technische universität
dortmund

fakultät
statistik

Analyzing Consistency and Statistical Inference in Random Forest Models

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of
DOCTOR RERUM NATURALIUM
at the Faculty of Statistics
of the Technical University of Dortmund

by

BURIM RAMOSAJ

March, 2020

Acting Dean:	Prof. Dr. Katja Ickstadt
Primary referee:	Prof. Dr. Markus Pauly
Secondary referee:	Prof. Dr. Jörg Rahnenführer
Commission chairperson:	Prof. Dr. Roland Fried
Assessor:	Dr. Uwe Ligges
Day of the oral examination:	July, 10, 2020

Acknowledgments

I started my PhD studies back in March, 2017 at the Institute of Statistics at Ulm University under the supervision of Markus Pauly. Although his professorship had initial problems in funding my PhD studies, he managed to receive third-party funds from the Daimler AG. Beside its commitment for enabling me the PhD studies back in Ulm, I also want to thank him for his professional support during these three years of academic education. Beyond the universities reputation, he is one of the main reasons I followed him to the Technical University of Dortmund. In this context, I also want to address additional thanks to the Daimler company for initially funding my PhD and Professor Rahnenführer for agreeing to be my second supervisor.

Furthermore, I want to express my gratitude to Gérard Biau and Erwan Scornet for fruitful discussions on Random Forest related issues during a scholary visit at the Sorbonne Université and the École Polythenique in Paris. Our discussions helped me to further organize theoretical thoughts and relations in Random Forest models. I also want to say thank you to my colleagues in Ulm and in Dortmund. We enjoyed funny moments during our coffee breaks talking about politics, social problems, science but rarely about statistics :-D

My father migrated to Germany as a skilled construction worker back in 1970. Unfortunately, he and my mother never had the opportunity to enjoy higher education, either because of financial or political reasons in their homecountry. Nevertheless, both of them motivated me as a young kid to learn and work hard for the goals I desire receiving the love and help to study Mathematics at the Syracuse University in NY, USA and Mathematics and Management at Ulm University. Sadly, my father passed away in 2015 therefore being not able to share with me important moments. This is the reason, I want to thank him, Imer B. Ramosaj, at the very first place for all his support and advices during the lifetime we shared together. My mother, Gjyle Ramosaj, has always been the voice of wisdom in my life. I want to thank her deeply for the support and encouragement she still gives me.

During my PhD studies, I got to know my wife, Albertina Ramosaj. I want to thank her for the endurance she had during my final PhD-phase. I also want to address special thanks to my brothers and sisters, Syzana Ramosaj, Jehona Ramosaj, Bujar Ramosaj and Dunjeta Ramosaj. You stood by my side even in difficult times. Last but not least, I want to thank my uncle, Smajl Zeqiraj, for having always fruitful discussions on life-related issues and all my friends and colleagues outside the university.

United we stand, divided we fall.
-Aesop-

Abstract

This thesis pays special attention to the Random Forest method as an ensemble learning technique using bagging and feature sub-spacing covering three main aspects: its behavior as a prediction tool under the presence of missing values, its role in uncertainty quantification and variable screening. In the first part, we focus on the performance of Random Forest models in prediction and missing value imputations while opposing it to other learning methods such as boosting procedures. Therein, we aim to discover potential modifications of Breiman's original Random Forest in order to increase imputation performance of Random Forest based models using the normalized root mean squared error and the proportion of false classification as evaluation measures. Our results indicated the usage of a mixed model involving the stochastic gradient boosting and a Random Forest based on kernel sampling. Regarding inferential statistics after imputation, we were interested if Random Forest methods do deliver correct statistical inference procedures, especially in repeated measures ANOVA. Our results indicated a heavy inflation of type-I-error rates for testing no mean time effects. We could furthermore show that the between imputation variance according to Rubin's multiple imputation rule vanishes almost surely, when repeatedly applying `missForest` as an imputation scheme. This has the consequence of less uncertainty quantification during imputation leading to scenarios where imputations are not *proper*. Closely related to the issue of valid statistical inference is the general topic of uncertainty quantification. Therein, we focused on consistency properties of several residual variance estimators in regression models and could deliver theoretical guarantees that Random Forest based estimators are consistent. Beside prediction, Random Forest is often used as a screening method for selecting *informative features* in potentially high-dimensional settings. Focusing on regression problems, we could deliver a formal proof that the Random Forest based internal permutation importance measure delivers on average correct results, i.e. is (asymptotically) unbiased. Simulation studies and real-life data examples from different fields support our findings in this thesis.

Contents

Abbreviations

CART	Classification and Regression Trees.
FCS	Fully-conditional specification; a class of imputation models.
JM	Joint-Modelling; a class of imputation models.
MAR	Missing at Random.
MCAR	Missing Completely at Random.
MNAR	Missing not at Random.
NORM	`mice` imputation procedure using Bayesian linear regression assuming normality.
NRMSE	Normalized Root Mean Squared Error; performance measure for evaluating imputation schemes.
OOB	Out-of-Bag sample; used especially in Bagging procedures.
PFC	Proportion of False Classification; performance measure for evaluating imputation schemes.
PMM	Predictive mean matching; `mice` imputation procedure.
RF MI	Random Forest Multiple Imputation; repeatedly applying the `missForest` method in R.
RF MICE	`mice` imputation procedure with a modified Random Forest scheme.
RFPIM	Random Forest Permutation Importance Measure.
a.s.	almost surely.
resp.	respectively.
w.l.o.g.	without loss of generality.
w.r.t.	with respect to.
$\|\mathbf{x}\|_r$	The r-norm of a p-dimensional vector $\mathbf{x} = [x_1, \ldots, x_p]^\top \in \mathbb{R}^p$ for some $r \in \mathbb{N}$, i.e. $\|\mathbf{x}\|_r = \left(\sum_{j=1}^{p} \|x_j\|^r \right)^{1/r}$.
$\Omega = \dot{\bigcup}_{j=1}^{p} A_j$	Disjoint separation of Ω for a sequence of sets $\{A_j\}_{j=1}^{p}$, i.e. $A_k \cap A_s = \emptyset$ for all $k \neq s$ and $\bigcup_{j=1}^{p} A_j = \Omega$.

$\lceil x \rceil$	Ceiling function, i.e. smallest integer greater than or equal to x.
$\mathcal{B}(\mathbb{R})$	The Borel $\sigma-$field.
$\mathbb{1}\{x \in A\}$	Indicator function. Attains value 1, if x lies in A, otherwise it attains the value 0.
$(M, n) \xrightarrow{seq} (a, b)$	The sequential limit in the sense that $\lim_{n \to b} \lim_{M \to a}$.
$Med(\mathbf{x})$	The median of a vector $\mathbf{x} = [x_1, \ldots, x_n]^\top$.
$Var(\mathbf{X})$; $Var(X)$	$Var(\mathbf{X})$ is the covariance matrix of a random vector $\mathbf{X} \in \mathbb{R}^p$, $p > 1$ and $Var(X)$ the variance of a random variable X, presuming that both exists.
$\mathbb{E}[\mathbf{X}]$; $\mathbb{E}[X]$	$\mathbb{E}[\mathbf{X}]$ is the expectation vector of a random vector $\mathbf{X} \in \mathbb{R}^p$, $p > 1$ and $\mathbb{E}[X]$ the expectation of a random variable X, presuming that both exists.
$\mathcal{S}_n$	Symmetric group, i.e. the set of all permutations $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$.
$a \propto b$	Proportional sign, i.e. there is a constant c such that $a = cb$.
$m_{n,M}$	Random Forest regression function trained on a training set $\mathcal{D}_n$ consisting of M decision trees.
$m_{n,M}^{OOB}(\mathbf{X}_i)$	Random Forest Out-of-Bag prediction at $\mathbf{X}_i$ trained on a training set $\mathcal{D}_n$ with M decision trees.
$m_{n,\infty}$	Random Forest regression function trained on a training set $\mathcal{D}_n$ with an infinite number of decision trees.
$m_{n,\infty}^{OOB}(\mathbf{X}_i)$	Random Forest Out-of-Bag prediction at $\mathbf{X}_i$ trained on a training set $\mathcal{D}_n$ with an infinite number of decision trees.
$tr(\mathbf{A})$	Trace of matrix a $\mathbf{A}$, i.e. the sum of the diagonal elements.
$\mathbf{X}_n \xrightarrow{\mathbb{P}} \mathbf{X}$	Convergence in probability, i.e. for all $\epsilon > 0$, we have $\lim_{n \to \infty} \mathbb{P}[\|\mathbf{X}_n - \mathbf{X}\|_1 > \epsilon] = 0$.
$\mathbf{X}_n \xrightarrow{L_r} \mathbf{X}$	Convergence in L_r-norm for some $r > 0$, i.e. $\lim_{n \to \infty} \mathbb{E}[\|\mathbf{X}_n - \mathbf{X}\|_r^r] = 0$.

