Statistical Inference in the Presence of Imputed Survey Data Through Regression Trees and Random Forests

Mehdi Dagdoug[(a)], Camelia Goga[(b)] and David Haziza[(c)]

(a) McGill University, Department of Mathematics and Statistics, Montréal, CANADA

(b) Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, Besançon, FRANCE

(c) University of Ottawa, Department of Mathematics and Statistics, Ottawa, CANADA

Address for correspondence: David Haziza, Department of mathematics and statistics, University of Ottawa, Ottawa, Canada. Email: dhaziza@uottawa.ca

Abstract

Item nonresponse in surveys is usually handled through some form of imputation. In recent years, imputation through machine learning procedures has attracted a lot of attention in national statistical offices. However, little is known of the theoretical properties of the resulting point estimators. In this paper, we study regression trees and random forests that provide flexible tools for obtaining a set of imputed values. Allowing the number of predictors to diverge, we lay out a set of conditions for establishing the mean square consistency of imputed estimators of a finite population mean. We propose a novel variance estimator based on a K-fold cross-validation procedure. The proposed point and variance estimation are assessed through a simulation study in terms of bias, efficiency, and coverage rate of normal-based confidence intervals. Finally, the choice of hyper-parameters in random forest algorithms is investigated through a mix of theoretical and empirical work.

Key words: Imputation; Item nonresponse; Missing data; Regression trees; Random forest; Variance

estimation; Cross-validation.

1 Introduction

Since the seminal paper of Breiman (2001), random forests have been used in a variety of applications including medicine (Fraiwan et al., 2012), time series analysis (Kane et al., 2014), agriculture (Grimm et al., 2008), missing data (Stekhoven and Buhlmann, 2011), genomics (Qi, 2012) and pattern recognition (Rogez et al., 2008). Random forests constitute a class of ensemble models based on B trees. Predictions through random forests are obtained by averaging the predictions obtained from each of the B trees of a forest. Unlike a number of nonparametric statistical procedures (e.g., kernel predictors, k-nearest neighbors, splines), random forests perform relatively well in a high-dimensional setting; see e.g., Hamza and Larocque (2005) and Díaz-Uriarte and de Andrés (2006). Some recent theoretical investigations (Biau, 2012; Scornet et al., 2015; Klusowski and Tian, 2022) also suggest that random forests adapt well to sparse situations.

In surveys, the problem of missing data is ubiquitous. Estimators of population means based on complete cases only tend to exhibit large biases when the proportion of missing data is appreciable, and the behavior of the responding units is different from that of the nonresponding units. In this article, we consider the problem of item nonresponse, a term used to describe the absence of information on some, but not all, survey variables for a sample unit. The missing values are imputed using a plausible value constructed on the basis of auxiliary variables available for both respondents and nonrespondents. Many imputation procedures have been developed, all sharing a common objective: reduce the potential nonresponse bias to the best possible extent. The reader is referred to Haziza (2009) and Chen and Haziza (2019) for comprehensive discussions of imputation procedures in survey sampling. Every imputation procedure relies on some implicit or explicit assumptions about the distribution of the survey variable requiring imputation. This set of assumptions is called an imputation model.

Tree-based methods such as random forests may prove useful for obtaining a set of imputed values. Because they are nonparametric in nature, random forests tend to be robust

against model misspecification. Also, with the emergence of large data sets in National Statistical Offices (NSO), random forests have attracted much attention in recent years and are currently being scrutinized as an alternative to traditional imputation procedures. However, there remain some important gaps in the literature that we aim to fill in this paper. First, little is known about the theoretical properties (e.g., consistency) of imputed estimators based on random forests for missing survey data. Also, to the best of our knowledge, how to estimate the variance of imputed estimators obtained through random forests while accounting for the sampling and nonresponse variances, has not been discussed in the literature. This is an important issue as NSOs publish point estimates as well as corresponding estimated coefficients of variation, defined as the ratio of the estimated standard error of the estimate to the point estimate. Treating imputed values as observed values and applying a complete data variance estimation procedure will typically result in serious underestimation of the true variance of imputed estimators. The resulting estimated coefficients of variation will thus be too small and the confidence intervals too narrow. As a result, inferences may be misleading. In this paper, we propose a novel variance estimator that is shown to perform well in a wide variety of settings.

The outline of the paper is as follows. In Section 2, we define the framework and introduce some notation. In Section 3, we provide an analysis of tree imputed estimators. The mean square consistency of the tree imputed estimator based on the CART algorithm (Breiman, 1984) is established. In Section 4, we focus on random forest imputed estimators. We begin by establishing the connection between tree imputed estimators, and random forests imputed estimators. As such, random forest estimators inherit a number of the properties of tree estimators. The mean square consistency of forest imputed estimators based on uniform random forests (Biau et al., 2008; Scornet, 2016) and Breiman's original algorithm (Breiman, 2001) is established. In Section 5, using the reverse approach of Shao and Steel (1999), we examine the problem of variance estimation. We investigate the properties of the proposed methods through a simulation study in Section 6. Before concluding in Section 8, the choice of some important hyper-parameters is studied, both theoretically and empirically, in Section 7. All proofs and further technical details are relegated to the Appendix.

2 The setup

Consider a finite population $U = \{1, 2, ..., N\}$ of size N. We are interested in estimating the finite population mean

$$\mu := \frac{1}{N} \sum_{k \in U} y_k,$$

of a survey variable Y. We select a sample S, of size n, according to a sampling design $\mathcal{P}$ with first-order inclusion probabilities $\{\pi_k\}_{k \in U}$ and second-order inclusion probabilities $\{\pi_{k\ell}\}_{k \neq \ell \in U}$. The sample S is completely characterized by the vector of sample selection indicators $\boldsymbol{I} = (I_1, \ldots, I_k, \ldots, I_N)^\top$, where $I_k = 1$ if $k \in S$ and $I_k = 0$ otherwise.

In the ideal case of 100% response, an estimator of μ is the Horvitz-Thompson estimator

$$\widehat{\mu}_\pi := \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}. \tag{1}$$

Provided that $\pi_k > 0$ for all $k \in U$, the estimator (1) is design-unbiased for μ.

In practice, the Y variable may be prone to missingness. Let $\boldsymbol{r} = (r_1, \ldots, r_k, \ldots, r_N)^\top$ denote the vector of response indicators such that $r_k = 1$ if y_k is observed, and $r_k = 0$ otherwise. Let $S_r := \{k \in S;\ r_k = 1\}$ be the set of respondents, of size n_r, and $S_m := \{k \in S;\ r_k = 0\}$ be the set of nonrespondents, of size n_m. We have $S_r \cup S_m = S$ and $n_r + n_m = n$. Let $\mathbf{x}_k = (x_{1k}, x_{2k}, ..., x_{pk})^\top$ be a vector of auxiliary variables attached to unit k and $\boldsymbol{X}$ denote the corresponding matrix. We assume that the $\mathbf{x}$-vector is observed for all $k \in S$. Available to the imputer are the data

$$D_{imp} := \Big\{ (\mathbf{x}_k, y_k)\,; k \in S_r \Big\} \bigcup \Big\{ \mathbf{x}_k; k \in S_m \Big\}.$$

We assume that the vectors $(r_1, y_1, \mathbf{x}_1)^\top, \ldots, (r_n, y_n, \mathbf{x}_n)^\top$ are independent and identically distributed. Further, we assume that (i) the data are missing are random (Rubin, 1976), i.e., $\mathbb{P}(r_k = 1 | y_k, \mathbf{x}_k,) = \mathbb{P}(r_k = 1 | \mathbf{x}_k)$; and (ii) the positivity assumption is satisfied; i.e., $\mathbb{P}(r_k = 1 | \mathbf{x}_k) > 0$. We postulate the following imputation model, describing the relationship

between the survey variable Y and the $\mathbf{x}$-vector:

$$\xi : \qquad y_k = m(\mathbf{x}_k) + \epsilon_k, \tag{2}$$

where $m(\mathbf{x}) := \mathbb{E}[Y|\mathbf{x}]$ denotes the regression function, assumed to be unknown, and the errors $\{\epsilon_k\}_{k\in U}$ are i.i.d. random variables such that $\mathbb{E}[\epsilon_k|\mathbf{x}_k] = 0$ and $\mathbb{V}(\epsilon_k|\mathbf{x}_k) = \sigma^2 < \infty$. In this article, we assume that: i) the regression function m is continuous; ii) the distribution of the covariates $\mathbb{P}_{\mathbf{x}}$ is supported on $Supp(\mathbb{P}_{\mathbf{x}})$, a compact subset the unit cube $[0;1]^p$; iii) the residuals $\{\epsilon_k\}_{k\in U}$ have compact support. Under these assumptions, note that the survey variable Y is bounded, almost surely.

Let $\widehat{m}$ be an estimator of m fitted on $D_r := \{(\mathbf{x}_k, y_k); k \in S_r\}$. An imputed estimator $\widehat{\mu}_{\widehat{m}}$ of μ, based on the imputation procedure $\widehat{m}$, is given by

$$\widehat{\mu}_{\widehat{m}} := \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{m}(\mathbf{x}_k)}{\pi_k}\right), \tag{3}$$

where $\widehat{m}(\mathbf{x}_k)$ denotes the imputed value associated with $k \in S_m$.

To establish the theoretical properties of $\widehat{\mu}_{\widehat{m}}$ in Sections 3 and 4, we consider the asymptotic framework of Isaki and Fuller (1982). Let $\{U_v\}_{v\in\mathbb{N}}$ denote a sequence of embedded finite populations of size $\{N_v\}_{v\in\mathbb{N}}$. In each finite population U_v, a sample S_v, of size n_v, is selected according to a sampling design $\mathcal{P}_v$ with inclusion probabilities $\pi_{k,v}$ and $\pi_{k\ell,v}$. While the finite populations are assumed to be embedded, we do not require this property to hold for the samples $\{S_v\}_{v\in\mathbb{N}}$. This asymptotic framework assumes that, as v goes to infinity, both the population size N_v and the sample size n_v increase to infinity. To improve readability, we use the subscript v only in the quantities U_v, N_v and n_v; quantities such as $\pi_{k,v}$ and $\pi_{k\ell,v}$ will simply be denoted by π_k and $\pi_{k\ell}$, respectively.

For an estimator $\widehat{\mu}$ of μ, we define its mean squared error (MSE) as

$$\mathrm{MSE}(\widehat{\mu}) := \mathbb{E}\left[(\widehat{\mu} - \mu)^2\right]. \tag{4}$$

The expectation in (4) is evaluated with respect to the joint distribution induced by the

imputation model, the sampling design and the nonresponse mechanism. A sequence of imputed estimators $\{\widehat{\mu}_{\widehat{m}_v}\}_{v\in\mathbb{N}}$ is mean square consistent for $\{\mu_v\}_{v\in\mathbb{N}}$ if

$$\lim_{v\to\infty} \text{MSE}\left(\widehat{\mu}_{\widehat{m}_v}\right) = \lim_{v\to\infty} \mathbb{E}\left[\left(\widehat{\mu}_{\widehat{m}_v} - \mu_v\right)^2\right] = 0. \tag{5}$$

We start by describing a set of regularity conditions under which the sequence of (complete data) Horvitz-Thompson estimators $\{\widehat{\mu}_{\pi,v}\}_{v\in\mathbb{N}}$ is mean square consistent.

(H1) We assume that the sampling design $\{\mathcal{P}_v\}_{v\in\mathbb{N}}$ is non-informative sampling and that

a) The sampling fraction is such that $\lim_{v\to\infty} \frac{n_v}{N_v} = \pi^* \in (0;1)$.

b) There exists positive constants λ and λ^* such that $\min_{k\in U_v} \pi_k \geqslant \lambda > 0$, $\min_{k,\ell\in U_v} \pi_{k\ell} \geqslant \lambda^* > 0$.

c) The sampling covariances, $Cov(I_k, I_\ell) = \pi_{k\ell} - \pi_k\pi_\ell$ for $k \neq \ell$, are such that $\limsup_{v\to\infty} n_v \max_{k\neq\ell\in U_v} |\pi_{k\ell} - \pi_k\pi_\ell| < \infty$.

Assumption (H1) is commonly used in the literature, see e.g., Robinson and Särndal (1983) and Breidt and Opsomer (2000). It is known to hold for commonly used sampling designs. The non-informativeness assumption means that, conditionally on the auxiliary variables, the sample selection indicators I_k are independent of the survey variable; see, e.g., Pfeffermann and Sverchkov (2009). Part (a) of (H1) requires that the sample sizes $\{n_v\}_{v\in\mathbb{N}}$ increase at the same rate as the population sizes $\{N_v\}_{v\in\mathbb{N}}$. Part (b) requires that both the first and second-order inclusion probabilities to be bounded away from zero, for all sampling designs $\{\mathcal{P}_v\}_{v\in\mathbb{N}}$. Finally, Part (c) states that the sampling covariances decrease to zero with a rate of at least $\mathcal{O}\left(n_v^{-1}\right)$.

3 Regression tree imputation

In this section, we use regression trees to predict the missing values. Predictions, denoted by $\widehat{m}_{tree}$, are obtained as follows:

Step 1: Create a partition $\mathcal{P} = \{A_1, A_2, ..., A_T\}$ of the predictor space based on D_r. The elements of $\mathcal{P}$ are called the terminal nodes.

Step 2: Use the partition $\mathcal{P}$ to make predictions: for a point $\mathbf{x}$, define

$$\widehat{m}_{tree}\left(\mathbf{x}, \mathcal{P}, D_r\right) := \sum_{k \in S_r} \frac{\mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}} \cdot y_k, \tag{6}$$

where $A(\mathbf{x})$ denotes the node of $\mathcal{P}$ containing $\mathbf{x}$.

From (6), the prediction $\widehat{m}_{tree}(\mathbf{x})$ is obtained by averaging the y-values that fall in the same node as point $\mathbf{x}$. The prediction in (6) can also be written as

$$\widehat{m}_{tree}\left(\mathbf{x}, \mathcal{P}, D_r\right) = \sum_{k \in S_r} \widehat{W}_k(\mathbf{x}, \mathcal{P}, D_r) \cdot y_k,$$

where

$$\widehat{W}_k(\mathbf{x}, \mathcal{P}, D_r) := \frac{\mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}}, \qquad k \in S_r. \tag{7}$$

The weights $\{\widehat{W}_k(\mathbf{x}, \mathcal{P}, D_r)\}_{k \in S_r}$ are called the prediction weights of $\widehat{m}_{tree}$ at $\mathbf{x}$. The predictor $\widehat{m}_{tree}$ is called a local averaging predictor (Györfi et al., 2006).

Given sample observations, different partitions of the predictor space may lead to different predictions. A partitioning predictor is fully characterized by both the set of observations D_r and the partition $\mathcal{P}$. Borrowing from the terminology of Devroye et al. (2013), when the partitioning algorithm does not make use of the survey variable Y, we say that the partitioning rule, and, by extension, the partitioning predictor, has the X-property.

