top$.

In contrast, missSUBART assigns independent univariate BART models to each response's missingness indicator, allowing for distinct sets of predictors used within the splitting rules of each $\boldsymbol{\mathcal{T}}^{m_j}$. For each response j, one possible specification of the missingness predictors is $\mathbf{Z}_i^{(j)} = \left(\mathbf{X}_i, \tilde{\mathbf{Y}}_i\right)^\top$, maintaining consistency with missBART2 while permitting response-specific associations. However, missSUBART can extend this approach by also allowing dependencies between a response's missingness and the missingness indicators of other responses, rather than just their values. This is formulated as

$$\mathbf{Z}_i^{(j)} = \left(\mathbf{X}_i, \tilde{\mathbf{Y}}_i, \mathbf{M}_i^{(-j)}\right)^\top,$$

where $\mathbf{M}_i^{(-j)}$ denotes the missingness indicators of all responses except the j^{th} indicator. This not only allows for the recovery of MNAR mechanisms, but also a form of MAR missingness where the missingness of one response depends on whether another response is missing or not.

5.3.1 Posterior Sampling of missSUBART

The joint posterior distribution of the missSUBART model is

$$p\left(\boldsymbol{\mathcal{T}}^{y}, \boldsymbol{Q}^{y}, \boldsymbol{\Sigma}, \mathbf{a}, \boldsymbol{\mathcal{T}}^{m}, \boldsymbol{Q}^{m}, \mathbf{M}^{*}, \mathbf{Y}^{mis} \mid \mathbf{X}, \mathbf{Y}^{obs}, \mathbf{M}\right),$$

where $\mathbf{M}^*$ is the latent variable introduced in the data augmentation scheme within probit BART to model the binary outcomes $\mathbf{M}$ (see Section 2.1.3). Retaining the prior specifications and model calibration techniques from Esser et al. [2024], as outlined in

Section 5.2, the sampling of $(\boldsymbol{\mathcal{T}}^{y}, \boldsymbol{Q}^{y}, \boldsymbol{\Sigma}, \mathbf{a})$ follows the steps from Section 5.2.1. Next, $\left(\boldsymbol{\mathcal{T}}^{m_j}, \boldsymbol{Q}^{m_j}, \mathbf{M}^{*(j)}\right)$ can be drawn for each j as per the univariate probit BART model, specifying default prior settings and calibrating the models as in Chipman et al. [2010]. More specifically, to sample M^{*}_{ij}, we make draws from a truncated normal distribution, such that

$$M^{*}_{ij} \mid \mathcal{T}^{m_j}, \boldsymbol{Q}^{m_j}, M_{ij} \sim \mathcal{TN}\left(\hat{M}^{*}_{ij}, 1, \gamma_{ij}\right), \tag{5.7}$$

where $\hat{M}^{*}_{ij} = \sum_{k=1}^{K_m} g\left(\mathbf{Z}_i^{(j)}; \boldsymbol{\mathcal{T}}^{m_j}, \boldsymbol{Q}^{m_j}\right)$ and γ_{ij} denotes the truncation points where $\gamma_{ij} = [0, \infty)$ if $M_{ij} = 1$ and $\gamma_{ij} = (-\infty, 0]$ if $M_{ij} = 0$.

Finally, we also require the sampling of missing responses $\mathbf{Y}^{mis}$. Denoting $\boldsymbol{\theta}^{(j)} = \{\boldsymbol{\mathcal{T}}^{y_j}, \boldsymbol{Q}^{y_j}, \boldsymbol{\Sigma}\}$ and $\boldsymbol{\psi}^{(j)} = \{\boldsymbol{\mathcal{T}}^{m_j}, \boldsymbol{Q}^{m_j}\}$ as model parameters for the j^{th} response and missingness indicator respectively, we require for each missing entry

$$p\left(Y^{mis}_{ij} \mid \tilde{\mathbf{Y}}_i^{(-j)}, \boldsymbol{\theta}^{(j)}\right) p\left(M^{\star}_{ij} \mid \tilde{\mathbf{Y}}_i^{(-j)}, Y^{mis}_{ij}, \boldsymbol{\psi}^{(j)}\right),$$

where, as was the case for missBART2, no known distributional forms exist. Thus, the posterior sampling of $\mathbf{Y}^{mis}$ requires the implementation of a Metropolis-Hastings algorithm. As specified in missBART2, we use a random walk proposal distribution with a standard deviation $\sigma_Y = 0.5$ to ensure that the proposed values do not exceed too far outside the $[-0.5, 0.5]$ range. The sampling steps for missSUBART are as follows:

(1) For all j responses,

   (a) For all K_y trees, propose a new tree via a grow, prune, change, or swap move and accept or reject using a Metropolis-Hastings step[a].

   (b) Update each terminal node parameter in $\left(\mu^{y_j}_{k1}, \ldots, \mu^{y_j}_{k\ell_{k(j)}}\right)$ of the k^{th} tree using Equation (5.3).

(2) Update $a_1, \ldots, a_j$ and update Σ as in Equations (5.4) and (5.5).

(3) For all j missing indicators,

   (a) Repeat Steps 1(a) and 1(b) for all K_m trees.

   (b) Update each terminal node parameter in $\left(\mu^{m_j}_{k1}, \ldots, \mu^{m_j}_{k\ell_{k(j)}}\right)$ of the k^{th} tree as in univariate probit BART from Chipman et al. [2010].

(4) Update each M^{*}_{ij} using Equation (5.7).

[a] See Kapelner and Bleich [2016] for details on the tree-proposal moves and Esser et al. [2024] on the Metropolis-Hastings acceptance probability.

(5) Update $\mathbf{Y}^{mis}$ via a Metropolis-Hastings step. For every missing entry, first propose a new value $Y^{mis}_{t,i,j}$ from $\mathcal{N}(Y^{mis}_{t-1,i,j}, \sigma^2_Y)$. Next, calculate the acceptance probability $\omega(Y^{mis}_{t,i,j}, Y^{mis}_{t-1,i,j})$, which is equal to

$$\frac{p\left(Y^{mis}_{t,i,j} \mid \tilde{\mathbf{Y}}_i^{(-j)}, \boldsymbol{\theta}^{(j)}\right) p\left(M^{\star}_{ij} \mid \tilde{\mathbf{Y}}_i^{(-j)}, Y^{mis}_{t,i,j}, \boldsymbol{\psi}^{(j)}\right) q(Y^{mis}_{t,i,j} \to Y^{mis}_{t-1,i,j})}{p\left(Y^{mis}_{t-1,i,j} \mid \tilde{\mathbf{Y}}_i^{(-j)}, \boldsymbol{\theta}^{(j)}\right) p\left(M^{\star}_{ij} \mid \tilde{\mathbf{Y}}_i^{(-j)}, Y^{mis}_{t-1,i,j}, \boldsymbol{\psi}\right) q(Y^{mis}_{t-1,i,j} \to Y^{mis}_{t,i,j})},$$

and accept or reject the proposed $Y^{mis}_{t,i,j}$ with probability $\min\left(1, \omega(Y^{mis}_{t,i,j}, Y^{mis}_{t-1,i,j})\right)$.

As discussed in Section 5.2, the total number of trees in the suBART model scales with the number of response variables, leading to increased computational complexity. In missSUBART, this scaling extends further, with the total number of trees given by $(K_y + K_m) \times p$, where both the regression trees and missingness trees grow in proportion to the number of responses and their corresponding missingness indicators. This presents a key limitation, particularly when dealing with a large number of responses, as it requires careful tuning of the number of trees to balance predictive accuracy and computational efficiency, making missBART2 a more practical choice in such cases. However, this does not pose a significant challenge when the number of responses is relatively small, as in the case of the *global Amax* dataset, where $p = 5$.

As discussed in previous chapters, this choice balances the need for effective variable selection with model stability. A reduced number of trees encourages selective variable usage by limiting unnecessary splits, improving the identification of key predictors of missingness, while too few trees can cause the MCMC sampler to get trapped in local modes, slowing convergence and degrading overall model performance. In Chipman et al. [2010], the default number of univariate trees was fixed at 200. However, Esser et al. [2024] set the default number of trees for each response in the suBART as 50 and noted that the model's performance remains largely unaffected by variations in this parameter.

5.4 Simulation Studies

To assess the performance of missSUBART, we compare it against our previous joint models, missBART1 and missBART2, as well as suBART applied to complete cases ('suBART_cc') and `missForest`-imputed data ('suBART_imp'). Comparisons with multivariate BART and univariate BART on both complete cases and imputed data are omitted here due to their comparatively weaker performances in earlier sections.

Before comparing with the two scenarios from the previous simulation studies in Chapter 4, we first show an example in Section 5.4.1, called "MNAR 3", where the data follows a seemingly unrelated structure and missingness is simulated independently following separate univariate BART models for each response. More specifically, we generate the data

using the "Friedman #1" example from Esser et al. [2024] with three response variables, outlined later in Equation (5.8), while missingness is simulated such that missingness of each response follows a distinct set of trees and varies between MAR and MNAR, outlined later in Equation (5.9). Following this, in Section 5.4.2, we further evaluate two scenarios from Section 4.2, MAR 2 and MNAR 2, where the bivariate responses share similar tree structures.

As in Section 4.2, we use 5000 burn-in and post-burn-in iterations and carry out 4-fold cross-validation. For the missSUBART models, we use Esser et al. [2024]'s default of $K_y = 50$ trees for each response in the data model. This setting is also used for data model trees in the suBART models.

For the missingness model, missBART2 previously used $K_m = 20$ probit trees in univariate and bivariate settings. However, in Section 4.2.3, where the number of responses increased to 5, the number of probit trees in missBART2 was adjusted to 50 to account for the greater number of response variables within a multivariate tree structure. However, as missSUBART employs univariate trees which are expected to maintain convergence stability even if the number of responses increase, we fix the number of probit trees in the missingness model of missSUBART as $K_m = 20$ probit trees throughout MAR 2, MNAR 2, and MNAR 3.

In MNAR 3, the number of regression and missingness trees for missBART1 and missBART2 are kept the same as in Section 4.2, i.e., $K_y = 100$ multivariate regression trees and $K_m = 20$ multivariate probit trees. For MAR 2 and MNAR 2, the previously reported results for missBART1 and missBART2 are reproduced here for comparison.

5.4.1 MNAR 3 Simulation Details and Results

We first generate a complete dataset with $n = 2000$ i.i.d. samples using the "Friedman #1" simulation recipe from Esser et al. [2024], given by

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{pmatrix} = \begin{pmatrix} 10 \sin \mathbf{X}_{i1}\mathbf{X}_{i2}\pi + 20\left(\mathbf{X}_{i3} - 0.5\right)^2 \\ 8\mathbf{X}_{i4} + 20 \sin(\mathbf{X}_{i1}\pi) \\ 10\mathbf{X}_{i5} - 5\mathbf{X}_{i2} - 5\mathbf{X}_{i4} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix}, \tag{5.8}$$

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix} \sim \mathcal{N}_p \left(\mathbf{0}, \begin{pmatrix} 1 & 2 & 2.5 \\ 2 & 6.25 & 3.125 \\ 2.5 & 3.125 & 25 \end{pmatrix} \right)$$

where the $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(5)} \sim \text{Unif}(0, 1)$. Next, we simulate an extra set of 5 covariates, $\mathbf{X}^{(6)}, \ldots, \mathbf{X}^{(10)} \sim \text{Unif}(0, 1)$ and induce missingness by first simulating the $\mathbf{M}^*$:

$$
\begin{aligned}
M_{i1}^{*} &\sim \mathcal{N}\left(g_1\left(X_{i2}, X_{i8}; \mathcal{T}^{m_1}, Q^{m_1}\right), 1\right) \\
M_{i2}^{*} &\sim \mathcal{N}\left(\sum_{k=1}^{5} g_2\left(X_{i1}, X_{i2}, Y_{i1}, Y_{i2}; \mathcal{T}_k^{m_2}, Q_k^{m_2}\right), 1\right) \\
M_{i3}^{*} &\sim \mathcal{N}\left(\sum_{k=1}^{10} g_3\left(Y_{i3}; \mathcal{T}_k^{m_3}, Q_k^{m_3}\right), 1\right),
\end{aligned}
\tag{5.9}
$$

followed by computing $\mathbf{M}$ where $M_{ij} = 0$ if $M_{ij}^{*} \leq 0$ and $M_{ij} = 1$ if $M_{ij}^{*} > 0$. In other words, $\mathbf{Y}^{(1)}$ has MAR missingness influenced only by two covariates, $\mathbf{Y}^{(2)}$ is MNAR where its missingness is associated with two covariates, $\mathbf{Y}^{(1)}$ and its own values $\mathbf{Y}^{(2)}$, while missingness of $\mathbf{Y}^{(3)}$ is MNAR but only depends on its own values. The resulting dataset has a missingness proportion of 18.95% in $\mathbf{Y}^{(1)}$, 41.75% in $\mathbf{Y}^{(2)}$, and 15.25% in $\mathbf{Y}^{(3)}$. Additionally, the covariates $\mathbf{X}^{(6)}, \mathbf{X}^{(7)}, \mathbf{X}^{(9)}$, and $\mathbf{X}^{(10)}$ have no importance in either the data or missingness model, but are included as covariates in the competing models.

The out-of-sample RMSE, CRPS, and Frobenius norms for the 5 different models are shown respectively in Figure 5.1, Figure 5.2, and Figure 5.3 below. While missBART2 has poor performance in terms of Frobenius norms, we note that missBART2 performed reasonably well in terms of RMSE and CRPS for $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ but performed poorly in $\mathbf{Y}^{(3)}$. Overall, missSUBART demonstrated strong performance compared to the other models, both in making accurate predictions for the observed data and imputations for the missing values.