Introduction

The 21st century is faced with several problems, where statistical science plays a key role in finding appropriate answers and solutions. In automotive engineering, for example, the research trend has drastically changed to autonomously driving cars. This, however, requires complex algorithmic models in order to identify potential objects during the video recording process such as neural nets in deep learning frameworks, which mimic the human brain activities. This is partially boosted by the enhanced technology in computer science, which has provided fast computational performance and enormous saving memories. A product of these developments are vast amounts of data with complex dependence structures, which require an appropriate modeling strategy. In social science, for example, companies such as *Facebook* or *Instagram* deliver platforms where people can be globally connected while saving personal information about each member of the platform. Analyzing these type of data have the severe effect of learning personal traits and preferences in order to provide, e.g., personalized commercial. As the examples show, the data being collected so far can result into mathematical issues, where a *traditional* modeling strategy can be either too time consuming for the specific problem at hand or even too complex in order to deliver satisfying results. This has partially pushed practitioners and scientists to change the way statistical modeling actually happens. Instead of modeling the complex data structure at hand and finding appropriate tools for analyzing them, simply develop an algorithm that solves the current problem at hand. This was the birth of the *Machine Learning* discipline. However, this comes with the cost of difficult interpretability and the lack of inferential statistics.

The flowchart diagram in Figure 1 illustrates the different approaches data is analyzed from a Machine Learner's perspective and the traditional approach using statistical modeling. The latter is focused on the correct statistical modeling of mathematical forces involved in the data generating process resp. the underlying analysis model. In Machine Learning, an algorithm is generated which is problem based and mostly led by intuition. Their difference is of conceptual nature while both approach's frontiers often gets blurred. Hence, there are data analyzing problems, which can be tackled by both methods. As an illustrative example, suppose that you want to measure the effect of a variable X on an outcome Y. From a statistical modeling point of view, one would assume a functional relationship f between X and Y - possibly of linear nature - and check whether e.g. the coefficients in a linear relationship between X and Y are vanishing. A Machine Learner, however, would first think of an algorithm, that might describe the relationship between X and Y without specifying any functional relationship between them. Therein, a possible approach would then be to permute the values of X, and check whether the predictive accuracy of the developed algorithm has drastically changed. However, the differences in both approaches have consequences for the later analysis. The statistical modeling approach enables an interpretation of its outcomes, for which statistical inference in terms of significance tests are possible. This is mainly based on the precedent work during the modeling phase. In Machine Learing, however, the under-

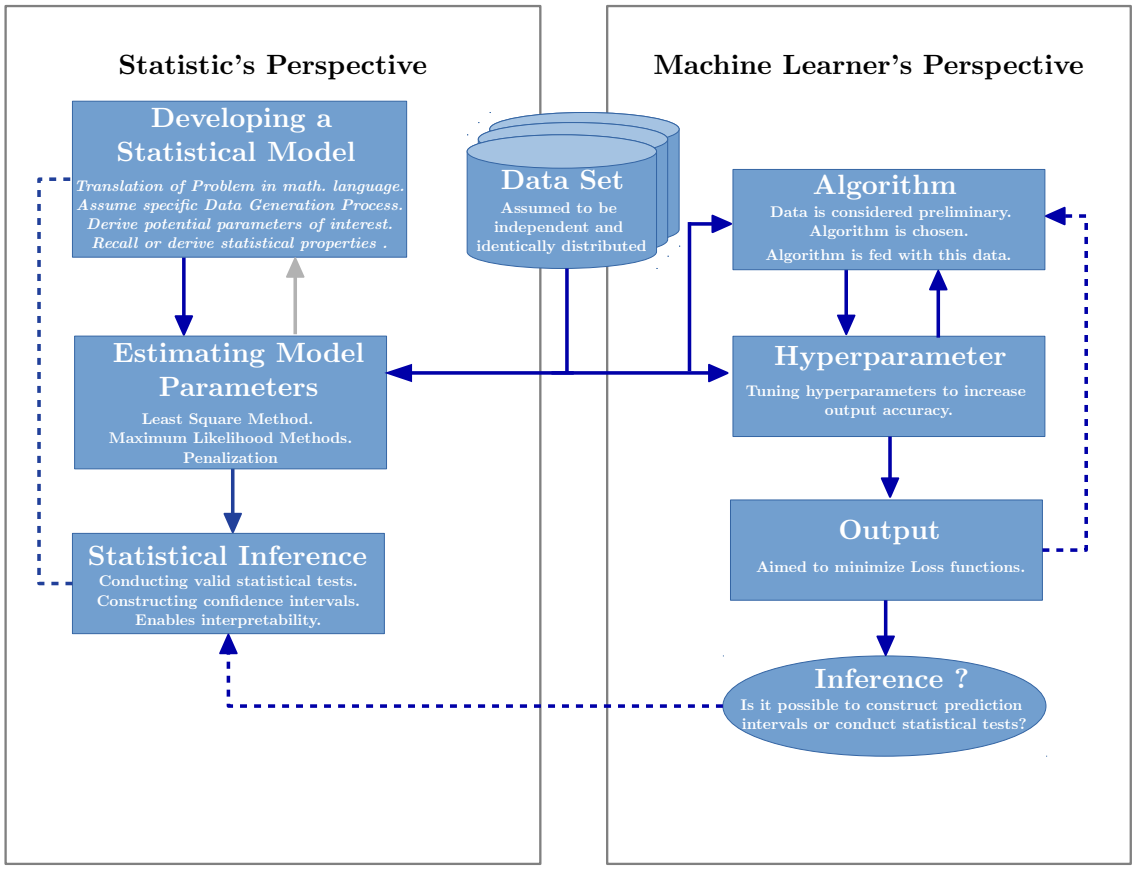


Figure 1: Systematic approach of analyzing data under the perspective of a statistician and a machine learner.

lying algorithm is often considered as a black box procedure, where direct modeling does not take place priorly. Therefore, valid statistical inference in terms of hypothesis tests or even deriving statistical properties such as consistency are rather difficult to tackle.

Recent statistical research has focused on the gap of statistical interpretability of black box procedures in Machine Learning. Luc Devroye, László Györfi, Gábor Lugosi, Gérard Biau, Erwan Scornet, Stefan Wager, Peter Bühlmann, Nicolai Meinshausen or Lucas Mentch are a couple of researchers around the globe, that have focused on topics such as consistency, uncertainty quantification and hypothesis tests of specific Machine Learning algorithms, see for example Meinshausen (2006), Biau et al. (2008),Devroye et al. (2013), Scornet et al. (2015), Scornet (2016), Mentch and Hooker (2016) and Wager and Athey (2018). The aim of this dissertation is to contribute to these tendencies by further enlightening statistical properties of Machine Learning algorithms making them more accessible and interpretable from a statistical perspective. In doing so, we mainly focus on algorithms containing decision trees as key learners, or often referred to as *weak learners* in the Machine Learning jargon. Therein, a more detailed look is taken towards Random Forest models. An ensemble of decision trees, that is based on the *bagging* principle described later in the dissertation aiming to increase model accuracy.

Depending on the underlying data, the Machine Learning approach can attain different forms. To be more precise, suppose one has a collection of independent and identically (iid) distributed random variables $[\mathbf{X}_i^\top, Y_i]^\top \in \mathbb{R}^{p+1}$, $i = 1, \ldots, n$ summarized in a data set of the form $\mathcal{D}_n = \{[\mathbf{X}_i^\top, Y_i]^\top \in \mathbb{R}^{p+1} : i = 1, \ldots, n\}$ and an independent copy $[\mathbf{X}^\top, Y]^\top \in \mathbb{R}^{p+1}$. Several questions can then be raised:

- One might be interested in detecting potential relationships between $\mathbf{X}$ and Y based on $\mathcal{D}_n$ for the purpose of predicting the unknown outcome Y, when new observations $\mathbf{X}' \stackrel{d}{=} \mathbf{X}_1$ are given.
- Instead of only predicting new outcomes based on $\mathcal{D}_n$, one might be interested in correctly specifying the relationship between $\mathbf{X}$ and Y by determining the scale of influence $\mathbf{X}$ has on Y.

For both type of problems, a learning algorithm requires two types of data sets: a set, which is used to feed the algorithm with information. The latter is often referred to as the training set. For evaluating performance measures, one requires observations, that have been *unseen* so far for the underlying algorithm. From a statistical perspective, this overcomes the initial problem of bias due to selection. The separation is often conducted using various forms of cross-validation procedures such as the k-fold cross-validation or the jackknife method. Furthermore, depending on the scale of the measurement, the learning problem can be different. For known outcomes Y in the training set, the learning problem is called a *supervised learning problem*, which can be separated into two types:

- In case of Y being interval or ratio scaled, the relationship between $\mathbf{X}$ and Y can be considered as a regression problem.
- In case of Y being either ordinal or nominal, the relationship aimed to find between $\mathbf{X}$ and Y can be considered as a classification problem.

