

Branching into Uncertainty: Theory and Simulations in Random Forest Imputation and Inference

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat. der Fakultät Statistik der
Technischen Universität Dortmund

Vorgelegt von

Nico Föge

geboren in Stade

Dortmund, Juni 2026

Amtierender Dekan:

Prof. Dr. Philipp Doebler

Gutachter:

Prof. Dr. Markus Pauly (Technische Universität Dortmund)

Prof. Dr. Philipp Doebler (Technische Universität Dortmund)

Tag der Prüfung:

04.09.2026

Abstract

Quantifying uncertainty in machine learning models has become a central task in modern statistical research, particularly in the context of Random Forests. Random Forests and related tree-based algorithms are widely used across a range of scientific domains. They are versatile and can be used in classification, regression, or imputation tasks. In all of these settings, assessing uncertainty is essential. In classification and regression tasks, inference for variable importance is a tool for interpretability, where uncertainty quantification is essential, for example, for the construction of confidence intervals. In imputation tasks, the usage of Random Forest (or other tree-based methods) raises the question of how imputation-induced uncertainty impacts inference procedures. In both inference and imputation, important gaps remain in the quantification of uncertainty, which this thesis aims to address.

While the Random Forest Permutation Importance Measure (RFPIM) is an established point estimator for (theoretical) variable importance, uncertainty quantification remains challenging, as the variance of RFPIM is difficult to estimate, and there are hardly any theoretical results on inference with RFPIM. Nevertheless, there are some approaches that estimate the variance via resampling methods. From this, confidence intervals have been derived for variable importance, which are based on an assumption of asymptotic normality. Here, the first research gap arises: (i) This assumption is only based on the results of a simulation study rather than a rigorous mathematical theory.

During the aforementioned construction of confidence intervals, the presence of missing values introduces additional challenges for statistical inference. Missing values can arise due to non-response in surveys, sensor failures, data entry errors, privacy restrictions, or participants dropping out of longitudinal studies. The aforementioned confidence intervals have not been investigated in the presence of missing values, which might affect the coverage rates. This raises the second open research question (ii) of which imputation method should be preferred when aiming for inference with the RFPIM.

The Random Forest itself also offers a potential imputation method implemented, for example, in `missRanger`. Other tree-based methods like Multiple Imputation through 'XGBoost' (e.g., implemented in `mixgb`) also offer imputation methods that might be more flexible than traditional imputation approaches like Multiple Imputation by Chained

Equations (MICE). In many practical applications, missing data arise in structured settings such as multilevel or hierarchical data. Yet, a comprehensive comparison between modern tree-based imputation methods and MICE has not been conducted so far for multilevel data (iii), which is the third gap in the research addressed by this thesis. In particular, adaptations of these tree-based methods to multilevel data structures may further improve their performance in realistic applications.

The first manuscript develops an asymptotic framework for RFPIM to address (i). By exploiting U-statistic theory and randomization arguments, it derives a Central Limit Theorem for the RFPIM under strict regularity conditions. To the best of our knowledge, this is the first Central Limit Theorem for this Importance Measure.

The second paper proposes modified confidence intervals for variable importance in the presence of missing data addressing (ii). It demonstrates that naively applying single imputation yields substantial undercoverage and introduces a multiple-imputation-based variance aggregation that integrates both within- and between-imputation uncertainty. Simulation results and real-data analyses show that the proposed approach often achieves nominal coverage.

The third manuscript focuses on multiple imputation in multilevel data structures and closes the gap (iii). Traditional tree-based imputation techniques implicitly assume independence between observations, which can lead to biased estimates in multilevel settings. In an extensive simulation study, multilevel extensions of chained equations, Random Forests, and gradient boosting methods are compared. The results demonstrate that classical MICE with predictive mean matching performs robustly for cluster-level variables, while modified boosting approaches outperform competitors for individual-level variables with high missingness.

Acknowledgments

First of all, I want to thank the late Marc Ditzhaus, who convinced me to do my PhD in the first place. In addition to being an excellent supervisor, he was a wonderful person. I'm glad that I had the opportunity to work with him for at least a short period of time. Even though he was not able to accompany me during the second half of my PhD, his ideas will remain the foundation of my scientific work.

Special thanks also to Markus Pauly, who caught me in a difficult time and was an equally great supervisor as Marc was. Also, thank you for your supervision and your positive and easygoing nature. In these difficult times, you made sure I could finish my PhD journey straightforwardly.

Also, I want to thank MathCoRe and especially Sebastian Sager. After Marc passed, you immediately checked on me and lifted the financial insecurities. I'm also very grateful for the funding, particularly when I returned to Dortmund and went through a transition period.

I would also like to thank Jakob Schwerter and Philipp Doeblner, who invited me to join the project on multilevel imputation, which is now the third article of this thesis.

I am also very grateful for all my friends, who were part of my journey in the last 3 years. Thank you, Merle, for your emotional support. Thank you, Stefan, for the idea of the s***** A*** M*** c*****. Thank you, Sheila, for your thorough proofreading of the body text. Thank you to all my Dortmund colleagues and friends for making my return joyful. And thank you to all my Magdeburg friends and colleagues for truly *forklifting* times. Further thanks to my parents, grandparents, and the rest of my family.

List of Publications

This cumulative thesis is based on the following three manuscripts:

Article 1: Föge, N., Schmid, L., Ditzhaus, M., & Pauly, M. (2024). A Central Limit Theorem for the permutation importance measure. arXiv preprint arXiv:2412.13020. <https://doi.org/10.48550/arXiv.2412.13020>

Manuscript submitted for publication to the Journal of Statistical Planning and Inference.

Contribution of the author:

For the results presented in this paper, Nico Föge developed the core proof idea for all Lemmas and Theorems. Nico Föge was primarily responsible for the detailed elaboration and formalization of the proofs. The co-authors provided valuable feedback, assisting in refining the presentation of the proofs and identifying errors in the preliminary proof version. Nico Föge prepared the first draft of the publication, which was then jointly polished by all authors.

The reuse of this article in the thesis is granted under the terms of the Creative Commons Attribution 4.0 International License.

Article 2: Föge, N., & Pauly, M. (2026). Confidence intervals for Random Forest permutation importance with missing data. Computational Statistics, 41(3), 45. <https://doi.org/10.1007/s00180-026-01722-w>

Contribution of the author:

Nico Föge prepared the first draft of the publication, including the writing of the simulation code. The conceptualization as well as the proofreading was done by the co-author. Nico Föge also prepared the revision of the manuscript, including several additional numerical experiments.

Article 3: Föge, N., Schwerter, J., Gurtskaia, K. et al. Adapting tree-based multiple imputation methods for multilevel data? A simulation study. Behav Res 58, 166 (2026). <https://doi.org/10.3758/s13428-026-03008-x>

Contribution of the author:

The initial version of this manuscript was prepared and submitted by Jakob Schwerter, Philipp Doeblner, and Ketevan Gurtskaia before Nico Föge joined the project. After the first submission was rejected, Nico Föge became involved and took the lead in substantially revising and extending the study. In particular, Nico Föge restructured and expanded the simulation framework, rewrote large parts of the analysis scripts, and implemented additional methodological components to address key reviewer comments and to improve clarity. Nico Föge also prepared the revision of the manuscript, including new analyses, figures, and textual revisions, and coordinated the resubmission process.

Contents

Abstract	iii
Acknowledgments	v
List of Publications	vi
Notation	viii
I Introduction	1
1 Motivation	3
2 Statistical Methods	7
2.1 CART splitting and Sufficient Impurity Decrease	7
2.2 Feature Importance	9
2.3 Generalized U -statistics and asymptotics of RFPIM	11
2.4 Missing Mechanisms and Imputation Methods	13
3 Summary of the Articles	19
3.1 Article 1: A Central Limit Theorem for the Permutation Importance Measure	19
3.2 Confidence Intervals for Random Forest Permutation Importance with Missing Data	21
3.3 Adapting tree-based multiple imputation methods for multi-level data? .	24
4 Discussion and Outlook	26
Bibliography	28
II Publications	31

Notation

Throughout the thesis, vectors and matrices are denoted by bold symbols, e.g., $\mathbf{M}$.

$\mathbb{N}$	Natural numbers
$\mathbb{R}$	Real numbers
$\mathbb{P}$	Probability measure
$\mathbb{E}(X)$	Expected value of a random variable X
$\mathbb{V}ar(X)$	Variance of real-valued random variable X
$\mathbb{E}(Y X = x)$	Conditional expectation of Y conditioned on the event that $X = x$
$\mathbf{x}^T$	Transposed vector of $\mathbf{x}$
$\times$	Cartesian product
$\xrightarrow{d}$	Convergence in distribution
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with expectation μ and variance σ^2

Part I

Introduction

1 Motivation

In modern empirical research, machine learning algorithms - e.g. Random Forests (Breiman, 2001) - have become indispensable tools for modeling complex relationships. Their ability to capture nonlinear effects, high-dimensional interactions, and mixed data types without strict parametric assumptions makes them a powerful alternative to traditional regression models (Strobl et al., 2009). Beyond predictive performance, the interpretability of such models has gained increasing importance, especially through feature importance measures that quantify the contribution of predictors to model accuracy Molnar (2020).

Among these, the Random Forest Permutation Importance Measure (RFPIM), originally introduced by Breiman (2001) and implemented in widely used software frameworks such as `ranger` (Wright and Ziegler, 2017) or `randomForestSRC` (Ishwaran and Lu, 2019), has become one of the most popular approaches. Despite its practical relevance, the inferential properties of permutation-based feature importance remain poorly understood. An important step in mathematical research for RFPIM was made by Gregorutti et al. (2017), who formalized theoretical Feature Importance, which is estimated by RFPIM, and investigated the influence of correlation between features on RFPIM. Later, Ramosaj and Pauly (2023) showed that RFPIM is an asymptotically unbiased and a (weakly) consistent estimator for the theoretical Variable Importance. For the asymptotic unbiasedness, the consistency of the Random Forest estimator is needed, which was established by Scornet et al. (2015) for additive and continuous regression functions. The consistency of the RFPIM requires the stronger assumption that the trees themselves are already consistent. Nevertheless, under i.i.d. zero-mean bounded errors, a bounded regression function, the Sufficient Impurity Decrease Condition, and some other technical assumptions, the trees are already consistent (Mazumder and Wang, 2023).

Ishwaran and Lu (2019) used RFPIM to create confidence intervals for the feature importance and allowing for a natural way of feature selection with Random Forests. While the empirical results looked promising, a formal proof of the validity of the confidence intervals remains an open task. One major challenge to close this theoretical

gap is to find the asymptotic distribution of RFPIM. Ishwaran and Lu (2019) assumed RFPIM to be asymptotically normally distributed, backed this up with simulations, and hoped for future proofs of the limiting distribution. This is where the first paper of this thesis comes in. Earlier theoretical works such as Mentch and Hooker (2016) and Wager and Athey (2018) provide asymptotic results focusing on prediction or average treatment effects rather than feature importance measures. Mentch and Hooker (2016) used generalized U -statistics theory to establish asymptotic normality of Random Forest Predictions, which was later refined for more common choices of hyperparameters by Peng et al. (2022).

In the first paper of this thesis, we use the theory for U -statistics from Peng et al. (2022) to establish a first Central Limit Theorem (CLT) for the RFPIM. While most assumptions (e.g., mutually independent features and additive regression function) are common in Random Forest Theory (Scornet et al., 2015; Ramosaj and Pauly, 2023), some more restrictive assumptions like bounded error terms or the randomly chosen number of trees are needed to prove the CLT.

Additionally, the first paper contains a simulation study that suggests that the assumptions used in the proof can be relaxed regarding the bounded errors and the additivity of the regression function. Larger deviations from the normal distribution occurred only when variables with interaction effects, but without marginal effects, were present. This is in line with Blum et al. (2025), who found that the Random Forest has problems capturing such interaction effects. On the other hand, polynomial or trigonometric regression functions didn't disturb the normality of the RFPIM in our simulation.

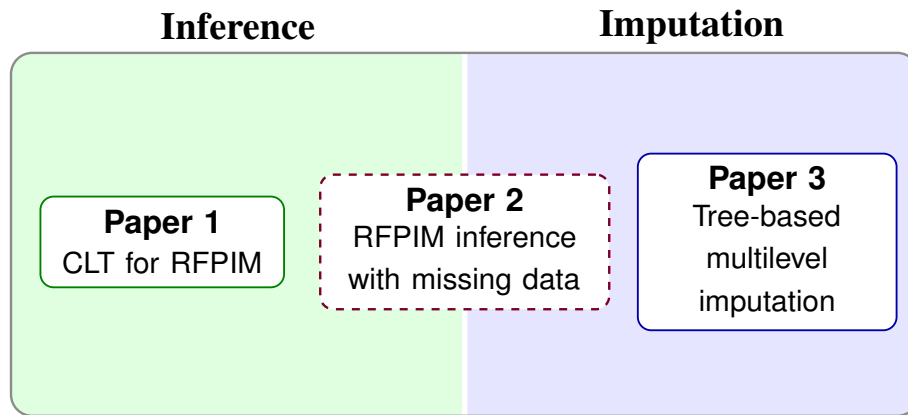
Our results contain the first theoretical guarantees for the asymptotic distribution of RFPIM, which proves the validity of the confidence intervals from Ishwaran and Lu (2019). In real-life applications of those confidence intervals, practitioners are often confronted with missing values due to non-response, technical failures, or other issues. This leads to the second gap in the research, which is closed by this thesis. The performance of the confidence intervals from Ishwaran and Lu (2019) in the case of missing values hasn't been investigated before. An obvious solution for addressing missing values is imputing the missing data with some classical approach like Multiple Imputation by Chained Equations (MICE) (Van Buuren and Groothuis-Oudshoorn, 2011). Alternatively, tree-based methods like missForest (Stekhoven, 2012) implemented in `missranger` (Mayer, 2024) or Multiple Imputation by XGBoost (Deng, 2023) implemented in `mixgb` can be used. While these methods aim for 'proper' imputation, i.e., conserving the distributional properties of the original data sets, it is not guaranteed that datasets with imputed data also lead to the same coverage rates of the confidence intervals for RFPIM.

The second paper of this thesis therefore provides an extensive simulation study comparing the mentioned imputation methods in this setting. It turns out that Ishwaran’s confidence intervals for RFPIM in combination with imputation lead to severe under-coverage in case of missing values. A heuristic fix for this issue is also provided in the second paper. By applying Rubin’s rules (Rubin, 1987) to the variance estimator proposed by Ishwaran and Lu (2019), the coverage rates can be improved substantially as long as missing rates are moderate ($\leq 30\%$).

Despite the usage of non-linear regression functions, MICE turned out to be competitive against its tree-based imputation alternatives. Nevertheless, tree-based methods for imputation might have an advantage over MICE, which has to assume some kind of underlying model, while tree-based methods are more flexible in that way. In simulations Kokla et al. (2019) found Random Forest-based imputation to be superior on real-life medical data in terms of normalized root mean squared error, when the missing values are generated with missing at random (MAR) or missing completely at random (MCAR). For these missing mechanisms, Random Forest imputation also shows strong results when judging subsequent classification on simulated datasets (Buczak et al., 2023).

This raises the question of whether there are other scenarios in which tree-based imputation can be beneficial in comparison to classical methods like MICE. A common data situation in social sciences, where observations are clustered in, e.g., classes, schools, or federal states, is multilevel data. So far, there has been no research on how tree-based imputation methods perform on multilevel data. This is where the third paper of this thesis contributes. The paper contains an extensive simulation study comparing the performance of the tests in the subsequent multilevel regression models for Random Forest imputation, MICE, and Gradient Boosting. Since the two tree-based approaches are not built for a multilevel structure, we also introduced modified versions via cluster dummy variables. The analysis contains bias and standard deviation of the regression estimators as well as the power of the associated t -test. It turns out that MICE performs best due to its conformity with the linear structure. The dummy adaptation did not make a huge difference for the tree-based methods. Although MICE is favored by the study structure, `mixgb` was able to outperform MICE when missing rates increased from 10% to 50%, underlying the potential of tree-based methods.

Uncertainty Quantification in Random Forests



This dissertation is structured as follows: Chapter 2 outlines the methodological foundation. Chapter 3 offers a summary of each of the three publications, emphasizing the innovative elements of every manuscript and its impact on the corresponding research area. The work ends with a discussion in Chapter 4, followed by the complete papers.

2 Statistical Methods

We consider a (non-)parametric regression model with an arbitrary measurable regression function $\tilde{m} : [0, 1]^p \rightarrow \mathbb{R}$ and a random error term ε with $\mathbb{E}[\varepsilon] = 0$ and $\text{Var}(\varepsilon) = \sigma^2 > 0$. Note that regression trees (and the Random Forest) are invariant under strict monotonic transformations of the features because they rely solely on the ordering of feature values for splitting, which is preserved under any monotonic transformation (Ramosaj and Pauly, 2023). Hence, the feature space of $[0, 1]^p$ is not a true restriction. For the response variable $Y \in \mathbb{R}$ and a feature vector $\mathbf{X} = (X^{(1)}, \dots, X^{(p)})$, we consider the model equation

$$Y = \tilde{m}(\mathbf{X}) + \varepsilon \quad (2.1)$$

with $\tilde{m}(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. We define $\mathcal{D}_n = \left\{ [\mathbf{X}_i^T, Y_i]^T \in [0, 1]^p \times \mathbb{R} : i = 1, \dots, n \right\}$ to be a dataset of independent and identically distributed copies of $(\mathbf{X}, Y)$. During the training of a Random Forest, two sources of randomness are involved: bootstrap sampling and feature subsampling. These can be represented by random vectors $\Theta_1, \dots, \Theta_M$, where each Θ_m determines the construction of the t -th tree in a forest of size $M \in \mathbb{N}$. The tree prediction for $\mathbf{x} \in [0, 1]^p$ of the t -th regression tree, $t \in \{1, \dots, M\}$, is denoted by $m_{n,1}(\mathbf{x}, \Theta_t, \mathcal{D}_n)$ and the corresponding Random Forest prediction by $m_{n,M}(\mathbf{x}, \Theta_1, \dots, \Theta_M, \mathcal{D}_n)$. In the original Random Forest introduced by Breiman (2001), Classification And Regression Trees (CART) are used. While CART can be used for both classification and regression, this thesis only focuses on the regression part.

2.1 CART splitting and Sufficient Impurity Decrease

For theoretical work with Random Forest, a closer look at the splitting mechanism is necessary. The CART algorithm starts with the whole feature space $[0, 1]^p$ and continues with recursive partitioning. Let $A \subset [0, 1]^p$ be a rectangle at a node, that is meant to be

split in two subsets

$$\begin{aligned} A_L &= A_L(j, b) := A \cap \{v \in [0, 1]^p : v^{(j)} \leq b\}, \\ A_R &= A_R(j, b) := A \cap \{v \in [0, 1]^p : v^{(j)} > b\}, \end{aligned}$$

for $b \in [0, 1]$, where $v^{(j)}$ denotes the j -th component of v . The CART algorithm chooses the split as

$$(\hat{j}, \hat{b}) \in \operatorname{argmax}_{j \in \{1, \dots, p\}; b \in [0, 1]} \hat{\Delta}(A, j, b),$$

where

$$\hat{\Delta}(A, j, b) = \frac{1}{N_n(A)} \sum_{i=1}^n \left\{ (Y_i - \bar{Y}_A)^2 - \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\{X_i^{(j)} \leq b\}} - \bar{Y}_{A_R} \mathbb{1}_{\{X_i^{(j)} > b\}} \right)^2 \right\} \mathbb{1}_{\{\mathbf{X}_i \in A\}}$$

is the CART splitting criterion, $\bar{Y}_A$ the average over all Y_i with feature vector falling into A (for A_L and A_R analogous) and $N_n(A)$ the number of observations in A (Breiman et al., 1984). Note, that $\operatorname{argmax}_{j \in \{1, \dots, p\}; b \in [0, 1]} \hat{\Delta}(A, j, b)$ is not necessarily unique.

The theoretical version of the CART splitting criterion is needed to later define the so-called 'Sufficient Impurity Decrease' (SID) condition and is introduced and used by Scornet et al. (2015) to formally prove L^2 -consistency of the Random Forest. Let

$$\begin{aligned} \Delta(A, j, b) &:= \mathbb{P}(\mathbf{X} \in A) \mathbb{V}\text{ar}(\tilde{m}(\mathbf{X}) | \mathbf{X} \in A) - \mathbb{P}(\mathbf{X} \in A_L) \mathbb{V}\text{ar}(\tilde{m}(\mathbf{X}) | \mathbf{X} \in A_L) \\ &\quad - \mathbb{P}(\mathbf{X} \in A_R) \mathbb{V}\text{ar}(\tilde{m}(\mathbf{X}) | \mathbf{X} \in A_R). \end{aligned}$$

denote the corresponding population version of the CART splitting criterion. To determine the rate of convergence of CART to the true regression function $\tilde{m}$, the SID condition is needed (Mazumder and Wang, 2023). For the definition of the SID condition, we need a constant $\lambda \in (0, 1)$ such that for all hyperrectangles $A \subset [0, 1]^p$

$$\sup_{1 \leq j \leq p, b \in [0, 1]} \Delta(A, j, b) \geq \lambda \cdot \mathbb{P}(\mathbf{X} \in A) \mathbb{V}\text{ar}(\tilde{m}(\mathbf{X}) | \mathbf{X} \in A). \quad (2.2)$$

Note that (2.2) is a condition on the regression function and the distribution of $\mathbf{X}$. The Sufficient Impurity Decrease (SID) condition can be interpreted as requiring a uniform lower bound on the amount of information (or impurity reduction) that can be achieved by an optimal split. Whenever a rectangle A still exhibits variation in the regression function, the SID condition enforces that at least a fixed proportion of this variation must be recoverable by some split. Mazumder and Wang (2023) showed that 2.2 is fulfilled for strictly increasing functions, smooth and strictly convex functions, and polynomials with

fixed degree, if $\mathbf{X}$ is uniformly distributed. Moreover, (2.2) is fulfilled if $\tilde{m}$ is additive (i.e. there are functions $m_1, \dots, m_p$ with $\tilde{m}(\mathbf{X}) = \sum_{j=1}^p m_j(X^{(j)})$) and the component functions fulfill (2.2). This requirement can be further relaxed (Mazumder and Wang, 2023): it suffices that each component function is defined piecewise, with every piece belonging to one of the aforementioned function classes (strongly increasing, smooth strongly convex, or fixed-degree polynomial).

2.2 Feature Importance

We are interested in quantifying the influence of the p features on Y in the setting (2.1). A theoretical measure of feature importance for $j \in \{1, \dots, p\}$ was defined by Gregorutti et al. (2017) as

$$I(j) := \mathbb{E} \left[\left(Y_1 - \tilde{m}(\mathbf{X}_{j,1}) \right)^2 \right] - \sigma^2,$$

where $\mathbf{X}_{j,1} = [X_1^{(1)}, \dots, X_1^{(j-1)}, X^{(j)}, X_1^{(j+1)}, \dots, X_1^{(p)}]$ and $X^{(j)}$ is an independent copy of $X_1^{(j)}$ (Gregorutti et al., 2017), i.e. $I(j)$ quantifies the increase of the mean squared error, after the influence of the j -th feature is deleted.

The set of observation indices $\mathcal{D}_n^{-(t)}(\Theta_t)$ that are not utilized in training the t -th tree is referred to as the out-of-bag (OOB) sample for that tree. This provides an internal validation mechanism, which can also be employed to (consistently) estimate $I(j)$ by quantifying the increase in mean squared error on the OOB sample after permuting its values along the j -th feature ((Ramosaj and Pauly, 2023)). Let $\mathbf{X}_i^{\pi_{j,t}} = (X_i^{(1)}, \dots, X_i^{(j-1)}, X_{\pi_{j,t}(i)}^{(j)}, X_i^{(j+1)}, \dots, X_i^{(p)})^\top$, where $\pi_{j,t}$ denotes a permutation of all observations in $\mathcal{D}_n^{-(t)}(\Theta_t)$. The Random Forest Permutation Importance Measure (RFPIM) for feature j is then defined as

$$\begin{aligned} I_{n,M}^{OOB}(j) \\ = \frac{1}{M\gamma_n} \sum_{t=1}^M \sum_{i \in \mathcal{D}_n^{-(t)}(\Theta_t)} \left\{ (Y_i - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \Theta_t, \mathcal{D}_n))^2 - (Y_i - m_{n,1}(\mathbf{X}_i, \Theta_t, \mathcal{D}_n))^2 \right\} \end{aligned} \quad (2.3)$$

for $j \in \{1, \dots, p\}$, where $\gamma_n = |\mathcal{D}_n^{-(t)}(\Theta_t)|$. We assume that sampling is performed without replacement (a_n times) in the bootstrap step, $\gamma_n = n - a_n$ remains constant.

Ramosaj and Pauly (2023) proved that $I_{n,M}^{OOB}(j)$ is asymptotically unbiased as $n \rightarrow \infty$ and $M \rightarrow \infty$ if there is at least one informative variable, the features are mutually

independent, and the following assumptions are fulfilled:

- (A1) We only consider permutations $\pi_{j;t} \in \mathcal{V}_n = \{\pi \in \mathcal{S}_{\gamma_n} : \pi(i) \neq i \text{ for all } i\}$, where $\mathcal{S}_{\gamma_n}$ is the symmetric group.
- (A2) $\tilde{m}$ is an additive function, i.e. there are functions $m_j : [0, 1] \rightarrow \mathbb{R}$, $j = 1, \dots, p$ such that $\tilde{m}(\mathbf{X}) = \sum_{j=1}^p m_j(X^{(j)})$.
- (A3) $\tilde{m}$ is bounded, i.e. there is a $K > 0$, such that $\sup_{\mathbf{x} \in [0,1]^p} |\tilde{m}(\mathbf{x})| \leq K$.

and

$$\lim_{n \rightarrow \infty} \lim_{M \rightarrow \infty} \mathbb{E} [(\tilde{m}(\mathbf{X}) - m_{n,M}(\mathbf{x}, \boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_M, \mathcal{D}_n))^2] = 0. \quad (2.4)$$

For the Random Forest from Breiman (2001), the consistency condition (2.4) is fulfilled under (A2) and some regularity conditions, which was formally proven by Scornet et al. (2015). Ramosaj and Pauly (2023) also showed (weak) consistency for the RFPIM, if instead of (2.4) the stronger condition

$$\lim_{n \rightarrow \infty} \mathbb{E} [(\tilde{m}(\mathbf{X}) - m_{n,1}(\mathbf{X}, \boldsymbol{\Theta}_1, \mathcal{D}_n))^2] = 0$$

is fulfilled, i.e., for consistency of RFPIM, the trees themselves need to be consistent. Intuitively speaking, in (2.3), the errors are measured by the trees and not the ensemble. Thus the trees themselves have to approximate the true regression function sufficiently well. While (A1), (A2), and (A3) are all needed for the consistency of RFPIM (Ramosaj and Pauly, 2023), (A2) and (A3) are also crucial assumptions for the consistency proof of Scornet et al. (2015).

For uncertainty quantification, we also need the variance. Computing $\text{Var}(I_{n,M}^{OOB}(j))$ on the other hand is more difficult because of the complex dependency structure. Ishwaran and Lu (2019) recommend estimating $\text{Var}(I_{n,M}^{OOB}(j))$ with a delete- d Jackknife approach. The delete- d Jackknife variance estimator with K subsampling repetitions is denoted by $v_{d,K}^2$. Based on the conjecture

$$\frac{I_{n,M}^{OOB}(j) - I(j)}{\sqrt{v_{d,K}^2}} \xrightarrow[n, M \rightarrow \infty]{d} \mathcal{N}(0, 1) \quad (2.5)$$

Ishwaran and Lu (2019) derived and recommended the approximate $(1 - \alpha)$ -confidence

interval for $I(j)$

$$\left[I_{n,M}^{OOB}(j) - z_{1-\alpha/2} \sqrt{v_{d,K}^2}, I_{n,M}^{OOB}(j) + z_{1-\alpha/2} \sqrt{v_{d,K}^2} \right], \quad (2.6)$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of the standard normal distribution. The estimator $v_{d,K}^2$ represented a good compromise between computational cost and performance during simulations according to Ishwaran and Lu (2019). Note that (2.5) is just a conjecture, and hence (2.6) is just a heuristic confidence interval and there are no mathematical guarantees that it is actually a confidence interval of asymptotic level $(1 - \alpha)$. Nevertheless, the confidence interval is implemented in the widely ¹ used R package `randomForestSRC` (Ishwaran and Kogalur, 2025). This motivates the following section, where we state a Central Limit Theorem for RFPIM under strong assumptions using U -statistic theory.

2.3 Generalized U -statistics and asymptotics of RFPIM

U -statistics form a broad class of estimators that arise naturally in many statistical problems and are defined as averages of a symmetric kernel applied to subsamples of the data. They possess desirable properties such as unbiasedness and, under mild conditions, consistency and asymptotic normality (Lee, 2019). In the context of the asymptotic distribution of the random forest permutation importance measure (RFPIM), this framework provides a way to derive its limiting behavior. For ease of notation define $\mathbf{Z}_i := (\mathbf{X}_i^T, Y_i)^T$, $i = 1, \dots, n$. Further, let h_{γ_n} be a permutation symmetric function such that $h_{\gamma_n}(\mathbf{Z}_1, \dots, \mathbf{Z}_{\gamma_n})$ is an unbiased estimator for some parameter of interest. The corresponding U -statistic is given by

$$U_{n,\gamma_n} = \binom{n}{\gamma_n}^{-1} \sum_{(t)} h_{\gamma_n}(\mathbf{Z}_{i_1}, \dots, \mathbf{Z}_{i_{\gamma_n}}), \quad (2.7)$$

where (t) denotes the summation over all subsamples of $\{1, \dots, n\}$ of size γ_n (Halmos, 1946; Hoeffding, 1992).

In the Random Forest setting, both the predictor and the RFPIM can be interpreted as averages over randomly drawn subsamples of the data. To account for the additional randomness induced by subsampling or permutation, we consider the generalized U -statistics framework of Peng et al. (2022), where subsamples are selected via i.i.d.

¹until March 2026 the `randomForestSRC` package has been downloaded more than 900'000 times.

Bernoulli random variables $(\rho_t)_t$, and the kernel includes additional randomness. This leads to the definition of the incomplete generalized U -statistics

$$U_{n,\gamma_n,M,\omega} = \frac{1}{\widehat{M}} \sum_{(t)} \rho_t h_{\gamma_n}(\mathbf{Z}_{t1}, \dots, \mathbf{Z}_{t\gamma_n}; \omega_t), \quad (2.8)$$

where ω_t denotes additional randomness in the kernel independent from $\mathcal{D}_n$ and ρ_t . Further we have $\widehat{M} := \sum_t \rho_t$. We refer to $U_{n,\gamma_n,M,\omega}$ as 'incomplete', because we choose only $\widehat{M}$ out of all possible $\binom{n}{\gamma_n}$ subsets of size γ_n . The $\widehat{M}$ is a result of the i.i.d. assumption of the Bernoulli variables ρ_t . Choosing a fixed M would induce dependencies that would invalidate the proofs in Peng et al. (2022).

In the first paper of this thesis, we also use generalized U -statistics to establish a Central Limit Theorem for RFPIM, which is a first step towards the validation of (2.6). To this end, Theorem 2 in Peng et al. (2022) is required.

Theorem 2.3.1 (Theorem 2 in Peng et al. (2022)). *Let $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ be i.i.d. from $F_{\mathbf{Z}}$ and $U_{n,\gamma_n,M,\pi}$ be a generalized incomplete U -statistic with kernel $h_{\gamma_n}(\mathbf{Z}_1, \dots, \mathbf{Z}_n; \pi_j)$. Let $\mu = \mathbb{E}[h_{\gamma_n}]$, $\zeta_{1,\gamma_n} = \text{Var}(\mathbb{E}[h_{\gamma_n}|\mathbf{Z}_1])$ and $\zeta_{\gamma_n} = \text{Var}(h_{\gamma_n})$. Suppose that*

$$\frac{\mathbb{E}[|h_{\gamma_n} - \mu|^{2k}]}{\mathbb{E}[|h_{\gamma_n} - \mu|^k]^2}$$

is uniformly bounded for all γ_n for $k = 2, 3$. If $\frac{\gamma_n}{n} \frac{\zeta_{\gamma_n}}{\gamma_n \zeta_{1,\gamma_n}} \rightarrow 0$, then

$$\frac{U_{n,\gamma_n,M,\pi} - \mu}{\sqrt{\gamma_n^2 \zeta_{1,\gamma_n}/n + \zeta_{\gamma_n}/M}} \xrightarrow{d} \mathcal{N}(0, 1)$$

as $M \rightarrow \infty$ and $n \rightarrow \infty$.

The RFPIM (2.3) is not a generalized U -statistic in the sense of (2.8), since in $I_{n,M}^{OOB}(j)$ each summand uses both the bootstrap sample (for the training of $m_{n,1}$) and the OOB-sample for error evaluation. If the trees are exchanged with the true regression function $\tilde{m}$, we achieve a simplified RFPIM version

$$\tilde{I}_{n,M}^{OOB}(j) = \frac{1}{M} \sum_{t=1}^M \frac{1}{\gamma_n} \sum_{i \in \mathcal{D}_n^-(t)} \left\{ (Y_i - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 - (Y_i - \tilde{m}(\mathbf{X}_i))^2 \right\}. \quad (2.9)$$

In the first paper of this thesis we developed a Central Limit Theorem for $\tilde{I}_{n,\widehat{M}}^{OOB}(j)$, where $\widehat{M}$ is a randomly chosen number of trees, and use this to establish asymptotic normality

of $I_{n,\widehat{M}}^{OOB}(j)$. To apply Theorem 2.3.1 to $\widetilde{I}_{n,M}^{OOB}(j)$ and get normality, assumption (A1)-(A3) as well as

(A4) The error terms have finite fourth moments, i.e. $\mathbb{E}[\varepsilon_1^4] < \infty$.

(A5) $\widehat{M} \sim \text{Bin}\left(\binom{n}{\gamma_n}, M/\binom{n}{\gamma_n}\right)$ is chosen at random.

(A6) The features are mutually independent.

are needed, where the randomly chosen number of trees (A5) is necessary to fit into the generalized incomplete U -statistic framework. The assumption (A5) is quite unusual, since the number of trees in a Random Forest is usually chosen as a fixed hyperparameter. Our model just allows us to choose the expected number of trees M . Since the number of trees is not too small, the effect of M on the predictive performance is negligible (Probst and Boulesteix, 2018). Hence, choosing the number of trees via $\text{Bin}\left(\binom{n}{\gamma_n}, M/\binom{n}{\gamma_n}\right)$ should not make a huge difference in practice.

2.4 Missing Mechanisms and Imputation Methods

So far, we assumed a model with completely observed data, which is not always the case for real-life data. For $i = 1, \dots, n$ and $j = 1, \dots, p$ let $Q_{ij} = 1$ if $X_i^{(j)}$ is missing and $Q_{ij} = 0$ else and for $j = 0$, let $Q_{i0} = 1$, if Y_i is missing and $Q_{i0} = 0$ else. The matrix $\mathbf{Q} = (Q_{ij})_{i=1,\dots,n; j=0,1,\dots,p} \in \mathbb{R}^{n \times (p+1)}$ contains all random missing indicators. Further, we denote the observed data by Z_{obs} , and the missing data as Z_{mis} . The missing data can be induced by: Missing Completely at Random (MCAR), Missing at Random (MAR) or Missing not at Random (MNAR). If ψ denotes all parameters of the missing data model (e.g. overall missing probability), the data is MCAR if

$$\mathbb{P}(Q_{ij} = 0 \mid Z_{\text{obs}}, Z_{\text{mis}}, \psi) = \mathbb{P}(Q_{ij} = 0 \mid \psi),$$

for $i = 1, \dots, n$ and $j = 0, \dots, p$ i.e., if the missing mechanism is independent of the data. When the missingness depends on the observed values, but not on the missing values, i.e.

$$\mathbb{P}(Q_{ij} = 0 \mid Z_{\text{obs}}, Z_{\text{mis}}, \psi) = \mathbb{P}(Q_{ij} = 0 \mid Z_{\text{obs}}, \psi),$$

for $i = 1, \dots, n$ and $j = 0, \dots, p$, the data is MAR. When the missingness additionally depends on Z_{mis} , the data is MNAR (Van Buuren, 2018).

In the second paper of this thesis, we are especially interested in the performance of the interval (2.6) in the presence of missing values. There are several algorithms to impute

missing data. We focus on three popular algorithms to impute missing values. For all of the three methods, we used Predictive Mean Matching (PMM). For PMM a number of donors `pmm.k` has to be specified. The donors of a prediction for a missing entry are the `pmm.k` observations, which are the nearest neighbors of the current prediction for the missing entry.

1. MICE

The first approach is Multiple Imputation by Chained Equations (MICE). Multiple Imputation means that R different imputations are generated from which the final imputation is computed. Let $\mathbf{Z}_i^{-(j)}$ be the vector $\mathbf{Z}_i$ without the j -th component $Z_i^{(j)}$. The MICE algorithm operates as follows (Van Buuren, 2018):

1. Specify a model (e.g, a linear model) $\mathbb{P}\left(Z_{\text{mis}}^{(j)} \mid Z_{\text{obs}}^{(j)}, \mathbf{Z}_i^{-(j)}, \mathbf{Q}\right)$ for $Z_i^{(j)}$ with $j = 1, \dots, p + 1$ and some model parameter β .
2. Generate initial guesses $\tilde{Z}_0^{(j)}$ with randomly drawn values from $Z_{\text{obs}}^{(j)}$ for $j = 0, \dots, p$

For $\ell = 1, \dots, L$ and $j = 0, \dots, p$:

3. Estimate the model parameter β_ℓ^j based on $\tilde{Z}_\ell^{(1)}, \dots, \tilde{Z}_\ell^{(j-1)}, \tilde{Z}_{\ell-1}^{(j+1)}, \dots, \tilde{Z}_{\ell-1}^{(p+1)}$ and $\mathbf{Q}, Z_{\text{obs}}^{(j)}$.
4. Draw $\tilde{Z}_\ell^{(j)} \sim \mathbb{P}\left(Z_{\text{mis}}^{(j)} \mid Z_{\text{obs}}^{(j)}, \tilde{Z}_\ell^{(1)}, \dots, \tilde{Z}_\ell^{(j-1)}, \tilde{Z}_{\ell-1}^{(j+1)}, \dots, \tilde{Z}_{\ell-1}^{(p+1)}, \mathbf{Q}, \beta_\ell^j\right)$.
5. **PMM Step:** Replace each drawn value from the previous step with the observed value of a randomly selected donor whose predicted value has the smallest euclidean distance to it.

6. End algorithm

The algorithm is implemented, for example, in the R package `mice` (Van Buuren and Groothuis-Oudshoorn, 2011).

2. Miss Forest

The second approach is the adaptation of the Random Forest by Stekhoven and Bühlmann (2012) for imputation tasks. The Random Forest for imputation is also called Miss Forest and has been enriched by PMM and implemented in the R package `missRanger` (Mayer, 2024). The structure of the algorithm is similar to that in MICE, but it doesn't rely on parametric assumptions like MICE. Before the algorithm starts, the features are sorted by number of missing values, setting the variable with the lowest missing rate to $Z^{(1)}$ and so on. The Miss Forest algorithm (in the version of Mayer (2024)) operates as follows

1. Generate initial guesses $\tilde{Z}_0^{(j)}$ for all j by simple imputation (e.g. randomly drawing values from $Z_{\text{obs}}^{(j)}$ or using mean/mode imputation).
2. Set $k = 1$
3. While stop criterion γ is not fulfilled

For $j = 0, \dots, p$:

4. Fit a Random Forest to the observed pairs

$$(Z_{\text{obs}}^{(j)}, \tilde{Z}_\ell^{(1)}, \dots, \tilde{Z}_\ell^{(j-1)}, \tilde{Z}_{\ell-1}^{(j+1)}, \dots, \tilde{Z}_{\ell-1}^{(p+1)}).$$

with $\tilde{Z}_\ell^{(j)}$ as target variable.

5. Predict the missing values and save them as $\tilde{Z}_\ell^{(j)}$ using the previously trained Random Forest.
6. For each missing entry, replace the predicted value from step 6 with a randomly drawn observed value from a donor whose own predicted value has the smallest euclidean distance to it.
7. Check γ
8. Set $\ell \rightarrow \ell + 1$
9. End algorithm

where the stop criterion γ is set to break the loop as soon as the difference between the newly imputed data matrix and the previous one increases for the first time.

3. Multiple Imputation via XGBoost Analogously to the Random Forest, the Multiple Imputation eXtreme Gradient Boosting (MIXGB) algorithm does not rely on any parametric assumptions, such as a linear model, since piecewise constant tree predictions are used. Unlike Random Forest, XGBoost algorithm (for details see Chen and Guestrin (2016)) doesn't generate the trees independently. XGBoost generates the trees sequentially, where each tree tries to correct the errors of the previous one. For the imputation, the features are sorted by the number of missing values, as they are for MissForest. The imputation algorithm itself, on the other hand, is very similar to MICE, with the main difference in how the predictions are generated. For MICE, a specified model generates the new imputed values, while MIXGB uses XGBoost. The algorithm (Deng, 2023) operates as follows:

1. Specify a (bootstrapped ensemble) XGBoost model $\mathbb{P} \left(Z_{\text{mis}}^{(j)} \mid Z_{\text{obs}}^{(j)}, \mathbf{Z}_i^{-(j)}, \mathbf{R} \right)$ for $Z^{(j)}$ with $j = 1, \dots, p + 1$.

2. Generate initial guesses $\tilde{Z}_0^{(j)}$ with randomly drawn values from $\mathbf{Z}_{\text{obs}}^{(j)}$ for $j = 1, \dots, p + 1$.
 - For $\ell = 1, \dots, L$ and $j = 1, \dots, p + 1$
 3. Estimate the XGBoost model (i.e. train an internal ensemble of bootstrapped trees) based on $\tilde{Z}_\ell^{(1)}, \dots, \tilde{Z}_\ell^{(j-1)}, \tilde{Z}_{\ell-1}^{(j+1)}, \dots, \tilde{Z}_{\ell-1}^{(p+1)}$ and $\mathbf{Q}, \mathbf{Z}_{\text{obs}}^{(j)}$.
 4. Draw $\tilde{Z}_\ell^{(j)} \sim \mathbb{P} \left(Z_{\text{mis}}^{(j)} \mid Z_{\text{obs}}^{(j)}, \tilde{Z}_\ell^{(1)}, \dots, \tilde{Z}_\ell^{(j-1)}, \tilde{Z}_{\ell-1}^{(j+1)}, \dots, \tilde{Z}_{\ell-1}^{(p+1)}, \mathbf{Q} \right)$ where the distribution $\mathbb{P}$ is the predictive distribution of the bootstrapped XGBoost ensemble
 5. **PMM Step:** Replace each drawn value from the previous step by the observed value of a randomly selected donor whose predicted value has the smallest euclidean distance to it.
6. End algorithm

The algorithm is implemented in the R package `mixgb` Deng (2023).

MICE and MIXGB repeat the described procedure multiple times and use the average of the test statistics on the imputed datasets as final test statistics. Because of this, these two methods are considered 'Multiple Imputation' methods. MissForest, on the other hand, as described, is a single imputation method, since only one complete dataset is generated (Stekhoven, 2012).

In the second paper of this thesis, it became clear that simply applying the imputation methods and their implementations `mice`, `missRanger`, and `mixgb` from the packages of the same name (Van Buuren and Groothuis-Oudshoorn, 2011; Mayer, 2024; Deng, 2023) to the data is problematic. Using the confidence intervals (2.6) afterwards leads to severe undercoverage. Hence some adaptation is needed, which we have done in a heuristic approach by applying Rubin's rules (Rubin, 1987, 2018). Let $R \in \mathbb{N}$ be the number of complete datasets for the Multiple Imputation Methods. On each of these, we run the Random Forest and compute the delete- d Jackknife estimator, denoted by $v_{d,K,r}^2$ ($r = 1, \dots, R$), and the Permutation Importance, denoted by $I_{n,M}^{OOB}(j, \mathcal{D}_n^r)$, where $\mathcal{D}_n^r$ is the r -th complete dataset and K is the number of subsamples used to compute the delete- d Jackknife estimator. Following Rubin's rules, the estimates can be combined by averaging

$$\overline{I_{n,M}^{OOB}}(j) = \frac{1}{R} \sum_{r=1}^R I_{n,M}^{OOB}(j, \mathcal{D}_n^r).$$

For the combined variance estimation, we must account for both the within-imputation

variance and the between-imputation variance. The within-imputation variance can also be estimated by averaging

$$\bar{v}_{d,K}^2 = \frac{1}{R} \sum_{r=1}^R v_{d,K,r}^2$$

while an estimator for the between-imputation variance is given by

$$B = \frac{1}{R-1} \sum_{r=1}^R \left(I_{n,M}^{OOB}(j, \mathcal{D}_n^r) - \overline{I_{n,M}^{OOB}}(j) \right)^2.$$

This leads to an estimator for the total variance

$$v_{d,K,\text{Rubin}}^2 = \bar{v}_{d,K}^2 + \left(1 + \frac{1}{R} \right) B.$$

Following Rubin's rules, the resulting confidence intervals for the theoretical feature importance scores $I(j)$ are given by

$$\left[\overline{I_{n,M}^{OOB}}(j) - z_{1-\alpha/2} \sqrt{v_{d,K,\text{Rubin}}^2}, \overline{I_{n,M}^{OOB}}(j) + z_{1-\alpha/2} \sqrt{v_{d,K,\text{Rubin}}^2} \right]. \quad (2.10)$$

In an extensive simulation study, we compared the interval (2.10) for `mice` and `mixgb` with (2.6) on data that includes imputed observation by `mice`, `missRanger` and `mixgb` in the second paper of this thesis. The results show that the adaptation with Rubin's rules manages to prevent the previously mentioned problems with under-coverage (see 3.2).

Imputations in multilevel data

The confidence intervals for feature importance are not the only interesting use case for `mice`, `missRanger`, and `mixgb`. While `mice` has been adapted to multilevel data and is widely used in this context, there has been a research gap regarding tree-based methods, which is explored in the third paper of this thesis.

Multilevel data is a very common phenomenon in social sciences, where observations follow a hierarchical structure. This occurs, for example, if the data contains students (level 1) from the same class, school, or region (level 2), which induces a complex structure of dependencies to the data. For these clustered observations, there are often in-cluster and between-cluster effects, and random slopes allow for varying effects of level 1 variables in different clusters. Hierarchical linear models are able to deal with

these problems (Hox, 1998; Hox and Roberts, 2010) and are easily interpretable for practitioners. Since practitioners are primarily interested in valid inference within such models, it is natural to assess imputation methods based on their impact on subsequent hierarchical analyses as we do in the third paper of this thesis. Thus, we evaluate the performance of the imputations based on their impact of the bias and standard deviation of the parameter estimation and the power of the tests in hierarchical linear models, because that is suitable as a take-home message for social scientists and other practitioners.

As a simple approach to adapt the tree-based methods to multilevel data, we include dummy variables for the level 2 variables. Assuming there are J clusters, the dummy variable for the j -th cluster, $j = 1, \dots, J$, equals 1 if the observation belongs to the j -th cluster and is 0 else. With dummies, the multilevel structure is considered even though it causes bias in the estimated variance and tends to overestimate the effect of clusters. Drechsler (2015) pointed out these weaknesses, but on the other hand, admitted that alternatives "might be too complex" or "rely on questionable model assumptions".

3 Summary of the Articles

3.1 Article 1: A Central Limit Theorem for the Permutation Importance Measure

The article develops a first Central Limit Theorem for the Random Forest Permutation Importance Measure (2.3) under the Model (2.1) presented in Section 2. Our work is based on the generalized U -statistics theory of Peng et al. (2022). As a first step towards the asymptotic normality of RFPIM, we start with the simplified version $\tilde{I}_{n,M}^{OOB}(j)$ from (2.9) and apply Theorem 2.3.1 to get the asymptotic normality of $\tilde{I}_{n,M}^{OOB}(j)$. For this, we need the assumptions (A1)-(A6). To establish asymptotic normality, we must also ensure that we are not in the case of $\tilde{I}(j) = 0$ almost surely. To this end, we additionally assume that the component functions from (A2) fulfill

$$\mathbb{V}\text{ar}(m_j(X_1^{(j)})) > 0.$$

Indeed, if $\mathbb{V}\text{ar}(m_j(X_1^{(j)})) = 0$, then $m_j(X_1^{(j)})$ is almost surely constant. This would imply that permuting the j -th component of $\mathbf{X}_1$ leaves $\tilde{m}(\mathbf{X}_1)$ unchanged, and hence

$$\tilde{I}_{n,M}^{OOB}(j) = 0 \quad \text{almost surely.}$$

The downside of this is that our theory does not enable hypothesis testing for hypotheses like $I(j) = 0$, since the asymptotic distribution of RFPIM in this case remains unknown. With the assumptions from the previous chapters, we first prove asymptotic normality for the simplified version of RFPIM

Theorem 3.1.1 (Theorem 4.2 in Article 1). *Let (A1)-(A6) be fulfilled and $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ independent and identically distributed. If $\mathbb{V}\text{ar}(m_j(X_1^{(j)})) > 0$, $\gamma_n \rightarrow \infty$ as $n \rightarrow \infty$, and $\gamma_n/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$, then*

$$\frac{\tilde{I}_{n,M}^{OOB}(j) - I(j)}{\sqrt{\gamma_n^2 \zeta_{1,\gamma_n}/n + \zeta_{\gamma_n}/M}} \xrightarrow{d} \mathcal{N}(0, 1)$$

3 Summary of the Articles

as $M \rightarrow \infty$ and $n \rightarrow \infty$.

In the next step Theorem 3.1.1 was used to prove the asymptotic normality of the RFPIM. For this the Assumption (A4) has to be replaced by the stronger assumption

(A4*) The error terms are bounded by some $K_\varepsilon > 0$.

Furthermore we need

(A7) The Sufficient Impurity Deacresae Condition (2.2) is fulfilled.

With these assumptions we are equipped to formulate the main result of the first article

Theorem 3.1.2 (Theorem 4.3 in Article 1). *Let (A1)-(A3), (A4*), (A5)-(A7) be fulfilled and $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ independent and identically distributed and $\text{Var}\left(m_j(X_1^{(j)})\right) > 0$. If $\gamma_n \rightarrow \infty$, $\frac{M_n \gamma_n \log(a_n)^2 \log(a_n p)}{a_n^{\phi(\lambda)}} \rightarrow 0$ for $\phi(\lambda) = \frac{-\log_2(1-\lambda)}{1-\log_2(1-\lambda)}$ as $M_n \rightarrow \infty$ and $n \rightarrow \infty$. Then*

$$\frac{I_{n,M}^{OOB}(j) - I(j)}{\sqrt{\zeta_{\gamma_n}/M_n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

as $M_n \rightarrow \infty$ and $n \rightarrow \infty$, if the trees are grown to depth $d = \lceil \log_2(n)/(1 - \log_2(1 - \lambda)) \rceil$.

In the examples for functions fulfilling (A7) provided in Mazumder and Wang (2023), the values chosen for λ are relatively small. Consequently, the target depth d is close to $\log_2 n$, which corresponds to the depth of a fully grown binary tree. The assumption

$$\frac{M_n \gamma_n \log(a_n)^2 \log(a_n p)}{a_n^{\phi(\lambda)}} \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

imposed in that work is quite restrictive and is not satisfied by typical choices such as $a_n \approx 0.632 n$. In contrast, our assumption is, for instance, fulfilled by sequences of the form

$$M_n = \lfloor (\log n)^{q_1} \rfloor \quad \text{and} \quad \gamma_n = \lfloor n - (\log n)^{q_2} \rfloor$$

for arbitrary exponents $q_1, q_2 > 0$. This is a relatively strict assumption and probably the greatest weakness of the paper, since it requires a large amount of the data to be in the bootstrap sample, which reduces tree diversity.

