Random Forests with Missing Data

Irving Gómez Méndez

20/10/2019

Contents

Chapter 1

Background

1.1 Statistical Learning

In statistical learning we have an outcome measurement $Y \in \mathcal{Y}$ that we wish to predict based on a set of input variables $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)}$ where $\mathbf{X}^{(j)} \in \mathcal{X}^{(j)}$ (for $j = 1, \ldots, p$). Together, the input variables $(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)})$ form a p dimensional input vector $\mathbf{X}$ taking its values in $\mathcal{X} = \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(p)}$.

The input space $\mathcal{X}$ and the output space $\mathcal{Y}$ are assumed, by definition, to respectively contain all possible input vectors and all possible output values. The input variables are sometimes known as features and the output variable as target. If Y is a categorical variable then the learning task is a classification problem. If Y is a numerical variable then the learning task is a regression problem.

Formally, the aim is to find a function $f : \mathcal{X} \to \mathcal{Y}$ that minimizes $\mathbb{E}_{\mathbf{X},Y}\left[\mathcal{L}(f(\mathbf{X}), Y)\right]$ where $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function that penalizes errors in the prediction.

By far, the most common loss function for regression problems is the squared error loss $\mathcal{L}(f(\mathbf{X}), Y) = (f(\mathbf{X}) - Y)^2$, and the function m which satisfies

$$m = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\min}\, \mathbb{E}_{\mathbf{X},Y}\left[(f(\mathbf{X}) - Y)^2\right]$$

is given by the regression function $m(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$.

For practical problems, the distribution of $(\mathbf{X}, Y)$ (and hence also the regression function) is unknown. But it is possible to collect data according to the distribution of $(\mathbf{X}, Y)$, summarize in a data set called learning set or training set $\mathcal{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$, where the observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ are considered independent.

The goal is to use the data $\mathcal{D}_n$ to construct a learning model, also called learner or predictor, $m_n : \mathcal{X} \to \mathcal{Y}$ which estimates the regression function m, and enables us to predict the outcome for new unseen objects.

In classification problems, it is usual to consider the 0-1 loss function $\mathcal{L}(f(\mathbf{X}, Y)) = \mathbb{1}_{f(\mathbf{X})\neq Y}$, the function g which satisfies

$$g = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\min}\, \mathbb{E}_{\mathbf{X},Y}\left[\mathbb{1}_{f(\mathbf{X})\neq Y}\right]$$

is called the Bayes classifier. To define it, let us suppose that $\mathcal{Y} = \{c_1, \ldots, c_J\}$, then $g(\mathbf{X}) = \arg\max_{y\in\{c_1,\ldots,c_J\}} \mathbb{P}\left[Y = y|\mathbf{X}\right]$.

As in the case of the regression function, the Bayes classifier is also unachievable for practical problems, so we must rely on data to estimate it. Depending on the data recorded in the training data set, the classification problems are divided into supervised and unsupervised. In the supervised classification the input variables and the corresponding target variable are available in the training data set. In the unsupervised classification, only the input variables are recorded.

1.2 Decision Trees

Decision trees part the input space into disjoint regions which usually are rectangles, and then fit a simple model (like a constant) in each one. They are conceptually simple yet powerful and attractive in practice by several factors:

- They can model arbitrarily complex relations between the input and the output space
- They handle categorical or numerical variables, or a mix of both.
- They can be used in regression or supervised classification problems.
- They are easy interpretable, even for non-statisticians.

More importantly, decision trees are the basis of many modern and state-of-the-art algorithms, including random forests (Breiman, 2001) (on which this works is about) or boosting (Schapire and Freund, 1995; Friedman, 2001), where they are used as building blocks for composing larger models.

To construct the partition of the input space, decision trees works in a recursive way. The root of the tree is the whole input space, $\mathcal{X}$, which is splitted into disjoint regions. Then each region is splitted into more regions, and this process is continued until some stopping rule is applied. Figure 1.1 exemplifies this procedure. At each step of the tree construction, the partition performed over a cell (or equivalently its corresponding node) is determined maximizing some split criterion.

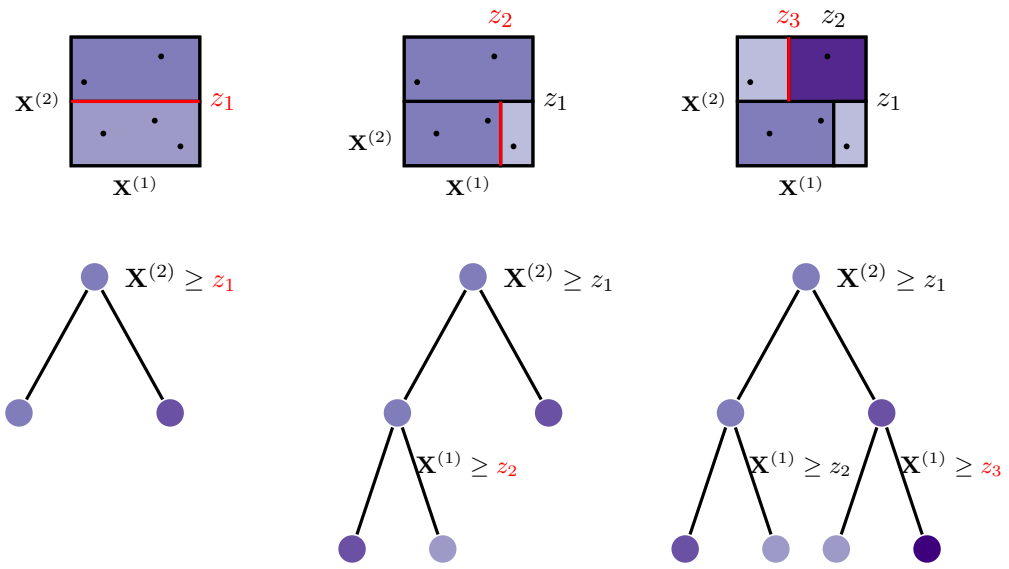

Figure 1.1: At each step of the tree construction a partition is performed over a cell (or equivalently its corresponding node) maximizing some split criterion.

Different split criteria have been proposed depending on the statistical problem and the nature of the input space. For regression problems, the most common split criterion is the CART criterion (Breiman et al., 1984). For supervised classification, the split criteria are proposed to minimize some impurity function, like the misclassification error, the Shannon entropy (Shannon, 1948) or the Gini index (Gini, 1912).

While binary splits are the most common to construct decision trees. Some split criteria, like ID3 and C4.5 (Quinlan, 1986; Quinlan, 1993), replace the binary splits on categorical variables with multiway splits.

1.3 Random Forest

The random forest, introduced by Breiman (2001), is a non-parametric and general-purpose algorithm that might be used in either regression problems or supervised classification.

A random forest is a predictor consisting of a collection of $M(>1)$ randomized decision trees. Randomization is introduced in two different parts of the tree construction. Prior to the construction of each tree, a_n observations are drawn at random with (or without) replacement from the learning data set $\mathcal{D}_n$. Only these a_n observations are taken into account in the tree construction. Then, at each cell of each tree, a split is performed by maximizing the split criterion over `mtry` input variables chosen uniformly at random among the original ones. In regression tasks, the split criterion presented in Breiman et al. (1984) is the CART criterion; while in supervised classification problems it is usual to consider the Gini impurity criterion (Gini, 1912; Friedman, Hastie, and Tibshirani, 2009). And the tree construction is stopped when each final node contains less or equal than `nodesize` points or when the tree has t_n final nodes.

Thus, the parameters in this algorithm are

- $M > 1$, the number of trees in the forest.
- $a_n \in \{1, \ldots, n\}$, the number of observations in each tree.
- `mtry` $\in \{1, \ldots, p\}$, the number of directions (features) chosen, candidate to be splitted. The features selected in each step are $\mathcal{M}_{try}$.
- `nodesize` $\in \{1, \ldots, a_n\}$, the maximum number of observations for a node to be a final cell.
- Instead of `nodesize` we can use the parameter $t_n \in \{1, \ldots, a_n\}$ which is the number of leaves (final nodes) in each tree.

For the kth tree in the collection, the predicted value at a query point $\mathbf{x}$ is denoted by $m_n(\mathbf{x}; \Theta_k, \mathcal{D}_n)$, where Θ_k characterizes the tree and $\Theta_1, \ldots, \Theta_M$ are independent random variables, distributed the same as a generic random variable Θ and independent of $\mathcal{D}_n$. The nature and dimensionality of Θ depends on its use in the tree construction. For us, consists of the observations selected for the tree and the variables candidates to be splitted.

In regression, the kth tree estimate is defined as

$$m_n(\mathbf{x}; \Theta_k, \mathcal{D}_n) = \sum_{i \in \mathcal{I}_{n,\Theta_k}} \frac{Y_i \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_k, \mathcal{D}_n)}}{N_n(\mathbf{x}; \Theta_k, \mathcal{D}_n)}$$

where $\mathcal{I}_{n,\Theta_k}$ is the set of the observations selected prior to the construction of the kth tree, $A_n(\mathbf{x}; \Theta_k, \mathcal{D}_n)$ is the cell that contains $\mathbf{x}$, and $N_n(\mathbf{x}; \Theta_k, \mathcal{D}_n)$ is the number of observations (between $\mathcal{I}_{n,\Theta_k}$) which belong to $A_n(\mathbf{x}; \Theta_k, \mathcal{D}_n)$.

The combination of the trees form the finite random forest estimate

$$m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{k=1}^{M} m_n(\mathbf{x}; \Theta_k, \mathcal{D}_n)$$

Breiman (2001) shows that the random forest does not overfit when M tends to infinity. Then M is a parameter just restrictive by the computer capability and it should be as larger as we could. Consequently, from a theoretical point of view, it is useful to consider the case when M tends to infinity, which generates the infinite random forest estimate

$$m_n(\mathbf{x}; \mathcal{D}_n) = \mathbb{E}_{\Theta|\mathcal{D}_n}[m_n(\mathbf{x}; \Theta, \mathcal{D}_n)]$$

1.4 Mechanisms for Missing Data

The data-missing mechanisms established the relationship between missingness and the data. This concept was introduced by Rubin (1976), which recognizes three different data-missing mechanisms: if the missingness and the data are independent, it is said to be Missing Completely at Random (MCAR); if the missingness only depends on observed data. it is Missing at Random (MAR); otherwise it is Not Missing at Random (NMAR).

However, as pointed out by other authors (Josse et al., 2019), there is little literature on missing data for supervised learning, and the nomenclature of the data-missing mechanisms has not been appropriate discussed.

It is common to define the data-missing mechanisms through the data matrix, even so, to have a useful definition from a theoretical point of view we formally define them through the random variables, $\mathbf{X}$ and Y.

Let us define a new variable, called the indicator of missing value as

$$\mathbf{M}^{(h)} = \begin{cases} 1 & \text{if } \mathbf{X}^{(h)} \text{ is missing} \\ 0 & \text{otherwise} \end{cases}, \quad 1 \leq h \leq p$$

We will assume throughout this work that the response Y has no missing values, so there is not necessary to define an indicator of missing variable for Y. Let us also define $\mathbf{Z} = (\mathbf{X}, Y)$. Then, the mechanisms are characterized by the conditional distribution of $\mathbf{M}^{(h)}$ given $\mathbf{Z}$.

Missing Completely at Random (MCAR) We say that the variable $\mathbf{X}^{(h)}$ is MCAR if $\mathbf{M}^{(h)} \perp \mathbf{Z}$. That is, every observation $\mathbf{X}_i$ has the same probability of have $\mathbf{X}^{(h)}$ missing.

Missing at Random (MAR) Let us define the set $h_o = \{h : \mathbb{P}\left[\mathbf{M}^{(h)} = 0\right]\}$, thus $\mathbf{X}^{(h_o)}$ is the vector conformed by those variables that are always observed.

The variable $\mathbf{X}^{(h)}$ is MAR if

$$\mathbb{P}\left[\mathbf{M}^{(h)}(\mathbf{X}) = 1|\mathbf{Z}\right] = \mathbb{P}\left[\mathbf{M}^{(h)}(\mathbf{X}) = 1|\mathbf{Z}^{(h_o)}\right]$$

where $\mathbf{Z}^{(h_o)} = (\mathbf{X}^{(h_o)}, Y)$

1.5 CART Criterion

In this section we present the CART criterion introduced by Breiman et al. (1984) and its theoretical version.

For ease of understanding, we consider a tree with no subsampling step, which uses the entire and original data $\mathcal{D}_n$.

From this point forward, we use the follow notation

- A denotes a general node.
- $N_n(A)$ is the number of points in A.
- (h, z) denotes a cut in A, where

  h is a direction, $h \in \{1, \ldots, p\}$, and

  z is the position of the cut in the hth direction, between the limits of A.
- $\mathcal{C}_A$ is the set of all possible cuts in the node A.
- $A_L = \{\mathbf{X}_i \in A : \mathbf{X}_i^{(h)} < z\}$, $A_R = \{\mathbf{X}_i \in A : \mathbf{X}_i^{(h)} \geq z\}$
- $\bar{Y}_A$ (resp. $\bar{Y}_{A_L}$, $\bar{Y}_{A_R}$) is the average of the Y_i such that $\mathbf{X}_i$ belongs to the cell A (resp. A_L, A_R).

We define the split CART criterion (in its empirical version) for a generic cell A and for the case of complete data, as

$$\begin{aligned} L_{n,A}(h,z) =& \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_A\right)^2 \mathbb{1}_{\mathbf{X}_i \in A} \\ & - \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(h)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(h)} \geq z}\right)^2 \mathbb{1}_{\mathbf{X}_i \in A} \end{aligned} \tag{1.1}$$

With the convention $0/0 = 0$.

Note that it might be written as

$$\begin{aligned} L_{n,A}(h,z) = & \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_A\right)^2 \mathbb{1}_{\mathbf{X}_i \in A} - \frac{1}{N_n(A)} \sum_{\mathbf{X}_i^{(h)} < z} \left(Y_i - \bar{Y}_{A_L}\right)^2 \mathbb{1}_{\mathbf{X}_i \in A} \\ & - \frac{1}{N_n(A)} \sum_{\mathbf{X}_i^{(h)} \geq z} \left(Y_i - \bar{Y}_{A_R}\right)^2 \mathbb{1}_{\mathbf{X}_i \in A} \end{aligned}$$

$$\begin{aligned} = & \left[\frac{n}{N_n(A)}\right] \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_A\right)^2 \mathbb{1}_{\mathbf{X}_i \in A} - \left[\frac{n}{N_n(A)}\right] \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L}\right)^2 \mathbb{1}_{\mathbf{X}_i \in A, \mathbf{X}_i^{(h)} < z} \\ & - \left[\frac{n}{N_n(A)}\right] \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_R}\right)^2 \mathbb{1}_{\mathbf{X}_i \in A, \mathbf{X}_i^{(h)} \geq z} \end{aligned}$$

We can use eq. (A.1) to the latter expression, and get its asymptotic version

$$\begin{aligned} L_{n,A}(h,z) \xrightarrow[n\to\infty]{a.s.} & \frac{1}{\cancel{\mathbb{P}[\mathbf{X} \in A]}} \mathbb{E}\left\{\left(Y - \mathbb{E}\left[Y|\mathbf{X} \in A\right]\right)^2 | \mathbf{X} \in A\right\} \cancel{\mathbb{P}[\mathbf{X} \in A]} \\ & - \frac{1}{\mathbb{P}[\mathbf{X} \in A]} \mathbb{E}\left\{\left(Y - \mathbb{E}\left[Y|\mathbf{X}^{(h)} < z, \mathbf{X} \in A\right]\right)^2 | \mathbf{X}^{(h)} < z, \mathbf{X} \in A\right\} \mathbb{P}[\mathbf{X}^{(h)} < z, \mathbf{X} \in A] \\ & - \frac{1}{\mathbb{P}[\mathbf{X} \in A]} \mathbb{E}\left\{\left(Y - \mathbb{E}\left[Y|\mathbf{X}^{(h)} \geq z, \mathbf{X} \in A\right]\right)^2 | \mathbf{X}^{(h)} \geq z, \mathbf{X} \in A\right\} \mathbb{P}[\mathbf{X}^{(h)} \geq z, \mathbf{X} \in A] \end{aligned}$$

$$\begin{aligned} = & \mathbb{V}[Y|\mathbf{X} \in A] - \mathbb{V}[Y|\mathbf{X}^{(h)} < z, \mathbf{X} \in A] \mathbb{P}[\mathbf{X}^{(h)} < z|\mathbf{X} \in A] \\ & - \mathbb{V}[Y|\mathbf{X}^{(h)} \geq z, \mathbf{X} \in A] \mathbb{P}[\mathbf{X}^{(h)} \geq z|\mathbf{X} \in A] \end{aligned}$$

Thus, the theoretical CART criterion is defined as

$$\begin{aligned} L_A^\star(h,z) =& \mathbb{V}[Y|\mathbf{X} \in A] - \mathbb{V}[Y|\mathbf{X}^{(h)} < z, \mathbf{X} \in A] \mathbb{P}[\mathbf{X}^{(h)} < z|\mathbf{X} \in A] \\ & - \mathbb{V}[Y|\mathbf{X}^{(h)} \geq z, \mathbf{X} \in A] \mathbb{P}[\mathbf{X}^{(h)} \geq z|\mathbf{X} \in A] \end{aligned} \tag{1.2}$$

For each cell A, the best (empirical) cut $(h_n^\star, z_n^\star)$ is selected maximizing $L_{n,A}(h,z)$ over $\mathcal{M}_{try}$ and $\mathcal{C}_A$, that is

$$(h_n^\star, z_n^\star) \in \underset{\substack{h \in \mathcal{M}_{try} \\ (h,z) \in \mathcal{C}_A}}{\arg\max}\, L_{n,A}(h,z)$$

Analogously, we define the best theoretical cut $(h^\star, z^\star)$ in A as

$$(h^\star, z^\star) \in \underset{\substack{h \in \mathcal{M}_{try} \\ (h,z) \in \mathcal{C}_A}}{\arg\max}\, L_A^\star(h, z)$$

Applying basic algebra, it is easy to get an equivalent expression for the empirical CART criterion, given by

$$L_{n,A}(h, z) = \frac{N_n(A_L)N_n(A_R)}{N_n(A)N_n(A)} \left(\bar{Y}_{A_L} - \bar{Y}_{A_R}\right)^2 \tag{1.3}$$

We can apply ones again eq. (A.1) to the latter, and get an alternative expression to the theoretical CART criterion, given by

$$L_A^\star(h, z) = \mathbb{P}[\mathbf{X}^{(h)} < z | \mathbf{X} \in A]\mathbb{P}[\mathbf{X}^{(h)} \geq z | \mathbf{X} \in A]\Big(\mathbb{E}[Y | \mathbf{X} < z, \mathbf{X} \in A] - \mathbb{E}[Y | \mathbf{X} \geq z, \mathbf{X} \in A]\Big)^2 \tag{1.4}$$

Chapter 2

Previous Approaches

2.1 CART criterion

The methods proposed in the literature to deal with missing data through imputation operate in a recursive way **poner citas!!**. First, they use the original training data set, $\mathcal{D}_n$, to fill the blank spaces in a roughly way. For example, with the median of the observed values in the variable. We denote this new data set as $\mathcal{D}_{n,t_1}$.