Both type of problems can be faced with difficulties in the modeling phase, when the information in $\mathbf{X} \in \mathbb{R}^p$, or often referred to as features, are of mixed type, i.e. they contain both, nominal or ordinal scaled data and metric data. In case of Y being unknown in the training phase, the learning problem is called *unsupervised* and falls under the category of clustering. Therefore, the demands on the algorithm used in the Machine Learning approach can be challenging: It should be able to treat mixed type data, for both, regression and classification issues as well as cluster methods. The Random Forest method is one possible solution to these type of questions, especially for supervised learning.

A common problem during the data collection process is the presence of missing values, i.e. some of the observations in the training set are not observable due to different reasons. This can be challenging for both, regression and classification problems, if missing values occur in the features of the underlying learning problem. Therefore, one part of this dissertation aims to tackle the problem of partially observed data from a Machine Learner's perspective by considering different imputation techniques, i.e. methods, that predict missing outcomes and evaluate their prediction accuracy. From a statistical modeling perspective, the effect of various Machine Learning techniques on statistical significance tests after imputing missing values is then analyzed. Therein, different statistical relations are discovered that led us to important theoretical research questions such as quantifying uncertainty in Random Forest models and other Machine Learning Algorithms. The issue of quantifying uncertainty can be considered as the theoretical basis of the development of valid statistical inference procedures,

that are also used in missing value problems or the construction of prediction intervals. In addition, we discovered different statistical properties of variable selection procedures involved in the Random Forest method. Therefore, one can separate the dissertation into two parts:

1. A practical part, in which prediction accuracy and valid statistical inference of Random Forest models are considered within the framework of missing values.

2. A theoretical part, in which different mathematical properties such as consistency and unbiasedness of different estimators arising from the Random Forest algorithm are derived.

Both parts are related by the overall research question how uncertainty can be quantified or statistical inference within the Random Forest model can be conducted. This is an important field which requires a deep theoretical analysis of the Random Forest algorithm, but has severe practical consequences. It illuminates various mathematical forces within black box procedures such as the Random Forest and discover potential problems that are scientifically motivated and not mainly driven by intuition.

The thesis is structured as follows: In the first chapter, we give a brief summary of theoretical results in regression and learning problems. Therein, we make a clear distinction between regression and classification problems and motivate it from a mathematical perspective. In addition, the principles of Bagging and Boosting are introduced, that will be required for understanding the Random Forest method, or other regression and classification methods used in this dissertation. A separate chapter is devoted to the Random Forest method including important theoretical results discovered so far for this learning algorithm. In the second Chapter, we also give a deeper insight of possible implications our results might have from a theoretical and methodological perspective. In addition, the underlying research questions of the articles $(P3)$ and $(P4)$ listed below are thoroughly motivated. The results can be found in Section 2.2. Note that the latter section does not cover a detailed summary of the articles. Instead, it extends the work of these articles leading to the preparation of additional papers yet to be published. In Chapter 3, we discuss missing value problems and explore ways, how these can be considered as either a regression or classification problem, which require the knowledge summarized in Chapter 1 and 2. Similarly, we extend our work in Section 3.3 by delivering theoretical guarantees on uncertainty related issues within the missing framework, such as the proof of a vanishing between imputation variance estimator when using the Random Forest as an imputation tool. This work will result in additional publications. The fourth Chapter summarizes the own contributions in the field of missing value imputation, uncertainty quantification and variable selection in Random Forest models. The articles $(P1)$ - $(P4)$ listed below are summarized in detail, where three of them has been published and one has been submitted:

(P1) Ramosaj, B. and Pauly M., *Predicting Missing Values: A comparative study on non-parametric approaches for imputation.*, Computational Statistics, 34 (4): $1741 - 1764$, 2019.

(P2) Ramosaj B., Amro L. and Pauly M., *A cautionary tale on using imputation methods for inference in matched pairs design.*, Bioinformatics, 2020, doi:10.1093/bioinformatics/btaa082.

(P3) Ramosaj, B. and Pauly M., *Consistent estimation of residual variance with random forest Out-Of-Bag errors.*, Statistics and Probability Letters, 151: $49 - 57$, 2019.

(P4) Ramosaj B. and Pauly M., *Asymptotic Unbiasedness of Permutation Importance in Random Forest Models.* arXiv preprint arXiv:1912.03306.

The four articles are included in the Appendix part (Appendix A) of the dissertation together with the supplementary material they have been published resp. submitted. The following diagram shows the timely development of the four articles together with a distinction between the practical and theoretical dimension making the integration of the four contributions into the thesis more accessible, while emphasizing potential relationships between them.

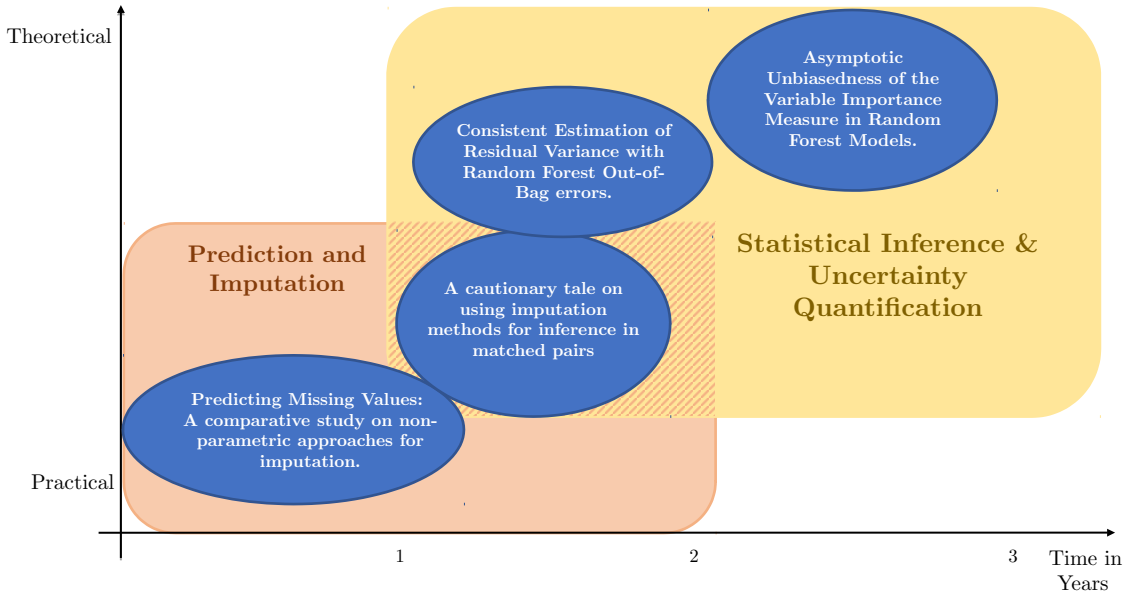


Figure 2: Timely development of the articles considered in this work together with their integration into the wole context including potential thematic dependencies.

Before digging deeper into the thesis, we want to stress out that vectors are indicated as bold, i.e. $\mathbf{x} = [x_1, \ldots, x_p]^\top \in \mathbb{R}^p$ indicates a p-dimensional vector, while x is simply a scalar value, i.e. $x \in \mathbb{R}$. Furthermore, slight notational deviations might exists between the thesis and the annexed articles $(P1)$ - $(P4)$ due to the time differences the articles have been written resp. published. For example in article $(P1)$, $\hat{\Gamma}(\mathbf{x}; \Theta_1)$ refers to a single tree within the Random Forest ensemble trained on $\mathcal{D}_n$, whereas in this thesis, we denoted the latter as $m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_1, \mathcal{D}_n)$. Based on the context, however, it should be clear what is meant while new notations have been introduced throughout the thesis.

Chapter 1

Supervised Learning Problems

Supervised learning accounts for the process of extracting information from a set of observations, mainly for the purpose of prediction, where the variable of interest, or simply the learning signal, is available. One usually distinguishes between classification and regression learning problems. Classification problems arise in various practical fields such as in image recognition, where the classification task is to recognize objects in particular images. Classification is often required in different forms rather than on images, for example, when financial institutions aim to classify people according to their credit worthiness, or in biomedical research, where the aim is to predict whether patients will suffer from a specific disease (see e.g. Keysers et al. (2007), Angelini et al. (2008) and Long et al. (2017)). Regression analysis is slightly different: instead of classifying objects or patients into given classes, i.e. assigning them to a finite set $\{1, \dots, K\}, K \in \mathbb{N}$, the aim is to predict possibly metric values to specific informations, where the target can attain uncountable or countable and infinitely many values. Predicting the average income based on the educational degree, the social status and other features are practical regression problems in econometrics, see for example Hoogerheide et al. (2012). From a mathematical perspective, both approaches differ based on the response variable, i.e. whether it is metric or not. To be more precise, suppose one has a set of iid random vectors $\mathcal{D}_n = \{[\mathbf{X}_i^\top, Y_i]^\top \in \mathbb{R}^p \times \mathcal{M} : i = 1, \dots, n\}$, $n, p \in \mathbb{N}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, i.e. $[\mathbf{X}_i^\top, Y_i]^\top : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^p \times \mathcal{M}, \mathcal{B}(\mathbb{R}^p) \otimes \mathcal{F}_{\mathcal{M}})$. We denote with $\mathbb{E}_{Y|\mathbf{X}}$ the expectation conditioned on $\mathbf{X}$, i.e. $\mathbb{E}_{Y|\mathbf{X}}[\cdot] = \mathbb{E}[\cdot|\mathbf{X}]$ and assume that $[\mathbf{X}^\top, Y]^\top \in \mathbb{R}^p \times \mathcal{M}$ is an independent copy of $[\mathbf{X}_1^\top, Y_1]^\top$.