We now describe the commonly used CART algorithm of Breiman (1984). Splits are created by a greedy algorithm that splits recursively the predictor space. Let A denote a node containing $\#(A)$ respondents considered for the next split, and let $\mathcal{C}_A$ be the set of possible splits in the node A, which corresponds to the set of all possible pairs $(j, z) = (\text{variable}, \text{position})$. Let

$$\text{mse}(A) := \frac{1}{\#(A)} \sum_{k \in S_r} \mathbb{1}_{\mathbf{x}_k \in A} \left(y_k - \bar{y}_A\right)^2,$$

where $\bar{y}_A$ denotes the average of the y-values of units belonging to the node A. The splitting process is performed by searching for the best split (j^*, z^*), i.e., the split for which the

following criterion is maximized:

$$L(j,z) = \text{mse}(A) - \text{mse}\,(A_L) - \text{mse}\,(A_R)$$

where $A_L = \{k \in A; x_{kj} < z\}$, $A_R = \{k \in A; x_{kj} \geqslant z\}$. This criterion searches for the split which would generate child nodes as homogeneous as possible, in terms of mean squared error. Splits are always performed in the middle of two points. The procedure continues until a stopping criterion is reached. Commonly used stopping criteria include specifying the minimum number of elements (n_0) in the terminal nodes, or the maximum depth (K) of the tree. For more details about trees and partitioning procedures, the reader is referred to Hastie et al. (2011) or Györfi et al. (2006).

An imputed estimator of μ based on regression trees is given by

$$\widehat{\mu}_{tree} := \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{m}_{tree}(\mathbf{x}_k)}{\pi_k}\right). \tag{8}$$

Before establishing the theoretical properties of $\widehat{\mu}_{tree}$ in a general setting, we consider two very simple but rather unrealistic settings: (i) Suppose that $y_k = C$ for all $k \in U$ and $\sum_{k\in S} d_k = N$, which holds, for instance, in the case of simple random sampling without replacement. In this case, $\widehat{\mu}_{tree}$ is a perfect estimator of μ. That is, $\widehat{\mu}_{tree} = \mu$ for all S. As a result, $\text{MSE}(\widehat{\mu}_{tree}) = 0$. (ii) Suppose that the true model is described by $y_k = C + \epsilon_k$,$k \in U$ such that $\mathbb{E}(\epsilon_k) = 0$, $\mathbb{E}(\epsilon_k\epsilon_\ell) = 0$ for $k \neq \ell$, and $\mathbb{V}(\epsilon_k) = \sigma^2$. We further assume that $\sum_{k\in S} d_k = N$ and that the tree predictor has the X-property. In this case, we have $\mathbb{E}\left[\widehat{\mu}_{tree} - \mu\right] = 0$ and $\mathbb{V}\left(\widehat{\mu}_{tree} - \mu \middle| \boldsymbol{X}\right) = \sigma^2 \sum_{k\in S_r} w_k^2 > 0$, where the weights $\{w_k\}_{k\in S_r}$ satisfy $N^{-1}\sum_{k\in S_r} w_k y_k = \widehat{\mu}_{tree}$. In this setting, $\widehat{\mu}_{tree}$ remains unbiased but it is no longer a perfect estimator of t_y.

For more general settings, the exact derivation of the bias or the variance are much more challenging. Result 3.1 below shows that, under some regularity conditions, the tree imputed estimator based on the CART criterion is mean square consistent. We adopt the following additional notations: Let $C^1([0;1]^p, \mathbb{R})$ be the set of differentiable functions defined on $[0;1]^p$ taking values in $\mathbb{R}$ with continuous first derivative and let $||\mathbf{x}||_1 := \sum_{j=1}^{p_v} |x_j|$ be the 1-vector

norm of $\mathbb{R}^{p_v}$. Also, let $\mathcal{A}_v$ be the class of additive functions of p_v variables:

$$\mathcal{A}_v := \Big\{ g(\mathbf{x}) = \sum_{j=1}^{p_v} g_j(x_j),\ g_j \in C^1([0;1],\mathbb{R}),\ j = 1,2,...p_v \Big\}.$$

Observe that, in particular, linear functions of p_v variables belong to $\mathcal{A}_v$. We restrict our attention to sequences of regression functions $\{m_v\}_{v\in\mathcal{A}_v}$ belonging to $\{\mathcal{A}_v\}_{v\in\mathbb{N}}$ with approximate sparsity. More precisely, we define the total variation norm $||\cdot||_{TV}$ for elements g in $C^1([0;1]^p,\mathbb{R})$ by

$$||g||_{TV} := \int_{[0;1]^{p_v}} ||\nabla g(\mathbf{x})||_1 d\mathbf{x},$$

where ∇g denotes the gradient of g. If $g_v \in \mathcal{A}_v$, then

$$||g_v||_{TV} := \sum_{j=1}^{p_v} \int_{[0;1]} |g_j^{'}(x_j)| dx_j$$

with $g^{'}$ denoting the derivative of a function g defined on $\mathbb{R}$. In the case of a linear function g such that $g(\mathbf{x}) := \mathbf{x}^\top \boldsymbol{\beta}$, then $||g||_{TV} = ||\boldsymbol{\beta}_v||_1$. Lastly, for real-valued functions defined on $\mathbb{R}^p$, we denote by $||g||_\infty := \sup_{\mathbf{x}\in\mathbb{R}^p} |g(\mathbf{x})|$ the sup-norm.

Result 3.1. *We assume that (H1) holds. Consider a sequence of tree imputed estimators $\{\widehat{\mu}_{tree,v}\}_{v\in\mathbb{N}}$ based on the CART criterion with maximal depths $\{K_v\}_{v\in\mathbb{N}}$. We further assume that*

1. *A node A at depth Q of a tree $\widehat{m}_{tree,v}$ is a terminal node if*
   - *i) the depth Q is equal to the maximal depth K_v,*
   - *ii) there is only one (respondent) element in A.*
2. *The sequence of regression functions $\{m_v\}_{v\in\mathbb{N}}$ satisfy $m_v \in \mathcal{A}_v$ and $\sup_{v\in\mathbb{N}} ||m_v||_\infty < \infty$.*

Moreover, assume that

$$\lim_{v\to\infty} K_v = +\infty, \qquad \lim_{v\to\infty} \frac{||m_v||_{TV}}{K_v} = 0, \qquad \textit{and } \lim_{v\to\infty} \frac{2^{K_v} \log^2(n_{r,v}) \log(n_{r,v} p_v)}{n_{r,v}} = 0.$$

Then, the sequence of tree imputed estimators $\{\widehat{\mu}_{tree,v}\}_{v\in\mathbb{N}}$ is mean-square consistent:

$$\lim_{v\to\infty} \mathrm{MSE}\left(\widehat{\mu}_{tree,v}\right) = 0.$$

The above conditions on the tree predictor states that the depth of the trees should increase as the sample and population sizes increase, but not too fast with respect to the number of respondents. Similarly, the relative sparsity of the models $\{m_v\}_{v\in\mathbb{N}}$ is controlled by the depth of the tree. These additional conditions are similar to those in Klusowski and Tian (2022); see also their article for a more thorough discussion about these conditions.

4 From trees to forests

4.1 Randomized predictors and random forests

In this section, we turn our attention to random forests. Using a deterministic partitioning rule as described in Section 4, would lead to B identical trees. We call such a forest a degenerate forest. To circumvent this problem and increase the diversity of models, Breiman (1996, 2001) suggested to introduce additional sources of randomness in the partitioning algorithm. The concept can be formalized through stochastic predictors. Let Θ be a random variable defined in a measurable space $(J, \mathcal{J})$. A stochastic predictor is a measurable function of the type $\widetilde{m} : \mathbb{R}^p \times J \to \mathbb{R}$. As a result, the prediction method $\widetilde{m}$ is random with respect to Θ.

Example 4.1. *Let $q \in]0;1[$ and Θ be a Bernoulli random variable with probability q. Define $\widetilde{m}(\mathbf{x}, \Theta) := \Theta||\mathbf{x}||_2$, where $||\cdot||_2$ denotes the Euclidean norm. Then, $\widetilde{m}$ is a stochastic predictor: for two different realizations of Θ, say $\Theta(j_1)$ and $\Theta(j_2)$ for $j_1, j_2 \in J$, the predictions $\widetilde{m}(\mathbf{x}, \Theta(j_1))$ and $\widetilde{m}(\mathbf{x}, \Theta(j_2))$ will likely be different. This leads to an additional source of variance: $\mathbb{V}_\Theta(\widetilde{m}(\mathbf{x}, \Theta)) = q(1-q)||\mathbf{x}||_2^2 > 0$ for $\mathbf{x} \neq \mathbf{0}_{\mathbb{R}^p}$.*

The stochastic predictor in Example 4.1 is only used for illustration purposes. Next, we give two more realistic examples of how randomization can be incorporated in the construction of regression trees.

Example 4.2. *Uniform random forest (Biau et al., 2008; Scornet, 2016).*
All the trees of a uniform forest are constructed as follows. Let $[0;1]^p$ be the initial leaf and L be the depth of the tree, specified by the user. The partition of the predictor space is created as follows:

1. *For levels $l = 1, \ldots, L$, do:*

(a) For each node G at depth l, do:

i. A splitting variable X_j is selected uniformly at random among the p predictors $X_1, X_2, \ldots, X_p$.

ii. A split is performed in the node G along X_j at a location chosen uniformly at random.

2. Output the partition.

Example 4.3. *Breiman's original algorithm.*
The algorithm proceeds as follows:

Step 1: Select B bootstrap samples from S_r, denoted by $\{S_r(\Theta_b)\}_{b=1}^{B}$.

Step 2: On each bootstrap sample $S_r(\Theta_b)$, fit a tree $\widehat{m}(\cdot, \Theta_b)$ using the CART algorithm on $D_{n_r}(\Theta_b)$. At each split, the CART criterion is optimized on only p_0 predictors chosen uniformly at random (without replacement) among the p available predictors.

Uniform random forests are mostly studied in the literature because the partitions of the trees are independent of the observed data, thus making their theoretical analysis simpler. However, because they do not use the data for building the partitions, they are of little practical interest. In practice, Breiman's original algorithm is typically used, but establishing its theoretical properties is much more challenging. In recent years, some important theoretical developments have been made by Scornet et al. (2015), who studied the mean square consistency of random forest predictors using results of Nobel (1996), assuming a fixed number of predictors. More recently, several authors have studied the high-dimensional mean square consistency of random forests, see e.g., Klusowski and Tian (2022); Chi et al. (2022). Random forests have also been studied through the theory of U-statistics, e.g., Mentch and Hooker (2016); Zhou et al. (2019); Xu et al. (2022).

Generally speaking, random forest predictions can be obtained as follows. Let $\{\Theta_b\}_{b=1}^{B}$ denote a sequence of i.i.d. random variables distributed according to some generic random variable Θ, and assumed to be independent of the observed data. Let $\{\widehat{m}_{tree}(\cdot, \Theta_b)\}_{b=1}^{B}$ be a sequence of randomized tree predictors. Then, the random forest prediction at $\mathbf{x}$ is given

by

$$\widehat{m}_{rf}\left(\mathbf{x}, \{\Theta_b\}_{b=1}^{B}\right) := \frac{1}{B}\sum_{b=1}^{B}\widehat{m}_{tree}\left(\mathbf{x}, \Theta_b\right) = \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in S_r(\Theta_b)}\frac{\mathbb{1}_{\mathbf{x}_k\in A(\mathbf{x},\Theta_b)}}{\sum_{\ell\in S_r(\Theta_b)}\mathbb{1}_{\mathbf{x}_\ell\in A(\mathbf{x},\Theta_b)}}\cdot y_k, \quad (9)$$

where $S_r(\Theta_b)$ denotes the subset of S_r selected for training the b-th tree. Note that $S_r(\Theta_b) = S_r$ in the absence of resampling mechanism in the forest algorithm. The predictions can be also be expressed as

$$\widehat{m}_{rf}\left(\mathbf{x}, \{\Theta_b\}_{b=1}^{B}\right) = \sum_{k\in S_r}\widehat{W}_k^{(B)}(\mathbf{x}, \{\Theta_b\}_{b=1}^{B})\cdot y_k,$$

where

$$\widehat{W}_k^{(B)}(\mathbf{x}, \{\Theta_b\}_{b=1}^{B}) = \frac{1}{B}\sum_{b=1}^{B}\frac{\psi_k^{(b)}\mathbb{1}_{\mathbf{x}_k\in A(\mathbf{x},\Theta_b)}}{\sum_{\ell\in S_r}\psi_\ell^{(b)}\mathbb{1}_{\mathbf{x}_\ell\in A(\mathbf{x},\Theta_b)}}, \quad (10)$$

with $\psi_k^{(b)} = 1$ if $k \in S_r(\Theta_b)$ and $\psi_k^{(b)} = 0$, otherwise. For any subset $D \subseteq U$, we write $N\left(\mathbf{x}_k, D\right) := \sum_{\ell\in D}\mathbb{1}_{\mathbf{x}_\ell\in\widehat{A}(\mathbf{x}_k)}$ to denote the number of elements of the set D belonging to the node of element k. Finally, we write $\widehat{N}\left(\mathbf{x}_k, D\right) := \sum_{\ell\in D} I_\ell\mathbb{1}_{\mathbf{x}_\ell\in\widehat{A}(\mathbf{x}_k)}\pi_\ell^{-1}$ to denote its Horvitz–Thompson estimator. For more details about random forests and their implementation, the reader is referred to Biau and Scornet (2016) and Genuer and Poggi (2019).

4.2 Random forest imputation

Let $\widehat{m}_{rf}^{(B)}$ be a random forest predictor built on B trees, with arbitrary partitioning rule and randomization procedure. An imputed estimator of μ based on random forests and denoted by $\widehat{\mu}_{rf}^{(B)}$, is defined as

$$\widehat{\mu}_{rf}^{(B)} := \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_k)}{\pi_k}\right). \quad (11)$$

We begin our analysis of $\widehat{\mu}_{rf}^{(B)}$ by establishing the link between forest estimators and tree estimators.

Proposition 4.1. *The forest imputed estimator $\widehat{\mu}_{rf}^{(B)}$ defined in (11) can be expressed as the average of (randomized) tree imputed estimators:*

$$\widehat{\mu}_{rf}^{(B)} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_{tree}^{(b)},$$

where $\widehat{\mu}_{tree}^{(b)}$ *is the imputed estimator of* μ *based on the bth tree of the forest* $\widehat{m}_{tree}^{(b)}$*:*

$$\widehat{\mu}_{tree}^{(b)} = \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)}{\pi_k}\right).$$

Therefore, many of the properties of tree imputed estimators are also shared by random forest imputed estimators.