The imputed data set $\mathcal{D}_{n,t_1}$ is used to build a random forest. Then, some structures of the forest are exploited, like the so-called proximity matrix, improving the imputation and resulting in a new data set $\mathcal{D}_{n,t_2}$. The procedure is continued iteratively until some stopping rule is applied, for example when there is little change between the imputed values or when a fix number of iterations is achieved.

More formally, let us define

$$\mathbf{X}_{i,t_\ell}^{(h)} = \begin{cases} \mathbf{X}_i^{(h)} & \text{if } \mathbf{M}_i^{(h)} = 0 \\ \widehat{\mathbf{X}}_{i,t_\ell}^{(h)} & \text{if } \mathbf{M}_i^{(h)} = 1 \end{cases}$$

where $\widehat{\mathbf{X}}_{i,t_\ell}^{(h)}$ is the imputation of $\mathbf{X}_i^{(h)}$ at time t_ℓ, $\ell \geq 1$

Also, we define

$$\mathbf{X}_{i,t_\ell} = \left(\mathbf{X}_{i,t_\ell}^{(1)}, \ldots, \mathbf{X}_{i,t_\ell}^{(p)}\right)$$

The CART criterion use for these methods to build the trees of the forest at time t_ℓ, $\ell \geq 1$, is

$$\begin{aligned} L_{n,A,t_\ell}(h,z) =& \frac{1}{N_{n,t_\ell}(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A,t_\ell}\right)^2 \mathbb{1}_{\mathbf{X}_{i,t_\ell} \in A} \\ & - \frac{1}{N_{n,t_\ell}(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L,t_\ell} \mathbb{1}_{\mathbf{X}_{i,t_\ell}^{(h)} < z} - \bar{Y}_{A_R,t_\ell} \mathbb{1}_{\mathbf{X}_{i,t_\ell}^{(h)} \geq z}\right)^2 \mathbb{1}_{\mathbf{X}_{i,t_\ell} \in A} \end{aligned}$$

$$(h_{n,t_\ell}^\star, z_{n,t_\ell}^\star) \in \underset{\substack{h \in \mathcal{M}_{try} \\ (h,z) \in \mathcal{C}_A}}{\arg\max}\, L_{n,A,t_\ell}(h,z)$$

where

$$N_{n,t_\ell}(A) = \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_{i,t_\ell} \in A}$$

$$N_{n,t_\ell}(A_L) = \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_{i,t_\ell} \in A,\, \mathbf{X}_{i,t_\ell}^{(h)} < z}$$
(resp. for $N_{n,t_\ell}(A_R)$)

$$\bar{Y}_{A,t_\ell} = \frac{1}{N_{n,t_\ell}(A)} \sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_{i,t_\ell} \in A}$$

$$\bar{Y}_{A_L,t_\ell} = \frac{1}{N_{n,t_\ell}(A_L)} \sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_{i,t_\ell} \in A,\, \mathbf{X}_{i,t_\ell}^{(h)} < z}$$
(resp. for $\bar{Y}_{A_R,t_\ell}$)

Analogously to the case of complete data, we define its theoretical version as

$$\begin{aligned} L^{\star}_{A,t_\ell}(h,z) =& \mathbb{V}[Y|\mathbf{X}_{t_\ell} \in A] - \mathbb{V}[Y|\mathbf{X}_{t_\ell}^{(h)} < z, \mathbf{X}_{t_\ell} \in A]\mathbb{P}[\mathbf{X}_{t_\ell}^{(h)} < z|\mathbf{X}_{t_\ell} \in A] \\ & - \mathbb{V}[Y|\mathbf{X}_{t_\ell}^{(h)} \geq z, \mathbf{X}_{t_\ell} \in A]\mathbb{P}[\mathbf{X}_{t_\ell}^{(h)} \geq z|\mathbf{X}_{t_\ell} \in A] \end{aligned}$$

where

$$\mathbf{X}_{t_\ell} = \left(\mathbf{X}_{t_\ell}^{(1)}, \ldots, \mathbf{X}_{t_\ell}^{(p)}\right)$$

$$\mathbf{X}_{t_\ell}^{(h)} = \begin{cases} \mathbf{X}^{(h)} & \text{if } \mathbf{M}^{(h)} = 0 \\ \widehat{\mathbf{X}}_{t_\ell}^{(h)} & \text{if } \mathbf{M}^{(h)} = 1 \end{cases}$$

and $\widehat{\mathbf{X}}_{t_\ell}^{(h)}$ is the imputation at time t_ℓ of the variable $\mathbf{X}^{(h)}$.

2.2 Connectivity and Proximity

In order to properly introduced previous methods that handle missing data through imputation, we need to define the connectivity between two points in a tree and the proximity matrix of the forest.

Let $K_{\Theta,n}(\mathbf{X}, \mathbf{X}'; \mathcal{D}_n) = \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}'}$ be the indicator that $\mathbf{X}$ is in the same leaf that $\mathbf{X}'$ in the tree designed with $\mathcal{D}_n$ and the parameter Θ. If $K_{\Theta,n}(\mathbf{X}, \mathbf{X}'; \mathcal{D}_n) = 1$ we say that $\mathbf{X}$ and $\mathbf{X}'$ are connected in the tree $m_n(\cdot; \Theta, \mathcal{D}_n)$.

The proximity of two points is the average of times in which they were connected, it measures the similarity of the observations in the eyes of the random forest. Formally, let us define the proximity between $\mathbf{X}$ and $\mathbf{X}'$ in the finite forest, $m_{M,n}(\cdot; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n)$, as

$$K_{M,n}(\mathbf{X}, \mathbf{X}'; \mathcal{D}_n) = \frac{1}{M} \sum_{k=1}^{M} K_{\Theta_k,n}(\mathbf{X}, \mathbf{X}'; \mathcal{D}_n)$$

2.3 Previous Approaches with Imputation

To simplify the notation, let $K_{M,t_\ell}(i,j)$ be the proximity between $\mathbf{X}_i$ and $\mathbf{X}_j$ at time t_ℓ. That is, $K_{M,t_\ell}(i,j) = K_{M,n}(\mathbf{X}_i, \mathbf{X}_j; \mathcal{D}_{n,t_\ell})$ is the proximity between $\mathbf{X}_i$ and $\mathbf{X}_j$ in the random forest constructed with the data set $\mathcal{D}_{n,t_\ell}$.

We also define $\mathbf{i}_{miss}^{(h)} \subseteq \{1, \ldots, n\}$ as the indexes where $\mathbf{X}^{(h)}$ is missing. And $\mathbf{i}_{obs}^{(h)} = \{1, \ldots, n\} \setminus \mathbf{i}_{miss}^{(h)}$ as the indexes were $\mathbf{X}^{(h)}$ is observed.

2.3.1 Breiman's Approach

This method appears in Breiman (2003). It is implemented in `R` by Liaw and Wiener (2002) (`rfImpute`).

If $\mathbf{X}^{(h)}$ is a continuous variable

$$\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}} = \frac{\sum_{i\in\mathbf{i}^{(h)}_{obs}} K_{M,t_\ell}(i,j)\mathbf{X}^{(h)}_i}{\sum_{i\in\mathbf{i}^{(h)}_{obs}} K_{M,t_\ell}(i,j)}, \qquad \begin{array}{l}\ell \geq 1\\ j \in \mathbf{i}^{(h)}_{miss}\end{array}$$

If $\mathbf{X}^{(h)}$ is a categorical variable

$$\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}} = \underset{\mathbf{x}\in\mathcal{X}^{(h)}}{\arg\max} \sum_{i\in\mathbf{i}^{(h)}_{obs}} K_{M,t_\ell}(i,j)\mathbb{1}_{\mathbf{X}^{(h)}_i=\mathbf{x}}, \qquad \begin{array}{l}\ell \geq 1\\ j \in \mathbf{i}^{(h)}_{miss}\end{array}$$

2.3.2 Ishioka's Approach

This method appears in Ishioka (2013).

If $\mathbf{X}^{(h)}$ is a continue variable

$$\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}} = \frac{\sum_{\substack{i\in\text{neigh}_k\\ i\neq j}} K_{M,t_\ell}(i,j)\mathbf{X}^{(h)}_{i,t_\ell}}{\sum_{\substack{i\in\text{neigh}_k\\ i\neq j}} K_{M,t_\ell}(i,j)}, \qquad \begin{array}{l}\ell \geq 1\\ j \in \mathbf{i}^{(h)}_{miss}\end{array}$$

If $\mathbf{X}^{(h)}$ is a categorical variable

$$\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}} = \underset{\mathbf{x}\in\mathcal{X}^{(h)}}{\arg\max} \sum_{i\neq j} K_{M,t_\ell}(i,j)\mathbb{1}_{\mathbf{X}^{(h)}_{i,t_\ell}=\mathbf{x}}, \qquad \begin{array}{l}\ell \geq 1\\ j \in \mathbf{i}^{(h)}_{miss}\end{array}$$

For continuous variables the missing value is not estimated with the observed values, but with the k closest.

The k closest values are chosen to make more robust the method and avoid values which are outliers.

For categorical variables, it is not necessary to see only the k closest values because the outliers of $\mathbf{X}$ will have few attention. Meanwhile the proximity with missing values should have more attention, specially when the missing rate is high.

2.3.3 MissForest Approach

This method appears in Stekhoven and Bühlmann (2011).

This algorithm the missing data as a supervised learning problem itself, where the target variable is the input variable with missing values. The MissForest consists in iteratively building a random forest from the observed data and the previous imputations to predict the missing values of the input variable.

2.4 Previous Approaches without Imputation

- "Missing" category

  For categorical variables a simple way to deal with the missing data is to create the new category "missing" (Quinlan, 1986; Friedman, Hastie, and Tibshirani, 2009).

  However this approach can lead to anomalous situations as exemplified by Quinlan (1986).

  It is important to remark that this option do not deal with missing data in continuous variables.

- Ternary decision trees

  We can change the structure of the trees and consider ternary splits instead of binary, where the third child contains all observations where the feature is missing (Louppe, 2014).

- Propagate in both child nodes

  We can weight the samples and give less weight to the missing data and propagate them in both child nodes dividing the observation into fractional objects. Of course, the issue with this alternative is that there is not an obvious methodology to calculate those weights. But as long as the method used to calculate the weights do not involve the use of the target variable, Y, this method can be used even for prediction with missing predictors (Quinlan, 1986; Louppe, 2014).

- Split the observations

  For methods that weight the data, they could assign the observations to the more probable node. We should have the same warnings with the weights as in the previous approach (Quinlan, 1986).

  However, we could split the observations between the child nodes assigning part of the missing data to one of the child nodes even without weighting the data (Ripley, 2007; Venables and Ripley, 2002).

- As far as it will go

  This method is for prediction with missing predictors and consists in dropping the case down the tree as far as it will go. We predict with the node reached by the observation (Ripley, 2007; Venables and Ripley, 2002).

- Surrogate splits

  It is a popular alternative. We order the directions that best split the node, if the first direction is missing we take the second surrogate split, if the second is missing take the third and so on. The surrogate split is such that maximizes the probability of making the same decision as the primary split (Breiman et al., 1984; Ripley, 2007; Venables and Ripley, 2002; Friedman, Hastie, and Tibshirani, 2009).

Chapter 3

Our CART Criterion

3.1 Empirical CART Criterion

Let us point out the problem on the original CART criterion when we are in a framework of missing values, to ease understanding, consider that $\mathbf{X} \sim \mathcal{U}[0,1]^2$ (we are going to consider always this case when exemplify) and suppose that we have two observations, $\mathbf{X}_1$ and $\mathbf{X}_2$. Also, suppose that we are interested in the cell A and that we have performed a cut, such that we have the cells A_L and A_R. The values of $\mathbf{X}_1$ and $\mathbf{X}_2$, as well as the region of the cells are given in table 3.1, fig. 3.1 shows a graphical representation, where the observations with the missing value in $\mathbf{X}^{(1)}$ are represented as a dashed line over the interval $[0,1]$.

	$\mathbf{X}^{(1)}$	$\mathbf{X}^{(2)}$
$\mathbf{X}_1$		0.5
$\mathbf{X}_2$	0.75	0.25
A	[0.3,0.9]	[0.2,0.7]
A_L	[0.3,0.6]	[0.2,0.7]
A_R	[0.6,0.9]	[0.2,0.7]

Table 3.1: Example data

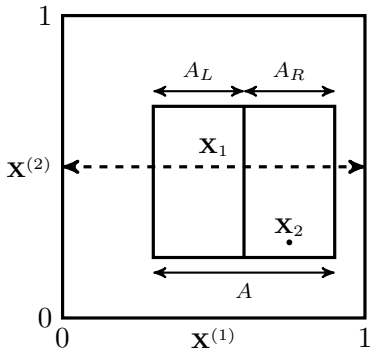

Figure 3.1: Graphical representation of the example data.

Note that, while we know to which child node $\mathbf{X}_2$ belongs, that is, we know that $\mathbf{X}_2 \in A_R$, we do not know to which child node $\mathbf{X}_1$ belongs, that is, we do not know if $\mathbf{X}_1 \in A_R$ or $\mathbf{X}_1 \in A_L$, actually we do not even know if $\mathbf{X}_1$ belongs to the cell A. While we continue in this uncertainty, the highlighted parts in eq. (3.1) of the CART criterion can not be calculated.

$$
\begin{aligned}
L_{n,A}(h,z) = & \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_A\right)^2 \mathbb{1}_{\mathbf{X}_i \in A} \\
& - \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L}\right)^2 \mathbb{1}_{\mathbf{X}_i \in A, \mathbf{X}_i^{(h)} < z} \\
& - \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_R}\right)^2 \mathbb{1}_{\mathbf{X}_i \in A, \mathbf{X}_i^{(h)} \geq z}
\end{aligned}
\tag{3.1}
$$

Our proposed method modify the CART criterion, substituting the parts that can not be calculated by mathematical objects that allow us to compute the split criterion and to maximize it. One important advantage of our proposal is that we only construct one random forest and it does not require imputation of the missing values or the computation of extra structures (like the proximity matrix).

The key idea of our approach is to notice that we are not really interested on calculate the CART criterion for every cell A, but just for those cells that we construct. As usual, the input space, $\mathcal{X}$, is the root of the tree, where all observations belong. We perform a cut and assign the missing values to the best child node, maximizing our CART criterion. Then we move on to the next cell and we act in the proper way, performing the best cut and making the best assignation for the missing values, and we continue in this way until a stopping rule is achieved.

More properly, let us define

$$\widehat{\mathbf{X}}_{i,1}^{(h)} = \begin{cases} \mathbf{X}_i^{(h)} & \text{if } \mathbf{M}_i^{(h)} = 0 \\ \mathcal{X}^{(h)} & \text{if } \mathbf{M}_i^{(h)} = 1 \end{cases}$$

and

$$\widehat{\mathbf{X}}_{i,1} = \left(\widehat{\mathbf{X}}_{i,1}^{(1)}, \ldots, \widehat{\mathbf{X}}_{i,1}^{(p)}\right)$$

That is, if we know the value of the variable $\mathbf{X}^{(h)}$ for the ith observation, we assign it to $\widehat{\mathbf{X}}_{i,1}^{(h)}$, otherwise we assign all the possible values for that variable, $\mathcal{X}^{(h)}$.

The root of our trees is denoted as

$$A_1 = \mathcal{X}$$

We use $\widehat{\mathbf{X}}_{i,1}$ and A_1 to define the CART criterion, $L_{n,A_1}\left(h, z, \mathbb{X}_{miss}^{(h)}\right)$, where $\mathbb{X}_{miss}^{(h)}$ are the missing values in the learning data set, $\mathcal{D}_n$, of the variable $\mathbf{X}^{(h)}$ is missing, i.e. $\mathbb{X}_{miss}^{(h)}$ are the blank spaces of the variable $\mathbf{X}^{(h)}$ in $\mathcal{D}_n$.

After a proper maximization of $L_{n,A_1}\left(h, z, \mathbb{X}_{miss}^{(h)}\right)$ we define $\widehat{\mathbf{X}}_{i,2}$ and A_2. Then, we continue in a recursive way to construct the decision trees.

Figure 3.2 exemplifies this method, in the example we have 4 points where $\mathbf{X}^{(1)}$ is missing $(\mathbf{X}_2, \mathbf{X}_5, \mathbf{X}_6)$. We perform a cut and assign two of these points $(\mathbf{X}_2, \mathbf{X}_6)$ to the left node and one to the right node $(\mathbf{X}_5)$. We continue until some stopping rule is achieved.