The aim in both, regression and classification problems is to predict the outcome Y based on the information in $\mathbf{X}$ or to find a potential relationship $\hookrightarrow$, such that $\mathbf{X} \hookrightarrow Y$. This can be usually conducted by having direct knowledge of the underlying multivariate distribution $\mathbb{P}$. Since in practice one does not have knowledge on $\mathbb{P}$, but instead random vectors given in $\mathcal{D}_n$, one usually estimates the relationship $\hookrightarrow$ based on $\mathcal{D}_n$. The difference between regression and classification, however, can be extracted from the support of the random variable Y described by $\mathcal{M}$. In case of $\mathcal{M}$ being finite countable, one usually refers to the learning problem as a classification problem, whereas if $\mathcal{M}$ is either infinitely countable or uncountable, the underlying learning problem is a regression problem. This has sever effects on both, practical and theoretical approaches learning problems are going to be tackled. Purposely, we did not yet specify the type of relation $\hookrightarrow$ one aims to find between $\mathbf{X}$ and Y, since this can cover different, yet very extensive fields in statistics. For example, the relation $\hookrightarrow$ might perhaps refer to a functional relationship, which aims to predict new outcomes. However, the relation $\hookrightarrow$ might also be an algorithm with the same purpose, namely prediction. The relation $\hookrightarrow$, can cover also hidden effects, such as detecting the *optimal* subset among the p features, that

truly describe the association to the response. The latter, for example, is known as *variable selection*, and falls under an optimal specification of the relation $\hookrightarrow$.

1.1 Regression

In regression learning problems, the support $\mathcal{M}$ of the response variable Y is any set that is not finitely countable. Most of the time, the relation $\hookrightarrow$ is then of functional nature and the aim is to find a measurable function $m : \mathbb{R}^p \longrightarrow \mathbb{R}$, that can describe Y in the *"best"* way possible. In order to understand what *best* means from a statistical perspective, it is required to have a look on loss functions. Although an official definition of the latter does not exist, we introduce a formal definition of it and usually distinguish between regression and classification losses.

Definition 1.1 (Regression Loss Function)**.** A loss function ψ for a regression problem is a measurable function $\psi : \mathcal{M} \times \mathcal{M} \longrightarrow \mathbb{R}_+$ such that $\psi(t_1, t_1) \leq \psi(t_2, t_3)$, for all $t_1, t_2, t_3 \in \mathcal{M}$.

The above definition is more general, since most of the considered loss functions require *convexity*, in order to ease the solution of specific minimization problems. With convexity, a local minimum results into a global minimum making the solution of an optimization problem easier. We will shortly give some examples of loss functions, that are often used in theory and practice. They are leaned on the examples given in Hastie et al. (2009b) on page 349.

1. The *squared error loss* $\psi(x, y) = (x - y)^2$
2. The *absolute loss* $\psi(x, y) = |x - y|$
3. The *Huber loss* given by $\psi(x, y) = \psi_\delta(x, y) = (y - x)^2 \mathbb{1}\{|y - x| \leq \delta\} + \{2\delta|y - x| - \delta^2\} \cdot \mathbb{1}\{|y - x| > \delta\}$ for some pre-defined $\delta > 0$. The Huber loss is more robust to outliers than the absolute loss or the squared error loss.

Then, considering the function class $\mathcal{G} = \{m : \mathbb{R}^p \longrightarrow \mathbb{R} \quad | \quad \mathbb{E}_{Y|\mathbf{X}}[\psi(m(\mathbf{X}), Y)] < \infty \text{ a.s.}\}$, *best* from a statistical perspective means that finding a function $m^* \in \mathcal{G}$, such that

$$m^* = \arg\min_{m \in \mathcal{G}} \mathbb{E}_{Y|\mathbf{X}}[\psi(m(\mathbf{X}), Y)]. \tag{1.1}$$

Note that the solution to (1.1) does not always admit an analytically tractable form. This clearly depends on the distributional law $\mathbb{P}$ of the random vector $[\mathbf{X}^\top, Y]^\top$ and the considered loss function ψ. However, for the squared error loss and the absolute loss, the solution m^* is a well-known theoretical result. Recalling the proof in Györfi et al. (2006), on page 2, one can obtain the following solution:

$$m^*(\mathbf{x}) = \begin{cases} \mathbb{E}[Y|\mathbf{X} = \mathbf{x}], & \text{if } \psi \text{ is the } \textit{squared error loss}, \\ Med[Y|\mathbf{X} = \mathbf{x}], & \text{if } \psi \text{ is the } \textit{absolute loss}. \end{cases} \tag{1.2}$$

In theory and practice, the squared error loss has prevailed due to its smoothness and differentiability property. In addition, the squared error loss simplifies later mathematical results. Therefore, unless not otherwise stated, we will assume that $\psi(x, y) = (x - y)^2$, such that $m^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. Since $\mathbb{P}$ is in practice often unknown, the optimal solution m^* to

equation (1.1) can only be approximated using the knowledge given by $\mathcal{D}_n$. Hence, we refer to m^* in the regression case as the *optimal predictor* or the *theoretical regression function*, and denote with m_n any estimator of m^*, that is based on the set $\mathcal{D}_n$ for approximation. The determination of m_n can happen differently. Most of the time, especially when prior knowledge on the data generating process $\mathbb{P}$ is available, it is assumed that m^* belongs to some restricted function classes. In that case, the derivation of the approximation m_n further depends on the support $\mathcal{M}$, and the assumption on which function class one restricts its view. This, because simply considering $\mathcal{G}$ does not directly deliver answers to the question, how exactly the functional relationship might look like. For example, if the support of Y, i.e. $\mathcal{M}$, is uncountable and real-valued such that $\mathcal{M} = \mathbb{R}$, one often refer to these regression problems by stating the equation of

$$Y = f(\mathbf{X}) + \epsilon, \tag{1.3}$$

where ϵ is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}[\epsilon|\mathbf{X}] = 0$ and $\mathbb{E}[\epsilon^2|\mathbf{X}] \in (0, \infty)$ almost surely and $f : \mathbb{R}^p \to \mathbb{R}$ is a function from a potentially smaller function class than $\mathcal{G}$ such that one assumes that $m^* = f$. Such a restricted function class might be all linear functions, i.e. $f \in \mathcal{F}_{linear} := \{f : \mathbb{R}^p \to \mathbb{R} | f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}, \boldsymbol{\beta} \in \mathbb{R}^p\}$ and an approximation to m^* is conducted by setting $f_n(\mathbf{x}) = \boldsymbol{\beta}_n^\top \mathbf{x}$, where

$$\boldsymbol{\beta}_n = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \psi(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}).$$

In case of more complex regression problems, for example, when the support is finite or countable, such as e.g. $\mathcal{M} = \mathbb{N}$, the relation between Y and $\mathbf{X}$ is not written in the form of (1.3), but is rather motivated by specific link-functions $g : \mathbb{R} \to \mathbb{R}$, that are closely related to the support given by $\mathcal{M}$. That is, one usually sets

$$\tilde{f}(\mathbf{X}) = g(m^*(\mathbf{X})). \tag{1.4}$$

This can be re-expressed in terms of $\mathbb{E}[Y|\mathbf{X}] = (g^{-1} \circ \tilde{f})(\mathbf{X})$. Relating this result to the minimization problem as given in (1.1), while focusing on the squared error loss $\psi(x, y) = (x-y)^2$, one can see that $g^{-1} \circ \tilde{f}$ is its optimal solution. Since g is chosen depending on the support of Y, the minimization in (1.1) is usually conducted over the function class to which $\tilde{f}$ belongs. If $\tilde{f} \in \mathcal{F}_{linear}$, then this is nothing else than a *generalized linear model* as given in McCullagh and Nelder (1989). Having a countable support of Y such as $\mathcal{M} = \mathbb{N}$ for example, then the link function g is given by $g(x) = \log(x)$ and can be found in McCullagh and Nelder (1989) on page 28 under the term *Poisson regression*.