Proposition 4.2. *Consider a random forest estimator* $\widehat{\mu}_{rf}^{(B)}$ *with an arbitrary number of trees* $B \in \mathbb{N}^*$*. Then, for any* $b \in \{1, ..., B\}$*, we have*

$$\text{MSE}\left(\widehat{\mu}_{rf}^{(B)}\right) \leqslant MSE\left(\widehat{\mu}_{tree}^{(b)}\right),$$

with equality if and only if either $B = 1$ *or the forest is degenerate.*

4.3 From finite to infinite forests

In this section, we introduce the notion of *infinite* random forest predictor, defined by

$$\widehat{m}_{rf}^{(\infty)} := \mathbb{E}\left[\widehat{m}_{rf}^{(B)}|\boldsymbol{X}, \boldsymbol{I}, \boldsymbol{r}, \boldsymbol{y}\right].$$

In practice, $\widehat{m}_{rf}^{(\infty)}$ cannot be computed but it will prove useful for establishing the theoretical properties of point estimators and for deriving variance estimators. It is called an infinite forest predictor because, by the strong law of large numbers, we have

$$\lim_{B\to\infty}\widehat{m}_{rf}^{(B)} = \lim_{B\to\infty}\frac{1}{B}\sum_{b=1}^{B}\widehat{m}_{tree}^{(b)} = \widehat{m}_{rf}^{(\infty)},$$

where the limit is taken in the almost sure sense. Accordingly, define the infinite forest imputed estimator as

$$\widehat{\mu}_{rf}^{(\infty)} := \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k}\right). \tag{12}$$

Since an imputed forest estimator can be expressed as an average of tree imputed estimators (see Proposition 4.1), it follows from the strong law of large numbers that

$$\lim_{B\to\infty} \widehat{\mu}_{rf}^{(B)} \overset{a.s.}{=} \mathbb{E}\left[\widehat{\mu}_{rf}^{(B)} | \boldsymbol{X}, \boldsymbol{I}, \boldsymbol{r}, \boldsymbol{y}\right] = \widehat{\mu}_{rf}^{(\infty)}. \tag{13}$$

Even though the infinite forest imputed estimator cannot be computed, there is hope to approach it with a finite forest imputed estimator based on a large number of trees B.

Lemma 1. *Consider sequences of finite* $\{\widehat{\mu}_{rf,v}^{(B)}\}_{v\in\mathbb{N}}$ *and infinite* $\{\widehat{\mu}_{rf,v}^{(\infty)}\}_{v\in\mathbb{N}}$ *forest imputed estimators. There exists* $C > 0$ *such that*

$$0 \leqslant \text{MSE}\left(\widehat{\mu}_{rf,v}^{(B)}\right) - \text{MSE}\left(\widehat{\mu}_{rf,v}^{(\infty)}\right) \leqslant \frac{C}{B}.$$

Moreover, the following asymptotic equivalence holds:

$$\sqrt{n_v}\ \left(\widehat{\mu}_{rf,v}^{(B)} - \mu_v\right) = \sqrt{n_v}\ \left(\widehat{\mu}_{rf,v}^{(\infty)} - \mu_v\right) + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{n_v}{B}}\right).$$

Corollary 4.1. *Consider a finite random forest imputed estimator* $\widehat{\mu}_{rf}^{(B)}$ *with an arbitrary number of trees* $B \in \mathbb{N}^*$ *and an infinite random forest estimator* $\widehat{\mu}_{rf}^{(\infty)}$. *Then, for any* $b \in \{1, ..., B\}$, *we have*

$$\text{MSE}\left(\widehat{\mu}_{rf}^{(\infty)}\right) \leqslant \text{MSE}\left(\widehat{\mu}_{rf}^{(B)}\right) \leqslant \text{MSE}\left(\widehat{\mu}_{tree}^{(b)}\right).$$

Corollary 4.1 suggests that the mean squared error of infinite forests is, at most, equal to the mean squared error of finite forests. It follows that infinite forests are more efficient than finite forests. Lemma 1 also suggests that the difference between the two errors is bounded, and even decreases to 0 if B diverges.

4.4 Consistency of random forest imputed estimators

We begin by considering the case of uniform random forests described in Example 4.2. An important part of our proof is based on the idea that the forests that we consider are, in some sense, large and stable: we assume that, without any rate requirement, the number of trees is strictly increasing: that is, for $v_1 < v_2$ positive integers, the number of trees B_{v_1}

used to obtain $\widehat{m}_{rf}^{(B_{v_1})}$ used to impute in S_{v_1} is strictly smaller than the number of trees B_{v_2} used to obtain $\widehat{m}_{rf}^{(B_{v_2})}$ of S_{v_2}; that is, $v_1 < v_2$ implies $B_{v_1} < B_{v_2}$. This construction implies that $\lim_{v\to\infty} B_v = +\infty$. The motivation for this requirement is that we will first prove the mean square consistency of infinite forest estimators; we then conclude by remembering that if the number of trees $\{B_v\}_{v\in\mathbb{N}}$ increases to infinity, the mean-square errors of both finite and infinite forests converge to the same quantity. It is enough that $\{B_v\}_{v\in\mathbb{N}}$ increases to infinity for finite and infinite forest to share the same (mean-square and probability) limit. However, to ensure that they share the same asymptotic variance, we require that $\{B_v\}_{v\in\mathbb{N}}$ and $\{n_v\}_{v\in\mathbb{N}}$ satisfy $\lim_{v\to\infty} B_v/n_v = 0$.

Result 4.1. *Suppose that (H1) holds. Consider a sequence of uniform forest imputed estimators* $\{\widehat{\mu}_{urf,v}^{(B)}\}_{v\in\mathbb{N}}$ *described in Example 4.2. We also assume that*

(1) The sequence of regression functions $\{m_v\}_{v\in\mathbb{N}}$ *satisfies* $\sup_{v\in\mathbb{N}}||m_v||_\infty < \infty$.

(2) The number of steps $\{L_v\}_{v\in\mathbb{N}}$ *increases as* v *increases such that*

$$(a)\ \lim_{v\to\infty} p_v\left(1-\frac{1}{4p_v}\right)^{L_v} = 0, \qquad \text{and} \qquad (b)\ \lim_{v\to\infty}\frac{2^{L_v}}{n_v} = 0.$$

(3) The number of trees in the forest increases i.e., $\lim_{v\to\infty} B_v = +\infty$.

Then, the forest estimator $\{\widehat{\mu}_{urf,v}^{(B)}\}_{v\in\mathbb{N}}$ *is mean-square consistent for* μ*, i.e.*

$$\lim_{v\to\infty} \mathrm{MSE}\left(\widehat{\mu}_{urf,v}^{(B)}\right) = 0.$$

The condition given in Part (1) of Result 4.1 follows from sufficient conditions for the mean square consistency in high-dimensional settings of $\{\widehat{m}_{urf,v}^{(B)} - \widehat{m}_v\}_{v\in\mathbb{N}}$ towards 0. If the number of covariates is fixed, this condition reduces to the conditions given in (Scornet, 2016, Corrolary 3.1). Condition (a) of Part (2) ensures the diameters of each node decreases to 0 as v increases. It is satisfied for instance if L_v increases fast enough compared to p_v. Condition (b) is sufficient to ensure that the probability to have an empty leaf converges to 0.

Result 4.2. *Suppose that (H1) holds. Consider a sequence of Breiman's random forest imputed estimators* $\{\widehat{\mu}_{brf,v}^{(B)}\}_{v\in\mathbb{N}}$ *described in Example 4.3. Assume also that:*

1. *A node A at depth Q of a tree $\widehat{m}_{tree,v}$ is a terminal node if*

   i) the depth Q is equal to the maximal depth K_v,

   ii) there is only one (respondent) element in A.

2. *The resampling mechanism is subsampling (without replacement) of size a_v.*

3. *The number of trees in the forest increases i.e., $\lim_{v\to\infty} B_v = +\infty$.*

4. *The sequence of regression functions $\{m_v\}_{v\in\mathbb{N}}$ satisfy $m_v \in \mathcal{A}_v$ and $\sup_{v\in\mathbb{N}}||m_v||_\infty < \infty$.*

Moreover, assume that

$$(a)\ \lim_{v\to\infty} K_v = +\infty, \qquad (b)\ \lim_{v\to\infty} \frac{\sqrt{p_v}||m_v||_{TV}}{\sqrt{p_{0v}}\sqrt{K_v}} = 0, \qquad (c)\ \lim_{v\to\infty} \frac{2^{K_v}\log^2(a_v)\log(a_v\ p_v)}{a_v} = 0.$$

Then, the forest estimator $\{\widehat{\mu}_{brf}^{(B)}\}$ is mean square consistent for μ_y, i.e.,

$$\lim_{v\to\infty} \mathrm{MSE}\left(\widehat{\mu}_{brf}^{(B)}\right) = 0.$$

Result 4.2 extends the consistency of imputed estimators based on CART regression trees to Breiman's random forests.

5 Variance estimation

In this section, we study the problem variance estimation in the context of imputed data through random forests. We start by describing the naive variance estimator, which is obtained by applying a complete data variance estimation procedure to the pseudo-values $\widetilde{y}_k = r_k y_k + (1-r_k)\widehat{m}_{rf}^{(B)}(\mathbf{x}_k)$. This leads to

$$\widehat{V}_{naive} := \frac{1}{N_v^2}\sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}\frac{\widetilde{y}_k}{\pi_k}\frac{\widetilde{y}_\ell}{\pi_\ell}, \tag{14}$$

with $\Delta_{k\ell} := \pi_{k\ell} - \pi_k\pi_\ell$ denoting the sampling covariances of elements k and ℓ of U. In general, the use of (14) may lead to severe underestimation of the total variance of $\widehat{\mu}_{rf}^{(B)}$. This is illustrated empirically in Section 7. To derive variance estimators that account for sampling and nonresponse, we first decompose the total variance of $\widehat{\mu}_{rf}^{(B)}$ as follows.

Proposition 5.1. *Consider sequences of finite* $\{\widehat{\mu}_{rf,v}^{(B)}\}_{v\in\mathbb{N}}$ *and infinite* $\{\widehat{\mu}_{rf,v}^{(\infty)}\}_{v\in\mathbb{N}}$ *forest estimators. We have*

$$\mathbb{V}\left(\widehat{\mu}_{rf}^{(B)} - \mu\right) = \mathbb{V}\left(\widehat{\mu}_{rf}^{(\infty)} - \mu\right) + \mathbb{E}\left[\mathbb{V}_{\Theta}\left(\widehat{\mu}_{rf}^{(B)}\right)\right], \tag{15}$$

where $\mathbb{V}_{\Theta}$ *denote the variance operator with respect to the random variables* $\{\Theta_b\}_{b=1}^{B}$*, conditionally on every other random quantities. Furthermore, there exists* $C > 0$ *such that*

$$\mathbb{E}\left[\mathbb{V}_{\Theta}\left(\sqrt{n_v}\,\widehat{\mu}_{rf}^{(B)}\right)\right] \leqslant \frac{C}{B_v}.$$

It follows from proposition 5.1 that the contribution of $\mathbb{E}\left[\mathbb{V}_{\Theta}\left(\widehat{\mu}_{rf}^{(\infty)}\right)\right]$ to the total variance $\mathbb{V}\left(\left(\widehat{\mu}_{rf}^{(B)}\right)\right)$ is negligible provided that $n_v/B_v = o(1)$. Proposition 5.1 suggests that the contribution of the randomization variance can be made arbitrarily small by choosing a large value of B. In Sections 6.1 and 6.2, we thus omit this term from the computations. The contribution of $\mathbb{E}\left[\mathbb{V}_{\Theta}\left(\widehat{\mu}_{rf}^{(\infty)}\right)\right]$ is assessed empirically in Section 8 for several values of B.

5.1 Variance estimation based on a first-order Taylor expansion

A variance estimator for $\widehat{\mu}_{rf}^{(B)}$ can be obtained through the so-called reverse approach of Fay (1991) and Shao and Steel (1999); see also Kim and Rao (2009) and Haziza and Vallée (2020). This approach leads to the following decomposition of the total variance of $\widehat{\mu}_{rf}^{(B)}$:

$$\begin{aligned}\mathbb{V}\left(\widehat{\mu}_{rf}^{(B)} - \mu_y|\boldsymbol{r}\right) &= \mathbb{E}\left[\mathbb{V}\left(\widehat{\mu}_{rf}^{(B)}|\boldsymbol{r},\boldsymbol{y},\boldsymbol{X}\right)|\boldsymbol{r},\boldsymbol{X}\right] + \mathbb{V}\left[\mathbb{E}\left(\widehat{\mu}_{rf}^{(B)} - \mu_y|\boldsymbol{r},\boldsymbol{y},\boldsymbol{X}\right)|\boldsymbol{r},\boldsymbol{X}\right] \\ &:= V_1 + V_2. \end{aligned} \tag{16}$$

As noted by various authors (Shao and Steel, 1999; Haziza and Vallée, 2020), the contribution of second term on the right hand-side of (16) to the total variance is negligible if the sampling fraction n/N is negligible. In this section, we assume that n/N is negligible, which is commonly encountered in practice.

Using a first-order Taylor expansion, an estimator of V_1 in (16) is given by

$$\widehat{V}_1 := \sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}\frac{\widehat{\xi}_k^{(rf,B)}}{\pi_k}\frac{\widehat{\xi}_\ell^{(rf,B)}}{\pi_\ell}, \tag{17}$$

where

$$\begin{aligned}\widehat{\xi}_k^{(rf,B)} &:= \frac{1}{B}\sum_{b=1}^{B}\widehat{\xi}_k^{(tree,b)} \\ &:= r_k y_k + (1-r_k)\,\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) + \frac{1}{B}\sum_{b=1}^{B}\psi_k^{(b)}\frac{\widehat{N}_b\left(\mathbf{x}_k, S_m\right)}{\widehat{N}_b\left(\mathbf{x}_k, S_r\left(\Theta_b\right)\right)}\cdot\left(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k)\right), \qquad k\in S.\end{aligned} \tag{18}$$

It is worth pointing out that the derivation of $\widehat{V}_1$ was made conditionally on the partition of the predictor space. This is a simplification of the reality as the partitions vary from one sample to another. Unfortunately, the use of $\widehat{V}_1$ may lead to significant underestimation of the total variance, especially for small values of n_0. This is illustrated empirically in Section 7. To overcome this issue, we propose a novel variance estimator based on a K-fold cross-validation procedure in Section 6.2.

Remark 5.1. *For a single deterministic regression tree $\widehat{m}_{tree}$, the linearized variable given by (18) reduces to*

$$\widehat{\xi}_k^{(tree)} = r_k y_k + (1-r_k)\,\widehat{m}_{tree}(\mathbf{x}_k) + \frac{\widehat{N}\left(\mathbf{x}_k, S_m\right)}{\widehat{N}\left(\mathbf{x}_k, S_r\right)}\cdot\left(y_k - \widehat{m}_{tree}(\mathbf{x}_k)\right), \qquad k\in S.$$

5.2 A variance estimator based on a K-fold cross-validation procedure

As mentioned in Section 6.1, the use the use of $\widehat{V}_1$ in (17) may lead to significant underestimation of the total variance, especially for small values of n_0. For the most part, this is due to the problem of overfitting. Indeed, the residuals, $y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k)$, in (18) can be made artificially small for small values of n_0. A similar issue was described by Dagdoug et al. (2023) in the context of model-assisted estimation.