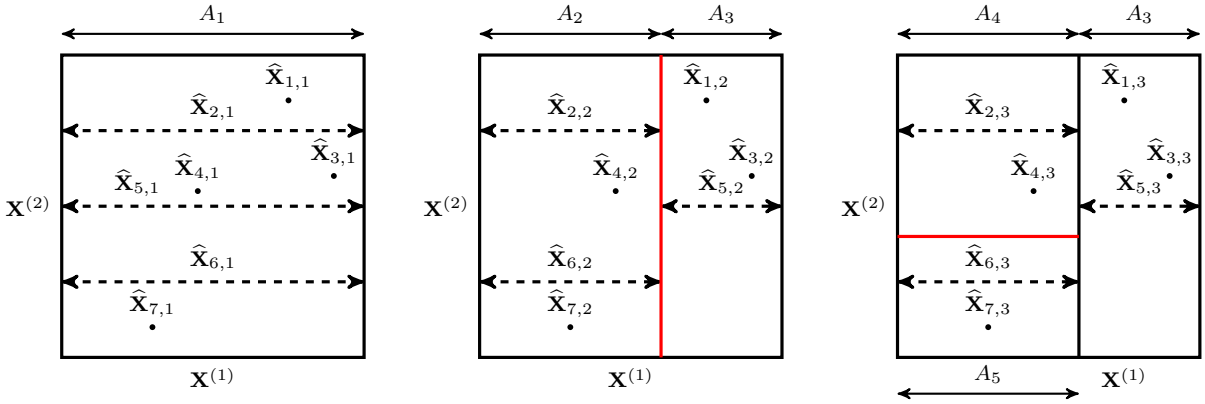

Figure 3.2: We perform a cut and assignation of points were the variable is missing, maximizing our CART criterion $L_{n,A_\ell}\left(h, z, \mathbb{X}_{miss}^{(h)}\right)$

Consider a generic cell $A = \prod_{j=1}^{p}[a^{(j)}, b^{(j)}]$, we say that $\widehat{\mathbf{X}}_{i,\ell} \in A$ if

$$\begin{cases} \widehat{\mathbf{X}}_{i,\ell}^{(h)} \in [a^{(h)}, b^{(h)}] & \forall\, h \text{ s.t. } \mathbf{M}_i^{(h)} = 0 \\ \widehat{\mathbf{X}}_{i,\ell}^{(h)} \subseteq [a^{(h)}, b^{(h)}] & \forall\, h \text{ s.t. } \mathbf{M}_i^{(h)} = 1 \end{cases}$$

Our empirical CART criterion for the cell A_ℓ is defined as

$$\begin{aligned} L_{n,A_\ell}\left(h, z, \mathbb{X}_{miss}^{(h)}\right) =& \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_\ell}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell} \\ &- \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,L}}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, a_\ell^{(h)} \leq \mathbf{X}_i^{(h)} < z} \\ &- \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,R}}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, z \leq \mathbf{X}_i^{(h)} \leq b_\ell^{(h)}} \end{aligned} \tag{3.2}$$

where

$$N_n(A_\ell) = \sum_{i=1}^{n} \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell}$$

$$N_n(A_{\ell,L}) = \sum_{i=1}^{n} \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, a_\ell^{(h)} \leq \mathbf{X}_i^{(h)} < z}$$

(resp. for $N_n(A_{\ell,R})$)

and

$$\bar{Y}_{A_\ell} = \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} Y_i \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell}$$

$$\bar{Y}_{A_{\ell,L}} = \frac{1}{N_n(A_{\ell,L})} \sum_{i=1}^{n} Y_i \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, a_\ell^{(h)} \leq \mathbf{X}_i^{(h)} < z}$$

(resp. for $\bar{Y}_{A_{\ell,R}}$)

Note that in the last two indicators of eq. (3.2) we have the conditions $a_\ell^{(h)} \leq \mathbf{X}_i^{(h)} < z$ and $z \leq \mathbf{X}_i^{(h)} \leq b_\ell^{(h)}$, without making distinction between observations with the variable $\mathbf{X}^{(h)}$ observed and those where it is missing, this also applies to $N_n(A_{\ell,L})$, $N_n(A_{\ell,R})$, $\bar{Y}_{A_{\ell,L}}$ and $\bar{Y}_{A_{\ell,R}}$. For those observations where the variable $\mathbf{X}^{(h)}$ is observed, we know which one of these conditions is satisfied, but for those where it is missing, we do not know which condition is satisfied, unless we assign a value to the variable.Therefore, $\mathbb{X}_{miss}^{(h)}$ is a variable in our CART criterion, along with the cut (h, z).

For every possible cut, we calculate our CART criterion over all feasible values for the observations where the variable $\mathbf{X}^{(h)}$ is missing.

That is, the best cut and assignation $\left(h_{n,\ell}^\star, z_{n,\ell}^\star, \mathbb{X}_{miss,\ell}^{\star\left(h_{n,\ell}^\star\right)}\right)$ is selected maximizing $L_{n,A_\ell}\left(h, z, \mathbb{X}_{miss}^{(h)}\right)$ over $\mathcal{M}_{try}$, $\mathcal{C}_{A_\ell}$ and the values of $\mathbb{X}_{miss}^{(h)}$ between the limits of A_ℓ, that is

$$\left(h^\star_{n,\ell}, z^\star_{n,\ell}, \mathbb{X}^{\star\left(h^\star_{n,\ell}\right)}_{miss,\ell}\right) \in \underset{\substack{h\in\mathcal{M}_{try}\\(h,z)\in\mathcal{C}_{A_\ell}\\ \mathbb{X}^{(h)}_{miss}\in\left[a^{(h)}_\ell, b^{(h)}_\ell\right]}}{\arg\max} L_{n,A_\ell}\left(h, z, \mathbb{X}^{(h)}_{miss}\right)$$

where

$\mathbb{X}^{(h)}_{miss} \in \left[a^{(h)}_\ell, b^{(h)}_\ell\right]$ means that $\mathbf{X}^{(h)}_i \in \left[a^{(h)}_\ell, b^{(h)}_\ell\right]$ $\forall\, \mathbf{M}^{(h)}_i = 1$ such that $\widehat{\mathbf{X}}_{i,\ell} \in A_\ell$

When we perform the cut (h, z), the specific value of the variable $\mathbf{X}^{(h)}$, for those observations in which it is missing, does not really matters for the CART criterion, but only takes into account if it is bigger or equal to z or if it is less than z. So, maximize our CART criterion over every possible cut and every possible value in $\mathbf{X}^{(h)}$ for those observations where it is missing, it is equivalent to maximize it over every possible cut and all the possible assign combinations.

Once we have selected the best cut and assignation, we move on to the next cell, and make the proper, more formally, let us define

$$\widehat{\mathbf{X}}^{\left(h^\star_{n,\ell}\right)}_{i,2\ell} = \begin{cases} \left[a^{\left(h^\star_{n,\ell}\right)}_\ell, z^\star_{n,\ell}\right] & \text{if } \mathbf{M}^{\left(h^\star_{n,\ell}\right)}_i = 1 \text{ and } a^{\left(h^\star_{n,\ell}\right)}_\ell \leq \widehat{\mathbf{X}}^{\star\left(h^\star_{n,\ell}\right)}_{i,\ell} < z^\star_{n,\ell} \text{ and } \widehat{\mathbf{X}}^\star_{i,\ell} \in A_\ell \\ \left[z^\star_{n,\ell}, b^{\left(h^\star_{n,\ell}\right)}_\ell\right] & \text{if } \mathbf{M}^{\left(h^\star_{n,\ell}\right)}_i = 1 \text{ and } z^\star_{n,\ell} \leq \widehat{\mathbf{X}}^{\star\left(h^\star_{n,\ell}\right)}_{i,\ell} \leq b^{\left(h^\star_{n,\ell}\right)}_\ell \text{ and } \widehat{\mathbf{X}}^\star_{i,\ell} \in A_\ell \\ \widehat{\mathbf{X}}^{\star\left(h^\star_{n,\ell}\right)}_{i,\ell} & \text{if } \mathbf{M}^{\left(h^\star_{n,\ell}\right)}_i = 1 \text{ and and } \widehat{\mathbf{X}}^\star_{i,\ell} \notin A_\ell \\ \mathbf{X}^{\left(h^\star_{n,\ell}\right)}_i & \text{if } \mathbf{M}^{\left(h^\star_{n,\ell}\right)}_i = 0 \end{cases}$$

$$\widehat{\mathbf{X}}^{(j)}_{i,2\ell} = \widehat{\mathbf{X}}^{(j)}_{i,\ell} \quad \forall\, j \neq h^\star_{n,\ell}$$

$$\widehat{\mathbf{X}}_{i,2\ell} = \left(\widehat{\mathbf{X}}^{(1)}_{i,2\ell}, \ldots, \widehat{\mathbf{X}}^{(p)}_{i,2\ell}\right)$$

$$\widehat{\mathbf{X}}_{i,2\ell+1} = \widehat{\mathbf{X}}_{i,2\ell}$$

That is, $\widehat{\mathbf{X}}^{\left(h^\star_{n,\ell}\right)}_{i,2\ell}$ will be the best assignation in the previous cut for the observations with the $h^\star_{n,\ell}$th variable missing such that $\widehat{\mathbf{X}}_{i,\ell} \in A_\ell$; it will be $\widehat{\mathbf{X}}^{\star\left(h^\star_{n,\ell}\right)}_{i,\ell}$ if the $h^\star_{n,\ell}$th variable is missing, but $\widehat{\mathbf{X}}_{i,\ell} \notin A_\ell$, i.e., we do not modify the assignation for those observations that were not assigned to the cell; and will be $\mathbf{X}^{h^\star_{n,\ell}}_i$ if the $h^\star_{n,\ell}$th variable is observed.

And, the child nodes are defined as

$$A_{2\ell} = A_\ell \cap \left(\prod_{j=1}^{h^\star_{n,\ell}-1}\left[a^{(j)}_\ell, b^{(j)}_\ell\right] \times \left[a^{\left(h^\star_{n,\ell}\right)}_\ell, z^\star_{n,\ell}\right] \times \prod_{j=h^\star_{n,\ell}+1}^{p}\left[a^{(j)}_\ell, b^{(j)}_\ell\right]\right)$$

$$A_{2\ell+1} = A_\ell \cap \left(\prod_{j=1}^{h^\star_{n,\ell}-1}\left[a^{(j)}_\ell, b^{(j)}_\ell\right] \times \left[z^\star_{n,\ell}, b^{\left(h^\star_{n,\ell}\right)}_\ell\right] \times \prod_{j=h^\star_{n,\ell}+1}^{p}\left[a^{(j)}_\ell, b^{(j)}_\ell\right]\right)$$

Figures 3.3 and 3.4 show a graphical representation of these definitions.

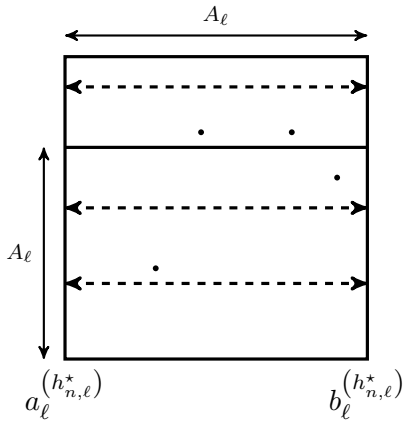

Figure 3.3: Definition of $\widehat{\mathbf{X}}_{i,\ell}$ and A_ℓ, right before the of best cut and assignation.

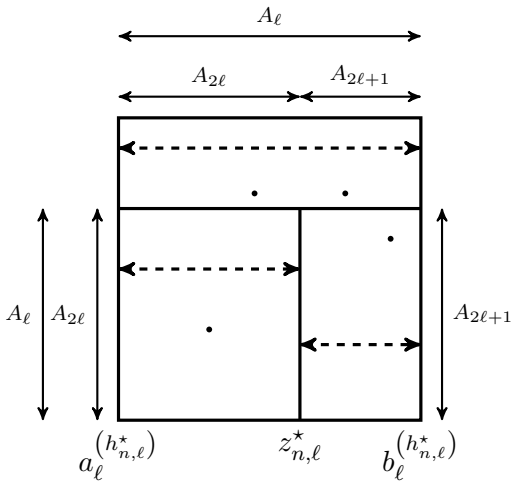

Figure 3.4: Definition of $\widehat{\mathbf{X}}_{i,2\ell}$, $\widehat{\mathbf{X}}_{i,2\ell+1}$, $A_{2\ell}$ and $A_{2\ell+1}$, when we have selected the best cut and assignation.

One more thing should be notice, we have said that we need to calculate our CART criterion for every possible cut and for all assign combinations for the observations where the variable considered in the cut is missing, but, actually we have to consider a less number of combinations.

Consider once again, that we have only three observations, say $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{X}_3$, whose variable considered in the cut is missing (as it is exemplified in fig. 3.2), suppose that their variable response, Y, is such that $Y_1 < Y_2 < Y_3$. Additionally, suppose that we assign two of the points together, assign $\mathbf{X}_1$ and $\mathbf{X}_2$ together or $\mathbf{X}_2$ and $\mathbf{X}_3$ together will always be better that assign $\mathbf{X}_1$ and $\mathbf{X}_3$ together.

So, if we have n_{miss} observations with the value missing on the variable considered, instead of searching over $2^{n_{miss}}$ possible combinations, we only have to check $n_{miss}+1$ combinations.

Long over, we have kept implicit $\mathbb{X}^{(h)}_{miss}$ in eq. (3.2), but we can make it explicit. To do this, consider the first term on the right side in eq. (3.2), and let us expand it as follows

$$\begin{aligned}\frac{1}{N_n(A_\ell)}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_{A_\ell}\right)^2\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell} =&\frac{1}{N_n(A_\ell)}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_{A_\ell}\right)^2\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\mathbf{M}_i^{(h)}=0}\\ &+\frac{1}{N_n(A_\ell)}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_{A_\ell}\right)^2\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\mathbf{M}_i^{(h)}=1}\\ =&\frac{1}{N_n(A_\ell)}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_{A_\ell,obs}+\bar{Y}_{A_\ell,obs}-\bar{Y}_{A_\ell}\right)^2\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\mathbf{M}_i^{(h)}=0}\\ &+\frac{1}{N_n(A_\ell)}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_{A_\ell,miss}+\bar{Y}_{A_\ell,miss}-\bar{Y}_{A_\ell}\right)^2\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\mathbf{M}_i^{(h)}=1}\end{aligned}$$

That is

$$\begin{aligned}\frac{1}{N_n(A_\ell)}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_{A_\ell}\right)^2\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell} =& \frac{1}{N_n(A_\ell)}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_{A_\ell,obs}\right)^2\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\mathbf{M}_i^{(h)}=0}\\ &+\frac{1}{N_n(A_\ell)}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_{A_\ell,miss}\right)^2\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\mathbf{M}_i^{(h)}=1}\\ &+\frac{N_{n,obs}(A_\ell)}{N_n(A_\ell)}\left(\bar{Y}_{A_\ell,obs}-\bar{Y}_{A_\ell}\right)^2\\ &+\frac{N_{n,miss}(A_\ell)}{N_n(A_\ell)}\left(\bar{Y}_{A_\ell,miss}-\bar{Y}_{A_\ell}\right)^2\end{aligned}$$

where

$$\begin{aligned}N_{n,obs}(A_\ell)&=\sum_{i=1}^{n}\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\,\mathbf{M}_i^{(h)}=0}\\ N_{n,miss}(A_\ell)&=\sum_{i=1}^{n}\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\,\mathbf{M}_i^{(h)}=1}\end{aligned}$$

and

$$\begin{aligned}\bar{Y}_{A_\ell,obs}&=\frac{1}{N_{n,obs}(A_\ell)}\sum_{i=1}^{n}Y_i\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\,\mathbf{M}_i^{(h)}=0}\\ \bar{Y}_{A_\ell,miss}&=\frac{1}{N_{n,miss}(A_\ell)}\sum_{i=1}^{n}Y_i\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\,\mathbf{M}_i^{(h)}=1}\end{aligned}$$

Additionally, let us define

$$\begin{aligned}N_{n,obs}(A_{\ell,L})&=\sum_{i=1}^{n}\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\,a_\ell^{(h)}\le\mathbf{X}_i^{(h)}<z,\,\mathbf{M}_i^{(h)}=0}\\ N_{n,miss}(A_{\ell,L})&=\sum_{i=1}^{n}\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\,a_\ell^{(h)}\le\mathbf{X}_i^{(h)}<z,\,\mathbf{M}_i^{(h)}=1}\\ &\text{(resp. for } N_{n,obs}(A_{\ell,R}) \text{ and } N_{n,miss}(A_{\ell,R}))\end{aligned}$$

and

$$\begin{aligned}\bar{Y}_{A_{\ell,L},obs}&=\frac{1}{N_{n,obs}(A_{\ell,L})}\sum_{i=1}^{n}Y_i\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\,a_\ell^{(h)}\le\mathbf{X}_i^{(h)}<z,\,\mathbf{M}_i^{(h)}=0}\\ \bar{Y}_{A_{\ell,L},miss}&=\frac{1}{N_{n,miss}(A_{\ell,L})}\sum_{i=1}^{n}Y_i\mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell}\in A_\ell,\,a_\ell^{(h)}\le\mathbf{X}_i^{(h)}<z,\,\mathbf{M}_i^{(h)}=1}\\ &\text{(resp. for } \bar{Y}_{A_{\ell,R},obs} \text{ and } \bar{Y}_{A_{\ell,R},miss})\end{aligned}$$

That is, $N_{n,obs}(A_{\ell,L})$ ($N_{n,obs}(A_{\ell,R})$) is the number of observations on the left (right) child node, between those satisfying $\widehat{\mathbf{X}}_{i,\ell}\in A_\ell$, with the variable $\mathbf{X}^{(h)}$ observed, and $\bar{Y}_{A_{\ell,L},obs}$ ($\bar{Y}_{A_{\ell,R},obs}$) is the average of the

response of those observations. While, $N_{n,miss}(A_{\ell,L})$ ($N_{n,miss}(A_{\ell,R})$) is the number of observations on the left (right) child node, between those satisfying $\widehat{\mathbf{X}}_{i,\ell} \in A_\ell$, with the variable $\mathbf{X}^{(h)}$ missing, and $\bar{Y}_{A_{\ell,L},miss}$ ($\bar{Y}_{A_{\ell,R},miss}$) is the average of the response of those observations. Once again, none of these quantities can be calculated until we assign values where the variable $\mathbf{X}^{(h)}$ is missing.