Evaluating whether an estimator f_n is *good* enough is then a central aspect in modern statistical learning problems, which is addressed by the term *consistency*. Since several definitions of consistency exist, we will mainly focus, similarly to Györfi et al. (2006), to the quantity

$$R_n := \mathbb{E}[(m_n(\mathbf{X}) - m^*(\mathbf{X}))^2 | \mathcal{D}_n], \tag{1.5}$$

which is nothing else than the L_2 error. Since R_n is random, different notions of consistency can be established. Mainly following the definitions in Györfi et al. (2006) on page 13, one can distinguish between the following:

Definition 1.2. (Consistency)

1. A sequence of regression function estimates $\{m_n\}$ is called **weakly consistent for a certain distribution of** $[\mathbf{X}^\top, Y]^\top$, if $\lim_{n\to\infty} \mathbb{E}[R_n] = 0$.

2. A sequence of regression function estimates $\{m_n\}$ is called **strongly consistent for a certain distribution of** $[\mathbf{X}^\top, Y]^\top$, if $\mathbb{P}[\lim_{n\to\infty} R_n = 0] = 1$.

The above definition refers to consistency properties, that rely on specific classes of distributions, such that the property 1. or 2. of Definition 1.2 might hold for a specific distribution, but not for all of them. In non-parametric regression, i.e. where the distribution of $[\mathbf{X}^\top, Y]^\top$ cannot be directly parametrized due to its complexity, it is important to which extent consistency properties of estimators m_n can be extended for all distributions, i.e. delivering a kind of *global* or *universal* consistency definition. Therefore, leaned on the definitions given in Györfi et al. (2006), page 13, one can set the following definition of *universal consistency.*

Definition 1.3. (Universal Consistency)

1. A sequence of regression function estimates $\{m_n\}$ is called **weakly universal consistent**, if it is weakly consistent for all distributions of $[\mathbf{X}^\top, Y]^\top$, with $\mathbb{E}[Y^2] < \infty$.

2. A sequence of regression function estimates $\{m_n\}$ is called **strongly universal consistent**, if it is strongly consistent for all distributions of $[\mathbf{X}^\top, Y]^\top$, with $\mathbb{E}[Y^2] < \infty$.

In this thesis, however, more emphasis is placed on consistency properties of the form given in Definition 1.2 for more special classes of distributions, since a special focus is placed on Random Forest models, for which, in order to obtain consistency, assumptions have to be made touching on the distribution of the pair $[\mathbf{X}^\top, Y]^\top$.

1.2 Classification

In classification problems, the support $\mathcal{M}$ is a countable set with finite cardinality $K = card(\mathcal{M})$. In case of $K > 2$, the underlying classification problem is called a *multi-class* classification. Similary to regression, the aim in classification is to predict the response class Y, based on the features $\mathbf{X}$ using the relation $\hookrightarrow$. However, the relation one aims to extract can be quite different to the regression case. In classification, one usually distinguishes between

(C1) The process of assigning class labels for feature inputs $\mathbf{X}$ using a classifier g, i.e. a measurable function $g : \mathbb{R}^p \longrightarrow \mathcal{M}$. In that case, the relation $\hookrightarrow$ is the classifier g.

(C2) The process of estimating class probabilities, i.e. instead of finding an explicit function g that assigns class labels, one is interested in extracting information regarding the probability vector $\boldsymbol{p}(\mathbf{x}) = [p_{\ell_1}(\mathbf{x}), \ldots, p_{\ell_K}(\mathbf{x})]^\top \in [0,1]^K$, where $p_{\ell_k}(\mathbf{x}) = \mathbb{P}[Y = \ell_k | \mathbf{X} = \mathbf{x}]$, with $\mathcal{M} = \{\ell_1, \ldots, \ell_K\}$. In that case, the relation $\hookrightarrow$ is the probability vector as a mapping $\boldsymbol{p} : \mathbb{R}^p \longrightarrow [0,1]^K$.

As mentioned in Andreas et al. (2005), the classification problem as given in (C1). is often addressed by Machine Learners, whereas class probability estimation is usually considered by statisticians. Regarding the first point, it is interesting to know , which measurable function achieves best results among a class of classifier $\mathcal{C} := \{g | g : \mathbb{R}^p \to \mathcal{M} \text{ is measurable}\}$. Differently to the regression case, at this stage of finding a *"best"* classifier, one usually does not consider a variety of loss functions. Instead, one makes use of the advantage that $\mathcal{M}$ has finite cardinality and is countable. Therefore, *best* in the sense of the classification problem given in (C1). is nothing else than a classifier, that makes minimal errors, i.e. where

$g(\mathbf{X}) \neq Y$ holds in some minimal sense. Since the event $\{\omega \in \Omega : g(\mathbf{X}(\omega)) \neq Y(\omega)\}$ accounts for uncertainty, one usually aims to find the optimal classifier g^* by setting

$$\begin{aligned} g_{opt} &= \arg\min_{g \in \mathcal{C}} \mathbb{E}[\mathbb{1}\{g(\mathbf{X}) \neq Y\}] \\ &= \arg\min_{g \in \mathcal{C}} \mathbb{P}[g(\mathbf{X}) \neq Y]. \end{aligned} \tag{1.6}$$

The underlying error considered here is nothing else than the *misclassification error*, i.e. the mapping $\psi(x, y) = \mathbb{1}\{x \neq y\}$, or in practice also referred to as the *0-1 loss*. At this stage of the problem, different loss functions are usually not considered. Explicitly stating g_{opt} can be conducted using the *Bayes rule* and basic probability theory, see e.g. Devroye et al. (2013) on page 10. Therein, the authors delivered a theoretical proof for the solution of (1.6) for the binary classification problem, i.e. when $K = 2$. For completeness, we extend the result to the multi-class problem and state that as a theorem. The proof can be found at the end of this section:

Theorem 1.1. *(Bayes Classifier) The optimal solution to the problem given in* (1.6) *is given by*

$$g_{opt}(\mathbf{x}) = g^*(\mathbf{x}) := \arg\max_{k \in \mathcal{M}} \mathbb{P}[Y = k | \mathbf{X} = \mathbf{x}].$$

The error probability of g^* as mentioned in Devroye et al. (2013), page 2 is then given by

$$L^* = \mathbb{P}[g^*(\mathbf{X}) \neq Y]. \tag{1.7}$$

Since g^* clearly depends on the distribution of the pair $[\mathbf{X}^\top, Y]^\top$, it is in practice often unknown, such that g^* remains unknown most of the time. Hence, the aim is to find an estimate g_n based on the data $\mathcal{D}_n$, that comes close enough to the response Y, given the data at hand $\mathcal{D}_n$, i.e. one aims to compute the performance of g_n by considering

$$L_n := \mathbb{P}[g_n(\mathbf{X}) \neq Y | \mathcal{D}_n] = \mathbb{E}[\mathbb{1}\{g_n(\mathbf{X}) \neq Y\} | \mathcal{D}_n], \tag{1.8}$$

which is clearly a random variable. Hence, performance assessment can be understood for a fixed data set $\mathcal{D}_n$, instead of an avaraged data set $\mathcal{D}_n$, since performance is a data specific issue rather than an averaged one. Similar to the regression case, we can call a sequence of classifier $\{g_n : n \geq 0\}$, i.e. a rule, as *good*, if it is consistent according to the following definition. The latter can be recalled in Devroye et al. (2013) on page 91:

Definition 1.4. (Consistency for Rules)

1. A rule $\{g_n : n \geq 1\}$ is called **weakly consistent for a certain distribution of** $[\mathbf{X}^\top, Y]^\top$, if $\lim_{n \to \infty} \mathbb{E}[L_n] = L^*$.

2. A rule $\{g_n : n \geq 1\}$ is called **strongly consistent for a certain distribution of** $[\mathbf{X}^\top, Y]^\top$, if $\mathbb{P}[\lim_{n \to \infty} L_n = L^*] = 1$.

3. A rule $\{g_n : n \geq 1\}$ is called **universally weakly consistent**, if it is weakly consistent for any distribution of $[\mathbf{X}^\top, Y]^\top$.

4. A rule $\{g_n : n \geq 1\}$ is called **universally strongly consistent**, if it is strongly consistent for any distribution of $[\mathbf{X}^\top, Y]^\top$.

The result in Theorem 1.1 also reveal some insights into the connection between the problems given in (C1). and (C2). Knowledge of the kind as given in (C2). is usually stronger than $(C1)$., since obtaining an estimate $\hat{\boldsymbol{p}}_n$ of $\boldsymbol{p}$ based on the set $\mathcal{D}_n$ can lead to the construction of a classifier g_n based on Theorem 1.1 using the plug-in principle. However, the extraction of an estimate $\hat{\boldsymbol{p}}_n$ or $\boldsymbol{p}$ from a classifier g_n or g is non-trivial.