In addition, a simulation study was conducted to investigate the asymptotic normality of the RFPIM under violations of the theoretical assumptions. The simulation design was inspired by Ishwaran and Lu (2019) with $N = 1000$ simulation runs, sample

size $n = 2500$, and $p = 10$ uniformly distributed covariates. Random Forests were trained using the `ranger` package with $M = 500$ trees and subsampling of size $a_n = \lfloor 2/3 \cdot n \rfloor$. To assess robustness, both truncated and standard normal error distributions were considered, together with four different regression functions ranging from additive polynomial models to highly non-additive trigonometric and interaction-based settings. The functions have been taken from Ramosaj and Pauly (2023) and Ishwaran and Lu (2019). Furthermore, both ordinary permutations and derangements were investigated in the computation of the RFPIM. The results suggest that violations of (A1) and the bounded-error assumption (A4*) have little effect on the asymptotic distribution. In contrast, violations of the additivity assumption (A2) led to noticeable deviations from normality when purely, i.e., interaction effects without marginal effects, were present, a phenomenon consistent with the known difficulties of Random Forests in detecting such effects (Blum et al., 2025).

A limitation of the first paper is that the theoretical assumptions required for the CLT are stronger than those typically used in Random Forest applications. In particular, the proof relies on additive regression functions, bounded error terms, derangements instead of arbitrary permutations, and a modified Random Forest framework with a random number of trees and large bootstrap samples relative to the out-of-bag sample size. Future research should therefore aim at extending the asymptotic theory to more general non-additive regression functions and to Random Forest settings closer to practical implementations. Another important direction is the consistent estimation of the variance to construct tests or confidence intervals. For this, our result should be extended to cases where $I(j) = 0$.

3.2 Article 2: Confidence Intervals for Random Forest Permutation Importance with Missing Data

The confidence interval (2.6) has been investigated in a simulation study by Ishwaran and Lu (2019) and achieved close-to-nominal coverage when all covariates are fully observed. The approach was implemented in the popular `randomForestSRC` R package, which by March 2026 had been downloaded more than 900,000 times.

We investigate how the confidence intervals for RFPIM proposed by Ishwaran and Lu (2019) behave after classical imputation methods such as MICE, MissForest, and MIXGB are applied. In addition, we propose an adaptation motivated by Rubin’s rules to account for the additional uncertainty introduced by imputation. Through a large-scale simulation study and complementary real-data benchmarks, we evaluate

the resulting coverage probabilities under realistic conditions. For this, the interval (2.10) is adopted for the Multiple Imputation methods MICE and MIXGB. The R implementations `mice` (Van Buuren and Groothuis-Oudshoorn, 2011), `missRanger` (Mayer, 2024) and `mixgb` (Deng, 2023) are chosen for the simulation.

Twelve diverse non-linear regression functions (taken from Ishwaran and Lu, 2019) are employed, ranging from smooth trigonometric and exponential relationships to sharp threshold effects and higher-order interactions, supplemented by pure noise variables to yield a realistic $p = 20$ covariate setting. Sample sizes of $n = 100, 250$, and 500 are considered, with $n = 250$ serving as the original benchmark. Missing values are introduced under the two most common and theoretically best-understood mechanisms: missing completely at random (MCAR) and missing at random (MAR), the latter constructed such that missingness in all variables except one depends on whether the observed value of the remaining variable falls below its median. After that, the missing data is imputed by `mice`, `missRanger`, or `mixgb`. On every completed dataset, a Random Forest is trained using the `randomForestSRC` implementation with 250 trees and otherwise standard hyperparameter choices as in Ishwaran and Lu (2019). RFPIM and its delete- d jackknife variance are computed. After that, the 90% corresponding confidence interval (2.6) as well as the adopted (2.10) is constructed.

The simulation comprises 216 experimental cells (12 functions, 3 sample sizes, 2 missingness mechanisms, 3 missing rates of 10 %, 30 % and 50 %) with 1000 independent replications each. Coverage probabilities are stratified by true signal strength (noise variables, moderate-importance variables, and the top 10 % high-importance variables) because earlier work had shown that coverage can depend heavily on whether a feature is truly irrelevant or strongly predictive (Ishwaran and Lu, 2019).

The results are consistent across the vast majority of settings. Single imputation, regardless of whether performed with MICE, `missRanger`, or `mixgb`, leads to dramatic undercoverage when the imputed data are naively fed into Ishwaran’s procedure. Even at a modest missing rate of 10%, coverage frequently drops below 80%, and at 30% or 50% missingness, it regularly collapses to values between 40% and 60% for moderately important variables. Treating imputed values as observed ignores the additional uncertainty introduced by the imputation step, resulting in variance estimates that are far too small and consequently in confidence intervals that are systematically too narrow.

The proposed Rubin intervals behave very differently. Whenever the sample size is at least moderately large ($n \geq 250$) and the missing rate does not exceed 30%, they regularly achieve or even slightly exceed the nominal 90% coverage level across all feature types and both missingness mechanisms. For noise variables, the intervals

remain mildly conservative, exactly as observed on complete data, while for moderate- and high-importance variables, coverage is restored to satisfactory levels. Only under extreme conditions, i.e., either small samples ($n = 100$) or very high missingness (50 %), the Rubin intervals also lose reliability, but even then they are substantially closer to nominal coverage than the naive single-imputation approach.

Among the imputation methods themselves, some differences emerge. Multiple imputation with `mice` and with `mixgb` perform very similarly, and both are markedly superior to `missRanger` with PMM. The latter regularly produces the shortest intervals and the most severe undercoverage, sometimes approaching zero for higher missing rates. Closer inspection reveals that `missRanger`, despite faithfully reproducing marginal distributions and pairwise correlations, systematically underestimates permutation importance in the presence of pure interaction effects, a known weakness of Random Forests (Blum et al., 2025). The resulting RFPIM values are therefore too small on average, and the naive variance estimator additionally underestimates uncertainty, yielding a doubly harmful effect on coverage.

Interval length, standardised relative to the complete-data benchmark, tells also a consistent story. Rubin intervals are wider than their naive counterparts, sometimes by a factor of 1.2 to 1.5 at higher missing rates, but this additional width is precisely what is required to achieve valid coverage. Single-imputation intervals, by contrast, are often slightly shorter than the complete-data reference, which directly explains their systematic undercoverage.

A complementary benchmark study on five real-world regression datasets with artificially induced 10% MCAR missingness confirms the simulation findings in a setting where ground truth is unknown. Standardised confidence intervals (centered at zero and scaled to unit mean length on the complete data) show that Rubin intervals are more stable in location and width across repeated imputations, whereas single-imputation intervals - particularly those based on MICE - can exhibit pronounced positional shifts and occasional extreme inflation for variables with heavily skewed or zero-inflated distributions.

Taken together, standard RFPIM confidence intervals cannot be trusted after single imputation. The additional uncertainty introduced by filling in missing values is simply ignored by the delete- d jackknife estimator in its current form, leading to severely liberal inference. The remedy proposed here, combining Multiple Imputation with classical Rubin pooling, is conceptually simple, and works with existing Random Forest implementations. Under realistic conditions, from our empirical studies, it restores close-to-nominal coverage at the cost of only moderately wider intervals.

Several limitations remain and point to important directions for future research. The Rubin-pooling strategy is heuristic; no theoretical proof of its validity for tree-based ensembles currently exists. The delete- d jackknife variance estimator itself still awaits rigorous consistency proofs, and the theoretical assumptions from the first article (additive regression functions, strictly positive true importance, slowly growing number of trees) are restrictive and don't apply to the settings and hyperparameters investigated in the second article. Finally, only one specific MAR mechanism and no MNAR scenarios were considered, and the behaviour in the presence of strongly correlated predictors also remains to be explored in theory and simulation.

3.3 Article 3: Adapting tree-based multiple imputation methods for multi-level data? A simulation study

The paper examines the performance of tree-based multiple imputation methods in hierarchically structured (multilevel) data, where observations are nested (e.g., students within classes). Conventional imputation techniques often assume independence among observations, which can lead to biased inferences in multilevel contexts. Although `mice` is widely applied to such data, the algorithm is sensitive to model specifications. Modern tree-based methods like `missRanger` and `mixgb` might be a viable alternative.

This simulation study compares traditional `mice` with `missRanger` and `mixgb`, both in their standard versions and adapted to the multilevel structure by incorporating dummy variables for cluster membership. Performance is assessed in terms of coefficient bias, Type-I error rates, and statistical power following imputation, using linear mixed-effects models (random intercept and random slope models). The simulation varies key factors: number of clusters (25 and 50), missingness rates (10%, 30%, 50%), and missingness mechanisms (MCAR and MAR). Data is generated with six continuous covariates, each at Level 1 and Level 2, some predictive and some noise variables, with a total sample size of $N = 1000$.

Key findings are as follows: `mice` (using `2l.norm` for Level-1 and `2lonly.pmm` for Level-2 variables) proves robust, particularly at low missingness rates (10%), and delivers accurate inference especially for Level-2 variables across most conditions. Random Forest-based imputation (`missRanger`) without predictive mean matching (PMM) produces severely inflated Type I errors and is not recommended; even with PMM, it generally underperforms relative to MICE and `mixgb`. The adapted XGBoost method (`mixgb` with cluster dummies) consistently outperforms all other methods for Level-1 variables at higher missingness rates (30% and 50%). At 50% missingness, it

also achieves superior or comparable power for Level-2 variables, irrespective of the missingness mechanism or cluster size. Coefficient bias patterns align with power results: MICE exhibits the lowest bias for Level-2 coefficients, whereas `mixgb` (particularly with dummies) shows lower bias for Level-1 coefficients at higher missing rates. Incorporating cluster dummies generally improves or maintains the performance of tree-based methods, though it does not always yield substantial gains.

The results demonstrate that properly adapted tree-based imputation, especially `mixgb` with cluster dummies, represents a potential alternative to `mixgb` in multilevel analyses, particularly when missingness is substantial. The effect of the dummies themselves seemed to be rather small. This is notable given MICE's closer alignment with the linear data-generating process. The study highlights the potential of non-parametric tree-based approaches in hierarchical data settings and encourages further research in more complex, non-linear scenarios, where the advantages of tree-based methods in comparison to `mice` come more into play.

A limitation of the third paper is that the simulation design is closely aligned with linear mixed models and therefore favors MICE, especially at low missingness rates. The data-generating process assumes linear relationships and relatively low-dimensional settings, which may not reflect the complexity of many real-world applications involving nonlinear effects, interaction structures, or a larger number of variables. Consequently, the advantages of flexible tree-based methods such as `missRanger` and `mixgb` may be underestimated in our study. Moreover, the multilevel structures considered are limited to random intercepts and random slopes, while more complex settings such as contextual effects, cross-level interactions, or systematically missing level-1 variables were not investigated. Future research should therefore study more complex and non-linear multilevel data-generating processes, alternative tree-based imputation strategies including random-effects trees, and the performance of imputation methods with respect to confidence interval coverage and additional inferential criteria. Despite these limitations, the results demonstrate that tree-based approaches, especially `mixgb`, can become competitive or even outperform MICE at higher missingness rates, highlighting their potential for future applications in multilevel data analysis.

4 Discussion and Outlook

This thesis consists of three manuscripts covering different aspects of Random Forest inference and imputation. Especially, this thesis addresses key theoretical and practical aspects of the Random Forest Permutation Importance Measure (RFPIM). The first contribution provides the first Central Limit Theorem for RFPIM under additive regression functions, independent features, and further restrictive assumptions, including the Sufficient Impurity Decrease (SID) condition and randomized number of trees. The established asymptotic normality provides a first theoretical foundation for constructing confidence intervals and hypothesis tests for the RFPIM. However, the current result still relies on restrictive assumptions, in particular a large bootstrap sample in comparison to the OOB-sample, which reduces tree diversity and therefore limits the direct applicability to standard Random Forest settings. If this restriction can be avoided is still unknown.

The second contribution tackles a more practical challenge: the impact of missing values on the validity of the heuristic confidence intervals for Feature Importance $I(j)$ proposed by Ishwaran and Lu (2019). Through extensive simulations, it is demonstrated that naive single imputation, regardless of whether performed with `mice`, `missRanger`, or `mixgb`, leads to severe undercoverage of the Ishwaran intervals, often dropping far below nominal levels even at moderate missing rates. The additional uncertainty introduced by imputation is ignored by the delete- d jackknife variance estimator, resulting in systematically too-narrow intervals. The proposed remedy, combining multiple imputation with Rubin’s rules for pooling point estimates and variances, proves to be more effective.

The third article focuses on imputing in multilevel data, which is common in the social sciences. The simulation study compares tree-based imputations (`missRanger`, `mixgb`) with `mice` and shows that a simple adaptation via inclusion of cluster dummies improves performance. In particular, `mixgb` with dummies consistently outperforms other methods for Level-1 variables at higher missing rates (30–50%) in terms of bias, Type I error rates control, and power, while MICE remains robust for Level-2 variables. Overall, `mice` remained the benchmark for the linear multilevel models considered in the study. This is not surprising, since the simulation design closely matches the assump-

tions built into `mice`, which explicitly models the underlying linear and hierarchical structure of the data. The fact that `mixgb` was able to outperform `mice` on higher missing rates (adapted) tree-based methods as a promising, non-parametric alternative to parametric approaches like `mice`.

Overall, the three articles provide some new insights into inference and imputation with Random Forests. The Central Limit Theorem starts to give some theoretical foundations for the inference provided by Ishwaran and Lu (2019). Our adapted Rubin intervals make the inference more robust in the presence of missing values, and we gave some guidance on how to work with missings in multilevel data settings.

Outlook Future theoretical work should investigate whether the assumptions of the CLT can be weakened and whether the consistency of the delete- d Jackknife variance estimator can be established. Another open theoretical problem is to provide a rigorous justification for the confidence intervals proposed by Ishwaran and Lu (2019) and for their adaptation via Rubin’s rules in the presence of missing data.

The simulation results for multilevel data suggest that tree-based imputation methods may have greater potential in settings with nonlinear effects or more complex dependency structures, where their flexibility can become advantageous. This motivates more simulation studies for broader settings and the development and evaluation of improved tree-based imputation approaches for more general multilevel settings.

Bibliography

- Blum, R., Hiabu, M., Mammen, E., and Meyer, J. T. (2025). Pure interaction effects unseen by Random Forests. *Computational Statistics & Data Analysis*, 212:108237.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.
- Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Chapman and Hall/CRC.
- Buczak, P., Chen, J.-J., and Pauly, M. (2023). Analyzing the effect of imputation on classification performance under MCAR and MAR missing mechanisms. *Entropy*, 25(3):521.
- Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Cornell University*.
- Deng, Y. (2023). *mixgb: Multiple Imputation Through 'XGBoost'*. R package version 1.0.2.
- Drechsler, J. (2015). Multiple imputation of multilevel missing data-Rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, 40(1):69–95.
- Gregorutti, B., Michel, B., and Saint-Pierre, P. (2017). Correlation and Variable Importance in Random Forests. *Statistics and Computing*, 27(3):659–678.
- Halmos, P. R. (1946). The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1):34–43.
- Hoeffding, W. (1992). *A class of statistics with asymptotically normal distribution*. Springer.
- Hox, J. (1998). Multilevel Modeling: When and Why. In Balderjahn, I., Mathar, R., and Schader, M., editors, *Classification, Data Analysis, and Data Highways*, pages 147–154, Berlin. Springer.

- Hox, J. and Roberts, J. K. (2010). *Handbook of Advanced Multilevel Analysis*. Routledge, New York, first edition.
- Ishwaran, H. and Kogalur, U. B. (2025). Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). *R package version 3.3.3*.
- Ishwaran, H. and Lu, M. (2019). Standard Errors and Confidence Intervals for Variable Importance in Random Forest Regression, Classification, and Survival. *Statistics in Medicine*, 38(4):558–582.
- Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., and Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. *BMC Bioinformatics*, 20(1):492.
- Lee, A. J. (2019). *U-statistics: Theory and Practice*. Routledge.
- Mayer, M. (2024). *missRanger: Fast Imputation of Missing Values*. R package version 2.6.0.
- Mazumder, R. and Wang, H. (2023). On the convergence of CART under sufficient impurity decrease condition. *Advances in Neural Information Processing Systems*, 36:57754–57782.
- Mentch, L. and Hooker, G. (2016). Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. *Journal of Machine Learning Research*, 17(26):1–41.
- Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.
- Peng, W., Coleman, T., and Mentch, L. (2022). Rates of convergence for random forests via generalized U-statistics. *Electronic Journal of Statistics*, 16:232–292.
- Probst, P. and Boulesteix, A. (2018). To tune or not to tune the number of trees in Random Forest. *Journal of Machine Learning Research*, 18(181):1–18.
- Ramosaj, B. and Pauly, M. (2023). Consistent and unbiased variable selection under independent features using Random Forest permutation importance. *Bernoulli*, 29:2101–2118.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, NY.

- Rubin, D. B. (2018). Multiple imputation. In *Flexible Imputation of Missing Data, Second Edition*, pages 29–62. Chapman & Hall/CRC, Boca Raton, FL.
- Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of Random Forests. *The Annals of Statistics*, pages 1716–1741.
- Stekhoven, D. J. (2012). Using the missForest Package. missForest Nonparametric Missing Value Imput. using Random For. *R Packag. version*, 1:1–11.
- Stekhoven, D. J. and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.
- Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4):323.
- Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.
- Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45:1–67.
- Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242.
- Wright, M. N. and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Part II

Publications

Article 1

Föge, N., Schmid, L., Ditzhaus, M., & Pauly, M. (2024). *A Central Limit Theorem for the permutation importance measure*. arXiv preprint arXiv:2412.13020.

<https://doi.org/10.48550/arXiv.2412.13020>

Manuscript submitted for publication to the Journal of Statistical Planning and Inference.

A Central Limit Theorem for the Permutation Importance Measure

Nico Föge¹, Lena Schmid², Marc Ditzhaus^{†1}, and
Markus Pauly^{2,3}

¹*Institut für Mathematische Stochastik, Otto-von-Guericke Universität Magdeburg, e-mail:*
nico.foege@ovgu.de;

²*Department of Statistics, TU Dortmund University, e-mail:*
lena.schmid@tu-dortmund.de; markus.pauly@tu-dortmund.de

³*Research Center Trustworthy Data Science and Security, University Alliance Ruhr, e-mail:*
markus.pauly@tu-dortmund.de

[†]*We would like to thank the late Marc Ditzhaus, a bright mind and wonderful mentor, collaboration partner and mentee for his contribution to that paper.*

Abstract: Random Forests have become a widely used tool in machine learning since their introduction in 2001, known for their strong performance in classification and regression tasks. One key feature of Random Forests is the Random Forest Permutation Importance Measure (RFPIM), an internal, non-parametric measure of variable importance. While widely used, theoretical work on RFPIM is sparse, and most research has focused on empirical findings. However, recent progress has been made, such as establishing consistency of the RFPIM, although a mathematical analysis of its asymptotic distribution is still missing.

In this paper, we provide a formal proof of a Central Limit Theorem for RFPIM using U-statistics theory. Our approach deviates from the conventional Random Forest model by assuming a random number of trees and imposing conditions on the regression functions and error terms, which must be bounded and additive, respectively. Our result aims at improving the theoretical understanding of RFPIM rather than conducting comprehensive hypothesis testing. However, our contributions provide a solid foundation and demonstrate the potential for future work to extend to practical applications which we also highlight with a small simulation study.

MSC2020 subject classifications: Primary 60F05; secondary 62G08.

Keywords and phrases: Random Forest, Permutation Importance, Central Limit Theorem.

1. Introduction

Since their introduction in 2001, Random Forests [2] quickly gained popularity due to their robust performance in both classification and regression tasks. Over the years, their application has been extended to right-censored data [9] and the imputation of missing values [14, 21]. Despite these empirical successes, the mathematical and statistical properties of Random Forests remained poorly

understood for some time. The complex structure and data-dependent weights of Random Forests pose significant challenges for rigorous mathematical analysis.

The first studies on the consistency of Random Forests began with simplified versions [3, 4] until [18] ultimately established the L^2 -consistency of the (almost) original Random Forest algorithm introduced by [2]. Further strengthening the theoretical foundation, [12] derived a Central Limit Theorem using a U -statistics framework, while [22] established asymptotic normality and confidence intervals for tree ensembles under the "honesty" condition, also providing an upper bound on their bias.

Beyond predictive performance, Random Forests offer interpretability in the form of an internal non-parametric measure of variable importance (VIMP). Among various importance measures, we focus on the Random Forest Permutation Importance (RFPI). The RFPI leverages the Out-of-Bag (OOB) sample, which includes all observations not used in training a given tree during the bootstrap step. This allows for an internal validation mechanism. To compute RFPI, the values of the feature of interest are randomly permuted within the OOB sample, and the resulting decrease in accuracy is averaged over all trees.

In contrast to the theoretical work on Random Forests themselves, results for RFPI are sparse. However, some foundational work has been done, starting with [7], which explored the setting of correlated features. [5] advanced the theoretical understanding by proving the asymptotic unbiasedness and weak consistency of RFPI, while [10] constructed confidence intervals for RFPI assuming asymptotic normality. Although simulations supported this assumption, a formal proof of asymptotic normality is still missing. This is where our work makes an important contribution. We apply U -statistic theory from [16] and leverage convergence rates for the CART algorithm from [11] to establish and rigorously prove a Central Limit Theorem for RFPI. Thus, providing formal support for the findings of [10].

The paper is organized as follows: In Section 2, we introduce our mathematical framework and formally define Permutation Importance. Section 3 presents the necessary theory of U -statistics and connects it to RFPI. Our main results, including the asymptotic normality of RFPI are provided in Section 4. In Section 5 we contribute a simulation study investigating the practical relevance of our findings. Finally, Section 6 discusses the implications of our findings.

2. Mathematical framework

We assume a sparse and non-parametric learning problem with regression function $\tilde{m} : [0, 1]^p \rightarrow \mathbb{R}$ and model equation

$$Y = \tilde{m}(\mathbf{X}) + \varepsilon, \quad (1)$$

with a uniformly distributed feature vector $\mathbf{X} = (X^{(1)}, \dots, X^{(p)}) \in [0, 1]^p$ and some random error ε with $\mathbb{E}[\varepsilon] = 0$ and $\text{Var}(\varepsilon) = \sigma^2 > 0$. Moreover, $\mathbf{X}$ and ε are assumed to be independent. With sparsity, we mean there is a subset

of features $\mathcal{S} \subseteq \{1, \dots, p\}$ that contains all information about Y . Assuming without loss of generality that $\mathcal{S} = \{1, \dots, s\}$, this means that there is a function $m : [0, 1]^s \rightarrow \mathbb{R}$ such that $m(\mathbf{X}_{\mathcal{S}}) = \tilde{m}(\mathbf{X})$, where $\mathbf{X}_{\mathcal{S}} = (X^{(1)}, \dots, X^{(s)})$, see [5] for the same assumption. It is widely known, that Random Forests adapt well to sparse frameworks because the choice of optimal splits at each node avoids splittings on non-informative features [18].

We are interested in the "Importance" of single features for Y . To extract the importance of a single feature, we need to exclude interactions that shed the isolated influence of a feature on Y . To this end, we assume an additive structure, i.e there are functions $m_j : [0, 1] \rightarrow \mathbb{R}$, $j = 1, \dots, p$ such that

$$\tilde{m}(\mathbf{X}) = \sum_{j=1}^p m_j(X^{(j)}). \quad (2)$$

Let

$$\mathcal{D}_n = \left\{ [\mathbf{X}_i^T, Y_i]^T \in [0, 1]^p \times \mathbb{R} : i = 1, \dots, n \right\},$$

with $[\mathbf{X}_i^T, Y_i]^T$ be a training set containing i.i.d. copies of $[\mathbf{X}^T, Y]^T$ also containing the i.i.d. errors $\varepsilon_1, \dots, \varepsilon_n$. A theoretical measure for the feature importance for $j \in \{1, \dots, p\}$ is given by

$$I(j) := \mathbb{E} \left[\left(Y_1 - \tilde{m}(\mathbf{X}_{j,1}) \right)^2 \right] - \sigma^2,$$

where $\mathbf{X}_{j,1} = [X_1^{(1)}, \dots, X_1^{(j-1)}, X^{(j)}, X_1^{(j+1)}, \dots, X_1^{(p)}]$ and $X^{(j)}$ an independent copy of $X_1^{(j)}$ introduced by [7]. In informal terms, $I(j)$ quantifies the contribution of the variability of the j -th feature to the variability of the response variable Y , by measuring the increase in prediction error when the feature $X^{(j)}$ is excluded from the model. Due to our model assumptions $I(j)$ simplifies to

$$\begin{aligned} I(j) &= \mathbb{E} \left[(Y_1 - \tilde{m}(\mathbf{X}_{j,1}))^2 \right] - \sigma^2 = \mathbb{E} \left[\left(m_j(X_1^{(j)}) + \varepsilon_1 - m_j(X^{(j)}) \right)^2 \right] - \sigma^2 \\ &= \text{Var} \left(m_j(X_1^{(j)}) + \varepsilon_1 - m_j(X^{(j)}) \right) - \sigma^2 = 2\text{Var} \left(m_j(X_1^{(j)}) \right). \end{aligned} \quad (3)$$

2.1. Random Forest Permutation Importance

A Random Forest is an ensemble of tree-based learners, where each tree is trained on a subsample of $\mathcal{D}_n$ of size $a_n \in \{1, \dots, n\}$. As in [5] we restrict ourselves to sampling without replacement. Moreover at each node of each tree $m_{try} \in \{1, \dots, p\}$ features are randomly chosen as candidates for the split direction. To apply the results of [16], we adapt to their framework and draw the number of trees $\widehat{M} \in \mathbb{N}$ at random assuming $\widehat{M} \sim \text{Bin} \left(\binom{n}{n-a_n}, p_M \right)$ with $\mathbb{E}[\widehat{M}] = M$ [16]. Let Θ be the generic random vector determining the subsample

construction and the feature sub-spacing. For a Random Forest, $\widehat{M}$ independent copies $\Theta_1, \dots, \Theta_{\widehat{M}}$ of Θ are drawn independent of $\mathcal{D}_n$. The piecewise constant estimation for $\widehat{m}$ of the t -th tree, $t \in \{1, \dots, \widehat{M}\}$, is denoted by $m_{n,1}(\cdot, \Theta_t, \mathcal{D}_n)$ and the Random Forest prediction for a fixed $\mathbf{x} \in [0, 1]^p$ is defined as

$$m_{n,\widehat{M}}(\mathbf{x}, \Theta_1, \dots, \Theta_{\widehat{M}}, \mathcal{D}_n) = \frac{1}{\widehat{M}} \sum_{t=1}^{\widehat{M}} m_{n,1}(\mathbf{x}, \Theta_t, \mathcal{D}_n).$$

The subsample determined by Θ_t is denoted by $\mathcal{D}_n^{(t)}$. To provide a more formal description of the selection mechanism, we define the generating vector

$$\Theta_t = \left[\left(\Theta_t^{(1)} \right)^\top, \left(\Theta_t^{(2)} \right)^\top \right]^\top,$$

which can be decomposed into subvectors $\Theta_t^{(1)}$ and $\Theta_t^{(2)}$. The components of $\Theta_t^{(1)} = \left[\Theta_{1,t}^{(1)}, \dots, \Theta_{n,t}^{(1)} \right]^\top$ indicate whether the i -th observation, $i = 1, \dots, n$ has been selected ($\Theta_{i,t}^{(1)} = 1$) or not ($\Theta_{i,t}^{(1)} = 0$), while $\Theta_t^{(2)}$ models the feature sup-spacing. The set of observation-indices $\mathcal{D}_n^{-(t)}$ not used for the training of the t -th tree is called out-of-bag (OOB) sample of the t -th tree. This delivers an internal validation, that can also be used for the estimation of $I(j)$ by measuring the increase of the mean squared error on the OOB-sample after permuting its observations along the j -th feature ([5]).

Let $\mathbf{X}_i^{\pi_{j,t}} = (X_i^{(1)}, \dots, X_i^{(j-1)}, X_{\pi_{j,t}(i)}^{(j)}, X_i^{(j+1)}, \dots, X_i^{(p)})^\top$, where $\pi_{j,t}$ is a permutation of all observations in $\mathcal{D}_n^{-(t)}$. The Random Forest Permutation Importance Measure (RFPIM) for a feature j then is defined as

$$\begin{aligned} I_{n,\widehat{M}}^{OOB}(j) &= \frac{1}{\widehat{M}\gamma_n} \sum_{t=1}^{\widehat{M}} \sum_{i \in \mathcal{D}_n^{-(t)}} \left\{ (Y_i - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \Theta_t))^2 - (Y_i - m_{n,1}(\mathbf{X}_i, \Theta_t))^2 \right\} \\ &= \frac{1}{\widehat{M}\gamma_n} \sum_{t=1}^{\widehat{M}} \sum_{i=1}^n \left\{ (Y_i - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \Theta_t))^2 - (Y_i - m_{n,1}(\mathbf{X}_i, \Theta_t))^2 \right\} \mathbb{1} \left\{ X_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)} \right\}, \end{aligned}$$

for $j \in \{1, \dots, p\}$, where γ_n is the cardinality of $\mathcal{D}_n^{-(t)}$. As we assume sampling without replacement (a_n times) during the bootstrap step $\gamma_n = n - a_n$ is constant. Furthermore, we have $\mathcal{D}_{n,\mathbf{X}}^{-(t)} = \{\mathbf{X}_i : i \in \mathcal{D}_n^{-(t)}\}$. Note that $\pi_{j,t}$ is independent of $\mathcal{D}_n$. While [5] proved, that $I_{n,\widehat{M}}^{OOB}(j)$ is a weakly consistent and asymptotically unbiased estimator for $I(j)$, we investigate the asymptotic distribution of $I_{n,\widehat{M}}^{OOB}(j)$. Simulations of [10] suggest that $I_{n,\widehat{M}}^{OOB}(j)$ is asymptotically normal. In the following, we are going to formally prove, that $I_{n,\widehat{M}}^{OOB}(j)$ is indeed asymptotically normal under some technical assumptions. To the best of our knowledge this is the first approach for a theoretical validation of the confidence intervals for $I(j)$ proposed in [10]. Note that our results are for a

randomly chosen number of trees, while the previous results for permutation importance were based on a fixed number of trees. The assumption of a randomly chosen number of trees is necessary to fit into the U statistic framework of [16], who already discussed this assumption.

3. U -statistics and permutation importance

We first prove that a simplified version of $I_{n,M}^{OOB}(j)$ is asymptotically normal via U -statistics and then show that $I_{n,M}^{OOB}(j)$ and its simplified version have the same asymptotic distribution.

We need a simplified version, because $I_{n,M}^{OOB}(j)$ does not fit into the general definition of a U -statistics, since the summation only goes over $M < \binom{n}{\gamma_n}$ subsets. Moreover, the permutation puts additional randomness to the kernel. Luckily there is a solution for both issues known as generalized U -statistics.

Definition 3.1 (Generalized U -statistics). Let $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ be i.i.d. random vectors and let $h : \mathbb{R}^s \rightarrow \mathbb{R}$ be a (possibly randomized) permutation symmetric function. Then U -statistics with kernel h_s are defined as estimators of the form,

$$U_{n,s,M,\omega} = \frac{1}{\widehat{M}} \sum_{(t)} \rho_t h_s(\mathbf{Z}_{i_1}, \dots, \mathbf{Z}_{i_s}; \omega_t) \quad (4)$$

where the summation is over all possible subsets from $\{\mathbf{Z}_1, \dots, \mathbf{Z}_n\}$ of size s . Moreover, ω_t denotes i.i.d. randomness independent of $(\mathbf{Z}_i)_i$ and $(\rho_t)_t$, where $\widehat{M} = \sum_t \rho_t$ and ρ_t are i.i.d. Bernoulli random variables with $\mathbb{P}(\rho_t = 1) = M/\binom{n}{s}$ determining which subsamples are selected. If $M = \binom{n}{s}$, then (4) is called a generalized complete U -statistics and if $M < \binom{n}{s}$ it is called generalized incomplete U -statistic [16].

U -statistics already were used to establish normality and rates of convergence for the Random Forest regression estimator $m_{n,M}$ in [13] and [16]. We want to apply the theory to the RFPIM. However, $I_{n,\widehat{M}}^{OOB}(j)$ cannot be put directly into a form like $U_{n,s,M,\omega}$. The situation is particularly complex because, for each tree, a subset of $\mathcal{D}n$ is used to train $m_{n,1}$, while the remaining observations are utilized to compute the OOB-error. That is why we start with a simplification of $I_{n,M}^{OOB}(j)$ by replacing the tree-estimation by the true regression function and thus define

$$\tilde{I}_{n,M}^{OOB}(j) = \frac{1}{M} \sum_{t=1}^M \frac{1}{\gamma_n} \sum_{i \in \mathcal{D}_n^{-(t)}} \left\{ (Y_i - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 - (Y_i - \tilde{m}(\mathbf{X}_i))^2 \right\}. \quad (5)$$

Our simplified version (5) of permutation importance fits into Definition 3.1, when we choose $s = \gamma_n$, $\omega_t = \pi_{j,t}$ and $(\mathbf{Z}_1, \dots, \mathbf{Z}_n) := ((\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n))$. Thus $\tilde{I}_{n,\widehat{M}}^{OOB}(j)$ is a generalized U -statistic with kernel

$$h_{\gamma_n}(\mathbf{Z}_{i_1}, \dots, \mathbf{Z}_{i_{\gamma_n}}; \pi_{j,t}) = \frac{1}{\gamma_n} \sum_{i \in \mathcal{D}_n^{-(t)}} \left\{ (Y_i - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 - (Y_i - \tilde{m}(\mathbf{X}_i))^2 \right\}.$$

Below we first show that $\tilde{I}_{n,M}^{OOB}(j)$ is asymptotically normal and then show that its difference with $I_{n,M}^{OOB}(j)$ asymptotically vanishes in probability to deduce the main result.

4. Asymptotic results

4.1. Central Limit Theorem for RFPIM

This section contains our main asymptotic results for the Permutation Importance Measure. To prove these, we need some additional assumptions

- (A1) We only consider permutations $\pi_{j;t} \in \mathcal{V}_n = \{\pi \in \mathcal{S}_{\gamma_n} : \pi(i) \neq i \text{ for all } i\}$, where $\mathcal{S}_{\gamma_n}$ is the symmetric group.
- (A2) $\tilde{m}$ is an additive function, i.e. there are functions $m_\ell : [0, 1] \rightarrow \mathbb{R}$, $\ell = 1, \dots, p$ such that $\tilde{m}(\mathbf{X}) = \sum_{\ell=1}^p m_\ell(X^{(\ell)})$.
- (A3) $\tilde{m}$ is bounded, i.e. there is a $K > 0$, such that $\sup_{\mathbf{x} \in [0,1]^p} |\tilde{m}(\mathbf{x})| \leq K$.
- (A4) The error terms have finite fourth moments, i.e. $\mathbb{E}[\varepsilon_1^4] < \infty$.
- (A5) $\widehat{M} \sim \text{Bin}\left(\binom{n}{\gamma_n}, M/\binom{n}{\gamma_n}\right)$ is chosen at random
- (A6) The features are mutually independent.

The Assumption (A1) is needed for technical reasons and was already used in [5] to establish consistency of the RFPIM. (A2) is needed to make an isolated measurement of the influence of each feature possible. Condition (A3) is needed for technical reasons, but we believe that the existence of an upper bound for the regression function is not too restrictive for most practical applications. The moment assumption (A4) is required for the consistency of the tree prediction [5]. For the proof of the consistency of $m_{n,M}$ in [18] assumption (A2) is also needed. [18] assumes the component functions to be continuous and instead of (A4) requires gaussian errors. The random number of trees (A5) is needed to fit in the generalized U -statistics framework of [16]. Condition (A6) also implies uncorrelated features, which is important since correlated features can lead to significant bias in RFPIM as discussed in [20]. To derive asymptotic normality, we need to make sure that we are in a non-degenerated case. To this end we additionally assume that $\text{Var}\left(m_j(X_1^{(j)})\right) > 0$. In fact, if $\text{Var}\left(m_j(X_1^{(j)})\right) = 0$, then $m_j(X_1^{(j)})$ would be constant almost surely meaning the permutation along the j -th component of $\mathbf{X}_1$ would not change $\tilde{m}(\mathbf{X}_1)$, which implies $\tilde{I}_{n,M}^{OOB}(j) = 0$ almost surely. Under these assumptions, $\tilde{I}_{n,M}^{OOB}(j)$ fits into the framework of generalized incomplete U -statistics and we can make use of Theorem 2 of [16] to achieve its asymptotic normality. To better follow the prove ideas, we repeat their theorem in

Theorem 4.1. *Let $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ be i.i.d. from $F_{\mathbf{Z}}$ and $U_{n,\gamma_n,M,\pi}$ be a generalized incomplete U -statistic with kernel $h_{\gamma_n}(\mathbf{Z}_1, \dots, \mathbf{Z}_n; \pi_j)$. Let $\mu = \mathbb{E}[h_{\gamma_n}]$, $\zeta_{1,\gamma_n} =$*

$\mathbb{V}ar(\mathbb{E}[h_{\gamma_n}|\mathbf{Z}_1])$ and $\zeta_{\gamma_n} = \mathbb{V}ar(h_{\gamma_n})$. Suppose that

$$\frac{\mathbb{E}[|h_{\gamma_n} - \mu|^{2k}]}{\mathbb{E}[|h_{\gamma_n} - \mu|^k]^2}$$

is uniformly bounded in γ_n for $k = 2, 3$. If $\frac{\gamma_n}{n} \frac{\zeta_{\gamma_n}}{\gamma_n \zeta_{1, \gamma_n}} \rightarrow 0$, then

$$\frac{U_{n, \gamma_n, M, \pi} - \mu}{\sqrt{\gamma_n^2 \zeta_{1, \gamma_n} / n + \zeta_{\gamma_n} / M}} \xrightarrow{d} \mathcal{N}(0, 1)$$

as $M \rightarrow \infty$ and $n \rightarrow \infty$ [16].

To apply Theorem 4.1 to $\tilde{I}_{n, M}^{OOB}(j)$, the main part of the work is to show that $\mathbb{E}[|h_{\gamma_n} - \mu|^{2k}] / \mathbb{E}[|h_{\gamma_n} - \mu|^k]^2$ is uniformly bounded in γ_n for $k = 2, 3$ and to prove that $\frac{\gamma_n}{n} \frac{\zeta_{\gamma_n}}{\gamma_n \zeta_{1, \gamma_n}} \rightarrow 0$. This is done in the Appendix (Lemma A.1 and Lemma A.2 and some lengthy calculations) and we thus receive the following

Theorem 4.2. *Let (A1)-(A6) be fulfilled and $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ independent and identically distributed. If $\mathbb{V}ar(m_j(X_1^{(j)})) > 0$, $\gamma_n \rightarrow \infty$ as $n \rightarrow \infty$, and $\gamma_n / \sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$, then*

$$\frac{\tilde{I}_{n, M}^{OOB}(j) - I(j)}{\sqrt{\gamma_n^2 \zeta_{1, \gamma_n} / n + \zeta_{\gamma_n} / M}} \xrightarrow{d} \mathcal{N}(0, 1)$$

as $M \rightarrow \infty$ and $n \rightarrow \infty$.

To transfer the result of Theorem 4.2 to $I_{n, M}^{OOB}(j)$, a sufficiently fast convergence of the base learner $m_{n, 1}$ is needed. [5] already pointed out, that the consistency of the individual trees implies the consistency of $I_{n, m}^{OOB}(j)$. We need the stronger assumption that the approximation error of $m_{n, 1}$ for $\tilde{m}$ is negligible. Therefore we need to make the so called 'Sufficient Impurity Decrease' assumption and use results from [11]. Let $A \subset [0, 1]^p$ be some hyperrectangle, that is partitioned into A_L and A_R via axis-aligned splitting. More formally for $b \in [0, 1]$ and $j \in \{1, \dots, p\}$

$$\begin{aligned} A_L &= A_L(j, b) := A \cap \{v \in [0, 1]^p : v^{(j)} \leq b\}, \\ A_R &= A_R(j, b) := A \cap \{v \in [0, 1]^p : v^{(j)} > b\}, \end{aligned}$$

where $v^{(j)}$ denotes the j -th component of v . Then the theoretical version of the CART splitting criterium [18] is defined as

$$\begin{aligned} \Delta(A, j, b) &:= \mathbb{P}(\mathbf{X} \in A) \mathbb{V}ar(\tilde{m}(\mathbf{X}) | \mathbf{X} \in A) - \mathbb{P}(\mathbf{X} \in A_L) \mathbb{V}ar(\tilde{m}(\mathbf{X}) | \mathbf{X} \in A_L) \\ &\quad - \mathbb{P}(\mathbf{X} \in A_R) \mathbb{V}ar(\tilde{m}(\mathbf{X}) | \mathbf{X} \in A_R). \end{aligned}$$

Equipped with this we can formulate the 'Sufficient Impurity Decrease' (SID) assumption

(A7) There is a constant $\lambda \in (0, 1)$ such that for all hyperrectangles $A \subset [0, 1]^p$

$$\sup_{1 \leq j \leq p, b \in [0, 1]} \Delta(A, j, b) \geq \lambda \cdot \mathbb{P}(\mathbf{X} \in A) \mathbb{V}\text{ar}(\tilde{m}(\mathbf{X}) | \mathbf{X} \in A).$$

Additionally we need a stricter version of (A4), to give a uniform upper bound for the tree predictions.

(A4*) The error terms are bounded by some $K_\varepsilon > 0$.

Moreover, we need to link the growth-rate of M to the sample size, i.e. $M = M_n$. Equipped with these assumptions we can now prove our main theorem, which delivers the asymptotic normality of $I_{n, \hat{M}}^{OOB}(j)$.

Theorem 4.3. *Let (A1)-(A3), (A4*), (A5)-(A7) be fulfilled and $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ independent and identically distributed and $\mathbb{V}\text{ar}(m_j(X_1^{(j)})) > 0$. If $\gamma_n \rightarrow \infty$, $\frac{M_n \gamma_n \log(a_n)^2 \log(a_n p)}{a_n^{\phi(\lambda)}} \rightarrow 0$ for $\phi(\lambda) = \frac{-\log_2(1-\lambda)}{1-\log_2(1-\lambda)}$ as $M_n \rightarrow \infty$ and $n \rightarrow \infty$. Then*

$$\frac{I_{n, \hat{M}}^{OOB}(j) - I(j)}{\sqrt{\gamma_n^2 \zeta_{1, \gamma_n} / n + \zeta_{\gamma_n} / M_n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

as $M_n \rightarrow \infty$ and $n \rightarrow \infty$, if the trees are grown to depth $d = \lceil \log_2(n) / (1 - \log_2(1 - \lambda)) \rceil$.

To our knowledge this is the first CLT for the RFPIM and thus the first mathematical reasoning for the asymptotic normality assumed in [10].

The assumption (A7) is an assumption on the regression function $\tilde{m}$, that is not very intuitive. Function classes satisfying this condition were investigated by [11]. Proposition 3.1 of [11] tells us that (A7) is fulfilled if $m_1, \dots, m_p$ are from the Locally Reverse Poincaré Class with parameter τ , i.e. $m_1, \dots, m_p$ are differentiable on $(0, 1)$ and for any subinterval $[a, b] \subseteq (0, 1)$ and

$$\left(\int_a^b |m'_j(t)| dt \right)^2 \leq \frac{\tau^2}{b-a} \inf_{w \in \mathbb{R}} \int_a^b |m_j(t) - w|^2 dt$$

for every $j \in \{1, \dots, p\}$. In this case $\lambda = 4/p\tau^2$. This is fulfilled if for instance each m_j is a piecewise defined function, whose pieces are polynomial, strictly monotonous functions or smooth and strongly convex.

The values for λ in the examples given in [11] are rather small. Thus the desired depth d is close to $\log_2(n)$ which corresponds to the depth of fully grown binary tree. The assumption of

$$\frac{M_n \gamma_n \log(a_n)^2 \log(a_n p)}{a_n^{\phi(\lambda)}} \rightarrow 0$$

as $n \rightarrow \infty$ is quite restrictive and is not fulfilled by typical choices like $a_n \approx 0.632 \cdot n$. Our assumption is for example fulfilled by choices like $M_n = \lfloor \log(n)^{q_1} \rfloor$ and $\gamma_n = \lfloor n - \log(n)^{q_2} \rfloor$ for some arbitrary $q_1, q_2 > 0$.

5. Simulation Study

In this section, we conduct a simulation study to evaluate the practical relevance of our central limit theorem. Specifically, we compare the empirical distribution of RFPIM under various conditions. Furthermore, we examine the robustness of our assumptions by analyzing how the distribution of RFPIM changes when these assumptions are violated. All simulations are performed using R version 4.4.1 [17].

5.1. Simulation settings

For each of the $N = 1000$ simulation runs we generate $n = 2500$ i.i.d. copies $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ of $(\mathbf{X}, Y)$ following Model (1). The sample size equals the sample size in [10], who already illustrated via simulation, that the distribution of RFPIM can often be approximated by a normal distribution. We choose $p = 10$, i.e. $\mathbf{X} \sim \mathcal{U}([0, 1]^{10})$.

The Random Forest algorithm is implemented in R using the **ranger** package [23]. For each $k = 1, \dots, N$, we train a Random Forest model with $M = 500$ trees and a subsample size of $a_n = \lfloor 2/3 \cdot n \rfloor$. The parameter m_{try} is set to the default value provided by the **ranger** package. Using the trained Random Forest, we compute the RFPIM $I_{n, \widehat{M}}^{OOB}(j)$ for $j = 1, \dots, 10$, and denote the results from the k -th simulation run by $I_{n, \widehat{M}}^{OOB, k}(j)$. Finally, we compare the empirical quantiles of the standardized values $I_{n, \widehat{M}}^{OOB, 1}(j), \dots, I_{n, \widehat{M}}^{OOB, N}(j)$ with the quantiles of the standard normal distribution using QQ-plots. To investigate robustness of our results to derivations from our assumptions, we vary the error distribution, the regression function $\tilde{m}$, and whether permutations or derrangements are used for the feature importance as described below.

Error Distribution. In Theorem 4.3 we require (A4*), which is that the error terms are bounded. This is fulfilled for a truncated normal distribution. We use the R function **rtrunc** from the package **truncdist** ([15]) to generate the truncated normal distributed error terms. We choose -2 as lower bound and 2 as upper bound. The mean is set to zero and the standard deviation is set to one for the truncated normal distribution. To test how crucial this stricter version of (A4) is, we also use standard normal distributed error terms fulfilling (A4), but not (A4*).

Regression Functions. For the regression function $\tilde{m} : [0, 1]^p \rightarrow \mathbb{R}$ our theory needs an additive structure (A2). We use a total of four different regression functions, where the first fulfills (A2) and the other three do not. The first two regression functions are inspired by [5], while the other non-additive functions are taken from [10]. In contrast to [10] we use less noisy features in order to stay consistent with the first function fulfilling (A2). The regression functions are specified as follows:

1. First, we use a polynomial regression function that satisfies (A2):

$$\tilde{m}(\mathbf{X}) = 2X^{(1)} + 4\left(X^{(2)}\right)^2 + 2\left(X^{(3)}\right)^3 - 3\left(X^{(4)}\right)^4 + \left(X^{(5)}\right)^5$$

2. We investigate a non-additive trigonometric function as link-function that connects Y with a linear/additive predictor:

$$\tilde{m}(\mathbf{X}) = \sin\left(2X^{(1)} + 4X^{(2)} + 2X^{(3)} - 3X^{(4)} + X^{(5)}\right)$$

3. We use a function with a trigonometric and a polynomial part:

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

4. Finally, we consider a function with three interaction effects:

$$\tilde{m}(\mathbf{X}) = X^{(1)} X^{(2)} + \left(X^{(3)}\right)^2 - X^{(4)} X^{(7)} + X^{(8)} X^{(10)} - \left(X^{(6)}\right)^2$$

Permutation and Derrangement. Our assumption (A1) restricts ourselves to derangements, which is needed for technical reasons and has previously been assumed in [5] to prove consistency of the RFPIM. To get an impression whether this is actually needed, we consider both scenarios and run the simulation under ordinary permutations and derangements.

5.2. Simulation Results

In Figure 1, we present the results for each of the four regression functions, focusing on the non-noise features, i.e., those where $I(j) > 0$. Under the assumption (A2), this corresponds to the condition $\text{Var}(m_j(X_1^{(j)})) > 0$ as stated in Theorem 4.2 and Theorem 4.3. We begin by examining potential violations of the additive structure of the regression function as per assumption (A2). The first plot in Figure 1 is the only one where all the assumptions of Theorem 4.3 hold. Here, we observe that the quantiles from the simulation closely align with the quantiles of a normal distribution, suggesting that the sample size of $n = 2500$ and the number of trees $M = 500$ are sufficient for a good approximation. For the second and third function, the normal distributions also appears to fit the distribution of the standardized RFPIM well, with the quantiles aligning closely. However, for the fourth function, the interaction effects introduce severe discrepancies. In this case, the normal distribution exhibits heavier tails than the RFPIM distribution, indicating a potential violation of the normality assumption. The usage of arbitrary permutations instead of derangements does not change the QQ-Plots much as we can see in Figure 2. One could even argue, that the empirical quantiles fit slightly better to the normal distributions compared to Figure 1.

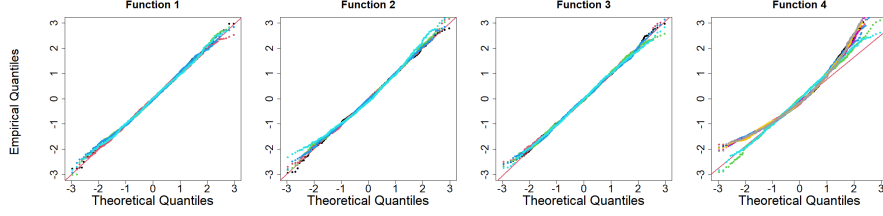


FIG 1. QQ-Plots for truncated normal errors and derrangements under the four different regression models.

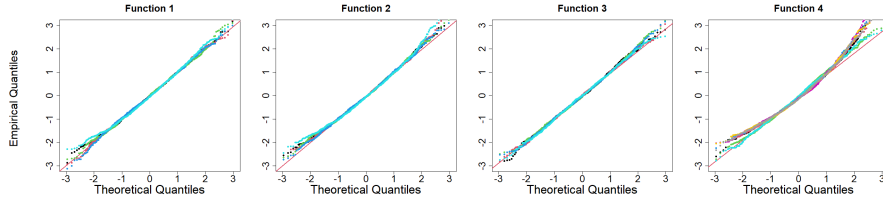


FIG 2. QQ-Plots for truncated normal errors and full permutation under the four different regression models.

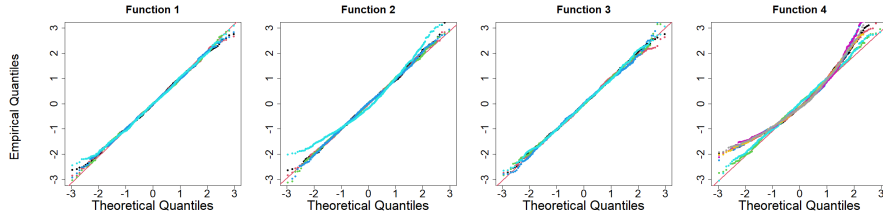


FIG 3. QQ-Plots for normal errors and derrangements under the four different regression models.

Thus the simulation suggests, that (A1) is not that crucial for the CLT given in Theorem 4.3. The results for normal (instead of truncated normal) errors combined with derangements are given in Figure 3. We again observe some deviation from the normal distribution for the fourth function and now also for the second one. For the two other regression functions, asymptotic normality again seems to be plausible.