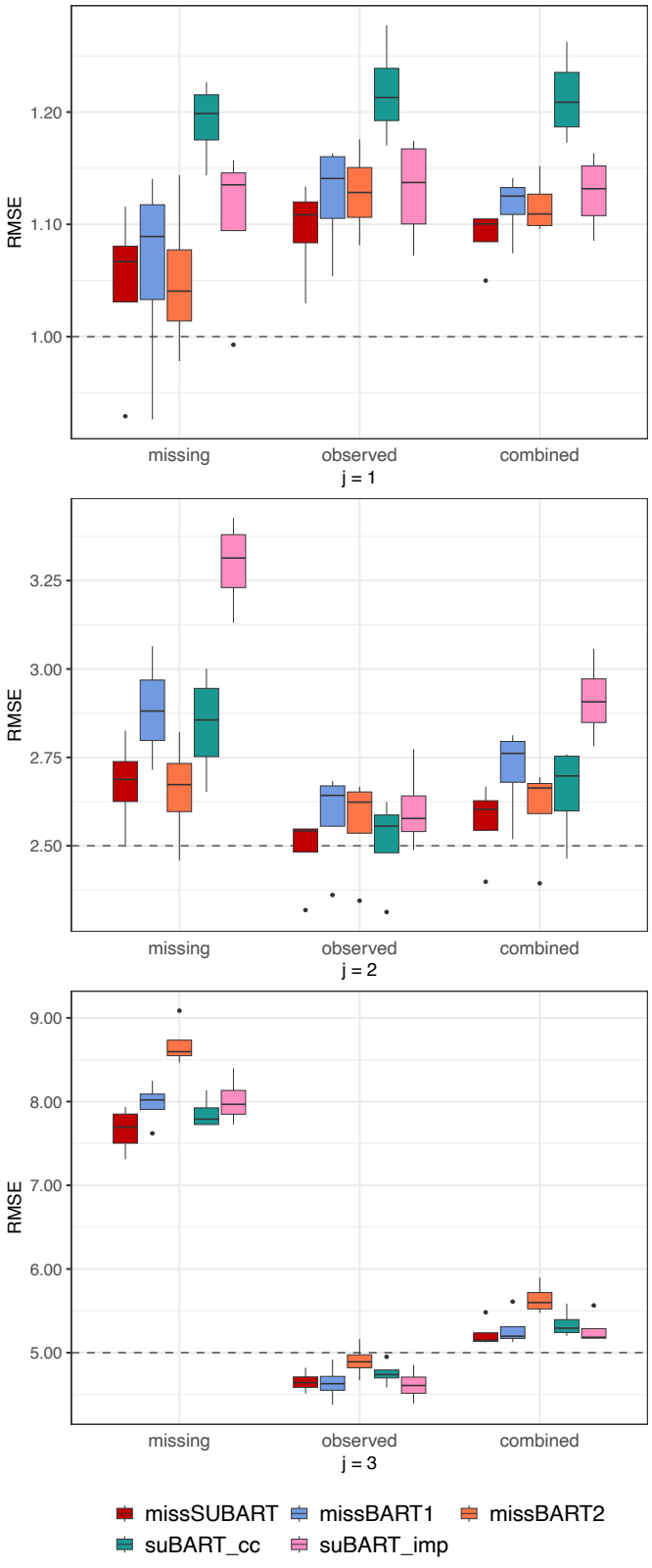


Figure 5.1: Out-of-sample RMSE for MNAR 3.

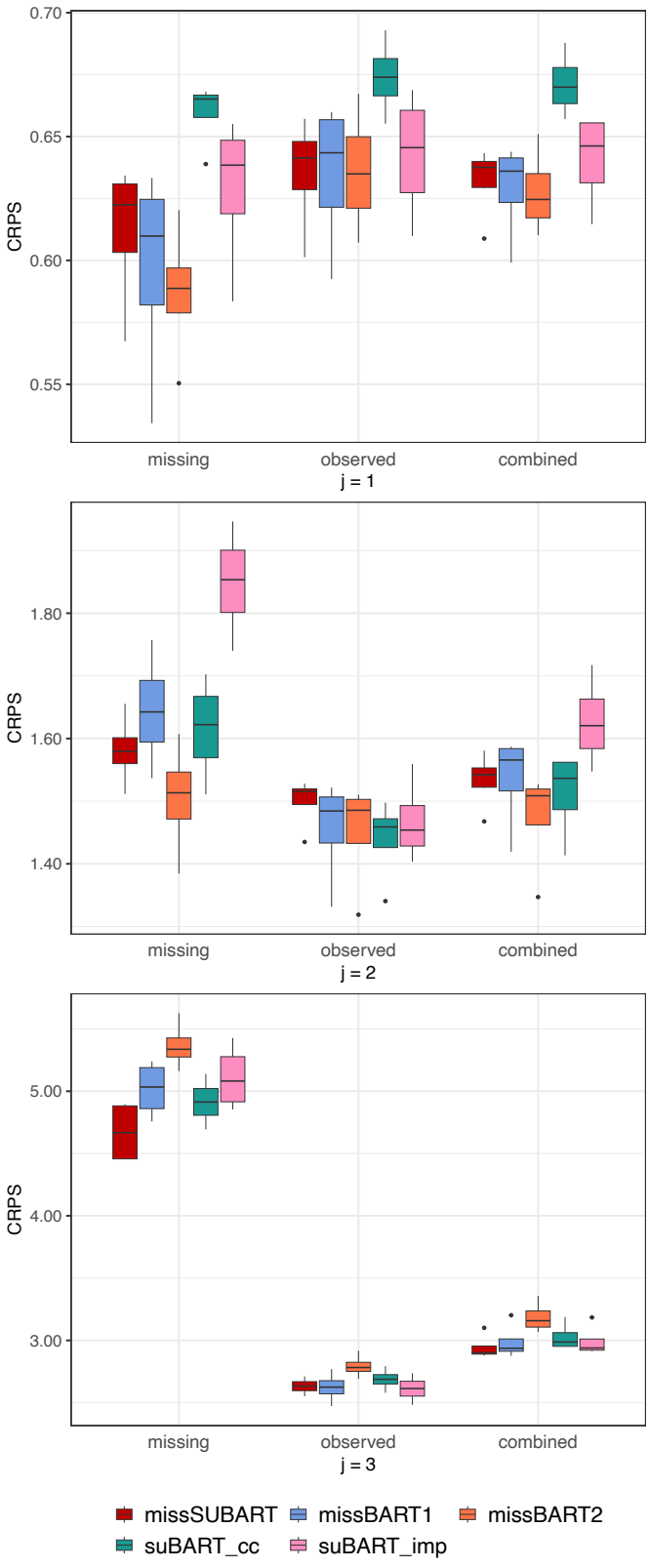


Figure 5.2: Out-of-sample CRPS for MNAR 3.

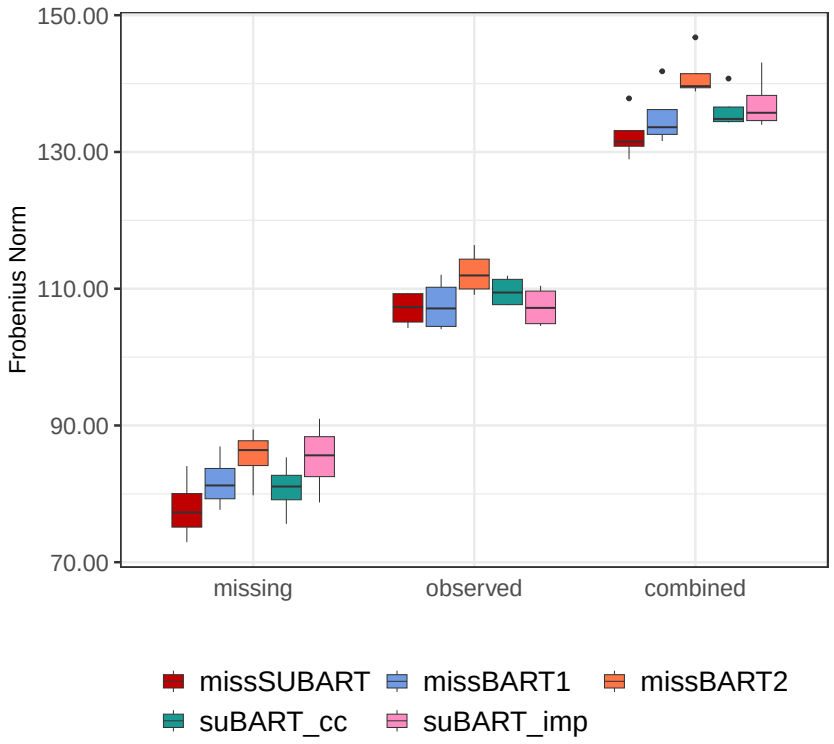
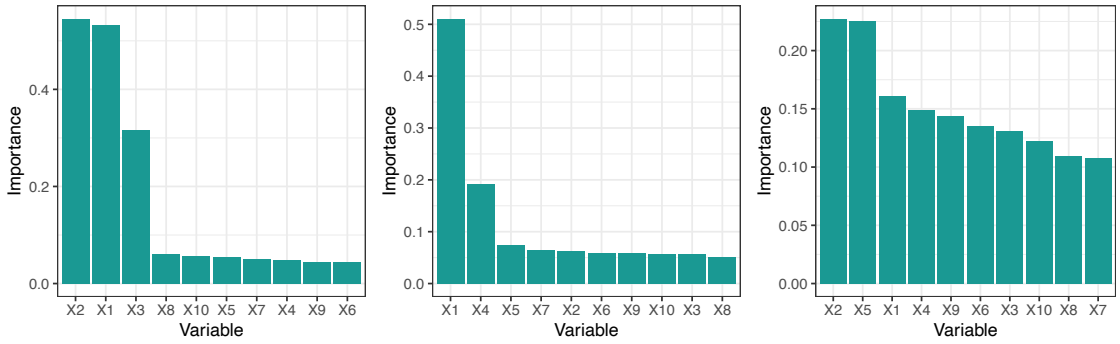
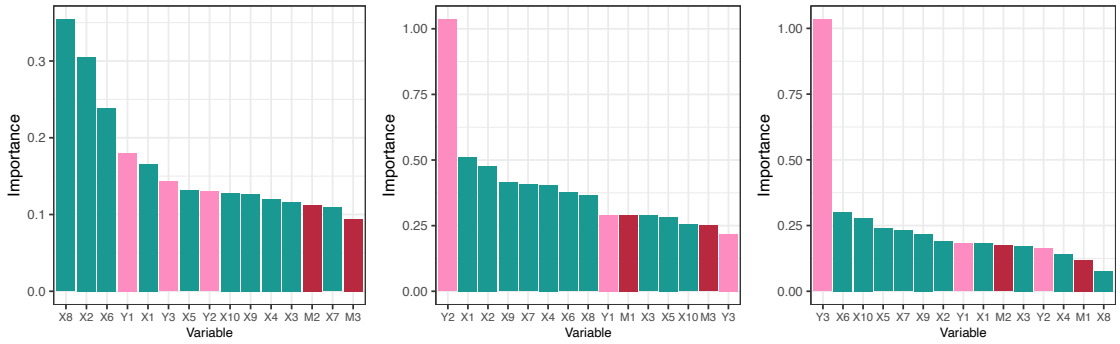


Figure 5.3: Out-of-sample Frobenius norms for MNAR 3.

As missSUBART employs separate tree structures for each response in both the data and missingness models, variable importance plots can be generated individually for each response and missingness indicator, rather than relying on a shared measurement across all responses as in missBART1 and missBART2. The variable importance plots from the regression and missingness trees of missSUBART, shown in Figure 5.4, confirm that the model performed well in identifying the covariates used in generating the complete responses in Friedman #1 from Equation (5.8), as well as the predictors used in simulating the missingness for each response from Equation (5.9). Specifically, $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, and $\mathbf{X}^{(3)}$ were the three most important covariates for predicting $\mathbf{Y}^{(1)}$; $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(4)}$ had the highest importance for $\mathbf{Y}^{(2)}$; and $\mathbf{X}^{(2)}$, $\mathbf{X}^{(4)}$, and $\mathbf{X}^{(5)}$ ranked among the top four most important covariates for $\mathbf{Y}^{(3)}$. In the missingness model, $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(8)}$ were the most important predictors for the missingness in $\mathbf{Y}^{(1)}$. For $\mathbf{Y}^{(2)}$, $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, and $\mathbf{Y}^{(2)}$ ranked among the top three most important predictors, whereas $\mathbf{Y}^{(1)}$ had comparatively lower importance. Finally, the missingness in $\mathbf{Y}^{(3)}$ was predominantly influenced by $\mathbf{Y}^{(3)}$ itself, which had the highest importance, far exceeding that of all other predictors.



(a) Variable importance from the 3 sets of regression trees in missSUBART.



(b) Variable importance from the 3 sets of missingness trees in missSUBART.

Figure 5.4: Variable importance of the regression and missingness trees for each of the three responses and missingness indicators modelled in missSUBART.

5.4.2 MAR 2 and MNAR 2 Results

The out-of-sample RMSE, CRPS, and Frobenius norms for MAR 2 and MNAR 2, shown for the observed, missing, and combined responses, are presented in Figure 5.5 and Figure 5.6 below, respectively. In both examples, missBART2 outperformed all other models, as expected, since the data and missingness mechanisms in MAR 2 and MNAR 2 were simulated using multivariate BART models with shared tree structures across responses. missSUBART demonstrated decent performance in both scenarios. Under MAR 2, missSUBART had comparable results with missBART2. Under MNAR 2, missSUBART outperformed missBART1 for the response $\mathbf{Y}^{(2)}$. Further, as anticipated, the complete-case and imputed suBART models performed poorly due to their inability to account for the missingness mechanisms. In fact, some suBART_cc results were omitted from Figure 5.6 (RMSE and CRPS, $j = 2$) due to poor performance.