Now, introducing loss functions into the framework of classification is usually a two-folded problem: Since classification problems can usually be considered from both perspectives, $(C1)$. and $(C2)$., one has to distinguish between losses in both cases. Therefore, let us start with the problem of estimating class probabilities. In that case, note that a multi-class problem can be transformed in such a way, that the inital support $\mathcal{M}$ can be transformed into an alternative support $\mathcal{M}_p$, that describes probability distributions. To be more precise, every class label $\ell_k \in \mathcal{M}$ for $k \in \{1, \ldots, K\}$ can be represented by the K-dimensional canonical vector $\mathbf{e}_k = [0, \ldots, 0, 1, 0, \ldots, 0]^\top \in \{0,1\}^K$, where the integer 1 is placed at the k-th position. This refers to saying that class label ℓ_k occurs with probability 1. Then the set $\mathcal{M}_p$ is of the form $\{\mathbf{e}_k : k = 1, \ldots, K\}$. Having in mind this situation, i.e. that the response Y is not a class label anymore, but a probability distribution, one can set up so called *probability losses* or shortly, *p-losses*. Therefore, similarly to Andreas et al. (2005), page 1, we introduce a novel and more general definition of probability losses:

Definition 1.5. (Loss Function for Probabilities) A loss function for probabilities is a measurable function ψ, such that $\psi : \mathcal{P}_1 \times \mathcal{P}_2 \longrightarrow \mathbb{R}_+$, where $\mathcal{P}_1 = \{\mathbf{y} \in \{0,1\}^K : \|\mathbf{y}\|_1 = 1\}$ and $\mathcal{P}_2 = \{\boldsymbol{p} \in [0,1]^K : ||\boldsymbol{p}||_1 = 1\}$ with $\psi(\mathbf{e}_k, \mathbf{e}_k) \leq \psi(\mathbf{y}, \mathbf{p})$ for all $[\mathbf{y}^\top, \boldsymbol{p}^\top]^\top \in \mathcal{P}_1 \times \mathcal{P}_2$ and $k \in \{1, \ldots, K\}$.

The choice of loss functions during the training of a classifier is important, since this can change the quality of the underlying classifier given the training set $\mathcal{D}_n$. For example in neural networks, where a famous practical problem is the recognition of handwritten digits, the choice of a suitable probability loss function at the output layer might be crucial for the performance of the whole network. Therefore, we will give short examples of probability losses, that are based on Andreas et al. (2005), page 7 and 8.

1. The *log loss* or also called as the *Kullback Leibler information*. It has the form $\psi(\mathbf{y}, \widehat{\boldsymbol{p}}_n(\mathbf{x})) = -\sum_{k=1}^{K} y_k \cdot \log(\hat{p}_{k,n}(\mathbf{x}))$, where $\hat{p}_{k,n}(\mathbf{x}) = \mathbb{P}[Y = \widehat{\ell_k|\mathbf{X} = \mathbf{x}}]$ is the k-th component of the class probability estimator $\widehat{\boldsymbol{p}}_n(\mathbf{x})$, which estimates the true class probability $\boldsymbol{p}(\mathbf{x})$ based on the training set $\mathcal{D}_n$ and $\mathbf{y} = [y_1, \ldots, y_K]^\top$.

2. The *squared error loss* given by $\psi(\mathbf{y}, \boldsymbol{p}) = ||\mathbf{y} - \boldsymbol{p}||_2^2$.

3. The *binary boosting loss* is defined for a binary classification problem and is given by $\psi(\mathbf{y}, \hat{\boldsymbol{p}}(\mathbf{x})) = y_1 \cdot \left(\frac{1-\hat{p}_1(\mathbf{x})}{\hat{p}_1(\mathbf{x})}\right)^{1/2} + y_2 \cdot \left(\frac{\hat{p}_1(\mathbf{x})}{1-\hat{p}_1(\mathbf{x})}\right)^{1/2}$, where $\hat{p}_1(\mathbf{x}) = \mathbb{P}[Y = \widehat{\ell_1|\mathbf{X} = \mathbf{x}}]$ is an estimate of $p_1(\mathbf{x}) = 1$.

The construction of classifiers, however, is not always conducted by using plug-in estimates for g^* as given in Theorem 1.1. Different algorithms do exist, that are not primarily developed based on the prediction of class probabilities. Such examples are boosting machines or support vector machines, to name a few. Regarding the class of classifiers, that do not directly rely on the estimation of class probabilities, the application of Definition 1.5 is not directly possible. Usually one makes use of link-functions to connect the class assignment of such classifiers to probability estimators. Since direct loss functions on this class of classifier clearly depend on

the domain of the considered function involved in the classifier, we drop a formal definition of it and refer to Andreas et al. (2005) on page 2 for an *informal* definition, called *F-losses*. Examples based on Hastie et al. (2009b), however, are shortly given. For those, it is assumed without loss of generality, that $\mathcal{M} = \{-1, 1\}$, i.e. the classification problem is binary.

1. The *0 - 1 loss*, or also called as the misclassification, is given by $\psi_f(\mathbf{x}, y) = \mathbb{1}\{y \cdot f(\mathbf{x}) < 0\}$.
2. The *exponential loss* refers to errors made by predictions f in binary classification problems. It is given by $\psi_f(\mathbf{x}, y) = \exp\{-y \cdot f(\mathbf{x})\}$ and f is the corresponding functional prediction, i.e. it is not necessarily the class prediction. As mentioned in Hastie et al. (2009b) on page 347, ψ_f can be considered as a monotone continuous approximation of the 0-1 loss.
3. The *binomial deviance loss* is given by $\psi_f(\mathbf{x}, y) = \log(1 + \exp\{-y \cdot f(\mathbf{x})\})$.
4. The *support vector loss* is given by $\psi_f(\mathbf{x}, y) = \max\{0, (1 - y \cdot f(\mathbf{x}))\}$.

In this work, both, classification and regression problems are considered from a practical perspective. In order to understand the correct usage of implemented functions in the statistical software R, it was important to have a deeper insight into the involvement of loss functions in both, regression and classification. However, from a theoretical perspective, classification problems have been a minor part of this thesis, for which regression was the main issue.

1.3 Bagging

The procedure of **Bagging** (**b**ootstrapped **agg**regat**ing**) was initially developed by Breiman (1996b) and aimed to reduce the variance of *unstable* estimators of predictors in order to increase predictive accuracy of various models. In the regression context, a predictor might refer to the quantity $\theta(\mathbf{x}) = m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$, whereas in the classification context, the predictor might refer to the classification of $\mathbf{x}$, into elements of $\mathcal{M}$, i.e. $\theta(\mathbf{x}) = g^*(\mathbf{x})$, where $g^* \in \mathcal{C}$ is the Bayes classifier. Suppose now that an estimator $\hat{\theta}_n(\mathbf{x})$ for the predictor $\theta(\mathbf{x})$ based on the training set $\mathcal{D}_n$ is given. Similar to Breiman (1996b), we first pay attention to the regression case and assume that prediction accuracy is measured by considering the squared error loss $\psi(x, y) = (x - y)^2$. Furthermore, denote with $\bar{\theta}(\mathbf{x}) = \mathbb{E}[\hat{\theta}_n(\mathbf{x})]$ the aggregated predictor of $\hat{\theta}_n(\mathbf{x})$. Then, Breiman (1996b) argued that the aggregated version of $\bar{\theta}(\mathbf{x})$ leads to a smaller mean squared error than the estimator $\hat{\theta}_n(\mathbf{x})$. This results from applying for a fixed point $[\mathbf{x}^\top, y]^\top \in \mathbb{R}^p \times \mathcal{M}$ Jensen's inequality to the following quantity:

$$\begin{aligned}
\mathbb{E}[(y - \hat{\theta}_n(\mathbf{x}))^2] &= y^2 - 2y\mathbb{E}[\hat{\theta}_n(\mathbf{x})] + \mathbb{E}[\hat{\theta}_n(\mathbf{x})^2] \\
&\geq y^2 - 2y\mathbb{E}[\hat{\theta}_n(\mathbf{x})] + \mathbb{E}[\hat{\theta}_n(\mathbf{x})]^2 \\
&= \mathbb{E}[(y - \mathbb{E}[\hat{\theta}_n(\mathbf{x})])^2].
\end{aligned}$$

Hence, we have for the corresponding mean squared error

$$MSE[\hat{\theta}_n(\mathbf{X})] = \mathbb{E}[(Y - \hat{\theta}_n(\mathbf{X}))^2] \geq \mathbb{E}[(Y - \bar{\theta}(\mathbf{X}))^2] = MSE[\bar{\theta}(\mathbf{X})].$$

As mentioned in Breiman (1996b), the decrease in mean squared error clearly depends on the difference between $\mathbb{E}[\hat{\theta}_n(\mathbf{x})^2]$ and $\mathbb{E}[\hat{\theta}_n(\mathbf{x})]^2$, such that in case of no difference, i.e. $\mathbb{E}[\hat{\theta}_n(\mathbf{x})^2] = \mathbb{E}[\hat{\theta}_n(\mathbf{x})]^2$, the estimator is called *stable*. This resulted into the idea that any estimator of

predictors can be improved in the sense of a squared error loss, if one considers the aggregated version of it, i.e. $\bar{\theta}(\mathbf{x})$. The issue in $\bar{\theta}$ is that knowledge on the data generating process is required, i.e. on $\mathbb{P}$, in order to compute the expectation. Therefore, an approximation of the latter quantity was proposed, that should mimic the behavior of $\bar{\theta}$ using bootstrapping, i.e. the general principle of *bagging*. It can be summarized in three steps:

- Draw B independent bootstrap samples of size n with replacement from $\mathcal{D}_n$ denoted by $\mathcal{D}^*_{n,1}, \dots, \mathcal{D}^*_{n,B}$.
- Compute B bootstrap estimators $\hat{\theta}^*_{n,1}(\mathbf{x}), \dots, \hat{\theta}^*_{n,B}(\mathbf{x})$ such that $\hat{\theta}^*_{n,j}(\mathbf{x}) = \hat{\theta}_n(\mathbf{x}; \mathcal{D}^*_{n,j})$.
- Aggregate the results by considering $\overline{\theta}^*_{n,B}(\mathbf{x}) = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}^*_{n,j}(\mathbf{x})$.