6. Discussion

In this paper, we employ U -statistics theory to establish, to the best of our knowledge, the first Central Limit Theorem for the Random Forest Permutation Importance (RFPIM). This result provides a mathematical guarantee that

the standardized RFPIM is asymptotically normally distributed, supporting the assumption made by [10]. Our theory applies to all bounded and additive regression functions whose components satisfy the ‘Sufficient Impurity Decrease’ assumption and bounded errors. This includes smooth and convex functions, piecewise polynomials, and piecewise monotone functions. However, the CLT needs a few other technical restrictions that deviate from the original Random Forest framework: we select the number of trees at random, to fit in the U -statistics framework of [16]. Another restriction is in the choice of hyperparameters because we need a huge sample size of the bootstrap sample in the bootstrap step of the Random Forests in comparison to the Out-Of-Bag sample, to control the approximation errors of the individual trees. Moreover, as in the consistency proofs of the RFPIM in [5], we restrict the results to derangements instead of all permutations.

In a simulation study we investigated how the asymptotic distribution behaves if these assumptions are violated. It turned out that the restriction to derangements do not seem to have a huge impact on the asymptotic RFPIM distribution. The same holds true for the assumption of bounded error terms. However, the additive structure of the regression function turned out to be important, since asymptotic normality does not seem to hold if there are multiple multiplicative effects. The result nevertheless delivers an important first step to the theoretical validation of the confidence intervals proposed in [10].

Appendix A: Proof of Theorem 3

Lemma A.1. *Let (A1)-(A4) be fulfilled. If $\text{Var}(m_j(X_1^{(j)})) > 0$, then there is a $\tilde{\zeta} > 0$, such that $\gamma_n^2 \zeta_{1, \gamma_n} \rightarrow \tilde{\zeta} > 0$ as $n \rightarrow \infty$.*

Proof. We start with

$$\begin{aligned} \mathbb{E}[h|\mathbf{Z}_1] &= \mathbb{E}\left[\frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \left\{ (Y_i - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 - (Y_i - \tilde{m}(\mathbf{X}_i))^2 \right\} \middle| \mathbf{Z}_1\right] \\ &= \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[(Y_i - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 \middle| \mathbf{Z}_1\right] - \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[(Y_i - \tilde{m}(\mathbf{X}_i))^2 \middle| \mathbf{Z}_1\right]. \end{aligned} \quad (6)$$

The second term can be rewritten as

$$\begin{aligned} &\frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[(Y_i - \tilde{m}(\mathbf{X}_i))^2 \middle| \mathbf{Z}_1\right] \\ &= \frac{1}{\gamma_n} \mathbb{E}\left[(Y_1 - \tilde{m}(\mathbf{X}_1))^2 \middle| \mathbf{Z}_1\right] + \frac{1}{\gamma_n} \sum_{i=2}^{\gamma_n} \mathbb{E}\left[(Y_i - \tilde{m}(\mathbf{X}_i))^2 \middle| \mathbf{Z}_1\right] \\ &= \frac{1}{\gamma_n} (Y_1 - \tilde{m}(\mathbf{X}_1))^2 + \frac{1}{\gamma_n} \sum_{i=2}^{\gamma_n} \mathbb{E}\left[(Y_i - \tilde{m}(\mathbf{X}_i))^2\right] \\ &= \frac{1}{\gamma_n} \varepsilon_1^2 + \frac{\gamma_n - 1}{\gamma_n} \mathbb{E}[\varepsilon_1^2] \end{aligned} \quad (7)$$

almost surely. We use (A2) and get

$$\begin{aligned} (Y_i - \tilde{m}(X_i^{\pi_{j,1}}))^2 &= \left(\sum_{\ell=1}^p m_\ell(X_i^{(\ell)}) + \varepsilon_i - \sum_{\ell=1, \ell \neq j}^p m_\ell(X_i^{(\ell)}) - m_j(X_{\pi_j(i)}^{(j)}) \right)^2 \\ &= \left(m_j(X_i^{(j)}) + \varepsilon_i - m_j(X_{\pi_j(i)}^{(j)}) \right)^2 \end{aligned}$$

for $i = 1, \dots, n$. Thus, for the first term in (6) we get

$$\begin{aligned} &\frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[(Y_i - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 \middle| \mathbf{Z}_1\right] \\ &= \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[\left(m_j(X_i^{(j)}) + \varepsilon_i - m_j(X_{\pi_j(i)}^{(j)})\right)^2 \middle| \mathbf{Z}_1\right] \\ &= \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[\left(m_j(X_i^{(j)}) + \varepsilon_i\right)^2 \middle| \mathbf{Z}_1\right] + \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[\left(m_j(X_{\pi_j(i)}^{(j)})\right)^2 \middle| \mathbf{Z}_1\right] \\ &\quad + \frac{2}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[m_j(X_{\pi_j(i)}^{(j)}) \left(m_j(X_i^{(j)}) + \varepsilon_i\right) \middle| \mathbf{Z}_1\right] \end{aligned} \quad (8)$$

almost surely. For the first sum we get

$$\begin{aligned}
& \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E} \left[\left(m_j \left(X_i^{(j)} \right) + \varepsilon_i \right)^2 \middle| \mathbf{Z}_1 \right] \\
&= \frac{1}{\gamma_n} \mathbb{E} \left[\left(m_j \left(X_1^{(j)} \right) + \varepsilon_1 \right)^2 \middle| \mathbf{Z}_1 \right] + \frac{1}{\gamma_n} \sum_{i=2}^{\gamma_n} \mathbb{E} \left[\left(m_j \left(X_i^{(j)} \right) + \varepsilon_i \right)^2 \middle| \mathbf{Z}_1 \right] \\
&= \frac{1}{\gamma_n} \left(m_j \left(X_1^{(j)} \right) + \varepsilon_1 \right)^2 + \frac{1}{\gamma_n} \sum_{i=2}^{\gamma_n} \mathbb{E} \left[\left(m_j \left(X_i^{(j)} \right) + \varepsilon_i \right)^2 \right] \\
&= \frac{1}{\gamma_n} \left(m_j \left(X_1^{(j)} \right) + \varepsilon_1 \right)^2 + \frac{\gamma_n - 1}{\gamma_n} \mathbb{E} \left[\left(m_j \left(X_1^{(j)} \right) + \varepsilon_1 \right)^2 \right] \tag{9}
\end{aligned}$$

almost surely. The law of total expectation delivers

$$\begin{aligned}
& \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E} \left[m_j \left(X_{\pi(i)}^{(j)} \right)^2 \middle| \mathbf{Z}_1 \right] \\
&= \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{P}(\pi(i) = 1) \mathbb{E} \left[m_j \left(X_1^{(j)} \right)^2 \middle| \mathbf{Z}_1 \right] \\
&+ \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{P}(\pi(i) \neq 1) \mathbb{E} \left[m_j \left(X_2^{(j)} \right)^2 \middle| \mathbf{Z}_1 \right] \\
&= \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{P}(\pi(i) = 1) m_j \left(X_1^{(j)} \right)^2 + \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{P}(\pi(i) \neq 1) \mathbb{E} \left[m_j \left(X_2^{(j)} \right)^2 \right]
\end{aligned}$$

almost surely, where the i.i.d. assumption was used in the first and second equation. For the remaining part of (8) we can see

$$\begin{aligned}
&= \frac{1}{\gamma_n - 1} m_j \left(X_1^{(j)} \right)^2 + \frac{\gamma_n - 2}{\gamma_n - 1} \mathbb{E} \left[m_j \left(X_1^{(j)} \right)^2 \right] \tag{10} \\
&+ \frac{2}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E} \left[m_j \left(X_{\pi_j(i)}^{(j)} \right) \left(m_j \left(X_i^{(j)} \right) + \varepsilon_i \right) \middle| \mathbf{Z}_1 \right] \\
&= \frac{2}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E} \left[m_j \left(X_{\pi_j(i)}^{(j)} \right) m_j \left(X_i^{(j)} \right) \middle| \mathbf{Z}_1 \right] + \frac{2}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E} \left[m_j \left(X_i^{(j)} \right) \varepsilon_i \middle| \mathbf{Z}_1 \right] \\
&= \frac{2}{\gamma_n} \left(\sum_{i=1}^{\gamma_n} \mathbb{E} \left[m_j \left(X_{\pi_j(i)}^{(j)} \right) m_j \left(X_i^{(j)} \right) \middle| \mathbf{Z}_1 \right] + \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \varepsilon_1 \middle| \mathbf{Z}_1 \right] \right. \\
&+ \left. \sum_{i=2}^{\gamma_n} \mathbb{E} \left[m_j \left(X_i^{(j)} \right) \varepsilon_i \right] \right) \\
&= \frac{2}{\gamma_n} \left(\sum_{i=1}^{\gamma_n} \mathbb{E} \left[m_j \left(X_{\pi_j(i)}^{(j)} \right) m_j \left(X_i^{(j)} \right) \middle| \mathbf{Z}_1 \right] + m_j \left(X_1^{(j)} \right) \varepsilon_1 \right.
\end{aligned}$$

$$\begin{aligned}
& + \sum_{i=2}^{\gamma_n} \mathbb{E} \left[m_j \left(X_i^{(j)} \right) \right] \mathbb{E} [\varepsilon_i] \Bigg) \\
& = \frac{2}{\gamma_n} \left(\sum_{i=2}^{\gamma_n} \mathbb{E} \left[m_j \left(X_{\pi_j(i)}^{(j)} \right) m_j \left(X_i^{(j)} \right) \middle| \mathbf{Z}_1 \right] \right. \\
& \quad \left. + \mathbb{E} \left[m_j \left(X_{\pi_j(1)}^{(j)} \right) m_j \left(X_1^{(j)} \right) \middle| \mathbf{Z}_1 \right] + m_j \left(X_1^{(j)} \right) \varepsilon_1 \right) \\
& = \frac{2}{\gamma_n} \left(\sum_{i=2}^{\gamma_n} \mathbb{E} \left[m_j \left(X_{\pi_j(i)}^{(j)} \right) m_j \left(X_i^{(j)} \right) \middle| \mathbf{Z}_1 \right] \right. \\
& \quad \left. + m_j \left(X_1^{(j)} \right) \mathbb{E} \left[m_j \left(X_{\pi_j(1)}^{(j)} \right) \middle| \mathbf{Z}_1 \right] + m_j \left(X_1^{(j)} \right) \varepsilon_1 \right)
\end{aligned} \tag{11}$$

almost surely. Next we have analogously to (10)

$$\begin{aligned}
\mathbb{E} \left[m_j \left(X_{\pi(1)}^{(j)} \right) \middle| \mathbf{Z}_1 \right] & = \frac{1}{\gamma_n - 1} \sum_{k=2}^{\gamma_n} \mathbb{E} \left[m_j \left(X_k^{(j)} \right) \middle| \mathbf{Z}_1 \right] + \frac{1}{\gamma_n - 1} m_j \left(X_1^{(j)} \right) \\
& = \frac{\gamma_n - 2}{\gamma_n - 1} \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] + \frac{1}{\gamma_n - 1} m_j \left(X_1^{(j)} \right)
\end{aligned}$$

and analogously for $i \geq 2$

$$\begin{aligned}
& \mathbb{E} \left[m_j \left(X_{\pi_j(i)}^{(j)} \right) m_j \left(X_i^{(j)} \right) \middle| \mathbf{Z}_1 \right] \\
& = \sum_{k=1, k \neq i}^{\gamma_n} \mathbb{P}(\pi_j(i) = k) \mathbb{E} \left[m_j \left(X_k^{(j)} \right) m_j \left(X_i^{(j)} \right) \middle| \mathbf{Z}_1 \right] \\
& = \frac{1}{\gamma_n - 1} \left(\mathbb{E} \left[m_j \left(X_k^{(j)} \right) m_j \left(X_i^{(j)} \right) \middle| \mathbf{Z}_1 \right] \right. \\
& \quad \left. + \sum_{k=2, k \neq i}^{\gamma_n} \mathbb{E} \left[m_j \left(X_k^{(j)} \right) m_j \left(X_i^{(j)} \right) \middle| \mathbf{Z}_1 \right] \right) \\
& = \frac{1}{\gamma_n - 1} \left(m_j \left(X_1^{(j)} \right) \mathbb{E} \left[m_j \left(X_i^{(j)} \right) \middle| \mathbf{Z}_1 \right] + \sum_{k=2, k \neq i}^{\gamma_n} \mathbb{E} \left[m_j \left(X_k^{(j)} \right) m_j \left(X_i^{(j)} \right) \right] \right) \\
& = \frac{1}{\gamma_n - 1} \left(m_j \left(X_1^{(j)} \right) \mathbb{E} \left[m_j \left(X_i^{(j)} \right) \right] + \sum_{k=2, k \neq i}^{\gamma_n} \mathbb{E} \left[m_j \left(X_k^{(j)} \right) \right] \mathbb{E} \left[m_j \left(X_i^{(j)} \right) \right] \right) \\
& = \frac{1}{\gamma_n - 1} \left(m_j \left(X_1^{(j)} \right) \mathbb{E} \left[m_j \left(X_i^{(j)} \right) \right] + (\gamma_n - 2) \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right]^2 \right) \\
& = \frac{1}{\gamma_n - 1} \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] m_j \left(X_i^{(j)} \right) + \frac{\gamma_n - 2}{\gamma_n - 1} \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right]^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
& \frac{2}{\gamma_n} \sum_{i=2}^{\gamma_n} \mathbb{E} \left[m_j \left(X_{\pi_j(i)}^{(j)} \right) \left(m_j \left(X_i^{(j)} \right) + \varepsilon_i \right) \middle| \mathbf{Z}_1 \right] \\
&= \frac{2}{\gamma_n} \left(\mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] m_j \left(X_1^{(j)} \right) + (\gamma_n - 2) \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right]^2 \right. \\
&\quad \left. + \frac{\gamma_n - 2}{\gamma_n - 1} m_j \left(X_1^{(j)} \right) \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] + \frac{1}{\gamma_n - 1} m_j \left(X_1^{(j)} \right)^2 + m_j \left(X_1^{(j)} \right) \varepsilon_1 \right) \quad (12)
\end{aligned}$$

holds true. Combining (8), (9), (10) and (12) leads to

$$\begin{aligned}
& \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E} \left[\left(Y_i - \tilde{m}(\mathbf{X}_i^{\pi_j, t}) \right)^2 \middle| \mathbf{Z}_1 \right] \\
&= \frac{1}{\gamma_n} \left(m_j \left(X_1^{(j)} \right) + \varepsilon_1 \right)^2 + \frac{\gamma_n - 1}{\gamma_n} \mathbb{E} \left[\left(m_j \left(X_1^{(j)} \right) + \varepsilon_1 \right)^2 \right] \\
&\quad + \frac{1}{\gamma_n - 1} m_j \left(X_1^{(j)} \right)^2 + \frac{\gamma_n - 2}{\gamma_n - 1} \mathbb{E} \left[m_j \left(X_1^{(j)} \right)^2 \right] \\
&\quad + \frac{2}{\gamma_n} \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] m_j \left(X_1^{(j)} \right) + 2 \frac{\gamma_n - 2}{\gamma_n} \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right]^2 \\
&\quad + 2 m_j \left(X_1^{(j)} \right) \frac{\gamma_n - 2}{\gamma_n (\gamma_n - 1)} \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] \\
&\quad + \frac{2}{\gamma_n (\gamma_n - 1)} m_j \left(X_1^{(j)} \right)^2 + \frac{2}{\gamma_n} m_j \left(X_1^{(j)} \right) \varepsilon_1. \quad (13)
\end{aligned}$$

Note that there are several constant summands in (13), that vanish inside the variance. Hence (6), (7) and (13) result in

$$\begin{aligned}
\zeta_{1, \gamma_n} = \mathbb{V} \text{ar} \left(\mathbb{E} [h | \mathbf{Z}_1] \right) &= \mathbb{V} \text{ar} \left(\frac{1}{\gamma_n} \left(m_j \left(X_1^{(j)} \right) + \varepsilon_1 \right)^2 + \frac{1}{\gamma_n - 1} m_j \left(X_1^{(j)} \right)^2 \right. \\
&\quad \left. + \frac{2}{\gamma_n} \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] m_j \left(X_1^{(j)} \right) \frac{2(\gamma_n - 2)}{\gamma_n (\gamma_n - 1)} m_j \left(X_1^{(j)} \right) \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] \right. \\
&\quad \left. + \frac{2}{\gamma_n (\gamma_n - 1)} m_j \left(X_1^{(j)} \right)^2 + \frac{2}{\gamma_n} m_j \left(X_1^{(j)} \right) \varepsilon_1 - \frac{1}{\gamma_n} \varepsilon_1^2 \right),
\end{aligned}$$

which implies with (A3), (A4) and the dominated convergence theorem, that

$\gamma_n^2 \zeta_{1, \gamma_n}$ is asymptotically equivalent to

$$\begin{aligned}
& \mathbb{V}\text{ar} \left(\left(m_j \left(X_1^{(j)} \right) + \varepsilon_1 \right)^2 + m_j \left(X_1^{(j)} \right)^2 + 4m_j \left(X_1^{(j)} \right) \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] \right. \\
& \quad \left. + 2m_j \left(X_1^{(j)} \right) \varepsilon_1 - \varepsilon_1^2 \right) \\
&= \mathbb{V}\text{ar} \left(2m_j \left(X_1^{(j)} \right)^2 + 4m_j \left(X_1^{(j)} \right) \mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] + 4m_j \left(X_1^{(j)} \right) \varepsilon_1 \right) \\
&= \mathbb{V}\text{ar} \left(m_j \left(X_1^{(j)} \right) \left(2m_j \left(X_1^{(j)} \right) + 4\mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] \right) + 4m_j \left(X_1^{(j)} \right) \varepsilon_1 \right) =: \tilde{\zeta}.
\end{aligned}$$

Due to the independence of ε_1 and $\mathbf{X}_1$,

$$\begin{aligned}
& \text{Cov} \left(m_j \left(X_1^{(j)} \right) \left(2m_j \left(X_1^{(j)} \right) + 4\mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] \right), 4m_j \left(X_1^{(j)} \right) \varepsilon_1 \right) \\
&= \mathbb{E} \left[4m_j \left(X_1^{(j)} \right)^2 \left(2m_j \left(X_1^{(j)} \right) + 4\mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] \right) \varepsilon_1 \right] \\
&= \mathbb{E} \left[4m_j \left(X_1^{(j)} \right)^2 \left(2m_j \left(X_1^{(j)} \right) + 4\mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] \right) \right] \mathbb{E} [\varepsilon_1] = 0
\end{aligned}$$

holds true and hence

$$\begin{aligned}
\tilde{\zeta} &= \mathbb{V}\text{ar} \left(m_j \left(X_1^{(j)} \right) \left(2m_j \left(X_1^{(j)} \right) + 4\mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] \right) + 4m_j \left(X_1^{(j)} \right) \varepsilon_1 \right) \\
&= \mathbb{V}\text{ar} \left(m_j \left(X_1^{(j)} \right) \left(2m_j \left(X_1^{(j)} \right) + 4\mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] \right) \right) + \mathbb{V}\text{ar} \left(4m_j \left(X_1^{(j)} \right) \varepsilon_1 \right) \\
&\geq 16\mathbb{V}\text{ar} \left(m_j \left(X_1^{(j)} \right) \varepsilon_1 \right) = 16\mathbb{E} \left[m_j \left(X_1^{(j)} \right)^2 \varepsilon_1^2 \right] = 16\sigma^2 \mathbb{E} \left[m_j \left(X_1^{(j)} \right)^2 \right] > 0,
\end{aligned}$$

which completes the proof. $\square$

Lemma A.2. *Let (A3) and (A4) be fulfilled. Then there is a $\tilde{K} > 0$, such that $\zeta_{\gamma_n} = \text{Var}(h) < \tilde{K}$ for all $n \in \mathbb{N}$.*

Proof. At first we apply Jensens inequality for sums and by noting that $(a+b)^2 \leq$

$2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$ we receive

$$\begin{aligned}
\mathbb{E}[h_{\gamma_n}^2] &= \mathbb{E}\left[\left(\frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \left\{ (Y_i - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 - (Y_i - \tilde{m}(\mathbf{X}_i))^2 \right\}\right)^2\right] \\
&\leq \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[\left(\left\{ (Y_i - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 - (Y_i - \tilde{m}(\mathbf{X}_i))^2 \right\}\right)^2\right] \\
&\leq \frac{2}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[(Y_i - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^4\right] + \frac{2}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[(Y_i - \tilde{m}(\mathbf{X}_i))^4\right] \\
&= \frac{2}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[\left(m_j(X_i^{(j)}) + \varepsilon_i - m_j(X_{\pi_j(i)}^{(j)})\right)^4\right] + \frac{2}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}[\varepsilon_i^4].
\end{aligned}$$

Because of (A4), it's sufficient to show that for all $i \in \{1, \dots, n\}$,

$$\mathbb{E}\left[\left(m_j(X_i^{(j)}) + \varepsilon_i - m_j(X_{\pi_j(i)}^{(j)})\right)^4\right]$$

is finite. Note that $|m_j(x)| < K$ for all $x \in [0, 1]$, due to (A3). Hence, by the independence of $\mathbf{X}_i$ and ε_i and (A4), we get

$$\begin{aligned}
&\mathbb{E}\left[\left(m_j(X_i^{(j)}) + \varepsilon_i - m_j(X_{\pi_j(i)}^{(j)})\right)^4\right] \\
&= \mathbb{E}\left[\left(m_j(X_i^{(j)}) - m_j(X_{\pi_j(i)}^{(j)})\right)^4\right] + 4\mathbb{E}\left[\left(m_j(X_i^{(j)}) - m_j(X_{\pi_j(i)}^{(j)})\right)^3 \varepsilon_i\right] \\
&+ 6\mathbb{E}\left[\left(m_j(X_i^{(j)}) - m_j(X_{\pi_j(i)}^{(j)})\right)^2 \varepsilon_i^2\right] + 4\mathbb{E}\left[\left(m_j(X_i^{(j)}) - m_j(X_{\pi_j(i)}^{(j)})\right) \varepsilon_i^3\right] \\
&+ \mathbb{E}[\varepsilon_i^4] \\
&= \mathbb{E}\left[\left(m_j(X_i^{(j)}) - m_j(X_{\pi_j(i)}^{(j)})\right)^4\right] + 6\sigma^2 \mathbb{E}\left[\left(m_j(X_i^{(j)}) - m_j(X_{\pi_j(i)}^{(j)})\right)^2\right] \\
&+ \mathbb{E}[\varepsilon_i^4] \\
&\leq 16K^4 + 24\sigma^2 K^2 + \mathbb{E}[\varepsilon_i^4] < \infty.
\end{aligned}$$

Therefore the proof is completed. $\square$

Lemma A.3. Let $\mathbf{X}_{j,1} = \left[X_1^{(1)}, \dots, X_1^{(j-1)}, X, X_1^{(j+1)}, \dots, X_1^{(p)}\right]^\top$, where X is an independent copy of $X_1^{(j)}$, then

$$\mathbb{E}[h] = \mathbb{E}[(Y_1 - \tilde{m}(\mathbf{X}_{j,1}))^2] - \sigma^2 =: I(j)$$

if (A1)-(A4) is fulfilled

Proof. Since $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ are i.i.d., we can conclude from (A2), that

$$\begin{aligned} & \mathbb{E}[h(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)] \\ &= \mathbb{E}\left[\frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \left\{ (Y_i - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 - (Y_i - \tilde{m}(\mathbf{X}_i))^2 \right\}\right] \\ &= \mathbb{E}\left[(Y_1 - \tilde{m}(\mathbf{X}_1^{\pi_{j,t}}))^2\right] - \mathbb{E}[\varepsilon_1^2] = \mathbb{E}\left[(Y_1 - \tilde{m}(\mathbf{X}_1^{\pi_{j,t}}))^2\right] - \sigma^2 \\ &= \mathbb{E}\left[\left(m_j(X_1^{(j)}) + \varepsilon_1 - m_j(X_{\pi_j(1)}^{(j)})\right)^2\right] - \sigma^2 \end{aligned}$$

holds true. With (A1) and the law of total expectation we get

$$\begin{aligned} & \mathbb{E}\left[\left(m_j(X_1^{(j)}) + \varepsilon_1 - m_j(X_{\pi_j(1)}^{(j)})\right)^2\right] \\ &= \sum_{k=2}^{\gamma_n} \mathbb{P}(\pi_j(1) = k) \mathbb{E}\left[\left(m_j(X_1^{(j)}) + \varepsilon_1 - m_j(X_k^{(j)})\right)^2\right] \\ &= \sum_{k=2}^{\gamma_n} \frac{1}{\gamma_n - 1} \mathbb{E}\left[\left(m_j(X_1^{(j)}) + \varepsilon_1 - m_j(X_k^{(j)})\right)^2\right] = \mathbb{E}\left[(Y_1 - \tilde{m}(\mathbf{X}_{j,1}))^2\right]. \end{aligned}$$

□

Proof of Theorem 3

Proof. The proof idea for the theorem is applying Theorem 2.1. The main task of the proof will be to check whether

$$\frac{\mathbb{E}[|h_{\gamma_n} - I(j)|^{2k}]}{\mathbb{E}[|h_{\gamma_n} - I(j)|^k]^2} \quad (14)$$

is uniformly bounded in n for $k = 2, 3$. Since we are dealing with permutations we start by establishing some properties. The elements $\pi_j \in \mathcal{V}_n$ are so called derangements. The number of derangements of an γ_n -element set is given as

$$|\mathcal{V}_n| = !\gamma_n := \gamma_n! \sum_{i=0}^n \frac{(-1)^i}{i!} = \left\lfloor \frac{\gamma_n!}{e} \right\rfloor = \left\lfloor \frac{\gamma_n! + 1}{e} \right\rfloor.$$

For $i \neq k$

$$\begin{aligned} & \mathbb{P}(\{\pi_j(k) = i\} \cup \{\pi_j(i) = k\}) \\ &= \mathbb{P}(\pi_j(k) = i) + \mathbb{P}(\pi_j(i) = k) - \mathbb{P}(\{\pi_j(k) = i\}, \{\pi_j(i) = k\}) \\ &= \frac{2}{\gamma_n - 1} - \frac{!(\gamma_n - 2)}{!\gamma_n} =: p_{\gamma_n} \end{aligned} \quad (15)$$

and

$$\begin{aligned}\mathbb{P}(\pi_j(k) = i, \pi_j(i) \neq k) &= \mathbb{P}(\pi_j(k) = i) - \mathbb{P}(\pi_j(k) = i, \pi_j(i) = k) \\ &= \frac{1}{\gamma_n - 1} - \frac{!(\gamma_n - 2)}{!\gamma_n} = \frac{1}{\gamma_n} + \mathcal{O}\left(\frac{1}{\gamma_n^2}\right).\end{aligned}\quad (16)$$

Using (A2), Lemma A.3, (15) and the i.i.d. assumption as well as $\mathbb{E}[\varepsilon_i] = 0$ (*) leads to

$$\begin{aligned}& \mathbb{E}\left[|h_{\gamma_n} - I(j)|^2\right] \\ &= \mathbb{E}\left[\left|\frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} (Y_i - \tilde{m}(X_i^{\pi_j}))^2 - \varepsilon_i^2 - I(j)\right|^2\right] \\ &= \mathbb{E}\left[\left|\frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \left\{ \left(m_j(X_i^{(j)}) - m_j(X_{\pi(i)}^{(j)})\right)^2 \right. \right. \right. \\ &\quad \left. \left. \left. + 2\varepsilon_i \left(m_j(X_i^{(j)}) - m_j(X_{\pi(i)}^{(j)})\right)\right\} - I(j) \right|^2\right] \\ &\stackrel{(*)}{=} \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \mathbb{E}\left[\left\{ \left(m_j(X_i^{(j)}) - m_j(X_{\pi(i)}^{(j)})\right)^2 \right. \right. \\ &\quad \left. \left. + 2\varepsilon_i \left(m_j(X_i^{(j)}) - m_j(X_{\pi(i)}^{(j)})\right) - I(j) \right\}^2\right] \\ &+ \frac{1}{\gamma_n^2} \sum_{i=1}^{\gamma_n} \sum_{k=1, k \neq i}^{\gamma_n} \mathbb{E}\left[\left\{ \left(m_j(X_i^{(j)}) - m_j(X_{\pi(i)}^{(j)})\right)^2 - I(j) \right\} \right. \\ &\quad \left. \left\{ \left(m_j(X_k^{(j)}) - m_j(X_{\pi(k)}^{(j)})\right)^2 - I(j) \right\} \right] \\ &= \frac{1}{\gamma_n} \mathbb{V}\text{ar}\left[\left(m_j(X_1^{(j)}) - m_j(X_2^{(j)})\right)^2 + 2\varepsilon_1 \left(m_j(X_1^{(j)}) - m_j(X_2^{(j)})\right)\right] \\ &+ \frac{\gamma_n(\gamma_n - 1)}{\gamma_n^2} \mathbb{P}(\pi_j(k) \neq i, \pi_j(i) \neq k) \mathbb{E}\left[\left(m_j(X_1^{(j)}) - m_j(X_2^{(j)})\right)^2 - I(j)\right]^2 \\ &+ \frac{\gamma_n(\gamma_n - 1)}{\gamma_n^2} \mathbb{P}(\pi_j(k) \neq i, \pi_j(i) = k) \\ &\cdot \mathbb{E}\left[\left\{ \left(m_j(X_1^{(j)}) - m_j(X_2^{(j)})\right)^2 - I(j) \right\} \left\{ \left(m_j(X_1^{(j)}) - m_j(X_3^{(j)})\right)^2 - I(j) \right\} \right] \\ &+ \mathcal{O}\left(\frac{1}{\gamma_n^2}\right) \\ &= \frac{1}{\gamma_n} \mathbb{V}\text{ar}\left[\left(m_j(X_1^{(j)}) - m_j(X_2^{(j)})\right)^2 + 2\varepsilon_1 \left(m_j(X_1^{(j)}) - m_j(X_2^{(j)})\right)\right]\end{aligned}$$

$$\begin{aligned}
& + \frac{\gamma_n(\gamma_n - 1)}{\gamma_n^2} \frac{2}{\gamma_n - 1} \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
& \quad \left. \left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& + \mathcal{O} \left(\frac{1}{\gamma_n^2} \right) \\
& = \frac{1}{\gamma_n} \mathbb{V} \text{ar} \left[\left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) \right] + 0 \\
& + \frac{2}{\gamma_n} \mathbb{C} \text{ov} \left[\left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2, \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 \right] \\
& + \mathcal{O} \left(\frac{1}{\gamma_n} \right). \tag{17}
\end{aligned}$$

Without loss of generality we can assume $\mathbb{E} \left[m_j \left(X_1^{(j)} \right) \right] = 0$, because as in previous calculations we get

$$\begin{aligned}
h_{\gamma_n} &= \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \left\{ \left(Y_i - \tilde{m} \left(X_i^{\pi_j} \right) \right)^2 - \varepsilon_i^2 \right\} \\
&= \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \left\{ \left(m_j \left(X_i^{(j)} \right) + \varepsilon_i - m_j \left(X_{\pi(i)}^{(j)} \right) \right)^2 - \varepsilon_i^2 \right\}.
\end{aligned}$$

Thus, the expectation of $m_j \left(X_i^{(j)} \right)$ and $m_j \left(X_{\pi(i)}^{(j)} \right)$ are cancelling each other. For random variables U, V and W with $U, V \perp\!\!\!\perp W$

$$\mathbb{C} \text{ov} (UW, V) = \mathbb{E} [UWV] - \mathbb{E} [UW] \mathbb{E} [V] = \mathbb{E} [W] \mathbb{C} \text{ov} (U, V) \tag{18}$$

holds, which implies

$$\begin{aligned}
& \mathbb{C} \text{ov} \left[\left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2, \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 \right] \\
&= \mathbb{C} \text{ov} \left[m_j \left(X_1^{(j)} \right)^2, m_j \left(X_1^{(j)} \right)^2 \right] - 4 \mathbb{C} \text{ov} \left[m_j \left(X_1^{(j)} \right)^2, m_j \left(X_1^{(j)} \right) m_j \left(X_2^{(j)} \right) \right] \\
&+ 4 \mathbb{C} \text{ov} \left[m_j \left(X_1^{(j)} \right) m_j \left(X_2^{(j)} \right), m_j \left(X_1^{(j)} \right) m_j \left(X_2^{(j)} \right) \right] \\
&= \mathbb{V} \text{ar} \left[m_j \left(X_1^{(j)} \right)^2 \right]
\end{aligned}$$

due to the i.i.d. assumption. Using (18) and $\mathbb{E}[\varepsilon_1] = 0$ leads to

$$\begin{aligned}
& \mathbb{V}\text{ar} \left[\left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) \right] \\
&= \mathbb{V}\text{ar} \left[\left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 \right] + \mathbb{V}\text{ar} \left[2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) \right] \\
&+ 2\mathbb{C}\text{ov} \left[\left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2, 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) \right] \\
&= \mathbb{V}\text{ar} \left[\left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 \right] + 4\mathbb{V}\text{ar} \left[\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) \right].
\end{aligned}$$

Hence,

$$\begin{aligned}
& \mathbb{E} \left[|h_{\gamma_n} - I(j)|^2 \right] \\
&= \frac{2}{\gamma_n} \mathbb{V}\text{ar} \left[m_j \left(X_1^{(j)} \right)^2 \right] + \frac{1}{\gamma_n} \mathbb{V}\text{ar} \left[\left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 \right] \\
&+ \frac{4}{\gamma_n} \mathbb{V}\text{ar} \left[\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) \right] + \mathcal{O} \left(\frac{1}{\gamma_n^2} \right). \tag{19}
\end{aligned}$$

Note that

$$\mathbb{E} \left[|h_{\gamma_n} - I(j)|^2 \right] \geq \frac{2}{\gamma_n} \mathbb{V}\text{ar} \left[m_j \left(X_1^{(j)} \right)^2 \right] + \mathcal{O} \left(\frac{1}{\gamma_n^2} \right). \tag{20}$$

Since $\mathbb{V}\text{ar} \left[m_j \left(X_1^{(j)} \right) \right] > 0$, it also holds true that $\mathbb{V}\text{ar} \left[m_j \left(X_1^{(j)} \right)^2 \right] > 0$ and with (20) there is a constant $c_1 > 0$ such that $\mathbb{E} \left[|h_{\gamma_n} - I(j)|^2 \right] \sim \frac{c_1}{\gamma_n}$ and $\mathbb{E} \left[|h_{\gamma_n} - I(j)|^2 \right]^2 \sim \frac{c_1^2}{\gamma_n^2}$. With similar calculations as before, we get

$$\begin{aligned}
& \mathbb{E} \left[|h_{\gamma_n} - I(j)|^4 \right] \\
&= \mathbb{E} \left[\left| \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \left\{ \left(m_j \left(X_i^{(j)} \right) - m_j \left(X_{\pi(i)}^{(j)} \right) \right)^2 + 2\varepsilon_i \left(m_j \left(X_i^{(j)} \right) - m_j \left(X_{\pi(i)}^{(j)} \right) \right) \right\} - I(j) \right|^4 \right] \\
&= \mathcal{O} \left(\frac{1}{\gamma_n^2} \right) + \frac{1}{\gamma_n^4} \sum_{i_1=1}^{\gamma_n} \sum_{i_2=1}^{\gamma_n} \sum_{i_3=1}^{\gamma_n} \sum_{i_2 \neq i_1, i_3 \neq i_2 \neq i_1} \\
&\mathbb{E} \left[\left\{ \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_1} \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right) - I(j) \right\}^2 \right]
\end{aligned}$$

$$\begin{aligned}
& \left\{ \left(m_j \left(X_{i_2}^{(j)} \right) - m_j \left(X_{\pi(i_2)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_2} \left(m_j \left(X_{i_2}^{(j)} \right) - m_j \left(X_{\pi(i_2)}^{(j)} \right) \right) - I(j) \right\} \\
& \left\{ \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_3)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_3} \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_3)}^{(j)} \right) \right) - I(j) \right\} \Big] \\
& + \frac{1}{\gamma_n^4} \sum_{i_1=1}^{\gamma_n} \sum_{i_2=1}^{\gamma_n} \sum_{i_3=1}^{\gamma_n} \sum_{\substack{i_4=1; i_4 \neq i_2 \\ i_2 \neq i_1, i_3 \neq i_2 \neq i_1, i_4 \neq i_1 \neq i_3}}^{\gamma_n} \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_1} \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right) - I(j) \right\} \right. \\
& \left\{ \left(m_j \left(X_{i_2}^{(j)} \right) - m_j \left(X_{\pi(i_2)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_2} \left(m_j \left(X_{i_2}^{(j)} \right) - m_j \left(X_{\pi(i_2)}^{(j)} \right) \right) - I(j) \right\} \\
& \left\{ \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_3)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_3} \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_3)}^{(j)} \right) \right) - I(j) \right\} \\
& \left. \left\{ \left(m_j \left(X_{i_4}^{(j)} \right) - m_j \left(X_{\pi(i_4)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_4} \left(m_j \left(X_{i_4}^{(j)} \right) - m_j \left(X_{\pi(i_4)}^{(j)} \right) \right) - I(j) \right\} \right] \\
& = \mathcal{O} \left(\frac{1}{\gamma_n^2} \right) + \frac{3! \gamma_n (\gamma_n - 1) (\gamma_n - 2)}{\gamma_n^4} \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_i \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \left. \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& + \frac{4! \gamma_n (\gamma_n - 1) (\gamma_n - 2) (\gamma_n - 3)}{\gamma_n^4} \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
& \left. \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \right],
\end{aligned}$$

where the second equality holds true because all summands of the remaining $\mathcal{O}(\gamma_n^2)$ summand are uniformly bounded. With the factor $1/\gamma_n^4$ the remainder is $\mathcal{O}(1/\gamma_n^2)$.

To calculate the expectations, we require some basis arguments of combinatorics. In particular, we only need to consider derangements such that we cannot isolate one single factor of the product regarding independence to the other factors. The latter is because each single factor itself is centered. Additionally, we can make use of the symmetry of the product leading to respective constant factor in front of the probabilities. To keep the proofs as simple as possible, we do not specify these factors more than that they are finite and positive. In this spirit,

for constants $c_2, \dots, c_7 \in (0, \infty)$ we have:

$$\begin{aligned}
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_i \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left. \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& = c_2 \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 1) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 + 2\varepsilon_i \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left. \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_1^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& + c_3 \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) > 3) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 + 2\varepsilon_i \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left. \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& + c_4 \mathbb{P}(\pi(1) = 2, \pi(2) > 3, \pi(3) = 1) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 + 2\varepsilon_i \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left. \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_1^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& + c_5 \mathbb{P}(\pi(1) > 3, \pi(2) > 3, \pi(3) = 2) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 + 2\varepsilon_i \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left. \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_5^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& = \mathcal{O} \left(\frac{1}{\gamma_n} \right), \tag{21}
\end{aligned}$$

since m_j , and therefore also the expectation, is bounded,

$$\mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 1) = \mathcal{O} \left(\frac{1}{\gamma_n^3} \right),$$

$$\mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) > 3) = \mathcal{O} \left(\frac{1}{\gamma_n^2} \right),$$

$$\mathbb{P}(\pi(1) = 2, \pi(2) > 3, \pi(3) = 1) = \mathcal{O} \left(\frac{1}{\gamma_n^2} \right)$$

and

$$\mathbb{P}(\pi(1) > 3, \pi(2) > 3, \pi(3) = 2) = \mathcal{O} \left(\frac{1}{\gamma_n} \right).$$

Furthermore

$$\begin{aligned}
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
& \quad \left. \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& = c_6 \mathbb{P}(\pi(1) = 2, \pi(2) \notin \{3, 4\}, \pi(3) = 4, \pi(4) \notin \{1, 2\}) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \middle| \pi(2) \notin \{3, 4\}, \pi(4) \notin \{1, 2\} \right] \\
& + c_7 \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) > 4) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
& \quad \left. \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_5^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& + \mathcal{O} \left(\frac{1}{\gamma_n^4} \right) \\
& = \mathcal{O} \left(\frac{1}{\gamma_n^2} \right), \tag{22}
\end{aligned}$$

because

$$\begin{aligned}
& \mathbb{P}(\pi(1) = 2, \pi(2) \notin \{3, 4\}, \pi(3) = 4, \pi(4) \notin \{1, 2\}) \\
& \leq \mathbb{P}(\pi(1) = 2, \pi(3) = 4, \pi(4) > 4) = \frac{!(\gamma_n - 2)}{!\gamma_n} = \frac{1}{\gamma_n^2} + o(1)
\end{aligned}$$

and

$$\begin{aligned}
& \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) > 4) \\
& \leq \mathbb{P}(\pi(1) = 2, \pi(3) = 4) = \frac{!(\gamma_n - 2)}{!\gamma_n} = \frac{1}{\gamma_n^2} + o(1).
\end{aligned}$$

Hence, with (21) and (22)

$$\mathbb{E} \left[|h_{\gamma_n} - I(j)|^4 \right] = \mathcal{O} \left(\frac{1}{\gamma_n^2} \right)$$

holds true. Since $\mathbb{E} \left[|h_{\gamma_n} - I(j)|^2 \right]^2 \sim \frac{c_1^2}{\gamma_n^2}$, it also holds that

$$\frac{\mathbb{E} \left[|h_{\gamma_n} - I(j)|^4 \right]}{\mathbb{E} \left[|h_{\gamma_n} - I(j)|^2 \right]^2} = \mathcal{O}(1)$$

is uniformly bounded. With Hölder inequality, we also have

$$\mathbb{E} \left[|h_{\gamma_n} - I(j)|^3 \right]^2 \leq \mathbb{E} \left[|h_{\gamma_n} - I(j)|^4 \right]^{3/2} \leq c_8 \left(\frac{1}{\gamma_n^2} \right)^{3/2} = \frac{c_8}{\gamma_n^3}$$

and

$$\mathbb{E} \left[|h_{\gamma_n} - I(j)|^3 \right]^2 \geq \mathbb{E} \left[|h_{\gamma_n} - I(j)|^2 \right]^3 \geq \frac{c_9}{\gamma_n^3}$$

for constants $c_8, c_9 > 0$. To show that (14) is uniformly bounded, it only remains to show that

$$\mathbb{E} \left[|h_{\gamma_n} - I(j)|^6 \right] = \mathcal{O} \left(\frac{1}{\gamma_n^3} \right).$$

In the same way as for the fourth moment, (A2), Lemma A.3, and the i.i.d assumptions can be used to receive

$$\begin{aligned} & \mathbb{E} \left[|h_{\gamma_n} - I(j)|^6 \right] \\ &= \mathbb{E} \left[\left| \frac{1}{\gamma_n} \sum_{i=1}^{\gamma_n} \left\{ \left(m_j \left(X_i^{(j)} \right) - m_j \left(X_{\pi(i)}^{(j)} \right) \right)^2 \right. \right. \right. \\ & \quad \left. \left. \left. + 2\varepsilon_i \left(m_j \left(X_i^{(j)} \right) - m_j \left(X_{\pi(i)}^{(j)} \right) \right) \right\} - I(j) \right|^6 \right] \\ &= \mathcal{O} \left(\frac{1}{\gamma_n^3} \right) + \frac{1}{\gamma_n^6} \sum_{\substack{i, i_2, i_3, i_4 \\ \text{pairwise distinct}}} \\ & \mathbb{E} \left[\left\{ \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_1} \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right) - I(j) \right\}^3 \right. \\ & \quad \left\{ \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_3)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_3} \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_4)}^{(j)} \right) \right) - I(j) \right\} \\ & \quad \left. \left\{ \left(m_j \left(X_{i_4}^{(j)} \right) - m_j \left(X_{\pi(i_4)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_4} \left(m_j \left(X_{i_4}^{(j)} \right) - m_j \left(X_{\pi(i_4)}^{(j)} \right) \right) - I(j) \right\} \right] \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{\gamma_n^6} \sum_{\substack{i_1, i_2, i_3, i_4 \\ \text{pairwise distinct}}} \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_1} \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left\{ \left(m_j \left(X_{i_2}^{(j)} \right) - m_j \left(X_{\pi(i_2)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_2} \left(m_j \left(X_{i_2}^{(j)} \right) - m_j \left(X_{\pi(i_2)}^{(j)} \right) \right) - I(j) \right\}^2 \\
& \quad \left\{ \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_3)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_3} \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_3)}^{(j)} \right) \right) - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_{i_4}^{(j)} \right) - m_j \left(X_{\pi(i_4)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_4} \left(m_j \left(X_{i_4}^{(j)} \right) - m_j \left(X_{\pi(i_4)}^{(j)} \right) \right) - I(j) \right\} \right] \\
& + \frac{1}{\gamma_n^6} \sum_{\substack{i_1, i_2, i_3, i_4, i_5 \\ \text{pairwise distinct}}} \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_1} \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left\{ \left(m_j \left(X_{i_2}^{(j)} \right) - m_j \left(X_{\pi(i_2)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_2} \left(m_j \left(X_{i_2}^{(j)} \right) - m_j \left(X_{\pi(i_2)}^{(j)} \right) \right) - I(j) \right\} \\
& \quad \left\{ \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_3)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_3} \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_3)}^{(j)} \right) \right) - I(j) \right\} \\
& \quad \left\{ \left(m_j \left(X_{i_4}^{(j)} \right) - m_j \left(X_{\pi(i_4)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_4} \left(m_j \left(X_{i_4}^{(j)} \right) - m_j \left(X_{\pi(i_4)}^{(j)} \right) \right) - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_{i_5}^{(j)} \right) - m_j \left(X_{\pi(i_5)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_5} \left(m_j \left(X_{i_5}^{(j)} \right) - m_j \left(X_{\pi(i_5)}^{(j)} \right) \right) - I(j) \right\} \right] \\
& + \frac{1}{\gamma_n^6} \sum_{\substack{i_1, i_2, i_3, i_4, i_5, i_6 \\ \text{pairwise distinct}}} \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_1} \left(m_j \left(X_{i_1}^{(j)} \right) - m_j \left(X_{\pi(i_1)}^{(j)} \right) \right) - I(j) \right\} \right. \\
& \quad \left\{ \left(m_j \left(X_{i_2}^{(j)} \right) - m_j \left(X_{\pi(i_2)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_2} \left(m_j \left(X_{i_2}^{(j)} \right) - m_j \left(X_{\pi(i_2)}^{(j)} \right) \right) - I(j) \right\} \\
& \quad \left\{ \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_3)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_3} \left(m_j \left(X_{i_3}^{(j)} \right) - m_j \left(X_{\pi(i_3)}^{(j)} \right) \right) - I(j) \right\} \\
& \quad \left\{ \left(m_j \left(X_{i_4}^{(j)} \right) - m_j \left(X_{\pi(i_4)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_4} \left(m_j \left(X_{i_4}^{(j)} \right) - m_j \left(X_{\pi(i_4)}^{(j)} \right) \right) - I(j) \right\} \\
& \quad \left\{ \left(m_j \left(X_{i_5}^{(j)} \right) - m_j \left(X_{\pi(i_5)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_5} \left(m_j \left(X_{i_5}^{(j)} \right) - m_j \left(X_{\pi(i_5)}^{(j)} \right) \right) - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_{i_6}^{(j)} \right) - m_j \left(X_{\pi(i_6)}^{(j)} \right) \right)^2 + 2\varepsilon_{i_6} \left(m_j \left(X_{i_6}^{(j)} \right) - m_j \left(X_{\pi(i_6)}^{(j)} \right) \right) - I(j) \right\} \right] \\
& = \mathcal{O} \left(\frac{1}{\gamma_n^3} \right)
\end{aligned}$$

$$\begin{aligned}
& + \frac{4!(\gamma_n-1)(\gamma_n-2)(\gamma_n-3)}{\gamma_n^5} \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^3 \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& + \frac{4!(\gamma_n-1)(\gamma_n-2)(\gamma_n-3)}{\gamma_n^5} \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 + 2\varepsilon_2 \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right) - I(j) \right\}^2 \\
& \quad \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& + \frac{5!(\gamma_n-1)(\gamma_n-2)(\gamma_n-3)(\gamma_n-4)}{\gamma_n^5} \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_{\pi(5)}^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& + \frac{6!(\gamma_n-1)(\gamma_n-2)(\gamma_n-3)(\gamma_n-4)(\gamma_n-5)}{\gamma_n^5} \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
& \quad \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_{\pi(5)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_6^{(j)} \right) - m_j \left(X_{\pi(6)}^{(j)} \right) \right)^2 - I(j) \right\} \right].
\end{aligned}$$