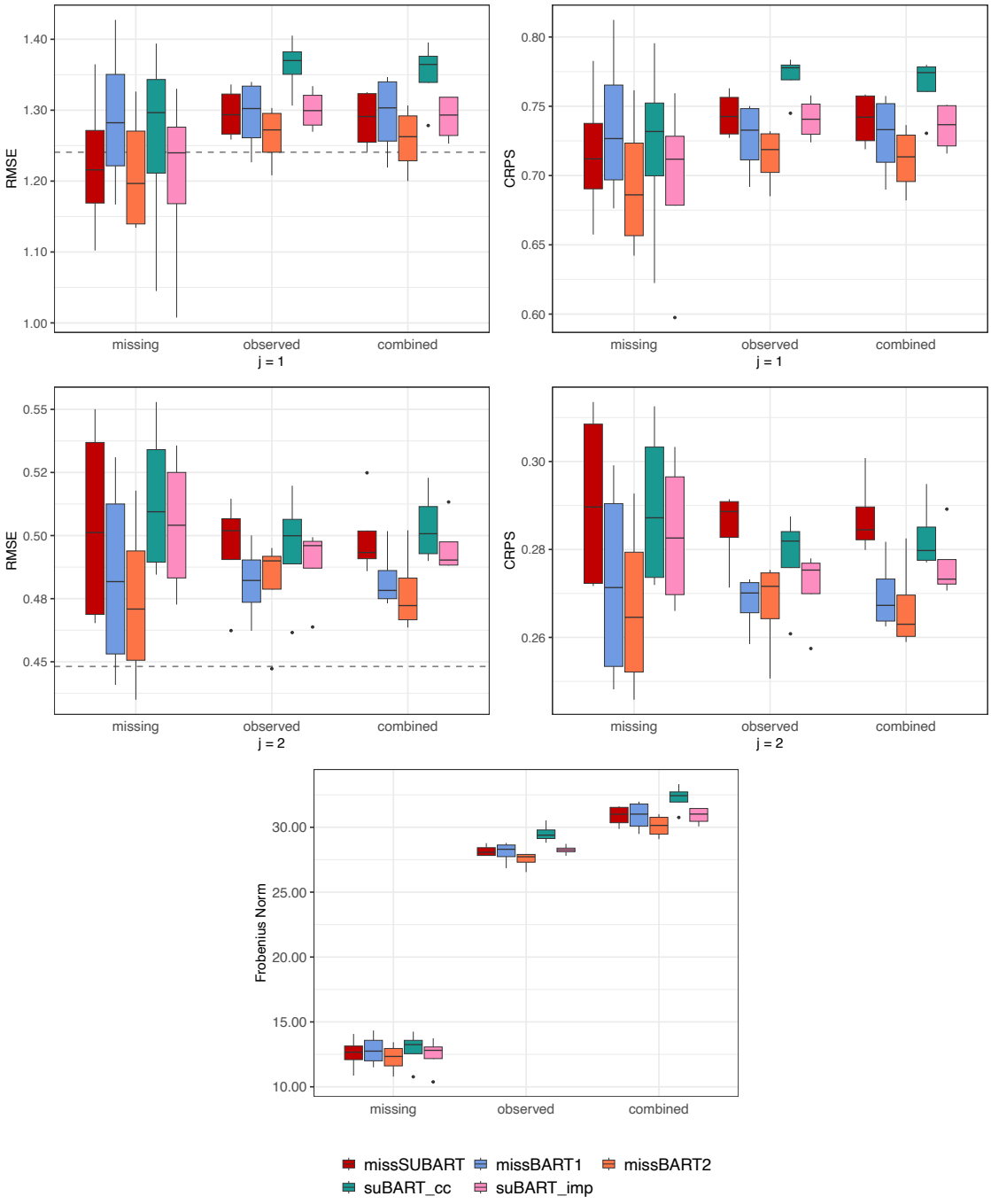


Figure 5.5: Out-of-sample RMSE, CRPS, and Frobenius norms for MAR 2. The joint models show comparable performances.

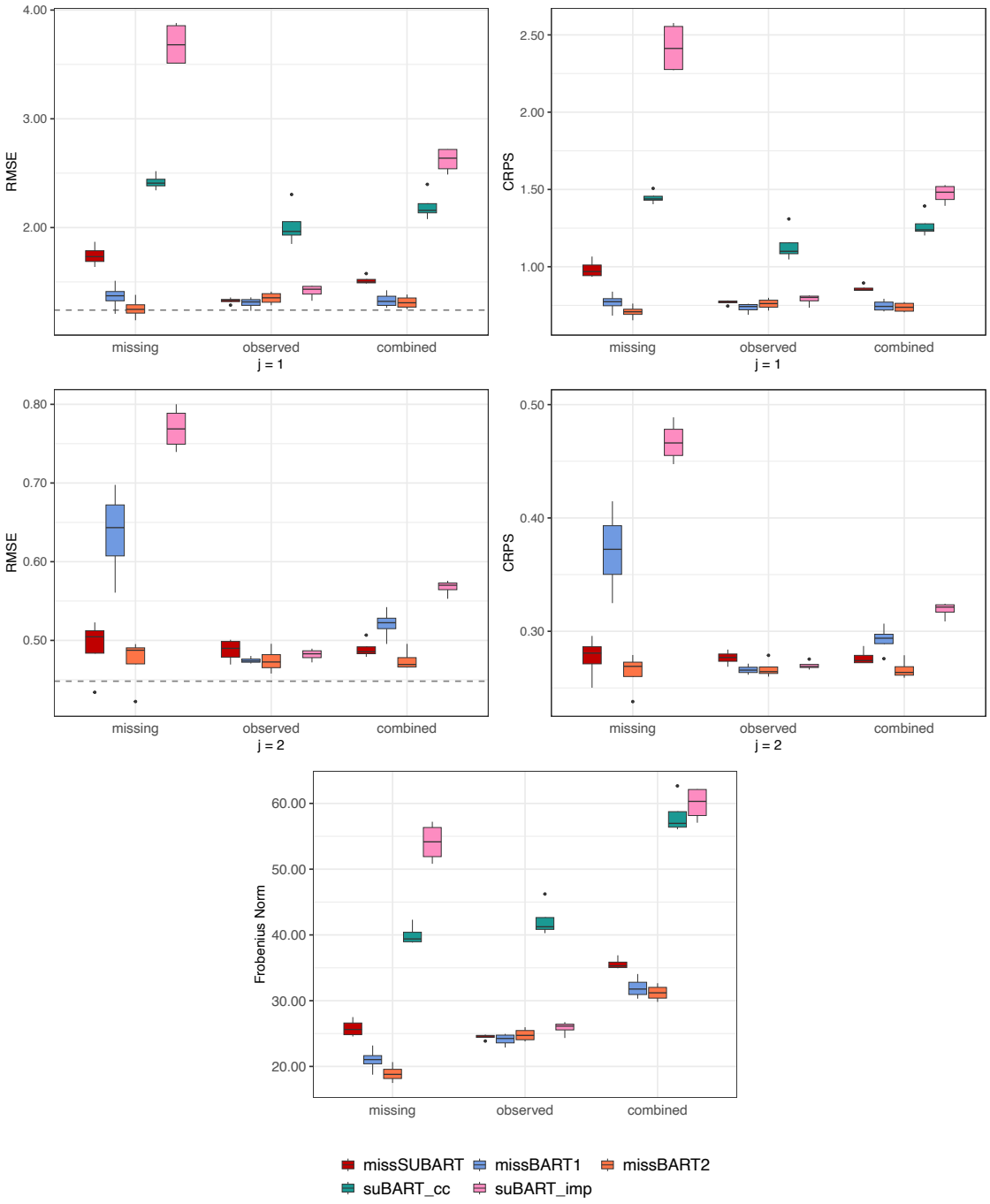


Figure 5.6: Out-of-sample RMSE, CRPS, and Frobenius norms for MNAR 2. missSUBART outperformed missBART1 for $\mathbf{Y}^{(2)}$ but not for $\mathbf{Y}^{(1)}$.

5.5 Application to *global Amax*

We now apply missSUBART to the *global Amax* data. For a more direct comparison with results from missBART2, we first present, in Section 5.5.1, results obtained from excluding the missingness indicators $\mathbf{M}^{(-j)}$ from the missingness model for each j^{th} response in missSUBART. Following this, in Section 5.5.2, we incorporate $\mathbf{M}^{(-j)}$ back into the missingness model and discuss results.

As in Section 4.3, the responses undergo a log-transformation to account for right-skew. The model is run for 5000 burn-in and post-burn-in iterations, and 50 regression and 20 missingness trees are used. While we increased the number of missingness trees in missBART2 in Sections 4.2 and 4.3, missSUBART assigns separate sets of trees to each response variable rather than fitting all responses within a shared tree structure. Consequently, we believe that the number of trees does not need to scale with the number of responses, as was necessary for missBART2.

5.5.1 missSUBART without Missing Indicators in Missingness Model

The predictions are shown for the observed data against their true log-transformed values in Figure 5.7, as well as rug plots showing the posterior mean imputations and vertical bars depicting the 95% prediction intervals. These predictions are comparable to those from missBART1 in Figure 4.13 and missBART2 in Figure 4.14. In terms of posterior imputations, missSUBART has a smaller range of values in comparison with that from the previous models, and seem to stay roughly around the mean of the observed data.

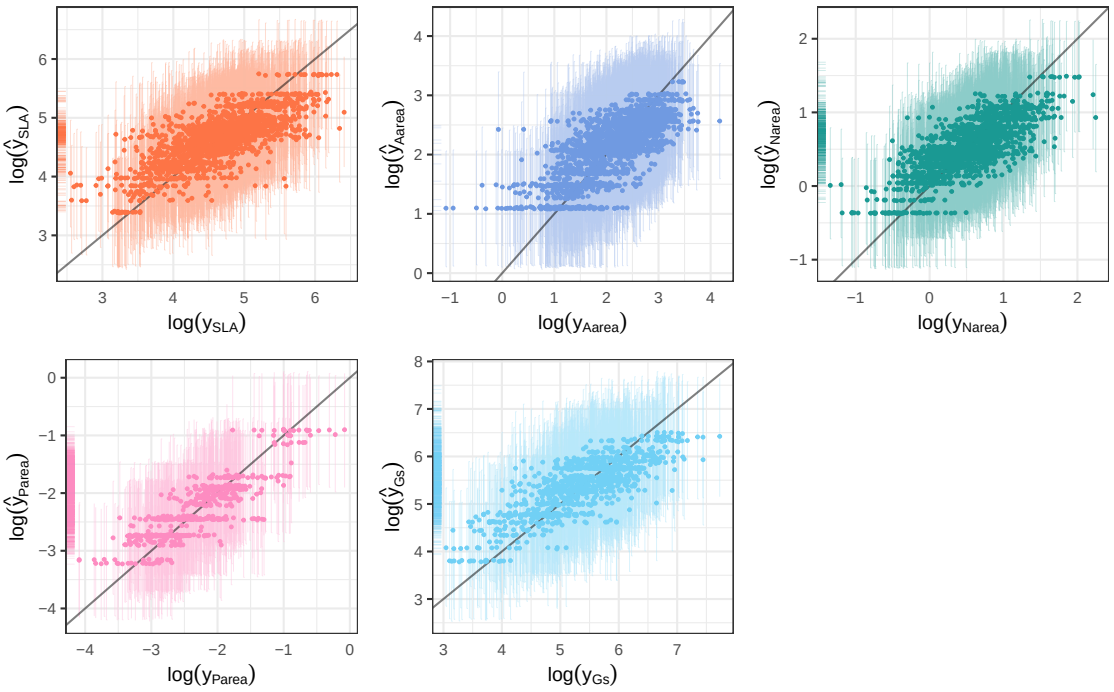


Figure 5.7: Predictions for the observed data from missSUBART without missingness indicators in the missingness model against their true log-transformed values. Vertical error bars represent the 95% prediction intervals for the observed data. Rug plots on the $y-$axes show the posterior means of the missing data imputations. Aside from *Gs*, the imputations mostly lie within the range of the observed data.

Unlike missBART2, where variable importance is shared across all responses, missSUBART allows for the evaluation of distinct variable importance measures for each response. Figure 5.8 presents the variable importance from this missSUBART model, including only the top 10 important covariates, derived from the five sets of regression trees. *ALU* appears in most plots, except for *SLA*, while *pH* is commonly used across responses, apart from *Parea* and *Gs*. Precipitation-related variables, including *PPTmax*, *PPTmin*, *PPTmean*, and *PPTseason*, frequently contribute to all responses. Similarly, fractional sunshine duration variables, such as *SUNmax*, *SUNmin*, and *SUNrange*, are important for all responses except *Parea*. Furthermore, all of the top 10 most important variables identified by missBART2 (see Figure 4.16) appear in at least one of the plots in Figure 5.8, reinforcing their relevance in the regression models.

To further investigate the differences in which variables are deemed to be relevant under missBART2 and missSUBART, the differences between the variable importance scores of the five sets of regression trees from missSUBART and the single shared set of regression trees from missBART2 (from Section 4.3) are illustrated in Figure 5.9. The differences are shown for each response variable as missSUBART assigns a distinct set of trees to each of the five responses, resulting in five sets of variable importance measures. Additionally, Figure 5.9 shows results only for the top 10 variables ranked most impor-

tant by the single set of missBART2 regression trees. Positive values indicate that the variable was more important under missSUBART, while negative values indicate higher importance under missBART2. To facilitate comparison, variable importance scores from each method were adjusted to a comparable scale prior to calculating the differences.

Notably, *ALU* was consistently more important in several of the missSUBART trees compared to missBART2, particularly for *Aarea*, *Narea*, *Parea*, and *Gs*. While *SUNmin* was considered more important for *SLA* and *Narea*, it was less important for *Aarea* and *Parea*. However, the magnitude of the differences in variable importance scores is consistently small, with absolute differences no greater than approximately 0.03 across all five responses and the top ten predictors.