With this extra notation, the last two terms on the right side of eq. (3.2) might be written as

$$\frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,L}}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, a_\ell^{(h)} \leq \mathbf{X}_i^{(h)} < z}$$

$$\begin{aligned}
=&\frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,L},obs}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, a_\ell^{(h)} \leq \mathbf{X}_i^{(h)} < z,\, \mathbf{M}_i^{(h)}=0} \\
&+ \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,L},miss}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, a_\ell^{(h)} \leq \mathbf{X}_i^{(h)} < z,\, \mathbf{M}_i^{(h)}=1} \\
&+ \frac{N_{n,obs}(A_{\ell,L})}{N_n(A_\ell)} \left(\bar{Y}_{A_{\ell,L},obs} - \bar{Y}_{A_{\ell,L}}\right)^2 \\
&+ \frac{N_{n,miss}(A_{\ell,L})}{N_n(A_\ell)} \left(\bar{Y}_{A_{\ell,L},miss} - \bar{Y}_{A_{\ell,L}}\right)^2
\end{aligned}$$

and

$$\frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,R}}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, z \leq \mathbf{X}_i^{(h)} \leq b_\ell^{(h)}}$$

$$\begin{aligned}
=&\frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,R},obs}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, z \leq \mathbf{X}_i^{(h)} \leq b_\ell^{(h)},\, \mathbf{M}_i^{(h)}=0} \\
&+ \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,R},miss}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, z \leq \mathbf{X}_i^{(h)} \leq b_\ell^{(h)},\, \mathbf{M}_i^{(h)}=1} \\
&+ \frac{N_{n,obs}(A_{\ell,R})}{N_n(A_\ell)} \left(\bar{Y}_{A_{\ell,R},obs} - \bar{Y}_{A_{\ell,R}}\right)^2 \\
&+ \frac{N_{n,miss}(A_{\ell,R})}{N_n(A_\ell)} \left(\bar{Y}_{A_{\ell,R},miss} - \bar{Y}_{A_{\ell,R}}\right)^2
\end{aligned}$$

So, finally, our CART criterion looks like

$$\begin{aligned}
L_{n,A_\ell}\left(h, z, \mathbb{X}_{miss}^{(h)}\right) =& L_{1,n,A_\ell}(h,z) + L_{2,n,A_\ell}\left(h, z, \mathbb{X}_{miss}^{(h)}\right) \\
&+ L_{3,n,A_\ell}\left(h, z, \mathbb{X}_{miss}^{(h)}\right) + L_{4,n,A_\ell}\left(h, z, \mathbb{X}_{miss}^{(h)}\right)
\end{aligned}$$

where

$$
\begin{aligned}
L_{1,n,A_\ell}(h,z) =& \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_\ell,obs}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, \mathbf{M}_i^{(h)}=0} \\
& - \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,L},obs}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, a_\ell^{(h)} \leq \mathbf{X}_i^{(h)} \leq z,\, \mathbf{M}_i^{(h)}=0} \\
& - \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,R},obs}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, z \leq \mathbf{X}_i^{(h)} \leq b_\ell^{(h)},\, \mathbf{M}_i^{(h)}=0}
\end{aligned}
$$

$$
\begin{aligned}
L_{2,n,A_\ell}\left(h,z,\mathbb{X}_{miss}^{(h)}\right) =& \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_\ell,miss}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, \mathbf{M}_i^{(h)}=1} \\
& - \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,L},miss}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, a_\ell^{(h)} \leq \mathbf{X}_i^{(h)} \leq z,\, \mathbf{M}_i^{(h)}=1} \\
& - \frac{1}{N_n(A_\ell)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\ell,R},miss}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,\ell} \in A_\ell,\, z \leq \mathbf{X}_i^{(h)} \leq b_\ell^{(h)},\, \mathbf{M}_i^{(h)}=1}
\end{aligned}
$$

$$
\begin{aligned}
L_{3,n,A_\ell}\left(h,z,\mathbb{X}_{miss}^{(h)}\right) =& \frac{N_{n,obs}(A_\ell)}{N_n(A_\ell)} \left(\bar{Y}_{A_\ell,obs} - \bar{Y}_{A_\ell}\right)^2 \\
& - \frac{N_{n,obs}(A_{\ell,L})}{N_n(A_\ell)} \left(\bar{Y}_{A_{\ell,L},obs} - \bar{Y}_{A_{\ell,L}}\right)^2 \\
& - \frac{N_{n,obs}(A_{\ell,R})}{N_n(A_\ell)} \left(\bar{Y}_{A_{\ell,R},obs} - \bar{Y}_{A_{\ell,R}}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
L_{4,n,A_\ell}\left(h,z,\mathbb{X}_{miss}^{(h)}\right) =& \frac{N_{n,miss}(A_\ell)}{N_n(A_\ell)} \left(\bar{Y}_{A_\ell,miss} - \bar{Y}_{A_\ell}\right)^2 \\
& - \frac{N_{n,miss}(A_{\ell,L})}{N_n(A_\ell)} \left(\bar{Y}_{A_{\ell,L},miss} - \bar{Y}_{A_{\ell,L}}\right)^2 \\
& - \frac{N_{n,miss}(A_{\ell,R})}{N_n(A_\ell)} \left(\bar{Y}_{A_{\ell,R},miss} - \bar{Y}_{A_{\ell,R}}\right)^2
\end{aligned}
$$

With the convention 0/0=0.

3.2 Interesting Formulas for the CART Criterion

Analogously to eq. (1.3) it can be shown that

$$
L_{1,n,A_\ell} = \left(\frac{N_{n,obs}(A_\ell)}{N_n(A_\ell)}\right) \left(\frac{N_{n,obs}(A_{\ell,L}) N_{n,obs}(A_{\ell,R})}{N_{n,obs}(A_\ell) N_{n,obs}(A_\ell)}\right) \left(\bar{Y}_{A_{\ell,L},obs} - \bar{Y}_{A_{\ell,R},obs}\right)^2 \tag{3.3}
$$

And

$$
L_{2,n,A_\ell} = \left(\frac{N_{n,miss}(A_\ell)}{N_n(A_\ell)}\right) \left(\frac{N_{n,miss}(A_{\ell,L}) N_{n,miss}(A_{\ell,R})}{N_{n,miss}(A_\ell) N_{n,miss}(A_\ell)}\right) \left(\bar{Y}_{A_{\ell,L},miss} - \bar{Y}_{A_{\ell,R},miss}\right)^2 \tag{3.4}
$$

3.3 The Effect of the Mechanism of Missingness

We have said that the best cut and assignation, $\left(h^\star_{n,\ell}, z^\star_{n,\ell}, \mathbb{X}^{\star\left(h^\star_{n,\ell}\right)}_{miss,\ell}\right)$, is given by

$$\left(h^\star_{n,\ell}, z^\star_{n,\ell}, \mathbb{X}^{\star\left(h^\star_{n,\ell}\right)}_{miss,\ell}\right) \in \underset{\substack{h\in\mathcal{M}_{try}\\(h,z)\in\mathcal{C}_{A_\ell}\\ \mathbb{X}^{(h)}_{miss}\in\left[a^{(h)}_\ell, b^{(h)}_\ell\right]}}{\arg\max} L_{n,A_\ell}\left(h, z, \mathbb{X}^{(h)}_{miss}\right)$$

We select to maximize over $\mathbb{X}^{(h)}_{miss} \in \left[a^{(h)}_\ell, b^{(h)}_\ell\right]$ because, usually we do not have access to the missing data mechanism. If we knew the mechanism, the values of $\mathbb{X}^{(h)}_{miss}$ where the CART criterion is maximized might change.

For example, consider that we have only missing values in $\mathbf{X}^{(1)}$, and it is missing when it is bigger than $\tau \in (0,1)$, that is, we have a right censoring mechanism. Also consider that we know the value of τ, so the missing mechanism is known, fig. 3.5 shows this case. Consider a cut in $\mathbf{X}^{(1)}$ at τ, then,the only admissible assignation for those observations whose $\mathbf{X}^{(1)}$ variable is missing is to the right child, as it is shown in fig. 3.6.

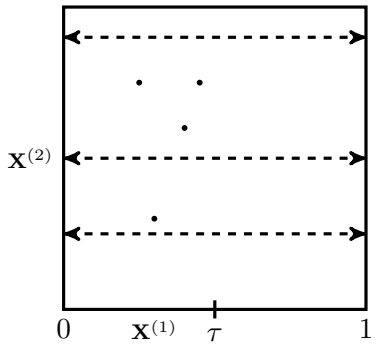

Figure 3.5: We have three observations whose value in $\mathbf{X}^{(1)}$ is missing, but bigger than τ.

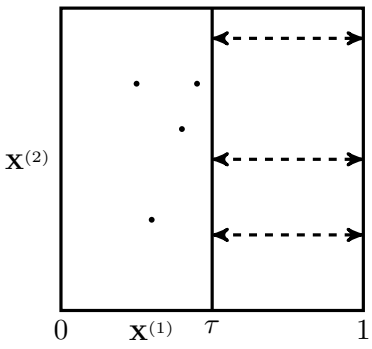

Figure 3.6: The only possible assignation for the missing values in $\mathbf{X}^{(1)}$.

As another example, consider that we have a MCAR mechanism in the variable $\mathbf{X}^{(h)}$, that is, every observation have the same probability that $\mathbf{X}^{(h)}$ is missing. Then we could assign at random, with the probability depending on the place of cut, z, the observations where $\mathbf{X}^{(1)}$ is missing. So, the maximization of the CART criterion, $L_{n,A_\ell}\left(h, z, \mathbb{X}^{(h)}_{miss}\right)$, over $\mathbb{X}^{(h)}_{miss}$ would be spurious.

3.4 Relation with MIA

Missing incorporated in attributes (MIA) is an approach developed by Twala, Jones, and Hand (2008) which seems very close to ours.

Consider the cases represented in figs. 3.7 and 3.8, in both of them a cut is performed in the $\mathbf{X}^{(1)}$ variable but the place were it is cut leave all the observations were $\mathbf{X}^{(1)}$ is observed on the same side, and we decide to assign all the observations where it is missing into the other child node.

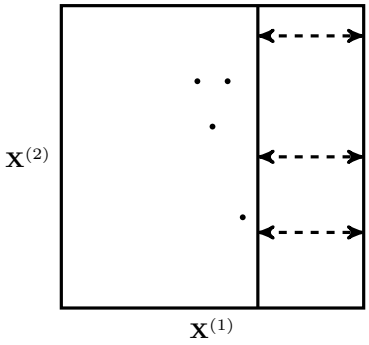

Figure 3.7: All the observations with the missing value are assigned to the right, while all observations with the variable observed keep on the left.

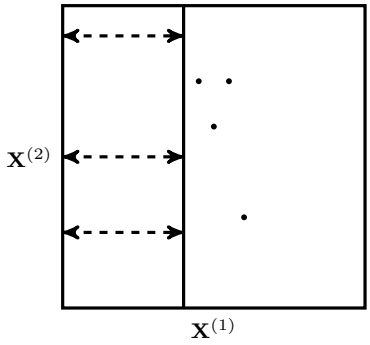

Figure 3.8: All the observations with the missing value are assigned to the left, while all observations with the variable observed keep on the right.

Since our CART criterion takes the same vale for the cases represented in figs. 3.7 and 3.8, we do not have any preference of one over another. These will happen no matter were the cut (h, z) is performed, along it satisfies $z \notin \left[\min\left\{\mathbf{X}_{obs}^{(h)}\right\}, \max\left\{\mathbf{X}_{obs}^{(h)}\right\}\right]$ and the assignation of the $\mathbf{X}_{miss}^{(h)}$ leaves the same observations together.

In fact, the case represented in fig. 3.7 (or fig. 3.8, since both are equivalent for our criterion) corresponds to make cells of the form $A_R = \{\mathbf{X}^{(h)}$ is missing$\}$ and $A_L = \{\mathbf{X}^{(h)}$ is observed$\}$ or vice versa (it does not really matter). This is exactly what the MIA does. MIA creates a cell with all the observations where $\mathbf{X}^{(h)}$ is missing and another with all the observations where it is observed.

Our method goes further than the MIA approach, since not only the assignations shown in fig. 3.7 and **??** are consider, but we have as candidates all possible assignations for the observations were $\mathbf{X}^{(h)}$ is missing. For example, the assignation shown in fig. 3.9 is also a candidate in our approach. That it is not allowed in the MIA approach.

With this observations in mind is clear that the MIA approach is a particular case of our method, in which the only assign combination consider is in which the observations with the variable $\mathbf{X}^{(h)}$ is missing are putting together.

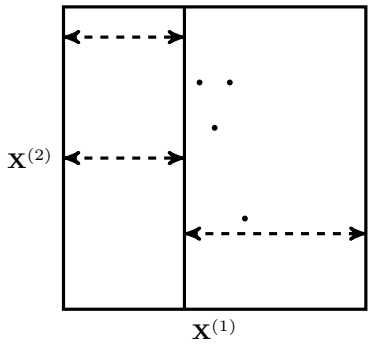

Figure 3.9: Another possible assignation, that is not allowed in the MIA approach.

3.5 Theoretical CART Criterion

From a theoretical point of view we do not longer have observations, but random variables, the assign combinations will be translated into the probabilities of belonging to the left or child node for those random variables with the missing value, which will be optimized every time we perform a cut. This observation allows us to define the theoretical CART criterion changing the role of $\mathbb{X}^h_{miss}$ by these probabilities.

Let us denote

$$\mathbb{P}_{A^\star_{\ell,L},obs} = \mathbb{P}\left[a^{\star(h)}_\ell \leq \mathbf{X}^{(h)} < z | \mathbf{X} \in A^\star_\ell, \mathbf{M}^{(h)} = 0\right]$$

$$\mathbb{P}_{A^\star_{\ell,L},miss} = \mathbb{P}\left[a^{\star(h)}_\ell \leq \mathbf{X}^{(h)} < z | \mathbf{X} \in A^\star_\ell, \mathbf{M}^{(h)} = 1\right]$$

$$\mu_{A^\star_\ell} = \mathbb{E}\left[Y | \mathbf{X} \in A^\star_\ell\right]$$

$$\mu_{A^\star_\ell,obs} = \mathbb{E}\left[Y | \mathbf{X} \in A^\star_\ell, \mathbf{M}^{(h)} = 0\right]$$
$$\mu_{A^\star_\ell,miss} = \mathbb{E}\left[Y | \mathbf{X} \in A^\star_\ell, \mathbf{M}^{(h)} = 1\right]$$

$$\mu_{A^\star_{\ell,L}} = \mathbb{E}\left[Y | \mathbf{X} \in A^\star_\ell, a^{\star(h)}_\ell \leq \mathbf{X}^{(h)} < z\right]$$
(resp. for $\mu_{A^\star_{\ell,R}}$)

$$\mu_{A^\star_{\ell,L},obs} = \mathbb{E}\left[Y | \mathbf{X} \in A^\star_\ell, a^{\star(h)}_\ell \leq \mathbf{X}^{(h)} < z, \mathbf{M}^{(h)} = 0\right]$$
$$\mu_{A^\star_{\ell,L},miss} = \mathbb{E}\left[Y | \mathbf{X} \in A^\star_\ell, a^{\star(h)}_\ell \leq \mathbf{X}^{(h)} < z, \mathbf{M}^{(h)} = 1\right]$$
(resp. for $\mu_{A^\star_{\ell,R},obs}$ and $\mu_{A^\star_{\ell,R},miss}$)

and note that

$$\mathbb{P}\left[z \leq \mathbf{X}^{(h)} \leq b^{\star(h)}_\ell | \mathbf{X} \in A^\star_\ell, M^{(h)} = 0\right] = 1 - \mathbb{P}_{A^\star_{\ell,L},obs}$$

$$\mathbb{P}\left[z \leq \mathbf{X}^{(h)} \leq b^{\star(h)}_\ell | \mathbf{X} \in A^\star_\ell, M^{(h)} = 1\right] = 1 - \mathbb{P}_{A^\star_{\ell,L},miss}$$

Let us define

$$\begin{aligned}
&L^\star_{1,A^\star_\ell}\left(h, z, \mathbb{P}_{A^\star_{\ell,L},obs}\right) = \\
&\mathbb{V}\left[Y | \mathbf{X} \in A^\star_\ell,\, \mathbf{M}^{(h)} = 0\right] \mathbb{P}\left[\mathbf{M}^{(h)} = 0 | \mathbf{X} \in A^\star_\ell\right] \\
&- \mathbb{V}\left[Y | \mathbf{X} \in A^\star_\ell,\, a^{\star(h)}_\ell \leq \mathbf{X}^{(h)} < z,\, \mathbf{M}^{(h)} = 0\right] \mathbb{P}\left[\mathbf{M}^{(h)} = 0 | \mathbf{X} \in A^\star_\ell\right] \mathbb{P}_{A^\star_{\ell,L},obs} \\
&- \mathbb{V}\left[Y | \mathbf{X} \in A^\star_\ell,\, z \leq \mathbf{X}^{(h)} \leq b^{\star(h)}_\ell,\, \mathbf{M}^{(h)} = 0\right] \mathbb{P}\left[\mathbf{M}^{(h)} = 0 | \mathbf{X} \in A^\star_\ell\right] \left(1 - \mathbb{P}_{A^\star_{\ell,L},obs}\right)
\end{aligned}$$

$$
\begin{aligned}
&L^{\star}_{2,A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},miss}\right) = \\
&\mathbb{V}\left[Y|\mathbf{X} \in A^{\star}_{\ell},\, M^{(h)} = 1\right]\mathbb{P}\left[\mathbf{M}^{(h)} = 1|\mathbf{X} \in A^{\star}_{\ell}\right] \\
&-\mathbb{V}\left[Y|\mathbf{X} \in A^{\star}_{\ell},\, a^{\star(h)}_{\ell} \leq \mathbf{X}^{(h)} < z,\, \mathbf{M}^{(h)} = 1\right]\mathbb{P}\left[\mathbf{M}^{(h)} = 1|\mathbf{X} \in A^{\star}_{\ell}\right]\mathbb{P}_{A^{\star}_{\ell,L},miss} \\
&-\mathbb{V}\left[Y|\mathbf{X} \in A^{\star}_{\ell},\, z \leq \mathbf{X}^{(h)} \leq b^{\star(h)}_{\ell},\, \mathbf{M}^{(h)} = 1\right]\mathbb{P}\left[\mathbf{M}^{(h)} = 1|\mathbf{X} \in A^{\star}_{\ell}\right]\left(1 - \mathbb{P}_{A^{\star}_{\ell,L},miss}\right)
\end{aligned}
$$

$$
\begin{aligned}
&L^{\star}_{3,A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},obs}, \mathbb{P}_{A^{\star}_{\ell,L},miss}\right) = \\
&\left(\mu_{A^{\star}_{\ell},obs} - \mu_{A^{\star}_{\ell}}\right)^2\mathbb{P}\left[\mathbf{M}^{(h)} = 0|\mathbf{X} \in A^{\star}_{\ell}\right] \\
&-\left(\mu_{A^{\star}_{\ell,L},obs} - \mu_{A^{\star}_{\ell,L}}\right)^2\mathbb{P}\left[\mathbf{M}^{(h)} = 0,\, a^{\star(h)}_{\ell} \leq \mathbf{X}^{(h)} < z|\mathbf{X} \in A^{\star}_{\ell}\right] \\
&-\left(\mu_{A^{\star}_{\ell,R},obs} - \mu_{A^{\star}_{\ell,R}}\right)^2\mathbb{P}\left[\mathbf{M}^{(h)} = 0,\, z \leq \mathbf{X}^{(h)} \leq b^{\star(h)}_{\ell}|\mathbf{X} \in A^{\star}_{\ell}\right]
\end{aligned}
$$

$$
\begin{aligned}
&L^{\star}_{4,A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},obs}, \mathbb{P}_{A^{\star}_{\ell,L},miss}\right) = \\
&\left(\mu_{A^{\star}_{\ell},miss} - \mu_{A^{\star}_{\ell}}\right)^2\mathbb{P}\left[\mathbf{M}^{(h)} = 1|\mathbf{X} \in A^{\star}_{\ell}\right] \\
&-\left(\mu_{A^{\star}_{\ell,L},miss} - \mu_{A^{\star}_{\ell,L}}\right)^2\mathbb{P}\left[\mathbf{M}^{(h)} = 1|\mathbf{X} \in A^{\star}_{\ell}\right]\mathbb{P}_{A^{\star}_{\ell,L},miss} \\
&-\left(\mu_{A^{\star}_{\ell,R},miss} - \mu_{A^{\star}_{\ell,R}}\right)^2\mathbb{P}\left[\mathbf{M}^{(h)} = 1|\mathbf{X} \in A^{\star}_{\ell}\right]\left(1 - \mathbb{P}_{A^{\star}_{\ell,L},miss}\right)
\end{aligned}
$$

We can take the original CART criterion, given in eq. (1.1) and expand it as we did with the eq. (3.2), then applying eq. (A.1) to the latter, it is easy to see that the theoretical version of our CART criterion is

$$
\begin{aligned}
L^{\star}_{A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},obs}, \mathbb{P}_{A^{\star}_{\ell,L},miss}\right) =& L^{\star}_{1,A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},obs}\right) + L^{\star}_{2,A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},miss}\right) \\
&+ L^{\star}_{3,A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},obs}, \mathbb{P}_{A^{\star}_{\ell,L},miss}\right) + L^{\star}_{4,A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},obs}, \mathbb{P}_{A^{\star}_{\ell,L},miss}\right)
\end{aligned}
$$

Using eqs. (3.3) and (3.4), it can be shown that $L^{\star}_{1,A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},obs}\right)$ and $L^{\star}_{2,A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},miss}\right)$ might be written as

$$
L^{\star}_{1,A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},obs}\right) = \mathbb{P}\left[\mathbf{M}^{(h)} = 0|\mathbf{X} \in A^{\star}_{\ell}\right]\mathbb{P}_{A^{\star}_{\ell,L},obs}\left(1 - \mathbb{P}_{A^{\star}_{\ell,L},obs}\right)\left(\mu_{A^{\star}_{\ell,L},obs} - \mu_{A^{\star}_{\ell,R},obs}\right)^2 \tag{3.5}
$$

$$
L^{\star}_{2,A^{\star}_{\ell}}\left(h, z, \mathbb{P}_{A^{\star}_{\ell,L},miss}\right) = \mathbb{P}\left[\mathbf{M}^{(h)} = 1|\mathbf{X} \in A^{\star}_{\ell}\right]\mathbb{P}_{A^{\star}_{\ell,L},miss}\left(1 - \mathbb{P}_{A^{\star}_{\ell,L},miss}\right)\left(\mu_{A^{\star}_{\ell,L},miss} - \mu_{A^{\star}_{\ell,R},miss}\right)^2 \tag{3.6}
$$

In the empirical case, we were looking for the best cut and assignation that maximize the empirical CART criterion. Remember that the assignation is made over all possible combinations for the observations were the hth variable is missing and $\widehat{\mathbf{X}}^{(h)}_{\cdot,\ell} \in A_{\ell}$. We pointed out before that this is a consequence from the

fact that we do not have access to the missing mechanism, where the possible assignations depend on the mechanism.