Continuing with the calculation of expectations, we again rely on some basic principles of combinatorics. In particular, we consider derangements to ensure that no single factor of the product can be isolated, as this would affect the independence of the other factors. This is because each individual factor is centered. Moreover, we exploit the symmetry of the product, which implies a constant factor in front of the probabilities. To keep the proofs simple, we only

specify these factors as finite and positive. In this context, the following applies to the constants $\tilde{c}_1, \dots, \tilde{c}_{31}$:

$$\begin{aligned}
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^3 \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
&= \tilde{c}_1 \mathbb{P}(\pi(1) = 2, \pi(2) \notin \{3, 4\}, \pi(3) = 4, \pi(4) \notin \{1, 2\}) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) - I(j) \right\}^3 \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \right] \Big| \pi(2) \notin \{3, 4\}, \pi(4) \notin \{1, 2\} \\
&+ \tilde{c}_2 \mathbb{P}(\pi(1) \notin \{3, 4\}, \pi(2) = 1, \pi(3) = 4, \pi(4) \notin \{1, 2\}) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^3 \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_1^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \right] \Big| \pi(1) \notin \{3, 4\}, \pi(4) \notin \{1, 2\} \\
&+ \tilde{c}_3 \mathbb{P}(\pi(1) \notin \{2, 3, 4\}, \pi(2) = 3, \pi(3) = 4, \pi(4) \neq 1) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^3 \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \right] \Big| \pi(1) \notin \{3, 4\}, \pi(4) \notin \{1, 2\} \\
&+ \tilde{c}_4 \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) \neq 4, \pi(4) = 1) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) - I(j) \right\}^3 \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \right]
\end{aligned}$$

$$\begin{aligned}
& \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_1^{(j)} \right) \right)^2 - I(j) \right\} \Big| \pi(3) \neq 4 \Big] \\
& + \tilde{c}_5 \mathbb{P}(\pi(1) = 4, \pi(2) = 3, \pi(3) = 1, \pi(4) \notin \{2, 3\}) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right) - I(j) \right\}^3 \right. \\
& \left. \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_1^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
& \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \Big| \pi(4) \notin \{2, 3\} \right] \\
& + \tilde{c}_6 \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) \notin \{2, 3\}) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) - I(j) \right\}^3 \right. \\
& \left. \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
& \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \Big| \pi(4) \notin \{2, 3\} \right] \\
& + \tilde{c}_7 \mathbb{P}(\pi(1) \neq 4, \pi(2) = 1, \pi(3) = 2, \pi(4) = 3) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^3 \right. \\
& \left. \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_1^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
& \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \Big| \pi(1) \neq 4 \right] \\
& + \tilde{c}_8 \mathbb{P}(\pi(1) = 4, \pi(2) = 1, \pi(3) = 2, \pi(4) = 3) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right) - I(j) \right\}^3 \right. \\
& \left. \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_1^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
& \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& = \mathcal{O} \left(\frac{1}{\gamma_n^2} \right),
\end{aligned}$$

since

$$\begin{aligned}
& \max \left\{ \mathbb{P}(\pi(1) = 2, \pi(2) \notin \{3, 4\}, \pi(3) = 4, \pi(4) \notin \{1, 2\}), \right. \\
& \mathbb{P}(\pi(1) \notin \{3, 4\}, \pi(2) = 1, \pi(3) = 4, \pi(4) \notin \{1, 2\}), \\
& \mathbb{P}(\pi(1) \notin \{2, 3, 4\}, \pi(2) = 3, \pi(3) = 4, \pi(4) \neq 1), \\
& \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) \neq 4, \pi(4) = 1), \\
& \mathbb{P}(\pi(1) = 4, \pi(2) = 3, \pi(3) = 1, \pi(4) \notin \{2, 3\}), \\
& \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) \notin \{2, 3\}), \\
& \mathbb{P}(\pi(1) \neq 4, \pi(2) = 1, \pi(3) = 2, \pi(4) = 3), \\
& \left. \mathbb{P}(\pi(1) = 4, \pi(2) = 1, \pi(3) = 2, \pi(4) = 3) \right\} \\
& \leq \mathbb{P}(\pi(1) = 2, \pi(2) = 3) = \frac{!(\gamma_n - 2)}{!\gamma_n} = \mathcal{O}\left(\frac{1}{\gamma_n^2}\right)
\end{aligned}$$

as in (15) or (16). Analogously, the second four fold sum is $\mathcal{O}\left(\frac{1}{\gamma_n}\right)$, as we can see with

$$\begin{aligned}
& \mathbb{E} \left[\left\{ \left(m_j(X_1^{(j)}) - m_j(X_{\pi(1)}^{(j)}) \right)^2 + 2\varepsilon_1 \left(m_j(X_1^{(j)}) - m_j(X_{\pi(1)}^{(j)}) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j(X_2^{(j)}) - m_j(X_{\pi(2)}^{(j)}) \right)^2 + 2\varepsilon_2 \left(m_j(X_2^{(j)}) - m_j(X_{\pi(2)}^{(j)}) \right) - I(j) \right\}^2 \\
& \left. \left\{ \left(m_j(X_3^{(j)}) - m_j(X_{\pi(3)}^{(j)}) \right)^2 - I(j) \right\} \left\{ \left(m_j(X_4^{(j)}) - m_j(X_{\pi(4)}^{(j)}) \right)^2 - I(j) \right\} \right] \\
& = \tilde{c}_{10} \mathbb{P}(\pi(1) = 3, \pi(2) = 4, \pi(3) \notin \{2, 4\}, \pi(4) \notin \{1, 3\}) \\
& \mathbb{E} \left[\left\{ \left(m_j(X_1^{(j)}) - m_j(X_3^{(j)}) \right)^2 + 2\varepsilon_1 \left(m_j(X_1^{(j)}) - m_j(X_3^{(j)}) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j(X_2^{(j)}) - m_j(X_4^{(j)}) \right)^2 + 2\varepsilon_2 \left(m_j(X_2^{(j)}) - m_j(X_4^{(j)}) \right) - I(j) \right\}^2 \\
& \left\{ \left(m_j(X_3^{(j)}) - m_j(X_{\pi(3)}^{(j)}) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j(X_4^{(j)}) - m_j(X_{\pi(4)}^{(j)}) \right)^2 - I(j) \right\} \right] \Big| \pi(3) \notin \{2, 4\}, \pi(4) \notin \{1, 3\} \Big] \\
& + \tilde{c}_{11} \mathbb{P}(\pi(1) = 3, \pi(2) \notin \{1, 3\}, \pi(3) \notin \{2, 4\}, \pi(4) = 2) \\
& \mathbb{E} \left[\left\{ \left(m_j(X_1^{(j)}) - m_j(X_3^{(j)}) \right)^2 + 2\varepsilon_1 \left(m_j(X_1^{(j)}) - m_j(X_3^{(j)}) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j(X_2^{(j)}) - m_j(X_{\pi(2)}^{(j)}) \right)^2 + 2\varepsilon_2 \left(m_j(X_2^{(j)}) - m_j(X_{\pi(2)}^{(j)}) \right) - I(j) \right\}^2 \\
& \left. \left\{ \left(m_j(X_3^{(j)}) - m_j(X_{\pi(3)}^{(j)}) \right)^2 - I(j) \right\} \right]
\end{aligned}$$

$$\begin{aligned}
& \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \Big| \pi(2) \notin \{1, 3\}, \pi(3) \notin \{2, 4\} \Big] \\
& + \tilde{c}_{12} \mathbb{P}(\pi(1) \notin \{2, 4\}, \pi(2) \notin \{1, 3\}, \pi(3) = 1, \pi(4) = 2) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 + 2\varepsilon_2 \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right) - I(j) \right\}^2 \\
& \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_1^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \Big| \pi(1) \notin \{2, 4\}, \pi(2) \notin \{1, 3\} \right] \\
& + \tilde{c}_{13} \mathbb{P}(\pi(1) \notin \{2, 3, 4\}, \pi(2) = 3, \pi(3) = 4, \pi(4) \neq 1) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 + 2\varepsilon_2 \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right) - I(j) \right\}^2 \\
& \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \Big| \pi(1) \notin \{2, 3, 4\}, \pi(4) \neq 1 \right] \\
& + \tilde{c}_{14} \mathbb{P}(\pi(1) \notin \{2, 3, 4\}, \pi(2) \neq 1, \pi(3) = 4, \pi(4) = 2) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 + 2\varepsilon_2 \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right) - I(j) \right\}^2 \\
& \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \Big| \pi(1) \notin \{2, 3, 4\}, \pi(2) \neq 1 \right] \\
& + \tilde{c}_{15} \mathbb{P}(\pi(1) \notin \{2, 3, 4\}, \pi(2) = 3, \pi(3) = 4, \pi(4) = 2) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 + 2\varepsilon_2 \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right) - I(j) \right\}^2 \\
& \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \Big| \pi(1) \notin \{2, 3, 4\} \right]
\end{aligned}$$

$$\begin{aligned}
& + \tilde{c}_{16} \mathbb{P}(\pi(1) \notin \{2, 3, 4\}, \pi(2) = 3, \pi(3) \notin \{1, 4\}, \pi(4) = 2) \\
& \mathbb{E} \left[\left\{ \left(m_j(X_1^{(j)}) - m_j(X_{\pi(1)}^{(j)}) \right)^2 + 2\varepsilon_1 \left(m_j(X_1^{(j)}) - m_j(X_{\pi(1)}^{(j)}) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j(X_2^{(j)}) - m_j(X_3^{(j)}) \right)^2 + 2\varepsilon_2 \left(m_j(X_2^{(j)}) - m_j(X_3^{(j)}) \right) - I(j) \right\}^2 \\
& \left\{ \left(m_j(X_3^{(j)}) - m_j(X_{\pi(3)}^{(j)}) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j(X_4^{(j)}) - m_j(X_2^{(j)}) \right)^2 - I(j) \right\} \middle| \pi(1) \notin \{2, 3, 4\}, \pi(3) \notin \{1, 4\} \right] \\
& + \tilde{c}_{17} \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) \notin \{1, 2\}) \\
& \mathbb{E} \left[\left\{ \left(m_j(X_1^{(j)}) - m_j(X_2^{(j)}) \right)^2 + 2\varepsilon_1 \left(m_j(X_1^{(j)}) - m_j(X_2^{(j)}) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j(X_2^{(j)}) - m_j(X_3^{(j)}) \right)^2 + 2\varepsilon_2 \left(m_j(X_2^{(j)}) - m_j(X_3^{(j)}) \right) - I(j) \right\}^2 \\
& \left\{ \left(m_j(X_3^{(j)}) - m_j(X_4^{(j)}) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j(X_4^{(j)}) - m_j(X_{\pi(4)}^{(j)}) \right)^2 - I(j) \right\} \middle| \pi(4) \notin \{1, 2\} \right] \\
& + \tilde{c}_{18} \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) = 1) \\
& \mathbb{E} \left[\left\{ \left(m_j(X_1^{(j)}) - m_j(X_2^{(j)}) \right)^2 + 2\varepsilon_1 \left(m_j(X_1^{(j)}) - m_j(X_2^{(j)}) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j(X_2^{(j)}) - m_j(X_3^{(j)}) \right)^2 + 2\varepsilon_2 \left(m_j(X_2^{(j)}) - m_j(X_3^{(j)}) \right) - I(j) \right\}^2 \\
& \left\{ \left(m_j(X_3^{(j)}) - m_j(X_4^{(j)}) \right)^2 - I(j) \right\} \left\{ \left(m_j(X_4^{(j)}) - m_j(X_1^{(j)}) \right)^2 - I(j) \right\} \right] \\
& = \mathcal{O} \left(\frac{1}{\gamma_n} \right),
\end{aligned}$$

because all the expectations are finite and

$$\begin{aligned}
& \max \left\{ \mathbb{P}(\pi(1) \notin \{3, 4\}, \pi(2) \notin \{3, 4\}, \pi(3) = 4, \pi(4) \notin \{1, 2\}), \right. \\
& \mathbb{P}(\pi(1) = 3, \pi(2) = 4, \pi(3) \notin \{2, 4\}, \pi(4) \notin \{1, 3\}), \\
& \mathbb{P}(\pi(1) = 3, \pi(2) \notin \{1, 3\}, \pi(3) \notin \{2, 4\}, \pi(4) = 2), \\
& \mathbb{P}(\pi(1) \notin \{2, 4\}, \pi(2) \notin \{1, 3\}, \pi(3) = 1, \pi(4) = 2), \\
& \left. \mathbb{P}(\pi(1) \notin \{2, 3, 4\}, \pi(2) = 3, \pi(3) = 4, \pi(4) \neq 1), \right\}
\end{aligned}$$

$$\begin{aligned}
& \mathbb{P}(\pi(1) \notin \{2, 3, 4\}, \pi(2) \neq 1, \pi(3) = 4, \pi(4) = 2) \\
& \mathbb{P}(\pi(1) \notin \{2, 3, 4\}, \pi(2) = 3, \pi(3) = 4, \pi(4) = 2), \\
& \mathbb{P}(\pi(1) \notin \{2, 3, 4\}, \pi(2) = 3, \pi(3) \notin \{1, 4\}, \pi(4) = 2) \\
& \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) \notin \{1, 2\}), \\
& \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) = 1) \Big\} \\
& \leq \mathbb{P}(\pi(3) = 4) = \mathcal{O}\left(\frac{1}{\gamma_n}\right).
\end{aligned}$$

In addition

$$\begin{aligned}
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_{\pi(5)}^{(j)} \right) \right)^2 - I(j) \right\} \right] \\
& = \tilde{c}_{19} \mathbb{P}(\pi(1) \notin \{2, 3, 4, 5\}, \pi(2) = 3, \pi(3) \notin \{1, 4, 5\}, \pi(4) = 5, \pi(5) \notin \{1, 2, 3\}) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_5^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_{\pi(5)}^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left. \left| \pi(1) \notin \{2, 3, 4, 5\}, \pi(3) \notin \{1, 4, 5\}, \pi(5) \notin \{1, 2, 3\} \right] \right. \\
& + \tilde{c}_{20} \mathbb{P}(\pi(1) \notin \{2, 3, 4, 5\}, \pi(2) = 3, \pi(3) = 4, \pi(4) = 5, \pi(5) \neq 1) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \quad \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_5^{(j)} \right) \right)^2 - I(j) \right\} \\
& \quad \left. \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_{\pi(5)}^{(j)} \right) \right)^2 - I(j) \right\} \left| \pi(1) \notin \{2, 3, 4, 5\}, \pi(5) \notin \{1\} \right] \right. \\
& + \tilde{c}_{21} \mathbb{P}(\pi(1) \notin \{3, 4, 5\}, \pi(2) = 1, \pi(3) = 4, \pi(4) = 5, \pi(5) \notin \{1, 2\}) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right.
\end{aligned}$$

$$\begin{aligned}
& \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_1^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_5^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_{\pi(5)}^{(j)} \right) \right)^2 - I(j) \right\} \left| \pi(1) \notin \{3, 4, 5\}, \pi(5) \notin \{1, 2, 3\} \right] \\
& + \tilde{c}_{22} \mathbb{P} \left(\pi(1) = 2, \pi(2) \notin \{3, 4, 5\}, \pi(3) = 4, \pi(4) = 5, \pi(5) \notin \{1, 2\} \right) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_5^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_{\pi(5)}^{(j)} \right) \right)^2 - I(j) \right\} \left| \pi(2) \notin \{3, 4, 5\}, \pi(5) \notin \{1, 2, 3\} \right] \right. \\
& + \tilde{c}_{23} \mathbb{P} \left(\pi(1) = 2, \pi(2) = 3, \pi(3) \notin \{4, 5\}, \pi(4) = 5, \pi(5) \notin \{1, 2, 3\} \right) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_5^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_{\pi(5)}^{(j)} \right) \right)^2 - I(j) \right\} \left| \pi(3) \notin \{4, 5\}, \pi(5) \notin \{1, 2, 3\} \right] \right. \\
& + \tilde{c}_{24} \mathbb{P} \left(\pi(1) \notin \{4, 5\}, \pi(2) = 1, \pi(3) = 2, \pi(4) = 5, \pi(5) \notin \{1, 2, 3\} \right) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_{\pi(1)}^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_1^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_5^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_{\pi(5)}^{(j)} \right) \right)^2 - I(j) \right\} \left| \pi(1) \notin \{4, 5\}, \pi(5) \notin \{1, 2, 3\} \right] \right. \\
& + \tilde{c}_{25} \mathbb{P} \left(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) = 5, \pi(5) \neq 1 \right) \\
& \mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 + 2\varepsilon_1 \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_5^{(j)} \right) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_{\pi(5)}^{(j)} \right) \right)^2 - I(j) \right\} \left| \pi(5) \neq 1 \right] \right.
\end{aligned}$$

$$\begin{aligned}
& + \tilde{c}_{26} \mathbb{P}(\pi(1) \notin \{2, 3, 4, 5\}, \pi(2) = 1, \pi(3) = 2, \pi(4) = 3, \pi(5) = 4) \\
& \mathbb{E} \left[\left\{ \left(m_j(X_1^{(j)}) - m_j(X_{\pi(1)}^{(j)}) \right)^2 + 2\varepsilon_1 \left(m_j(X_1^{(j)}) - m_j(X_{\pi(1)}^{(j)}) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j(X_2^{(j)}) - m_j(X_1^{(j)}) \right)^2 - I(j) \right\} \left\{ \left(m_j(X_3^{(j)}) - m_j(X_2^{(j)}) \right)^2 - I(j) \right\} \\
& \left\{ \left(m_j(X_4^{(j)}) - m_j(X_3^{(j)}) \right)^2 - I(j) \right\} \\
& \left. \left\{ \left(m_j(X_5^{(j)}) - m_j(X_4^{(j)}) \right)^2 - I(j) \right\} \middle| \pi(1) \notin \{2, 3, 4, 5\} \right] \\
& + \tilde{c}_{27} \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) = 5, \pi(5) = 1) \\
& \mathbb{E} \left[\left\{ \left(m_j(X_1^{(j)}) - m_j(X_2^{(j)}) \right)^2 + 2\varepsilon_1 \left(m_j(X_1^{(j)}) - m_j(X_2^{(j)}) \right) - I(j) \right\}^2 \right. \\
& \left\{ \left(m_j(X_2^{(j)}) - m_j(X_3^{(j)}) \right)^2 - I(j) \right\} \left\{ \left(m_j(X_3^{(j)}) - m_j(X_4^{(j)}) \right)^2 - I(j) \right\} \\
& \left\{ \left(m_j(X_4^{(j)}) - m_j(X_5^{(j)}) \right)^2 - I(j) \right\} \left\{ \left(m_j(X_5^{(j)}) - m_j(X_1^{(j)}) \right)^2 - I(j) \right\} \Big] \\
& = \mathcal{O}\left(\frac{1}{\gamma_n^2}\right),
\end{aligned}$$

because

$$\begin{aligned}
& \max \left\{ \mathbb{P}(\pi(1) \notin \{2, 3, 4, 5\}, \pi(2) = 3, \pi(3) \notin \{1, 4, 5\}, \pi(4) = 5, \pi(5) \notin \{1, 2, 3\}), \right. \\
& \quad \mathbb{P}(\pi(1) \notin \{2, 3, 4, 5\}, \pi(2) = 3, \pi(3) = 4, \pi(4) = 5, \pi(5) \neq 1), \\
& \quad \mathbb{P}(\pi(1) \notin \{3, 4, 5\}, \pi(2) = 1, \pi(3) = 4, \pi(4) = 5, \pi(5) \notin \{1, 2\}), \\
& \quad \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) \notin \{4, 5\}, \pi(4) = 5, \pi(5) \notin \{1, 2, 3\}), \\
& \quad \mathbb{P}(\pi(1) \notin \{4, 5\}, \pi(2) = 1, \pi(3) = 2, \pi(4) = 5, \pi(5) \notin \{1, 2, 3\}), \\
& \quad \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) = 5, \pi(5) \neq 1) \\
& \quad \mathbb{P}(\pi(1) \notin \{2, 3, 4, 5\}, \pi(2) = 1, \pi(3) = 2, \pi(4) = 3, \pi(5) = 4), \\
& \quad \left. \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) = 5, \pi(5) = 1) \right\} \\
& \leq \mathbb{P}(\pi(2) = 3, \pi(4) = 5) = \frac{!(\gamma_n - 2)}{!\gamma_n} = \mathcal{O}\left(\frac{1}{\gamma_n}\right)
\end{aligned}$$

as in (16) and (15). For the expected value resulting from the sixfold sum, we get

$$\begin{aligned}
& \mathbb{E} \left[\left\{ \left(m_j(X_1^{(j)}) - m_j(X_{\pi(1)}^{(j)}) \right)^2 - I(j) \right\} \left\{ \left(m_j(X_2^{(j)}) - m_j(X_{\pi(2)}^{(j)}) \right)^2 - I(j) \right\} \right. \\
& \left. \left\{ \left(m_j(X_3^{(j)}) - m_j(X_{\pi(3)}^{(j)}) \right)^2 - I(j) \right\} \left\{ \left(m_j(X_4^{(j)}) - m_j(X_{\pi(4)}^{(j)}) \right)^2 - I(j) \right\} \right]
\end{aligned}$$

$$\begin{aligned}
& \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_{\pi(5)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_6^{(j)} \right) - m_j \left(X_{\pi(6)}^{(j)} \right) \right)^2 - I(j) \right\} \\
&= \tilde{c}_{28} \mathbb{P}(\pi(1) = 2, \pi(2) \notin \{3, 4, 5, 6\}, \pi(3) = 4, \\
&\quad \pi(4) \notin \{1, 2, 5, 6\}, \pi(5) = 6, \pi(6) \notin \{1, 2, 3, 4\}) \\
&\mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_{\pi(2)}^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
&\quad \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \\
&\quad \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_6^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_6^{(j)} \right) - m_j \left(X_{\pi(6)}^{(j)} \right) \right)^2 - I(j) \right\} \\
&\quad \left. \left| \pi(2) \notin \{3, 4, 5, 6\}, \pi(4) \notin \{1, 2, 5, 6\}, \pi(6) \notin \{1, 2, 3, 4\} \right| \right] \\
&+ \tilde{c}_{29} \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) \notin \{4, 5, 6\}, \pi(4) = 5, \pi(5) = 6, \pi(6) \notin \{1, 2, 3\}) \\
&\mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
&\quad \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_{\pi(3)}^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_5^{(j)} \right) \right)^2 - I(j) \right\} \\
&\quad \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_6^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_6^{(j)} \right) - m_j \left(X_{\pi(6)}^{(j)} \right) \right)^2 - I(j) \right\} \\
&\quad \left. \left| \pi(3) \notin \{4, 5, 6\}, \pi(6) \notin \{1, 2, 3\} \right| \right] \\
&+ \tilde{c}_{30} \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) \notin \{5, 6\}, \pi(5) = 6, \pi(6) \notin \{1, 2, 3, 4\}) \\
&\mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
&\quad \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_{\pi(4)}^{(j)} \right) \right)^2 - I(j) \right\} \\
&\quad \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_6^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_6^{(j)} \right) - m_j \left(X_{\pi(6)}^{(j)} \right) \right)^2 - I(j) \right\} \\
&\quad \left. \left| \pi(4) \notin \{5, 6\}, \pi(6) \notin \{1, 2, 3, 4\} \right| \right] \\
&+ \tilde{c}_{31} \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) = 5, \pi(5) = 6, \pi(6) \notin \{1, 2, 3, 4, 5\}) \\
&\mathbb{E} \left[\left\{ \left(m_j \left(X_1^{(j)} \right) - m_j \left(X_2^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_2^{(j)} \right) - m_j \left(X_3^{(j)} \right) \right)^2 - I(j) \right\} \right. \\
&\quad \left\{ \left(m_j \left(X_3^{(j)} \right) - m_j \left(X_4^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_4^{(j)} \right) - m_j \left(X_5^{(j)} \right) \right)^2 - I(j) \right\} \\
&\quad \left\{ \left(m_j \left(X_5^{(j)} \right) - m_j \left(X_6^{(j)} \right) \right)^2 - I(j) \right\} \left\{ \left(m_j \left(X_6^{(j)} \right) - m_j \left(X_{\pi(6)}^{(j)} \right) \right)^2 - I(j) \right\} \\
&\quad \left. \left| \pi(6) \notin \{1, 2, 3, 4, 5\} \right| \right] \\
&= \mathcal{O} \left(\frac{1}{\gamma_n^3} \right),
\end{aligned}$$

since

$$\begin{aligned}
& \max \left\{ \mathbb{P}(\pi(1) = 2, \pi(2) \notin \{3, 4, 5, 6\}, \pi(3) = 4, \right. \\
& \quad \left. \pi(4) \notin \{1, 2, 5, 6\}, \pi(5) = 6, \pi(6) \notin \{1, 2, 3, 4\}), \right. \\
& \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) \notin \{4, 5, 6\}, \pi(4) = 5, \pi(5) = 6, \pi(6) \notin \{1, 2, 3\}), \\
& \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) \notin \{5, 6\}, \pi(5) = 6, \pi(6) \notin \{1, 2, 3, 4\}), \\
& \left. \mathbb{P}(\pi(1) = 2, \pi(2) = 3, \pi(3) = 4, \pi(4) = 5, \pi(5) = 6, \pi(6) \notin \{1, 2, 3, 4, 5\}) \right\} \\
& \leq \mathbb{P}(\pi(1) = 2, \pi(3) = 4, \pi(5) = 6) = \frac{!(\gamma_n - 3)}{\gamma_n} = \frac{1}{\gamma_n^3} + o(1).
\end{aligned}$$

This results in

$$\mathbb{E} \left[|h_{\gamma_n} - I(j)|^6 \right] = \mathcal{O} \left(\frac{1}{\gamma_n^3} \right),$$

which implies that

$$\frac{\mathbb{E} [|h_{\gamma_n} - I(j)|^6]}{\mathbb{E} [|h_{\gamma_n} - I(j)|^3]^2}$$

is uniformly bounded. Hence,

$$\frac{\mathbb{E} [|h_{\gamma_n} - I(j)|^{2k}]}{\mathbb{E} [|h_{\gamma_n} - I(j)|^k]^2}$$

is uniformly bounded for $k = 2, 3$. Due to (A1)-(A3), Lemma A.2 can be applied delivering

$$\frac{\gamma_n}{n} \frac{\zeta_{\gamma_n}}{\gamma_n \zeta_{1, \gamma_n}} \leq \frac{\gamma_n^2}{n} \frac{\tilde{K}}{\gamma_n^2 \zeta_{1, \gamma_n}}$$

for a $\tilde{K} > 0$. Lemma A.1 and $\gamma_n/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$ result in

$$\lim_{n \rightarrow \infty} \frac{\gamma_n}{n} \frac{\zeta_{\gamma_n}}{\gamma_n \zeta_{1, \gamma_n}} = 0.$$

Hence, Theorem 4.1 can be applied, which completes the proof. $\square$

Appendix B: Proof of Theorem 4.3

Proof. We start with the decomposition

$$\frac{I_{n, \hat{M}}^{OOB}(j) - I(j)}{\sqrt{\gamma_n^2 \zeta_{1, \gamma_n}/n + \zeta_{\gamma_n}/M_n}} = \frac{I_{n, \hat{M}}^{OOB}(j) - \tilde{I}_{n, \hat{M}}^{OOB}(j)}{\sqrt{\gamma_n^2 \zeta_{1, \gamma_n}/n + \zeta_{\gamma_n}/M_n}} + \frac{\tilde{I}_{n, \hat{M}}^{OOB}(j) - I(j)}{\sqrt{\gamma_n^2 \zeta_{1, \gamma_n}/n + \zeta_{\gamma_n}/M_n}}.$$

Due to Theorem 4.2, we already know that the second summand is asymptotically normal. Hence it remains to show that the first term vanishes for growing M and n . The first term in the denominator is asymptotically equivalent to κ/n by Lemma A.1. From the proof of Theorem 4.2, we can observe that there is a constant c_1 such that $\zeta_{\gamma_n} = \text{Var}(h_{\gamma_n}) = \mathbb{E}[|h_{\gamma_n} - I(j)|^2] \sim \frac{c_1}{\gamma_n}$. Thus $\frac{\zeta_{\gamma_n}}{M} \sim \frac{c_1}{\gamma_n M}$ holds true. The assumption $\frac{M\gamma_n \log(a_n)^2 \log(a_n p)}{a_n^{\phi(\lambda)}} \rightarrow 0$ implies $\frac{M\gamma_n}{a_n} \rightarrow 0$ as well as $a_n \sim n$, since $a_n = n - \gamma_n$. Hence,

$$\sqrt{\frac{\gamma_n^2 \zeta_{1, \gamma_n}}{n} + \frac{\zeta_{\gamma_n}}{M}} \sim \sqrt{\frac{c_1}{\gamma_n M}}.$$

For the numerator, we observe

$$\begin{aligned} & \left(I_{n,M}^{OOB}(j) - \tilde{I}_{n,M}^{OOB}(j) \right) \\ &= \frac{1}{M\gamma_n} \sum_{t=1}^M \sum_{i \in \mathcal{D}_n^{-(t)}} \left\{ (Y_i - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \boldsymbol{\Theta}_t, \mathcal{D}_n))^2 - (Y_i - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n))^2 \right\} \\ & \quad - \frac{1}{M\gamma_n} \sum_{t=1}^M \sum_{i \in \mathcal{D}_n^{-(t)}} \left\{ (Y_i - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 - (Y_i - \tilde{m}(\mathbf{X}_i))^2 \right\} \\ &= \frac{1}{M\gamma_n} \sum_{t=1}^M \sum_{i \in \mathcal{D}_n^{-(t)}} \left\{ m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \boldsymbol{\Theta}_t, \mathcal{D}_n)^2 - \tilde{m}(\mathbf{X}_i^{\pi_{j,t}})^2 - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)^2 \right. \\ & \quad \left. + \tilde{m}(\mathbf{X}_i)^2 + 2Y_i (\tilde{m}(\mathbf{X}_i^{\pi_{j,t}}) - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \boldsymbol{\Theta}_t, \mathcal{D}_n)) \right. \\ & \quad \left. - 2Y_i (\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)) \right\}. \quad (23) \end{aligned}$$

We will handle the terms in (23) separately and start with (A3), (A4) to get

$$\begin{aligned} & |\tilde{m}(\mathbf{X}_i)^2 - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)^2| \\ &= |\tilde{m}(\mathbf{X}_i) + m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)| \cdot |\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)| \\ &\leq (2K + K_\varepsilon) \cdot |\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)| \end{aligned}$$

with K from (A3) and K_ε from (A4). Let $\xi > 0$, then, using the Markov inequality for the joint probability measure $\mathbb{P} = \mathbb{P}_{\mathbf{X}, \boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$ and $\mathbf{X}$, we get

$$\begin{aligned} & \mathbb{P} \left(\sqrt{\gamma_n M} |\tilde{m}(\mathbf{X}_i)^2 - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)^2| > \xi \right) \\ &\leq \mathbb{P} \left((2K + K_\varepsilon) \cdot \sqrt{\gamma_n M} |\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)| > \xi \right) \\ &= \mathbb{P} \left(\sqrt{\gamma_n M} |\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)| > \frac{\xi}{2K + K_\varepsilon} \right) \\ &\leq (2K + K_\varepsilon)^2 \cdot \gamma_n M \cdot \frac{\mathbb{E}_{\mathbf{X}, \boldsymbol{\Theta}} [|\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)|^2]}{\xi^2}. \quad (24) \end{aligned}$$

To handle this, we are going to use Theorem 2.3 from [11], which, under (A3), (A4) and (A7), states that for $\mathbf{X}_i \in [0, 1]^p$, there is a $\delta \in (0, 1)$ such that with probability at least $1 - \delta$,

$$\mathbb{E}_{\mathbf{X}} \left[|\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)|^2 \right] \leq C_{\lambda, K, K_\varepsilon} \frac{\log(a_n)^2 \log(a_n p) + \log(a_n) \log(1/\delta)}{a_n^{\phi(\lambda)}} \quad (25)$$

holds for a_n large enough if the tree depth $d = \lceil \log_2(a_n)/(1 - \log_2(1 - \lambda)) \rceil$, where $\phi(\lambda) := \frac{-\log_2(1-\lambda)}{1-\log_2(1-\lambda)}$. $C_{\lambda, K, K_\varepsilon}$ is a constant that only depends on K, K_ε and λ from (A7). In this case 'a_n large enough' means that

$$\max \left\{ \frac{2 \exp(1)^2 d}{a_n}, \frac{\log(72 p^d (n + 1^{2d}/\delta))}{a_n} \right\} < \frac{3}{4}$$

holds true. Now, by choosing $\delta = \frac{1}{a_n^{\phi(\lambda)}}$, combining (24), (25) and the independence of $\boldsymbol{\Theta}_t$ and $\mathbf{X}_i$ delivers

$$\begin{aligned} & \gamma_n M \cdot \mathbb{E}_{\mathbf{X}, \boldsymbol{\Theta}} \left[|\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)|^2 \right] \\ &= \gamma_n M \int |\tilde{m}(\mathbf{x}_i) - m_{n,1}(\mathbf{x}_i, \theta, \mathcal{D}_n)|^2 d\mathbb{P}_{\mathbf{X}, \boldsymbol{\Theta}}(\mathbf{x}, \theta) \\ &= \gamma_n M \int \int |\tilde{m}(\mathbf{x}_i) - m_{n,1}(\mathbf{x}_i, \theta, \mathcal{D}_n)|^2 d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) d\mathbb{P}_{\boldsymbol{\Theta}}(\theta) \\ &\leq \gamma_n M C_{\lambda, K, K_\varepsilon} \frac{\log(a_n)^2 \log(a_n p) - \phi(\lambda) \log(a_n) \log(a_n)}{a_n^{\phi(\lambda)}} + \frac{\gamma_n M (2K + K_\varepsilon)}{a_n^{\phi(\lambda)}} \\ &\leq \gamma_n M C_{\lambda, K, K_\varepsilon} \frac{\log(a_n)^2 \log(a_n p)}{a_n^{\phi(\lambda)}} + \frac{\gamma_n M (2K + K_\varepsilon)}{a_n^{\phi(\lambda)}} \end{aligned} \quad (26)$$

for a_n large enough. The assumption $\frac{M \gamma_n \log(a_n)^2 \log(a_n p)}{a_n^{\phi(\lambda)}} \rightarrow 0$ implies that (26) becomes arbitrary small. Thus, with (26) we can see that $\gamma_n M (\tilde{m}(\mathbf{X}_i)^2 - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}, \mathcal{D}_n)^2) \xrightarrow{\mathbb{P}} 0$. The calculations in (24) can also be conducted for

$$\begin{aligned} & \mathbb{P}_{\pi, \mathbf{X}, \boldsymbol{\Theta}} \left(\sqrt{\gamma_n M} |\tilde{m}(\mathbf{X}_i^{\pi_{j,t}})^2 - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \boldsymbol{\Theta}_t, \mathcal{D}_n)^2| > \xi \right) \\ &\leq (2K + K_\varepsilon)^2 \cdot \gamma_n M \cdot \frac{\mathbb{E}_{\pi, \mathbf{X}, \boldsymbol{\Theta}} \left[|\tilde{m}(\mathbf{X}_i^{\pi_{j,t}}) - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \boldsymbol{\Theta}_t, \mathcal{D}_n)|^2 \right]}{\xi^2}. \end{aligned}$$

for the joint probability measure $\mathbb{P}_{\pi, \mathbf{X}, \boldsymbol{\Theta}}$ of $\pi, \mathbf{X}$ and $\boldsymbol{\Theta}$. Let $\mathfrak{D}_{\gamma_n}$ be the set of all derangements of $\{1, \dots, \gamma_n\}$ and for $(i_1, \dots, i_{\gamma_n}) \in \mathfrak{D}_{\gamma_n}$ denote $\mathbf{X}_i^{\pi_{j,t}}$ by $\mathbf{X}_i^{(i_1, \dots, i_{\gamma_n})}$ when $\pi_{j,t}$ takes the realization $(i_1, \dots, i_{\gamma_n})$. Then, the independence

of $\pi_{j,t}$, $\mathbf{X}_i$ and $\boldsymbol{\Theta}_t$ leads to

$$\begin{aligned}
& \gamma_n M \cdot \mathbb{E}_{\mathbf{X}, \boldsymbol{\Theta}} \left[\left| \tilde{m}(\mathbf{X}_i^{\pi_{j,t}}) - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \boldsymbol{\Theta}_t, \mathcal{D}_n) \right|^2 \right] \\
&= \gamma_n M \int \left| \tilde{m}(\mathbf{x}_i^{\pi_j}) - m_{n,1}(\mathbf{x}_i^{\pi_j}, \theta, \mathcal{D}_n) \right|^2 d\mathbb{P}_{\pi, \mathbf{X}, \boldsymbol{\Theta}}(\pi_j, \mathbf{x}_i, \theta) \\
&= \gamma_n M \int \int \int \left| \tilde{m}(\mathbf{x}_i^{\pi_j}) - m_{n,1}(\mathbf{x}_i^{\pi_j}, \theta, \mathcal{D}_n) \right|^2 d\mathbb{P}_{\pi}(\pi_j) d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) d\mathbb{P}_{\boldsymbol{\Theta}}(\theta) \\
&= \gamma_n M \int \int \sum_{(i_1, \dots, i_{\gamma_n}) \in \mathfrak{D}_{\gamma_n}} \frac{1}{\gamma_n!} \\
&\quad \cdot \left| \tilde{m}(\mathbf{x}_i^{(i_1, \dots, i_{\gamma_n})}) - m_{n,1}(\mathbf{x}_i^{(i_1, \dots, i_{\gamma_n})}, \theta, \mathcal{D}_n) \right|^2 d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) d\mathbb{P}_{\boldsymbol{\Theta}}(\theta) \\
&= \frac{\gamma_n M}{\gamma_n!} \sum_{(i_1, \dots, i_{\gamma_n}) \in \mathfrak{D}_{\gamma_n}} \int \int \left| \tilde{m}(\mathbf{x}_i^{(i_1, \dots, i_{\gamma_n})}) - m_{n,1}(\mathbf{x}_i^{(i_1, \dots, i_{\gamma_n})}, \theta, \mathcal{D}_n) \right|^2 d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) d\mathbb{P}_{\boldsymbol{\Theta}}(\theta) \\
&= \gamma_n M \mathbb{E}_{\mathbf{X}, \boldsymbol{\Theta}} \left[\left| \tilde{m}(\mathbf{X}_{j,i}) - m_{n,1}(\mathbf{X}_{j,i}, \boldsymbol{\Theta}, \mathcal{D}_n) \right|^2 \right].
\end{aligned}$$

Since $\mathbf{X}_{j,i} \in [0, 1]^p$, the inequality (25) can also be applied to

$$\mathbb{E}_{\mathbf{X}, \boldsymbol{\Theta}} \left[\left| \tilde{m}(\mathbf{X}_{j,i}) - m_{n,1}(\mathbf{X}_{j,i}, \boldsymbol{\Theta}, \mathcal{D}_n) \right|^2 \right].$$

Hence (26) holds for $\mathbb{E}_{\mathbf{X}, \boldsymbol{\Theta}} \left[\left| \tilde{m}(\mathbf{X}_{j,i}) - m_{n,1}(\mathbf{X}_{j,i}, \boldsymbol{\Theta}, \mathcal{D}_n) \right|^2 \right]$ as well and thus

$$\gamma_n M (\tilde{m}(\mathbf{X}_i^{\pi_{j,t}})^2 - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \boldsymbol{\Theta}, \mathcal{D}_n)^2) \xrightarrow{\mathbb{P}} 0.$$

It remains to show that

$$\sqrt{\gamma_n M} 2Y_i(\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}, \mathcal{D}_n))$$

and

$$\sqrt{\gamma_n M} 2Y_i(\tilde{m}(\mathbf{X}_i^{\pi_{j,t}}) - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \boldsymbol{\Theta}, \mathcal{D}_n))$$

are also vanishing for growing n . We use (A3) and (A4) to get

$$|2Y_i(\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}, \mathcal{D}_n))| \leq 2(K + K_\varepsilon) \cdot |\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}, \mathcal{D}_n)|.$$

As before, Markov inequality can be applied, which leads to

$$\begin{aligned}
& \mathbb{P} \left(\sqrt{\gamma_n M} \cdot 2|Y_i| \cdot |\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)| > \xi \right) \\
& \leq \mathbb{P} \left(2(K + K_\varepsilon) \cdot \sqrt{\gamma_n M} |\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)| > \xi \right) \\
& = \mathbb{P} \left(\sqrt{\gamma_n M} |\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)| > \frac{\xi}{2(K + K_\varepsilon)} \right) \\
& \leq 4(K + K_\varepsilon)^2 \gamma_n M \frac{\mathbb{E}_{\mathbf{X}, \boldsymbol{\Theta}} \left[\left| \tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n) \right|^2 \right]}{\xi^2}. \tag{27}
\end{aligned}$$

Since (24) tends to 0 as n grows large, (27) also converges to 0. Therefore,

$$2Y_i(\tilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t, \mathcal{D}_n)) \xrightarrow{\mathbb{P}} 0$$

holds true. Note that (27) also holds for $\mathbf{X}_i^{\pi_{j,t}}$ instead of $\mathbf{X}_i$ with the exact same arguments, which implies $2Y_i(\tilde{m}(\mathbf{X}_i^{\pi_{j,t}}) - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}, \boldsymbol{\Theta}_t, \mathcal{D}_n)) \xrightarrow{\mathbb{P}} 0$. Hence,

$$\frac{I_{n,M}^{OOB}(j) - \tilde{I}_{n,M}^{OOB}(j)}{\sqrt{\gamma_n^2 \zeta_{1,\gamma_n}/n + \zeta_{\gamma_n}/M}} \xrightarrow{\mathbb{P}} 0$$

holds and completes the proof. $\square$

Funding

The first author was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 314838170, GRK 2297 MathCoRe.

References

- [1] ATHEY, S. and WAGER, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113** 1228-1242.
- [2] BREIMAN, L. (2001). Random forests. *Machine Learning* **45** 5-32.
- [3] BIAU, G., DEVROYE, L. and LUGOSI, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research* **9** 9-32.
- [4] BIAU, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research* **13** 1063-1095. JMLR.org.
- [5] RAMOSAJ, B. and PAULY, M. (2023). Consistent and unbiased variable selection under independent features using Random Forest permutation importance. *Bernoulli* **29** 2101-2118.
- [6] RAMOSAJ, B. and PAULY, M. (2023). Supplementary Material to: Consistent and unbiased variable selection under independent features using Random Forest permutation importance. *Bernoulli*.
- [7] GREGORUTTI, B., MICHEL, B. and SAINT-PIERRE, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing* **27** 659-678.
- [8] ISHWARAN, H. (2007). Variable importance in binary regression trees and forests. *Machine Learning*. 519-537.
- [9] ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. 841-860. *Annals of Applied Statistics*.
- [10] ISHWARAN, H. and LU, M. (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine* **38** 558-582.

- [11] MAZUMDER, R. and WANG, H. (2024). On the convergence of CART under sufficient impurity decrease condition *Advances in Neural Information Processing Systems* **36**.
- [12] MENTCH, L. and HOOKER, G. (2014). Ensemble trees and CLTs: Statistical inference for supervised learning. *Stat* **1050** 25.
- [13] MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research* **17** 1-41.
- [14] STEKHOVEN, DANIEL J and BÜHLMANN, PETER (2016). MissForest: non-parametric missing value imputation for mixed-type data. *Bioinformatics* **17** 112-118.
- [15] NOVOMESTKY, F. and NADARAJAH, S. (2016). *truncdist*: Truncated Random Variables. R package version 1.0-2. URL: <https://CRAN.R-project.org/package=truncdist>.
- [16] PENG, W., COLEMAN, T. and MENTCH, L. (2022). Rates of convergence for random forests via generalized U-statistics. *Electronic Journal of Statistics* **16** 232–292.
- [17] R CORE TEAM (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- [18] SCORNET, E., BIAU, G. and VERT, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*. 1716-1741
- [19] SHAO, J. (1991). Consistency of jackknife variance estimators. *Statistics: A Journal of Theoretical and Applied Statistics* **22** 49-57.
- [20] STROBL, C., BOULESTEIX, A.-L., KNEIB, T., AUGUSTIN, T. and ZEILEIS, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics* **9** 1-11.
- [21] TANG, F. and ISHWARAN, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **10** 363-377.
- [22] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113** 1228-1242.
- [23] WRIGHT, M. N. and ZIEGLER, A. (2017). *ranger*: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77** 1–17. DOI: 10.18637/jss.v077.i01.

Article 2

Föge, N., & Pauly, M. (2026). Confidence intervals for Random Forest permutation importance with missing data. *Computational Statistics*, 41(3), 45.
<https://doi.org/10.1007/s00180-026-01722-w>



Confidence intervals for Random Forest permutation importance with missing data

Nico Föge^{1,2} · Markus Pauly^{2,3}

Received: 31 July 2025 / Accepted: 23 January 2026
© The Author(s) 2026

Abstract

Random Forests are renowned for their predictive accuracy, but valid inference – particularly about permutation-based feature importances – remains challenging. Existing methods, such as Ishwaran et al.’s (2019) confidence intervals (CIs), are promising but assume complete feature observation. However, real-world data often contains missing values. In this paper, we investigate how common imputation techniques affect the validity of Random Forest permutation-importance CIs when data are incomplete. Through an extensive simulation and real-world benchmark study, we compare state-of-the-art imputation methods across various missing-data mechanisms and missing rates. Our results show that single-imputation strategies, when paired with naive variance estimators, lead to low CI coverage due to underestimation of imputation uncertainty. As a remedy, we adapt Rubin’s rule to aggregate feature-importance estimates and their variances over several imputed datasets and account for imputation uncertainty. Our numerical results indicate that the adjusted CIs achieve better nominal coverage for moderate sample sizes ($n \geq 250$) and missingness up to 30%, whereas they are statistically unstable for small samples ($n = 100$) or when missingness exceeds 50%.

Keywords Permutation importance · Missing values · Random Forest · Mixgb · Mice · Confidence intervals

✉ Nico Föge
nico.foege@ovgu.de

Markus Pauly
markus.pauly@tu-dortmund.de

¹ Institute for Mathematical Stochastics, Otto-von-Guericke University Magdeburg, Magdeburg 39106, Germany

² Research Center Trustworthy Data Science and Security, University Alliance Ruhr, Dortmund 44227, Germany

³ Department of Statistics, TU Dortmund University, Dortmund 44227, Germany

1 Introduction

Random Forests, introduced by Breiman (2001), are powerful tree-based ensemble methods, widely known for their robustness and predictive accuracy across diverse applications. In contrast to (generalized) linear regression models – which offer interpretable parameter and variance estimates that support formal statistical inference – Random Forests rely on feature-importance measures that lack well-developed inferential statistics, such as CIs and hypothesis tests. The absence of analytical solutions for variance estimation and the intricate dependencies within tree-based predictions make statistical inference far less straightforward.

Recent advancements have sought to address these challenges by developing rigorous frameworks for inference in Random Forests. Scornet et al. (2015) and Blum et al. (2024) proved consistency for Random Forest predictions. Moreover, Wager et al. (2014) and Mentch and Hooker (2016) established asymptotic normality and variance estimators for Random Forest predictions, which can be used to construct prediction intervals. However, these advances address the prediction level rather than inference on the feature level.

But feature importances are crucial for practitioners who want to understand which covariates drive predictions. Ramosaj and Pauly (2023) demonstrated the consistency of Random Forest Permutation Importance (RFPI), affirming its reliability as a measure of feature significance - though under rather strong assumptions. Based on RFPIs, Ishwaran and Lu (2019) proposed confidence intervals (CIs) for interpretable population feature importance scores. The approach is implemented in the R package `randomForestSRC`, that has been downloaded more than 900.000 times (18th July 2025). Although these numbers and the simulation results on the validity of these RFPI-based CIs presented in Ishwaran and Lu (2019) are promising, the approach lacks a formal proof of the implied asymptotic normality of RFPIs. The recent preprint (Föge et al. 2024) made a first theoretical contribution to close this gap by proving a central limit theorem for RFPIs. However, the results rely on strong conditions, which may limit their applicability in practice. Together, these developments highlight a growing body of research aimed at bridging the gap between the predictive power of Random Forests and the need for interpretable statistical inference.

Despite these promising developments, existing inference procedures assume that the features are fully observed. However, real-world datasets are rarely complete, and missing data pose a challenge to the applicability of these inferential methods. The performance of imputation methods in the context of Random Forest-based inference, particularly for Ishwaran's CIs for population feature importance scores, remains underexplored. This gap raises two research questions: (i) How do existing imputation methods affect the coverage and validity of Ishwaran's feature-importance CIs when data are missing? (ii) Can one devise an imputation-inference strategy that preserves nominal coverage?

To address these questions, we conduct an extensive simulation study to compare different state-of-the-art imputation techniques – such as `mice` (Van Buuren and Groothuis-Oudshoorn 2011), `missforest` (Stekhoven 2012), and `mixgb`, an XGBoost-based imputation (Deng and Lumley 2024) – and evaluate their impact and

the robustness of Ishwaran's CIs under various missing data scenarios. Our results show that single imputation, when combined with naive variance estimators that treat imputed values as observed, leads to substantial undercoverage of CIs due to unaccounted imputation uncertainty. To address this, we propose an inference approach based on Rubin's rule (1987a, 2018), which combines estimates and variances across multiple imputed datasets to explicitly account for imputation uncertainty. Our simulations demonstrate that this approach achieves better nominal coverage, suggesting that multiple imputation is practically more convenient than single imputation. We also note that an adjusted variance estimator for single imputation could, in principle, mitigate undercoverage, though developing such an estimator for RFPIM is challenging due to the complex dependency structure induced by Random Forests.

By treating each imputed dataset as a separate realization and aggregating the resulting importance estimates and variances, we derive adjusted CIs that better account for imputation uncertainty. Through extensive simulation, we show that this approach leads to improved coverage rates, suggesting that combining multiple imputation with Rubin's rule is essential for valid RFPIM-based CIs in the presence of missing data.

The proposed methodology for constructing confidence intervals for Random Forest Permutation Importance with missing data is particularly relevant for research fields that analyze complex, high-dimensional datasets affected by incomplete observations. Examples include medical research (e.g., predicting disease outcomes from electronic health records), environmental science (e.g., modeling climate impacts with sensor data), and the social sciences (e.g., analyzing survey data with non-response), where missing data are common and interpretability and reliable inference is essential.

The rest of the paper is structured as follows. In section 2 we formally introduce the RFPIM and the CIs. section 3 presents the investigated imputation methods, including their adaptation with Rubin's rule. section 4 describes the settings of our simulation study and its results are presented in section 5. To also investigate the performance on real-world data we conduct a benchmark study in section 6. The paper closes with a discussion (section 7).

2 Permutation importance and ishwarans confidence intervals

We consider a regression setting with a response variable $Y \in \mathbb{R}$ and a feature vector $X = (X_1, \dots, X_p) \in \mathbb{R}^p$. The underlying regression function is denoted by $f(x) = \mathbb{E}[Y|X = x]$. A Random Forest consists of M tree predictors, each generated on a randomly drawn bootstrap sample. Additionally, each tree uses feature subsampling, whereby a random subset of features is selected at each split. We denote this algorithmic randomness – arising from bootstrapping and feature subsampling – by the random vectors $\Theta_1, \dots, \Theta_M$. We write $f_{n,1}(x, \Theta_t, \mathcal{D}_n)$ for the prediction of tree t on an input covariate vector $x \in \mathbb{R}^p$. Here $\mathcal{D}_n$ is the training dataset consisting of n i.i.d. copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of (X, Y) . The *Random Forest prediction* on x is then defined as the ensemble prediction

$$f_{n,M}(\mathbf{x}, \Theta_1, \dots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{t=1}^M f_{n,1}(\mathbf{x}, \Theta_t, \mathcal{D}_n).$$

Each tree is trained on a bootstrap subsample drawn from $\mathcal{D}_n$. The excluded points form the Out-of-Bag (OOB) set $\mathcal{D}_n^{-(t)}(\Theta_t)$ of tree t .

We now begin by introducing the Tree Permutation Importance, and based on this definition we then introduce the Random Forest Permutation Importance Measure. For the j -th feature, $j \in \{1, \dots, p\}$, we want to quantify the contribution of $X^{(j)}$ to the predictive accuracy of a single tree and, subsequently, of the full Random Forest. To compute it, the values of the j -th feature in the OOB sample $\mathcal{D}_n^{-(t)}(\Theta_t)$ are randomly permuted. More formally let, $\tilde{X}_i^{(j)}$ be a randomly selected value from the set of j -th feature values in the OOB sample. By replacing the j -th component of X_i with $\tilde{X}_i^{(j)}$ we obtain the permuted feature vector

$$\tilde{X}_i^j = (X_i^{(1)}, \dots, X_i^{(j-1)}, \tilde{X}_i^{(j)}, X_i^{(j+1)}, \dots, X_i^{(p)})$$

for each observation (X_i, Y_i) with $i \in \mathcal{D}_n^{-(t)}(\Theta_t)$. Ishwaran and Lu (2019) define the *Tree Permutation Importance* for the t -th tree and a loss-function ℓ as

$$I(j, \Theta_t, \mathcal{D}_n) = \frac{1}{|\mathcal{D}_n^{-(t)}(\Theta_t)|} \sum_{i \in \mathcal{D}_n^{-(t)}(\Theta_t)} \left\{ \ell(Y_i, f_{n,1}(\tilde{X}_i^j, \Theta_t, \mathcal{D}_n)) - \ell(Y_i, f_{n,1}(X_i, \Theta_t, \mathcal{D}_n)) \right\}, \quad (1)$$

where $|\mathcal{D}_n^{-(t)}(\Theta_t)|$ is the size of the OOB set for tree t . Since we focus on a regression setting, ℓ is chosen as the squared loss $\ell(x, y) = (x - y)^2$. Averaging (1) across all trees $t = 1, \dots, M$ yields the *Random Forest Permutation Importance Measure (RFPIM)*

$$I_{n,M}^{OOB}(j, \mathcal{D}_n) = I_{n,M}^{OOB}(j, \Theta_1, \dots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{t=1}^M I(j, \Theta_t, \mathcal{D}_n). \quad (2)$$

Due to the dependence structure of the OOB-samples, its rather complex to estimate the variance of (2). This makes it hard to construct CIs for the true population *Feature Importance (Scores)*

$$I(j) := \mathbb{E} \left[\left(Y_1 - f(X_{j,1}) \right)^2 \right] - \mathbb{E} \left[\left(Y_1 - f(X_1) \right)^2 \right]. \quad (3)$$

Here, $X_{j,1} = [X_1^{(1)}, \dots, X_1^{(j-1)}, X^{(j)}, X_1^{(j+1)}, \dots, X_1^{(p)}]$ and $X^{(j)}$ is an independent copy of $X_1^{(j)}$ (Gregorutti et al. 2017). Ishwaran and Lu (2019) compared several variance estimators and recommend the delete- d Jackknife estimator, which has been implemented in the RandomForestSRC R package (Ishwaran and Kogalur 2007;

Ishwaran et al. 2008; Ishwaran and Kogalur 2025; R Core Team 2023). The delete- d Jackknife is computed as in Ishwaran and Lu (2019) and the corresponding Algorithm 1 is displayed below.