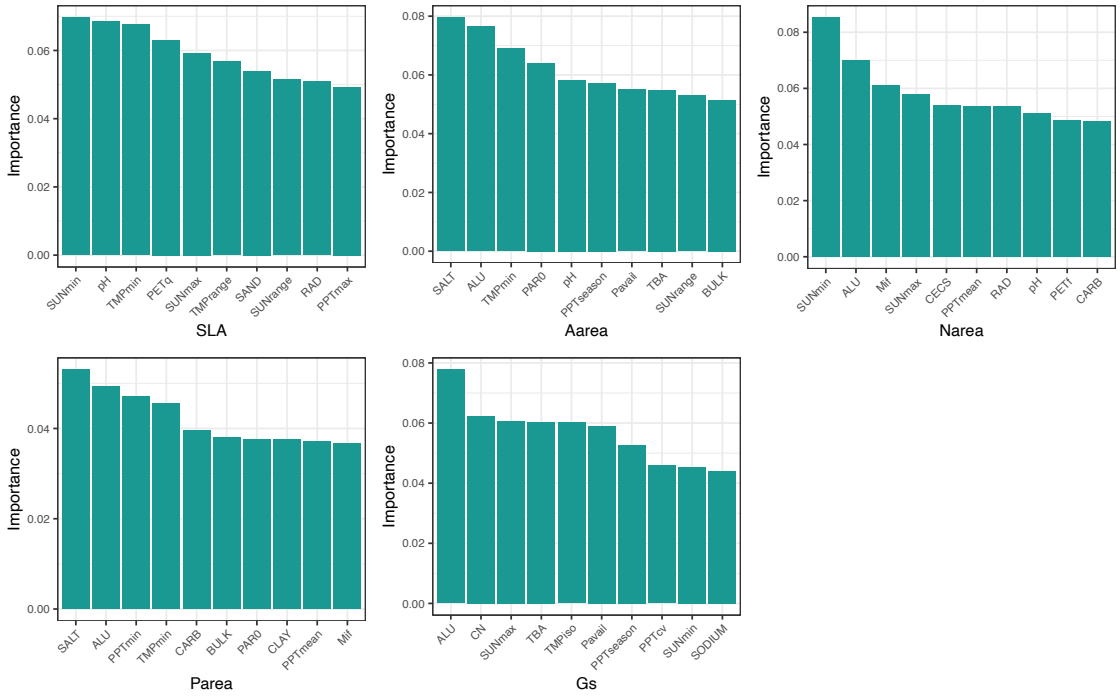


Figure 5.8: Variable importance from the regression trees of the missSUBART model which excludes $\mathbf{M}^{(-j)}$ from the missingness model predictors.

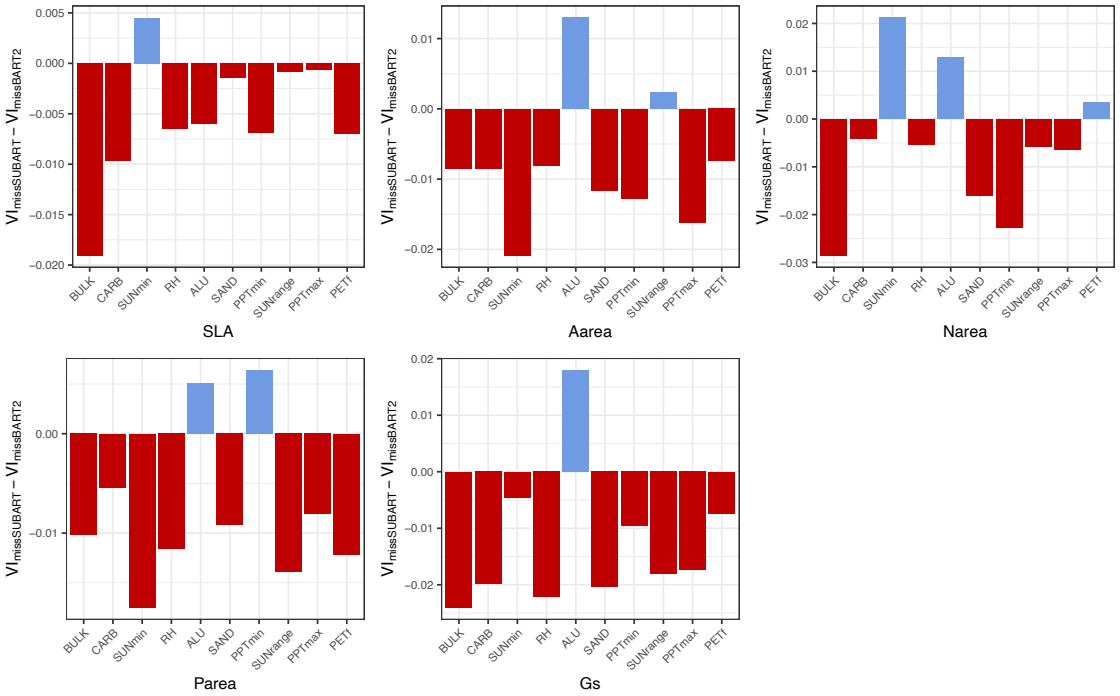


Figure 5.9: Differences between variable importance from the 5 sets of missSUBART trees and single set of missBART2 trees, for the top 10 most important variables from missBART2. Positive values indicate higher importance under missSUBART; negative values indicate higher importance under missBART2.

As before, we further examine the relationships between responses and key variables identified in Maire et al. [2015] using PDP and ICE curves. According to Maire et al. [2015], *Aarea*, *Narea*, and *Parea* increased, while *SLA* decreased as *pH* increased and *Miq* decreased. Additionally, *Parea* increased, whereas *Gs* decreased with increasing *Pavail*. From missBART2, the results shown in Figure 4.19 largely align with Maire et al. [2015], except for the relationship between *Parea* and *Pavail*. missBART2 also reveals additional associations not detected in the original analysis of the *global Amax* data, where *Gs* increased with increasing *pH* and *Miq*, which was not observed in the previous study.

The PDP and ICE curves from this missSUBART model are shown in Figure 5.10 for *pH*, *Miq*, and *Pavail*. As *pH* increases, *Aarea*, *Narea*, and *Gs* exhibit a general upward trend. *SLA* shows a slight dip between *pH* values of 6.5 and 7.5, before stabilising, while *Parea* demonstrates a minimal increase at lower *pH* levels but remains relatively unchanged thereafter. With increasing *Miq*, *Aarea* decreases, while *SLA*, *Narea*, and *Parea* show only marginal increases. The relationship between *Miq* and *Gs* is non-linear, with values between 1 and 5 corresponding to slightly higher stomatal conductance before decreasing again. Finally, as *Pavail* increases from 0 to 300, *Narea* exhibits a slight upward trend, whereas *Gs* declines steadily. Other responses, including *SLA*, *Aarea*, and *Parea*, remain largely unaffected across the range of *Pavail* values.

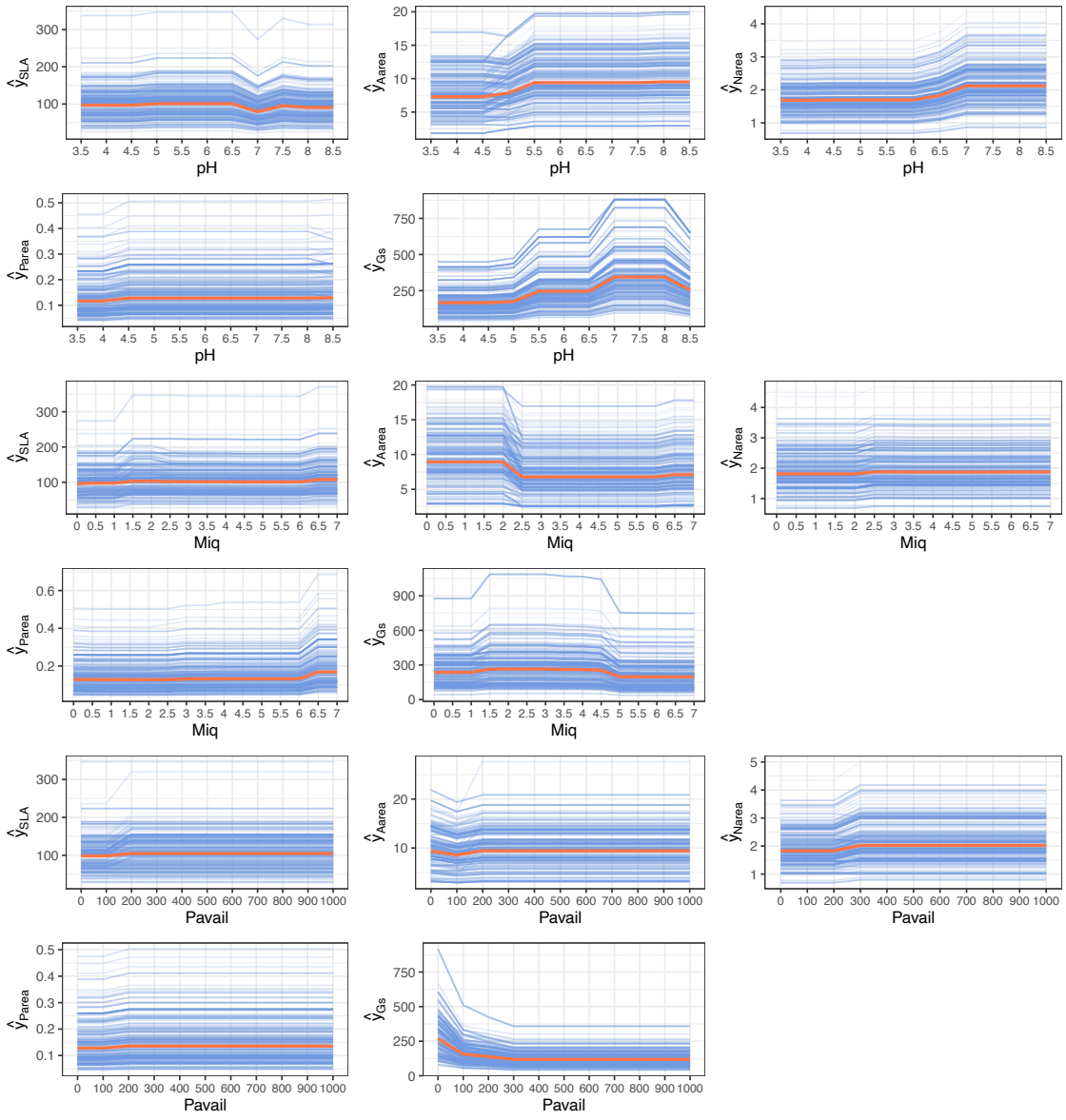


Figure 5.10: PDP + ICE curves from the data model of missSUBART without $\mathbf{M}^{(-j)}$ in the missingness model predictors, across different levels of *pH*, *Miq*, and *Pavail*. The PDP curves (in orange) show the overall average effect of the covariates on the response variables, while the ICE curves (in blue) illustrate how the covariate influences individual predictions across different observations.

The missSUBART analysis largely aligns with Maire et al. [2015], confirming that *Aarea* and *Narea* increase with *pH*, *Aarea* decreases as *Miq* increases, and *Parea* increases while *Gs* decreases with *Pavail*. However, missSUBART reveals further relationships such as an increasing trend in *Narea* as *Pavail* increases, as well as non-linear trends between *Gs* and two covariates *pH* and *Miq*. This underscores the greater flexibility of the tree-based model compared to the linear relationships described by Maire et al. [2015], while also enabling additional inferences through the joint model's simultaneous data imputation.

The variable importance of the 5 sets of missingness trees in missSUBART, using only $\mathbf{X}$ and $\mathbf{Y}$ as predictors in the missingness model, are shown in Figure 5.11. Previously, variable importance in missBART2 indicated that *Parea* played a significant role in the missingness model. Here, *Parea* appears only in the importance plot for *Aarea*, yet it is the second most influential variable. No other response variables are present in the importance plots for the remaining responses, suggesting that the missingness mechanisms for all responses except *Aarea* are more likely to follow a MAR pattern. This highlights the benefit of separately modelling the missingness mechanisms, allowing for a more nuanced understanding of how missing data patterns vary across responses.

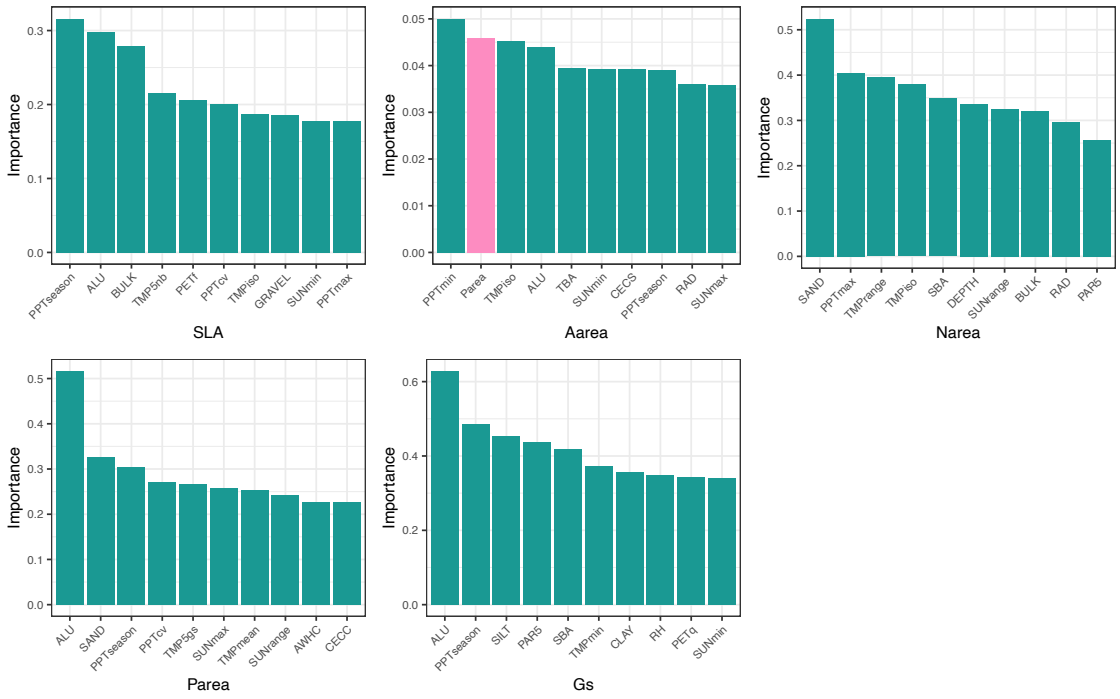


Figure 5.11: Variable importance from the missingness trees of the missSUBART model which excludes $\mathbf{M}^{(-j)}$ from the missingness model predictors.