To circumvent this problem, we suggest replacing the residuals, $y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k)$, by new residuals obtained by using a K-fold cross-validation procedure. The reader is referred to Arlot and Celisse (2010) for a general review of cross-validation and to Dagdoug et al. (2023)

for a discussion in a finite population setting. Let $\{\widehat{\epsilon}_k^{(cv,b)}; k \in S_r,\ b \in \{1,\ldots,B\}\}$ denote a set of tree residuals obtained through K-fold cross-validation, for $k \in \{2,\ldots,n_r\}$. An alternative variance estimator of V_1 is given by

$$\widehat{V}_1^{(cv)} := \sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}\frac{\widehat{\xi}_k^{(cv)}}{\pi_k}\frac{\widehat{\xi}_\ell^{(cv)}}{\pi_\ell}, \tag{19}$$

where

$$\widehat{\xi}_k^{(cv)} := r_k y_k + (1-r_k)\,\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) + \frac{1}{B}\sum_{b=1}^{B}\psi_k^{(b)}\frac{\widehat{N}_b\left(\mathbf{x}_k, S_m\right)}{\widehat{N}_b\left(\mathbf{x}_k, S_r\left(\Theta_b\right)\right)}\cdot\widehat{\epsilon}_k^{(b,cv)}, \qquad k \in S. \tag{20}$$

The residuals $\widehat{\epsilon}_k^{(b,cv)}$ are readily obtained through `R` packages such as `RandomForest` (Liaw and Wiener, 2002).

6 Simulation studies

In this section, we conduct a simulation study to compare the performance of several imputation procedures in terms of bias and efficiency and to compare the performance of several variance estimation procedures in terms of bias and coverage rate of normal-based confidence intervals.

6.1 Point estimation

We repeated $R = 5,000$ iterations of the following process:

(i) A finite population of size $N = 5,000$ was generated. The population consisted of a set of $p = 90$ explanatory variables $X_1,\ldots,X_{90}$, and 5 survey variables $Y_1,\ldots,Y_5$. To generate the X-variables, we considered two scenarios: (i) The explanatory variables were generated independently from a normal distribution with mean equal to 5 and variance equal to 1. (ii) The explanatory variables were generated from a multivariate normal distribution with a mean vector equal to $5\times\mathbf{1}^\top$ and variance-covariance matrix whose diagonal elements were equal to 1 and the off-diagonal elements were equal to 0.7, where $\mathbf{1}$ denotes the vector of ones. That is, the explanatory variables were correlated. Given the values of $X_1,\ldots,X_{90}$, we generated 5 survey variables according to the

following models:

$$
\begin{aligned}
Y_1 &= 2 + X_1 + 3X_2 + 4X_5 + \mathcal{N}(0,5)\,, \\
Y_2 &= 10^{-3} X_1^6 X_2^3 + \mathcal{N}(0,1)\,, \\
Y_3 &= 1.5 + \cos(X_1 + X_2 + X_3 + X_4) + \mathcal{N}\left(0, 10^{-2}\right), \\
Y_4 &= 2 + \mathbb{1}_{\{X_1>7\}} - \mathbb{1}_{\{X_1<4\}} + 2\mathbb{1}_{\{X_4>6\}} + \mathcal{N}(0,1)\,, \\
Y_5 &= 2 + X_1 + 10\exp\left(2\mathbb{1}_{\{X_5>5\}} - \mathbb{1}_{\{X_5<6\}}\right).
\end{aligned}
$$

Note that the survey variables $Y_1, \ldots, Y_5$ were generated using a subset of the first five explanatory variables X_1-X_5.

(ii) From the finite population generated in Step (i), a sample, of size $n = 250$, was selected according to simple random sampling without replacement.

(iii) In each sample, the response indicators r_k, $k \in S$, were independently generated according to a Bernoulli distribution with probability

$$
p_k = \text{logit}\left(0.15 \times \{-30 + X_1 + X_2 + X_3 + X_4 + 2X_5\}\right). \tag{21}
$$

This led to a response rate approximately equal to 50%.

(iv) The missing values in each sample were imputed by five imputation procedures:

(1) Deterministic linear regression imputation;

(2) Regression tree (CART) imputation, with $n_0 = 10$ and a complexity parameter[1] `cp` $= 0.01$. The package `Rpart` (Therneau and Atkinson, 2022) was used for the implementation of CART imputation.

(3) Random forest (RF) imputation with $B = 1000$ trees, $n_0 = 10$ elements in each terminal node and $p_0 = p$. We used the bootstrap as the resampling algorithm. We implemented RF imputation using the package `Ranger` (Wright and Ziegler, 2015).

(4) Nearest neighbour (NN) imputation.

[1]The complexity parameter `cp` is a parameter available in the `Rpart` package, whereby, as per the `Rpart` documentation, "any split that does not decrease the overall lack of fit by a factor of `cp` is not attempted."

(5) K-nearest neighbour (KNN) imputation with $K = 5$. The package `caret` (Kuhn, 2022) was used for the implementation of both NN and KNN.

Because we were interested in understanding the impact of the number of explanatory variables on the behaviour of the resulting imputed estimator, we considered two scenarios: (a) the case where only the first 5 variables were included in each model; b) the case where the 90 variables were included.

(v) In each completed data set and for each imputation procedure, we computed the estimators $\widehat{\mu}_{\widehat{m}}$ given by (3).

As a measure of bias, we used the Monte-Carlo percent relative bias (RB) defined as

$$RB(\widehat{\mu}_{\widehat{m},j}) := 100 \times \frac{1}{R}\sum_{r=1}^{R}\frac{\left(\widehat{\mu}_{\widehat{m},j}^{(r)} - \mu_j^{(r)}\right)}{\mu_j^{(r)}}, \qquad j = 1, 2, \ldots, 5,$$

where $\hat{\mu}_{\widehat{m},j}^{(r)}$ denotes an estimator of μ_j at the rth iteration, $r = 1, \ldots, R$. As a measure of relative efficiency (RE) with respect to the Horvitz-Thompson estimator, we used

$$RE(\hat{\mu}_{\widehat{m},j}) := 100 \times \frac{\sum_{r=1}^{R}\left(\hat{\mu}_{\widehat{m},j}^{(r)} - \mu_j^{(r)}\right)^2}{\sum_{r=1}^{R}\left(\hat{\mu}_{j\pi}^{(r)} - \mu_j^{(r)}\right)^2}, \qquad j = 1, 2, \ldots, 5.$$

The results for $p = 5$ with independent and correlated covariates are given in Table 1 and Table 2, respectively. In both cases, LR was, as expected, the most efficient estimator for the variable Y_1 with a value of RE of about 158% for $p = 5$ independent predictors. RF performed better than the other procedures with a value of RE equal to 186% for $p = 5$ independent predictors and 160% for $p = 5$ correlated predictors. For the survey variables Y_2-Y_5, LR was generally biased, as expected. The biases were larger in the case of correlated predictors. RF, on the other hand, exhibited negligible bias across all scenarios. Its relative efficiency was always either close to that of the best procedure or was the most efficient procedure. The procedures NN, 5NN and CART were also efficient in most scenarios. In some scenarios, NN and 5NN exhibited a slight bias. This is likely due to the curse of dimensionality; e.g., Abadie and Imbens (2006) and Yang and Kim (2019). RF outperformed CART in all the scenarios.

Survey variable	MC measure	Imputed estimators				
		LR	NN	5NN	CART	RF
Y_1	RB	0.0	0.5	0.7	0.6	0.4
	RE	158	222	227	220	186
Y_2	RB	-1.4	-2.8	-3.1	3.4	1.0
	RE	137	123	121	168	124
Y_3	RB	0.4	1.1	1.3	-0.4	-0.1
	RE	213	205	206	217	178
Y_4	RB	-0.1	-0.3	-0.3	0.2	0.1
	RE	197	229	202	187	181
Y_5	RB	-2.1	-2.1	-2.6	1.0	0.0
	RE	164	157	153	105	104

Table 1: Monte Carlo Simulation Results for $p = 5$ and independent covariates.

Survey variable	MC measure	Imputed estimators				
		LR	NN	KNN	CART	RF
Y_1	RB	0.0	0.4	0.8	0.9	0.5
	RE	137	178	185	209	160
Y_2	RB	-19.2	-1.0	-1.1	7.4	1.3
	RE	368	105	103	179	108
Y_3	RB	-0.1	-0.4	-0.8	-0.2	-0.5
	RE	247	188	200	239	196
Y_4	RB	-1.7	0.0	0.4	0.3	0.1
	RE	213	234	202	193	186
Y_5	RB	-10.0	-0.9	-1.0	2.8	0.1
	RE	310	124	123	113	103

Table 2: Monte Carlo Simulation Results for $p = 5$ and correlated covariates.

The results for $p = 90$ with independent and correlated covariates are given in Table 3 and Table 4, respectively. In most scenarios, RF exhibited negligible bias and was the most efficient. For the survey variable Y_1, RF even outperformed LR both with independent and correlated covariates. This is not surprising as the performance of linear regression imputation tends to deteriorate as the dimension of the $\mathbf{x}$-vector increases. Again, NN and 5NN suffered from the curse of dimensionality. This was especially evident in the case

Survey variable	MC measure	Imputed estimators				
		LR	NN	5NN	CART	RF
Y_1	RB	0.0	1.5	1.6	0.8	0.9
	RE	305	588	536	259	263
Y_2	RB	-1.5	2.4	4.0	2.7	3.2
	RE	286	286	256	158	150
Y_3	RB	0.5	-0.9	-1.1	-1.3	-1.7
	RE	485	396	315	259	231
Y_4	RB	-0.1	1.2	1.3	0.4	0.6
	RE	461	386	293	193	177
Y_5	RB	-1.9	8.6	9.4	1.0	0.2
	RE	348	52	8457	104	104

Table 3: Monte Carlo Simulation Results for $p = 90$ and independent covariates.

Survey variable	MC measure	Imputed estimators				
		LR	NN	5NN	CART	RF
Y_1	RB	0.0	0.8	1	1.1	0.9
	RE	249	231	226	237	193
Y_2	RB	-23.7	0.8	0.8	8.5	2.4
	RE	790	112	107	210	111
Y_3	RB	0.0	0.1	-0.4	0.1	-0.1
	RE	615	275	226	268	205
Y_4	RB	-1.6	0.8	1.1	0.6	0.5
	RE	519	293	233	213	190
Y_5	RB	-8.8	2.3	2.4	2.5	0.1
	RE	507	181	155	111	102

Table 4: Monte Carlo Simulation Results for $p = 90$ and correlated covariates.

of independent covariates, where NN and KNN displayed relative biases up to 8.6% and 9.4%, respectively. Comparing the results in Table 1 and Table 3, it is worth mentioning that, unlike the other estimators, the performance of RF was moderately impacted by the dimension of the $\mathbf{x}$-vector. This observation suggests that RF have a tendency to maintain good performance even in high-dimensional settings.

we note that, unlike the other estimators, the performance of RF was moderately affected

by the dimension of the **x**-vector. This suggests that the performance of random forests tends to hold well in a high-dimensional setting.

6.2 Performance of variance estimators

We investigated the performance of both the linearized variance estimator (Section 6.1) and the variance estimator based on a K-fold cross validation procedure (Section 6.2). We considered the same models as the one used for point estimation; see Section 7.1.

As a measure of bias of a variance estimator $\widehat{V}$, we computed its Monte-Carlo percent relative bias (RB) given by

$$RB(\widehat{V}) := 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{\widehat{V}^{(r)} - V_{MC}\left(\widehat{\mu}\right)}{V_{MC}\left(\widehat{\mu}\right)},$$

with $V_{MC}\left(\widehat{\mu}\right)$ denoting the Monte-Carlo variance of $\widehat{\mu}$. We also computed the Monte Carlo coverage rate of 95% normal-based confidence intervals of the form

$$IC_r\left(\widehat{\mu}^{(r)}, \widehat{V}^{(r)}\right) := \left[\widehat{\mu}^{(r)} - 1.96 \times \sqrt{\widehat{V}^{(r)}}\ ;\ \widehat{\mu}^{(r)} + 1.96 \times \sqrt{\widehat{V}^{(r)}}\right].$$

The Monte-Carlo coverage rate is then defined as

$$\text{Coverage}\left(\widehat{\mu}^{(r)}, \widehat{V}^{(r)}\right) := \frac{100}{R} \sum_{r=1}^{R} \mathbb{1}_{\mu \in \left\{IC_r\left(\widehat{\mu}_t^{(r)}, \widehat{V}^{(r)}\right)\right\}}.$$

As in Section 7.1, the sample size n was set to 250, which corresponded to a sampling fraction, n/N, of 5%. This can be viewed as a small sampling fraction.

6.2.1 Variance estimation for regression trees

Results for the case of $p = 5$ independent covariates and correlated covariates are presented in Table 5 and Table 6, respectively. Results for the case of $p = 90$ with independent covariates and correlated covariates are presented in Table 7 and Table 8, respectively.

As expected, the naive variance estimator suffered from large negative biases, leading to substantial undercoverage in most scenarios. The linearized variance estimator $\widehat{V}_1$ given by (17) exhibited noticeable negative bias as well, although not as prominently as the naive

Survey variable	Estimator	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
		RB	Coverage	RB	Coverage	RB	Coverage
Y_1	Naïve	-56.4	77.9	-59.9	74.0	-66.4	65.0
	Linearized	-27.4	88.6	-18.9	89.4	-16.7	86.0
	CV	5.0	94.0	0.1	92.4	-5.8	88.2
Y_2	Naïve	-34.4	88.2	-42.5	86.0	-48.9	82.5
	Linearized	-9.8	92.0	-6.6	93.5	-2.8	94.1
	CV	6.2	93.8	4.5	94.8	4.3	95.0
Y_3	Naïve	-66.0	74.1	-68.8	72.0	-71.6	69.4
	Linearized	-33.0	88.8	-21.5	91.5	-13.1	92.6
	CV	3.6	95.1	0.3	95.0	-1.6	94.2
Y_4	Naïve	-58.9	79.0	-59.9	78.6	-63.4	76.7
	Linearized	-29.5	90.1	-18.7	92.3	-15.6	92.7
	CV	1.7	95.3	-2.4	94.8	-5.2	94.0
Y_5	Naïve	-3.9	94.0	-2.9	94.4	-7.7	93.0
	Linearized	-3.7	94.0	-2.6	94.4	-1.9	94.8
	CV	-0.7	94.4	0.4	94.7	6.6	95.5

Table 5: Monte-Carlo simulation results for tree variance estimators for $p = 5$ and independent covariates.

variance estimator. The bias was especially appreciable for small values of n_0. For instance, for Y_1, the relative bias of the linearized variance estimator ranged between -33.5% and -3.3% for $p = 5$ correlated predictors, whereas the values of the coverage rate lied between 88.4% and 94.1%. This pattern was observed in every other scenario. As mentioned in Section 6.2, this is most likely due to overfitting: small values of n_0 tend to produce artificially small sample residuals, which in turn, produces variance estimates that are too small. It is worth noting that, in the extreme case $n_0 = 1$, the linearized variance estimator $\widehat{V}_1$ given by (17) reduces to the naive variance estimator $\widehat{V}_{naive}$ given by (14), which, as mentioned above, suffers from substantial underestimation. In contrast, the proposed variance estimator based on 10-fold cross-validation procedure, performed well. For instance, for $p = 5$ correlated predictors (See Table 6) and $n_0 = 5$, the values of relative bias ranged from -0.7% to 6.2% in the case of Y_1-Y_5, and the coverage rate ranged from 93.8% to 95.3%. Similar results were

Survey variable	Estimator	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
		RB	Coverage	RB	Coverage	RB	Coverage
Y_1	Naive	-46.3	82.3	-50.0	76.6	-57.0	53.2
	Linearized	-23.0	89.2	-17.5	86.4	-14.5	72.7
	CV	4.5	93.9	-0.2	89.7	-2.1	76.4
Y_2	Naive	-14.3	91.5	-17.6	91.8	-26.3	90.5
	Linearized	-3.6	92.8	0.2	94.2	-0.5	94.7
	CV	3.3	93.6	5.8	94.8	4.0	95.2
Y_3	Naive	-68.4	72.7	-71.5	70.2	-75.3	65.3
	Linearized	-33.5	88.4	-23.6	91.0	-13.8	91.4
	CV	4.7	94.8	1.6	94.8	-2.8	93.3
Y_4	Naive	-60.2	78.5	-61.1	78.0	-62.4	76.5
	Linearized	-30.0	89.9	-21.0	91.8	-14.5	92.5
	CV	1.9	94.8	-3.1	94.4	-5.5	93.8
Y_5	Naive	-3.5	94.0	-3.4	94.1	-1.3	94.5
	Linearized	-3.3	94.0	-3.2	94.1	-0.5	94.8
	CV	-1.6	94.1	-1.7	94.4	2.1	95.4

Table 6: Monte-Carlo simulation results for tree variance estimators for $p = 5$ and correlated covariates.

obtained for $p = 90$ predictors.