Input: data set $\mathcal{D}_n$, number of repetitions K , subsample size b (so delete $d = n - b$), trained Random Forest generated by $\Theta_1, \dots, \Theta_M$,
Output: The delete- d variance estimator $v_{d,K}^2$
 Use the entire dataset $\mathcal{D}_n$ to compute $I_{n,M}^{OOB}(j, \mathcal{D}_n)$
for $k \leftarrow 1$ **to** K **do**
 Draw a subsample s_k of size b from $\{1, \dots, n\}$;
 Denote the subsample of $\mathcal{D}_n$ containing only the indices of s_k by
 $\mathcal{D}_{b,s_k} = \{(\mathbf{X}_i, Y_i) : i \in s_k\}$;
 Calculate and save RFPIM $I_{n,M}^{OOB}(j, \mathcal{D}_{b,s_k})$ using only $\mathcal{D}_{b,s_k}$;
end
return

$$v_{d,K}^2 = \frac{b}{(n-b)K} \sum_{k=1}^K (I_{n,M}^{OOB}(j, \mathcal{D}_{b,s_k}) - I_{n,M}^{OOB}(j, \mathcal{D}_n))^2 \quad (4)$$

Algorithm 1 Delete- d Jackknife estimator ($d = n - b$) for $\mathbb{V}\text{ar}(I_{n,M}^{OOB}(j))$

Using the variance estimator $v_{d,K}^2$, Ishwaran and Lu (2019) propose the $(1 - \alpha)$ CI

$$\left[I_{n,M}^{OOB}(j, \mathcal{D}_n) - z_{1-\alpha/2} \sqrt{v_{d,K}^2}, I_{n,M}^{OOB}(j, \mathcal{D}_n) + z_{1-\alpha/2} \sqrt{v_{d,K}^2} \right]. \quad (5)$$

Here, $z_{1-\alpha/2}$ is the $(1 - \frac{\alpha}{2})$ -quantile of the standard normal distribution. This normal approximation was justified by promising simulations (Ishwaran and Lu 2019). In Föge et al. (2024) a theoretical justification - though under restrictive assumptions on the hyperparameters - is also given. In summary, this approach constructs CIs via RFPIM and the delete- d Jackknife variance estimator, under the assumption that all features are fully observed. We next investigate the challenge of missing data, and how imputation before training may impact the validity of these CIs.

3 Missing data and imputation approaches

Missing data are ubiquitous in empirical research across disciplines. Discarding incomplete observations in the analysis (complete-case or listwise deletion) can introduce bias estimates/predictions or lead to a substantial loss of information (Schafer 1997; Little and Rubin 2019). As a remedy, (multiple) imputation is recommended, particularly when data is missing at random (MAR) or missing completely at random (MCAR) (Pepinsky 2018). In multiple imputation, missing values are replaced by draws from the predictive (imputation) distribution multiple times, generating R complete datasets. Final estimates and standard errors are then combined across these

datasets using Rubin's rule to reflect uncertainty from both sampling and imputation (Rubin 1987a, 2018).

In general, a variety of imputation techniques exist, ranging from simple single-imputation approaches (e.g., mean imputation or regression imputation) to more advanced techniques such as `mice` (Van Buuren and Groothuis-Oudshoorn 2011; White et al. 2011). The choice of method depends on the study design, the assumed missingness mechanism, the scale of the features, and the primary inferential goals. For comprehensive overviews and statistical foundations, we refer to Rubin (1987a); Schafer (1997); Van Buuren and Groothuis-Oudshoorn (2011); White et al. (2011); Rubin (2018); Little and Rubin (2019); Thurow et al. (2021); Buczak et al. (2023).

For our selection of imputation methods, we deliberately focus on modern, state-of-the-art approaches that have demonstrated favorable properties in a wide range of scenarios and are available in robust and well-maintained **R** packages. Moreover, as we particularly investigate the effect of both single and multiple imputation, we wanted complementary imputation techniques offering both imputation variants, or at least allow for predictive-mean-matching (PMM). PMM ensures that only observed values are used for imputation, thus preserving the original data distribution and avoiding implausible values (Rubin 1987b; Little 1988; White et al. 2011). Moreover, PMM is less sensitive to model misspecification compared to purely parametric approaches and is well-suited for non-normally distributed variables (Morris et al. 2014). We thus selected three imputation that fulfill these requirements:

I. Multiple imputation by chained equations. `mice` iteratively models each feature with missing values using regression based on other features to produce imputations (Van Buuren and Groothuis-Oudshoorn 2011; White et al. 2011). Imputation can be performed using single imputation with PMM or multiple imputation, where multiple plausible values are generated to account for uncertainty. In fact, `mice` with PMM exhibits favorable distributional imputation accuracy (Thurow et al. 2021, 2024) and is still the gold standard imputation method in many quantitative sciences (Schwerter et al. 2024; Romero et al. 2025).

The implementation of `mice` is done with the **R** package from Van Buuren and Groothuis-Oudshoorn (2011).

II. Chained Random Forests (missRanger). This approach leverages the Random Forest algorithm to sequentially predict each feature with missing values using out-of-bag estimates from Random Forests trained on the other features (Stekhoven 2012). Though it doesn't support multiple imputation, it can deal with mixed-type data and its extension by Mayer (2024) also supports PMM: after obtaining a predicted value for a missing observation, `missRanger` samples from a small set of observed values whose out-of-bag predictions are closest. Empirical studies (Stekhoven and Bühlmann 2011; Waljee et al. 2013; Ramosaj and Pauly 2019; Buczak et al. 2023) have shown that `missRanger` often outperforms parametric-imputation approaches with respect to imputation accuracy and other predictive tasks. However, when applied without PMM, it was also seen to inflate type-I-errors of statistical tests in linear models (Ramosaj et al. 2020; Schwerter et al. 2024). For the chained Random Forest Imputation we use the `missRanger` package which is based upon the fast `ranger` implementation of Random Forests. We chose the options `num.trees=200`, `pmm.k=5` and the default values for the rest of the hyperparameters.

III. XGBoost-based Imputation (mixgb). This algorithm uses the XGBoost algorithm, a gradient boosting method, to estimate missing values within a multiple imputation framework (Deng and Lumley 2024). Implemented in the `mixgb` package, `mixgb` captures complex relationships efficiently, using PMM and subsampling to enhance imputation accuracy and variability (Deng 2023). Multiple imputation generates several plausible datasets to account for uncertainty (Deng and Lumley 2024).

These are also the same imputation algorithms as considered by Schwerter et al. (2024) who also evaluate inferential properties (estimation bias, type-I-error and power of tests), but in traditional linear-models. Our goal is to assess how these methods perform when paired with nonparametric Random Forest-based CIs that should better deal with nonlinear data structures. By including both single- and multiple-imputation variants of `mice` and `mixgb`, we can directly compare:

(A) Single-Imputations as Quick Fixes that fill in missing values once (with PMM) and proceed with the construction of (5) as if the data were fully observed. In parametric linear models, single imputation often underestimates uncertainty, which can lead to excessively narrow CIs for parameters (Van Buuren and Groothuis-Oudshoorn 2011; White et al. 2011). As we will see later, this approach also causes serious problems for Random Forest-based CIs.

(B) More complex Multiple-Imputations that generate R complete datasets via repeated draws from each imputation model (applying PMM on each draw) and then pool estimates via Rubin's rule. In parametric linear models, this approach leads to valid inference procedures when data is MAR. However, there are neither theoretical reasons nor numerical studies that support this for Random Forest-based CIs. We will close this gap in the present paper.

To explain the heuristic multiple imputation approach based on Rubin's rule for combining estimators (Rubin 2018) recall that `mice` and `mixgb` are multiple Imputation methods. Thus, instead of a single complete dataset, they can generate $R \in \mathbb{N}$ complete datasets. On each of those we run the Random Forest and compute the delete- d Jackknife estimator, denoted by $v_{d,K,r}^2$ ($r = 1, \dots, R$), and the Permutation Importance, denoted by $I_{n,M}^{OOB}(j, \mathcal{D}_n^r)$, where $\mathcal{D}_n^r$ is the r -th complete dataset. Following Rubin's rule the estimations can be combined by averaging

$$I_{n,M}^{\bar{O}OB}(j) = \frac{1}{R} \sum_{r=1}^R I_{n,M}^{OOB}(j, \mathcal{D}_n^r).$$

For the combined variance estimation, we have to consider the within-imputation variance and the between-imputation variance. The within-imputation variance can also be estimated by averaging

$$\bar{v}_{d,K}^2 = \frac{1}{R} \sum_{r=1}^R v_{d,K,r}^2$$

while an estimator for the between-imputation variance is given by

$$B = \frac{1}{R-1} \sum_{r=1}^R \left(I_{n,M}^{OOB}(j, \mathcal{D}_n^r) - I_{n,M}^{\bar{O}OB}(j) \right)^2.$$

This leads to an estimator for the total variance given by

$$v_{d,K,\text{Rubin}}^2 = v_{d,K}^2 + \left(1 + \frac{1}{m} \right) B.$$

Following Rubin's rule, the resulting CIs for (3) scores are given by

$$\left[I_{n,M}^{OOB}(j, \mathcal{D}_n) - z_{1-\alpha/2} \sqrt{v_{d,K,\text{Rubin}}^2}, I_{n,M}^{OOB}(j, \mathcal{D}_n) + z_{1-\alpha/2} \sqrt{v_{d,K,\text{Rubin}}^2} \right]. \quad (6)$$

We will refer to them as 'Rubin intervals'. As stated above, this approach is heuristic as it is not backed by theoretical results.

The application of Rubin's Rule to combine multiple imputations for RFPIM inference, as described above, relies on several theoretical assumptions that warrant careful scrutiny. First, Rubin's Rule requires that the imputation model (e.g., `mice` or `mixgb` with PMM) is congenial with the analysis model (Meng 1994). In our setting, this means that the imputation method must adequately capture the relationships and patterns in the data, such as non-linear interactions, that the Random Forest uses to compute $I_{n,M}^{OOB}(j, \mathcal{D}_n^r)$. Hence, in our situation, `mice` and `mixgb` probably cause uncongeniality, which can lead to biased estimates of $I_{n,M}^{\bar{O}OB}(j)$ or incorrect variance estimates $v_{d,K,\text{Rubin}}^2$ (Reiter 2007). Second, Rubin's Rule assumes that the pooled estimator $I_{n,M}^{\bar{O}OB}(j)$ is approximately normally distributed, as this is required for the resulting Rubin intervals to be valid. So far, asymptotic normality of RFPIM has only been established under restrictive conditions: additive regression function f , true Feature Importance $I(j) > 0$, randomly chosen trees, and a very slowly growing number of trees M (Föge et al. 2024). Third, consistency of the delete- d Jackknife variance estimator in (4) would be needed to prove the validity of the Rubin intervals. While delete- d Jackknife estimators are consistent under the assumption of Fréchet differentiability for simpler point estimators (Shao 1991), it remains unknown whether this also holds for RFPIM., especially as existing consistency results for $I_{n,M}^{OOB}(j)$ require $M \rightarrow \infty$ and $n \rightarrow \infty$ (Ramosaj and Pauly 2023). Establishing consistency for the variance estimator would likely require a coupling of n , M and K , which may interact unfavorably with conditions needed for asymptotic normality (Föge et al. 2024).

Nonetheless, we will later see in our numeric experiments that it leads to significant improvements compared to the quick fixes by single imputations.

4 Simulation setting

To evaluate the performance of our six imputation strategies in conjunction with the Random Forest-based CIs for the population feature importance scores by Ishwaran and Lu (2019), we adopt their data-generating framework. This choice serves two purposes: first, it ensures comparability with prior work on tree-based inference; second, it provides a diverse collection of nonlinear regression models to test the six imputation procedures. Below, we first describe the twelve regression models, including the form of regression function, feature space, error distributions, and sample sizes before we explain the mechanisms for inducing missingness and the operational details of our study.

Regression models. We simulate $\mathbf{x} = (x_1, \dots, x_p)$ and responses y according to twelve distinct regression functions. Following Ishwaran and Lu (2019), our simulations assume independent features to isolate imputation effects and ensure comparability with prior work. Moreover, independent features are one of the structural conditions under which consistency of RFPIM can currently be guaranteed (Ramosaj and Pauly 2023). The settings span smooth non-linear terms (e.g., trigonometric or exponentials), higher-order interactions, indicator thresholds, and pure noise models, representing a broad spectrum of realistic data-generating processes (see Table 1 in Appendix A for detailed functional forms). To mimic realistic scenarios, the number of features is set to $p = 20$ for each model, by simply adding noise features, independent of y . That is, if $y = f(x_1, \dots, x_q) + \epsilon$ with random error ϵ holds with $q < p$, we simply add iid features $x_{q+1}, \dots, x_p$ unrelated to the response y . We vary the sample size as follows $n \in \{100, 250, 500\}$. The choice $n = 250$ matches the original setting from Ishwaran and Lu (2019), whereas $n = 100$ and $n = 500$ allow us to assess the impact of smaller resp. larger sample sizes on CI coverage.

Missing mechanisms. For the introduction of missings, we apply two different missing mechanisms. MCAR randomly inserts missing values across all features in the dataset, with no connection to the data itself, ensuring that the missingness is entirely by chance. For the MAR algorithm, the 50th percentile (median) of $x^{(j)}$ is computed, and the subset of rows with values below this threshold is identified

$$\mathcal{I}_{\text{MAR}} = \{i \in \{1, \dots, n\} \mid x_i^{(j)} < \text{median}(x^{(j)})\}.$$

From this subset, a proportion determined by the desired missing rate is sampled uniformly at random. In the selected rows, missing values are introduced independently in all features $x^{(k)}$ where $k \neq j$. This aims to mimic a pattern often seen in real-world datasets (Enders 2010).

Within each MCAR or MAR scenario, we vary the missing rates (10%, 30%, and 50%). MCAR implies that missingness is independent of observed or unobserved data, with a fixed number of rows per column randomly set to NA, proportional to the missing rate. MAR means missingness depends on observed data (e.g., values in the second column below its median determine missingness in other columns), with a fixed number of NAs per affected column, proportional to the missing rate and the number of selected rows, or a fixed number via MCAR fallback if no rows are selected.

We focus on MCAR and MAR missing mechanisms, which are the standard assumptions in most imputation literature and primary settings emphasized in Schafer (1997) for robust analysis. We do not consider missing not at random (MNAR), as modeling MNAR requires specifying and validating explicit missingness mechanisms, meriting a distinct research effort beyond the scope of this study.

Conducting the experiments. For each of the $12 \cdot 3 \cdot 2 \cdot 3 = 216$ combinations of Model (Functions 1–12), sample size ($\in \{100, 250, 500\}$), missing mechanisms (MAR or MCAR) and missing rates ($\in \{10\%, 30\%, 50\%\}$), we conduct $n_{sim} = 1,000$ independent simulation runs. Within each run the following steps are carried out:

1. **Data generation:** Generate a dataset with the specified sample size of $n = 250$ and $p = 20$ features according to the chosen regression models, including noise features independent of the response, as described in section 4.
2. **Introducing missing values:** Introduce missing values under the MCAR or MAR mechanism with the specified missing rates.
3. **Imputation:** Apply each of the three imputation methods:
 - Using the `mice` package with predictive mean matching, `m=5`, `maxit = 5`,
 - Chained Random Forests (`missRanger` package, with 250 trees, a node size of 5 and `pmm.k=5`),
 - XGBoost-based imputation (`mixgb` package, supporting multiple imputation). `mice` and `mixgb` yield $m = 5$ completed data sets; `missRanger` returns a single completed data set.
4. **Compute RFPIM:** Fit a Random Forest via `rfsrc` (Ishwaran and Kogalur 2025) with `ntree = 250`, `nodesize = 5`, `samptype = "swr"`, and `block.size = 1`. Compute the RFPIM $I_{n,M}(j)$, for each feature j in each complete dataset, as defined in section 2.
5. **CI construction:** Construct 95% CIs ($\alpha = 0.05$) for the true feature importances using the two approaches: Imputing the data by (multiple) imputation and getting a single complete dataset on which the delete- d -Jackknife (4) is determined. And as an alternative for the multiple Imputation methods, the delete- d Jackknife is determined on the $m=5$ complete datasets and then combined by Rubin's rules. The computation of the delete- d Jackknife variance estimator is done with `subsample.rfsrc` and `extract.subsample` from Ishwaran and Kogalur (2025). We choose $B=100$ and `subratio=sqrt(n)/n` as Ishwaran and Lu (2019).
6. **Ground truth:** We estimate the population feature importance scores by generating 1,000 complete data sets of size $n = 250$ (no missingness) and averaging their OOB importances $I_{n,M}^{OOB}(j)$. For the noise features the true feature importance scores (3) are set to 0.
7. **Evaluation:** For each replicate and each imputation approach, record whether the corresponding CI covers the ground truth feature importance. Coverage rates are summarised over the 1,000 replicates per experimental condition.

All simulations and computations are conducted with R Version 4.4.0 (R Core Team 2023). To assess runtime and resource requirements, we conducted an empirical benchmark on a Windows 10 workstation (x86_64, 12 logical cores, 16 GB RAM). For each combination of sample size $n \in \{100, 250, 500\}$, missing-data mechanism, and imputation method, the full simulation pipeline – including missingness generation, single or multiple imputation, Random Forest fitting (200 trees, permutation importance), and delete- d jackknife subsampling – was executed once and timed end-to-end. Median wall-clock times per setting ranged from 0.07–0.5 s for the non-imputed baseline and 0.2–0.7 s for `mice`, to 0.5–1.2 s for `missRanger`, and approximately 3 s for `mixgb` at $n = 500$. For fully grown trees, the computational cost for the training of the Random Forest with `ranger` is $\mathcal{O}(m_{\text{try}} \cdot M \cdot n \cdot \log(n))$ (Breiman et al. 1984; Breiman 2001; Wright and Ziegler 2015). The computation of RFPIM for all variables requires $\mathcal{O}(p \cdot M \cdot n \cdot \log(n))$ operations when the size of the OOB-sample is linear in n . The cost of the delete- d Jackknife from Algorithm 1 is $\mathcal{O}(K \cdot p \cdot M \cdot (n - d) \cdot \log(n))$. Additionally, $R \cdot p$ times the computation complexity of the imputation method is needed (if there are missing values in all variables).

5 Simulation results

We evaluate the CIs on two criteria: coverage and interval length. Following Ishwaran and Lu (2019), we stratify the features by their 'true' feature importance score to account for signal strength. In our experiment, there are 240 features (12 functions with $p = 20$ features). Their ground truth feature importance scores induce three groups:

- **Noise (zero importance)-features:** 185 features with no effect on y and true importance 0.
- **High-importance-features:** the top 10% of all features (24 features) by true importance.
- **Moderate-importance-features:** the remaining 31 non-noise features. This group corresponds to the 75th-90th percentile group in Ishwaran and Lu (2019).

Throughout the paper we display results for MAR and discuss them in detail. Analogous plots for MCAR are given in the supplement and their results are only summarized here.

5.1 CI coverage under MAR

Figure 1 shows boxplots of empirical coverage for each imputation method, stratified by feature group (noise, moderate, high), sample size ($n = 250$) and missing rate (10%, 30%, 50% from top to bottom). The analogous plots for $n = 100, 500$ are shown in the Appendix in Figs. 6 and 7. For detailed results see Table 2. Within each subplot the methods are displayed from left to right in the following order: 1. single-imputation variants (`mice`, `missRanger`, `mixgb`; all with $\text{PMM}=5$), 2. multiple imputation variants with Rubin's rule (`mice` and `mixgb`, each with $R = 5$ imputations),

and 3. the results without missings ('none'; for $n = 250$), that serve as benchmark and are equal for all subplots.

Zero importance. CIs are quite conservative and attain or exceed, the nominal 95% level at a missing rate of 10% for all studied sample sizes. This also confirms the results from Ishwaran and Lu (2019). For growing missing rates, only the Rubin intervals can conserve this behavior. In particular, at 50% missing rate, the coverage of the Ishwaran intervals on the imputed data fall below 90%.

Moderate importance. For $n = 100$, coverage collapses for all methods - even at 10% missing rate. Therefore CIs for Feature Importance based on Random Forest are not reliable at this sample size. For the larger sample sizes ($n = 250$ and $n = 500$)

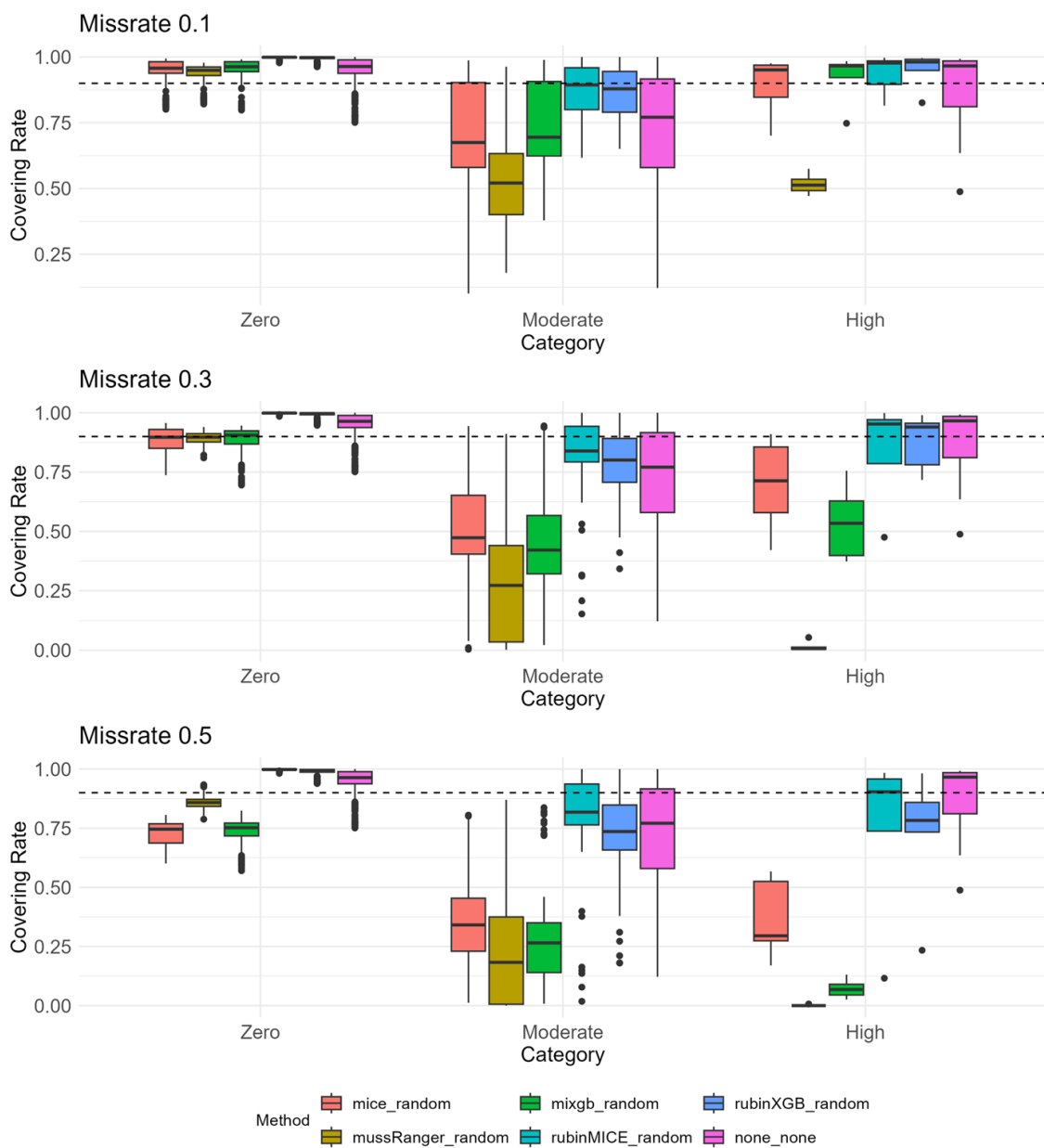


Fig. 1 Boxplot-Matrix of coverage rates for MAR, stratified by Missrate (0.1, 0.3, 0.5), sample size 250 and category (Zero, Below 90%, Above 90%)

the Rubin Intervals are superior in comparison to the Ishwaran intervals on imputed data. Especially if the imputation is made with `missranger` we find extremely low coverage rates of under 50%, which decrease even further for growing missing rates. The Rubin intervals, on the other hand have even larger coverage rates than the intervals without imputation but they pay for this robustness with increased CI length (fig. 2).

High importance. In comparison to the moderate effects, CIs for high-effect features have larger coverage rates except for the scenario with 50% missing rate. In this case `missranger` and `mixgb` produce worse coverages. The effect of `missrate` and n is similar as for the moderate effect features. At 30% missing rate and $n \geq 250$ only the Rubin intervals maintain coverages of at least 90%.

Interim conclusion: Applying Rubin's rule to multiply imputed data consistently outperforms single imputation with naive variance estimation, which leads to under-coverage due to unaccounted imputation uncertainty. The two Rubin-based methods perform similarly, and neither `mice` nor `mixgb` shows a systematic advantage in terms of coverage. In contrast, the poor coverage rates of `missRanger` are notable. To understand this behavior, we examined the variability of the imputed values and the correlations between features and the target variable. These quantities do not differ at all between `mice` and `missRanger` (see Figs. 14 and 15 in the Appendix). However, the resulting RFPIM is often smaller for the `missRanger`-imputed datasets (see Fig. 16). Since the imputation with `missRanger` seems to capture pairwise correlation well, we attribute the underestimation of permutation importance to the known difficulty of Random Forests in capturing pure interaction effects (Blum et al. 2025).

5.2 CI lengths under MAR

To compare interval lengths across different experimental conditions, the CI lengths are standardized relative to a reference interval from the complete dataset (see below). For each experimental condition, the mean CI length is calculated. The standardized CI length for each observation is obtained by dividing the length of the interval by the mean length of the corresponding reference interval. This normalization ensures that the standardized CI lengths are dimensionless and expressed as ratios relative to the baseline condition without missingness, allowing for a consistent comparison across different scales of features.

The results are again summarized in boxplots as above (Fig. 2 for $n = 250$ and 6, 7 for $n = 100, 500$) and discussed below.

Zero importance. We observe very low variance in the CI lengths with missings/imputation in comparison to the complete data intervals. The Rubin intervals are the largest, with the `mixgb` version being a little shorter than the `mice` version. The difference in lengths gets smaller with growing n . Moreover, the median length of all imputation methods is closer to the median of the CIs constructed on the complete data, when n is larger.

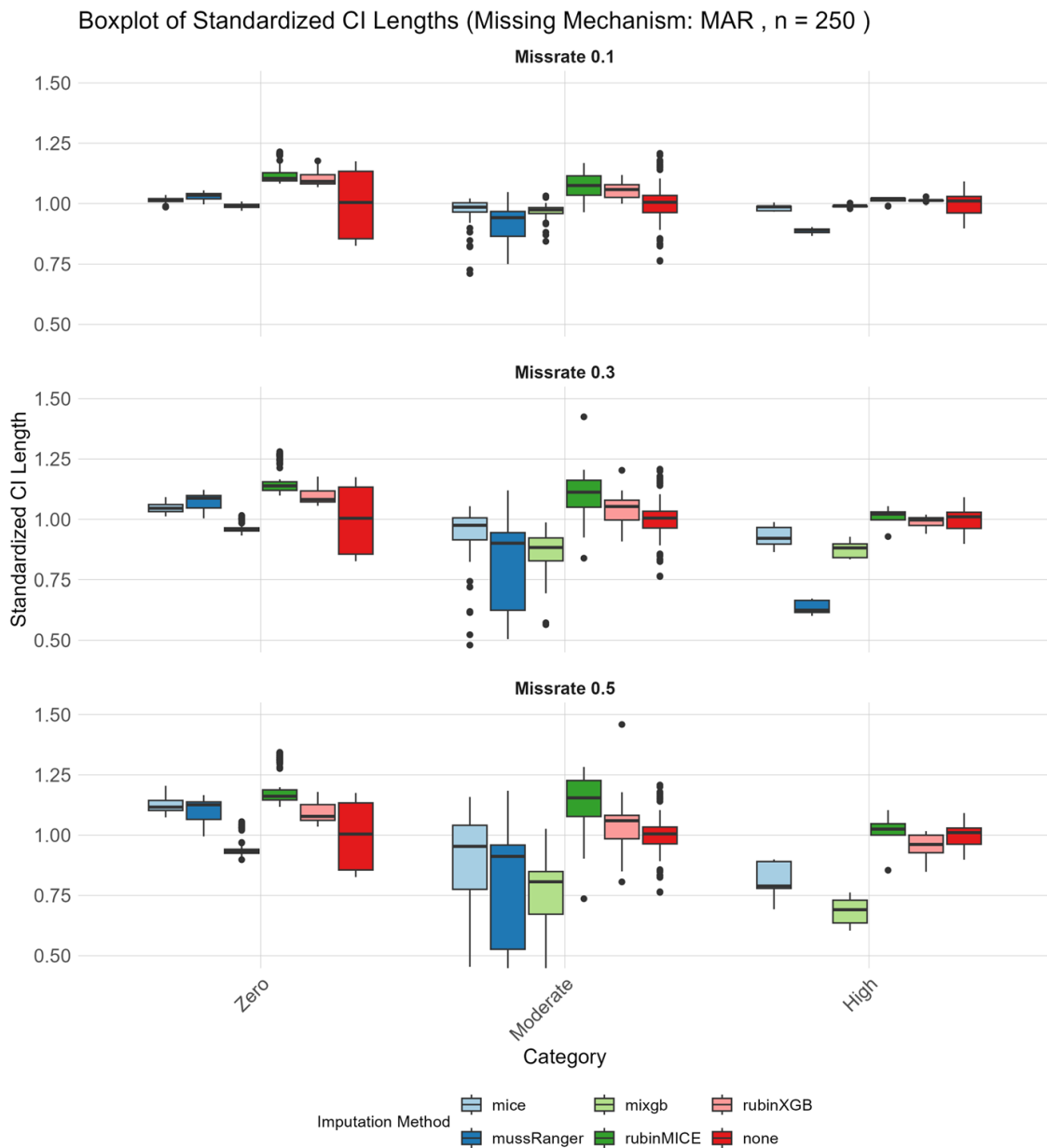


Fig. 2 Boxplot-Matrix of standardized CI lengths for MAR, stratified by Missrate (0.1, 0.3, 0.5), sample size (250) and category (Zero, Below 90%, Above 90%)

Moderate importance. The behaviour mirrors the one from the noise group. The Rubin intervals are still the longest, but with an increased gap between the `mixgb` and the `mice` version. For 30% and 50% missing rate and $n = 500$ the `mixgb` Rubin-Interval seems to be the only interval whose box is shifted upwards. Another difference of moderate effect features and noise features is the dispersion of the length of the complete data interval. For 10% missing rate the dispersions of the six intervals don't show notable differences. With larger missing rates, however, the dispersion of lengths increases sharply for `mice`- and `missRanger`.

High importance. The Rubin CIs again lead to the longest intervals. `missRanger` has by far the shortest interval for high effect features across all settings.

For $n = 500$ all intervals based on imputation datasets are shorter than the intervals on the complete dataset.

Conclusion for MAR: The Rubin-based CIs are wider but more conservative, achieving better coverage by accounting for both within- and between-imputation variability. Single imputation CIs, when paired with naive variance estimators, are too narrow due to unaccounted imputation uncertainty, resulting in systematic undercoverage.

5.3 CI coverage and length under MCAR

The results under the MCAR framework are given in Figs. 8, 9 and 10 (coverage) and 11, 12 and 13 (lengths) in the supplement. Qualitative conclusions mostly match MAR and some key takeaways are discussed below:

Sample size. With $n = 100$ all methods - including the complete data benchmark - exhibit undercoverage. RFPI inference is therefore not recommended at such small samples. With growing sample sizes ($n = 250, 500$), however, the coverages improve.

Missing rate. As expected, coverage decreases as the missing rate increases. For the approaches without Rubin's rule, coverage can even fall below 50% when the missing rate is 30% or 50%. If $n = 500$, the Rubin Interval with `mice` delivers the most robust coverage. However, for 50% missing rate it also struggles and none of the considered approaches have reliable coverages.

Feature type. For the noise features, the Rubin intervals are as conservative as the intervals from the complete data. For the features with an effect, the moderate effect features are hardest to cover as CIs become more liberal. Whether this stems from the effect size per se or the specific choice of functions.

Imputation approach. Under MAR, both Rubin approaches (`mice` and `mixgb`) had similar coverage; under MCAR, Rubin with `mice` is superior to Rubin with `mixgb`. Nevertheless, both lead to much better coverage than the Ishwaran intervals on single imputed data. Thereby, imputation with `missRanger` performs worst with coverage rates close to zero for larger missing rates.

Conclusion for MCAR: Again, Rubin intervals were the most promising approach. However, they only exhibit reliable coverages for the lowest missing rate of 10%, particularly in combination with `mice`. For larger missing rate, no method lead to reliable coverages, while Rubin with `mice` stayed more robust.

Nevertheless, the wider intervals are less informative for practitioners. As shown in Table 2, relative interval width ratios (Rubin intervals vs. complete-data CIs) range from 0.87 to 1.25 under MAR, with the largest ratios occurring at smaller sample sizes ($n = 100$) or higher missingness levels (50%). Single-imputation CIs have ratios closer to 1 but much lower coverage (e.g., 0.804 vs. 0.910 for `mice` with Rubin at $n = 100$, 10% missingness). The slight increase in width of Rubin intervals ensures more reliable coverage while keeping the intervals interpretable for practical use.

Summary: Overall, these results demonstrate that Ishwaran and Lu (2019) CIs for permutation importance scores are not suitable for imputed data and small samples. In fact, with sample size of $n = 100$, these CIs shouldn't be used at all. For moderate missing rate (30% or 10%) and adequate sample size (n at least 250), the combination of Ishwaran and Lu (2019) CIs with one of the two multiple imputation approaches via Rubin's rules was more promising. Thereby, multiple imputation with `mice` was mostly preferable as it led to more reliable coverages in case of 10% (both, for MAR and MCAR) and 30% (only for MAR) missing rate. For the highest missing rate of 50%, the results are not reliable.

5.4 Model uncertainty and Random Forest hyperparameters

The robustness of our RFPIM procedure could, in principle, depend on Random Forest hyperparameters. However, extensive prior work on Random Forest tuning has shown that the algorithm is comparatively insensitive to most of their tuning parameters, often yielding strong performance with default settings (Probst and Boulesteix 2018; Probst et al. 2019). A larger effect by tuning can be observed for datasets with many irrelevant features or complex interaction effects. The most influential hyperparameters are `mtry` (the number of variables randomly sampled at each split) and `nodesize`, which have the strongest impact on the bias–variance trade-off and thus represent the most relevant sources of model uncertainty (Scornet et al. 2015; Probst et al. 2019). In contrast, the hyperparameter `ntree` (number of trees) mainly affects computational cost but not inferential stability, as prediction and variable importance estimates converge rapidly once a sufficient number of trees is reached (typically ≥ 200) (Probst and Boulesteix 2018). Based on this evidence, we restricted our sensitivity analysis to `mtry` and `nodesize`, varying them around their standard defaults (`mtry` = $\sqrt{p}$, `nodesize` = 5). To assess robustness, we additionally repeated a small subset of simulations ($n = 250$) and Function 1 in the Appendix (cf. Table 1) while varying the Random Forest hyperparameters (`mtry` $\in \{\sqrt{p}/2, \sqrt{p}, 2\sqrt{p}\}$, `nodesize` $\in \{3, 5, 10\}$). This yields a total of 480 000 individual variable-importance values (4 `mtry` values $\times$ 3 `nodesizes` $\times$ 2 missingness mechanisms $\times$ 3 imputation methods $\times$ 250 repetitions $\times$ 20 variables). For the five truly relevant variables, the variance (i.e., the variance of the mean importance across all `mtry`–`nodesize` combinations) between hyperparameter settings accounted for only 43–56% of the natural within-setting variability (i.e., the across-replication variance of importance estimates for a fixed hyperparameter configuration), whereas for the irrelevant variables the between-setting contribution was below 5% of the total variation. Thus, the uncertainty introduced by Random Forest tuning is substantially smaller than the inherent bootstrap-based variability of the permutation importance estimator.

6 Benchmark study

To evaluate how the competing CI approaches behave on real data, we analyse five publicly available regression datasets whose features are fully observed. Below, we briefly describe each dataset used in the experiments.

- **Boston housing:** This dataset contains information about housing prices in the Boston area, with 506 instances and 13 features, including the median house value (MEDV) as the target (Harrison jr. and Rubinfeld 1978).
- **California housing:** Comprising 20,640 instances and 8 features, this dataset provides data on housing prices in California, again with the median house value as the target (Pace and Barry 1997).
- **Concrete compressive strength:** This dataset includes 1,030 instances with 8 features related to concrete mixtures, aiming to predict the compressive strength of concrete (Yeh 1998).
- **Wine quality:** The red wine quality dataset contains 1,599 instances and 11 features, with the quality score (class) as the target (Cortez et al. 2009).
- **Bike sharing:** This dataset records bike rental counts with 17,379 instances and 14 features, targeting the total rental count (count) (Fanaee-T and Gama 2013).

Experimental framework: For each dataset we delete 10% of all entries MCAR, then impute the missing values using `mice`, `missRanger`, and `mixgb`. On every completed data set, we compute (5), choosing the subsample size for the deleted Jackknife estimator $b = \sqrt{n}$ to reduce runtime on the larger datasets. The Rubin intervals are also obtained for the two multiple imputation approaches `mice` and `mixgb`. This process is repeated 100 times.

Since there is no known ground truth, we cannot compute coverage rates. Hence, we compare the length and location of the CIs. Because features live on different scales, each CI is standardized, such that the interval without imputation has center 0 and mean length=1.

Results: In Fig. 3, the standardized CIs for each feature across the five datasets are presented. The x-axes are scaled individually for each dataset, because of some outlier features. Across all five datasets, the `mice`-based methods exhibit the largest shifts in the position of the CIs (i.e., changes in both lower and upper bounds) compared to the non-imputed dataset. In contrast, Random Forest imputation with `missRanger` results in smaller positional shifts. To quantify method stability, we also computed the variance of CI midpoint shifts relative to the 'No Imputation' baseline across 100 replications. As shown in Table 3 in the Appendix, the methods minimizing the variance depends on the dataset, further illustrating the dataset-dependent behavior of RFPIIM.

For the “Boston Housing” dataset, CIs after `mice` imputation generally encompass those from the non-imputed dataset but have significantly wider upper bounds for some features. A similar pattern is observed for the Rubin intervals with `mixgb`. In the “California Housing” dataset, intervals from `mixgb` and `missRanger` show less pronounced positional shifts compared to the `mice` methods. The “Concrete Compressive Strength” data set exhibit the smallest changes across all five datasets. In the “Bike Sharing” dataset, one feature exhibits significant shifts in the position of the CIs for both `missRanger` and `mice` methods, resulting in no overlap with the intervals from the non-imputed dataset. For this feature, only the two `mixgb` methods remain relatively robust, maintaining intervals with minimal positional

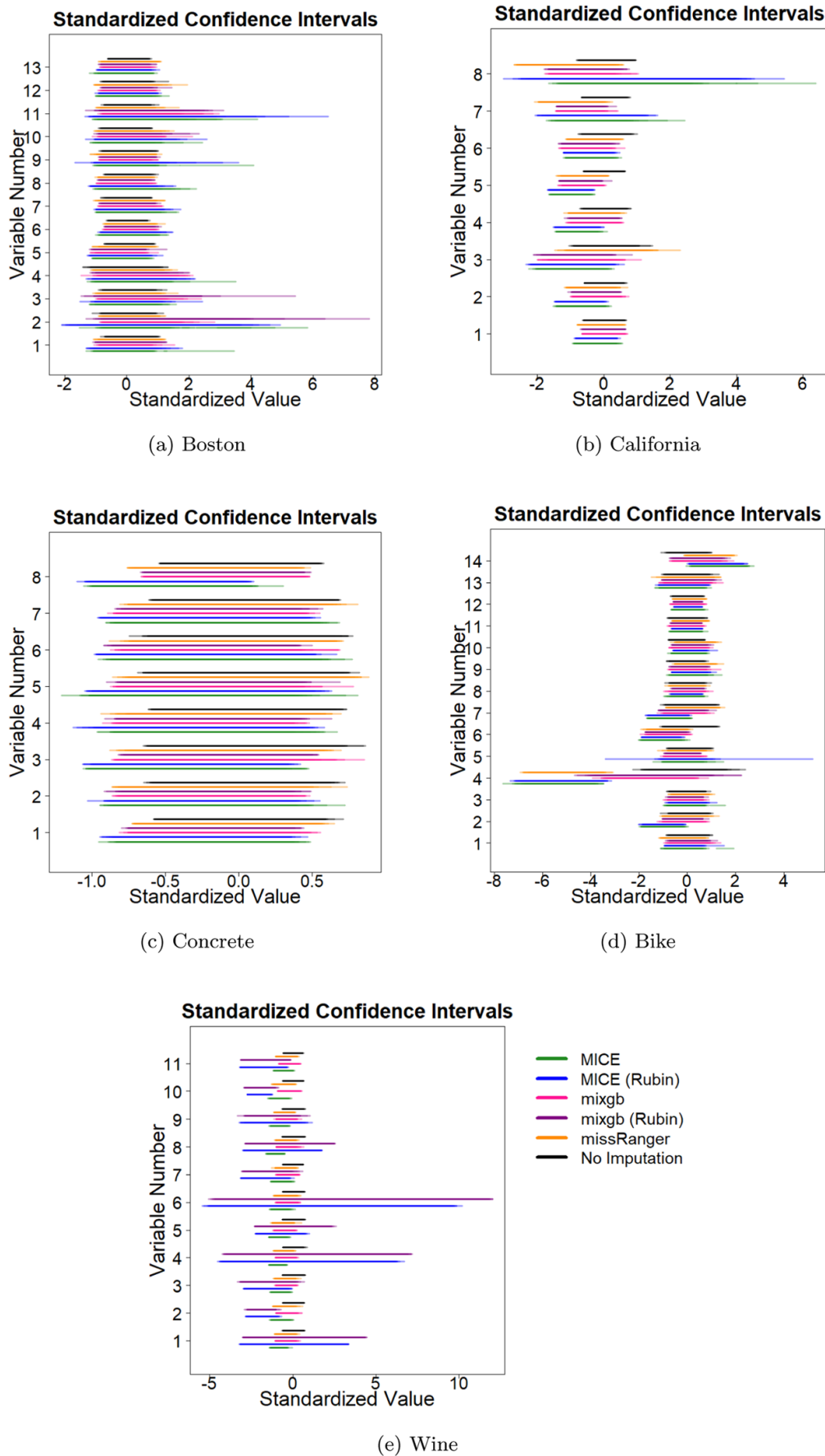


Fig. 3 Standardized CIs (mean length=1, center=0) relative to the complete data benchmark. Thereby x-axes are scaled individually for each dataset

shifts. The most notable results are observed for the “Wine Quality” dataset. Here, `mice_rubin` intervals are significantly shifted in their position compared to the non-imputed intervals and are, for some features, two to three times wider than those from other methods. In contrast, `missRanger` and `mixgb` methods show more moderate shifts and narrower intervals.

Regarding interval lengths, several features exhibit intervals that are substantial – sometimes several times – longer for particular imputation methods than for the others. The most extreme examples occur for variables 2 and 11 from the “Boston Housing” dataset, variable 8 from the “California Housing” dataset, and variables 4 and 6 from the “Wine Quality” dataset. A closer inspection of these variables gives a potential explanation for these inflated intervals (for detailed plots see Fig. 17). For variable 2 of the “Boston Housing” dataset, the Rubin intervals and the `mice`-based interval are much longer than those from the other methods. Over 70% of this feature’s values are zero with a small number of large outliers. These zeros span the entire range of the target variable, creating a complex, non-linear relationship. Since `mice` relies on linear models, it struggles to capture this non-linearity, leading to unstable imputations and increased variance in the feature importance estimates. Multiple imputation generates several imputed datasets to account for missing data. The high proportion of zeros and sparse outliers leads to diverse imputations, as imputation models struggle to consistently predict the rare non-zero values. This variability across imputed datasets increases the between-imputation variance, which Rubin’s rules incorporate into the total variance estimate thereby inflating the resulting confidence intervals. For variable 11, `mice` intervals (with and without Rubin’s Rules) are strongly inflated in size. Here, the dependence between variable 11 and the target variable exhibits a rather logarithmic pattern (see Fig. 17), which again violates `mice`’s linearity assumptions, yielding unstable imputations. Variable 8 from the “California Housing” dataset also displays a non-linear relationship 17. Consequently, `mice` produces longer intervals, similar to the patterns observed in the “Boston Housing” dataset. At the “Wine Quality” dataset, variables 4 and 6 exhibit almost no correlation with the target variable (0.014 resp. -0.051), but include some outliers for the middle range of the target variable. As with variable 2 in the “Boston Housing” dataset, the outliers induce high between-imputation variability for multiple imputation, translates into substantially wider Rubin intervals.

Conclusions: This benchmark study illustrates qualitative patterns in the performance of imputation methods across the “Concrete”, “Wine Quality”, “Boston Housing”, “California Housing”, and “Bike Sharing” datasets. Since the true feature importance $I(j)$ is unknown, the coverage of Rubin intervals and Ishwaran CIs cannot be verified, limiting our conclusions to descriptive trends rather than method validation. Notably, Rubin intervals are generally larger than the others. These patterns align with simulation results (section 5), where Rubin intervals show robust coverage for moderate missing rates ($\leq 30\%$). Importantly, no single method uniformly minimises deviations from the complete-data CIs. However, for certain variables – especially on the “Wine Quality” dataset – the Rubin intervals can be much wider than the corresponding complete-data intervals. This reduces interpretability for these features and highlights the need for caution when applying Rubin intervals in settings with complex distributional patterns or influential outliers

7 Discussion

Missing data are ubiquitous in applied data analyses. This also holds for machine-learning applications. However, the properties of post-imputation inference remain poorly understood. In this study, we investigated the robustness of Ishwaran's (Ishwaran and Lu 2019) CIs for Random Forest Permutation Importance (RFPIM) in the presence of missing data. Through an extensive simulation study, we evaluated the performance of these intervals under various missing data scenarios (MCAR and MAR) and sample sizes ($n = 100, 250, 500$), comparing three different imputation methods: `mice`, Random Forest-based imputation (`missRanger`), and XGBoost-based imputation (`mixgb`). Additionally, we proposed a heuristic approach based on Rubin's rules to combine multiple imputations, aiming at improving coverage. An additional benchmark study on five real-world datasets complemented our simulations, examining the behavior of the CIs in practical settings.

Our **key findings** from the simulation study were as follows:

1. Single imputation with naive variance estimation leads to severe undercoverage. Direct application of Ishwaran's (Ishwaran and Lu 2019) CIs to singly imputed datasets – while treating imputed values as if they were observed – results in substantial undercoverage, particularly for moderate to high missing rates or small sample sizes ($n = 100$). This is consistent with findings in parametric settings, where single imputation underestimates uncertainty (Rubin 1987b; Van Buuren and Groothuis-Oudshoorn 2011). Our Rubin-inspired approach mitigates these issues by combining multiple imputations to account for both within- and between-imputation variability, offering a practical solution for RFPIM inference. While an adjusted variance estimator for single imputation could theoretically achieve comparable coverage, developing such an estimator for RFPIM is challenging due to the complex dependencies in OOB samples and the dependent nature of Random Forests. Thus, multiple imputation with Rubin's rule appears to be a practical and convenient alternative.

2. Rubin-based CIs substantially improve coverage but are wider. Our simulations show that Rubin-based CIs substantially improve coverage, often keeping the nominal 95% level when n is not too small ($n \geq 250$) or missing rates not too large ($\leq 30\%$). The price is wider intervals.

3. Imputation methods differ considerably. Among the methods examined, `missRanger` often leads to the shortest intervals with severe under-coverage, `mice` and `mixgb` behave more stably, particularly `mice` with Rubin's rule. In the benchmark study, however, no single method consistently minimized deviations from complete-data intervals, and performance varied substantially across datasets.

Implications for practitioners: Our results can be seen as a warning that imputation can destroy the nominal properties and thus the interpretability of RFPIM CIs. If imputation cannot be avoided, multiple imputation in combination with Rubin's rules is recommended. In general, the CIs have to be treated interval estimates with caution, particularly for moderate-signal features or high missing rate.

The introduced methodology is broadly applicable across research domains where Random Forests are used to interpret complex datasets with missing values. In bio-

informatics, it may support the identification of key genetic markers from incomplete genomic measurements. In finance, it can support risk modeling with partially observed transactional or behavioural data. In industrial applications – such as predictive maintenance based upon sensor data – our approach offers more reliable feature importance estimates despite missing values. These domains benefit from the ability of Rubin intervals, which accommodate non-linear interactions and imputation uncertainty, enhancing interpretability in real-world settings with moderate missing rates ($\leq 30\%$).

Limitations and future work: The Rubin-inspired procedures are heuristic and beside our numerical results no theorem guarantees its validity for RFPIM. Extending central limit theorems such as Föge et al. (2024) to incorporate imputation uncertainty thus is an important theoretical task.

Ishwaran’s (Ishwaran and Lu 2019) CIs themselves need theoretical investigations. In particular, the theoretical assumptions underlying the asymptotic normality of RFPIM, as established by Föge et al. (2024), are restrictive (independent features, i.i.d. setting, fixed dimensions etc.) and may limit the generalizability of our findings. Relaxing these and proving consistency of (4) for $\text{Var}(I_{n,M}^{OOB}(j))$ remains unclear. Moreover, only one specific MAR mechanism was studied and other realistic scenarios as well as MNAR frameworks are of interest.

A practical limitation is the statistical instability of both Rubin-based and Ishwaran CIs for small sample sizes ($n = 100$) or high missingness ($\geq 50\%$), conditions prevalent, e.g., in medical or social science datasets with limited observations or extensive non-response. This instability restricts the method’s reliability in scenarios with small n and high missingness. However, these scenarios are also inherently challenging: Random Forests are generally more reliable and stable for moderate to large sample sizes, and imputation methods themselves perform best when missingness is not extreme. Our empirical results are thus only in line with these expectations. Future work should nevertheless explore alternative or hybrid imputation strategies designed for small-sample or high-missingness settings, as well as the development of variance estimators tailored to single imputation for Random Forest-based permutation importance. In addition, settings with correlated predictors warrant closer examination if the theoretical properties of feature importance measures in the presence of multicollinearity become better understood.

Summary: Until more results, particularly theoretical, we recommend to use the proposed multiple imputation with Rubin’s rule while nevertheless interpreting RFPIM CIs from imputed data with appropriate scepticism.

Appendix A Tables

See Tables 1, 2 and 3 .