In the theoretical case, the missing mechanism has a direct effect over $\mathbb{P}_{A^\star_{\ell,L},obs}$ and $\mathbb{P}_{A^\star_{\ell,L},miss}$, if we do not have access to the missing mechanism we were looking for the best cut and the best value for the probabilities $\mathbb{P}_{A^\star_{\ell,L},obs}$ and $\mathbb{P}_{A^\star_{\ell,L},miss}$. While if we have access to the missing mechanism, it will affect the best value that we can assign to those probabilities.

In other words, the best theoretical cut and assignation, $\left(h^\star_\ell, z^\star_\ell, \mathbb{P}^\star_{A^\star_{\ell,L},obs}, \mathbb{P}^\star_{A^\star_{\ell,L},miss}\right)$, is selected maximizing $L^\star_{A^\star_\ell}\left(h, z, \mathbb{P}^\star_{A^\star_{\ell,L},obs}, \mathbb{P}_{A^\star_{\ell,L},miss}\right)$ over $\mathcal{M}_{try}$, $\mathcal{C}_{A^\star_\ell}$ and the allowed values by the missing mechanism.

$$\left(h^\star_\ell, z^\star_\ell, \mathbb{P}^\star_{A^\star_{\ell,L},obs}, \mathbb{P}^\star_{A^\star_{\ell,L},miss}\right) \in \underset{\substack{h\in\mathcal{M}_{try}\\(h,z)\in\mathcal{C}_{A^\star_\ell}\\ \mathbb{P}_{A^\star_{\ell,L},obs}\\ \mathbb{P}_{A^\star_{\ell,L},miss}}}{\arg\max}\ L^\star_{A^\star_\ell}\left(h, z, \mathbb{P}_{A^\star_{\ell,L},obs}, \mathbb{P}_{A^\star_{\ell,L},miss}\right)$$

The theoretical cells are defined as

$$A^\star_1 = \mathcal{X}$$

and

$$A^\star_{2\ell} = A^\star_\ell \cap \left(\prod_{j=1}^{h^\star_\ell - 1}\left[a^{\star(j)}_\ell, b^{\star(j)}_\ell\right] \times \left[a^{\star(h^\star_\ell)}_\ell, z^\star_\ell\right] \times \prod_{j=h^\star_\ell+1}^{p}\left[a^{\star(j)}_\ell, b^{\star(j)}_\ell\right]\right)$$

$$A^\star_{2\ell+1} = A^\star_\ell \cap \left(\prod_{j=1}^{h^\star_\ell - 1}\left[a^{\star(j)}_\ell, b^{\star(j)}_\ell\right] \times \left[z^\star_\ell, b^{\star(h^\star_\ell)}_\ell\right] \times \prod_{j=h^\star_\ell+1}^{p}\left[a^{\star(j)}_\ell, b^{\star(j)}_\ell\right]\right)$$

We want to point out an interesting fact, let us denote the distribution of $\mathbf{X}^{(h)}|\widehat{\mathbf{X}}_\ell \in A^\star_\ell, M^{(h)} = 1$ as $\nu^{(h)}_\ell$

When we select the best cut and assignation, $\left(h^\star_\ell, z^\star_\ell, \mathbb{P}^\star_{A^\star_{\ell,L},miss}\right)$, we estimate $\nu^{(h^\star_\ell)}_\ell$ as

$$\nu^{(h^\star_\ell)}_\ell = \left(\mathbb{P}^\star_{A^\star_{\ell,L},miss}\right)\nu^{(h^\star_\ell)}_{2\ell} + \left(1 - \mathbb{P}^\star_{A^\star_{\ell,L},miss}\right)\nu^{(h^\star_\ell)}_{2\ell+1}$$

where $\nu^{(h^\star_\ell)}_{2\ell}$ is the distribution of $\mathbf{X}^{(h^\star_\ell)}|a^{\star(h^\star_\ell)}_\ell \leq \mathbf{X}^{(h^\star_\ell)} < z^\star_\ell \mathbf{X} \in A^\star_\ell, M^{(h^\star_\ell)} = 1$ and $\nu^{(h^\star_\ell)}_{2\ell+1}$ is the distribution of $\mathbf{X}^{(h^\star_\ell)}|z^\star_\ell \leq \mathbf{X}^{(h^\star_\ell)} < b^{\star(h^\star_\ell)}_\ell \mathbf{X} \in A^\star_\ell, M^{(h^\star_\ell)} = 1$

An important consequence is that we can estimate the distribution of $\mathbf{X}^{(h)}|M^{(h)} = 1$ in the empirical case, since

$$\begin{aligned}\nu^{(h)}_1 &= \nu\left(\mathbf{X}^{(h)}|\widehat{\mathbf{X}}_1 \in A^\star_1, M^{(h)} = 1\right)\\ &= \nu\left(\mathbf{X}^{(h)}|M^{(h)} = 1\right)\end{aligned}$$

From the empirical perspective, $\mathbb{P}^\star_{A^\star_{\ell,L},miss}$ is estimated as

$$\widehat{\mathbb{P}}^\star_{A^\star_{\ell,L},miss} = \frac{N^\star_{n,miss}(A_{\ell,L})}{N_{n,miss}(A_\ell)}$$

where $N^\star_{n,miss}(A_{\ell,L})$ is the number of observations assigned to the left child node after the selection of the best (empirical) cut and assignation.

And the distribution $\nu^{(h^\star_{n,\ell})}_\ell$ is estimated as

$$\hat{\nu}_{\ell}^{(h^{\star}_{n,\ell})} = \left(\widehat{\mathbb{P}}^{\star}_{A^{\star}_{\ell,L},miss}\right)\hat{\nu}_{2\ell}^{(h^{\star}_{n,\ell})} + \left(1-\widehat{\mathbb{P}}^{\star}_{A^{\star}_{\ell,L},miss}\right)\hat{\nu}_{2\ell+1}^{(h^{\star}_{n,\ell})}$$

Since, we do not have access to the distribution of $\mathbf{X}^{(h^{\star}_{\ell})}|a_{\ell}^{\star\,(h^{\star}_{\ell})} \leq \mathbf{X}^{(h^{\star}_{\ell})} < z^{\star}_{\ell}\mathbf{X} \in A^{\star}_{\ell}, M^{(h^{\star}_{\ell})} = 1$ and $\mathbf{X}^{(h^{\star}_{\ell})}|z^{\star}_{\ell} \leq \mathbf{X}^{(h^{\star}_{\ell})} < b_{\ell}^{\star\,(h^{\star}_{\ell})}\mathbf{X} \in A^{\star}_{\ell}, M^{(h^{\star}_{\ell})} = 1$, when we arrive to a the final nodes, we can estimate these distributions as uniform distributions in $\left[a_{\ell}^{(h^{\star}_{n,\ell})}, z^{\star}_{n,\ell}\right]$ and $\left[z^{\star}_{n,\ell}, b_{\ell}^{(h^{\star}_{n,\ell})}\right]$.

3.6 Analysis of the CART Criterion

To be written!

In this section I explain that L1 and L2 are related to a variance introduced by the missing mechanism, L3 and L4 are related to the bias introducedby the missing mechanism. This explain why in the MCAR case L3 and L4 are zero, because MCAR does not introduced bias (as we can see in the simulation study).

Chapter 4

Prediction with our Approach

When considering missing data, special care must be taken with the prediction. If consider that both the train data and the test data come from the same distribution and missing data mechanism, it is probable that we will have missing values not only in the train data set but also in the test set. Therefore, an appropriate approach to handle missing data must be able to predict when there is no complete information in a new point.

We dedicate this section to explain how our method deals with missing values in the test data set.

If there is no missing values in a new observations we follow the standard approach to predict from a tree, that is, we drop down the new observation till a final cell. However, when the new observation has missing the value of a variable use at some point in the tree, it is no possible to drop down the observation to a final cell.

To deal with this phenomenon several techniques have been proposed, which can be classified in three different groups.

To be written!

Chapter 5

Scornet's Results

5.1 Hypothesis 1 (MCAR)

The response Y follows

$$Y = \sum_{j=1}^{p} m_j(\mathbf{X}^{(j)}) + \varepsilon$$

where $\mathbf{X}$ is uniformly distributed over $[0,1]^p$, there is a MCAR mechanism for all variables, ε is an independent centered Gaussian noise with finite variance $\sigma^2 > 0$ and each component m_j is continuous.

5.2 Technical Lema 1

Assume that Hypothesis 1 (MCAR) is satisfied, and that $L^\star_{A^\star_\ell} \equiv 0$ for all cuts. Then the regression function m is constant on $A^\star_\ell$.

5.2.1 Proof

Because of the MCAR assumption,

$$\mu_{A^\star_\ell,obs} = \mu_{A^\star_\ell,miss} = \mu_{A^\star_\ell}$$

$$\mu_{A^\star_{\ell,L},obs} = \mu_{A^\star_{\ell,L},miss} = \mu_{A^\star_{\ell,L}}$$

$$\mu_{A^\star_{\ell,R},obs} = \mu_{A^\star_{\ell,R},miss} = \mu_{A^\star_{\ell,R}}$$

So, $L^\star_{3,A^\star_\ell}$ and $L^\star_{4,A^\star_\ell}$ are trivially equal to zero.

Also, because of the same assumption, the only variable in $L^\star_{1,A^\star_\ell}$ and $L^\star_{2,A^\star_\ell}$ is the cut (h,z), taking advantage of eqs. (3.5) and (3.6), it is directly that

$$L^\star_{1,A^\star_\ell}(h,z) = \mathbb{P}\left[\mathbf{M}^{(h)} = 0\right] \mathbb{P}\left[a^{\star\,(h)}_\ell \leq \mathbf{X}^{(h)} < z | \mathbf{X} \in A^\star_\ell\right] \mathbb{P}\left[z \leq \mathbf{X}^{(h)} \leq b^{\star\,(h)}_\ell | \mathbf{X} \in A^\star_\ell\right] \left(\mu_{A^\star_{\ell,L}} - \mu_{A^\star_{\ell,R}}\right)^2$$

and

$$L^\star_{2,A^\star_\ell}(h,z) = \mathbb{P}\left[\mathbf{M}^{(h)} = 1\right] \mathbb{P}\left[a^{\star\,(h)}_\ell \leq \mathbf{X}^{(h)} < z | \mathbf{X} \in A^\star_\ell\right] \mathbb{P}\left[z \leq \mathbf{X}^{(h)} \leq b^{\star\,(h)}_\ell | \mathbf{X} \in A^\star_\ell\right] \left(\mu_{A^\star_{\ell,L}} - \mu_{A^\star_{\ell,R}}\right)^2$$

So, our CART criterion, $L^\star_{A^\star_\ell}\left(h, z, \mathbb{P}_{A^\star_{\ell,L},obs}, \mathbb{P}_{A^\star_{\ell,L},miss}\right)$, is nothing but the original CART criterion on cell $A^\star_\ell$, that is

$$L^\star_{A^\star_\ell}(h,z) = \mathbb{P}\left[a^\star_\ell \leq \mathbf{X}^{(h)} < z | \mathbf{X} \in A^\star_\ell\right] \mathbb{P}\left[z \leq \mathbf{X}^{(h)} \leq b^\star_\ell | \mathbf{X} \in A^\star_\ell\right] \left(\mu_{A^\star_{\ell,L}} - \mu_{A^\star_{\ell,R}}\right)^2$$

and Scornet's Technical Lema 1 follows.

To prove this lemma we use ideas of Scornet, Biau, and Vert (2015), but we change a little bit the proof. Because $\mathbf{X}$ is uniformly distributed over $[0,1]^p$,

$$\mathbb{P}\left[a^\star_\ell \leq \mathbf{X}^{(h)} < z | \mathbf{X} \in A^\star_\ell\right] = \frac{z - a_\ell^{\star(h)}}{b_\ell^{\star(h)} - a_\ell^{\star(h)}}$$

and

$$\mathbb{P}\left[z \leq \mathbf{X}^{(h)} \leq b^\star_\ell | \mathbf{X} \in A^\star_\ell\right] = \frac{b_\ell^{\star(h)} - z}{b_\ell^{\star(h)} - a_\ell^{\star(h)}}$$

The same assumption implies that

$$\begin{aligned}
\mu_{A^\star_{\ell,L}} &= \mathbb{E}\left[\sum_{j=1}^p m_j(\mathbf{X}^{(j)}) | a_\ell^{\star(h)} \leq \mathbf{X}^{(h)} < z, \mathbf{X} \in A^\star_\ell\right] \\
&= \sum_{j \neq h} \mathbb{E}\left[m_j(\mathbf{X}^{(j)}) | a_\ell^{\star(j)} \leq \mathbf{X}^{(h)} \leq b_\ell^{\star(j)}\right] + \mathbb{E}\left[m_h(\mathbf{X}^{(h)}) | a_\ell^{\star(h)} \leq \mathbf{X}^{(h)} < z\right] \\
&= \sum_{j \neq h} \frac{1}{b_\ell^{\star(j)} - a_\ell^{\star(j)}} \int_{a_\ell^{\star(j)}}^{b_\ell^{\star(j)}} m_j(t)dt + \frac{1}{z - a_\ell^{\star(h)}} \int_{a_\ell^{\star(h)}}^{z} m_h(t)dt
\end{aligned}$$

Let $C_j = \int_{a_\ell^{\star(j)}}^{b_\ell^{\star(j)}} m(t)dt$ and $M_h(z) = \int_{a_\ell^{\star(h)}}^{z} m(t)dt$, then

$$\mu_{A^\star_{\ell,L}} = \sum_{j \neq h} \frac{1}{b_\ell^{\star(j)} - a_\ell^{\star(j)}} C_j + \frac{1}{z - a_\ell^{\star(h)}} M_h(z)$$

Analogously, it is easy to show that

$$\mu_{A^\star_{\ell,R}} = \sum_{j \neq h} \frac{1}{b_\ell^{\star(j)} - a_\ell^{\star(j)}} C_j - \frac{1}{b_\ell^{\star(h)} - z} M_h(z) + \frac{1}{b_\ell^{\star(h)} - z} C_h$$

Therefore,

$$\begin{aligned}
L^\star_{A^\star_\ell}(h,z) &= \left(\frac{z - a_\ell^{\star(h)}}{b_\ell^{\star(h)} - a_\ell^{\star(h)}}\right)\left(\frac{b_\ell^{\star(h)} - z}{b_\ell^{\star(h)} - a_\ell^{\star(h)}}\right)\left(\frac{1}{z - a_\ell^{\star(h)}} M_h(z) + \frac{1}{b_\ell^{\star(h)} - z} M_h(z) - \frac{1}{b_\ell^{\star(h)} - z} C_h\right)^2 \\
&= \frac{1}{(z - a_\ell^{\star(h)})(b_\ell^{\star(h)} - z)}\left(M_h(z) - \frac{z - a_\ell^{\star(h)}}{b_\ell^{\star(h)} - a_\ell^{\star(h)}} C_h\right)^2
\end{aligned}$$

Since $L^\star_{A^\star_\ell} \equiv 0$ by assumption, we obtain

$$M_h(z) = \frac{z - a_\ell^{\star(h)}}{b_\ell^{\star(h)} - a_\ell^{\star(h)}} C_h, \quad \forall\, h$$

This prove that $M_h(z)$ is linear in z therefore m_h is constant on $[a_\ell^{\star(h)}, b_\ell^{\star(h)}]$ and it is straightforward that m is constant in $A_\ell^\star$.

5.3 Lemma 1

Assume that Hypothesis 1 (MCAR) is satisfied. Then, for all $\mathbf{x} \in [0,1]^p$,

$$\Delta(m, A_k^\star(\mathbf{x}, \Theta)) \to 0 \quad \text{almost surely, as } k \to \infty$$

where $\Delta(m, A) = \sup_{\mathbf{x},\mathbf{x}' \in A} |m(\mathbf{x}) - m(\mathbf{x}')|$.

5.3.1 Proof

Fix $\mathbf{x} \in [0,1]^p$ and let θ be a realization of the random variable Θ. We also let $A_k^\star(\mathbf{x}, \theta)$ be a cell of the theoretical random tree at level k, containing $\mathbf{x}$ and designed with θ.