6.2.2 Variance estimation for random forests

In this section, we present results for imputation through random forests. Again, we used the same setup as the one described in Section 7.1. The forests were based on 100 trees and the value of p_0 was set to p. We revisit the choice of p_0 in Section 8.3. Results for $p = 5$ correlated predictors are presented in Table 9. Results for $p = 90$ predictors were very similar and are thus omitted. From Table 9, we note that the proposed variance estimator based on a 10-fold cross-validation procedure performed well, especially for $n_0 = 5$ and $n_0 = 10$. In contrast, the naive variance estimator and the linearized variance estimator suffered from substantial bias in most scenarios.

Survey variable	Estimator	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
		RB	Coverage	RB	Coverage	RB	Coverage
Y_1	Naïve	-60.9	71.6	-61.8	70.6	-64.7	64.8
	Linearized	-44.4	80.0	-28.1	85.1	-14.5	85.6
	CV	2.5	91.7	-0.4	90.8	-1.8	88.5
Y_2	Naïve	-37.5	87.2	-41.0	86.0	-46.2	83.4
	Linearized	-17.6	91.4	-7.3	93.4	-0.5	94.5
	CV	2.4	94.7	4.4	94.9	6.9	95.3
Y_3	Naïve	-66.8	72.2	-68.7	70.7	-72.1	66.8
	Linearized	-48.7	82.0	-33.5	87.3	-20.9	89.8
	CV	6.1	94.6	1.9	93.8	-3.9	92.7
Y_4	Naïve	-58.7	78.9	-59.7	78.7	-62.0	77.4
	Linearized	-42.0	86.3	-27.4	90.8	-16.6	92.7
	CV	3.5	95.3	-1.1	94.7	-0.5	94.9
Y_5	Naïve	-4.1	94.1	-4.8	94.1	-6.8	93.6
	Linearized	-4.0	94.1	-4.5	94.2	-4.7	94.3
	CV	-1.4	94.4	-2.2	94.5	0.3	95.1

Table 7: Monte-Carlo simulation results for tree variance estimators for $p = 90$ and independent covariates.

7 Choice of hyper-parameters

Random forests algorithms require the specification of several hyper-parameters. In this section, we discuss the choice of three hyper-parameters: the number of trees B, the number of observations in each terminal node, n_0, and the number, p_0, of predictors randomly selected at each split.

7.1 Choice of B

Selecting the number of trees B to be used is likely the simplest parameter to decide on: the more, the better. Indeed, choosing a large value of B leads to more efficient point estimators of a population mean; see Proposition 5.1 and Corollary **??**. Also, it simplifies the variance estimation process, as the second term on the right hand-side of (15) can be safely omitted from the computations. The contribution of the randomization variance $\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf,}^{(B)}\right)\right]$ to

Survey variable	Estimator	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
		RB	Coverage	RB	Coverage	RB	Coverage
Y_1	Naïve	-50.1	78.3	-50.5	74.3	-57.7	53.9
	Linearized	-36.3	83.6	-24.6	84.3	-20.3	71.8
	CV	3.2	91.8	0.8	89.4	-4.1	76.7
Y_2	Naïve	-19.8	90.6	-23.2	91	-33.1	91
	Linearized	-3.8	91.5	0.3	93.1	-1.7	94.9
	CV	6.1	92.5	6.5	93.7	2.3	95.3
Y_3	Naïve	-70.7	71.7	-72.6	69.2	-75.9	64.8
	Linearized	-50.8	82.9	-37.7	87	-24.3	89.6
	CV	4.2	94.6	-1	94.2	-7.1	92.9
Y_4	Naïve	-61.5	77.6	-61.6	77.3	-63.1	76.2
	Linearized	-44.2	85.9	-31	89.2	-22.3	90.9
	CV	1.9	94.8	-3.1	94.1	-8.8	93.1
Y_5	Naïve	-2.5	94.2	-1.7	94.4	-4	94.1
	Linearized	-2.4	94.2	-1.4	94.4	-3.6	94.2
	CV	-0.6	94.4	0.1	94.7	-1.9	94.7

Table 8: Monte-Carlo simulation results for tree variance estimators for $p = 90$ and correlated covariates.

the total variance $\mathbb{V}\left(\widehat{\mu}_{rf}^{(B)}\right)$ was assessed through a simulation study. The Monte-Carlo contribution of $\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf}^{(B)}\right)\right]$ is given by

$$\text{Contribution}_{MC}\left(\widehat{\mu}_{rf}^{(B)}\right) := 100 \times \frac{\frac{1}{R}\sum_{r=1}^{R} V_{MC,\Theta}^{(r)}\left(\widehat{\mu}_{rf}^{(B)}\right)}{V_{MC}\left(\widehat{\mu}_{rf}^{(B)}\right)},$$

where $V_{MC}\left(\widehat{\mu}_{rf}^{(B)}\right)$ denotes the usual Monte-Carlo variance of $\widehat{\mu}_{rf}^{(B)}$ and $V_{MC,\Theta}^{(r)}\left(\widehat{\mu}_{rf}^{(B)}\right)$ denotes the Monte-Carlo conditional variance of $\widehat{\mu}_{rf}^{(B)}$ computed by, conditionally on the rth population, the rth sample and the rth set of respondents, resampling from $\mathbb{P}_\Theta$ a number R_Θ of iterations to compute the Monte-Carlo variance of the estimator $\widehat{\mu}_{rf}^{(B)}(\Theta)$. Results for the survey variables Y_1-Y_5 (see Section 7), are shown in Figure 1. From Figure 1, we note that the contribution of the randomization variance decreases rapidly as B increases. With $B = 50$ trees, the contribution of the randomization variance was below 3% for all survey

Survey variable	Estimator	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
		RB	Coverage	RB	Coverage	RB	Coverage
Y_1	Naïve	-45.2	83.1	-45.1	81.9	-48.4	77.2
	Linearized	-40.8	84.6	-33.2	85.9	-26.0	85.4
	CV	6.4	93.9	5.3	93.3	0.3	90.8
Y_2	Naïve	-10.8	91.6	-9.4	91.9	-13.8	92.0
	Linearized	-9.2	92.0	-5.4	92.6	-7.3	93.4
	CV	-0.7	93.2	2.7	93.6	0.4	94.3
Y_3	Naïve	-67.2	73.8	-67.5	73.4	-70.3	71.5
	Linearized	-60.4	78.4	-48.2	83.9	-37.6	87.6
	CV	4.2	94.6	6.8	95.3	6.1	95.2
Y_4	Naïve	-60.7	78.0	-60.7	78.4	-60.7	78.3
	Linearized	-54.5	81.3	-45.2	85.5	-34.8	88.9
	CV	6.6	95.4	4.4	95.1	1.8	95.0
Y_5	Naïve	-3.0	94.1	-3.9	94.2	-1.7	94.2
	Linearized	-2.9	94.1	-3.8	94.2	-1.5	94.2
	CV	-0.7	94.4	-1.7	94.4	0.5	94.5

Table 9: Monte-Carlo simulation results for random forest variance estimators for $p = 5$ and correlated covariates.

variables. The results of this experiment suggest that we can safely omit the randomization variance from the computations for large B, say $B = 1,000$.

Next, we provide a concentration inequality that highlights that, with high probability, the random forest imputed estimator based on a finite number of trees B can be made arbitrarily close to the (unknown) infinite forest imputed estimator.

Proposition 7.1. *Fix $B \in \mathbb{N}$ and $\epsilon > 0$. The probability that the finite forest imputed estimator is not in an ϵ-neighbourhood of the infinite forest estimator is bounded by*

$$\mathbb{P}_\Theta\left(|\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}| > \epsilon\right) \leqslant 2\exp\left(\frac{-B\epsilon^2}{2n_m^2\left(\frac{\sup_{\omega\in\Omega_Y} Y(\omega) - \inf_{\omega\in\Omega_Y} Y(\omega)}{\min_{k\in U}\pi_k}\right)^2}\right), \tag{22}$$

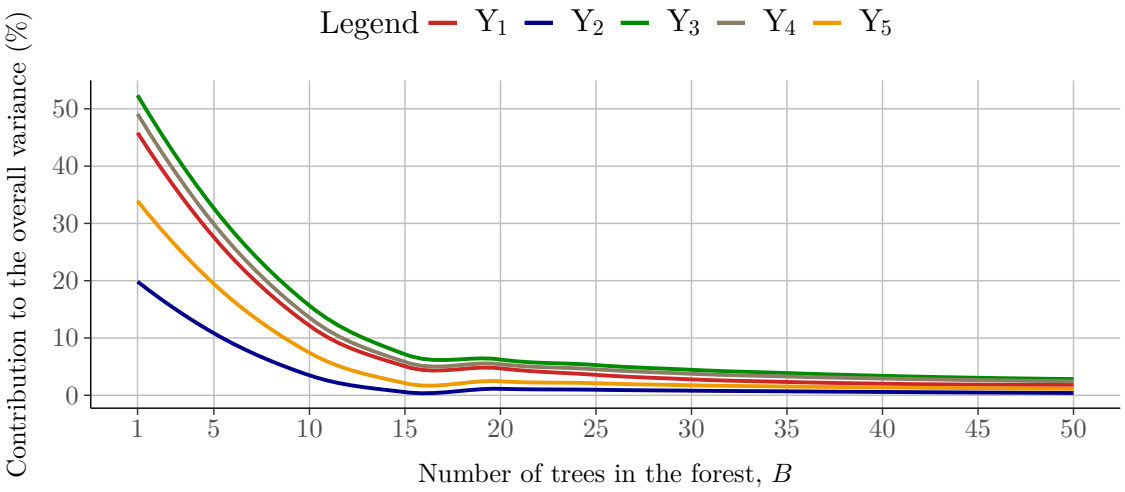

Figure 1: Contribution of the randomization variance $\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf,j}^{(B)}\right)\right]$ to the overall variance $\mathbb{V}\left(\widehat{\mu}_{rf,j}^{(B)}\right)$ as a function of B, with $p = 5$ correlated predictors.

where Ω_Y *denotes the sample space of the random variable* Y*.*

Since the bound given decreases to 0 as B increases, it follows from (22) that $\widehat{\mu}_{rf}^{(B)}$ converges in probability to $\widehat{\mu}_{rf}^{(\infty)}$. This result is not surprising as almost sure convergence (see (13)) implies convergence in probability. The bound (22) can be used to select the number of trees in practice. For simple random sampling without replacement, the denominator on the right hand-side of (22) can be expressed as a function of the unweighed response rate, $\bar{p} := n_r/n$:

$$\mathbb{P}_\Theta\left(|\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}| > \epsilon\right) \leqslant 2\exp\left(\frac{-B\epsilon^2}{2N^2\left(1-\bar{p}\right)^2\left(\sup_{\omega\in\Omega_Y} Y(\omega) - \inf_{\omega\in\Omega_Y} Y(\omega)\right)^2}\right),$$

The above expression suggests that, to obtain sufficient closeness between $\widehat{\mu}_{rf}^{(B)}$ and $\widehat{\mu}_{rf}^{(\infty)}$, a large number of trees is required when the nonresponse rate $1-\bar{p}$ is appreciable.

7.2 Choice of n_0

The number of observations, n_0, in each terminal node of a tree determines its complexity: A small value of n_0 tends to produce flexible predictions, exhibiting low bias but potentially a high variance. To avoid overfitting and to reduce the unnecessary complexity of a tree, it is common practice to perform some form of pruning (Hastie et al., 2011). To illustrate the impact of n_0 on the properties of imputed estimators, we conducted a limited simulation

study, using the same setup as the one described in Section 7. The values of n_0 varied from 1 to $(\mathbb{E}[n_r]+1)/2$, the latter most often leading to a single node in each tree. We computed the Monte-Carlo bias, variance and mean squared error of the imputed estimator of $\mu_1, \ldots, \mu_5$, the population means of $Y_1, \ldots, Y_5$, respectively. The results are shown in Figure 2.

From Figure (2), the behaviour of the tree imputed estimator was similar across all survey variables except Y_3. In every scenario, small values of n_0 led to the best results in terms of bias and variance. As n_0 increased, the bias increased. This can be explained by the fact that a large value of n_0 leads to shallow trees and somewhat heterogeneous terminal nodes in terms of the survey variable requiring imputation. For the variable Y_3, both the bias and the variance were essentially identical for all values of n_0. This is an uncommon scenario. Our findings indicate that selecting values for n_0 in the range of 5 to 15 seems to be a safe choice.

7.3 Choice of p_0

In Section 7, the number of predictors considered at each split was set to $p_0 = p$. In this section, we dig deeper into this aspect. We start with the following proposition.