Table 1 Overview of simulated regression functions used in the experiments

Function	Formula
Function 1 (Sine Model)	$x \sim U([0, 1]^p), \epsilon \sim N(0, 1),$ $y = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon$
Function 2 (Radial-Distance Model)	$x_1 \sim U([0, 100]), x_2 \sim U([40\pi, 560\pi]), x_3 \sim U([0, 1]),$ $x_4 \sim U([1, 11]), \text{others} \sim U([0, 1]), \epsilon \sim N(0, 125),$ $y = \sqrt{x_1^2 + (x_2 x_3 - (x_2 x_4)^{-1})^2} + \epsilon$
Function 3 (Arctan Model)	x as in Function 2, $\epsilon \sim N(0, 0.1),$ $y = \arctan\left(\frac{x_2 x_3 - (x_2 x_4)^{-1}}{x_1}\right) + \epsilon$
Function 4 (Mixed Polynomial Model)	$x \sim U([-1, 1]^p), \epsilon \sim N(0, 0.1),$ $y = x_1 x_2 + x_3^2 + x_4 x_7 + x_8 x_{10} - x_6^2 + \epsilon$
Function 5 (Threshold Polynomial–Exponential Model)	$x \sim U([-1, 1]^p), \epsilon \sim N(0, 0.1),$ $y = \mathbb{1}_{\{x_1 > 0\}} + x_2^3 + \mathbb{1}_{\{x_4 + x_6 - x_8 - x_9 > 1 + x_{10}\}} + e^{-x_2^2} + \epsilon$
Function 6 (Mixed Polynomial–Exponential Model)	$x \sim U([-1, 1]^p), \epsilon \sim N(0, 0.1),$ $y = x_1^2 + 3x_2^2 x_3 e^{- x_4 } + x_6 - x_8 + \epsilon$
Function 7 (Threshold Model)	$x \sim U([-1, 1]^p), \epsilon \sim N(0, 0.1),$ $y = \mathbb{1}(x_1 + x_4^3 + x_9 + \sin(x_2 x_8) + \epsilon > 0.38)$
Function 8 (Mixed Log–Exponential Model)	$x \sim U([0.5, 1]^p), \epsilon \sim N(0, 0.1),$ $y = \log(x_1 + x_2 x_3) - e^{x_4/x_5 - x_6} + \epsilon$
Function 9 (Mixed Polynomial–Root–Floor Model)	$x \sim U([-1, 1]^p), \epsilon \sim N(0, 0.1),$ $y = x_1 x_2^2 \sqrt{ x_3 } + \lfloor x_4 - x_5 x_6 \rfloor + \epsilon$
Function 10 (Mixed Power–Root Model)	$x \sim U([-1, 1]^p), \epsilon \sim N(0, 0.1),$ $y = x_3(x_1 + 1)^{ x_2 } - \sqrt{\frac{x_5^2}{ x_4 + x_5 + x_6 }} + \epsilon$
Function 11 (Mixed Trigonometric Model)	$x \sim U([-1, 1]^p), \epsilon \sim N(0, 0.1),$ $y = \cos(x_1 - x_2) + \arcsin(x_1 x_3) - \arctan(x_2 - x_3^2) + \epsilon$
Function 12 (Noise Model)	$x_j \sim N(0, 1), \epsilon \sim N(0, 1),$ $y = \epsilon$

$U([a, b]^p)$ denotes a p -dimensional uniform distribution on $[a, b]$. $\mathbb{1}(\cdot)$ is the indicator function. ϵ denotes Gaussian noise with the specified variance

Table 2 Simulation results: coverage and CI length (with standard error) across imputation methods

Mechanism	N	Imputation Method	Missrate	Coverage	SE (Coverage)	CI Length	SE (Length)
MCAR	100	None	0.1	0.903	0.005	14.464	2.300
MCAR	100	None	0.3	0.903	0.005	14.464	2.300
MCAR	100	None	0.5	0.903	0.005	14.464	2.300
MCAR	100	mice	0.1	0.804	0.013	15.883	4.087
MCAR	100	mice	0.3	0.719	0.014	15.530	3.746
MCAR	100	mice	0.5	0.619	0.014	14.840	3.200
MCAR	100	mixgb	0.1	0.811	0.013	15.782	4.088
MCAR	100	mixgb	0.3	0.724	0.015	14.293	3.497
MCAR	100	mixgb	0.5	0.621	0.015	12.781	2.863
MCAR	100	missRanger	0.1	0.774	0.015	15.392	3.759
MCAR	100	missRanger	0.3	0.703	0.018	14.425	3.162
MCAR	100	missRanger	0.5	0.670	0.018	13.810	2.898
MCAR	100	mice (Rubin)	0.1	0.907	0.013	17.543	4.375
MCAR	100	mice (Rubin)	0.3	0.899	0.014	19.049	4.557
MCAR	100	mice (Rubin)	0.5	0.891	0.017	22.765	5.022
MCAR	100	mixgb (Rubin)	0.1	0.892	0.013	17.004	4.258
MCAR	100	mixgb (Rubin)	0.3	0.860	0.016	16.307	3.899
MCAR	100	mixgb (Rubin)	0.5	0.846	0.019	16.912	3.792
MCAR	250	None	0.1	0.903	0.005	14.464	2.300
MCAR	250	None	0.3	0.903	0.005	14.464	2.300
MCAR	250	None	0.5	0.903	0.005	14.464	2.300
MCAR	250	mice	0.1	0.887	0.010	14.449	3.944
MCAR	250	mice	0.3	0.794	0.013	14.330	3.778
MCAR	250	mice	0.5	0.644	0.012	14.119	3.422
MCAR	250	mixgb	0.1	0.901	0.008	14.358	3.936
MCAR	250	mixgb	0.3	0.779	0.014	13.162	3.439
MCAR	250	mixgb	0.5	0.625	0.015	11.857	2.840
MCAR	250	missRanger	0.1	0.845	0.013	14.174	3.643
MCAR	250	missRanger	0.3	0.747	0.019	13.288	3.005
MCAR	250	missRanger	0.5	0.712	0.020	12.565	2.649
MCAR	250	mice (Rubin)	0.1	0.953	0.008	15.480	4.103
MCAR	250	mice (Rubin)	0.3	0.928	0.012	16.113	4.138
MCAR	250	mice (Rubin)	0.5	0.917	0.012	19.252	4.630
MCAR	250	mixgb (Rubin)	0.1	0.951	0.006	15.170	4.038
MCAR	250	mixgb (Rubin)	0.3	0.886	0.013	14.480	3.691
MCAR	250	mixgb (Rubin)	0.5	0.840	0.018	14.492	3.447
MCAR	500	None	0.1	0.903	0.005	14.464	2.300
MCAR	500	None	0.3	0.903	0.005	14.464	2.300
MCAR	500	None	0.5	0.903	0.005	14.464	2.300
MCAR	500	mice	0.1	0.922	0.007	12.621	3.625
MCAR	500	mice	0.3	0.826	0.014	12.457	3.469
MCAR	500	mice	0.5	0.672	0.012	12.474	3.279
MCAR	500	mixgb	0.1	0.935	0.005	12.514	3.616
MCAR	500	mixgb	0.3	0.807	0.012	11.541	3.191
MCAR	500	mixgb	0.5	0.602	0.014	10.449	2.648
MCAR	500	missRanger	0.1	0.888	0.009	12.383	3.309
MCAR	500	missRanger	0.3	0.756	0.021	11.789	2.736

Table 2 (continued)

Mechanism	N	Imputation Method	Missrate	Coverage	SE (Coverage)	CI Length	SE (Length)
MCAR	500	missRanger	0.5	0.722	0.021	11.220	2.386
MCAR	500	mice (Rubin)	0.1	0.973	0.005	13.333	3.722
MCAR	500	mice (Rubin)	0.3	0.937	0.012	13.658	3.705
MCAR	500	mice (Rubin)	0.5	0.926	0.012	15.663	3.993
MCAR	500	mixgb (Rubin)	0.1	0.975	0.003	13.146	3.692
MCAR	500	mixgb (Rubin)	0.3	0.909	0.011	12.427	3.343
MCAR	500	mixgb (Rubin)	0.5	0.826	0.017	12.355	3.074
MAR	100	None	0.1	0.903	0.005	14.464	2.300
MAR	100	None	0.3	0.903	0.005	14.464	2.300
MAR	100	None	0.5	0.903	0.005	14.464	2.300
MAR	100	mice	0.1	0.804	0.013	15.971	4.111
MAR	100	mice	0.3	0.717	0.014	15.456	3.732
MAR	100	mice	0.5	0.620	0.014	14.883	3.210
MAR	100	mixgb	0.1	0.810	0.013	15.702	4.058
MAR	100	mixgb	0.3	0.724	0.015	14.379	3.514
MAR	100	mixgb	0.5	0.617	0.015	12.732	2.842
MAR	100	missRanger	0.1	0.775	0.015	15.470	3.779
MAR	100	missRanger	0.3	0.703	0.018	14.403	3.145
MAR	100	missRanger	0.5	0.671	0.017	13.896	2.915
MAR	100	mice (Rubin)	0.1	0.910	0.012	17.390	4.374
MAR	100	mice (Rubin)	0.3	0.909	0.013	17.667	4.386
MAR	100	mice (Rubin)	0.5	0.910	0.013	18.041	4.434
MAR	100	mixgb (Rubin)	0.1	0.902	0.012	17.223	4.344
MAR	100	mixgb (Rubin)	0.3	0.888	0.013	16.845	4.196
MAR	100	mixgb (Rubin)	0.5	0.877	0.015	16.527	4.062
MAR	250	None	0.1	0.903	0.005	14.464	2.300
MAR	250	None	0.3	0.903	0.005	14.464	2.300
MAR	250	None	0.5	0.903	0.005	14.464	2.300
MAR	250	mice	0.1	0.887	0.010	14.460	3.948
MAR	250	mice	0.3	0.796	0.013	14.296	3.769
MAR	250	mice	0.5	0.643	0.012	14.211	3.453
MAR	250	mixgb	0.1	0.899	0.008	14.312	3.929
MAR	250	mixgb	0.3	0.779	0.014	13.169	3.441
MAR	250	mixgb	0.5	0.625	0.015	11.838	2.834
MAR	250	missRanger	0.1	0.844	0.013	14.096	3.617
MAR	250	missRanger	0.3	0.746	0.019	13.263	3.000
MAR	250	missRanger	0.5	0.711	0.019	12.639	2.667
MAR	250	mice (Rubin)	0.1	0.968	0.005	15.405	4.112
MAR	250	mice (Rubin)	0.3	0.952	0.008	15.482	4.087
MAR	250	mice (Rubin)	0.5	0.943	0.010	15.730	4.125
MAR	250	mixgb (Rubin)	0.1	0.966	0.005	15.320	4.102
MAR	250	mixgb (Rubin)	0.3	0.945	0.007	15.055	3.987
MAR	250	mixgb (Rubin)	0.5	0.923	0.010	14.648	3.829
MAR	500	None	0.1	0.903	0.005	14.464	2.300
MAR	500	None	0.3	0.903	0.005	14.464	2.300
MAR	500	None	0.5	0.903	0.005	14.464	2.300
MAR	500	mice	0.1	0.922	0.008	12.596	3.614

Table 2 (continued)

Mechanism	N	Imputation Method	Missrate	Coverage	SE (Coverage)	CI Length	SE (Length)
MAR	500	mice	0.3	0.824	0.014	12.461	3.470
MAR	500	mice	0.5	0.672	0.012	12.458	3.277
MAR	500	mixgb	0.1	0.935	0.005	12.481	3.607
MAR	500	mixgb	0.3	0.809	0.012	11.580	3.205
MAR	500	mixgb	0.5	0.601	0.014	10.405	2.644
MAR	500	missRanger	0.1	0.889	0.009	12.387	3.309
MAR	500	missRanger	0.3	0.756	0.021	11.773	2.731
MAR	500	missRanger	0.5	0.722	0.021	11.174	2.375
MAR	500	mice (Rubin)	0.1	0.981	0.003	13.324	3.743
MAR	500	mice (Rubin)	0.3	0.968	0.006	13.362	3.714
MAR	500	mice (Rubin)	0.5	0.946	0.011	13.461	3.717
MAR	500	mixgb (Rubin)	0.1	0.973	0.004	13.261	3.744
MAR	500	mixgb (Rubin)	0.3	0.970	0.004	12.961	3.620
MAR	500	mixgb (Rubin)	0.5	0.948	0.007	12.629	3.480

Table 3 Mean variance of confidence interval midpoint shifts relative to “No Imputation” across variables and 100 replications

Dataset	Method	Mean Var(Shift)
Bike	mice	0.1353
	mice (Rubin)	0.0870
	missRanger	0.1378
	mixgb	0.1579
	mixgb (Rubin)	0.0979
Boston	mice	0.1709
	mice (Rubin)	0.1025
	missRanger	0.0789
	mixgb	0.1023
	mixgb (Rubin)	0.0968
California	mice	0.2279
	mice (Rubin)	0.1099
	missRanger	0.1578
	mixgb	0.0873
	mixgb (Rubin)	0.0514
Concrete	mice	0.0293
	mice (Rubin)	0.0182
	missRanger	0.0211
	mixgb	0.0178
	mixgb (Rubin)	0.0129
Wine Quality	mice	0.0273
	mice (Rubin)	0.0216
	missRanger	0.0276
	mixgb	0.0251
	mixgb (Rubin)	0.0263

Appendix B MAR Plots

See Figs. 4, 5, 6 and 7.

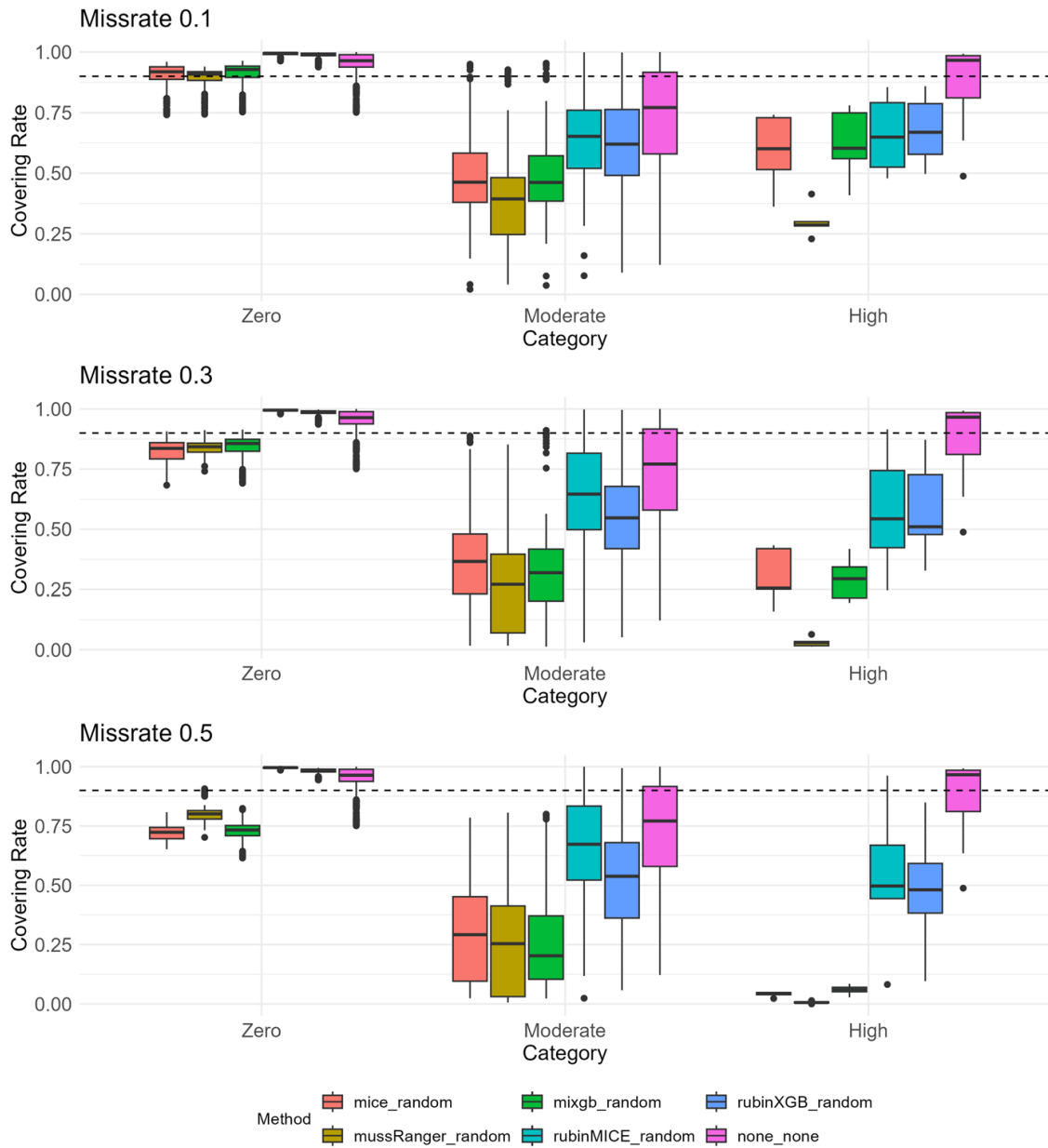


Fig. 4 Boxplot-Matrix of coverage rates for MAR, stratified by Missrate (0.1, 0.3, 0.5), sample size 100 and category (Zero, Below 90%, Above 90%)

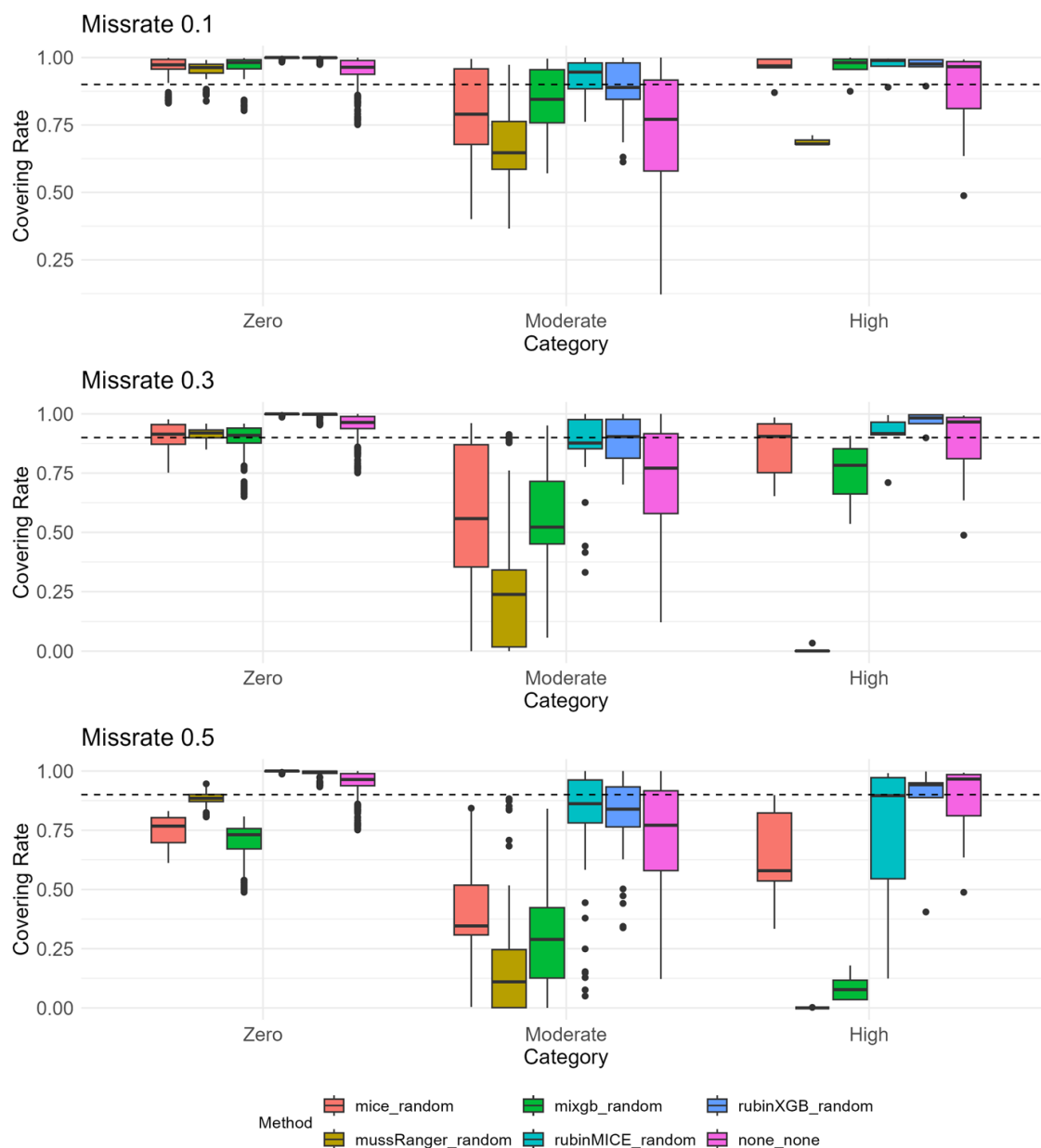


Fig. 5 Boxplot-Matrix of coverage rates for MAR, stratified by Missrate (0.1, 0.3, 0.5), sample size 500 and category (Zero, Below 90%, Above 90%)

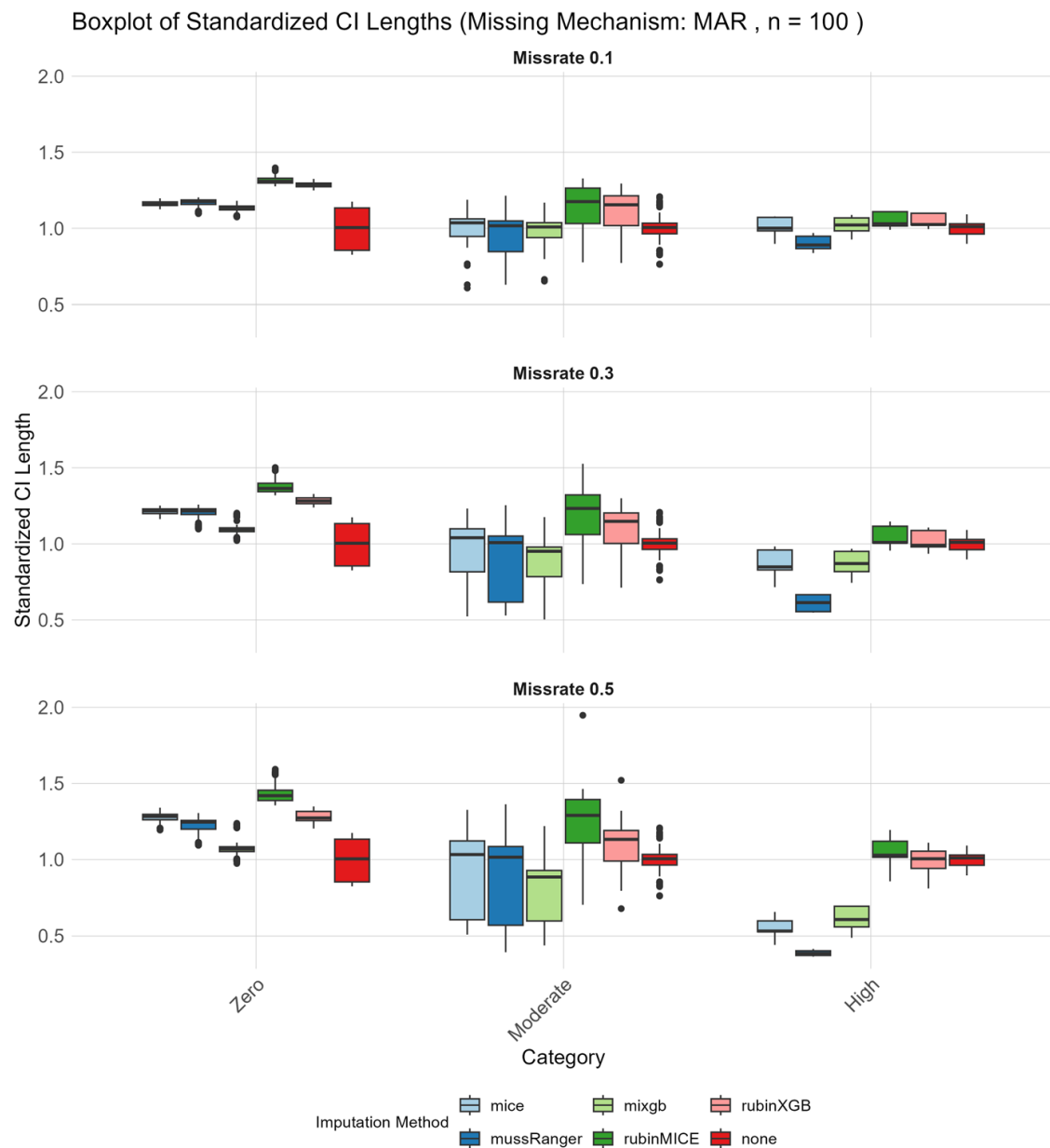


Fig. 6 Boxplot-Matrix of standardized CI lengths for MAR, stratified by Missrate (0.1, 0.3, 0.5), sample size (500) and category (Zero, Below 90%, Above 90%)

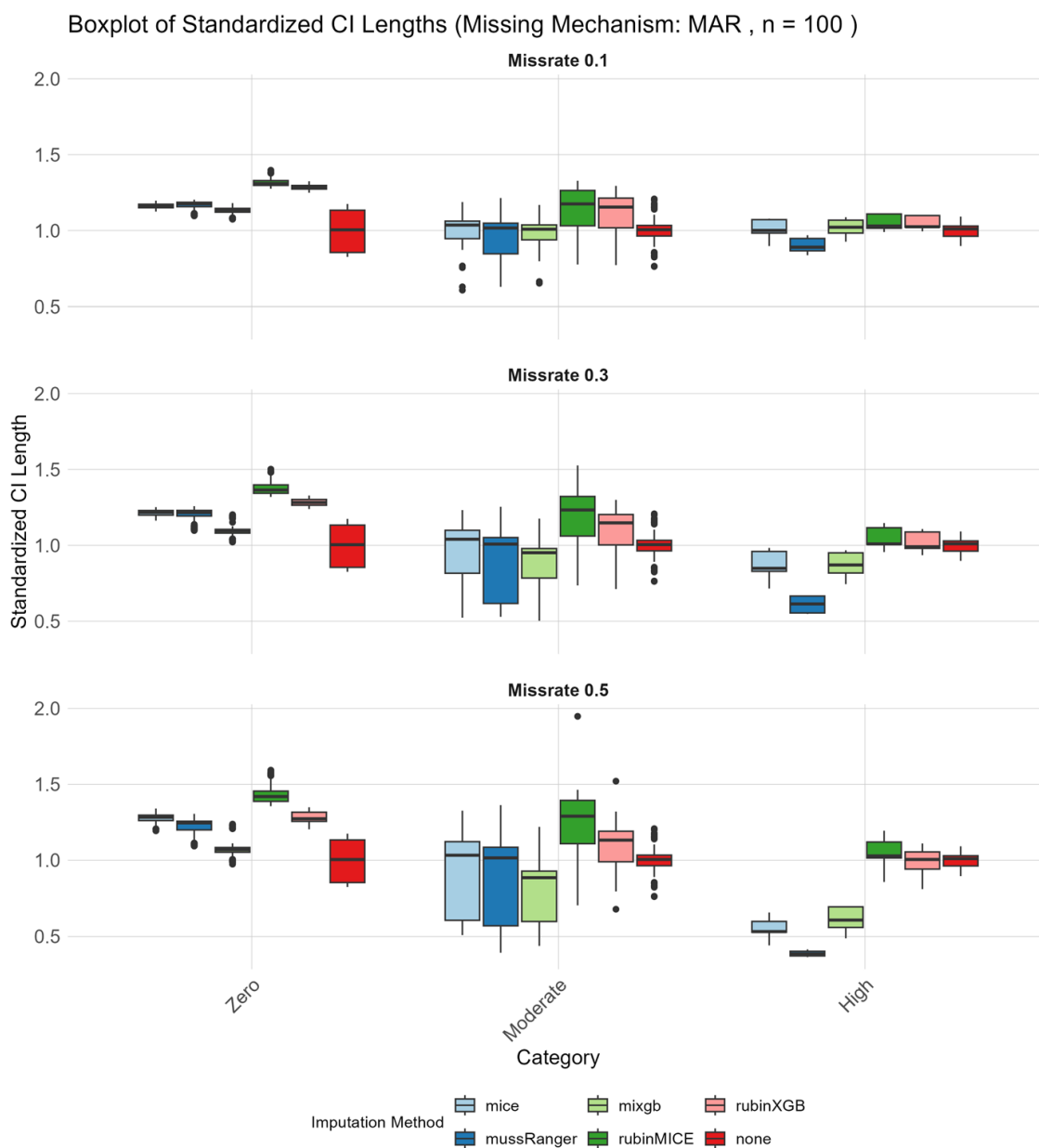


Fig. 7 Boxplot-Matrix of standardized CI lengths for MAR, stratified by Missrate (0.1, 0.3, 0.5), sample size (500) and category (Zero, Below 90%, Above 90%)

Appendix C MCAR plots

See Figs. 8, 9, 10, 11, 12 and 13.

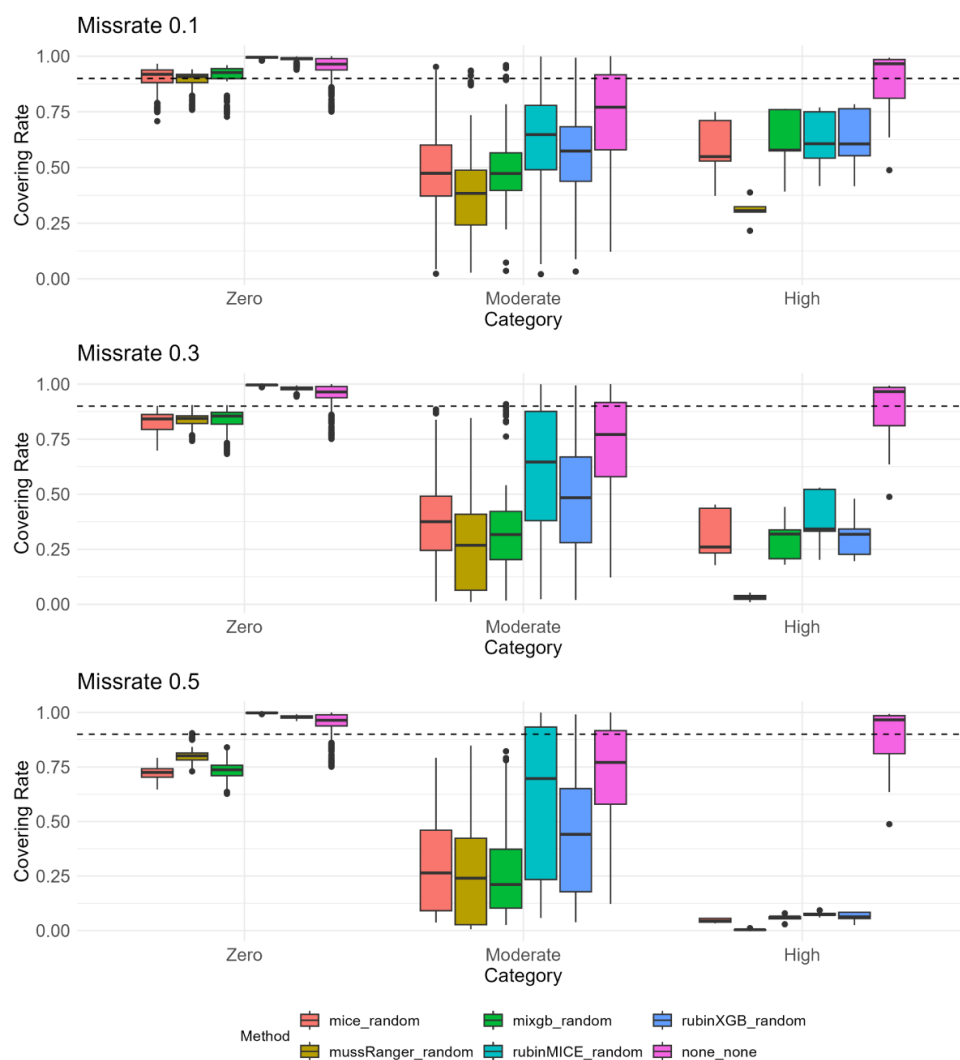


Fig. 8 Boxplot-Matrix of coverage rates for MCAR, stratified by Missrate (0.1, 0.3, 0.5), sample size 100 and category (Zero, Below 90%, Above 90%)

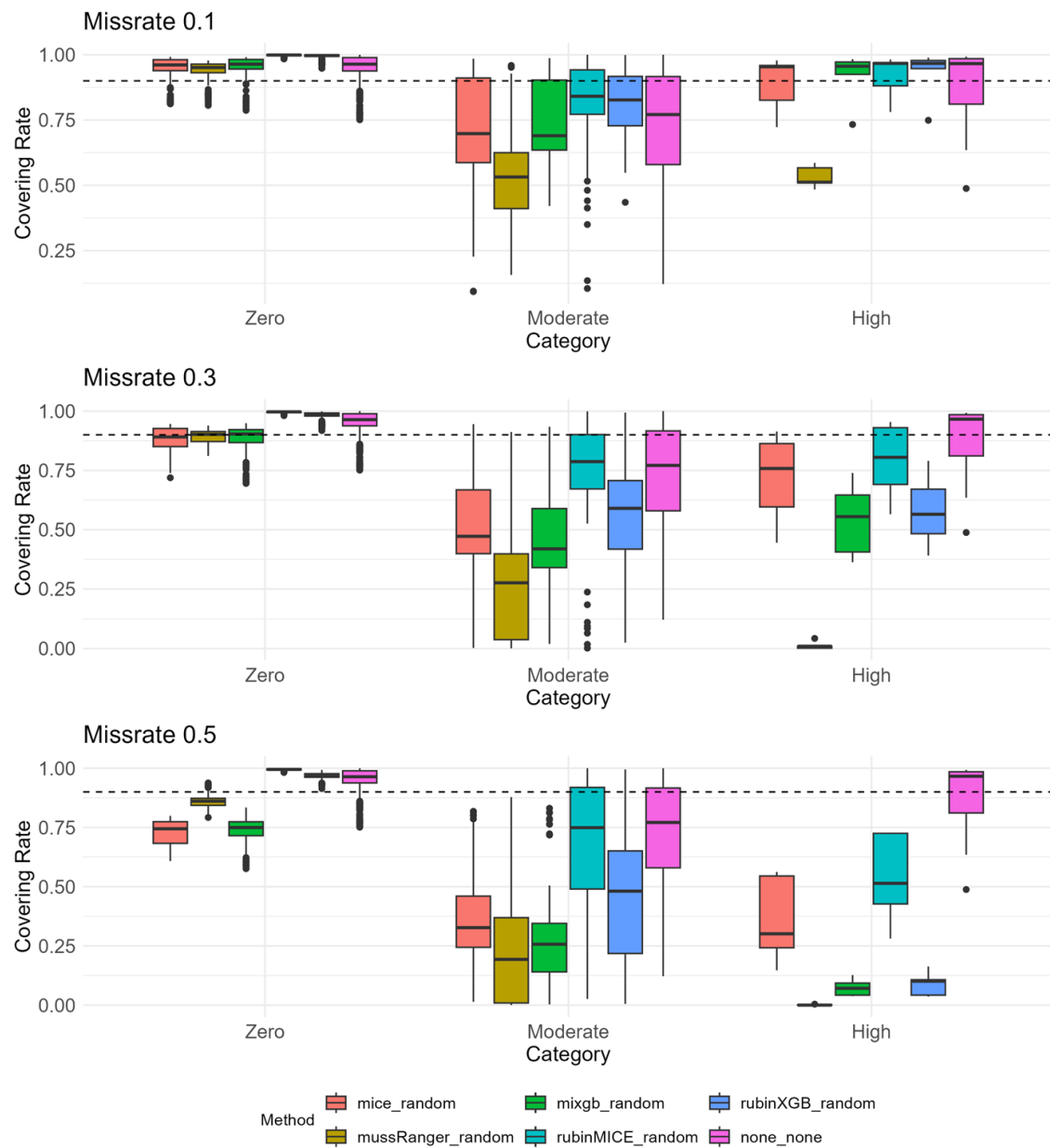


Fig. 9 Boxplot-Matrix of coverage rates for MCAR, stratified by Missrate (0.1, 0.3, 0.5), sample size 250 and category (Zero, Below 90%, Above 90%)

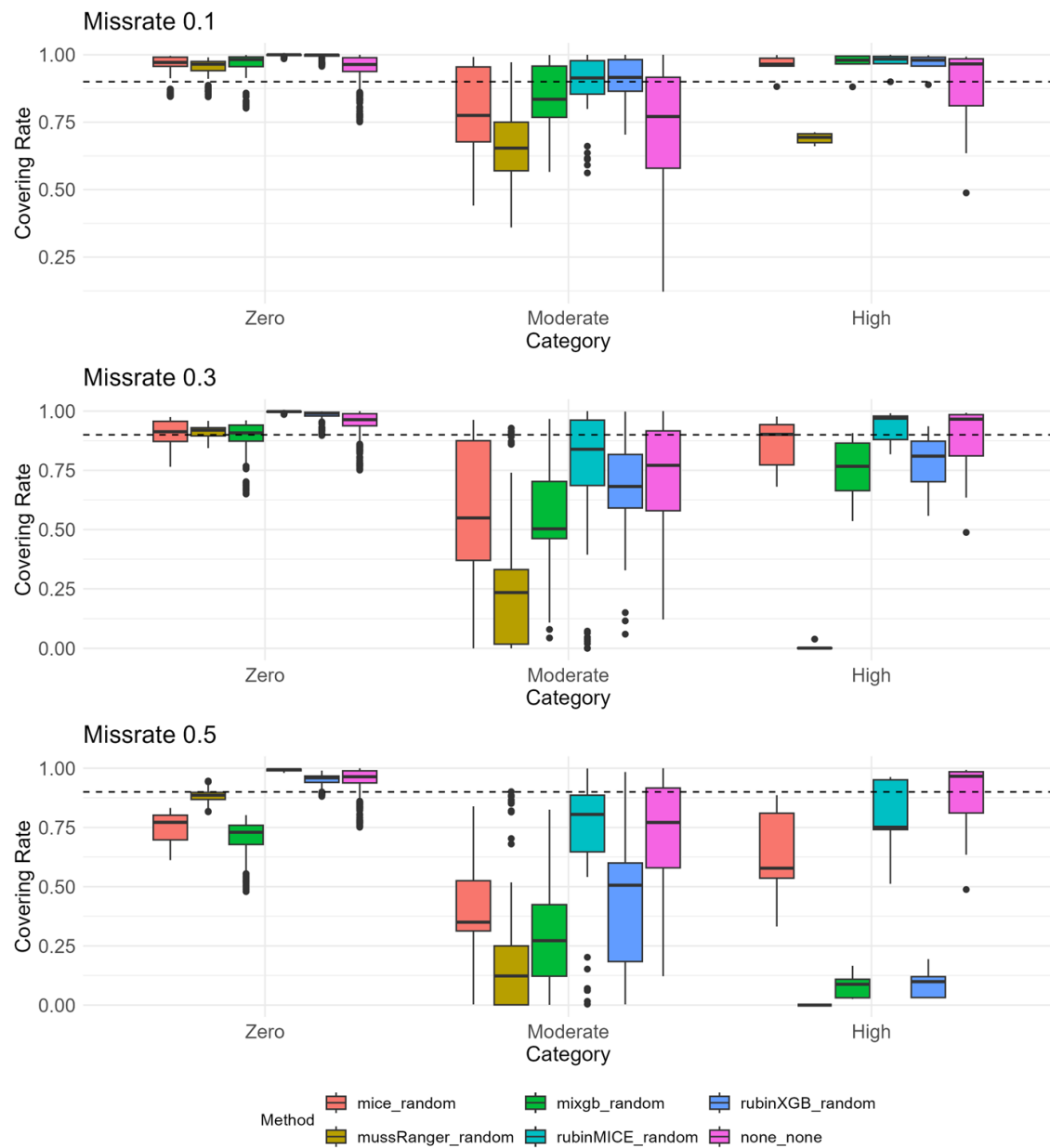


Fig. 10 Boxplot-Matrix of coverage rates for MCAR, stratified by Missrate (0.1, 0.3, 0.5), sample size 500 and category (Zero, Below 90%, Above 90%)

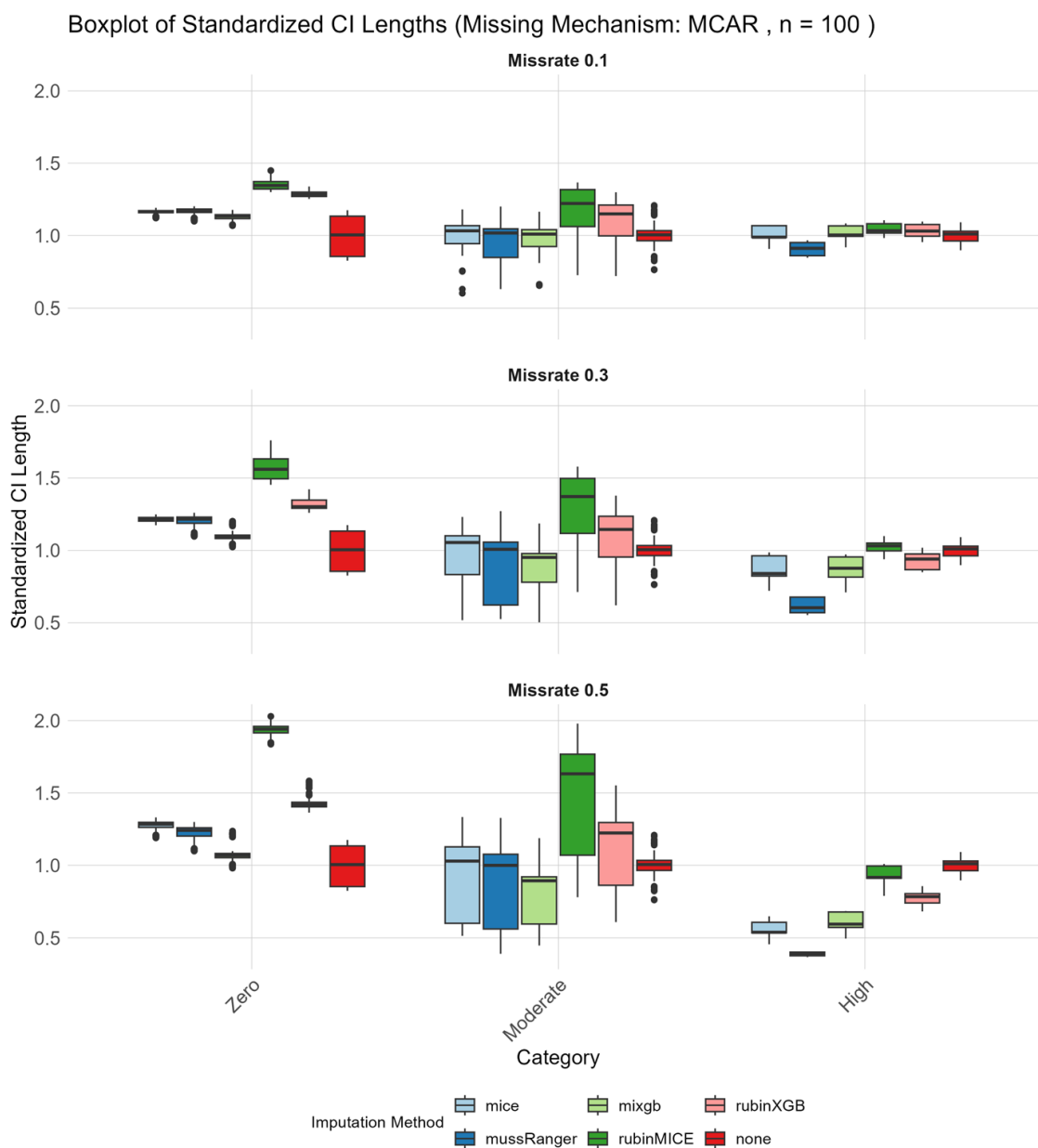


Fig. 11 Boxplot-Matrix of standardized CI lengths for MCAR, stratified by Missrate (0.1, 0.3, 0.5), sample size 100 and category (Zero, Below 90%, Above 90%)

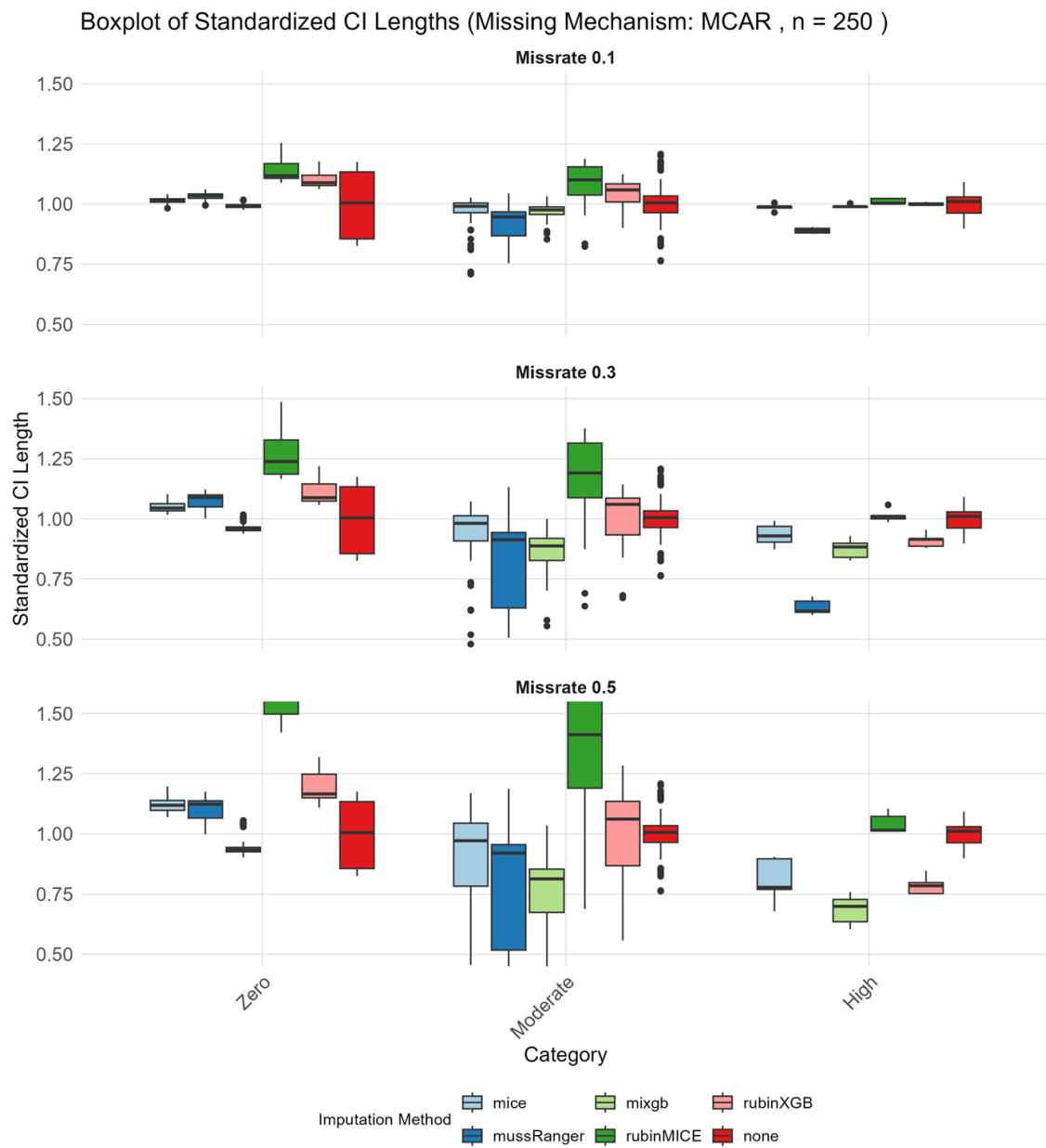


Fig. 12 Boxplot-Matrix of standardized CI lengths for MCAR, stratified by Missrate (0.1, 0.3, 0.5), sample size 250 and category (Zero, Below 90%, Above 90%)

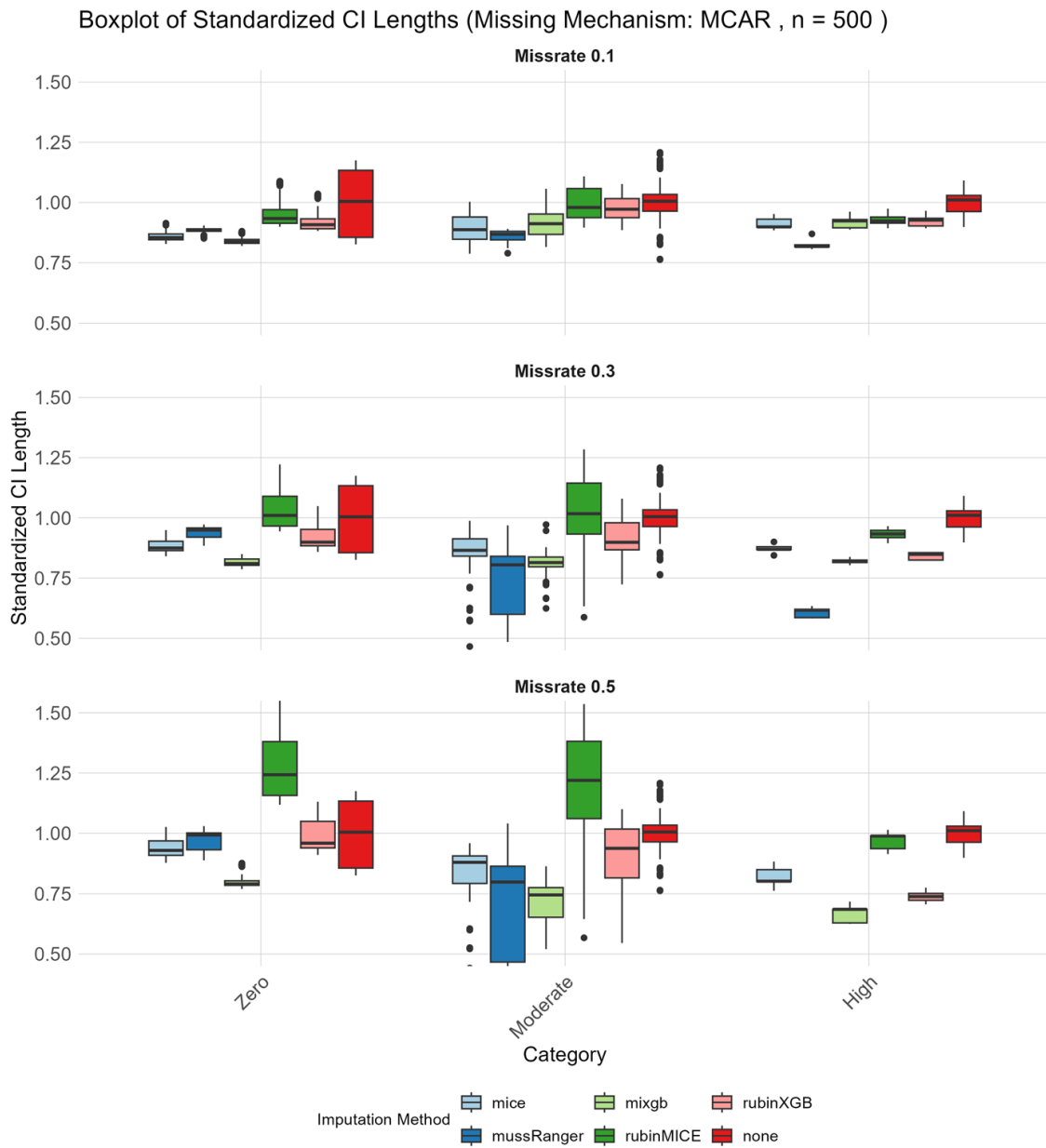


Fig. 13 Boxplot-Matrix of standardized CI lengths for MCAR, stratified by Missrate (0.1, 0.3, 0.5), sample size 500 and category (Zero, Below 90%, Above 90%)

Appendix D Explanatory Plots

See Figs. 14, 15, 16 and 17.