Since $(A_k^\star(\mathbf{x}, \theta))_k$ is a decreasing sequence of compact sets, there exist $\mathbf{a}_\infty(\mathbf{x}, \theta) = (a_\infty^{(1)}(\mathbf{x}, \theta), \ldots, a_\infty^{(p)}(\mathbf{x}, \theta)) \in [0,1]^p$ and $\mathbf{b}_\infty(\mathbf{x}, \theta) = (b_\infty^{(1)}(\mathbf{x}, \theta), \ldots, b_\infty^{(p)}(\mathbf{x}, \theta)) \in [0,1]^p$ such that

$$\bigcap_{k=1}^{\infty} A_k^\star(\mathbf{x}, \theta) = \prod_{j=1}^{p} \left[a_\infty^{(j)}(\mathbf{x}, \theta), b_\infty^{(j)}(\mathbf{x}, \theta) \right] \stackrel{\text{def}}{=} A_\infty^\star(\mathbf{x}, \theta)$$

We first show that

$$\lim_{k \to \infty} \Delta(m, A_k^\star(\mathbf{x}, \theta)) = \Delta(m, A_\infty^\star(\mathbf{x}, \theta))$$

Since $A_k^\star(\mathbf{x}, \theta) \downarrow A_\infty^\star(\mathbf{x}, \theta)$, for every $\delta > 0$ there exists k_0 such that, for all $k \geq k_0$

$$\max\left(\|\mathbf{a}_k(\mathbf{x}, \theta) - \mathbf{a}_\infty(\mathbf{x}, \theta)\|_\infty, \|\mathbf{b}_k(\mathbf{x}, \theta) - \mathbf{b}_\infty(\mathbf{x}, \theta)\|_\infty \right) \leq \delta$$

In the sequel we assume that $k \geq k_0$. Let

$$\pi_{A_k}(\mathbf{y}) = \underset{\mathbf{z} \in A_\infty^\star(\mathbf{x}, \theta)}{\arg\min} \|\mathbf{z} - \mathbf{y}\|_\infty, \quad \mathbf{y} \in A_k^\star(\mathbf{x}, \theta)$$

Note that (see fig. 5.1 for an illustration in dimension 2),

$$\|\mathbf{y} - \pi(\mathbf{y})\|_\infty \begin{cases} = 0 & \text{if} \quad \mathbf{y} \in A_\infty^\star(\mathbf{x}, \theta) \\ \leq \delta & \text{if} \quad \mathbf{y} \in A_k^\star(\mathbf{x}, \theta) \setminus A_\infty^\star(\mathbf{x}, \theta) \end{cases}$$

Let $\xi > 0$ be arbitrarily small, since m is uniformly continuous $m(\mathbf{y} - m(\pi_{A_k}(\mathbf{y}))) \leq \xi$
Thus,

$$\begin{aligned}
\sup_{\mathbf{y},\mathbf{y}' \in A_k^\star(\mathbf{x},\theta)} |m(\mathbf{y}) - m(\mathbf{y}')| &= \sup_{\mathbf{y},\mathbf{y}' \in A_k^\star(\mathbf{x},\theta)} |m(\mathbf{y}) - m(\pi_{A_k}(\mathbf{y})) + m(\pi_{A_k}(\mathbf{y})) - m(\pi_{A_k}(\mathbf{y}')) + m(\pi_{A_k}(\mathbf{y}')) - m(\mathbf{y}')| \\
&\leq 2 \sup_{\mathbf{y} \in A_k^\star(\mathbf{x},\theta)} |m(\mathbf{y}) - m(\pi_{A_k}(\mathbf{y}))| + \sup_{\mathbf{y},\mathbf{y}' \in A_k^\star(\mathbf{x},\theta)} |m(\pi_{A_k}(\mathbf{y})) - m(\pi_{A_k}(\mathbf{y}'))| \\
&\leq 2\xi + \sup_{\mathbf{y},\mathbf{y}' \in A_\infty^\star(\mathbf{x},\theta)} |m(\mathbf{y}) - m(\mathbf{y}')|
\end{aligned}$$

Therefore,

$$\sup_{\mathbf{y},\mathbf{y}'\in A^\star_\infty(\mathbf{x},\theta)} |m(\mathbf{y}) - m(\mathbf{y}')| \leq \sup_{\mathbf{y},\mathbf{y}'\in A^\star_k(\mathbf{x},\theta)} |m(\mathbf{y}) - m(\mathbf{y}')| \leq 2\xi + \sup_{\mathbf{y},\mathbf{y}'\in A^\star_\infty(\mathbf{x},\theta)} |m(\mathbf{y}) - m(\mathbf{y}')|$$

Which implies that

$$\left| \sup_{\mathbf{y},\mathbf{y}'\in A^\star_k(\mathbf{x},\theta)} |m(\mathbf{y}) - m(\mathbf{y}')| - \sup_{\mathbf{y},\mathbf{y}'\in A^\star_\infty(\mathbf{x},\theta)} |m(\mathbf{y}) - m(\mathbf{y}')| \right| \leq 2\xi$$

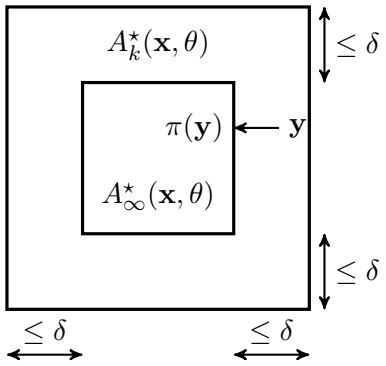

Figure 5.1: Illustration in $p = 2$.

On the other hand, if $L^\star$ is identically zero for all cuts in $A^\star_\infty$, then m is constant on $A^\star_\infty(\mathbf{x},\theta)$ according to Tecnical Lemma 1. This implies that $\Delta(m, A^\star_\infty(\mathbf{x},\theta)) = 0$, and

$$\lim_{k\to\infty} \Delta(m, A^\star_k(\mathbf{x},\theta)) = \Delta(m, A^\star_\infty(\mathbf{x},\theta)) = 0$$

Let us now show by contradiction that $L^\star$ is almost surely necessarily zero for every cut in $A^\star_\infty(\mathbf{x},\theta)$

5.4 Lemma 2

To clarify the notation **To be witten!!**

5.4.1 Proof

To be witten!!

Chapter 6

Simulation

We have developed a simulation study for our proposed approach. We study its mean squared error (MSE) and its bias, taking the previous work of Rieger, Hothorn, and Strobl (2010) to create datasets with missing values under different mechanisms of missingness.

The regression function in this simulation study is the so-called "friedman1" (Friedman, 1991), given by

$$m(\mathbf{x}) = 10 \sin\left(\pi \mathbf{x}^{(1)} \mathbf{x}^{(2)}\right) + 20\left(\mathbf{x}^{(3)} - 0.5\right)^2 + 10\mathbf{x}^{(4)} + 5\mathbf{x}^{(5)}$$

We simulate 5 independent uniformly distributed variables.

6.1 Missing Mechanisms

We introduce missing values in $\mathbf{X}^{(1)}, \mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$, to do so, we consider 7 different missing mechanisms, to be MCAR, MAR or NMAR.

6.1.1 Missing Completely at Random (MCAR)

We select many locations as desired sampled out of the n observations and replaced them by `NA`.

6.1.2 Missing at Random

Methods for generating values missing at random are more complicated. The choice of the locations that are replaced by missing values in the "missing" variable now depends on the value of a second variable, that we will term "determining" variable. Therefore, the values of the "determining" variable now have influence on whether a value in the "missing" variable is missing or not..

Creation of ranks (MAR1) Let be $tot = n(n+1)/2$, the probability for a missing value in a certain location in the "missing" variable is computed by dividing the rank of the location in the "determining" variable by tot. The locations for `NA` in the "missing" variable are then sampled with the resulting probability vector.

Creation of two groups (MAR2) We divide the data set in two groups defined by the "determining" variable. A value belongs to the first group if the value in the "determining" variable is greater than or equal to the median of the "determining" variable, otherwise it belongs to the second group. An observation in the respective group has a missing vale probability of 0.9 or 0.1 divided by the number of members in the group. The locations for `NA` in the "missing" variable are then sampled with the resulting probability vector.

Dexter truncation (MAR3) The observations with the biggest values in the "determining" variable have the "missing" variable replaced by `NA` until the desired fraction of `NA` has been achieved.

Symmetric truncation (MAR4) This method is similar to the previous one but the we replace by `NA` the values in the "missing" variable with the biggest and the smallest values in the "determining" variable.

Missings depending on Y (DEPY) The missing values depend on the value of the response, the probability is 0.1 for values $Y \geq 13$, otherwise it is 0.4. The locations for `NA` in the "missing" variable are then sampled with the resulting probability vector.

6.1.3 Not Missing at Random

Logit modelling (LOG) In this method the probability for `NA` no longer depends on a single "determining" variable but in all the others variables.

It is modelled as

$$\text{logit}\left(\mathbb{P}\left[M^{(h)}=1\right]\right) = -0.5 + \sum_{\substack{k=1\\k\neq h}}^{5} \mathbf{X}^{(k)}$$

Since more than one variable will have missing values the probability of missingness depends on variables with missing values too.

6.2 Fraction of Missingness

The fraction of missingness is the same in each learning data set. In the first variable, 20% of data are missing, in the third variable, the amount is 10% of missing, in the fourth variable there are again 20% missing.

Additionally, for the MAR-functions "determining" variables are needed. For missingness in $\mathbf{X}^{(1)}$ the "determining" variable is $\mathbf{X}^{(2)}$. The variable $\mathbf{X}^{(5)}$ is used as the "determining" variable for $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$.

We create one test data set, and 10 learning data sets for each missing mechanism, also, we keep the 10 learning data sets with no missing values for the simulation analysis. Each learning data set contains 200 observations and the test data set contains 2000 observations. For each learning data set we create missing values accordingly with the fraction of missingness and missing mechanisms described previously, while we do not create missing values in the test data set.

6.3 Random Forests' Parameters

To build the random forests, we use $M = 50$ trees, which has been seeing by simulation to be sufficient to stabilized the error in the case of complete learning data sets.

For the other parameters we use the default values in the regression mode of the `R` package `randomForests`, the paramter `mtry` is set to $\lfloor p/3 \rfloor$ in the case of sample without replace a_n is set to $\lceil 0.632n \rceil$, that is `mtry` $= 1$, $a_n = 127$ and `nodesize` is set to 5.

6.4 Results

Figure 6.1 shows the violin plot (Hintze and Nelson, 1998) of the squared errors in the test data set, using one random forest for each missing mechanism. While figs. 6.2 and 6.3 show the box plot (Tukey, 1977) of the mean squared error and bias, respectively, for the ten random forests built in each missing mechanism.

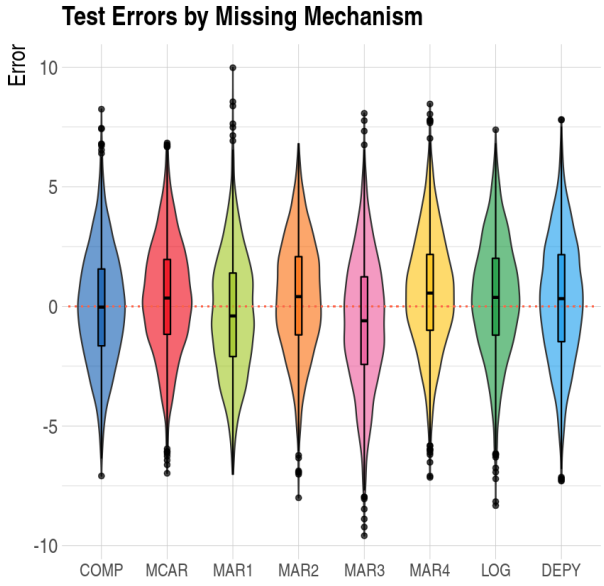

Figure 6.1: MSE of the test data set using one random forest, varying the mechanism of missingness.

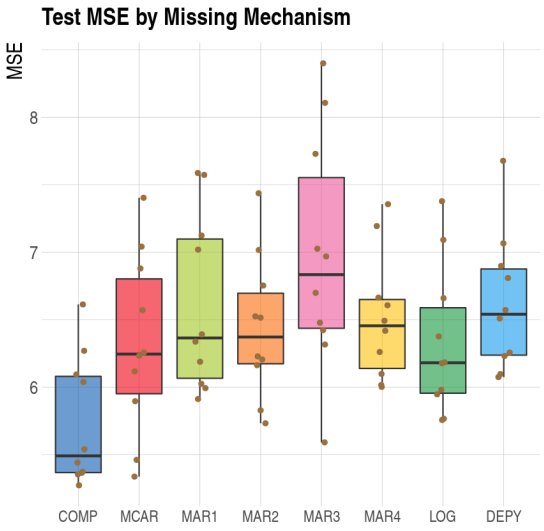

Figure 6.2: MSE of the test data set for random forests, varying the mechanism of missingness.

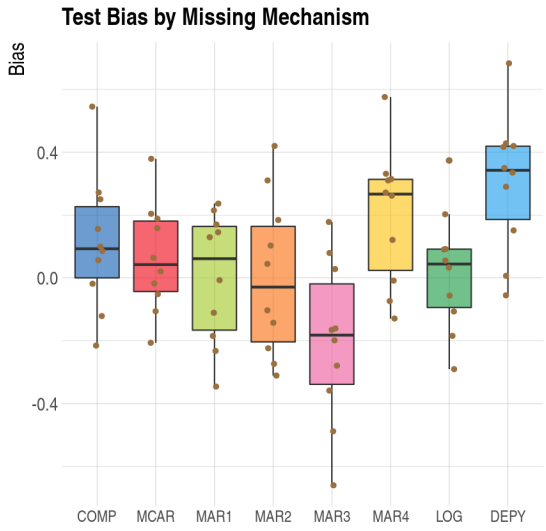

Figure 6.3: Bias of the test data set for random forests, varying the mechanism of missingness.

6.5 Changing the Fraction of Missingness

Using the ten learning data sets with no missing values, we have calculated the importance of the variables using the `R` package `randomForests` by percentage of increase in mean squared error and by increase in node purity (Breiman, 2001; Breiman, 2003).

With this in mind we would like to see the effect of different fraction of missing values in the three variables $\mathbf{X}^{(1)}, \mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$ by order of importance. To do this, we changed the fraction of missing values to 5%, 10%, 20%, 40%, 60%, 80%, 90% and 95% leaving the others with the fraction of missing values unchanged.

6.5.1 Variables Importance

Figures 6.4 and 6.5 show the lollipop charts (atributed to Andy Cotgreave in Cairo (2016)) for the mean squared error and the increase in node purity, respectively. Variables where missing values were introduced are marked with an orange diamond, while complete variables are marked with a blue circle. We can see a consistently order, in both measures of importance, for the variables with missing values, given by $\mathbf{X}^{(4)}$, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$.

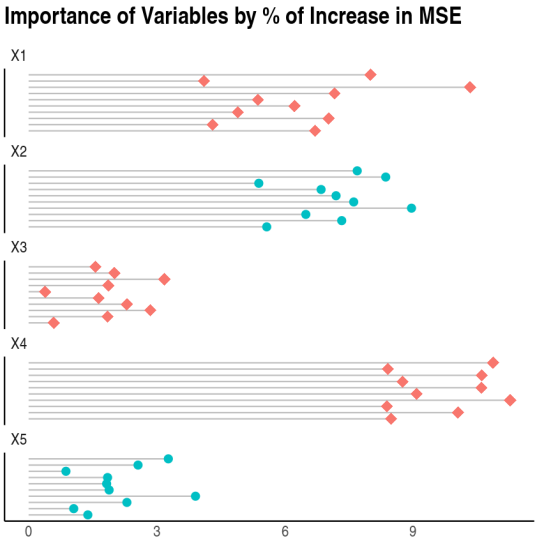

Figure 6.4: Importance Variable accordingly to percentage increase in MSE.

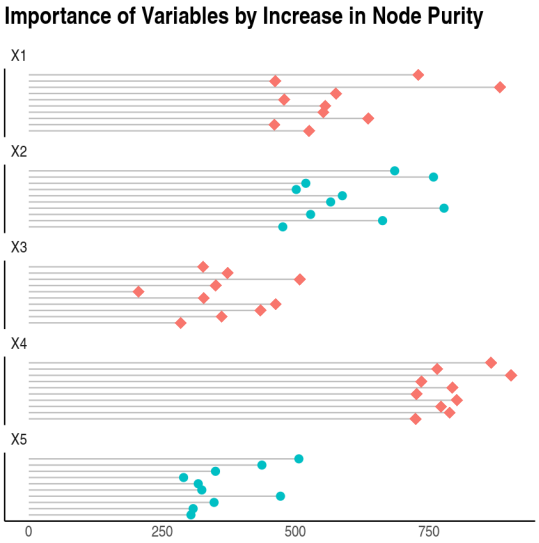

Figure 6.5: Importance Variable accordingly to increase in node purity.

6.5.2 Mean Squared Error

Figures 6.6 to 6.8 show the MSE of our approach varying the percentage for $\mathbf{X}^{(4)}$, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$, respectively. We observe that for some missing mechanisms the MSE has less increase than in others. Also, even for a high percentage of missing values, the MSE seems under control.

Figures 6.9 to 6.11 show the same results, but this time we put together all the different missing mechanisms, highlighting the *MCAR, MAR3, LOG* and *DEPY* mechanisms. We observe the MSE for *MAR3* and *DEPY* is consistently higher to the rest while the MSE for *MCAR* and *LOG* is consistently lower.