Proposition 7.2. *Let T_b denote the number of nodes in the b-th tree and X denote an arbitrary predictor among $X_1, ..., X_p$. Then,*

$$\mathbb{P}\left("X \text{ not considered in } \widehat{m}_{rf}^{(B)} " \right) = \prod_{b=1}^{B} \left\{ 1 - \frac{p_0}{p} \right\}^{T_b - 1}.$$

Based on Proposition 7.2, when the value of p_0 is relatively small compared to p, for a given fixed B, there is a high probability that a predictor X will not be considered. Nevertheless, in order to effectively decrease the potential nonresponse bias, it is crucial for the predictions to include at least the predictors that are associated with both the survey variable requiring imputation and the response indicators. However, if p_0 is small compared to p, the predictions will likely fail to incorporate these important predictors. Ultimately, this may result in an inadequate reduction of the potential bias caused by nonresponse. To cope with this issue, we suggest performing a set of univariate analyses to determine which predictors among the available predictors are related to the response indicator. The selected predictors would then be considered at each split with probability one. For the non-selected

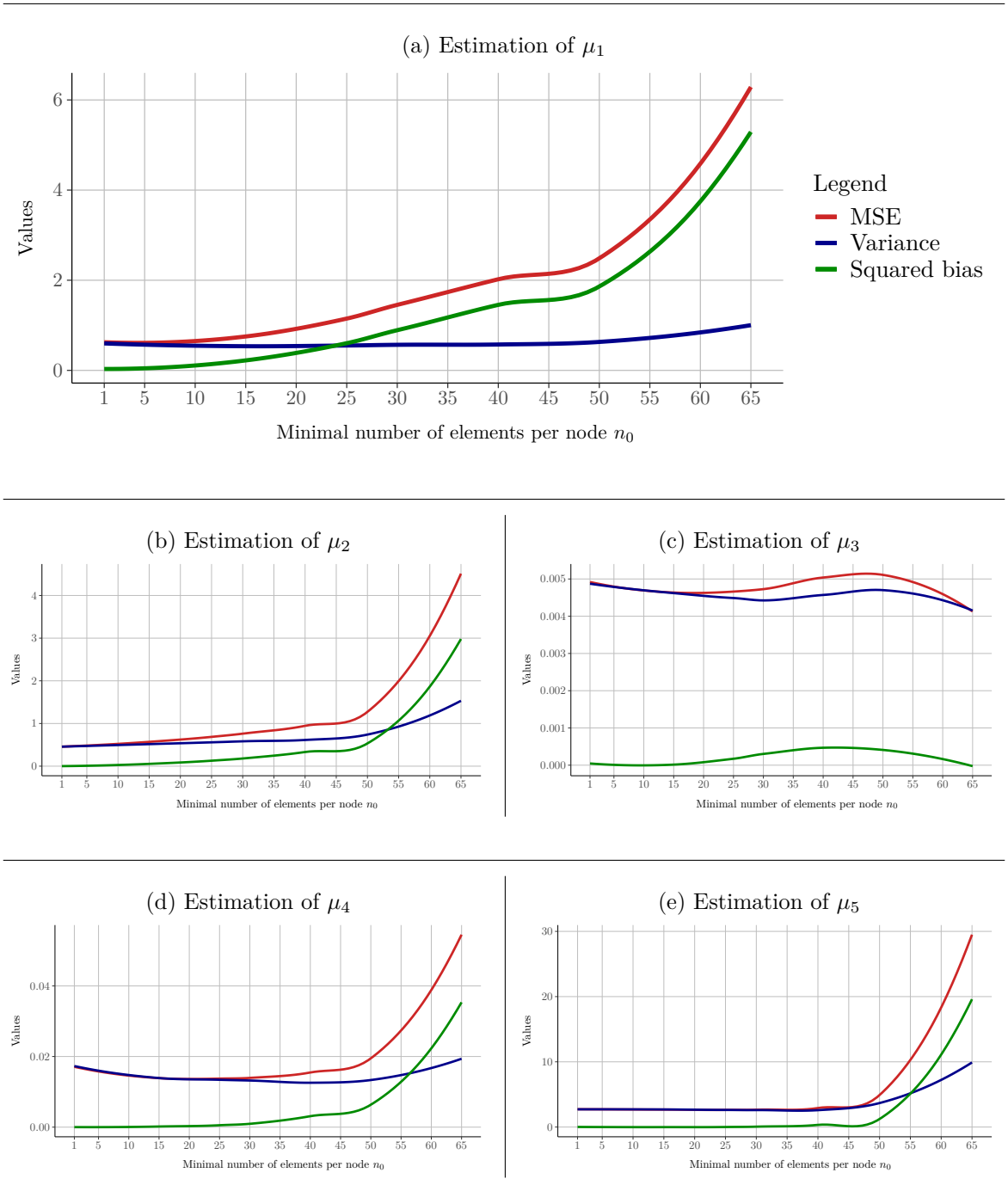

Figure 2: Square bias (green curve), variance (red curve) and mean square error (red curve) of tree imputed estimators as a function of n_0 with $p = 5$ correlated predictors.

predictors, we select, as usual, a subset of predictors at random.

To illustrate the effect of p_0 on the quality of the resulting estimators, we conducted a simulation study, using the same setup as the one described in Section 7. Recall from (21) that

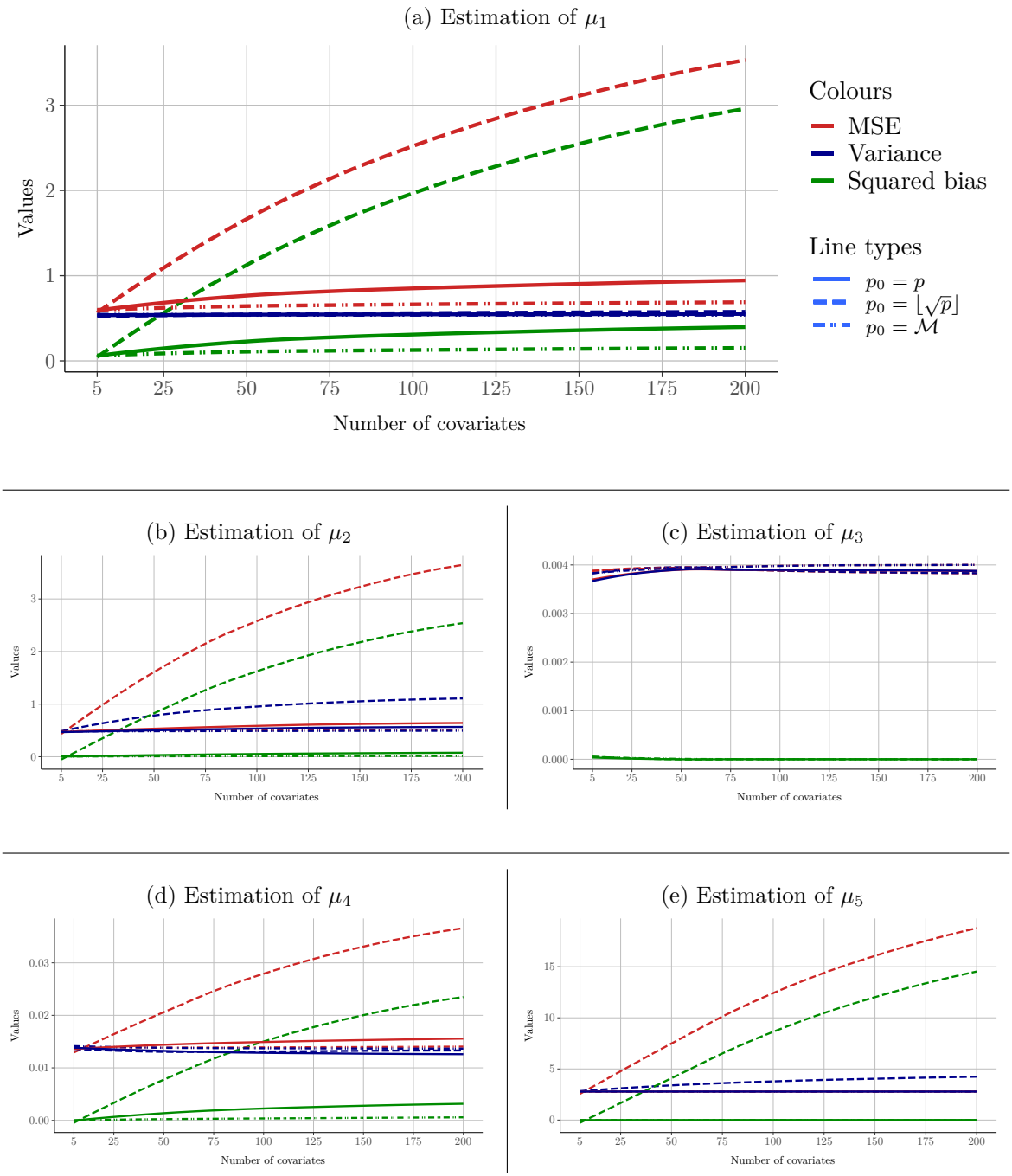

Figure 3: Evolution of the mean squared errors (red curves), squared biases (green curves) and variances (blue curves) of random forests estimators as the number of covariates p increases. Dotted lines represent the choice $p_0 = \lfloor\sqrt{p}\rfloor$, full lines represent $p_0 = p$, combination of both indicates the choice $p_0 = \mathcal{M} := \lfloor\sqrt{p}\rfloor + \text{MAR}$ variables.

the predictors X_1-X_5 were related to the response indicators. Again, we were interested in estimating $\mu_1, \mu_2, \ldots, \mu_5$, the population means of the survey variables Y_1-Y_5, respectively. We

considered three choices for p_0: $p_0 = \lfloor\sqrt{p}\rfloor$, $p_0 = p$, and $p_0 = \mathcal{M} := \lfloor\sqrt{p-5}\rfloor + \{X_1, \ldots, X_5\}$. For the latter choice, the predictors X_1-X_5 were considered at each split with probability one as they are associated with the probability of response, while the $p-5$ remaining predictors were subject to a random selection. In our experiments, the number of predictors p ranged between 5 and 200.

Results are shown in Figure (3). We start by noting that the default choice in most software packages, $p_0 = \lfloor\sqrt{p}\rfloor$, produced biased estimators, in general. The bias increased as the number of predictors p increased. This can be explained that, as the number of predictors increased, the predictors X_1-X_5 had a tiny chance of being considered at each split. As a result, we anticipate that a significant proportion of the predictions did not incorporate X_1-X_5. The choice $p_0 = p$ led to good results in all the scenarios, as expected. Finally, our proposal, $p_0 = \mathcal{M}$, led to results as good or better than the ones obtained with $p_0 = p$. Note that this choice led to good results even when p was larger than the number of respondents n_r. For these reasons, we recommend the choice $p_0 = \mathcal{M}$ in practice.

8 Final remarks

In this article, we have studied the theoretical properties of imputed estimators obtained after regression tree imputation and random forest imputation. In particular, we have established the mean square consistency of imputed estimator in a high-dimensional setting, whereby the number of covariates was allowed to diverge. Also, we have proposed a novel variance estimator based on a K-fold cross-validation procedure. Unlike the customary variance estimator based on a first-order Taylor expansion, the simulation results suggest that the proposed variance estimator performs well in terms of bias and coverage rate of normal-based confidence intervals.

It would be desirable to study the theoretical properties of alternative tree-based procedures such as boosting. For instance, the algorithm XGboost (Chen and Guestrin, 2016) is known to perform very well in terms of bias and efficiency in the context of imputation for missing survey data. However, to the best of our knowledge, the properties of XGboost in a finite population setting remains yet to be studied. This is a topic of future research.

The performance of the proposed variance estimator based on a K-fold cross-validation procedures is promising. Using the same idea, it would be desirable to develop a generic variance estimator that could be applied for any machine learning procedure. This will be treated elsewhere.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr):1063–1095.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(Sep):2015–2033.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

Breidt, F.-J. and Opsomer, J.-D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4):1023–1053.

Breiman, L. (1984). *Classification and regression trees*. Routledge.

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

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chen, S. and Haziza, D. (2019). Recent developments in dealing with item non-response in surveys: a critical review. *International Statistical Review*, 87:S192–S218.

Chen, T. and Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*. ACM Press.

Chi, C.-M., Vossler, P., Fan, Y., and Lv, J. (2022). Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438.

Dagdoug, M., Goga, C., and Haziza, D. (2023). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542):1234–1251.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Díaz-Uriarte, R. and de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3.

Fay, R. (1991). *A design-based perspective on missing data variance*. US Census Bureau.

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., and Dickhaus, H. (2012). Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, 108(1):10–19.

Genuer, R. and Poggi, J.-M. (2019). *Les forêts aléatoires avec R*. Presses universitaires de Rennes.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on barro colorado island — digital soil mapping using random forests analysis. *Geoderma*, 146(1-2):102–113.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hamza, M. and Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, 75(8):629–643.

Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C., editors, *Handbook of statistics*, volume 29A, pages 215–246. Elsevier.

Haziza, D. and Vallée, A.-A. (2020). Variance estimation procedures in the presence of singly imputed survey data: a critical review. *Japanese Journal of Statistics and Data Science*, 3(2):583–623.

Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, 77:49–61.

Kane, M., Price, N., Scotch, M., and Rabinowitz, P. (2014). Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC Bioinformatics*, 15(1).

Kim, J. K. and Rao, J. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96(4):917–932.

Klusowski, J. and Tian, P. (2022). Large scale prediction with decision trees. *Journal of the American Statistical Association*.

Kuhn, M. (2022). *caret: Classification and Regression Training*. R package version 6.0-93.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881.

Nobel, A. (1996). Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084–1105.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. In *Handbook of statistics*, volume 29, pages 455–487. Elsevier.

Qi, Y. (2012). *Random forests for bioinformatics*, pages 307–323. Springer.

Robinson, P. M. and Särndal, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā Ser. B*, 45(2):240–248.

Rogez, G., Rihan, J., Ramalingam, C., Orrite, C., and Torr, P. (2008). Randomized trees for human pose detection. In *Computer Vision and Pattern Recognition, CVPR, IEEE Conference on.*, pages 1–8.

Scornet, E. (2016). On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72–83.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94(445):254–265.

Stekhoven, D. J. and Buhlmann, P. (2011). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Therneau, T. and Atkinson, B. (2022). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1.16.

Wright, M. and Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*.

Xu, T., Zhu, R., and Shao, X. (2022). On variance estimation of random forests. *arXiv preprint arXiv:2202.09008*.

Yang, S. and Kim, J. K. (2019). Nearest neighbor imputation for general parameter estimation in survey sampling. In *The Econometrics of Complex Survey Data*. Emerald Publishing Limited.

Zhou, Z., Mentch, L., and Hooker, G. (2019). Asymptotic normality and variance estimation for supervised ensembles. *arXiv preprint arXiv:1912.01089*.

A Appendix

We start be presenting a preliminary result that will prove useful in establishing the mean square consistency of imputed estimators obtained through regression trees and random forests.

Result A.1. *We assume that (H1) holds. Let* $\{\widetilde{m}_v\}_{v\in\mathbb{N}}$ *be a sequence of regression function estimates fitted on* $D_{U_v} := \{(\mathbf{x}_k, y_k)\,; k \in U_v\}$ *and let* $\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}$ *be independent of* D_{U_v}*. Let* $\{\widehat{m}_v\}_{v\in\mathbb{N}}$ *be the corresponding estimates fitted on* $D_{r_v} = \{(\mathbf{x}_k, y_k); k \in S_{r,v}\}$*. If*

i) The sequence of population predictors $\{\widetilde{m}_v\}_{v\in\mathbb{N}}$ *satisfies*

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\widetilde{m}_v(\mathbf{x}) - m(\mathbf{x})\right)^2\right] = 0,$$

with a convergence rate denoted γ_{N_v}.

ii) There exists a positive constant C*, independent of* v*, such that*

$$\mathbb{E}\left\{\left(\widehat{m}_v(\mathbf{x}) - m(\mathbf{x})\right)^2 \;\middle|\; \boldsymbol{r}_v, \boldsymbol{X}_v, \boldsymbol{I}_v\right\} \leqslant C. \qquad a.s.$$

Then, the sequence of imputed estimators $\{\widehat{\mu}_{\widehat{m}_v}\}_{v\in\mathbb{N}}$ *is mean square consistent with rate*

$$\text{MSE}\left(\widehat{\mu}_{\widehat{m}_v}\right) = \mathcal{O}\left(\gamma_{n_v}\right). \tag{23}$$

Note that Result A.1 holds in a high-dimensional setting in which the number of covariates $\{p_v\}_{v\in\mathbb{N}}$ is allowed to increase to infinity, provided that conditions i) and ii) of Result A.1 are satisfied. Condition (i) is satisfied for a large number of (parametric and nonparametric) estimators of the regression function including k-nearest neighbors and kernel regression, among others; see Györfi et al. (2006). Result A.1 suggests that, in order to build a consistent imputed estimator, it is enough to construct a consistent predictor for imputing the missing values.