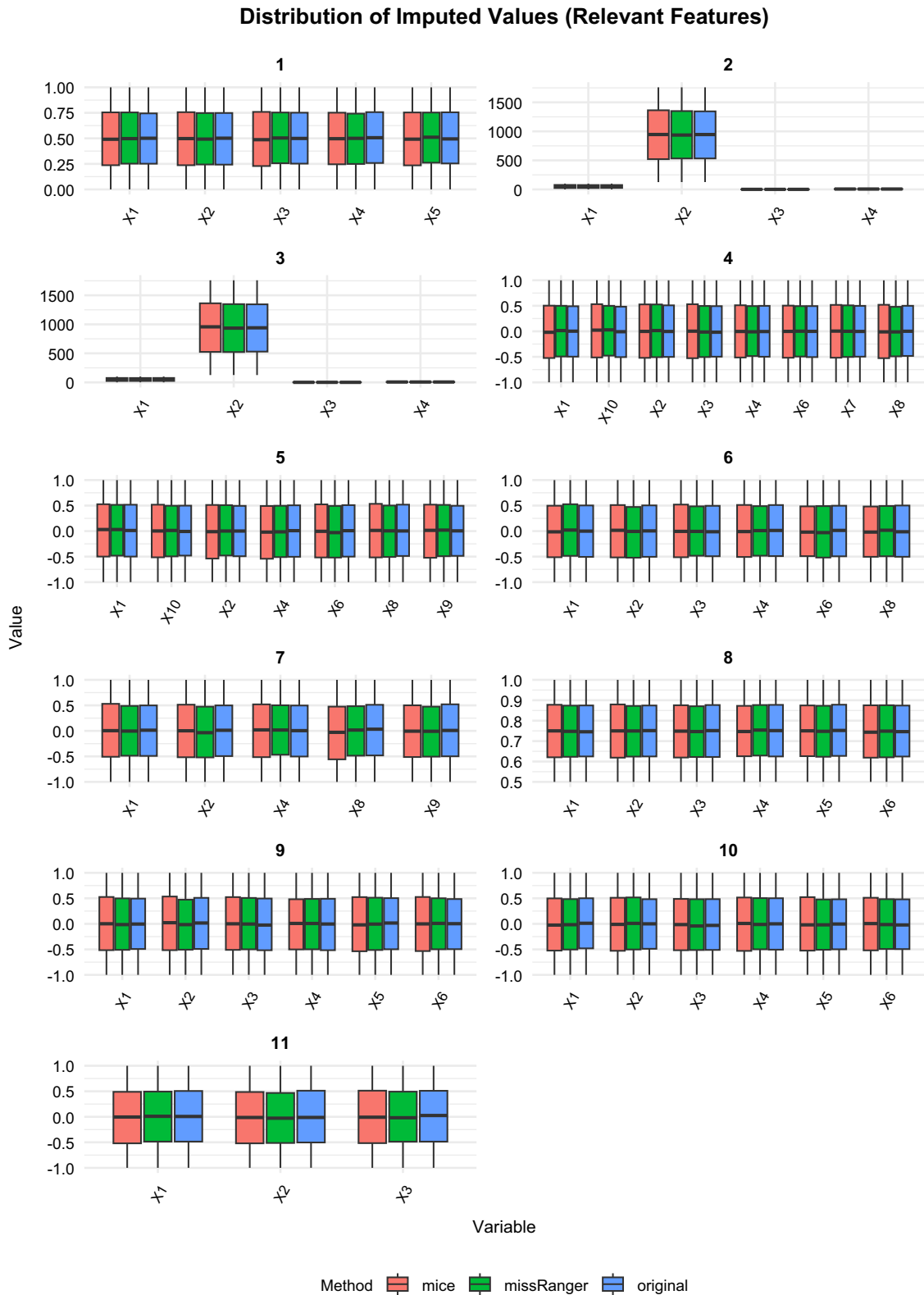
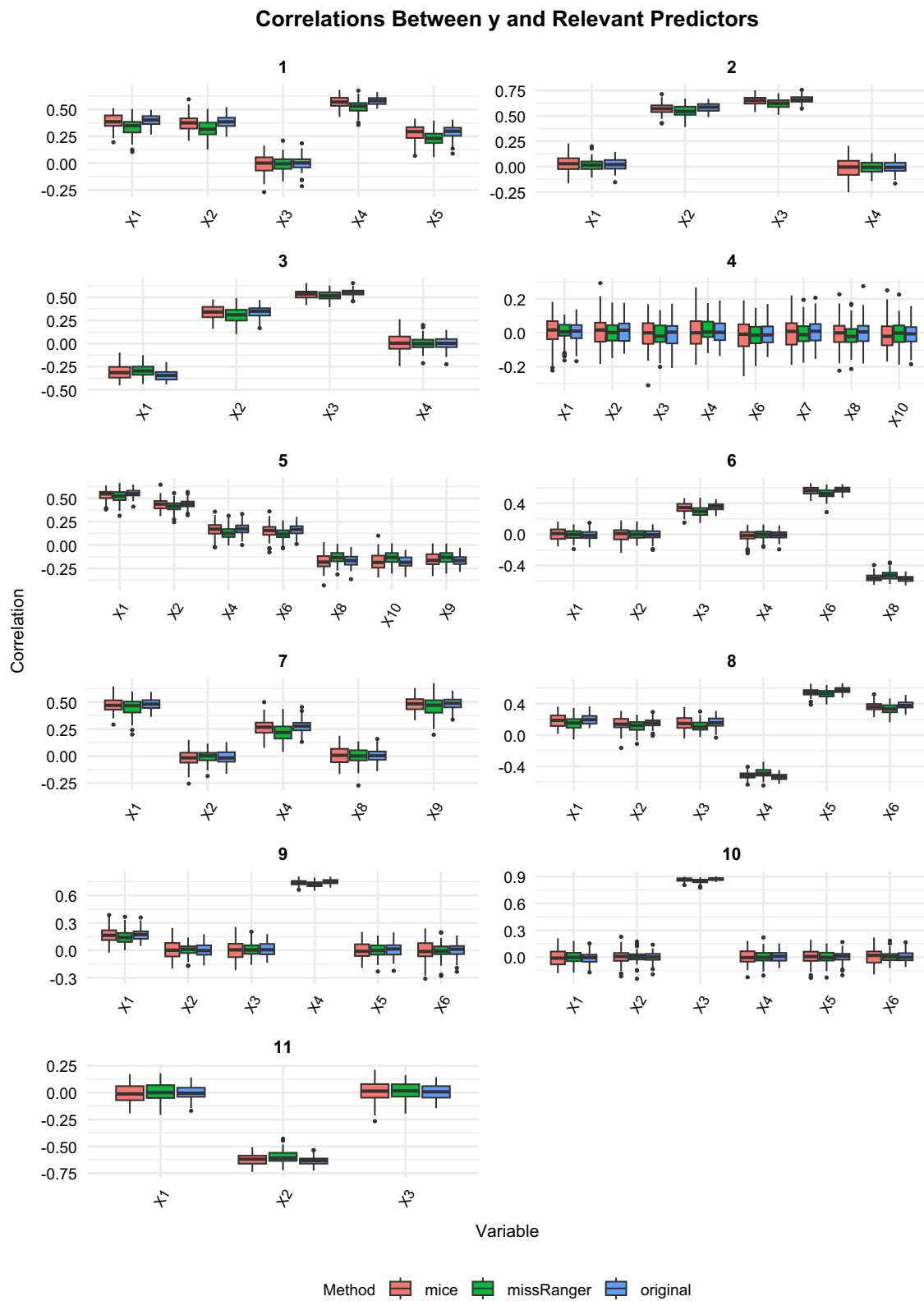


Fig. 14 Distribution of imputed values (relevant features)

**Fig. 15** Correlations between y and relevant predictors

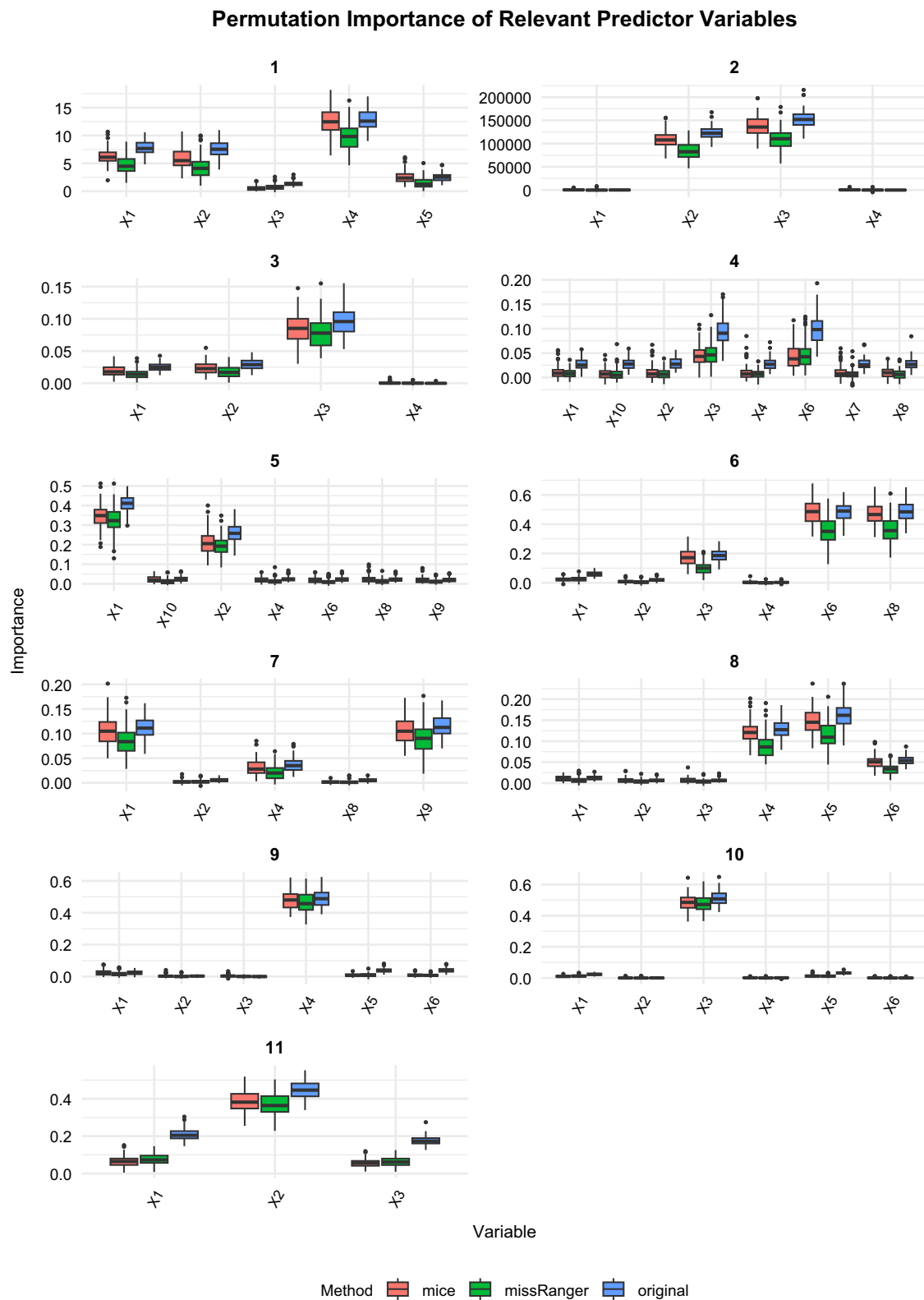
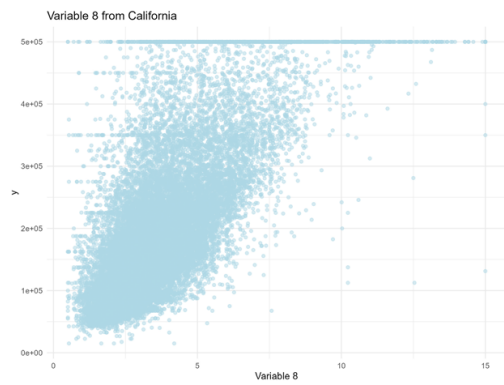
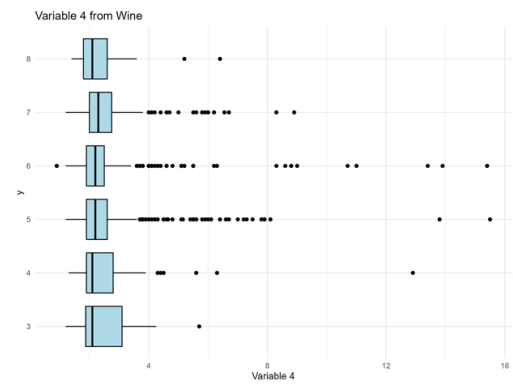


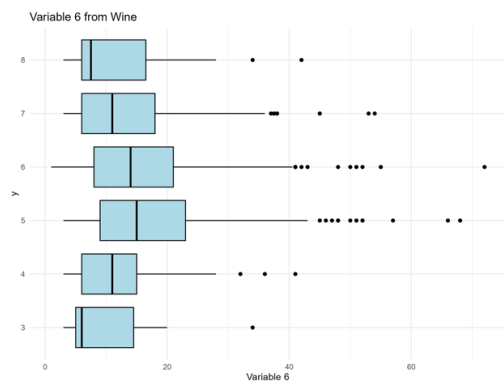
Fig. 16 Permutation importance of relevant predictor variables



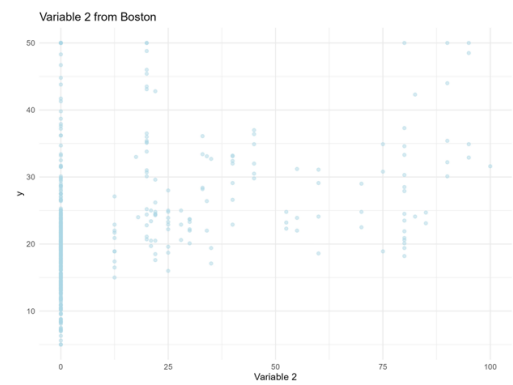
(a) California: Variable 8 vs. House Value



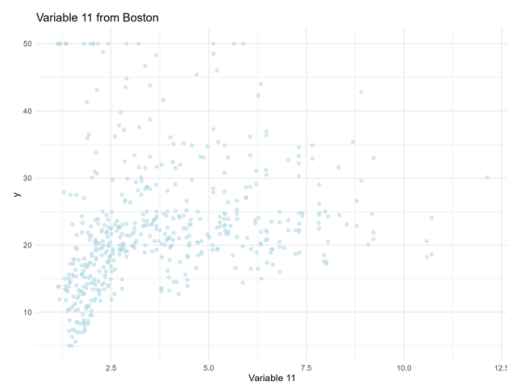
(b) Wine: Variable 4 vs. Class



(c) Wine: Variable 6 vs. Class



(d) Boston: Variable 2 vs. MEDV



(e) Boston: Variable 11 vs. MEDV

Fig. 17 Vergleich verschiedener Variablen in den Datensätzen “California Housing”, Wine und Boston Housing

Acknowledgements This work has been supported by the Research Center Trustworthy Data Science and Security (<https://rc-trust.ai>), one of the Research Alliance Centers within the <https://uaruhr.de>. The first author was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 314838170, GRK 2297 MathCoRe. The simulations were done on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest There is no conflict of interest of any kind.

Other During the preparation of this work, the authors used generative AI to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Buczak P, Chen J-J, Pauly M (2023) Analyzing the effect of imputation on classification performance under mcar and mar missing mechanisms. *Entropy* 25(3):521
- Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth and Brooks/Cole Monterey, CA, USA, ???
- Blum R, Hiabu M, Mammen E, Meyer JT (2024) Consistency of random forest type algorithms under a probabilistic impurity decrease condition. [arxiv:2309.01460](https://arxiv.org/abs/2309.01460)
- Blum R, Hiabu M, Mammen E, Meyer JT (2025) Pure interaction effects unseen by random forests. *Comput Stat & Data Anal* 212:108237. <https://doi.org/10.1016/j.csda.2025.108237>
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Cortez P, Cerdeira A, Almeida F, Mattos T, Reis J (2009) Wine quality dataset. UCI Machine Learning Repository. Available at: <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>. <https://doi.org/10.24432/C56S3T>
- Deng Y (2023) Mixgb: multiple imputation through 'XGBoost'. R package version 1.0.2. <https://CRAN.R-project.org/package=mixgb>
- Deng Y, Lumley T (2024) Multiple imputation through xgboost. *J Comput Graph Stat* 33(2):352–363
- Enders CK (2010) Applied missing data analysis. Guilford Press, New York, NY
- Föge N, Schmid L, Ditzhaus M, Pauly M (2024) A central limit theorem for the permutation importance measure. *arXiv preprint* [arXiv:2412.13020](https://arxiv.org/abs/2412.13020)
- Fanaee-T H, Gama J (2013) Bike sharing dataset. UCI Machine Learning Repository. Available at: <https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>. <https://doi.org/10.24432/C5W894>
- Gregorutti B, Michel B, Saint-Pierre P (2017) Correlation and variable importance in random forests. *Stat Comput* 27:659–678

- Harrison jr D, Rubinfeld DL (1978) Boston housing dataset. UCI Machine Learning Repository. Available at: <https://archive.ics.uci.edu/ml/datasets/housing>. <https://doi.org/10.24432/C5437P>
- Ishwaran H, Kogalur UB (2007) Random survival forests for r. *R News* 7(2):25–31
- Ishwaran H, Kogalur UB (2025) Fast unified random forests for survival, regression, and classification (rf-src). R package version 3.3.3
- Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. *Annals Appl Stat* 2(3):841–860
- Ishwaran H, Lu M (2019) Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Stat Med* 38(4):558–582
- Little RJ (1988) Missing-data adjustments in large surveys. *J Bus & Econ Stat* 6(3):287–296
- Little RJA, Rubin DB (2019) Statistical analysis with missing data. John Wiley & Sons, Hoboken, NJ
- Mayer M (2024) MissRanger: fast imputation of missing values. R package version 2.6.0. <https://CRAN.R-project.org/package=missRanger>
- Meng X-L (1994) Multiple-imputation inferences with uncongenial sources of input. *Stat Sci* 9(4):538–558
- Mentch L, Hooker G (2016) Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J Mach Learn Res* 17(26):1–41
- Morris TP, White IR, Royston P (2014) Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol* 14(1):75
- Pace RK, Barry R (1997) California housing dataset. UCI Machine Learning Repository. Available at: <http://archive.ics.uci.edu/ml/datasets/California+Housing>. <https://doi.org/10.24432/C5NS39>
- Probst P, Boulesteix A (2018) To tune or not to tune the number of trees in random forest. *J Mach Learn Res* 18(181):1–18
- Pepinsky TB (2018) A note on listwise deletion versus multiple imputation. *Polit Anal* 26(4):480–488
- Probst P, Wright MN, Boulesteix A-L (2019) Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip Rev Data Mining Knowl Discov* 9(3):1301
- R Core Team: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2023). R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ramosaj B, Amro L, Pauly M (2020) A cautionary tale on using imputation methods for inference in matched-pairs design. *Bioinformatics* 36(10):3099–3106
- Reiter JP (2007) Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika* 94(1):87–99
- Ramosaj B, Pauly M (2019) Predicting missing values: a comparative study on non-parametric approaches for imputation. *Comput Stat* 34(4):1741–1764
- Ramosaj B, Pauly M (2023) Consistent and unbiased variable selection under independent features using random forest permutation importance. *Bernoulli* 29(3):2101–2118
- Romero A, Schwerter J, Dumpert F, Pauly M (2025) Which imputation fits which feature selection method? a survey-based simulation study. *Data Sci Sci* 4(1):2562209
- Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York, NY
- Rubin DB (1987) Multiple imputation for nonresponse in surveys. John Wiley & Sons, New York, NY
- Rubin DB (2018) Multiple imputation. In: *Flexible Imputation of Missing Data*, Second Edition, pp. 29–62. Chapman & Hall/CRC, Boca Raton, FL
- Stekhoven DJ, Bühlmann P (2011) Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 112–118
- Scornet E, Biau G, Vert J-P (2015) Consistency of random forests. *Ann Stat* 43(4):1716–1741
- Schafer JL (1997) Analysis of incomplete multivariate data. CRC Press, Boca Raton, FL
- Schwerter J, Gurtškaia K, Romero A, Zeyer-Gliozzo B, Pauly M (2024) Evaluating tree-based imputation methods as an alternative to MICE PMM for drawing inference in empirical studies. [arXiv:2401.09602](https://arxiv.org/abs/2401.09602)
- Shao J (1991) Consistency of jackknife variance estimators jun. *Statistics* 22(1):49–57. <https://doi.org/10.1080/02331889108802282>
- Stekhoven DJ (2012) Using the missforest package. missforest nonparametric missing value imput. using random for. R Packag. version 1, 1–11
- Thurrow M, Dumpert F, Ramosaj B, Pauly M (2021) Imputing missings in official statistics for general tasks-our vote for distributional accuracy. *Stat J IAOS* 37(4):1379–1390
- Thurrow M, Dumpert F, Ramosaj B, Pauly M (2024) Assessing the multivariate distributional accuracy of common imputation methods. *Stat J IAOS* 40(1):99–108

- Van Buuren S, Groothuis-Oudshoorn K (2011) Mice: multivariate imputation by chained equations in r. *J Stat Softw* 45:1–67
- Wager S, Hastie T, Efron B (2014) Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *J Mach Learn Res* 15(1):1625–1651
- Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, Marrero J, Zhu J, Higgins PD (2013) Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open* 3(8):002847
- White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 30(4):377–399
- Wright M, Ziegler A (2015) ranger: a fast implementation of random forests for high dimensional data in c++ and r. *J Stat Softw* 77(1):1–17
- Yeh I-C (1998) Concrete compressive strength dataset. UCI Machine Learning Repository. Available at: <https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength>. <https://doi.org/10.24432/C5PK67>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Article 3

Föge, N., Schwerter, J., Gurtskaia, K. et al. Adapting tree-based multiple imputation methods for multilevel data? A simulation study. *Behav Res* 58, 166 (2026).
<https://doi.org/10.3758/s13428-026-03008-x>



Adapting tree-based multiple imputation methods for multilevel data? A simulation study

Nico Föge^{1,4} · Jakob Schwerter^{2,3} · Ketevan Gurtskaia² · Markus Pauly^{2,4} · Philipp Doeblér²

Received: 23 March 2025 / Accepted: 18 March 2026
© The Author(s) 2026

Abstract

When data have a hierarchical structure, such as students nested within classrooms, ignoring dependencies between observations can compromise the validity of imputation procedures. Standard (tree-based) imputation methods implicitly assume independence between observations, limiting their applicability in multilevel data settings. Although multivariate imputation by chained equations (MICE) is widely used for hierarchical data, it has limitations, including sensitivity to model specification and computational complexity. Alternative tree-based approaches have shown promise for individual-level data, but remain largely unexplored for hierarchical contexts. In this simulation study, we systematically evaluate the performance of novel tree-based methods—chained random forests (missRanger) and extreme gradient boosting (mixgb)—explicitly adapted for multilevel data by incorporating dummy variables indicating cluster membership. We compare these tree-based methods and their adapted versions with traditional MICE imputation in terms of coefficient estimation bias, type I error rates, and statistical power under different cluster sizes (25 and 50), missingness mechanisms (missing completely at random [MCAR], missing at random [MAR]), and missingness rates (10%, 30%, 50%), using both random intercept and random slope data generation models. The results show that MICE provides robust and accurate inference for level 2 variables, especially at low missingness rates (10%). However, the adapted boosting approach (mixgb with cluster dummies) consistently outperforms other methods for level 1 variables at higher missingness rates (30%, 50%). For level 2 variables, while MICE retains better power at moderate missingness (30%), adapted boosting becomes superior at high missingness (50%), regardless of the missingness mechanism or cluster size. These findings highlight the potential of appropriately adapted tree-based imputation methods as effective alternatives to conventional MICE in multilevel data analyses.

Keywords Multiple imputation · Multilevel data · Bias · Hierarchical data · MICE · MissRanger · Mixgb · Power · Type I error

Introduction

Multiple imputation (MI) is a prominent and widely used technique for dealing with missing data. MI involves generating multiple plausible complete versions of the dataset to reflect the uncertainty about the missing values. Each dataset is then analyzed separately using standard statistical methods, and the results are pooled to obtain estimates and standard errors that account for both within- and between-imputation variability. While these steps are essential components of MI, “proper” imputations additionally require that uncertainty about the missing values is adequately propagated within the imputation model itself (Carpenter et al., 2023; Rubin, 1987, 1996). MI is particularly useful when the missingness mechanism is missing at random or missing completely at random (Little & Rubin, 2002; Schafer, 2000).

Nico Föge and Jakob Schwerter share first authorship.

✉ Nico Föge
nico.foege@ovgu.de

¹ Department of Mathematics, Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany

² Department of Statistics, TU Dortmund University, Dortmund, Germany

³ Center for Research on Education and School Development, TU Dortmund University, Dortmund, Germany

⁴ Research Center for Trustworthy Data Science and Security, UA Ruhr, Ruhr, Germany

Ensuring that imputation models preserve the underlying relationships in the data and account for the missing data mechanism is crucial (van Buuren & Groothuis-Oudshoorn, 2011). This is a particular challenge in complex hierarchical datasets with multiple levels. Examples include clustering at individual and higher-level units like classes or schools. Preserving key multilevel features—such as contextual effects of level 1 variables, varying effects across clusters, and level 2 predictors—is crucial for valid inference, as failure to maintain the hierarchical structure can lead to biased analyses; however, this is frequently neglected in imputation (van Buuren, 2011).

Although fully conditional specification (FCS) or multivariate imputation by chained equations (MICE) is a prevalent approach in the social sciences, it presents several limitations, including its considerable complexity due to the challenges in the specification of the imputation model and computational intensity (van Buuren & Groothuis-Oudshoorn, 2011). Since MICE relies heavily on model specifications, it can lead to issues like overfitting and convergence errors, especially when dealing with multicollinearity and other instability problems (van Buuren & Oudshoorn, 1999).

Exploring alternatives to MICE, such as nonparametric tree-based methods, offers the potential for handling missing data in hierarchical structures, particularly in settings where their flexibility may be advantageous (Deng & Lumley, 2024; Stekhoven & Bühlmann, 2012). However, their performance in multilevel contexts—especially under linear mixed model assumptions common in the social sciences—remains largely unexplored.

Therefore, this simulation study addresses the following *research question*: Do tree-based imputation methods exhibit similar performance in terms of bias, type I error, and power as the standard level 2 imputation method? The specific tree-based methods are chained random forests (Stekhoven & Bühlmann, 2012; Tang & Ishwaran, 2017) and extreme gradient boosting with trees as base learners (Deng, 2023). To account for the multilevel data structure, we use dummy variables for each cluster, adapting tree-based imputation methods in the same way that Lüdtke et al. (2017) did for non-tree-based methods.

Although many different MI methods have been proposed and investigated for this setting—including multilevel MI with joint modeling, multilevel MI with fully conditional specifications, multilevel substantive-model-compatible MI with sequential modeling, or model-based treatment with Bayesian estimation, as recently reviewed by Grund et al. (2024)—tree-based methods have not been evaluated for multilevel data.

Our simulation study evaluated tree-based methods in a controlled setting using linear mixed models with linear effects and parametric data-generating processes. After that, missingness was introduced, and the resulting datasets

were imputed using different methods. The imputed datasets were analyzed with a linear mixed model to evaluate type I error rates, power, standard deviation, and coefficient bias, thereby assessing the impact of imputation on subsequent multilevel analysis. This setup allows a direct comparison with MICE and deliberately focuses on linear multilevel scenarios. In particular, the data-generating process aligns with the assumptions of the linear mixed model used for analysis and does not include additional multilevel features such as contextual (between-cluster) effects of level 1 predictors, decomposed within- and between-cluster relations, or cross-level interactions to focus on a well-defined linear multilevel setting. While linear model-based imputation approaches are naturally well suited for settings with linear functional forms, tree-based methods are known to perform particularly well in the presence of nonlinearities and interaction effects, which can be more difficult to approximate using parametric models (Breiman, 2001; Chen & Guestrin, 2016). Accordingly, the present study focuses on linear functional forms to assess the performance of tree-based imputation methods in scenarios that align with standard linear multilevel analyses.

Readers should therefore interpret our findings with these limitations in mind: our results are most relevant to linear multilevel scenarios, while the advantages of tree-based approaches may become more apparent in nonlinear or interaction-heavy contexts that were not included in our study.

The remainder of the paper is organized as follows: In the next section, we briefly review multilevel data and existing (parametric) MI approaches for multilevel data structures. Next, the simulation study is detailed, starting with a review of the imputation approaches and an explanation of the simulated missingness mechanisms. We present simulation results in the following section with the help of various plots for the relative performance, and finally, we close with remarks on the current study and potential future work.

Multilevel data

Multilevel data structures are common in social and behavioral sciences research (Hox & Roberts, 2010). This is often seen, for example, in educational research where students (level 1) are nested within classes, schools, or regions (level 2). Variation of higher-level variables can significantly influence the outcome variable. This calls for robust analysis methods that account for the complexity introduced by these hierarchical structures (Grund et al., 2024; Steenbergen & Jones, 2002), since simply ignoring the dependencies or aggregating everything to a single level can be deceptive (Aitkin & Longford, 1986; Grund et al., 2024). To maintain statistical integrity, it is crucial to employ appropriate imputation methods tailored for multilevel data. These methods should consider both within-cluster and between-cluster variability for a more accurate representation of the

underlying data (Audigier et al., 2018; Grund et al., 2024; van Buuren, 2011).

Multilevel data often involve within- and between-cluster effects, where level 1 variables (e.g., student test scores) may have distinct effects at the individual level within clusters and contextual effects at the cluster level (McNeish et al., 2016). Additionally, random slopes allow the effects of level 1 variables to vary across clusters, capturing heterogeneity in relationships. Level 2 variables, such as cluster characteristics (e.g., class size or teacher experience), further influence outcomes and must be accurately modeled to preserve hierarchical relationships. Ignoring these features and the nested structure in imputation can lead to biased parameter estimates and incorrect variance components, particularly in multilevel MI, where preserving these relationships is essential for valid inference (Aitkin & Longford, 1986; McNeish et al., 2016; van Buuren, 2011). Hierarchical linear models, including random intercept and random slope models, provide a sophisticated statistical framework for analyzing such data structures, ensuring that both fixed and random effects are appropriately captured (Hox, 1998; Hox & Roberts, 2010).

Existing imputation approaches for multilevel data

For a valid analysis, the imputation model must account for the dependency between observations inherent in the multilevel structure. Otherwise, inferences might be biased even when the statistical methods are otherwise appropriate (Audigier et al., 2018; Hox & Roberts, 2010). The bias of random-effect variance estimates as well as global fixed-effects confidence intervals depends on the cluster size, the relation of within- and between-cluster variance, and the missing data mechanism (Speidel, Drechsler, & Sakshaug, 2018). Two common imputation frameworks for multilevel data are joint modeling (JM) and fully conditional specification (FCS). In JM, imputations are generated simultaneously from a joint multivariate distribution for all variables (Hox & Roberts, 2010). In contrast, FCS imputes missing values iteratively on a variable-by-variable basis, using univariate conditional models based on other variables (van Buuren, 2011). MICE uses the conditional imputation approach from FCS and chains the variables by iteratively imputing one variable at a time based on the others (van Buuren & Groothuis-Oudshoorn, 2011). In this study, we use multivariate imputation by chained equations (MICE), implemented via the mice package, as a specific application of FCS, employing 2l.norm for level 1 variables and 2lonly.pmm for level 2 variables (van Buuren & Groothuis-Oudshoorn, 2011). This terminology reflects MICE as our chosen FCS implementation, with no aspects applying exclusively to FCS or MICE. A study from Enders et al. (2016) compared two imputation frameworks for multilevel data: JM and chained equation

imputation (MICE with a two-level normal model 2l.norm). The joint model turned out to be better for analyses postulating distinct within- and between-cluster relations, while chained equations imputation turned out to be superior in random slope analysis (Enders et al., 2016).

For clustered data, both of these methods are effective in the broad context of *random intercept* models, even with variables at higher levels, as simulation results by Grund et al. (2018b) indicate. However, for *random slopes*, FCS appears to be more flexible than JM, but still has some limitations and is not all that reliable when data in an explanatory variable are missing (Grund et al., 2018b). Similar results were shown by Enders et al. (2016) for *random intercepts*. Additionally, Grund et al. (2018a) show that both approaches (FCS and JM) provide useful tools for dealing with missing data at level 2 in most applications in practice (especially for balanced data).

Recent studies highlight limitations of conventional multilevel MI approaches, such as FCS and JM, particularly for random-slope models. When explanatory variables have missing values, FCS can produce biased random slope estimates due to misspecification of conditional models (Erler et al., 2017; Erler et al., 2016). Model-based methods, such as substantive-model-compatible MI and Bayesian approaches, align imputation and analysis models more closely, improving accuracy for random slopes and complex hierarchical structures (Enders, 2023; Enders et al., 2020; Grund et al., 2021). These methods better capture within- and between-cluster effects in the presence of random slopes.

One advantage of choosing FCS over JM is that FCS grants more flexibility in creating multilevel models by splitting a k -dimensional problem into k one-dimensional problems. That is, for each of the variables with missing data, a regression model with a univariate outcome conditional on the other $k - 1$ variables is constructed. Furthermore, it is easier to avoid logical inconsistencies in the imputed data and incorporate methods to preserve unique features in the data (van Buuren et al., 2006); for example, temporal dependency can be taken into account in longitudinal data.

Multiple imputation methods

In this simulation study, we implemented three main imputation methods: As baseline, we use multivariate imputation by chained equations (MICE) via the mice function from the R package of the same name (van Buuren & Groothuis-Oudshoorn, 2011). For level 1 variables, the level 1 normal model (2l.norm) is used, and for level 2 variables the level 2 class predictive mean matching (2lonly.pmm) is employed. Additionally, we include recent tree-based imputation methods: a fast implementation of imputation via random forests from the missRanger package (Mayer, 2021), and MI via gradient boosting implemented in the mixgb package (Deng,

2023). For MICE, the imputation methods and the prediction matrix for each variable are carefully adjusted to take into account the multilevel data structure. Two factors were varied for missRanger: We run the missRanger algorithm both with and without predictive mean matching (using 5 donors)¹ and the variant (standard or an adapted implementation of missRanger). The adapted variant respects the multilevel structure of the data by including dummy variables for the respective cluster, while the standard implementation does not. In line with Drechsler (2015), who shows that multilevel models at the imputation stage can provide superior inference—especially for random effects—we acknowledge that running such models can be computationally demanding and that researchers often resort to simpler, dummy-based approaches. Indeed, representing cluster membership via dummies biases random-effects estimates more than fixed-effects estimates, yet remains popular because it is straightforward to implement with many conventional imputation tools. As we pioneer the adaption of tree-based methods to hierarchical data, we follow Drechsler's rationale and incorporate dummy variables for clusters in our adjusted version of missRanger. While this does not capture the complete random-effects structure, it delivers a manageable, first-tier solution for incorporating level 2 distinctions without the overhead of fully specified multilevel modeling. Similarly, mixgb was adapted with additional dummy variables. We also used five imputations.² For the pooling, Rubin's rules were applied. Details of each method are given in the supplement.

Multivariate imputation by chained equations

Multivariate imputation by chained equations (MICE) is a flexible and efficient imputation method that can treat missingness in a wide range of data types and analysis models. Its applications can be found in diverse research fields, including medicine, epidemiology, psychology, management, politics, and sociology (van Buuren & Groothuis-Oudshoorn, 2011).

MICE operates by iteratively drawing missing values from univariate conditional distributions given the other observed features. By repeatedly updating these imputations over several iterations, the method approximates the underlying conditional

distribution, not unlike the burn-in phase of a Markov chain. The goal is not to produce increasingly accurate imputations but to generate “proper” imputations that reflect the uncertainty about the missing values and the parameters of the underlying model (Carpenter et al., 2023; Rubin, 1987). Once the iterative process stabilizes, multiple imputations are generated. The variability between imputations arises from independent draws from this stable conditional distribution (Kistner et al., 2024; van Buuren & Groothuis-Oudshoorn, 2011).

Imputing at level 2 requires additional considerations. One possibility is to have a separate model for level 2 variables, for example, a regression model that includes other level 2 variables as well as cluster-level components (e.g., mean, median) of the level 1 variables. As described by Grund et al. (2018a), imputations at level 2 are generated based on a probability distribution that considers multiple factors: aggregated information from level 1 (such as group means of relevant variables), the observed values of the other variables within the cluster, and specific model parameters.

As mentioned, mice is an R package for chained equation imputations using a function of the same name. To impute level 1 missing values, `2l.norm`³ is used, which uses univariate missing data imputation with a two-level normal model. For level 2 variables, a level 2 class predictive mean matching (`2lonly.pmm`) is applied. Although Enders et al. (2016) used `2lonly.norm`, we use `2lonly.pmm` because `2lonly.norm` follows the normality assumption, which does not hold in our case because we transformed some variables to aggregate them at level 2. `2lonly.pmm` is a semi-parametric method, which is why it works in a wider range of cases.

Chained random forest imputation

The idea of the random forest is to combine predictions from multiple decision trees to improve accuracy and reduce overfitting in both classification and regression tasks. In a random forest, each tree independently makes a prediction, and the final result is based on the majority vote (in classification) or the average prediction (in regression) of all the trees. Random forest can handle mixed types of data, is capable of addressing interactions and nonlinearity, and is robust to overfitting (Breiman, 2001). Because of its flexibility, random forest can also be adapted to imputation tasks (Golino & Gomes, 2016; Stekhoven & Bühlmann, 2012), transferring useful properties such as the lack of assumptions about normality or homoscedasticity.

As for MICE, the chained random forest imputation also starts with an initial guess that replaces all missing values in

¹ The default value for missRanger is 0, and the default value in MICE is 5. Not using PMM for missRanger showed inflated type I errors in another context (Ramosaj et al., 2020).

² To choose a number of imputations, a mini-simulation on randomly selected design was run with 5 and 20 imputations with MICE and missRanger (standard and with dummies with three donors). The results suggested very little to no differences between the number of imputations in terms of coefficient estimation bias and H_0 rejection rates, but imputing 20 times takes approximately four times longer than performing 5 imputations. Consequently, the whole simulation is run with 5 imputations for each method, making the simulation considerably faster.

³ 2l indicates that the data structure is multilevel; it does not mean the calculation happens at a higher level, but it considers that there are variables at another level.

the data. The initial guess can be made by mean (default for metric data), median, mode (default for categorical data), or any other univariate imputation method. At this point, every variable in the dataset is complete, although the imputations are just placeholders. Next, the columns in the dataset are ranked according to the proportion of missing data, starting with the column containing the fewest missing values. For each column, the algorithm builds a predictive model using random forests, where the observed (non-missing) values in that column serve as the outcome variable, and the values from other columns for the same rows are used as predictors. Once predictions are generated, the initial guesses are replaced by these predictions. After cycling through all variables with missing values (one full iteration), the algorithm repeats the process several times (Stekhoven & Bühlmann, 2012), with a default of 10 iterations. During each iteration, the imputed values from the previous step are updated using the predictions from the random forest models trained in the current iteration. As the iterations proceed, the imputations are refined, as the predictions are based on increasingly accurate imputed values for the other variables. The algorithm stops after a predefined number of iterations.

A faster and extended version of the algorithm is given by the *missRanger* package of Mayer (2021). It uses the run-time efficient *ranger* implementation of random forest and additionally extends *missForest* by offering the option of *predictive mean matching* (PMM; Mayer, 2021).

PMM enhances random forest imputation by ensuring that imputed values are more realistic. Instead of directly using predicted values from the random forest, PMM matches each missing entry to one of the k closest observed values, the so-called donors. The actual observed value from this match is then used for imputation, maintaining the variability of the original data. PMM is particularly useful when it is important to faithfully reflect the underlying data distribution and impute values that are plausible within the observed range (Mayer, 2021). The number of donors is denoted by k . The default value is $k = 5$, which we also use in our study.

MI through extreme gradient boosting (XGBoost)

Extreme gradient boosting (XGBoost) is a machine learning algorithm that belongs to the family of gradient boosting methods (Chen & Guestrin, 2016). Instead of simply averaging multiple decision trees as in random forests, XGBoost uses gradient boosting to combine multiple regression trees. It also employs regularization and shrinkage techniques. Typically, the depth of the regression trees used in sequential boosting are not very deep. One advantage of XGBoost over random forest is that XGBoost selects only one of a set of highly correlated features, because in sequential boosting, features are added to the model incrementally. If a feature

is selected early, other highly correlated features can only improve the prediction if they provide new information.

XGBoost uses gradient descent optimization techniques to iteratively minimize the loss function, making it highly efficient and effective in finding the optimal solution. L1 and L2 regularization prevent overfitting and improve generalization. Tree pruning removes unnecessary branches and reduces model complexity, further enhancing XGBoost's predictive performance.

Lastly, XGBoost can take advantage of parallel processing capabilities, making it suitable for large datasets and reducing training time. Overall, XGBoost is known for its ability to handle complex datasets and provide accurate predictions. Its popularity is evident in various domains, including Kaggle competitions and real-world applications (Chen & Guestrin, 2016).

The *mixgb* R package uses XGBoost to implement missing value imputation in a scalable and efficient manner (Deng & Lumley, 2024). It addresses the challenge of missing data in large datasets with complex structures with an iterative or non-iterative strategy parallel to chained random forest imputation. Initial values for imputation are filled with random values drawn from the observed data to initialize the imputation process (Suh & Song, 2023). *mixgb* imputes variables with fewer missing values first. This strategic approach aims to prioritize variables with more available information early in the imputation process. *mixgb* provides a versatile approach to missing data imputation, leveraging XGBoost, subsampling, and PMM to enhance imputation accuracy, especially for continuous data. The default imputation in *mixgb* is non-iterative but the package allows users to set the number of iterations for imputation (Deng & Lumley, 2024). For this simulation we used five iterations to be in line with the other imputation methods.

Adapting tree-based imputation methods with cluster membership dummy variable

By default, *missRanger* and *mixgb* do not account for the multilevel structure of hierarchical data. To address this, we adapted these tree-based methods by incorporating dummy variables for cluster membership (denoted *dummies* in the following), following the approach of Lüdtke et al. (2017). For J clusters, J dummy variables were added, with the j th variable equal to 1 if the observation belongs to cluster j and 0 otherwise. Unlike regression-based methods, tree-based methods do not require a reference group, as their decision trees can model cluster-specific effects without explicit parameterization (Breiman, 2001; Chen & Guestrin, 2016). Cluster-specific predictions in *missRanger* and *mixgb*, such as those using level 2 dummy variables or cluster means, are

appropriate in our simulation, which uses the entire dataset to approximate the complete data distribution. However, in analyses with a training/test split, such predictions risk data leakage if test set information (e.g., cluster-level predictors) is used in the imputation of training data, and strict separation of training and test sets is required. Out-of-sample predictions for previously unobserved clusters are not naturally supported by this approach, as no cluster-specific effects can be learned or extrapolated beyond the observed clusters.

Multilevel structure The dummy variables enable *missRanger* and *mixgb* to capture between-cluster effects by modeling cluster-specific differences in predictor values (in our case means of level 2 variables). Within-cluster effects are addressed by including level 1 variables as predictors in the imputation formula, which account for individual-level relationships. The flexibility of random forests (*missRanger*) and gradient boosting (*mixgb*) allows these methods, in principle, to capture nonlinear effects and interactions between predictors and dummy variables, potentially approximating variations akin to random slopes in more complex settings. However, unlike MICE, which explicitly models random intercepts and slopes using 2l.norm and a tailored predictor matrix (Grund et al., 2018a), tree-based methods do not directly estimate random effects but rely on algorithmic approximations, which differ conceptually from parametric multilevel specifications. The multilevel structures considered in this study are limited by the data-generating process and the analysis model. In particular, level 1 predictors are not decomposed into within- and between-cluster components, such that contextual effects of level 1 variables are not included. Cross-level interactions and other more complex multilevel dependency structures are likewise not part of the simulation design.

Level 2 variables Tree-based methods, such as *missRanger* and *mixgb*, and MICE handle missing values in multilevel data differently, particularly for level 2 variables. For *missRanger* and *mixgb*, level 2 variables are imputed at the individual level using tree-based prediction models that include level 1 variables and cluster membership indicators as predictors (Deng, 2023; Mayer, 2021). As a result, cluster-level information enters the imputation process indirectly through the predictor set rather than through an explicit multilevel imputation model. In contrast, MICE's 2lonly.pmm method imputes level 2 variables directly at the cluster level, generating a single imputed value per cluster and thereby explicitly enforcing within-cluster consistency and the hierarchical structure of the data (Grund et al., 2018a).

If regression trees have sufficient depth, the flexibility gained by the dummy variable approach can, in principle, exceed that of procedures that include only random intercepts, since some interactions with dummy variables are

possible, although not all kinds of interactions (Wright et al., 2016). Potentially, more dummy variables are added than there are variables originally, so the dummy variable procedure relies on imputation methods that have some form of regularization. Tree-based ensemble methods like random forest and XGBoost do have this property by design (Breiman, 2001; Chen & Guestrin, 2016).

The tree-based methods *missRanger* and *mixgb* incorporate stochastic elements to introduce variability across imputations and thereby approximate aspects of “proper” imputations in the sense of Rubin (1996). Specifically, *missRanger* relies on random forests with bootstrapping and random feature selection, while *mixgb* introduces randomness through stochastic gradient boosting and subsampling. These mechanisms induce variation between imputations and partially propagate uncertainty related to both the missing values and the fitted imputation models. However, unlike fully parametric or Bayesian MI approaches, the tree-based procedures considered here do not generate imputations via explicit draws from a posterior predictive distribution. As a result, uncertainty propagation is achieved heuristically rather than through a formally specified probabilistic model.

Simulation study design

Simulated data

For the simulation setup, four general factors are varied, resulting in a total of 24 simulation designs. The varying factors are the *number of clusters*, *data generation model*, *missing rate*, and *missing mechanism*. We consider two numbers of clusters: 25 and 50. These choices reflect typical applied scenarios, with 25 clusters representing a small-sample setting (<30 clusters) and 50 clusters approximating asymptotic conditions (Cameron & Miller, 2015). This distinction matters practically, as inference methods perform differently depending on whether the number of clusters is small or large. This results in two cluster sizes of 40 and 20 for a balanced design with a sample size of $N = 1,000$. This variation in the number of clusters allows us to examine the effect of cluster size on the performance of the imputation methods. This fixed sample size limits the design to two combinations (25 clusters $\times$ 40 units, 50 clusters $\times$ 20 units), rather than a fully crossed design. This choice balances computational feasibility with scenarios common in applied multilevel research. The data generation models included a random intercept model and a random intercept and slope model, capturing varying degrees of multilevel complexity. Missingness rates were set to 10%, 30%, and 50% to cover a range of practical scenarios, and missingness mechanisms

were restricted to missing completely at random (MCAR) and missing at random (MAR), following standard practices (Little & Rubin, 2002). We evaluated each condition using bias, standard deviation, type I error rates, and power.

Data generation

To assess the impact of these factors, we simulate data in a controlled manner.

Specifically, we generate data by using the fungible package in R (Waller, 2022). The monte function generates clustered data with predefined characteristics. We generate 12 continuous predictor variables: six level 1 variables (X_{ij1} to X_{ij6}) and six level 2 variables (L_{ij1} to L_{ij6}). The level 1 variables were drawn from a multivariate normal distribution, with correlations and intraclass correlation coefficients (ICCs) specified by $\eta^2 \sim \text{Uniform}(0, 0.85)$ to reflect varying clustering strength typical in social science research (Snijders & Bosker, 2012). Level 2 variables were computed as the cluster means of the corresponding variables generated by fungible::monte. This process ensures that level 2 variables are correlated with level 1 variables, facilitating imputation by capturing hierarchical relationships (van Buuren, 2011).

Finally, two data generation models are considered for the outcome variable: *random intercept* and *random intercept and random slope* (hereafter denoted as *random slope* model). The data-generating process assumes a linear multilevel model, aligning with the assumptions of the linear mixed model used for analysis and the MICE imputation method (2l.norm), which may favor MICE over other methods like missRanger or mixgb that do not explicitly assume linearity (van Buuren, 2018). Based on the model, the output variable is entered as follows.

- Random intercept model:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij4} + \beta_5 X_{ij5} + \beta_6 X_{ij6} + \beta_7 L_{ij1} + \beta_8 L_{ij2} + \beta_9 L_{ij3} + \beta_{10} L_{ij4} + \beta_{11} L_{ij5} + \beta_{12} L_{ij6} + \delta_{0j} + \epsilon_{ij},$$

where $\delta_{0j} \sim N(0, 10)$, $\epsilon_{ij} \sim (0, 1)$, and fixed coefficients are $\beta = (0.3, 0.5, 1, 1.5, 0, 0, 0.5, 1, 1.5, 0, 0, 0)$.

- Random intercept and slope model: where δ_{0j} , δ_{1j} , $\delta_{2j} \sim N(0, 1)$, $\epsilon_{ij} \sim N(0, 10)$, and the same fixed coefficients apply.

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij4} + \beta_5 X_{ij5} + \beta_6 X_{ij6} + \beta_7 L_{ij1} + \beta_8 L_{ij2} + \beta_9 L_{ij3} + \beta_{10} L_{ij4} + \beta_{11} L_{ij5} + \beta_{12} L_{ij6} + \delta_{0j} + \delta_{0j} X_{ij1} + \delta_{2j} X_{ij2} + \epsilon_{ij},$$

In the case of the random slope model, δ_{0j} , δ_{1j} , δ_{2j} , and ϵ_{ij} are randomly generated. The six noise features X_{ij4} , X_{ij5} , X_{ij6} , L_{ij4} , L_{ij5} , and L_{ij6} are uncorrelated with the output variable

Y_{ij} . They are included to enable evaluation of the type I error and to assess the ability of the imputation methods to distinguish between important and unimportant variables. The variance of the error terms was set to 10 to achieve a realistic marginal R^2 . We have a marginal R^2 of 0.484 for the random intercept model and 0.442 for the random intercept and random slope model. The fixed coefficients ($\beta_1 = 0.5$, $\beta_2 = 1$, $\beta_3 = 1.5$, $\beta_7 = 0.5$, $\beta_8 = 1$, $\beta_9 = 1.5$, others zero) were chosen to reflect small to moderate effect sizes typical in social science multilevel studies according to Maas and Hox (2005). Our values, with standardized coefficients ranging from approximately 0.10 to 0.33, provide a realistic testbed for evaluating the ability of imputation methods to recover subtle effects in hierarchical data, where clustering often reduces effect magnitudes. The variance components for random effects were set to 1 to standardize their contributions and simplify comparisons across conditions, though this may not reflect the heterogeneity of real-world data (Snijders & Bosker, 2012).

Missing values were introduced at both level 1 and level 2. After generating the data, missing values were introduced in the level 2 predictors ($L_{ij1}, \dots, L_{ij6}$) at rates under MCAR and MAR mechanisms at the cluster level. These missing values were imputed using five methods: MICE, missRanger (with and without dummies) and mixgb (with and without dummies). The imputed datasets were analyzed using the linear mixed model with random intercepts, matching the data-generating model. We evaluated the performance of the imputation methods by computing the type I error rate (the proportion of significant results for predictors with true coefficients $\beta_4 = \beta_5 = \beta_6 = \beta_{10} = \beta_{11} = \beta_{12} = 0$ at $\alpha = 0.05$), power, coefficient bias and standard deviation of the parameter estimation. These metrics assess the impact of imputation on the subsequent multilevel analysis, rather than directly measuring the accuracy of imputed values. This choice was made because practitioners are usually more interested in inference multilevel analysis.

Missingness mechanism

Missingness in the generated data is induced at three levels from moderately low to relatively high (10%, 30%, and 50%). This range of missingness levels reflects real-world scenarios. Two missingness mechanisms (MAR, MCAR) are considered. The introduction of missing data is based on the selected missingness mechanism and the specified missing rate. For MCAR, a simple function is implemented in R that randomly sets each data point to NA with probability equal to the missingness level for each variable. For MAR, missingness in level 1 variables ($X_{ij1}, \dots, X_{ij6}$, Y_{ij}) was implemented using a custom function based on Thurow, Dumpert, Ramosaj, and Pauly (2021a, b). The algorithm operates as follows: Starting with one variable, it generates

missing values under MCAR with the overall missing rate. The values of the selected variable were binned into 25 equal-width intervals based on the variable's range, and missingness in other level 1 variables ($X_{ij2}, \dots, X_{ij6}, Y_{ij}$) was assigned with probabilities proportional to the frequency of non-missing values in each interval, adjusted by random weights drawn from $S(0, 1)$. This ensures that missingness depends on the observed values of the selected variable, satisfying the MAR condition. For 50% missingness, the miss rate of the reference column was set to 30% in order to avoid intervals only containing missing values. The missingness of the level 2 variables was induced via MCAR in both scenarios. The resulting simulation has $2 \times 2 \times 2 \times 3 = 24$ conditions. Each combination of the above factors is replicated 1,000 times.

Treatment of missing data and choice of hyperparameters

Missing data were imputed using three methods: MICE, *missRanger*, and *mixgb*. For MICE, a multilevel approach was employed using the *mice* package (van Buuren & Groothuis-Oudshoorn, 2011), applying *2l.norm* for level 1 variables ($X_{ij1}, \dots, X_{ij6}, Y_{ij}$) to model random intercepts by cluster. For the level 2 variables ($L_{ij1}, \dots, L_{ij6}$) we used *2lonly.pmm*, i.e., predictive mean matching at the cluster level with other level 2 variables as predictors. This approach accounts for clustering but does not include random slopes. The exclusion of random slopes was intentional, as it avoids further favoring MICE, which was already advantaged by the simulation's linear mixed model design compared to the tree-based methods. After imputation, we fitted the analysis model to each completed dataset and combined the results using Rubin's rules (Little & Rubin, 2002).