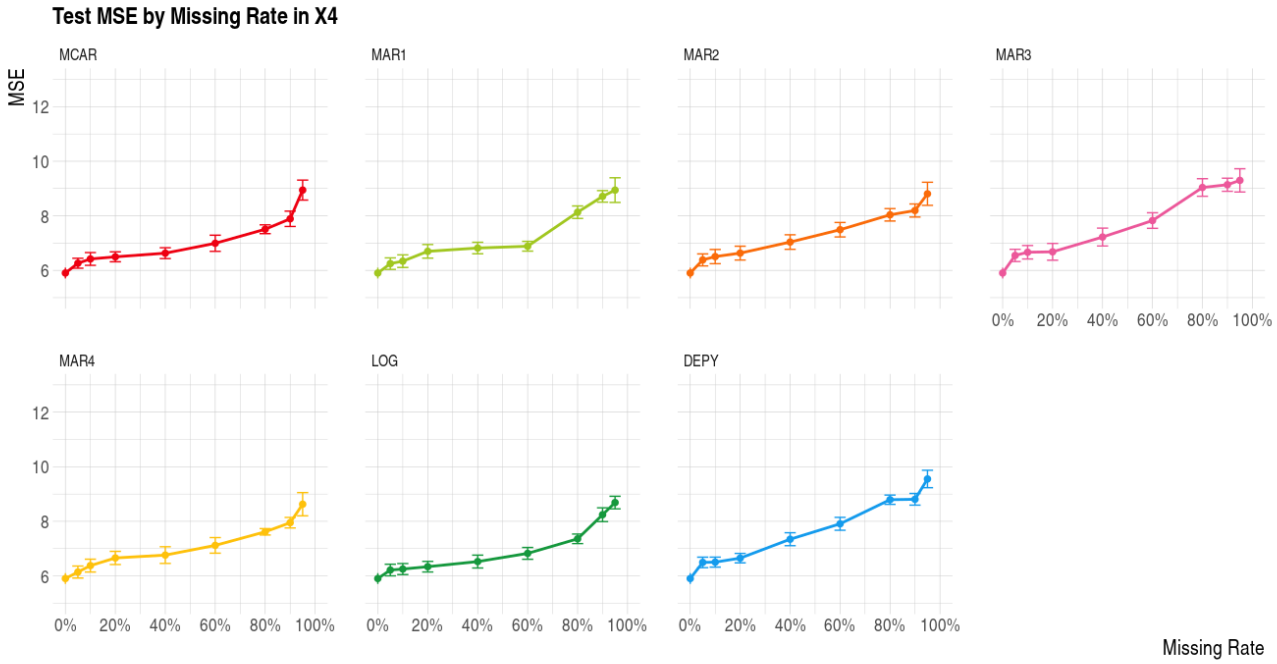

Figure 6.6: MSE of the test data set, varying the percentage of missing values in $\mathbf{X}^{(4)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$ are set to 20% and 10%, respectively.

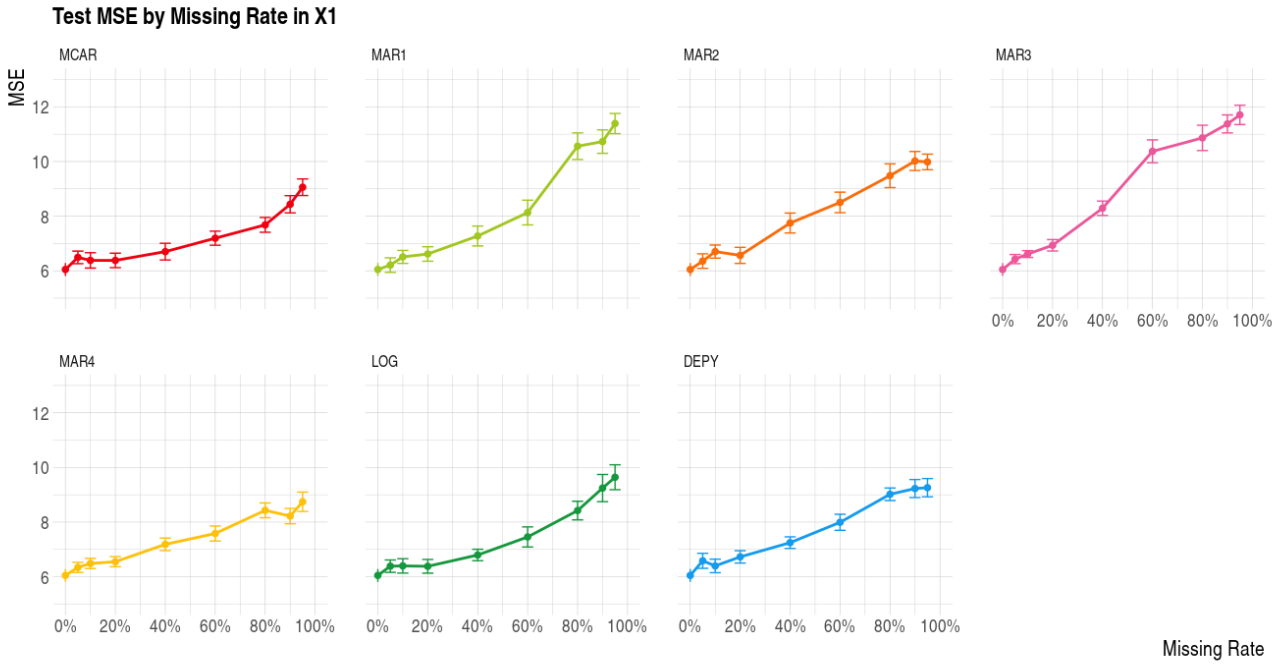

Figure 6.7: MSE of the test data set, varying the percentage of missing values in $\mathbf{X}^{(1)}$, missing values percentage for $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$ are set to 10% and 20%, respectively.

Figure 6.8: MSE of the test data set, varying the percentage of missing values in $\mathbf{X}^{(3)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(4)}$ are set to 20% and 20%, respectively.

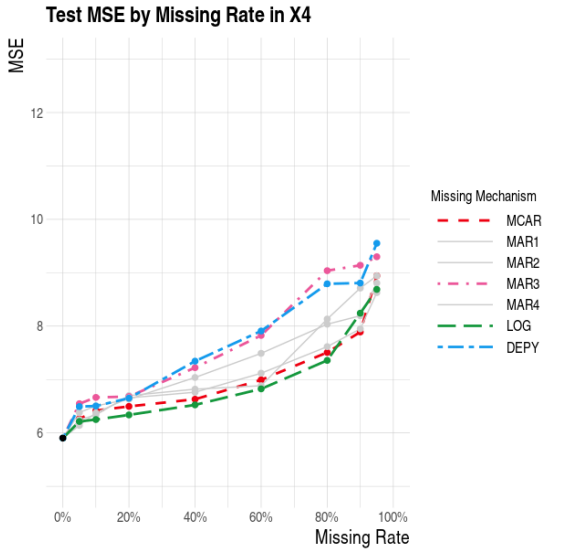

Figure 6.9: MSE of the test data set, varying the percentage of missing values in $\mathbf{X}^{(4)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$ are set to 20% and 10%, respectively.

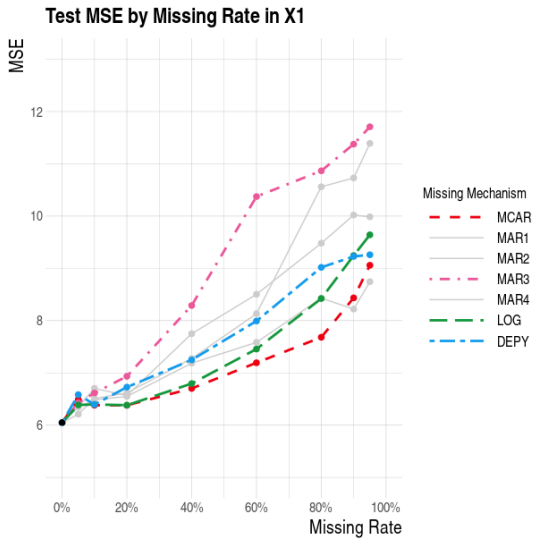

Figure 6.10: MSE of the test data set, varying the percentage of missing values in $\mathbf{X}^{(1)}$, missing values percentage for $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$ are set to 10% and 20%, respectively.

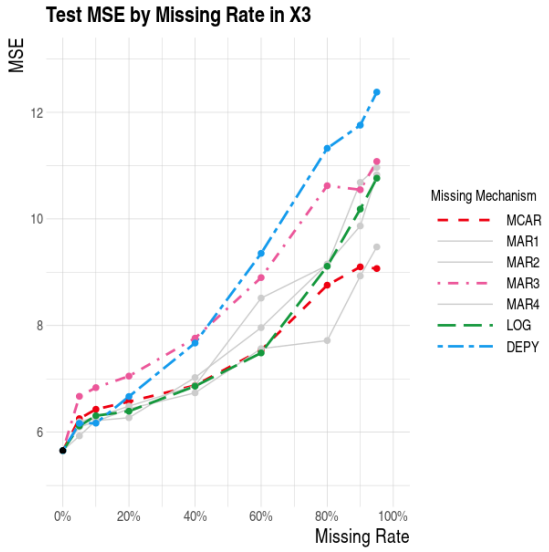

Figure 6.11: MSE of the test data set, varying the percentage of missing values in $\mathbf{X}^{(3)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(4)}$ are set to 20% and 20%, respectively.

6.5.3 Bias

Figures 6.12 to 6.14 show the bias of our approach varying the percentage for $\mathbf{X}^{(4)}$, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$, respectively. We observe that the bias is highly dependent on the missing mechanism.

Figures 6.15 to 6.17 show the same results, but this time we put together all the different missing mechanisms, highlighting the *MCAR, MAR3, LOG* and *DEPY* mechanisms. We observe *MAR3* and *DEPY* are the less unbiased, while *MCAR* and *LOG* are the more unbiased.

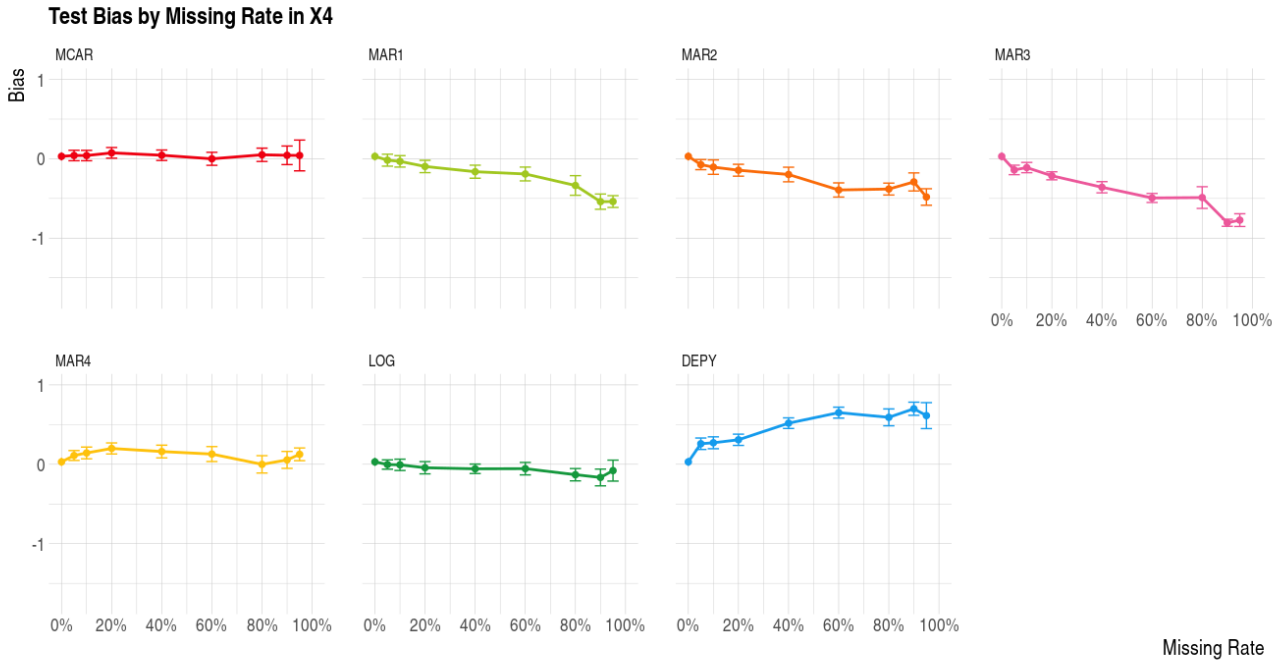

Figure 6.12: Bias of the test data set, varying the percentage of missing values in $\mathbf{X}^{(4)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$ are set to 20% and 10%, respectively.

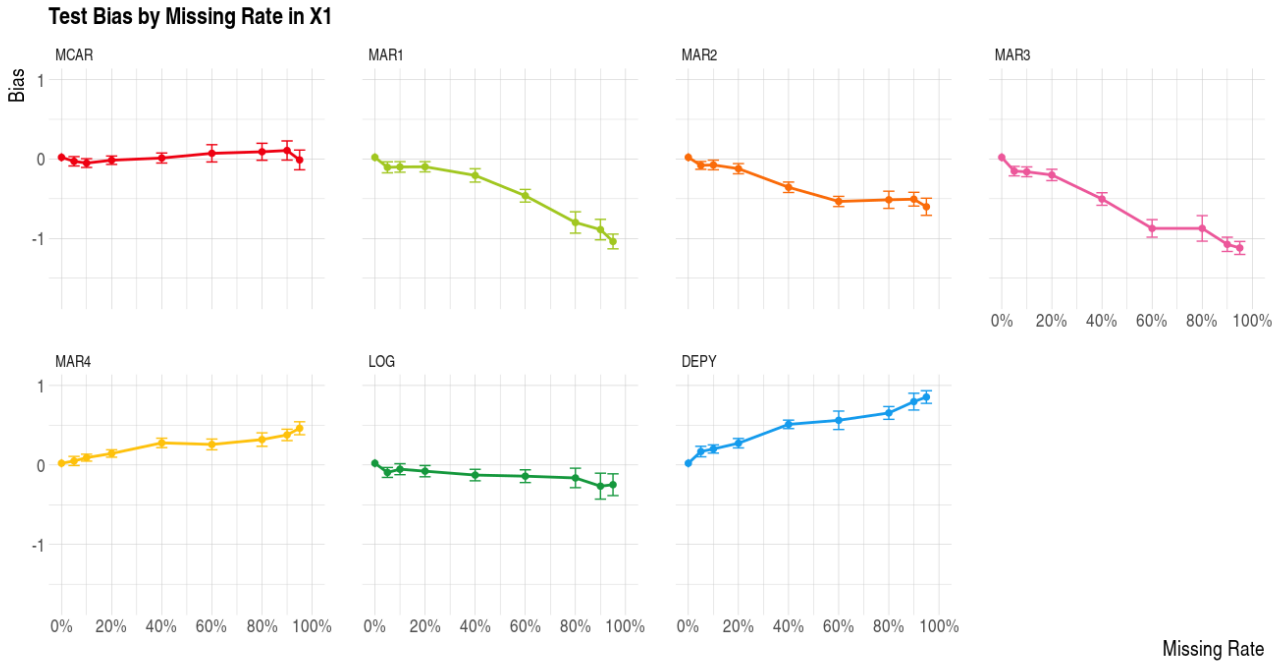

Figure 6.13: Bias of the test data set, varying the percentage of missing values in $\mathbf{X}^{(1)}$, missing values percentage for $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$ are set to 10% and 20%, respectively.

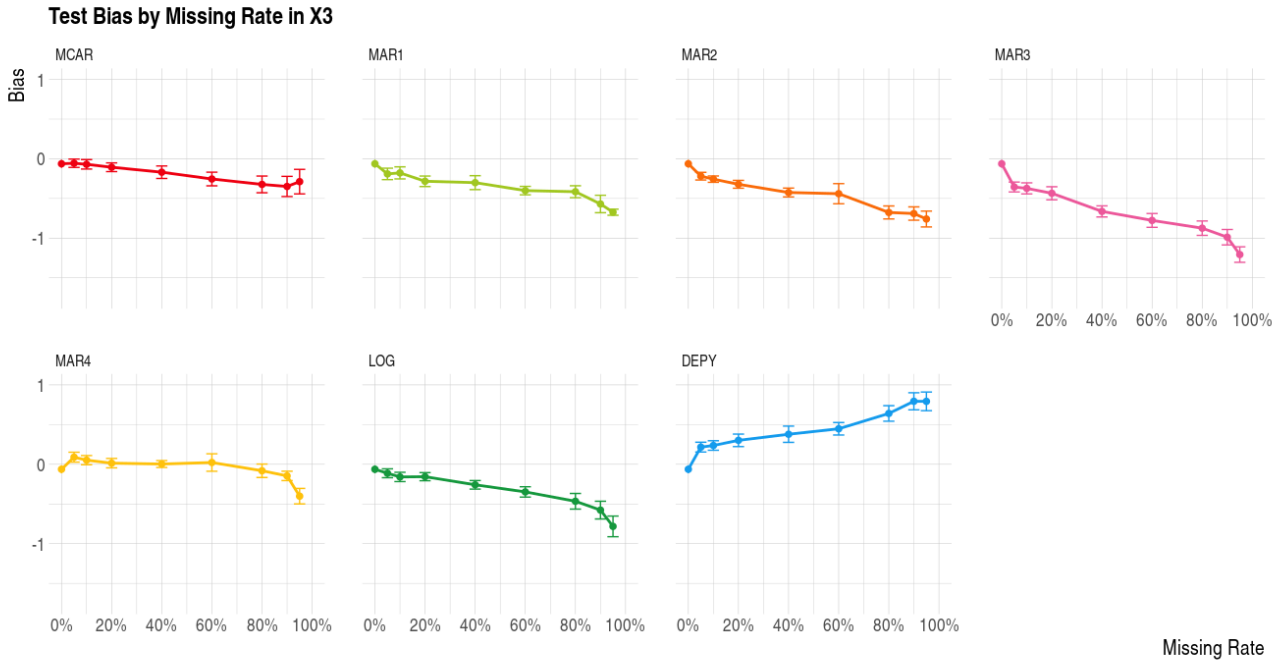

Figure 6.14: Bias of the test data set, varying the percentage of missing values in $\mathbf{X}^{(3)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(4)}$ are set to 20% and 20%, respectively.

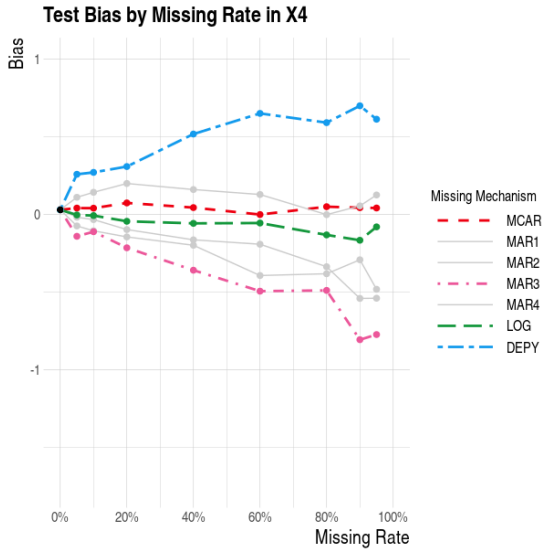

Figure 6.15: Bias of the test data set, varying the percentage of missing values in $\mathbf{X}^{(4)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$ are set to 20% and 10%, respectively.