Proof. We write

$$\mathbb{E}\left[\left(\widehat{\mu}_{\widehat{m}_v} - \mu_v\right)^2\right] \leqslant 2\mathbb{E}\left[\left(\widehat{\mu}_{\widehat{m}_v} - \widehat{\mu}_{\pi,v}\right)^2\right] + 2\mathbb{E}\left[\left(\widehat{\mu}_{\pi,v} - \mu_v\right)^2\right], \tag{24}$$

where $\widehat{\mu}_{\pi,v}$ denotes the complete data estimator given by (1). The second term of the right hand-side of (24) is the mean squared error of $\widehat{\mu}_{\pi,v}$ that can be expressed as

$$\mathbb{E}\left[\left(\widehat{\mu}_{\pi,v} - \mu_v\right)^2\right] = \mathbb{E}\left[\frac{1}{N_v^2}\sum_{k\in U_v}\alpha_k^2 y_k^2\right] + \mathbb{E}\left[\frac{1}{N_v^2}\sum_{\substack{k,\ell\in U_v\\ k\neq\ell}}\alpha_k\alpha_\ell y_k y_\ell\right],$$

where $\alpha_k := I_k\pi_k^{-1} - 1$. Under (H1) each of these term converge to zero with the rate $\mathcal{O}(n_v^{-1})$. It remains to show that the first term on the right hand-side of (24) converges to 0 with the

rate $\mathcal{O}(\gamma_v)$. Recall that

$$\left(\widehat{\mu}_{\widehat{m}_v} - \widehat{\mu}_{\pi,v}\right) = \frac{1}{N_v} \sum_{k \in S_v} \left\{ \frac{(1-r_k)}{\pi_k} \left(\widehat{m}_v(\mathbf{x}_k) - y_k\right) \right\}.$$

Hence,

$$\begin{aligned} \mathbb{E}\left[\left(\widehat{\mu}_{\widehat{m}_v} - \widehat{\mu}_{\pi,v}\right)^2\right] \leqslant 2\mathbb{E}\left[\left(\frac{1}{N_v} \sum_{k \in S_v} \frac{(1-r_k)}{\pi_k} \cdot \{\widehat{m}_v(\mathbf{x}_k) - m(\mathbf{x}_k)\}\right)^2\right] \\ + 2\mathbb{E}\left[\left(\frac{1}{N_v} \sum_{k \in S_v} \frac{(1-r_k)}{\pi_k} \left(m(\mathbf{x}_k) - y_k\right)\right)^2\right]. \end{aligned} \tag{25}$$

We now establish the consistency of the second term on the right hand-side of (25) with the rate $\mathcal{O}(n_v^{-1})$. Write

$$\begin{aligned} \mathbb{E}\left[\left(\frac{1}{N_v} \sum_{k \in S_v} (1-r_k)\left(m(\mathbf{x}_k) - y_k\right)\right)^2\right] = \mathbb{E}\left[\frac{1}{N_v^2} \sum_{\substack{k,\ell \in S_v \\ \ell \neq k}} \frac{(1-r_k)}{\pi_k} \frac{(1-r_\ell)}{\pi_l} \times \epsilon_k \epsilon_\ell\right] \\ + \mathbb{E}\left[\frac{1}{N_v^2} \sum_{k \in S_v} \left(\frac{(1-r_k)}{\pi_k}\right)^2 \epsilon_k^2\right]. \end{aligned} \tag{26}$$

For the first term on the right hand-side of (26), we use the law of total expectation to obtain

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{N_v^2} \sum_{\substack{k,\ell \in S_v \\ l \neq k}} \frac{(1-r_k)}{\pi_k} \frac{(1-r_l)}{\pi_l} \times \epsilon_k \epsilon_\ell\right] \\ &= \mathbb{E}\left[\frac{1}{N_v^2} \sum_{\substack{k,\ell \in S_v \\ \ell \neq k}} \frac{(1-r_k)}{\pi_k} \frac{(1-r_\ell)}{\pi_\ell} \mathbb{E}\left[\epsilon_k \epsilon_\ell \middle| \boldsymbol{X}_v, \boldsymbol{I}_v, \boldsymbol{r}_v\right]\right] \end{aligned}$$

Since the random variables ϵ_k and ϵ_ℓ are independent for all $k \neq \ell$ and $\mathbb{E}\left[\epsilon_k | \mathbf{x}_k\right] = 0, k \in U$, it follows that

$$\mathbb{E}\left[\epsilon_k \epsilon_\ell \middle| \boldsymbol{X}_v, \boldsymbol{I}_v, \boldsymbol{r}_v\right] = \mathbb{E}\left[\epsilon_k \middle| \boldsymbol{X}_v, \boldsymbol{I}_v, \boldsymbol{r}_v\right] \mathbb{E}\left[\epsilon_\ell \middle| \boldsymbol{X}_v, \boldsymbol{I}_v, \boldsymbol{r}_v\right] = 0. \tag{27}$$

Therefore, the first term on the right hand-side of (26) is 0. For the second term, we have

that

$$\mathbb{E}\left[\frac{1}{N_v^2}\sum_{k\in S_v}\left(\frac{(1-r_k)}{\pi_k}\right)^2\epsilon_k^2\right] \leqslant \frac{N_v}{\lambda^2 N_v^2}\max_{k\in U_v}\mathbb{E}\left[\epsilon_k^2\right] = \frac{\sigma^2}{\lambda^2 N_v} = O(N_v^{-1}).$$

It remains to show that the first term on the right hand-side of (25) is $\mathcal{O}(\gamma_v)$. Bounding arguments ensure that

$$\mathbb{E}\left[\left(\frac{1}{N_v}\sum_{k\in S_v}\frac{(1-r_k)}{\pi_k}\left(\widehat{m}_v(\mathbf{x}_k)-m(\mathbf{x}_k)\right)\right)^2\right] \leqslant \frac{n_v}{\lambda^2 N_v}\mathbb{E}\left[\frac{1}{N_v}\sum_{k\in S_{m,v}}\left(\widehat{m}_v(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2\right].$$

Now, Condition ii) implies that there exists a positive constant $C>0$, independent of v, such that

$$\mathbb{E}\left[\left(\widehat{m}_v(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2 \,\middle|\, \boldsymbol{r}_v, \boldsymbol{X}_v, \boldsymbol{I}_v\right] \leqslant C, \qquad \text{a.s.}$$

It follows from Condition ii) and Lemma 2 that, uniformly,

$$\mathbb{E}\left[\left(\widehat{m}_v(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2 \,\middle|\, \boldsymbol{r}_v, \boldsymbol{X}_v, \boldsymbol{I}_v\right] \xrightarrow{\mathbb{P}} 0.$$

Hence, by the Lebesgues dominated convergence theorem,

$$\lim_{v\to\infty}\mathbb{E}\left[\frac{1}{N_v}\sum_{k\in S_{m,v}}\mathbb{E}\left[\left(\widehat{m}_v(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2\middle|\boldsymbol{r}_v, \boldsymbol{X}_v, \boldsymbol{I}_v\right]\right] = 0,$$

with the rate $O(\gamma_v)$. Moreover, $\max(\gamma_v, 1/n_v) = \gamma_v$ since, by assumption, $\{\gamma_v\}_{v\in\mathbb{N}}$ is the functional rate of convergence of $\{\widehat{m}_v\}_{v\in\mathbb{N}}$ towards m. The result follows. ■

A.1 Proof of Result 4.1.

We begin by noting that, from Corollary 4.3 of Klusowski and Tian (2022), it follows that the sequence $\{\widetilde{m}_{tree,v}\}_{v\in\mathbb{N}}$ of tree predictors fitted on D_{N_v} is universally consistent in L^2 for m, meaning

$$\lim_{v\to\infty}\mathbb{E}\left[\left(\widetilde{m}_{tree,v}(\mathbf{x})-m_v(\mathbf{x})\right)^2\right] = 0,$$

which is Condition ii) of Result A.1. Since Y is assumed to be almost surely bounded, it follows that, there exists $C > 0$, satisfying

$$\mathbb{E}\left\{\left(\widehat{m}_{tree,v}(\mathbf{x}) - m_v(\mathbf{x})\right)^2 \;\middle|\; \boldsymbol{r}_v, \boldsymbol{X}_v, \boldsymbol{I}_v\right\} \leqslant C. \qquad \text{a.s.}$$

Therefore, Condition ii) of Result A.1 holds as well. Hence, Result A.1 ensures that the mean-square consistency of $\{\widehat{\mu}_{tree,v}\}_{v\in\mathbb{N}}$ is established.

A.2 Proof of Proposition 5.1.

We can write

$$\begin{aligned}
\widehat{\mu}_{rf}^{(B)} &= \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_k)}{\pi_k}\right) \\
&= \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{1}{B}\sum_{b=1}^{B}\frac{\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)}{\pi_k}\right) \\
&= \frac{1}{B}\sum_{b=1}^{B}\underbrace{\frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)}{\pi_k}\right)}_{\widehat{\mu}_{tree}^{(b)}}.
\end{aligned}$$

A.3 Proof of Proposition 5.2.

We can rite

$$\begin{aligned}
\mathrm{MSE}\left(\widehat{\mu}_{rf}^{(B)}\right) &= \mathbb{E}\left[\left(\widehat{\mu}_{rf}^{(B)} - \mu\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{B^2}\left(\sum_{b=1}^{B}\left\{\widehat{\mu}_{tree}^{(b)} - \mu\right\}\right)^2\right] \qquad \text{(By Proposition 5.1).} \\
&\leqslant \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}\left[\left(\widehat{\mu}_{tree}^{(b)} - \mu\right)^2\right] \qquad \text{(By CS ineq. on Euclidean inner product).} \\
&= \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}\left[\mathbb{E}_\Theta\left[\left(\widehat{\mu}_{tree}^{(b)} - \mu\right)^2\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}_\Theta\left[\left(\widehat{\mu}_{tree}^{(b)} - \mu\right)^2\right]\right] \\
&= \mathrm{MSE}\left(\widehat{\mu}_{tree}^{(b)}\right).
\end{aligned}$$

using that $\{\Theta_b\}_{b=1,...,B}$ are i.i.d.; moreover, equality holds if and only if there exists C such that $\widehat{\mu}_{tree}^{(b)} = C$, almost surely, for all $b = 1, ..., B$, which is equivalent to

$$\widehat{\mu}_{tree}^{(1)} = \widehat{\mu}_{tree}^{(2)} = ... = \widehat{\mu}_{tree}^{(B)} = \widehat{\mu}_{rf}^{(B)}, \qquad a.s.$$

which means that the forest is degenerate.

A.4 Proof of Lemma 1.

The proof essentially follows ideas described in Scornet (2016). Write

$$\begin{aligned}\left(\widehat{\mu}_{rf}^{(B)} - \mu_y\right)^2 &= \left(\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)} + \widehat{\mu}_{rf}^{(\infty)} - \mu_y\right)^2 \\ &= \left(\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}\right)^2 + \left(\widehat{\mu}_{rf}^{(\infty)} - \mu_y\right)^2 + 2\left(\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}\right)\left(\widehat{\mu}_{rf}^{(\infty)} - \mu_y\right). \quad (28)\end{aligned}$$

Next, note that

$$\mathbb{E}\left[\left(\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}\right)\left(\widehat{\mu}_{rf}^{(\infty)} - \mu\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\left(\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}\right)\middle|\boldsymbol{r}, \boldsymbol{X}, \boldsymbol{I}, \boldsymbol{y}\right]\left(\widehat{\mu}_{rf}^{(\infty)} - \mu\right)\right] = 0.$$

Taking expectations on both sides of (28) leads to

$$\mathbb{E}\left[\left(\widehat{\mu}_{rf}^{(B)} - \mu\right)^2\right] = \mathbb{E}\left[\left(\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}\right)^2\right] + \mathbb{E}\left[\left(\widehat{\mu}_{rf}^{(\infty)} - \mu\right)^2\right], \quad (29)$$

so that

$$\mathbb{E}\left[\left(\widehat{\mu}_{rf}^{(B)} - \mu\right)^2\right] - \mathbb{E}\left[\left(\widehat{\mu}_{rf}^{(\infty)} - \mu\right)^2\right] = \mathbb{E}\left[\left(\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}\right)^2\right] \geqslant 0. \quad (30)$$

Next, write

$$\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)} = \frac{1}{N_v}\sum_{k \in S_m} \frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k},$$

so that

$$\mathbb{E}\left[\left(\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}\right)^2\right] = \frac{1}{N_v^2} \cdot \mathbb{E}\left[\left(\sum_{k \in S_m} \frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k}\right)^2\right]$$

$$\leqslant \frac{n_v}{N_v^2} \cdot \mathbb{E}\left[\sum_{k \in S_m} \frac{\left(\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)\right)^2}{\pi_k^2}\right]$$
$$\leqslant \frac{n_v N_v}{N_v^2 \lambda^2} \cdot \max_{k \in U} \mathbb{E}\left[\left(\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)\right)^2\right]$$

Now, using Theorem 3.3 of Scornet (2016), there exists a positive constant C such that, uniformly,

$$\mathbb{E}\left[\left(\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)\right)^2\right] \leqslant \frac{C}{B_v},$$

leading to

$$\mathbb{E}\left[\left(\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}\right)^2\right] \leqslant \frac{C n_v N_v}{N_v^2 \lambda^2 B_v} = \mathcal{O}\left(\frac{1}{B_v}\right).$$

A.5 Proof of Result 5.1.