We went with the default choices of hyperparameters if not stated elsewhere (Deng, 2023; Mayer, 2021; van Buuren & Groothuis-Oudshoorn, 2011). For *mice*, the number of multiple imputations was thus set to $m=5$, and the number of iterations was set to $\text{maxit}=10$. Moreover, PMM was used with five donors. For the random forest and the *ranger* package, there are more hyperparameters that the user has to choose. We used fully grown trees corresponding to $\text{max.depth}=\text{NULL}$ in the *ranger* function. For *mtry* and *sample.fraction* we also stayed at the default value, which was the rounded-down root of the number of features and 0.632 times the sample size, respectively. Since the random forest is computationally expensive, we reduce the number of trees to speed up the simulation and set $\text{num.trees}=300$ instead of the default value of 500, because the expected performance loss is negligible

(Probst & Boulesteix, 2018). As a splitting rule for *ranger*, we used the traditional variance minimization from the CART (Classification and Regression Trees) algorithm (Breiman, 2001; Breiman et al., 1984). An alternative conditional inference-based splitting criterion was introduced by Hothorn et al. (2006), which can reduce variable selection bias toward variables with many cut points (Strobl et al., 2007). This bias is inherent in traditional CART-based splitting rules and can also affect permutation importance measures. Such bias could, in principle, also affect downstream analyses. However, in our imputation context, the primary objective is to obtain plausible values that enable valid subsequent inferential analysis rather than estimation of variable importance. To our knowledge, no adaptation of the conditional inference-based splitting criterion for use in imputation currently exists. Moreover, as the standard random forest as implemented in *ranger* has great predictive performance (Grinsztajn et al., 2022), we expect that these properties are beneficial for imputation as well. For the predictive mean matching step, the number of candidates to sample from is set to 5. For the *mixgb* imputation method, we also use mostly default hyperparameter settings with two exceptions: The number of multiple imputations is set to $m = 5$ for better comparability with the other imputation approaches. For *mixgb* we use the default of histogram-based splitting (Deng, 2023). To balance computational efficiency with sufficient convergence, we limit the number of iterations for the imputation process to $\text{maxit} = 5$. These values reflect a compromise between computational cost and the stability of imputations needed for robust estimates. Prior to the main simulation, we conducted a tuning study to assess whether hyperparameter optimization could meaningfully improve the performance of *mixgb*. We performed a grid search across key parameters (including *num_boost_round*, *max_depth*, *eta*, and *subsample*) in a reduced simulation setting with 100 replications, optimizing root mean square error (RMSE). For each condition, we compared the statistical power obtained with *mixgb*'s default parameters to the power obtained with the best-performing (according to RMSE) hyperparameter configuration identified through a structured tuning grid search.

However, the results showed that optimized settings improved power by at most 0.02 on average, and the mean absolute change in power across all variables and scenarios was approximately 0.01 (see Table 1 and Table 2 in the Appendix). Given the negligible benefit relative to the substantial additional computational cost, we decided to use the default hyperparameter settings for *mixgb* in the main simulation study.

Simulation study results

In this simulation study, various imputation methods were evaluated and compared to address missing data in multi-level designs. The primary objective was to assess the performance of these methods with respect to subsequent inference across different settings. To this end, the methods were evaluated using rejection rates and coefficient bias. Rejection rates assess type I error control (in the case of a true null hypothesis) and statistical power (in the case of a false null hypothesis), while coefficient bias provides insight into the precision and reliability of the estimated parameters. Comparing the standard deviation of the parameter estimation in the imputed data with that in the complete data quantifies the uncertainty of the estimation caused by imputation. Evaluating different settings helps to assess the overall robustness of the imputation methods.

Since the overall conclusion from the results of the random intercept model aligns with that of the random slope model, we focus on discussing the former. The results for the random slope model are provided in the supplement.

Rejection rates

The Monte Carlo standard error (MCSE) for power and type I error rates was calculated following Morris et al. (2019) as

$$\sqrt{\frac{p(1-p)}{N}}$$

where p is the fraction of significant results and $N = 1,000$ is the number of replications, ranging from 0.0010 to 0.0158.

Type I error Figure 1 shows the average rejection rates of H_0 across replications for random intercept models, categorized by the missingness mechanism. Here, H_0 states that the variable (for level 1) or the cluster variable (for level 2) has no effect on the dependent variable. Only coefficients that are truly zero are included in the figure, thus reflecting the estimated type I errors. As seen in the subfigures of Fig. 1, missranger without predictive mean matching (displayed in the figures as ranger) exceeds the significance level (with and without dummies) for every coefficient across all combinations of missingness mechanism, missing rate, and cluster size. In some cases, the rejection rates are larger than 10%, despite the significance level being set to 5%, highlighting the importance of PMM. Consequently, we exclude random forests without PMM from the subsequent power analysis. For a cluster size of 25 and 50% missingness, the type I error with missranger and predictive mean matching (displayed as ranger5) is larger than desired for level 2 features, though the test maintains the significance level when using dummies. This

suggests that dummies can be crucial. However, for a cluster size of 50 and 50% missingness, all random forest approaches result in estimated type I errors exceeding 0.05 for the level 2 features. This holds for both MCAR and MAR. Notably, for the level 1 variables without an effect, the null hypothesis (H_0) was rejected in almost none of the cases for MICE and missranger with PMM, i.e., the type I error rate is close to zero, indicating that these methods are overly conservative.

As shown in Fig. 1, among the different imputation methods evaluated, only MICE, ranger.dummies with PMM, and mixgb (both with and without dummies) consistently achieved a type I error rate below 5%, while MICE was the only method that was always below 5%.

Power Figure 2 displays the rejection rates for the nonzero coefficients, i.e., the estimated power. While the plots for MAR and MCAR are very similar, the missingness rate clearly plays an important role: For 10% missingness, the level 1 effects are almost always detected with only minor differences between the imputation methods. However, for the level 2 variables, MICE outperformed the other methods, with a power more than 0.3 higher than the others for a cluster size of 25. For 30% missingness, MICE remains the benchmark for detecting effects on level 2, but for the level 1 variables, differences between the imputation methods start to occur: The power function at 0.5 for mixgb and mixgb with dummies is better than the power of MICE, where the variants with dummies seem to have a slight advantage. This holds independently of the cluster size.

This remains consistent across all cluster sizes. The variants of missranger performed worse than MICE and the mixgb methods. At 50% missingness, the advantage of mixgb with dummies for level 1 variables becomes more pronounced. For each cluster size and missingness mechanism, mixgb with dummies sets the benchmark. Particularly for MCAR, mixgb with dummies significantly outperforms MICE. For level 2 variables, mixgb with dummies performs similarly to MICE at a cluster size of 25, whereas at a cluster size of 50, MICE has a slight advantage. This pattern is consistent for both missingness mechanisms. Therefore, at 50% missingness, mixgb with dummies emerges as the preferred method, even though MICE benefits from knowing the linear structure of the model equation.

Coefficient bias

For the bias of zero coefficients, the MCSE was estimated as $\frac{SD}{\sqrt{N}}$ using the standard deviation (SD) from simulation

data, with values between 0.0044 and 0.0601. The MCSE for the relative bias of nonzero coefficients was computed similarly, ranging from 0.0407 to 0.2353, where all MSCE errors of more than 0.2 occurred, when missingness was 50%. Figure 3 shows the coefficient bias for each method under random intercept models with MAR. For the true zero coefficients at both level 1 and level 2, we observe that the bias values are consistently close to zero. Even with increasing missingness or decreasing cluster size, no significant deviations from zero are found. A more interesting observation is the relative bias (adjusted for effect size) for the nonzero coefficients. The relative bias is used to normalize the bias by the true parameter value, making it easier to interpret and compare across different coefficient sizes. It appears that all imputation methods induce negative bias. These findings regarding relative bias align closely with the rejection rates observed earlier. *mixgb*, both with and without dummies, exhibits relatively high relative bias for the level 2 variables. In contrast, MICE produces the lowest bias among the tested imputation methods for level 2 variables, regardless of the missingness rate. On the other hand, for level 1 variables, MICE shows larger bias than the boosting methods, and this difference increases with increasing missing rate. The random forest methods also show relatively large bias for both level 1 and level 2 variables.

Standard deviation

For the standard deviation of nonzero coefficients, the MCSE was determined as $SD/\sqrt{2N}$ (Morris et al., 2019), based on the standard deviation, ranging from 0.0023 to 0.22236 (the larger values only occurred occasionally for 50% missingness). Figure 4 shows the ratio of the standard deviation of parameter estimates from the imputed data to the standard deviation of parameter estimates from the complete dataset. The imputation increases the standard deviation for (almost) every combination of parameter, cluster size, missing mechanism, miss rate, and imputation method.

For the level 1 features, the increase in the standard deviation seems rather small at a 10% miss rate. At a miss rate of 30%, the imputation with *mixgb* (with or without boosting) doubles the standard deviation of the parameter estimation. In this case, MICE performs best at preserving the variance of the complete dataset. The difference becomes even more drastic at a 50% miss rate: the standard deviation for MICE stays robust for the true-zero features, while the other imputation methods multiply the standard deviation of parameter estimation. For the nonzero features, MICE remains best at preserving the standard deviation, but the difference relative to the other methods is smaller. At a 50% miss rate, MICE's standard deviation is not as robust for the nonzero

features as for the true-zero features, but is still better than the tree-based alternatives. Thus, the power advantage of *mixgb* for level 1 features at the larger miss rates comes from a smaller bias (Fig. 3).

For the level 2 features, the inflation of standard deviation is much higher than for the level 1 features, especially for larger miss rates. For a cluster size of 25 and 50% missingness, the standard deviation on the MICE imputed data is more than 10 times that on the full dataset. This also explains why *mixgb* has more power on level 2 (Fig. 1) despite having larger negative bias (Fig. 3). In this scenario, *ranger* is the best at preserving the original standard deviation. The standard deviation of MICE remains worse than the other imputation methods for the cluster size of 50. At a miss rate of 10%, MICE preservation is superior.

In addition to rejection rates and empirical standard deviations, we evaluated the calibration of model-based standard errors by examining the ratio of the average reported standard errors to the empirical standard deviation of the parameter estimates across replications (Morris et al., 2019). Ratios close to 1 indicate well-calibrated uncertainty estimates, whereas deviations reflect under- or overestimation of uncertainty. Figure 5 shows that *boost* and *boost.dummies* yield ratios close to 1 in the random intercept model. Results for the random intercept and random slope model are similar and are shown in Fig. A5 in the Appendix. For *mice*, model-based standard errors tend to be larger than the empirical standard deviations for level 1 parameters, particularly for higher missing rates and $J = 50$. In contrast, the variants of *ranger* (with and without dummy variables) show substantial deviations from 1, indicating both under- and overestimation of standard errors across simulation conditions also depending on the use of PMM.

Discussion

This paper critically evaluates the performance of novel tree-based imputation methods for handling missing data in hierarchical data structures, focusing on random forest (*missranger*)- and XGBoost (*mixgb*)-based approaches with and without dummy modeling.

Through a comprehensive simulation study, we contrast these novel techniques with the state-of-the-art MICE approach, focusing in particular on bias and inference. The evaluation considers level 1 and level 2 variables across two different data generation processes (random intercept and random slope models), three missingness rates (10%, 30%, and 50%), and two missingness mechanisms (MCAR and MAR). Our results indicate that MICE is characterized by consistent accuracy in rejection rates when using the 2l.norm method for level 1 variables and 2lonly.pmm for level 2

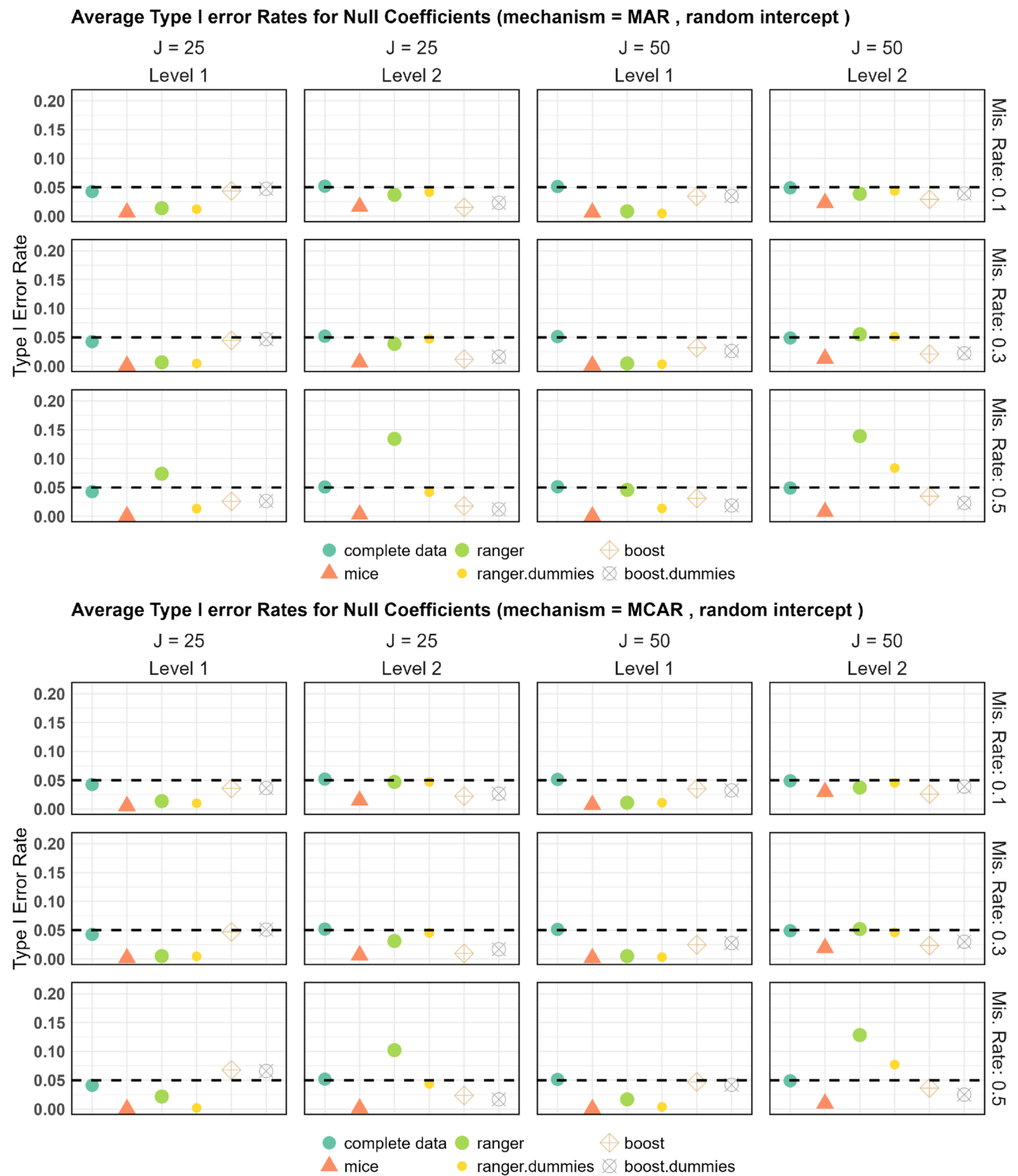


Fig. 1 Estimated type I errors for random intercept designs

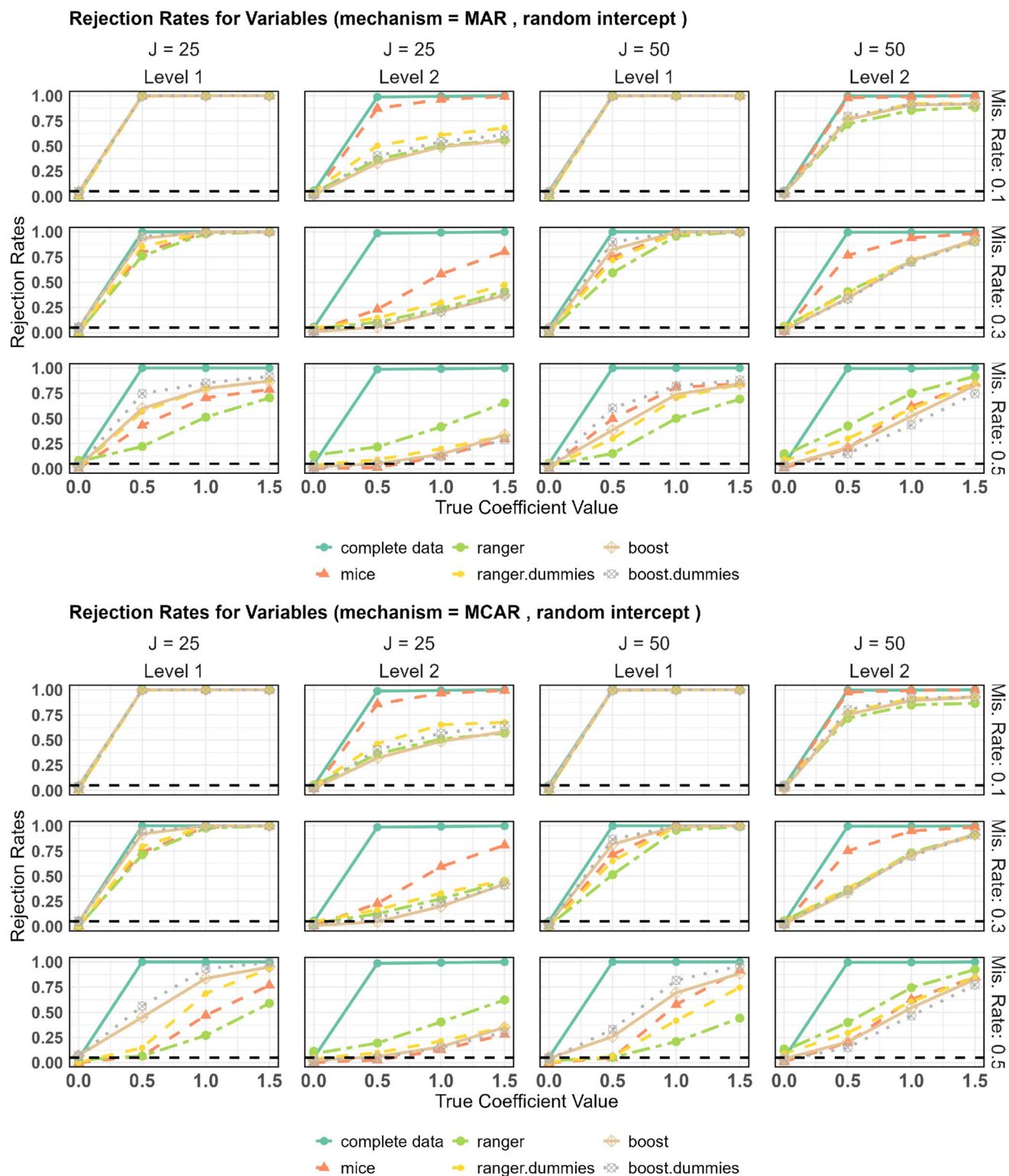


Fig. 2 Power for random intercept designs

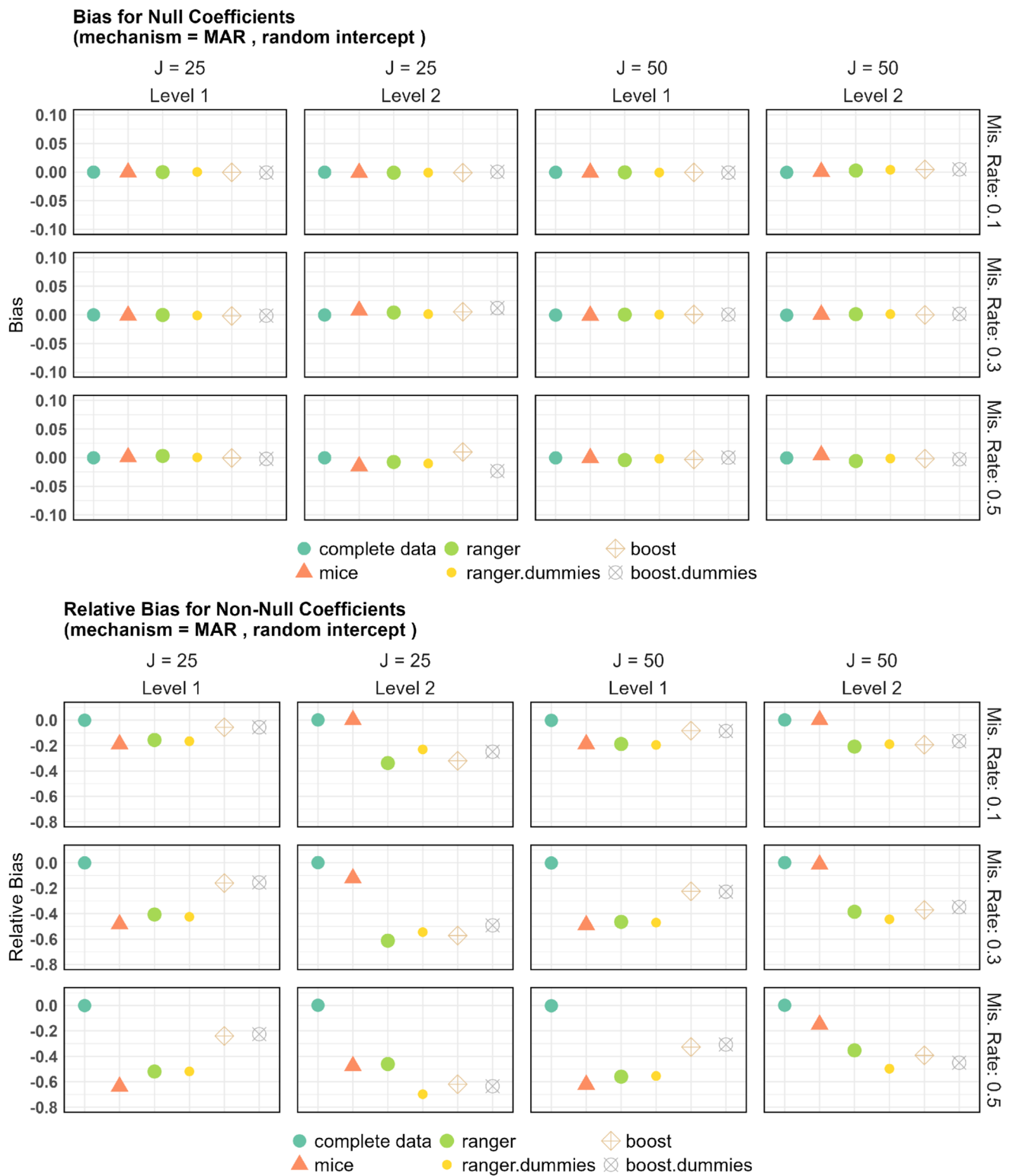


Fig. 3 Bias for the true zero coefficients and relative bias for the nonzero coefficients

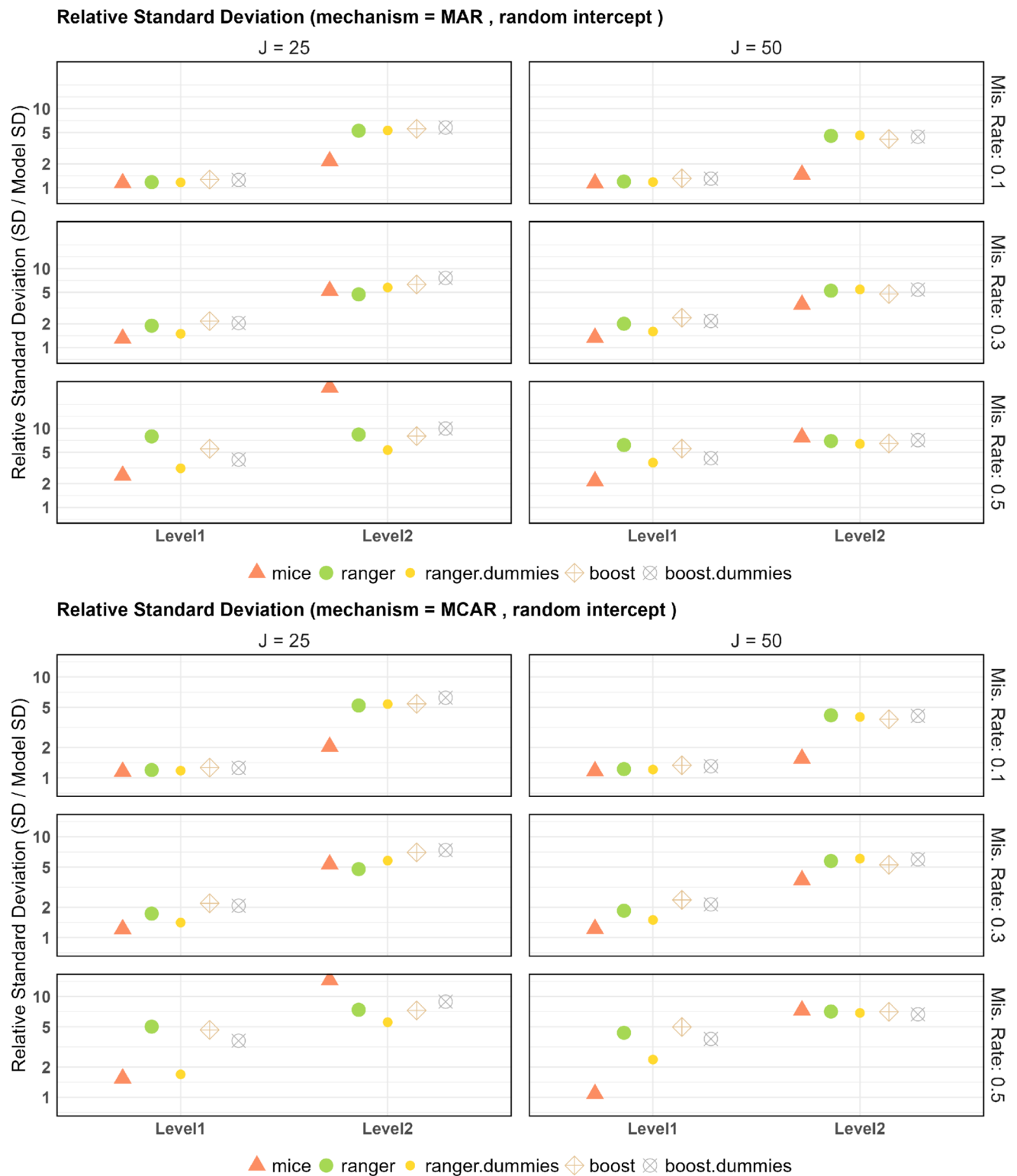


Fig. 4 Ratio of standard deviation (imputed data vs. complete data) for the random intercept model

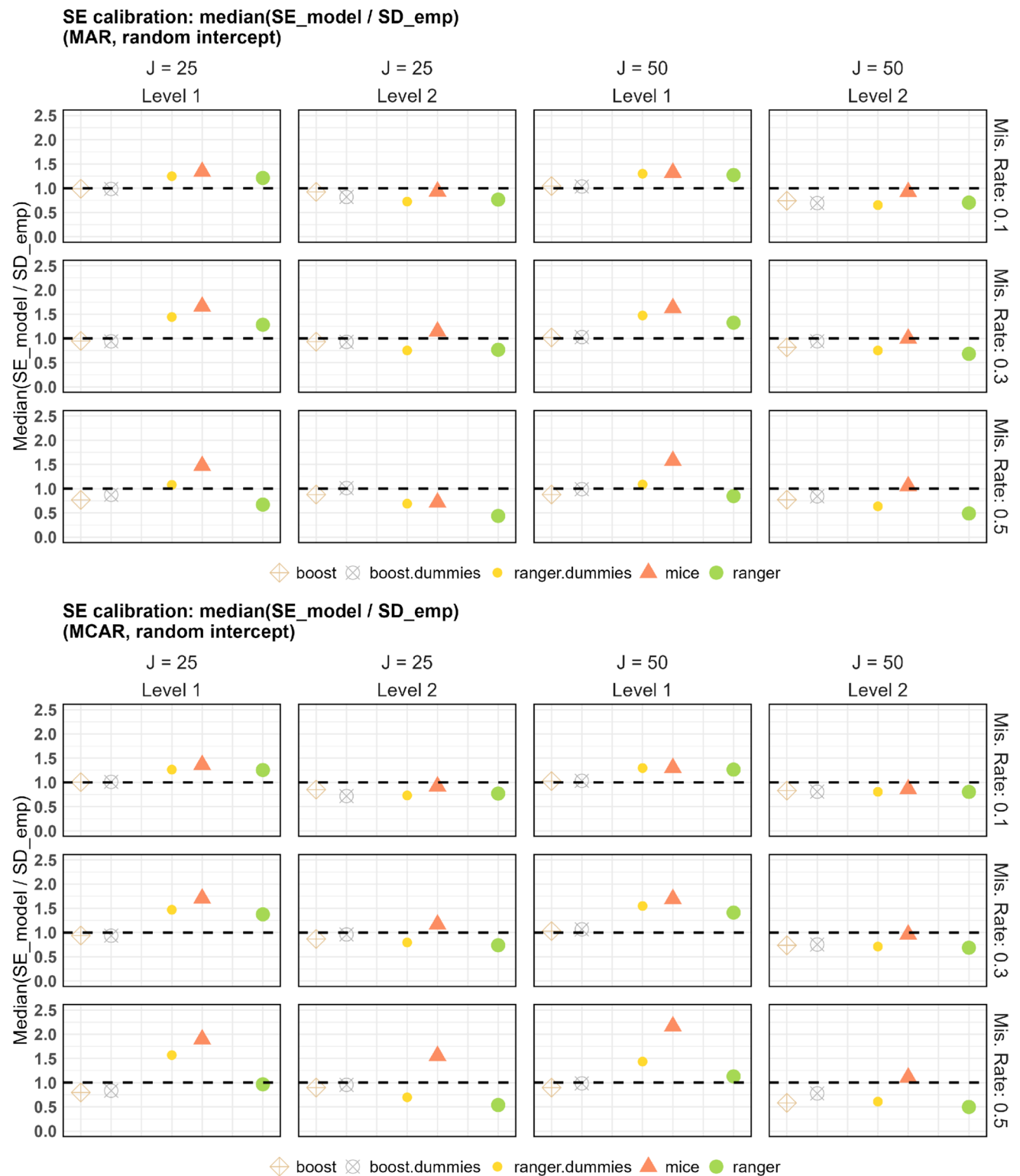


Fig. 5 Calibration of model-based standard errors for the random intercept model. Ratios of average model-based standard errors to empirical standard deviations across replications are shown

variables. This consistency underscores MICE's robustness to hierarchical data and closely approximates true rejection rates (i.e., type I error and power). Our results are consistent with the findings in Grund et al. (2018a), demonstrating the practical usefulness of MICE when dealing with missingness in multilevel structures, especially in the case of missingness at level 2. Similar to Enders et al. (2016), we also found that MICE was superior not only for random slope but for random intercept models, provided that the imputation model was specified according to the data generation process and the missing rate was low (10%).

For level 1 coefficients, boosting (mixgb) with dummies consistently outperforms other methods, including MICE, across all missingness mechanisms when the missingness rate is 30% or 50%. At 30% missingness, however, MICE has more power for level 2 variables, making the choice of imputation method dependent on the practitioner's priorities and the relative importance of level 1 versus level 2 variables in the specific application. At 50%, mixgb (with and without dummies) proved to be the preferable alternative in comparison to MICE regardless of missing mechanism or cluster size. The analysis of (relative) coefficient bias corroborates the findings from the rejection rates, revealing larger relative bias for MICE in level 1 variables and greater relative bias for the boosting methods in level 2 variables. The missranger methods, on the other hand, fell slightly behind the boosting methods and MICE for almost all settings and cannot be recommended as an imputation method in our multilevel settings. In particular, missranger without PMM showed substantially increased type I errors.

While the rejection rate differences between imputation methods vary slightly across the random intercept and random slope models, the overall trends remain consistent regardless of the data-generating process. The results do not necessarily indicate that tree-based methods with dummies (also including dummies for the level 2 cluster) outperform their standard counterpart without dummies. For example, under MCAR and MAR, standard mixgb showed slightly higher power than mixgb with dummies for level 2 coefficients when the missingness rate was 50% in both random intercept and random slope. Meanwhile, adjusted mixgb sometimes exhibited higher power for level 1 coefficients. In some situations (e.g., 50% missingness and cluster size 25), the introduction of dummies helped to decrease the type I error to the desired level. But for most scenarios the introduction of dummies did not alter the results much.

Limitations, strengths, and outlook

Our simulation study provides valuable insights into imputation methods for hierarchical data, but its generalizability is limited by the design of the data-generating process

(DGP). The DGP assumes a linear multilevel model, aligning closely with the assumptions of the linear mixed model used for analysis and the MICE imputation method (2l.norm) (van Buuren, 2018). This alignment likely favors MICE, particularly at low missingness rates (10%), as 2l.norm explicitly models linear and hierarchical relationships, unlike missRanger or mixgb, which are more flexible but do not assume linearity. Consequently, our results may not generalize to settings with nonlinear relationships (e.g., jumps or polynomial terms) or higher-order interaction effects, where tree-based methods could leverage their ability to capture complex patterns. The use of a low number of variables in our simulation further limits generalizability to complex, variable-rich datasets typical in real-world social science applications. The fact that mixgb was still able to outperform MICE for increasing missing rates underlines the potential of tree-based imputation methods for multilevel data.

Future research investigating the impact of nonlinear variable effects and a larger number of variables on both rejection rate and bias could provide a more complete understanding of the methods' performance in more complex data settings. In particular, simulating data beyond a linear mixed model (e.g., with jumps or non-monotonic effects) could be insightful, such as using a semi- or nonparametric data-generating model that remains plausible enough for a linear mixed model analysis. In addition, the tree-based methods missRanger and mixgb rely on traditional splitting algorithms (variance minimization and histogram-based, respectively), which may introduce bias in variable selection for certain tasks. Alternative approaches, such as conditional inference trees (Hothorn et al., 2006), could reduce such biases but are not yet adapted for imputation, presenting a promising direction for future research.

For empirical researchers, the alignment between DGP, imputation method, and analysis model is critical for imputation quality. When the underlying DGP is likely linear and hierarchical, as is common in the social sciences (e.g., education, psychology) (Maas & Hox, 2005), MICE with 2l.norm or 2lonly.pmm is recommended due to its robustness and ability to preserve multilevel structures. For larger missing rates, mixgb can be a viable alternative. For other DGPs, further research is needed.

In addition, the multilevel structures considered in this study are limited. Although random intercepts and random slopes are included, more complex multilevel features such as contextual effects of level 1 predictors, cross-level interactions, or alternative centering strategies are not part of the DGP. Addressing these more complex multilevel structures would require substantially different simulation designs and is therefore left for future research. A further limitation of the present simulation study concerns the assessment of inferential uncertainty. While we evaluated rejection rates,

empirical standard deviations, and the calibration of model-based standard errors, we did not directly assess confidence interval coverage.

The strength of the current study, however, lies in its novelty. We introduce and evaluate the combination of the latest tree-based imputation methods and a simple multi-level adjustment via dummy variables against established techniques such as MICE. This approach provides new insights into the evolving landscape of data imputation methods, especially in the context of hierarchical data structures.

Another interesting area to explore would be to improve the performance of tree-based methods. This could involve experimenting with multivariate tree-based methods (Schmid et al., 2023; Sega & Xiao, 2011). Tree-based methods with random effects have been developed (Bergonzoli et al., 2024; Buczak, 2024; Fokkema et al., 2020; Fokkema et al., 2018; Hajjem et al., 2014, 2017; Sela & Simonoff, 2012), but still have to be adapted to missing data. Another possible approach could be to first average all data at level 2 and impute missing data at this aggregate level, and then merge only the level 2 variables with the individual level dataset. A further idea for future research is a comparison of (potentially improved) tree-based methods with model-based approaches of (Enders, 2023; Enders et al., 2020; Grund et al., 2021) or the evaluation of different regression models like an all-noise model or a model with more coefficients. Future research should also explore imputation methods for systematically missing level 1 features, which were not included in our simulation, focusing on MCAR and MAR mechanisms, to address more complex missing data scenarios (Jolani et al., 2015; Resche-Rigon & White, 2018).

In conclusion, our study underscores the continued effectiveness of MICE in dealing with hierarchical data, particularly in terms of rejection rates. However, the emerging tree-based methods, especially mixgb, show potential for larger missing rates, suggesting their usefulness as alternatives in certain contexts. This dual finding opens new avenues for future research and practical applications in data imputation, highlighting the dynamic nature of the field.

Appendix

Table 1 Average power for default and tuned mixgb (model: random intercept) by missing rate

Mechanism	Miss rate	Variable	Mean default	Mean tuned	Difference
MAR	0.10	L21	0.33	0.33	0.00
MAR	0.10	L22	0.59	0.59	−0.00
MAR	0.10	L23	0.79	0.79	0.00
MAR	0.10	X ₁	0.52	0.52	0.00
MAR	0.10	X ₂	0.97	0.97	0.00
MAR	0.10	X ₃	1.00	1.00	0.00
MAR	0.30	L21	0.32	0.33	0.00
MAR	0.30	L22	0.60	0.60	−0.00
MAR	0.30	L23	0.78	0.79	0.02
MAR	0.30	X ₁	0.52	0.52	−0.00
MAR	0.30	X ₂	0.96	0.96	−0.00
MAR	0.30	X ₃	1.00	1.00	0.00
MAR	0.50	L21	0.32	0.33	0.00
MAR	0.50	L22	0.61	0.61	−0.01
MAR	0.50	L23	0.78	0.79	0.01
MAR	0.50	X ₁	0.50	0.52	0.01
MAR	0.50	X ₂	0.95	0.96	0.01
MAR	0.50	X ₃	1.00	1.00	0.00
MCAR	0.10	L21	0.30	0.30	0.00
MCAR	0.10	L22	0.56	0.56	−0.00
MCAR	0.10	L23	0.79	0.79	−0.00
MCAR	0.10	X ₁	0.48	0.49	0.01
MCAR	0.10	X ₂	0.96	0.96	0.00
MCAR	0.10	X ₃	1.00	1.00	0.00
MCAR	0.30	L21	0.27	0.28	0.01
MCAR	0.30	L22	0.56	0.55	−0.01
MCAR	0.30	L23	0.78	0.79	0.01
MCAR	0.30	X ₁	0.43	0.46	0.03
MCAR	0.30	X ₂	0.91	0.94	0.03
MCAR	0.30	X ₃	0.99	1.00	0.00
MCAR	0.50	L21	0.30	0.32	0.01
MCAR	0.50	L22	0.52	0.51	−0.01
MCAR	0.50	L23	0.74	0.75	0.01
MCAR	0.50	X ₁	0.46	0.49	0.03
MCAR	0.50	X ₂	0.81	0.83	0.02
MCAR	0.50	X ₃	0.98	0.99	0.01

Table 2 Average power for default and tuned mixgb (model: random intercept and random slope) by missing rate

Mechanism	Miss rate	Variable	Mean default	Mean tuned	Difference
MAR	0.10	L21	0.25	0.26	0.00
MAR	0.10	L22	0.51	0.52	0.00
MAR	0.10	L23	0.73	0.74	0.00
MAR	0.10	X ₁	0.49	0.49	-0.00
MAR	0.10	X ₂	0.92	0.92	0.01
MAR	0.10	X ₃	1.00	1.00	0.00
MAR	0.30	L21	0.26	0.24	-0.02
MAR	0.30	L22	0.52	0.52	0.01
MAR	0.30	L23	0.75	0.75	0.00
MAR	0.30	X ₁	0.49	0.50	0.01
MAR	0.30	X ₂	0.91	0.91	0.00
MAR	0.30	X ₃	0.99	0.99	0.00
MAR	0.50	L21	0.28	0.29	0.01
MAR	0.50	L22	0.52	0.52	0.00
MAR	0.50	L23	0.76	0.76	0.00
MAR	0.50	X ₁	0.48	0.49	0.01
MAR	0.50	X ₂	0.88	0.90	0.01
MAR	0.50	X ₃	0.99	0.99	0.00
MCAR	0.10	L21	0.23	0.23	0.00
MCAR	0.10	L22	0.56	0.55	-0.01
MCAR	0.10	L23	0.72	0.73	0.01
MCAR	0.10	X ₁	0.42	0.44	0.01
MCAR	0.10	X ₂	0.92	0.93	0.00
MCAR	0.10	X ₃	0.99	0.99	0.00
MCAR	0.30	L21	0.25	0.23	-0.01
MCAR	0.30	L22	0.56	0.56	-0.00
MCAR	0.30	L23	0.69	0.68	-0.01
MCAR	0.30	X ₁	0.41	0.47	0.06
MCAR	0.30	X ₂	0.84	0.88	0.04
MCAR	0.30	X ₃	0.99	0.99	0.00
MCAR	0.50	L21	0.24	0.24	-0.00
MCAR	0.50	L22	0.49	0.50	0.01
MCAR	0.50	L23	0.70	0.69	-0.01
MCAR	0.50	X ₁	0.42	0.43	0.01
MCAR	0.50	X ₂	0.75	0.80	0.05
MCAR	0.50	X ₃	0.97	0.98	0.01

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13428-026-03008-x>.

Author note The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as project 271512359.

Authors' contributions Conceptualization: JS, PD; Data curation: NF, JS, KG; Formal analysis: NF, JS, KG; Funding acquisition: MP, PD; Investigation: NF, JS, KG; Methodology: NF, JS, KG, MP, PD; Project

administration: NF, JS; Resources: PD, MP; Software: NF, KG; Supervision: JS, MP, PD; Validation: JS, KG; Visualization: NF, JS, KG; Writing—original draft: NF, JS, KG; Writing—review & editing: NF, JS, MP, PD

Funding Open Access funding enabled and organized by Projekt DEAL. The project “From Prediction to Agile Interventions in the Social Sciences (FAIR)” receives funding from the “Profilbildung 2020” program, an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia. The sole responsibility for the content of this publication lies with the authors.

Nico Föge was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 314838170, GRK 2297 MathCoRe.

Availability of data and materials The simulated datasets during the current study are available under https://drive.google.com/file/d/1BD1EhQtyBGYArkZpgOUR51eIQsHaqO5Z/view?usp=drive_link.

Code availability The code used in this study is available under <https://osf.io/8256p/overview>.

Declarations

Conflict of interest/Competing interests The authors declare that they have no conflicts of interest related to this work.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication All authors have given their consent for publication of this manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A, General*, 149(1), 1–26. <https://doi.org/10.2307/2981882>
- Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., ..., & Resche-Rigon, M. (2018). Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science*, 33 (2), 160–183. <https://doi.org/10.1214/18-STS646>
- Bergonzoli, G., Rossi, L., & Masci, C. (2024). Ordinal mixed-effects random forest. *arXiv preprint arXiv:2406.03130*. <https://doi.org/10.48550/arXiv.2406.03130>
- Breiman, L. (2001). Random forest. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>

- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Chapman and Hall/CRC. (Published by Wadsworth).
- Buczak, P. (2024). Mixed-effects frequency-adjusted borders ordinal forest: A tree ensemble method for ordinal prediction with hierarchical data. *OSF Preprint*. <https://doi.org/10.31219/osf.io/ny6we>
- Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50, 317–372. <https://doi.org/10.3368/jhr.50.2.317>
- Carpenter, J. R., Bartlett, J. W., Morris, T. P., Wood, A. M., Quartagno, M., & Kenward, M. G. (2023). *Multiple imputation and its application* (2nd ed). Wiley.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785–794). ACM, San Francisco, CA, USA. <https://doi.org/10.1145/2939672.2939785>
- Deng, Y. (2023). mixgb: Multiple imputation through 'xgboost' [Computer software manual]. (R package version 1.0.2) Retrieved September 29, 2025, from <https://CRAN.R-project.org/package=mixgb>
- Deng, Y., & Lumley, T. (2024). Multiple imputation through XGBoost. *Journal of Computational and Graphical Statistics*, 33(2), 352–363. <https://doi.org/10.1080/10618600.2023.2252501>
- Drechsler, J. (2015). Multiple imputation of multilevel missing data—Rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, 40(1), 69–95. <https://doi.org/10.3102/1076998614563393>
- Enders, C. K. (2023). *Applied missing data analysis* (2nd ed.). Guilford Press.
- Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, 21(2), 222–240. <https://doi.org/10.1037/met0000063>
- Enders, C. K., Keller, B. T., & Du, H. (2020). A comparison of full-information and multiple imputation strategies for handling missing data in repeated measures studies with random slopes. *Behavior Research Methods*, 52(4), 1406–1422. <https://doi.org/10.3758/s13428-019-01322-9>
- Erler, N. S., Rizopoulos, D., van Rosmalen, J., Jaddoe, V. W. V., Franco, O. H., & Lesaffre, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full bayesian approach. *Statistics in Medicine*, 35(17), 2955–2974. <https://doi.org/10.1002/sim.6944>
- Erler, N. S., Rizopoulos, D., Jaddoe, V. W. V., Franco, O. H., & Lesaffre, E. M. E. H. (2017). Bayesian imputation of time-varying covariates in linear mixed models. *Statistical Methods in Medical Research*, 28(4), 1295–1312. <https://doi.org/10.1177/0962280217730851>
- Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, 50(5), 2016–2034. <https://doi.org/10.3758/s13428-017-0971-x>
- Fokkema, M., Edbrooke-Childs, J., & Wolpert, M. (2020). Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research*, 1–13. <https://doi.org/10.1080/10503307.2020.1785037>
- Golino, H. F., & Gomes, C. M. (2016). Random forest as an imputation method for education and psychology research: Its impact on item fit and difficulty of the rasch model. *International Journal of Research & Method in Education*, 39(4), 401–421. <https://doi.org/10.1080/1743727X.2016.1168798>
- Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35, 507–520.
- Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data at level 2: A comparison of fully conditional and joint modeling in multilevel designs. *Journal of Educational and Behavioral Statistics*, 43(3), 316–353. <https://doi.org/10.3102/1076998617738087>
- Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, 21(1), 111–149. <https://doi.org/10.1177/1094428117703686>
- Grund, S., Lüdtke, O., & Robitzsch, A. (2021). Pooling imputed datasets with random effects: Substantive model compatible multiple imputation. *Methodology*, 17(2), 121–142. <https://doi.org/10.5964/meth.6459>
- Grund, S., Lüdtke, O., & Robitzsch, A. (2024). Missing data in the analysis of multilevel and dependent data. In M. Stemmler, W. Wiedermann, & F. Huang (Eds.), *Dependent data in social sciences research: Forms, issues, and methods of analysis* (2nd ed.). Springer.
- Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6), 1313–1328. <https://doi.org/10.1080/00949655.2012.741599>
- Hajjem, A., Larocque, D., & Bellavance, F. (2017). Generalized mixed effects regression trees. *Statistics and Probability Letters*, 126, 114–118. <https://doi.org/10.1016/j.spl.2017.02.033>
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674. <https://doi.org/10.1198/106186006X133933>
- Hox, J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). Springer.
- Hox, J., & Roberts, J. K. (2010). *Handbook of advanced multilevel analysis* (1st ed.). Routledge. <https://doi.org/10.4324/9780203848852>
- Jolani, S., Debray, T. P. A., Koffijberg, H., van Buuren, S., & Moons, K. G. M. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: A generalized approach using mice. *Statistics in Medicine*, 34, 1841–1863. <https://doi.org/10.1002/sim.6451>
- Kistner, F., Tulowitzki, J., Slaney, L., Alibhai, S., Jewell, Z., Ramosaj, B., & Pauly, M. (2024). Enhancing endangered species monitoring by lowering data entry requirements with imputation techniques as a preprocessing step for the footprint identification technology (FIT). *Ecological Informatics*, 102676. <https://doi.org/10.1016/j.ecoinf.2024.102676>
- Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed). Wiley & Sons.
- Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, 22(1), 141–165. <https://doi.org/10.1037/met0000096>
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3), 86–92. <https://doi.org/10.1027/1614-2241.1.3.86>
- Mayer, M. (2021). missRanger: Fast imputation of missing values [Computer software manual]. (R package version 2.1.3). Comprehensive R Archive Network (CRAN). Retrieved September 29, 2025 <https://CRAN.R-project.org/package=missRanger>
- McNeish, D., Stapleton, L. M., & Silverman, R. D. (2016). On the unnecessary ubiquity of hierarchical linear modeling in social and behavioral sciences. *Psychological Methods*, 22(1), 114–140. <https://doi.org/10.1037/met0000078>
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. <https://doi.org/10.1002/sim.8086>
- Probst, P., & Boulesteix, A.-L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18(181), 1–18.

- Ramosaj, B., Amro, L., & Pauly, M. (2020). A cautionary tale on using imputation methods for inference in matched-pairs design. *Bioinformatics*, 36(10), 3099–3106. <https://doi.org/10.1093/bioinformatics/btaa082>
- Resche-Rigon, M., & White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 27, 1634–1649. <https://doi.org/10.1177/0962280216666564>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (2nd ed.). Wiley & Sons, UK.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473–489. <https://doi.org/10.1080/01621459.1996.10476908>
- Schafer, J. L. (2000). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.
- Schmid, L., Gerharz, A., Groll, A., & Pauly, M. (2023). Tree-based ensembles for multi-output regression: Comparing multivariate approaches with separate univariate ones. *Computational Statistics & Data Analysis*, 179, 107628. <https://doi.org/10.1016/j.csda.2022.107628>
- Sega, M., & Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 80–87. <https://doi.org/10.1002/widm.12>
- Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, 86(2), 169–207. <https://doi.org/10.1007/s10994-011-5258-3>
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE Publications.
- Speidel, M., Drechsler, J., & Sakshaug, J. W. (2018). Biases in multilevel analyses caused by cluster-specific fixed-effects imputation. *Behavior Research Methods*, 50(5), 1824–1840. <https://doi.org/10.3758/s13428-017-0951-1>
- Steenbergen, M. R., & Jones, B. S. (2002). Modeling multilevel data structures. *American Journal of Political Science*, 46(1), 218–237. <https://doi.org/10.2307/3088424>
- Stekhoven, D. J., & Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118. <https://doi.org/10.1093/bioinformatics/btr597>
- Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1), Article 25.
- Suh, H., & Song, J. (2023). A comparison of imputation methods using machine learning models. *Communications for Statistical Applications and Methods*, 30(3), 331–341. <https://doi.org/10.29220/CSAM.2023.30.3.331>
- Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6), 363–377. <https://doi.org/10.1002/sam.11348>
- Thurrow, M., Dumpert, F., Ramosaj, B., & Pauly, M. (2021a). Goodness (of fit) of Imputation Accuracy: The GoodImpact Analysis. (arXiv:2101.07532) [arXiv preprint]. arXiv. <https://doi.org/10.48550/arXiv.2101.07532>
- Thurrow, M., Dumpert, F., Ramosaj, B., & Pauly, M. (2021b). Imputing missings in official statistics for general tasks—our vote for distributional accuracy. *Statistical Journal of the IAOS*, 37(4), 1379–1390. <https://doi.org/10.3233/SJI-210798>
- van Buuren, S. (2011). Multiple imputation of multilevel data. In J. Hox & J. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173–196). Routledge.
- van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9780429492259>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://doi.org/10.18637/jss.v045.i03>
- van Buuren, S., & Oudshoorn, K. (1999). *Flexible multivariate imputation by mice*. TNO.
- van Buuren, S., Brand, J., Groothuis-Oudshoorn, C., & Rubin, D. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064. <https://doi.org/10.1080/10629360600810434>
- Waller, N. G. (2022). fungible: Psychometric functions from the Waller lab (Version 2.2.1) [R package]. Comprehensive R Archive Network (CRAN). Retrieved September 29, 2025, from <https://CRAN.R-project.org/package=fungible>
- Wright, M. N., Ziegler, A., & König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics*, 17, 1–10. <https://doi.org/10.1186/s12859-016-0995-8>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.