5.5.2 missSUBART with Missing Indicators in Missingness Model

As before, Figure 5.12 presents the predictions for the observed data against their true log-transformed values, along with posterior mean imputations illustrated using rug plots, obtained from missSUBART with missingness indicators incorporated in the missingness model. We see no observable difference between the predictions and imputations from this model as compared to the previous model shown in Figure 5.7.

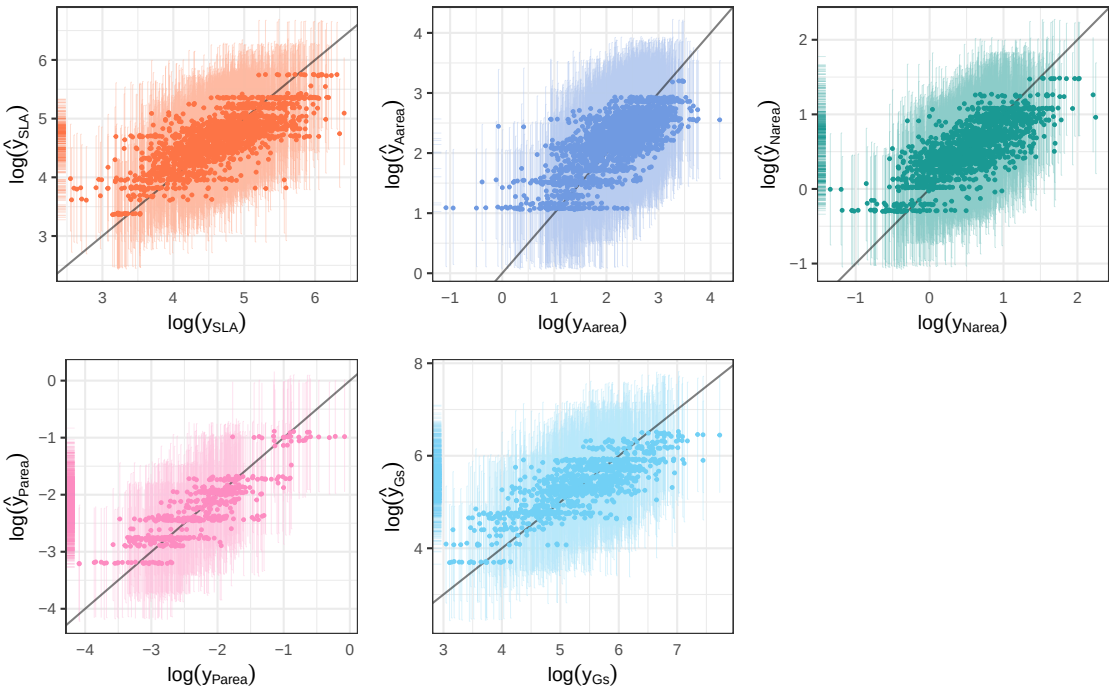


Figure 5.12: Predictions for the observed data from missSUBART with missingness indicators in the missingness model against their true log-transformed values. There is no observable difference between these results and that from the previous missSUBART model shown in Figure 5.7.

As before, we assess variable importance using the variable importance plots presented in Figure 5.13. *pH* ranks among the top 10 most important variables for all five responses. Variables related to fractional sunshine duration, including *SUNmax*, *SUNmean*, *SUNmin*, and *SUNrange*, frequently appear across all responses except for *Gs*. *CN* emerges as the most influential predictor for both *Aarea* and *Narea*, while also ranking among the top 10 variables for *Gs*. Compared to the previous model without missingness indicators in the missingness model (Figure 5.8), ALU now appears only in the importance plots for *SLA* and *Parea*, and precipitation-related variables are less frequently used.

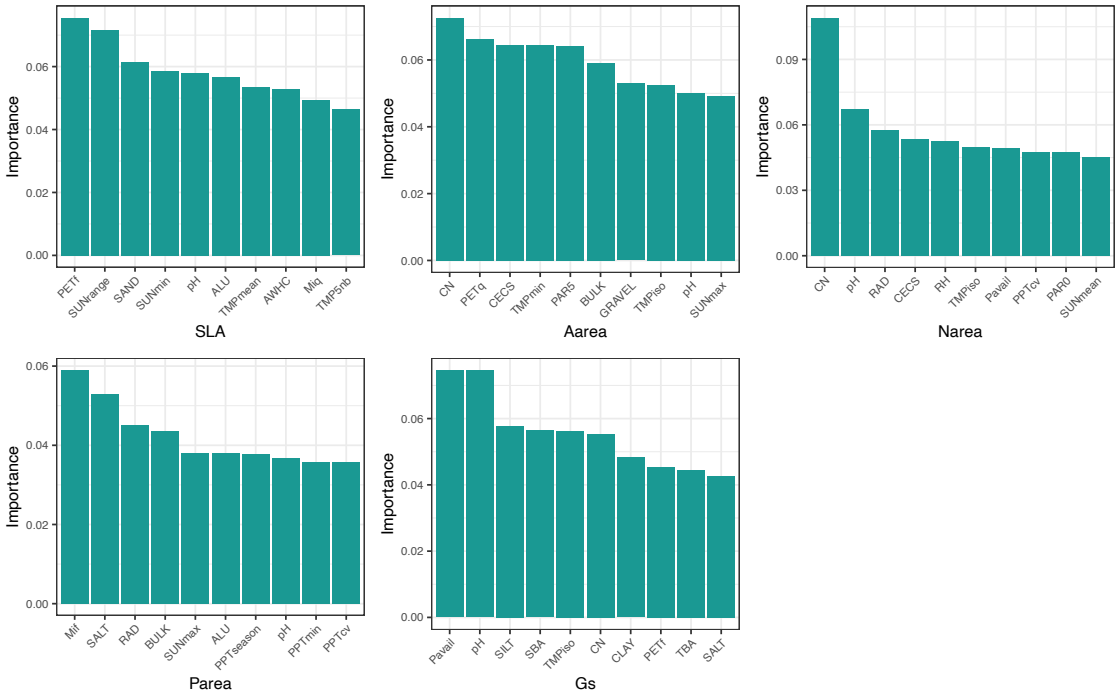


Figure 5.13: Variable importance from the regression trees of the missSUBART model with $\mathbf{M}^{(-j)}$ included as missingness model predictors.

In terms of PDP and ICE curves, shown for this model in Figure 5.14, *Narea*, *Parea*, and *Gs* show an increase with increasing *pH*, *SLA* increases and decreases between *pH* values of 4.5 and 7, while *Aarea* shows virtually no difference. As *Miq* increases, *SLA*, *Aarea*, and *Parea* show some sort of increasing trend, *Gs* decreases, while *Narea* remains unchanged. Finally, *SLA* and *Aarea* show little change with increasing *Pavail*, *Parea* increases while *Gs* decreases as *Pavail* increases from 0 to 200.

The results from this missSUBART model are largely consistent with the previous missSUBART analysis and Maire et al. [2015], confirming that *Narea* and *Parea* increase with *pH*, while *Gs* decreases as *Pavail* increases. However, some notable differences emerge. Unlike the previous missSUBART model, which showed a dip in *SLA* between *pH* values of 6.5 and 7.5, this model finds a rise and fall in *SLA* between *pH* values of 4.5 and 7. Additionally, while both missSUBART models identified a declining trend in *Gs* with *Miq*, the previous model suggested a non-linear response with a mid-range peak, whereas this model indicates a more consistent decreasing trend. Furthermore, *Aarea* showed no significant response to *pH* in this analysis, differing from both Maire et al. [2015] and the previous missSUBART results.

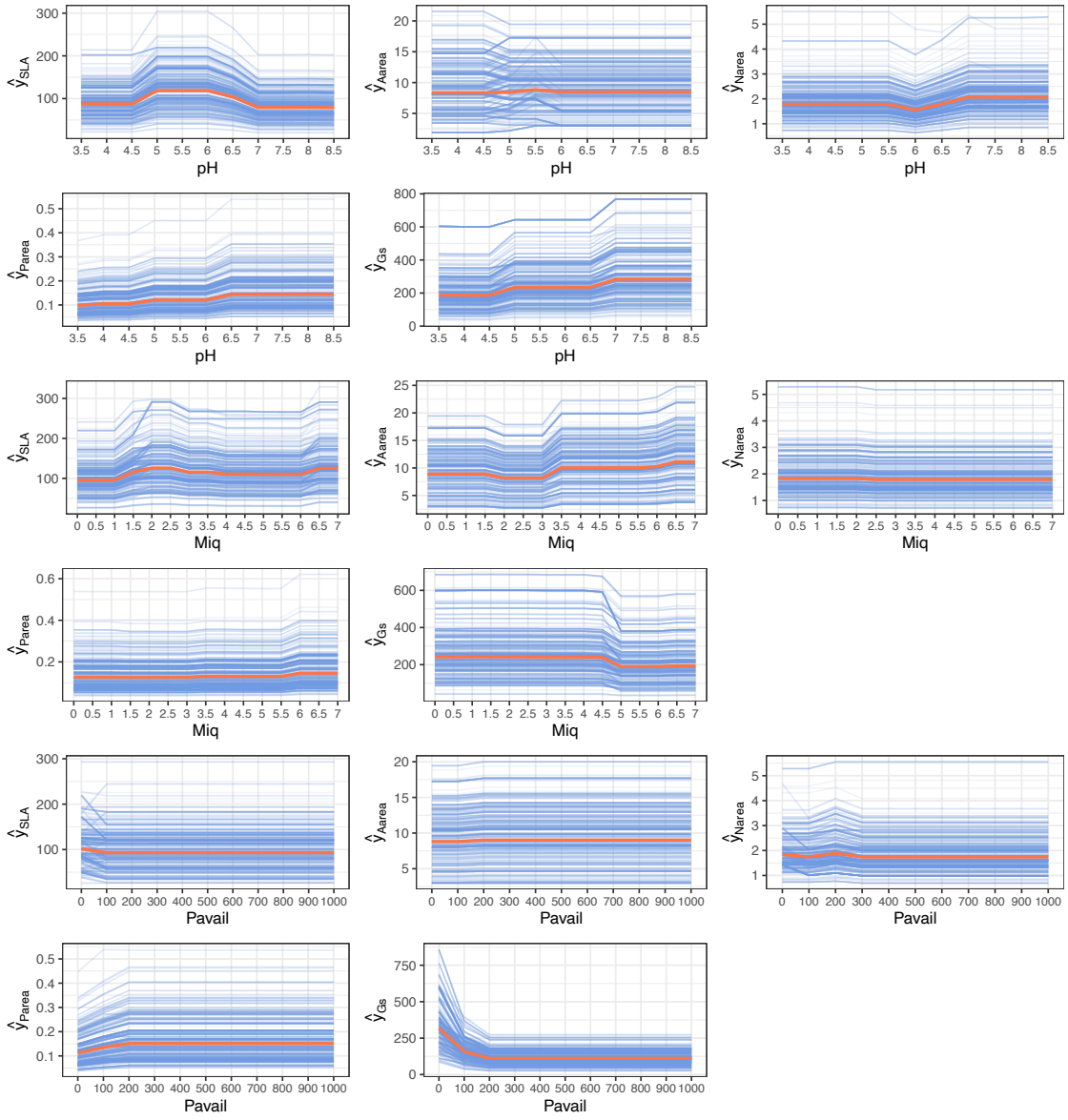


Figure 5.14: PDP + ICE curves from the data model of missSUBART when $\mathbf{M}^{(-j)}$ is incorporated into the missingness model predictors, across different levels of *pH*, *Miq*, and *Pavail*. The PDP curves (in orange) show the overall average effect of the covariates on the response variables, while the ICE curves (in blue) illustrate how the covariate influences individual predictions across different observations.

Figure 5.15 shows the top 10 important variables obtained from the missingness trees in missSUBART where $\left(\mathbf{X}, \mathbf{Y}, \mathbf{M}^{(-j)}\right)$ are used as missingness model predictors. Notably, none of the responses are included amongst the top 10 important variables, while the missingness indicators of all responses, apart from *Aarea*, are commonly used to predict the missingness in other responses. The missingness in *Narea*, denoted as 'M(Narea)', was the most important predictor in predicting the missingness in all other responses. Missingness in *Gs* was also an important predictor for most responses other than *Aarea*.