Corollary 1 of Scornet (2016) leads to the consistency of the infinite forest estimator $\{\widehat{m}_{urf,v}^{(\infty)}\}_{v \in \mathbb{N}}$ in a framework in which the dimension p is fixed. We extend their proof to a high-dimensional asymptotic framework. To that aim, we shall make use of Stone's Theorem, see e.g., Györfi et al. (2006), page 56. We begin by noting that, in our framework, verifying Stone's theorem conditions is enough to ensure consistency. That is, as shown by Biau et al. (2008) and Scornet (2016), we must prove that

$$\text{For all } K > 0, \quad \lim_{v \to \infty} \mathbb{P}\left\{\text{Card}\left(A_v\left(\mathbf{x}, \Theta\right)\right) > K\right\} = 1, \tag{31}$$

$$\text{For all } \epsilon > 0, \quad \lim_{v \to \infty} \mathbb{P}\left\{\text{diam}\left(A_v\left(\mathbf{x}, \Theta\right)\right) > \epsilon\right\} = 0, \tag{32}$$

where $\text{diam}\left(A_v\left(\mathbf{x}, \Theta\right)\right)$ is used to denote the diameter of the hyper-rectangle $A_v\left(\mathbf{x}, \Theta\right)$, i.e., the maximal distance between two points in the rectangle. The proof given by Scornet (2016) of (31) continues to hold in a high-dimensional asymptotic framework. It is thus enough to prove (32). To that aim, let $d_{v,j}$ denote the length of the j-th side of the rectangle containing $\mathbf{x}$, and $\mathbf{d} := [d_{1,v}, d_{2,v}, ..., d_{p_v,v}]^\top$. Let $\epsilon > 0$ and write

$$\begin{aligned}
\mathbb{P}\left\{\text{diam}\left(A_v\left(\mathbf{x}, \Theta\right)\right) > \epsilon\right\} &\leqslant \mathbb{P}\left\{||\mathbf{d}||_2 > \epsilon\right\} \\
&\leqslant \mathbb{P}\left\{||\mathbf{d}||_1 > \epsilon\right\} \\
&\leqslant \frac{\mathbb{E}\left[\sum_{j=1}^{p_v} d_{j,v}\right]}{\epsilon}
\end{aligned}$$

$$= p_v \times \frac{\mathbb{E}\left[d_{1,v}\right]}{\epsilon},$$

using norms' inequality, Markov's inequality and symmetry of the dimensions. Let $K_{1,v}$ denote the number of times the leaf containing $\mathbf{x}$ has been cut along the first coordinate. Then, as noted by Biau et al. (2008), we can use the following inequality,

$$\mathbb{E}\left[d_{1,v}\right] \leqslant \mathbb{E}\left[\left(\frac{3}{4}\right)^{K_{1,v}}\right]$$

to obtain that

$$\mathbb{E}\left[d_{1,v}\right] \leqslant \sum_{l=0}^{L_v} \binom{L_v}{l} \left(\frac{3}{4}\right)^l \left(\frac{1}{p_v}\right)^l \left(1-\frac{1}{p_v}\right)^{L_v-l} = \left(1-\frac{1}{4p_v}\right)^{L_v}.$$

Therefore, combining these inequalities lead to

$$\mathbb{P}\left\{\operatorname{diam}\left(A_v\left(\mathbf{x},\Theta\right)\right) > \epsilon\right\} \leqslant \frac{p_v}{\epsilon}\left(1-\frac{1}{4p_v}\right)^{L_v}.$$

Hence, under our conditions, both (31) and (32) hold, leading to

$$\lim_{v\to\infty} \mathbb{E}\left[\left\{\widehat{m}_v^{(\infty)}(\mathbf{x}) - m_v(\mathbf{x})\right\}^2\right] = 0.$$

Applying Result A.1 gives the consistency of the infinite uniform forest estimator. Moreover, from Lemma 1, we have

$$0 \leqslant \mathbb{E}\left[\left(\,\widehat{\mu}_{urf,v}^{\,(B)} - \mu_v\,\right)^2\right] - \mathbb{E}\left[\left(\,\widehat{\mu}_{urf,v}^{\,(\infty)} - \mu_v\,\right)^2\right] \leqslant \frac{C}{B_v}.$$

Thus, if we consider large forests (i.e., with an increasing number of trees), the sequences $\mathbb{E}\left[\left(\widehat{\mu}_{urf,v}^{\,(B)} - \mu_v\right)^2\right]$ and $\mathbb{E}\left[\left(\widehat{\mu}_{urf,v}^{\,(\infty)} - \mu_v\right)^2\right]$ must have the same limit. Hence,

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\,\widehat{\mu}_{urf,v}^{\,(B_v)} - \mu_v\,\right)^2\right] = 0,$$

which concludes the proof.

A.6 Proof of Section 6.1.

By the law of iterated variance,

$$\mathbb{V}\left(\widehat{\mu}_{rf}^{(B)} - \mu\right) = \mathbb{V}\left(\mathbb{E}_\Theta\left[\widehat{\mu}_{rf}^{(B)} - \mu\right]\right) + \mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf}^{(B)} - \mu\right)\right].$$

From (13), it follows that

$$\mathbb{V}\left(\widehat{\mu}_{rf}^{(B)} - \mu\right) = \mathbb{V}\left(\widehat{\mu}_{rf}^{(\infty)} - \mu\right) + \mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf}^{(B)} - \mu\right)\right].$$

Relation (15) is proved. Next, using Proposition 4.1, we have

$$\mathbb{V}_\Theta\left(\widehat{\mu}_{rf}^{(B)} - \mu\right) = \mathbb{V}_\Theta\left(\widehat{\mu}_{rf}^{(B)}\right) = \mathbb{V}_\Theta\left(\frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_{tree}^{(b)}\right) \overset{(4)}{=} \frac{1}{B}\cdot\mathbb{V}_\Theta\left(\widehat{\mu}_{tree}^{(1)}\right),$$

where equality (4) follows from the fact that, as detailed in the proof of Proposition 7.1, conditionally on everything but $\{\Theta_b\}_{b=1}^B$, $\{\widehat{\mu}_{tree}^{(b)}\}_{b=1}^B$ is a sequence of i.i.d. random variables. Now, for any $b \in \{1, 2, ..., B\}$,

$$\mathbb{V}_\Theta\left(\widehat{\mu}_{tree}^{(b)}\right) = \mathbb{V}_\Theta\left(\frac{1}{N_v}\sum_{k\in S_m}\frac{\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)}{\pi_k}\right) \leqslant \mathbb{E}_\Theta\left[\left(\frac{1}{N_v}\sum_{k\in S_m}\frac{\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)}{\pi_k}\right)^2\right] \leqslant \frac{n_v^2}{N_v^2}\left(\sup_{\omega\in\Omega_Y}|Y(\omega)|\max_{k\in U}d_k\right)^2.$$

This concludes the proof.

A.7 Proof of Proposition 8.1.

Observe that

$$\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)} = \frac{1}{N_v}\sum_{k\in S_m}\frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k},$$

so that

$$\mathbb{P}_\Theta\left(|\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}| > \epsilon\right) = \mathbb{P}_\Theta\left(\left|\frac{1}{B}\sum_{b=1}^{B}\left\{\frac{1}{N_v}\sum_{k\in S_m}\frac{\widehat{m}_{tree}^{(b)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k}\right\}\right| > \epsilon\right).$$

Define $\widehat{d}^{\,(b)} := \frac{1}{N_v}\sum_{k\in S_m}\pi_k^{-1}\left(\widehat{m}_{tree}^{(b)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)\right)$. Note that, given the covariates, the sample membership indicators, the survey variable and the nonresponse indicators, the

sequence $\{\widehat{m}_{tree}^{(b)}\}_{b=1}^{B}$ is a sequence of independently and identically distributed (according to $\mathbb{P}_\Theta$) random variables. The same holds therefore for the sequence $\{\widehat{d}^{\,(b)}\}_{b=1}^{B}$. Moreover, in our framework, these are zero mean bounded random variables. To see that, first note that $\inf_{\omega\in\Omega_Y} Y(\omega)$ and $\sup_{\omega\in\Omega_Y} Y(\omega)$ are finite constants. Hence, for all $b \in \{1, 2, ..., B\}$ and $k \in S_m$,

$$\inf_{\omega\in\Omega_Y} Y(\omega) - \sup_{\omega\in\Omega_Y} Y(\omega) \leqslant \widehat{m}_{tree}^{(b)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k) \leqslant \sup_{\omega\in\Omega_Y} Y(\omega) - \inf_{\omega\in\Omega_Y} Y(\omega). \qquad a.s.$$

Therefore, noting that $\inf_{\omega\in\Omega_Y} Y(\omega) - \sup_{\omega\in\Omega_Y} Y(\omega) < 0$, it follows that

$$\frac{n_m}{N_v} \cdot \frac{\inf_{\omega\in\Omega_Y} Y(\omega) - \sup_{\omega\in\Omega_Y} Y(\omega)}{\min_{k\in U} \pi_k} \leqslant \widehat{d}^{\,(b)} \leqslant \frac{n_m}{N_v} \cdot \frac{\sup_{\omega\in\Omega_Y} Y(\omega) - \inf_{\omega\in\Omega_Y} Y(\omega)}{\min_{k\in U} \pi_k}, \qquad a.s.$$

Thus, for $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}_\Theta\left(|\widehat{\mu}_{rf}^{(B)} - \widehat{\mu}_{rf}^{(\infty)}| > \epsilon\right) &= \mathbb{P}_\Theta\left(\left|\sum_{b=1}^{B} \widehat{d}^{\,(b)}\right| > B\epsilon\right) \\ &\overset{(3)}{\leqslant} 2\exp\left(\frac{-2B\epsilon^2}{4\frac{n_m^2}{N_v^2}\left(\frac{\sup_{\omega\in\Omega_Y} Y(\omega) - \inf_{\omega\in\Omega_Y} Y(\omega)}{\min_{k\in U} \pi_k}\right)^2}\right),\end{aligned}$$

where (3) follows from Hoeffding inequality for bounded random variables.

A.8 Proof of Proposition 8.2.

Let X be an arbitrary covariate among $X_1, X_2, ..., X_p$. Denote by $\mathcal{S}$ the set of covariates considered (at least once) for splitting in $\widehat{m}_{rf}^{(B)}$, and $\mathcal{S}_b$ those considered in $\widehat{m}_{tree}^{(b)}$. Basic graph theory reveals that, if T_b denotes the number of terminal nodes of the b-th tree $\widehat{m}_{tree}^{(b)}$, then the number of splits in $\widehat{m}_{tree}^{(b)}$ is $T_b - 1$. Finally, let $P_{b,j}$ denote the set of covariates considered for splitting in the j-th split $\widehat{m}_{tree}^{(b)}$. We may then write

$$\begin{aligned}\mathbb{P}\left\{X \notin \mathcal{S}\right\} &= \mathbb{P}\left\{\bigcap_{b=1}^{B} (X \notin \mathcal{S}_b)\right\} \\ &= \mathbb{P}\left\{\bigcap_{b=1}^{B}\bigcap_{j=1}^{T_b-1} (X \notin P_{b,j})\right\} \\ &= \prod_{b=1}^{B}\left\{1 - \frac{p_0}{p}\right\}^{T_b-1},\end{aligned}$$

by independence between each draw of covariates.

A.9 Proof of Lemma 2.

Lemma 2. *Assume (H1). Let $\{\widetilde{m}_v\}_{v\in\mathbb{N}}$ be a sequence of regression function estimates fitted on $D_{U_v} := \{(\mathbf{x}_k, y_k)\,; k \in U_v\}$ and let $\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}$ be independent of D_{U_v}. Let $\{\widehat{m}_v\}_{v\in\mathbb{N}}$ be the corresponding estimates fitted on $D_{r_v} = \{(\mathbf{x}_k, y_k); k \in S_{r,v}\}$. Then, $\{\widehat{m}_v\}_{v\in\mathbb{N}}$ is such that,*

$$\mathbb{E}\left[\left(\widehat{m}_v(\mathbf{x}) - m(\mathbf{x})\right)^2 \middle| \boldsymbol{X}_v, \mathbf{x}, \boldsymbol{I}_v, \boldsymbol{r}_v\right] \xrightarrow{\mathbb{P}} 0.$$

Proof. By assumption,

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\widetilde{m}_v(\mathbf{x}) - m(\mathbf{x})\right)^2\right] = 0.$$

It follows that

$$\lim_{v\to\infty} \mathbb{E}\left[\left|\mathbb{E}\left[\left(\widetilde{m}_v(\mathbf{x}) - m(\mathbf{x})\right)^2 \middle| \boldsymbol{X}_v, \mathbf{x}\right] - 0\right|\right] = 0.$$

In other words, the random variable $\mathbb{E}\left[\left(\widetilde{m}_v(\mathbf{x}) - m(\mathbf{x})\right)^2 \middle| \boldsymbol{X}_v, \mathbf{x}\right] := g(\boldsymbol{X}_v, \mathbf{x})$ converges in L^1 towards 0, which implies that $g(\boldsymbol{X}_v, \mathbf{x}) \xrightarrow{\mathbb{P}} 0$. Furthermore, note that, under MAR and (H1), we have almost sure equality of the two random measures $\mathbb{P}_{Y|X}$ and $\mathbb{P}_{Y|X,I,r}$. Therefore, fixing the sample of respondents and using the equality of conditional distribution gives

$$\mathbb{E}\left[\left(\widehat{m}_v(\mathbf{x}) - m(\mathbf{x})\right)^2 \middle| \boldsymbol{X}_v, \mathbf{x}, \boldsymbol{I}_v, \boldsymbol{r}_v\right] \xrightarrow{\mathbb{P}} 0.$$

■

A.10 Proof of Technical Lemma 1.

Technical lemma 1. *Consider the weights of a regression tree as defined in (7). The following hold:*

i) If there is at least one element per terminal node, then, for all $\mathbf{x} \in \mathbb{R}^p$,

$$\sum_{k\in S_r} \widehat{W}_k(\mathbf{x}) = 1.$$

ii) The weights of the tree can be seen as a the images of a weight function from $\mathbb{R}^p \times \mathbb{R}^p$

to $[0;1]$*, that is,*

$$\widehat{W}_k(\mathbf{x}_\ell) := \widehat{W}\left(\mathbf{x}_k, \mathbf{x}_\ell\right).$$

iii) If there is at least n_0 *elements per terminal node, then the range of* $\widehat{W}$ *reduces to* $[0; n_0^{-1}]$*.*

iv) The weight function is symmetrical in its arguments, that is, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$*,*

$$\widehat{W}\left(\mathbf{x}, \mathbf{y}\right) = \widehat{W}\left(\mathbf{y}, \mathbf{x}\right).$$

Proof. For i), fix $\mathbf{x} \in \mathbb{R}^p$. Using the definition of (7), we have

$$\sum_{k \in S_r} \widehat{W}_k(\mathbf{x}) = \sum_{k \in S_r} \frac{\mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}} = \frac{\sum_{k \in S_r} \mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}} = 1.$$

Point ii) follows directly from the definition. To prove, iii), write

$$\frac{\mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}} \leqslant \frac{1}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}} \leqslant \frac{1}{n_{0v}}$$

by noting that $\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})} \geqslant n_0$. To see iv), let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Observe that

$$\widehat{W}\left(\mathbf{x}, \mathbf{y}\right) = \begin{cases} \dfrac{1}{\sum_{k \in S_r} \mathbb{1}_{\mathbf{x}_k \in A(\mathbf{y})}} & \text{if } \mathbf{x} \in A(\mathbf{y}), \\ 0 & \text{otherwise.} \end{cases}$$

Noting that the conditions $\mathbf{x} \in A(\mathbf{y})$ and $\mathbf{y} \in A(\mathbf{x})$ are the same, it is enough to split cases to prove the equality. Assuming that $\mathbf{x} \in A(\mathbf{y})$, it follows that $A(\mathbf{y}) = A(\mathbf{x})$, so that

$$\widehat{W}\left(\mathbf{x}, \mathbf{y}\right) = \frac{1}{\sum_{k \in S_r} \mathbb{1}_{\mathbf{x}_k \in A(\mathbf{y})}} = \frac{1}{\sum_{k \in S_r} \mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}} = \widehat{W}\left(\mathbf{y}, \mathbf{x}\right).$$

In cases where, $\mathbf{x} \notin A(\mathbf{y})$, then $\widehat{W}\left(\mathbf{x}, \mathbf{y}\right) = 0$ and $\widehat{W}\left(\mathbf{x}, \mathbf{y}\right) = 0$ so that the equality also holds. ■