Note that the sampling strategy is not always restricted to sampling n points with replacement from $\mathcal{D}_n$. Deviations to that do exists, such as Breiman's Random Forest model (Breiman, 2001) or later in this thesis, were we proposed different sampling schemes within the bagging procedure of Random Forest models. The quantity $\overline{\theta}^*_{n,B}(\mathbf{x})$ can be considered as a Monte-Carlo approximation of the optimal, bootstrapped-bagged predictor given by $\hat{\theta}_{n,\infty}(\mathbf{x}) = \mathbb{E}^*[\hat{\theta}^*_{n,1}(\mathbf{x}; \mathcal{D}^*_n)] = \mathbb{E}[\hat{\theta}^*_{n,1}(\mathbf{x}; \mathcal{D}^*_n)|\mathcal{D}_n]$. The expectation is taken over the bootstrap measure $\mathbb{P}^*$. The latter quantity should be preferred, if its extraction is analytically possible, and therefore practically available. Since otherwise, additional bias in $\overline{\theta}^*_{n,B}(\mathbf{x})$ will be included, due to the finite choice of B, i.e. $Bias^*(\overline{\theta}^*_{n,B}(\mathbf{x})) = \overline{\theta}^*_{n,B}(\mathbf{x}) - \hat{\theta}_{n,\infty}(\mathbf{x}) = o_{\mathbb{P}^*}(1)$. For increasing B, this finite-B bias will vanish due to the strong law of large numbers.

Now returning to the classification case, where errors are measured using the $0-1$ loss, i.e. $\psi(x, y) = \mathbb{1}\{x \neq y\}$, the optimal classifier, i.e. the Bayes classifier, results into the lowest error rate, which was proven in the previous section. Then for any fixed point $\mathbf{x} \in \mathbb{R}^p$, one can then deduce that

$$\begin{aligned}
r^*(\mathbf{x}) &= \mathbb{P}[g^*(\mathbf{x}) \neq Y] = 1 - \mathbb{P}[g^*(\mathbf{x}) = Y] \\
&= 1 - \sum_{k=1}^{K} \mathbb{P}[\arg\max_{i=1,\dots,k} \mathbb{P}[Y = \ell_i | \mathbf{X} = \mathbf{x}] = \ell_k | Y = \ell_k, \mathbf{X} = \mathbf{x}] \cdot \mathbb{P}[Y = \ell_k | \mathbf{X} = \mathbf{x}] \\
&= 1 - \sum_{k=1}^{K} \mathbb{1}\{\arg\max_{i=1,\dots,k} P(i|\mathbf{x}) = k\} P(k|\mathbf{x}) \\
&= 1 - \max_{k=1,\dots,K} P(k|\mathbf{x}),
\end{aligned}$$

where $P(k|\mathbf{x}) = \mathbb{P}[Y = \ell_k | \mathbf{X} = \mathbf{x}]$ with $k = 1, \dots K$ such that the Bayes misclassification rate can then be rewritten into $L^* = \mathbb{E}[\psi(Y, g^*(\mathbf{X}))] = \mathbb{P}[Y \neq g^*(\mathbf{X})]) = 1 - \mathbb{E}[\max_{k=1,\dots,K} P(k|\mathbf{X})]$. The above computations have been conducted by our own making some additional clarifications in comparison to the work of Breiman (1996b). The latter has been mainly used for motivational purposes. Based on these results and similar observations in Breiman (1996b) on page 131, the following inequality can be obtained

$$L^* = \mathbb{E}[r^*(\mathbf{X})] = 1 - \mathbb{E}[\max_{k=1,\dots,K} P(k|\mathbf{X})] \leq \mathbb{P}[\hat{\theta}_n(\mathbf{X}) \neq Y] = \mathbb{E}[\psi(Y, \hat{\theta}_n(\mathbf{X}))] = \mathbb{E}[L_n].$$

Breiman proposed for classification problems an aggregated predictor of the form $\bar{\theta}(\mathbf{x}) = \arg\max_{k=1,\dots,K} Q(k|\mathbf{x})$, where $Q(k|\mathbf{x}) = \mathbb{P}[\hat{\theta}_n(\mathbf{X}) = \ell_k | \mathbf{X} = \mathbf{x}]$, with $\mathcal{M} = \{\ell_1, \dots, \ell_K\}$. Then,

the probability of falsely classifying a fixed input $\mathbf{x}$ for the aggregated predictor is given by

$$\begin{aligned}\mathbb{P}[\bar{\theta}(\mathbf{x}) \neq Y] &= 1 - \sum_{k=1}^{K} \mathbb{P}[\bar{\theta}(\mathbf{x}) = \ell_k | Y = \ell_k, \mathbf{X} = \mathbf{x}] \cdot \mathbb{P}[Y = \ell_k | \mathbf{X} = \mathbf{x}] \\ &= 1 - \sum_{k=1}^{K} \mathbb{1}\{\arg \max_{j=1,\ldots,K} Q(j|\mathbf{x}) = k\} \cdot \mathbb{P}[Y = \ell_k | \mathbf{X} = \mathbf{x}] \\ &= 1 - \max_{k=1,\ldots,K} P(k|\mathbf{x}),\end{aligned}$$

if the classifier $\hat{\theta}_n(\mathbf{x})$ is *order-correct*, i.e. $\arg \max_{k=1,\ldots,K} Q(k|\mathbf{x}) = \arg \max_{k=1,\ldots,K} P(k|\mathbf{x})$ as mentioned in Breiman (1996b). Verifying order-correctness based on a data set $\mathcal{D}_n$ is a delicate issue, since knowledge on the true conditional distribution $\mathbb{P}[Y = \ell_k|\mathbf{X} = \mathbf{x}]$ and $\mathbb{P}[\hat{\theta}_n(\mathbf{X}) = \ell_k|\mathbf{X} = \mathbf{x}]$ is required. There is no literature for finding appropriate strategies for *testing* or exploring alternative conditions for order-correctness of a classifier. Under the case of an order-correct classifier, one receives for the aggregated predictor $\bar{\theta}(\mathbf{x})$ the same misclassification rate as for the Bayes classifier, i.e.

$$\mathbb{E}[\psi(\bar{\theta}(\mathbf{X}), Y)] = \mathbb{E}[1 - \max_{k=1,\ldots,K} P(k|\mathbf{X})] = L^* \leq \mathbb{E}[\psi(\hat{\theta}_n(\mathbf{X}), Y)]. \tag{1.9}$$

That means, any estimator $\hat{\theta}_n(\mathbf{x})$ of the predictor $\theta(\mathbf{x})$ can be improved, if the estimated predictor $\hat{\theta}_n(\mathbf{x})$ is order-correct and one considers the aggregated predictor $\bar{\theta}(\mathbf{x})$. This is different to the regression case, where the principle of bagging always worked in the sense of resulting into lower expected losses. For classification problems, bagging might result into even worse classifiers, if $\hat{\theta}_n(\mathbf{x})$ is not order-correct for almost all points $\mathbf{x}$. Hence, introducing bagging into classification problems does not always lead to better prediction performance. This has also been mentioned in Breiman (1996b), on page 131. Note that the definition of order-correctness of a classifier is not directly related to the notion of stability according to Breiman (1996b). We introduced the term *order-correctness* for a classifier, that can be obtained when the underlying theoretical distribution generating $\mathcal{D}_n$ is known. Hence, *order-correctnes* is a property, that is based on $\mathbb{P}[\hat{\theta}_n(\mathbf{X}) = \ell_k|\mathbf{X} = \mathbf{x}]$, whereas *stability* is based on $\hat{\theta}_n(\mathbf{X})$ as an estimator. Therefore, we might be in possession of a stable classifier, which is not order correct. However, if a classifier $\hat{\theta}_n(\mathbf{x})$ is not stable, it makes sense to use the bagging principle to stabilize it. The natural question arises why? The reason is that *stability* as defined in Bühlmann and Yu (2002) is a weaker property compared to (pointwise) consistency. However, if a classifier is consistent, then it is also *stable* and in addition, (pointwise) consistency implies *order-correctness*. Therefore, bagging unstable classifiers is a good idea.