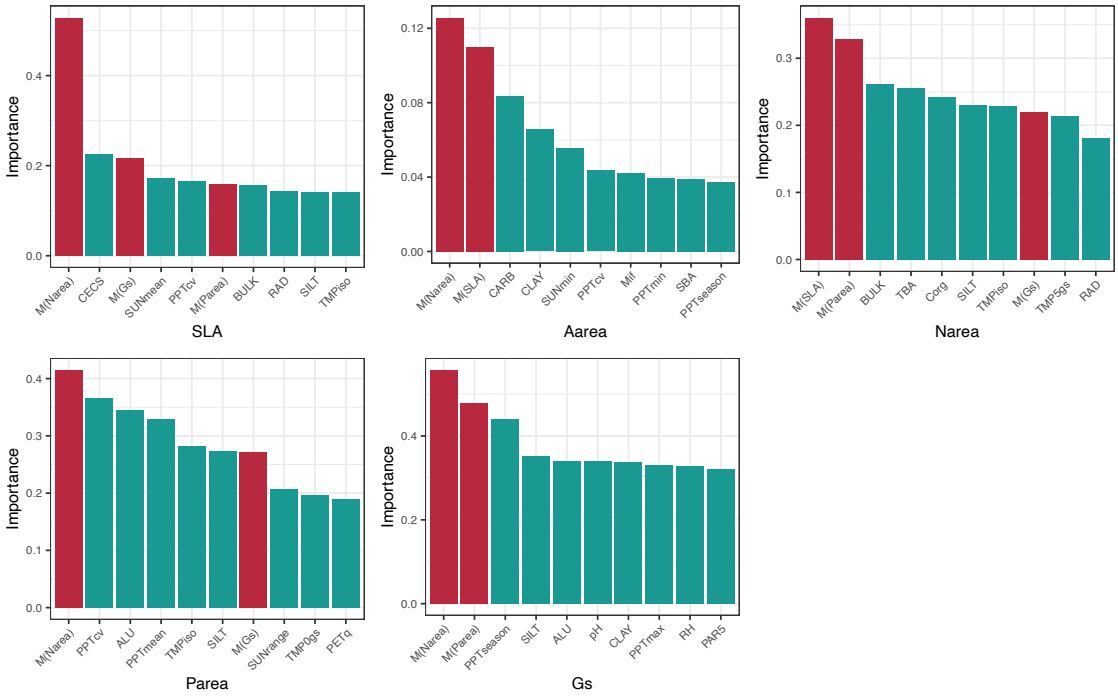


Figure 5.15: Variable importance from the missingness trees of the missSUBART model with $\mathbf{M}^{(-j)}$ included as missingness model predictors.

Since the missingness indicators for each response is used to model the missingness of other responses, we can further investigate this relationship to explore how the missingness of a response variable affects whether or not a different response will be observed. Using PDP and ICE curves, we first look at the most influential predictor, M(Narea), and how it influences the missingness probabilities of *SLA*, *Aarea*, *Parea*, and *Gs*. This is shown in Figure 5.16. Overall, all four of these responses are more likely to be observed when *Narea* is observed.

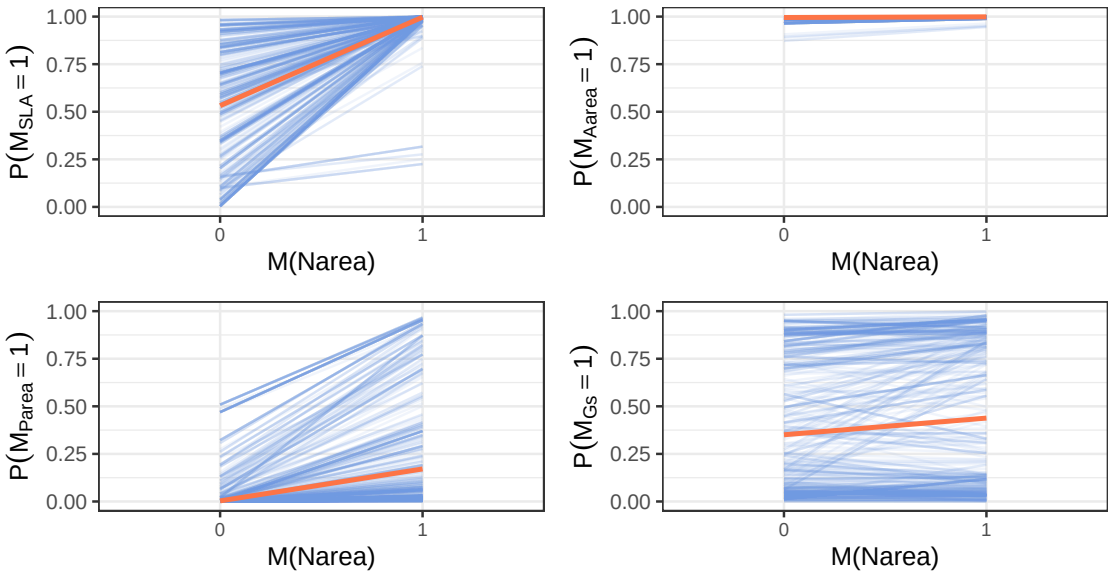
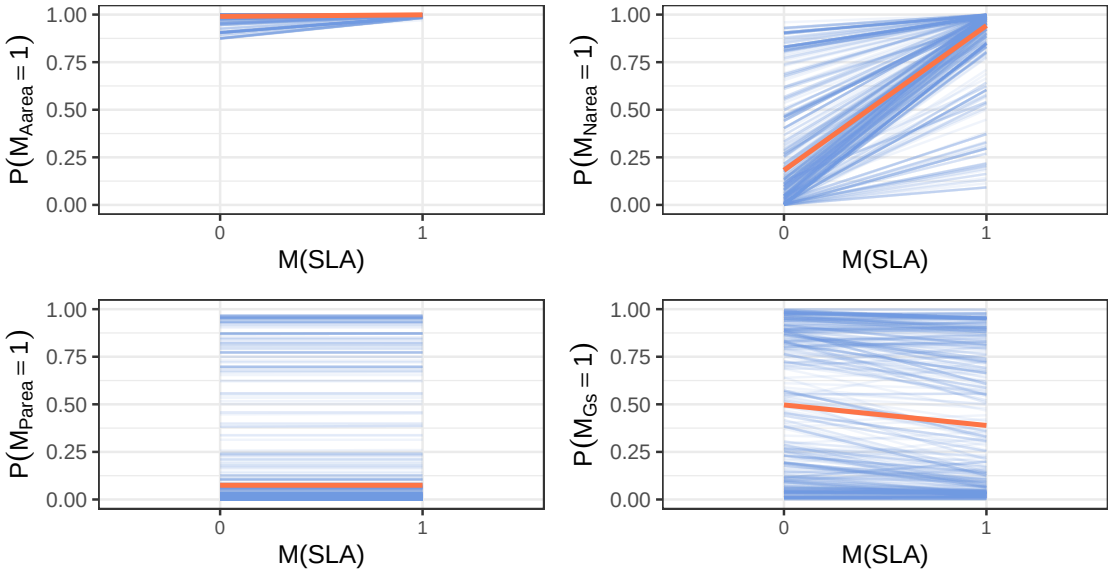
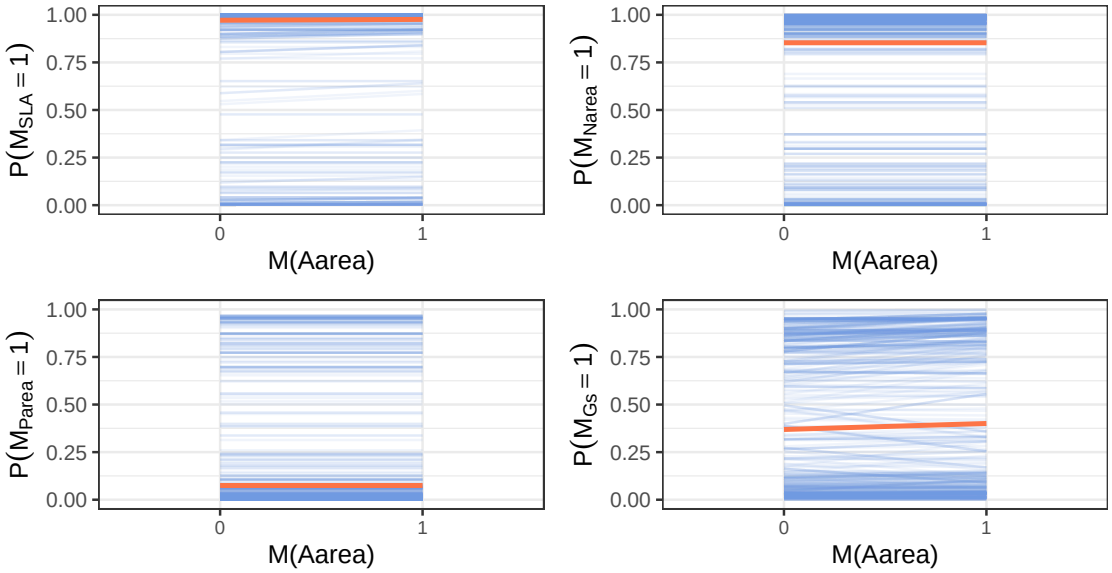


Figure 5.16: PDP + ICE curves from the missingness model in missSUBART when $\mathbf{M}^{(-j)}$ is incorporated into the missingness model predictors, with respect to the binary missingness indicator for *Narea*. $P(M_{\text{Narea}})$ is omitted as it cannot depend on M(Narea).

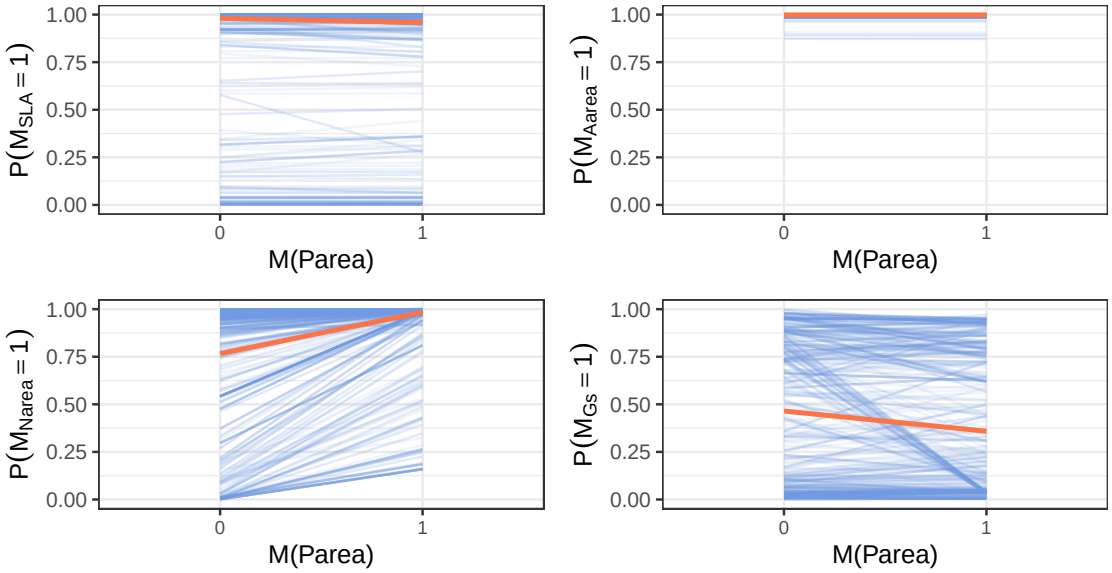
The PDP and ICE curves with respect to the missingness indicators of all other responses are shown in Figure 5.17. When *SLA* is observed, *Narea* is more likely to be observed, whereas *Gs* has a lower probability of being observed. *Aarea* appears to have no significant effect on the missingness probabilities of other responses, which is consistent with its absence from the variable importance plots in Figure 5.15. Similarly, when *Parea* is observed, *Narea* is more likely to be observed, while *Gs* is less likely. Finally, *Gs* has minimal influence on most missingness probabilities, except for a slightly increased chance of *Narea* being observed when *Gs* is also observed.



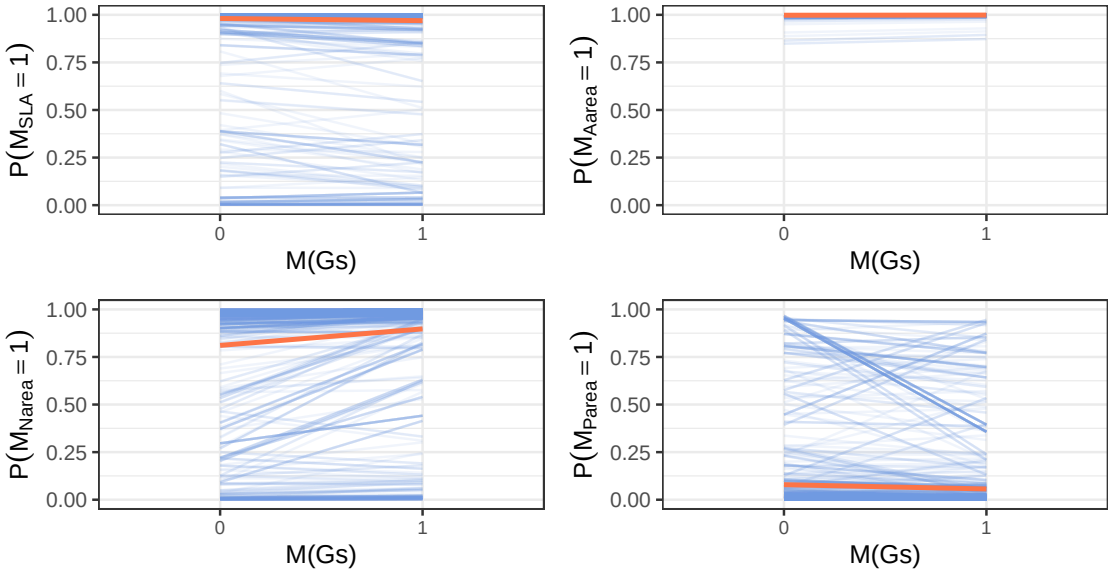
(a) PDP + ICE curves with respect to the binary missingness indicator for *SLA*. When *SLA* is observed, *Narea* is more likely to be observed, whereas *Gs* has a lower probability of being observed.



(b) PDP + ICE curves with respect to the binary missingness indicator for *Aarea*. *Aarea* appears to have no significant effect on the missingness probabilities of other responses.



(c) PDP + ICE curves with respect to the binary missingness indicator for *Parea*. When *Parea* is observed, *Narea* is more likely to be observed, while *Gs* is more likely to be missing.



(d) PDP + ICE curves with respect to the binary missingness indicator for *Gs*. *Gs* has minimal influence on most missingness probabilities, except for a slightly increased chance of *Narea* being observed when *Gs* is also observed.

Figure 5.17: PDP + ICE curves from the missingness model in missSUBART when $\mathbf{M}^{(-j)}$ is incorporated into the missingness model predictors, with respect to the binary missingness indicators for *SLA*, *Aarea*, *Parea*, and *Gs*.