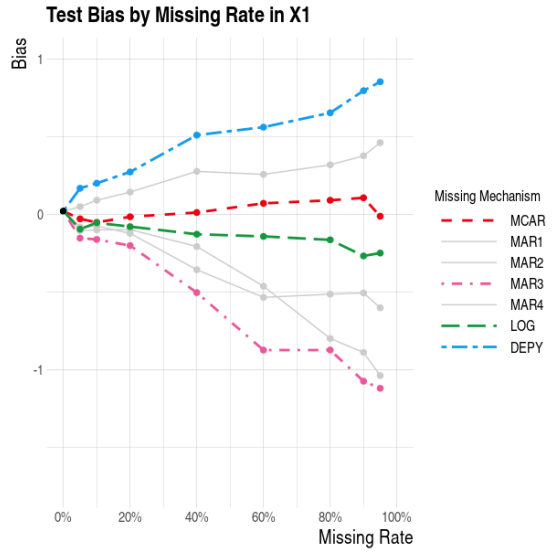

Figure 6.16: Bias of the test data set, varying the percentage of missing values in $\mathbf{X}^{(1)}$, missing values percentage for $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$ are set to 10% and 20%, respectively.

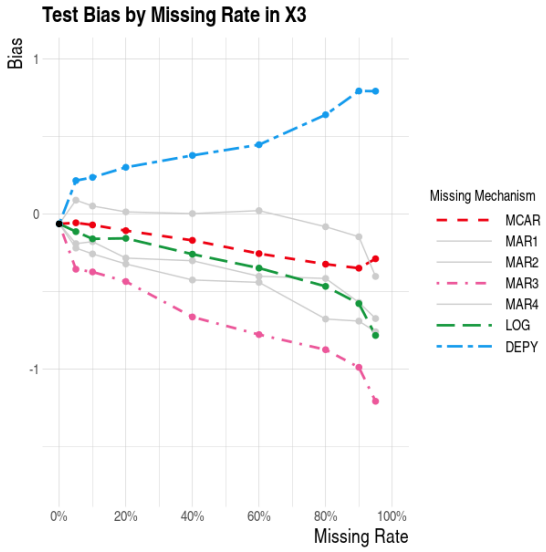

Figure 6.17: Bias of the test data set, varying the percentage of missing values in $\mathbf{X}^{(3)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(4)}$ are set to 20% and 20%, respectively.

6.5.4 Variance

Figures 6.18 to 6.20 show the variance of our approach varying the percentage for $\mathbf{X}^{(4)}$, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$, respectively. This variance has been calculated as the difference between the MSE and the squared of the bias. We observed a similar behavior to the case when MSE was studied.

Figures 6.21 to 6.23 show the same results, but this time we put together all the different missing mechanisms, highlighting the *MCAR, MAR3, LOG* and *DEPY* mechanisms.

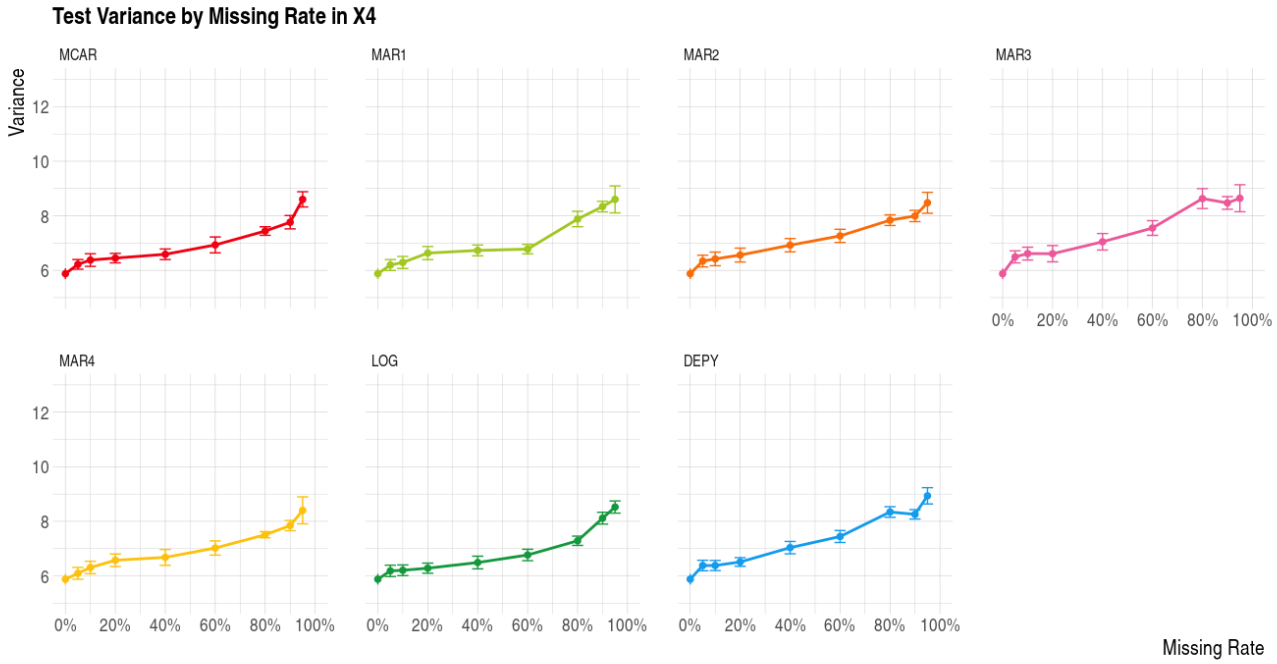

Figure 6.18: Variance of the test data set, varying the percentage of missing values in $\mathbf{X}^{(4)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$ are set to 20% and 10%, respectively.

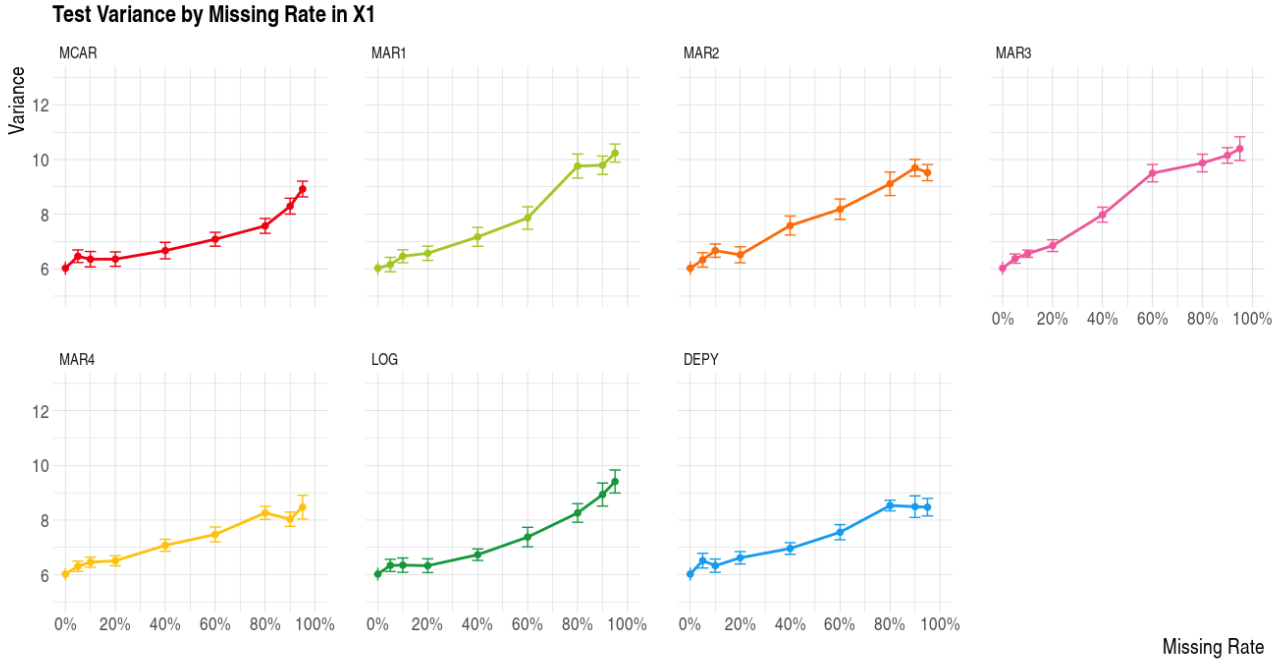

Figure 6.19: Variance of the test data set, varying the percentage of missing values in $\mathbf{X}^{(1)}$, missing values percentage for $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$ are set to 10% and 20%, respectively.

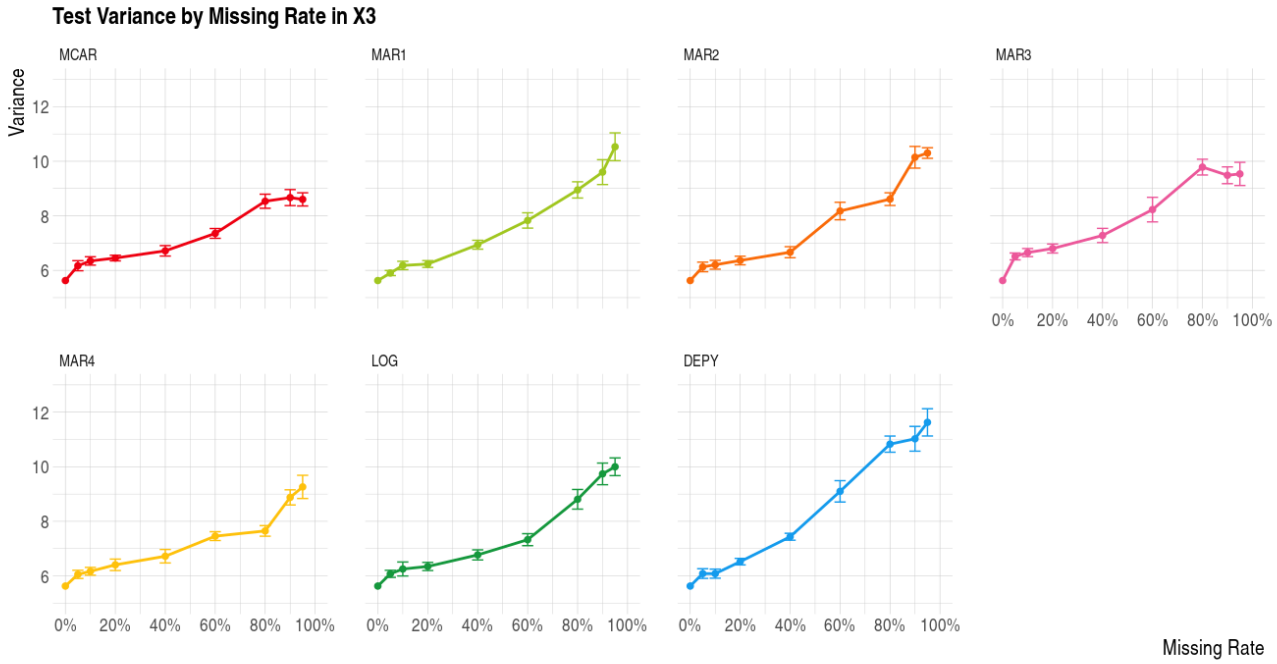

Figure 6.20: Variance of the test data set, varying the percentage of missing values in $\mathbf{X}^{(3)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(4)}$ are set to 20% and 20%, respectively.

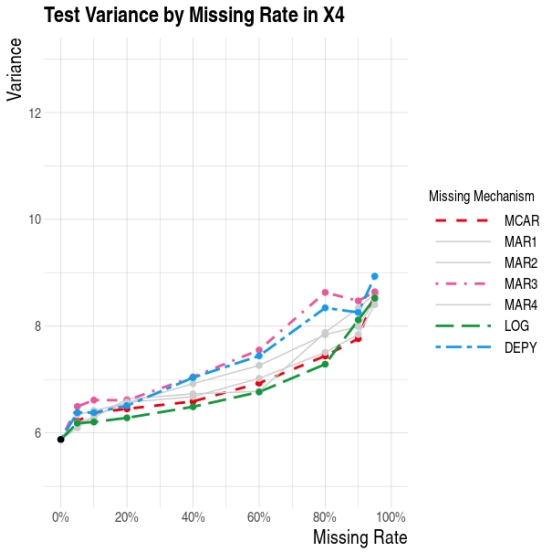

Figure 6.21: Variance of the test data set, varying the percentage of missing values in $\mathbf{X}^{(4)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$ are set to 20% and 10%, respectively.

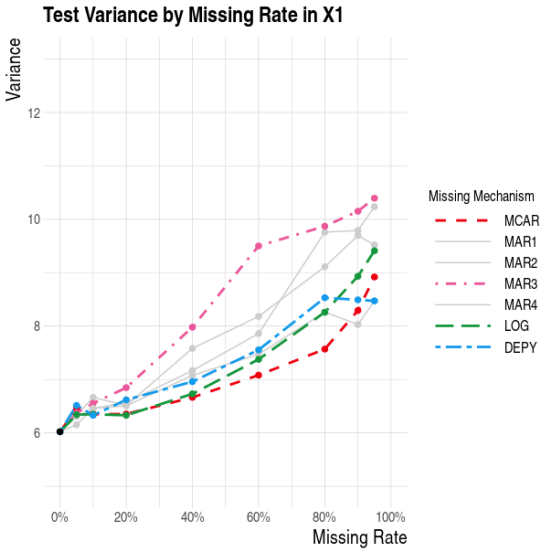

Figure 6.22: Variance of the test data set, varying the percentage of missing values in $\mathbf{X}^{(1)}$, missing values percentage for $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$ are set to 10% and 20%, respectively.

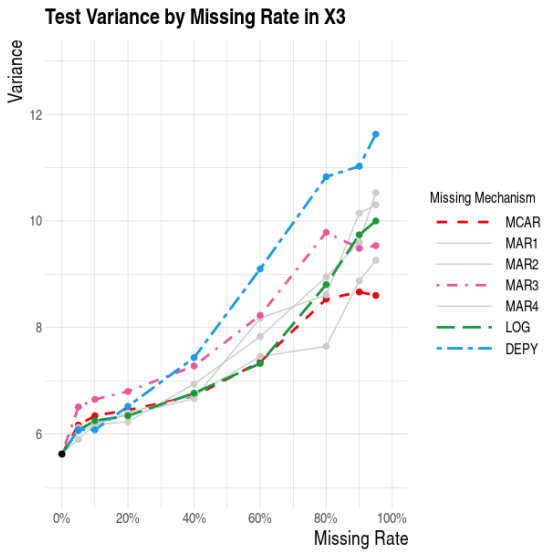

Figure 6.23: Variance of the test data set, varying the percentage of missing values in $\mathbf{X}^{(3)}$, missing values percentage for $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(4)}$ are set to 20% and 20%, respectively.

Bibliography

Biau, Gérard and Erwan Scornet (2016). "A random forest guided tour". In: *Test* 25.2, pp. 197–227.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

— (2003). *Setting up, using, and understanding random forests V4.0*. URL: `https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf`.

Breiman, Leo and Adele Cutler. *Random Forests*. URL: `https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#missing1`.

Breiman, Leo et al. (1984). "Classification and regression trees. Wadsworth & Brooks". In: *Cole Statistics/Probability Series*.

Cairo, Alberto (2016). *The truthful art: Data, charts, and maps for communication*. New Riders.

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*, pp. 1189–1232.

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

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2009). *The elements of statistical learning*. 2nd ed. Springer series in statistics New York.

Gini, Corrado (1912). "Variabilità e mutabilità". In: *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*.

Hintze, Jerry L and Ray D Nelson (1998). "Violin plots: a box plot-density trace synergism". In: *The American Statistician* 52.2, pp. 181–184.

Ishioka, Tsunenori (2013). "Imputation of missing values for unsupervised data using the proximity in random forests". In: *International Conference on Mobile, Hybrid, and On-line Learning. Nice*, pp. 30–36.

Josse, Julie et al. (2019). "On the consistency of supervised learning with missing values". In: *arXiv preprint arXiv:1902.06931*.

Liaw, Andy, Matthew Wiener, et al. (2002). "Classification and regression by randomForest". In: *R news* 2.3, pp. 18–22.

Louppe, Gilles (2014). "Understanding random forests: From theory to practice". In: *arXiv preprint arXiv:1407.7502*.

Quinlan, J. Ross (1986). "Induction of decision trees". In: *Machine learning* 1.1, pp. 81–106.

— (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rieger, Anna, Torsten Hothorn, and Carolin Strobl (2010). "Random forests with missing values in the covariates". In:

Ripley, Brian D (2007). *Pattern recognition and neural networks*. Cambridge university press.

Rubin, Donald B (1976). "Inference and missing data". In: *Biometrika* 63.3, pp. 581–592.

Schapire, RE and Y Freund (1995). "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Second European Conference on Computational Learning Theory*, pp. 23–37.

Scornet, Erwan, Gérard Biau, and J Vert (2015). "Supplement to "Consistency of random forests."" In: *Annals of Statistics* 43.4.

Shannon, Claude Elwood (1948). "A mathematical theory of communication". In: *Bell system technical journal* 27.3, pp. 379–423.

Stekhoven, Daniel J and Peter Bühlmann (2011). "MissForest—non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1, pp. 112–118.

Tukey, John W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Twala, BETH, MC Jones, and David J Hand (2008). "Good methods for coping with missing data in decision trees". In: *Pattern Recognition Letters* 29.7, pp. 950–956.

Venables and Ripley (2002). "Modern Applied Statistics with S". In: *Springer, New York* 1228, p. 1229.

Appendix A

From Empirical to Theoretical Versions

$$\begin{aligned}\mathbb{E}[Y|\mathbf{X}\in A] &= \int Y\mathbb{P}[Y|\mathbf{X}\in A]dY \\ &= \int Y\frac{\mathbb{P}[Y,\mathbf{X}\in A]}{\mathbb{P}[\mathbf{X}\in A]}dY \\ &= \frac{\int\int Y\mathbb{1}_{\mathbf{X}\in A}\mathbb{P}[Y,\mathbf{X}]d\mathbf{X}dY}{\mathbb{P}[\mathbf{X}\in A]} \\ &= \frac{\mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A}]}{\mathbb{P}[\mathbf{X}\in A]}\end{aligned}$$

That is,

$$\mathbb{E}[Y|\mathbf{X}\in A] = \frac{\mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A}]}{\mathbb{P}[\mathbf{X}\in A]}$$

On the other hand, by law of large numbers,

$$\begin{aligned}\mathbb{E}_{F_n}[Y\mathbb{1}_{\mathbf{X}\in A}] = \frac{1}{n}\sum_{i=1}^{n} Y_i\mathbb{1}_{\mathbf{X}_i\in A} &\xrightarrow[n\to\infty]{a.s.}\mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A}] \\ &=\mathbb{E}[Y|\mathbf{X}\in A]\mathbb{P}[\mathbf{X}\in A] \qquad (A.1)\end{aligned}$$