Maire et al. [2015] compiled photosynthetic traits from various studies, with *Aarea* as the primary focus, resulting in minimal missing data for this trait. This is consistent with the observation that *Aarea* has virtually no influence on the missingness of other responses and is itself unaffected by their missingness. Additionally, *SLA*, being relatively

easy to measure, is commonly included in studies assessing *Aarea*, while *Narea*, a key trait in photosynthesis, is also frequently quantified in research related to plant function. The observed missingness patterns further support these explanations. When *SLA* is recorded, *Narea* is more likely to be observed, reinforcing their frequent co-measurement in photosynthesis studies. Likewise, *Parea* and *Narea* exhibit a similar relationship, suggesting that studies measuring phosphorus content often also quantify nitrogen due to their interconnected roles in plant metabolism. Notably, *Gs* is less likely to be recorded when *SLA* or *Parea* is measured, implying that stomatal conductance is not prioritised in studies focused on leaf morphological or nutrient-related traits.

5.6 Discussion

In this chapter, we introduced missSUBART, a novel joint modelling approach for multivariate response data with missingness that avoids imposing strict assumptions on missingness mechanisms, enforcing strong linear relationships, or constraining responses to a shared tree structure. Building on the seemingly unrelated BART framework, missSUBART addresses key limitations of previous models, missBART1 and missBART2, by allowing each response and missingness indicator to follow distinct tree structures while simultaneously modelling their correlations. This flexibility enhances the accuracy of missing data imputation and improves the identification of missingness mechanisms, particularly in cases where responses have unique predictor associations. However, this increased flexibility comes at the cost of greater computational complexity. The need for separate sets of trees for both the response and missingness models results in a higher number of parameters and updates per iteration, making missSUBART computationally demanding, especially in high-dimensional settings. While this trade-off is manageable for datasets with a moderate number of responses, it may limit the scalability of the model in applications with a large number of response variables.

Through extensive simulation studies, we demonstrated that missSUBART outperforms competing models in terms of predictive accuracy and missingness recovery. When responses share similar tree structures (e.g., MAR 2 and MNAR 2), missSUBART performs comparably to missBART2, which enforces a shared tree structure across responses. However, missSUBART provides additional flexibility, allowing it to model heterogeneous predictor-response relationships more effectively. In MNAR 3, where each response follows distinct missingness patterns, missSUBART outperformed all other methods, recovering the underlying missingness structures while making accurate predictions and imputations. When applied to the global Amax dataset, missSUBART produced results largely consistent with Maire et al. [2015], while uncovering additional relationships that were not detected in previous analyses.

A key advantage of missSUBART is its ability to incorporate missingness indicators as predictors in the missingness model—something that was not feasible in missBART2 due to its shared tree structure. This feature enables a more nuanced representation of missingness dependencies, offering a comprehensive approach to handling missing data. By allowing missingness indicators to influence the model, missSUBART provides deeper insights into how missingness propagates across variables, potentially revealing mechanisms that might otherwise remain undetected.

Further extensions could involve enhancing missSUBART's robustness in handling missing covariate data. Previously, missBART1 and missBART2 leveraged the `BARTm` framework from Kapelner and Bleich [2016] to accommodate missing covariates within BART trees, an approach that could similarly be integrated into missSUBART.

Additionally, while missSUBART does not explicitly account for correlations between missingness indicators, the parameter-expanded data augmentation strategy from Zhang [2020], as implemented in Esser et al. [2024], could be incorporated to estimate a non-diagonal correlation structure within a seemingly unrelated BART missingness model.

Finally, while our simulation studies considered various scenarios—including responses sharing tree structures (MAR 2 and MNAR 2), responses with distinct structures but correlated residuals (MNAR 3), and missingness mechanisms following either a linear (MAR 2) or tree-based model (MNAR 2 and MNAR 3)—further simulations could be conducted to assess missSUBART's performance under broader conditions and larger datasets.

6 Conclusion

This thesis introduced three novel joint models for handling multivariate response data with non-ignorable partial missingness which address key limitations of existing missing data methods that predominantly assume data are missing completely at random (MCAR) or missing at random (MAR). Unlike traditional approaches that either discard missing observations or rely on explicit assumptions about the missing data mechanism, our models learn the missingness process directly from the data, allowing for the recovery of non-ignorable MNAR mechanisms, as well as MCAR and MAR.

The motivation for developing these models stemmed from the *global Amax* dataset, which exhibits substantial missingness in the response variables while retaining fully observed covariates. The partial missingness in the response variables leads to complex, overlapping missingness patterns and a limited number of complete cases. The original analysis conducted by Maire et al. [2015] relied on complete-case analysis and excluded certain responses when missingness levels were too high, implicitly assuming an MCAR mechanism. However, even after excluding the most frequently missing response variable, the number of complete cases remained low, likely introducing biases and information loss. Additionally, several univariate models were implemented to separately model each response in the 5-dimensional set of responses.

Our work aimed to provide a robust alternative by jointly modelling both the responses and the missingness process while maintaining flexibility in handling different missingness mechanisms within the multivariate framework. We developed three novel Bayesian joint models that extend the selection model framework from Heckman [1976] while leveraging Bayesian additive regression trees (BART) to flexibly model complex relationships in multivariate response data. Instead of the traditional 'two-step' approach from Heckman [1976] which accounts for univariate response data, we use a joint modelling approach and extend it to the multivariate framework.

First, we introduced two new joint models, missBART1 and missBART2, which integrate multivariate BART with selection models to handle multivariate missing responses.

These models enable simultaneous recovery of MCAR, MAR, and MNAR data without imposing restrictive assumptions about the missingness mechanism. In both models, the responses follow a multivariate BART structure that captures dependencies between responses while capturing flexible, non-linear relationships with covariates. For the missingness model, missBART1 employs a multivariate Bayesian probit regression model, which enables incorporation of prior knowledge when available. In contrast, missBART2 extends this to a non-parametric setting using multivariate probit BART, leveraging BART's variable selection capabilities to identify the most influential predictors of missingness.

Although the covariates in our motivating *global Amax* dataset are fully observed, real-world data often have missingness in the covariates as well as the responses. Thus, missBART1 and missBART2 can also handle missing covariates, provided that the missingness is ignorable. missBART1 requires prior imputation on the covariates before model fitting, while missBART2 incorporates a technique from Kapelner and Bleich [2016] where missing covariates are integrated within the splitting rules of the BART trees.

The third model, missSUBART, further extends this framework by incorporating seemingly unrelated BART (suBART), where each response variable is modelled with its own tree structure while still capturing dependencies across responses through correlated error terms. In missSUBART, the multivariate responses are modelled using suBART, while missingness indicators for each response are modelled with individual sets of univariate BART models. This enables distinct relationships between the missingness of each response and its associated predictors, offering greater flexibility in modelling heterogeneous response-covariate relationships. Additionally, the missingness of a single response can also be informed by the missingness of other responses, rather than solely relying on observed or imputed values.

Through extensive simulation studies, we demonstrated that these models accurately predict multivariate responses under various missingness conditions while effectively recovering the true missingness mechanisms. Comparisons with alternative BART-based missing data methods—such as complete-case analysis with multivariate or univariate BART or imputation followed by model fitting with multivariate or univariate BART—highlighted the robustness of our models, particularly in MNAR settings where standard techniques tend to fail. Furthermore, we applied our models to the *global Amax* dataset, demonstrating their practical utility. While some of our findings aligned with the original analyses by Maire et al. [2015], our models also uncovered additional insights that were previously undetected, possibly due to the limiting assumptions of ignorable missingness implicit in the modelling approaches undertaken in the previous analyses.

The development and evaluation of missBART1, missBART2, and missSUBART led to several key findings. Our models improve the handling of missingness in multivariate data by explicitly modelling the missingness process rather than assuming an ignorable

mechanism. This is particularly important in cases of correlated multivariate responses and when missingness is non-ignorable. Unlike traditional methods that assume MCAR or MAR, our models accommodate MNAR data without requiring explicit assumptions about the missingness mechanism. Additionally, our models simultaneously impute missing responses within their modelling framework, rather than adopting a two-step approach, enhancing efficiency. Through evaluation of the posterior intervals within the parametric probit regression model and variable importance in the non-parametric probit BART model, we demonstrated that these models can effectively recover the true underlying missing mechanism without relying on strong priors.

6.1 Future Work

To improve the identification of missingness mechanisms, future research could explore more sophisticated variable selection techniques. As noted by Bleich et al. [2014], relying solely on raw variable inclusion proportions in BART to determine variable importance is insufficient, as these values do not directly reflect posterior probabilities. A key challenge is determining appropriate thresholds for classifying a predictor as important. A potential solution is the use of permutation-based methods, as proposed by Bleich et al. [2014], to establish optimal thresholds for variable inclusion, thereby enhancing the robustness of variable selection in the missingness model.

Another potential avenue for future work is exploring different approaches to handling missing covariates under the ignorable MCAR or MAR assumptions. Currently, missBART1 relies on prior covariate imputation using external methods such as `mice` or `missForest` before model fitting, while missBART2 incorporates the `BARTm` strategy, which accounts for missing covariates directly within the splitting rules of the trees, eliminating the need for explicit imputation. The same approach can also be applied to missSUBART. An alternative strategy for all three models is to explicitly model the ignorable missing covariates by assigning priors to the missing $\mathbf{X}$ variables and sampling from their full conditionals within the MCMC framework. This approach would eliminate the need for prior imputation, as required in missBART1, while still imputing missing values rather than leaving them unknown, as in missBART2. However, this method introduces additional computational complexity, and assigning appropriate priors can be challenging, particularly when prior information is limited or when covariates exhibit complex dependencies. Future research could explore the feasibility of this approach, assessing its computational trade-offs and evaluating its effectiveness in comparison to existing strategies for handling missing covariates.

In missBART1, the missingness model is based on a parametric multivariate probit regression framework, where priors are placed on the probit parameters to incorporate

prior information about the missingness mechanism. However, alternative priors, such as spike-and-slab or horseshoe priors, could be introduced to improve variable selection, thereby potentially enhancing the recovery of missingness mechanisms, particularly in high-dimensional settings with a large number of covariates and responses. As part of this approach, it will be of interest to explore differing priors for the covariates $\mathbf{X}$ and the responses $\mathbf{Y}$ which form the predictors of the missingness models.

In contrast, the non-parametric missingness models in missBART2 and missSUBART rely on uninformative and uniform splitting rules, as in standard BART. These approaches leverage BART's automatic variable selection capabilities to identify influential predictors, but can be inefficient when the number of predictors is large. An alternative approach is to adopt the strategy proposed by Linero [2018] which applies a sparsity-inducing Dirichlet prior to the splitting proportions of the regression trees. This encourages the model to prioritise a smaller subset of predictors, potentially improving variable selection and increasing efficiency in the presence of a large number of predictors. Within this framework, it will be of interest to investigate prior elicitation strategies, for the missingness trees in particular, that allow for splitting rules on $\mathbf{X}$ or $\mathbf{Y}$ to be prioritised, in order to express a degree of belief in the ignorability of the missingness mechanism.

Next, missBART1 employs the parameter-expanded data augmentation strategy from Talhouk et al. [2012] to estimate a non-diagonal correlation matrix in the missingness model, allowing dependencies between the missingness indicators. However, incorporating multiple latent variables adds considerable computational complexity. Due to the flexibility of the BART trees, missBART2 and missSUBART assume conditionally uncorrelated binary missingness indicators, with missBART2 permitting a non-diagonal covariance structure only within its terminal nodes. Future extensions could refine these models by explicitly estimating the correlation matrix in the missingness model using parameter-expanded data augmentation techniques, such as those implemented in Esser et al. [2024]. However, this approach presents additional challenges, including the need for the parameter-expanded Metropolis-Hastings algorithm from Zhang [2020], requiring careful model tuning to mitigate the excessive computational burden.

Our models were initially designed to handle continuous responses and covariates, as seen in the *global Amax* data, while also accommodating binary missingness indicators (via probit BART) and binary covariates (through BART's splitting rules). A key avenue for future research is extending these models to incorporate categorical covariates and mixed-type responses, broadening their applicability to more diverse datasets. Methods such as those of Deshpande [2022] for categorical predictors and Papageorgiou et al. [2015] for mixed-type responses offer promising directions for enhancing model flexibility. Notably, the DRYAD Digital Repository[a], from which the *global Amax* data were obtained, contains

[a]https://datadryad.org/stash/dataset/doi:10.5061/dryad.j42m7.

several categorical covariates that were not included in the original analysis by Maire et al. [2015] and were therefore omitted from our study. Incorporating these variables could potentially improve the predictive performance of our models and provide deeper insights into the relationships between leaf traits and environmental conditions.

Finally, we address the issue of computational efficiency. All models are implemented in `R` and are available on GitHub[b]. While our models efficiently integrate response modelling, missingness handling, and imputation within a unified framework—eliminating the need for separate imputation steps—computational demands remain a concern, particularly for multivariate models with a large number of responses. Several enhancements could be made to improve efficiency.

One potential extension is to adopt an adaptive tree selection strategy, as those proposed by Chakraborty [2016], to dynamically optimise the number of trees used. In missBART1 and missBART2, this would apply to the set of multivariate BART trees in the regression model and/or the missingness model. In missSUBART, this would adjust the number of univariate BART trees for each response, similarly in the response and/or missingness models. However, implementing this approach would introduce reversible jump MCMC steps [Green, 1995], requiring careful consideration of its impact on computational performance and potential trade-offs.

Additionally, the efficiency of our `R` implementation could be significantly improved by integrating `C++` via `Rcpp` [Eddelbuettel and François, 2011]. The `Rcpp` package provides a streamlined interface for incorporating `C++` into `R`, which has been successfully used in various BART-based packages, including the `BART` package from Sparapani et al. [2021a], `flexBART` from Deshpande [2022], and the suBART implementation of Esser et al. [2024][c].

6.2 Final Remarks

This thesis was motivated by the widespread challenge of missing data in multivariate response settings, as exemplified by the *global Amax* dataset, which exhibits substantial missingness in key response variables. Handling missing data appropriately is crucial for drawing valid inferences, yet many existing methods rely on restrictive assumptions about the missingness mechanism or suffer from inefficiencies due to separate imputation and analysis steps. To address these limitations, we developed three novel Bayesian joint models—missBART1, missBART2, and missSUBART—that integrate the selection model framework with Bayesian additive regression trees. These models provide a flexible and unified framework for simultaneously modelling responses, estimating missingness mechanisms, and imputing missing values. Through extensive simulation studies and

[b] Available at https://github.com/yongchengoh/missBART.
[c] Available at https://github.com/MateusMaiaDS/subart.

application to the *global Amax* dataset, we demonstrated the robustness and adaptability of our models, particularly in MNAR settings where standard techniques often fail.

In practice, the choice between the three proposed models depends on the practitioner's knowledge of the missingness process and the characteristics of the dataset. missBART1 is best-suited when strong prior knowledge about the missingness mechanism is available, as it allows the specification of informative priors within the parametric missingness model. However, it requires prior imputation for any missing covariates and may struggle to capture complex non-linearities or interactions within the missingness process. missBART2 is generally a more robust choice, as it does not require strong prior information for the missingness structure, can flexibly capture complex non-linearities and interactions in both the response and missingness models, and can also accommodate missing covariates without prior imputation.

However, missSUBART may be preferable when distinct sets of factors are believed to affect the values and/or missingness status of the different responses, rather than having shared tree structures, as per missBART2, whereby covariates affect all responses (and associated missingness indicators) simultaneously. In addition, missSUBART may also be preferable when the missingness of one response is expected to depend on the missingness or values of other responses, or when different responses are believed to follow distinct missingness structures rather than a shared one. Finally, missSUBART can, in principle, accommodate missing covariates without prior imputation (though further methodological work is needed in this regard, much like missBART2). Its added flexibility, however, comes at the cost of substantially greater computational burden, particularly as the number of responses increases.

While this work represents a significant advancement in missing data methodology, several avenues for future refinement remain, as discussed in this chapter. Further improvements, such as incorporating categorical covariates, enhancing variable selection techniques, and optimising computational efficiency, would expand the applicability and scalability of joint modelling approaches for missing data. Ultimately, this work contributes to the growing field of Bayesian statistical machine learning models for predicting multivariate data with non-ignorable partial missingness, offering a principled and flexible framework that ensures missingness is treated not as an obstacle, but as an integral part of statistical modelling.

Bibliography

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

Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011.

Amanda N Baraldi and Craig K Enders. An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37, 2010.

John Barnard, Robert McCulloch, and Xiao-Li Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311, 2000.

NH Batjes. ISRIC-WISE derived soil properties on a 5 by 5 arc-minutes global grid (ver. 1.2). Technical report, Isric, 2012.

Herman J Bierens. *Introduction to the Mathematical and Statistical Foundations of Econometrics*, volume 186 of *Themes in Modern Econometrics*. Cambridge University Press, 2004.

Justin Bleich, Adam Kapelner, Edward I George, and Shane T Jensen. Variable selection for BART: an application to gene regulation. *The Annals of Applied Statistics*, 8(3): 1750–1781, 2014.

Chester I Bliss. The method of probits. *Science*, 79(2037):38–39, 1934.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Leo Breiman, Jerome H Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees*. CRC Press, 1984.

Shawn Bushway, Brian D Johnson, and Lee Ann Slocum. Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. *Journal of Quantitative Criminology*, 23:151–178, 2007.

S Chakraborty. Bayesian additive regression tree for seemingly unrelated regression with automatic tree selection. In Venkat N Gudivada, Vijay V Raghavan, Venu Govindaraju, and CR Rao, editors, *Cognitive Computing: Theory and Applications*, volume 35 of *Handbook of Statistics*, chapter 7, pages 229–251. Elsevier, 2016.

Tianqi Chen and Carlos Guestrin. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

Siddhartha Chib and Edward Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2):347–361, 1998.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Paul Damien and Stephen G Walker. Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, 10(2):206–215, 2001.

A Philip Dawid and Steffen L Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, pages 1272–1317, 1993.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

David GT Denison, Bani K Mallick, and Adrian FM Smith. A Bayesian CART algorithm. *Biometrika*, 85(2):363–377, 1998.

Sameer K Deshpande. flexBART: flexible Bayesian regression trees with categorical predictors. *arXiv preprint arXiv:2211.04459*, 2022.

Vincent Dorie. *dbarts: discrete Bayesian additive regression trees sampler*, 2024. URL `https://CRAN.R-project.org/package=dbarts`. R package version 0.9-30.

Dirk Eddelbuettel and Romain François. Rcpp: seamless R and C++ integration. *Journal of Statistical Software*, 40:1–18, 2011.

Iris Eekhout, R Michiel de Boer, Jos WR Twisk, Henrica CW De Vet, and Martijn W Heymans. Missing data: a systematic review of how they are reported and handled. *Epidemiology*, 23(5):729–732, 2012.

Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008.

Jonas Esser, Mateus Maia, Andrew C Parnell, Judith Bosmans, Hanneke van Dongen, Thomas Klausch, and Keefe Murphy. Seemingly unrelated Bayesian additive regression trees for cost-effectiveness analyses in healthcare. *arXiv preprint arXiv:2404.02228*, 2024.

FAO, IIASA, ISRIC, ISSCAS, and JRC. Harmonized World Soil Database (version 1.2). FAO, Rome and IIASA, Laxenburg, Austria., 2012. URL `https://iiasa.ac.at/`.

Denzil G Fiebig. *Seemingly Unrelated Regression*, chapter 5, pages 101–121. John Wiley & Sons, Ltd, 2003.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Jerome H Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, pages 1189–1232, 2001.

Jacques-Emmanuel Galimard, Sylvie Chevret, Camelia Protopopescu, and Matthieu Resche-Rigon. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in Medicine*, 35(17):2907–2920, 2016.

Jacques-Emmanuel Galimard, Sylvie Chevret, Emmanuel Curis, and Matthieu Resche-Rigon. Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. *BMC Medical Research Methodology*, 18:1–13, 2018.

A Ronald Gallant. Seemingly unrelated nonlinear regressions. *Journal of Econometrics*, 3(1):35–50, 1975.

Andrew Gelman and Jennifer L Hill. Opening windows to the black box. *Journal of Statistical Software*, 40, 2011.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

Robert J Glynn, Nan M Laird, and Donald B Rubin. Selection modelling versus mixture modelling with nonignorable nonresponse. In Howard Wainer, editor, *Drawing Inferences from Self-Selected Samples*, pages 115–142. Springer, New York, NY, U.S.A., 1986.

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

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Trevor J Hastie and Robert Tibshirani. Bayesian backfitting (with comments and a rejoinder by the authors). *Statistical Science*, 15(3):196–223, 2000.

WK Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

James J Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement*, volume 5, number 4 of *NBER Chapters*, pages 475–492. National Bureau of Economic Research, Inc., 1976.

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

Alan Huang and Matthew P Wand. Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8(2):439–452, 2013.

Alan Inglis, Andrew C Parnell, and Catherine B Hurley. Visualizing variable importance and variable interaction effects in machine learning models. *Journal of Computational and Graphical Statistics*, 31(3):766–778, 2022.

ISRIC. Soilgrids: an automated system for global soil mapping, 2013. URL `https://www.isric.org/explore/soilgrids`.

Niko A Kaciroti and Trivellore Raghunathan. Bayesian sensitivity analysis of incomplete data: bridging pattern-mixture and selection models. *Statistics in Medicine*, 33(27): 4841–4857, 2014.

Adam Kapelner and Justin Bleich. Prediction with missing data via Bayesian additive regression trees. *Canadian Journal of Statistics*, 43(2):224–239, 2015.

Adam Kapelner and Justin Bleich. bartMachine: machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40, 2016.

Michael G Kenward and Geert Molenberghs. Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, 8(1):51–83, 1999.

Alexander Kowarik and Matthias Templ. Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16, 2016.

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

Antonio R Linero. In nonparametric and high-dimensional models, Bayesian ignorability is an informative prior. *Journal of the American Statistical Association*, 119(548):2785–2798, 2024.

Antonio R Linero and Yun Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1087–1110, 2018.

Roderick JA Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

Roderick JA Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121, 1995.

Roderick JA Little. Selection and pattern-mixture models. In Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs, editors, *Longitudinal Data Analysis*, chapter 18, pages 423–446. Chapman and Hall/CRC press, New York, NY, U.S.A., 2008.

Roderick JA Little and Donald B Rubin. *Statistical Analysis with Missing Data*, volume 793 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, Hoboken, NJ, U.S.A., 3rd edition, 2019.

Jun S Liu and Ying Nian Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274, 1999.

Vincent Maire, Ian J Wright, I Colin Prentice, Niels H Batjes, Radika Bhaskar, Peter M van Bodegom, Will K Cornwell, David Ellsworth, Ülo Niinemets, Alejandro Ordonez, et al. Global effects of soil and climate on leaf photosynthetic traits and rates. *Global Ecology and Biogeography*, 24(6):706–717, 2015.

Carolina Gil Marcelino, Gabriel MC Leite, P Celes, and Carlos Eduardo Pedreira. Missing data analysis in regression. *Applied Artificial Intelligence*, 36(1):2032925, 2022.

Giampiero Marra, Rosalba Radice, Till Bärnighausen, Simon N Wood, and Mark E McGovern. A simultaneous equation approach to estimating HIV prevalence with nonignorable missing responses. *Journal of the American Statistical Association*, 112(518): 484–496, 2017.

Nathan McJames, Ann O'Shea, Yong Chen Goh, and Andrew C Parnell. Bayesian causal forests for multivariate outcomes: application to Irish data from an international large scale education assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnae049, 2024.

Bart Michiels, Geert Molenberghs, and Stuart R Lipsitz. Selection models and pattern-mixture models for incomplete data with covariates. *Biometrics*, 55(3):978–983, 1999.

Jared S Murray. Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. *Journal of the American Statistical Association*, 116(534):756–769, 2021.

Mark New, David Lister, Mike Hulme, and Ian Makin. A high-resolution data set of surface climate over global land areas. *Climate Research*, 21(1):1–25, 2002.

Hanne Oberman, Thom Volker, Gerko Vink, Pepijn Vink, and Jamie Wallis. ggmice: Visualizations for "mice" with "ggplot2.", 2022. URL `https://github.com/amices/ggmice`.

Georgios Papageorgiou, Sylvia Richardson, and Nicky Best. 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, 2015.

Therese D Pigott. A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383, 2001.

Estevão B Prado, Rafael A Moral, and Andrew C Parnell. Bayesian additive regression trees with model trees. *Statistics and Computing*, 31(3):1–13, 2021.

Patrick Puhani. The heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14(1):53–68, 2000.

Veronika Ročková and Enakshi Saha. On theory for BART. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2839–2848. PMLR, 2019.

Veronika Ročková and Stephanie Van der Pas. Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48(4):2108–2131, 2020.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Rianne Margaretha Schouten, Peter Lugtig, and Gerko Vink. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930, 2018.

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

Rodney Sparapani, Charles Spanbauer, and Robert E McCulloch. Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software*, 97(1):1–66, 2021b.

Rodney A Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine*, 35(16):2741–2753, 2016.

Virendera K Srivastava and David EA Giles. *Seemingly unrelated regression equations models: estimation and inference*. CRC Press, 1987.

Daniel J Stekhoven and Peter Bühlmann. missForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Aline Talhouk, Arnaud Doucet, and Kevin Murphy. Efficient Bayesian inference for multivariate probit models with sparse inverse correlation matrices. *Journal of Computational and Graphical Statistics*, 21(3):739–757, 2012.

Yaoyuan Vincent Tan and Jason Roy. Bayesian additive regression trees and the general bart model. *Statistics in Medicine*, 38(25):5048–5069, 2019.

Niansheng Tang and Yuanyuan Ju. Statistical inference for nonignorable missing-data problems: a selective review. *Statistical Theory and Related Fields*, 2(2):105–133, 2018.

Herbert Thijs, Geert Molenberghs, Bart Michiels, Geert Verbeke, and Desmond Curran. Strategies to fit pattern-mixture models. *Biostatistics*, 3(2):245–265, 2002.

Nicholas Tierney and Dianne Cook. Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *Journal of Statistical Software*, 105(7):1–31, 2023.

Nicholas J Tierney, Fiona A Harden, Maurice J Harden, and Kerrie L Mengersen. Using decision trees to understand structure in missing data. *BMJ Open*, 5(6):e007450, 2015.

Seungha Um, Antonio R Linero, Debajyoti Sinha, and Dipankar Bandyopadhyay. Bayesian additive regression trees for multivariate skewed responses. *Statistics in Medicine*, 42(3):246–263, 2023.

Stef Van Buuren. *Flexible Imputation of Missing Data*. Interdisciplinary Statistics Series. Chapman and Hall/CRC press, New York, NY, U.S.A., 2nd edition, 2018.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

Mark van der Loo. simputation: simple imputation. R package version 0.2.9, 2022. URL `https://github.com/markvanderloo/simputation`.

Andrea C Westerband, Ian J Wright, Vincent Maire, Jennifer Paillassa, Iain Colin Prentice, Owen K Atkin, Keith J Bloomfield, Lucas A Cernusak, Ning Dong, Sean M Gleason, et al. Coordination of photosynthetic traits across soil and climate gradients. *Global Change Biology*, 29(3):856–873, 2023.

Ian J Wright, Peter B Reich, Mark Westoby, David D Ackerly, Zdravko Baruch, Frans Bongers, Jeannine Cavender-Bares, Terry Chapin, Johannes HC Cornelissen, Matthias Diemer, et al. The worldwide leaf economics spectrum. *Nature*, 428(6985):821–827, 2004.

Arnold Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298): 348–368, 1962.

Xiao Zhang. Parameter-expanded data augmentation for analyzing correlated binary data using multivariate probit models. *Statistics in Medicine*, 39(25):3637–3652, 2020.