Similarly to the regression case, a direct extraction of the bagged predictor $\bar{\theta}(\mathbf{x})$ is often not possible, since a direct knowledge on the distributional law is required. Therefore, one approximates similarly to the regression case the aggregated predictor $\bar{\theta}(\mathbf{x})$ by introducing bootstrapping schemes, but aggregation is then put differently, i.e. we obtain the following procedure:

- Draw B independent bootstrap samples of size n with replacement from $\mathcal{D}_n$ denoted by $\mathcal{D}^*_{n,1}, \ldots, \mathcal{D}^*_{n,B}$.
- Compute B bootstrap classifier $\hat{\theta}^*_{n,1}(\mathbf{x}), \ldots, \hat{\theta}^*_{n,B}(\mathbf{x})$ such that $\hat{\theta}^*_{n,j}(\mathbf{x}) = \hat{\theta}_n(\mathbf{x}; \mathcal{D}^*_{n,j})$.
- Aggregate the results by considering $\hat{\theta}^{*,+}_{n,B}(\mathbf{x}) = Mode(\hat{\theta}^*_{n,1}(\mathbf{x}), \ldots, \hat{\theta}^*_{n,B}(\mathbf{x}))$. In case of ambiguities, one draws randomly an element among potential mode-candidates.

Similary, $\hat{\theta}^{*,+}_{n,B}$ can be considered as a Monte-Carlo approximation of $\arg\max_{k=1,\dots,K} \mathbb{P}^*[\hat{\theta}^*_{n,1}(\mathbf{X}) = \ell_k|\mathbf{X}=\mathbf{x}] =: \arg\max_{k=1,\dots,K} Q^*(k|\mathbf{x})$.

The reason for a more detailed motivation of bagging procedures is its important role in stabilizing predictors such as classification and regression trees (CART), see e.g. Breiman (1996a), Breiman (1996b), Breiman (2001), Bühlmann and Yu (2002) and Sutton (2005). The latter can be considered as an algorithm, which separates the feature space into hyper-rectangular regions and assigns to each region a constant value. That is, a CART predictor can be formalized as

$$\hat{\theta}_n(\mathbf{x}) = \sum_{\ell=1}^{J} \hat{c}_{n,\ell} \cdot \mathbb{1}\{\mathbf{x} \in \hat{A}_{n,\ell}\}, \tag{1.10}$$

where $\hat{A}_{n,\ell} = \bigotimes_{j=1}^{p} [a^{(j)}_{n,\ell}, b^{(j)}_{n,\ell})$, with $a^{(i)}_{n,\ell} < b^{(i)}_{n,\ell}$ such that $\mathbb{R}^p \supseteq supp(\mathbf{X}) = \dot{\bigcup}_{\ell=1}^{J} \hat{A}_{n,\ell}$, and $J \in \mathbb{N}$ being the number of leaves in a CART tree. We refer to Subsection 1.5 for a more detailed description of CART. The estimated constant $\hat{c}_{n,\ell}$ is then, depending on the underlying learning problem, either the mean of all responses Y_i or the mode of all Y_i, for which $\mathbf{X}_i \in \hat{A}_{n,\ell}$. As mentioned in Bühlmann and Yu (2002), the instability problem in estimators of predictors of the form as given in (1.10) arises due to hard decisions represented by the indicator function. Therefore, bagging in this case can smooth the effect of hard decisions such as in (1.10) by averaging over indicators. Corresponding results regarding the variance reduction effect of bagged predictors of the form (1.10), mainly for regression problems, can be found in Bühlmann and Yu (2002), for example. The Random Forest model is an example of a bagged predictor of the form (1.10), which will be introduced in the next chapter.
In the Machine Learning community, the bootstrapping schemes *sampling $a_n < n$ data points without replacement* and *sampling with replacement* dominate within the bagging framework, due to its easy and fast implementation. In this thesis, however, extensions to other bootstrapping schemes within Random Forest models are going to be considered as well.

1.4 Boosting

Boosting was initially developed by Freund and Schapire (1997) for binary classification problems. It aims, similarly to the bagging procedure, to increase the performance of classifiers $\hat{\theta}_n(\mathbf{x}) \in \mathcal{M} = \{-1, 1\}$ for all $\mathbf{x} \in \mathbb{R}^p$ through the consideration of a set of classifiers, say $\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,T}$, $T \in \mathbb{N}$, where each classifier is not necessarily better than random guessing. Differently to bagging, where the extraction of a set of classifiers was conducted using bootstrapping, boosting combines classifiers $\hat{\theta}_{n,1}, \dots \hat{\theta}_{n,T}$ by weighted voting, where more weight is put to observational points that have been misclassified leading to a stronger influence of these points. That is, the boosted classifier for a binary classification problem $\mathcal{M} = \{-1, 1\}$ is given by

$$\hat{\theta}_{n,Boost}(\mathbf{x}) = sign\left(\sum_{t=1}^{T} \alpha_t \cdot \hat{\theta}_{n,t}(\mathbf{x})\right).$$

The weights $\alpha_1, \dots, \alpha_T$ are chosen such that a kind of optimality criterion is met, i.e. the boosted classifier $\hat{\theta}_{n,Boost}$ as a function of $\alpha_1, \dots . \alpha_T$ minimizes the conditional probability

error, i.e.

$$[\alpha_1, \ldots, \alpha_T]^\top = \arg \min_{\alpha_1,\ldots,\alpha_T} \mathbb{E}\left[\psi_f\left(\sum_{t=1}^{T} \alpha_t \cdot \hat{\theta}_{n,t}(\mathbf{X}), Y\right)\middle| \mathcal{D}_n\right], \tag{1.11}$$

where ψ_f is an F-loss as mentioned on page 12 of this thesis. Since the data generating process of $[\mathbf{X}^\top, Y]^\top$ is usually unknown, the approximation of (1.11) is usually conducted by considering its Monte-Carlo approximation

$$[\hat{\alpha}_1, \ldots, \hat{\alpha}_n] = \arg \min_{\alpha_1,\ldots,\alpha_T} \frac{1}{n_{test}} \sum_{i \in \mathcal{I}_{test}} \psi_f\left(\sum_{t=1}^{T} \alpha_t \cdot \hat{\theta}_{n,t}(\mathbf{X}_i), Y_i\right), \tag{1.12}$$

where the set $\{[\mathbf{X}_i^\top, Y_i]^\top\}_{i \in \mathcal{I}_{test}}$ is a test set separated initially from $\mathcal{D}_n$, i.e. $\mathcal{I}_{test} \dot{\cup} \mathcal{I}_{train} = \{1, \ldots, n\}$ such that the estimators $\{\hat{\theta}_{n,t}\}_{t=1}^T$ are trained on the training set $\{[\mathbf{X}_i^\top, Y_i]^\top\}_{i \in \mathcal{I}_{train}}$ and the approximation of (1.11) is based on the test set. The separation is usually conducted using cross validation methods and aims to not underestimate the performance of classifiers by separating $\mathcal{D}_n$ into a train and test set, see Quinlan (1996), Efron and Tibshirani (1997) and Ridgeway (2004). The choice of the loss function ψ_f clearly affects the solution to (1.12). In Freund and Schapire (1997), the loss function resulting into better classification performance in simulation experiments was the exponential loss, i.e. $\psi_f(\mathbf{x}, y) = \exp\{-y \cdot f(\mathbf{x})\}$ leading to the basis of the well-known **AdaBoost** algorithm. The specific choice of the exponential loss function has several reasons, but one can consider it as a monotone continuous approximation of the $0-1$ loss leading to a smooth and convex surrogate of the latter, as mentioned in Hastie et al. (2009b) and Schapire (2013). Due to computational efficiency reasons, a direct solution of the minimization problem given in (1.12) under the exponential loss is not directly executed, but the additive aggregation structure within the classifier $\hat{\theta}_{n,Boost}$ is used in order to obtain a greedy algorithm, i.e. an algorithm for solving problems in a stage-wise manner. Thus, given the current states $t = 1, \ldots, T-1$, find the next best vote α_{t+1} in terms of the exponential loss, i.e. $f_{n,t+1} = f_{n,t} + \alpha_{t+1} \cdot \hat{\theta}_{n,t+1}$, where $f_{n,t} = \sum_{\ell=1}^{t} \alpha_\ell \cdot \hat{\theta}_{n,\ell}$. The logic of the greedy minimization under the exponential loss led to the well-known AdaBoost algorithm with a solution of the form

$$\alpha_t = \frac{1}{2} \cdot \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right), \quad t \in \{1, \ldots, T\}, \tag{1.13}$$

where $\epsilon_t = \sum_{\substack{i \in \mathcal{I}_{test} \\ \hat{\theta}_{n,t}(\mathbf{X}_i) \neq Y_i}} w_i^{(t-1)} \cdot \left(\sum_{i \in \mathcal{I}_{test}} w_i^{(t-1)}\right)^{-1}$ with $w_i^{(t)} = w_i^{(t-1)} \exp\{-Y_i \cdot \alpha_t \cdot \hat{\theta}_{n,t}(\mathbf{X}_i)\}$ for some initial weights $\{w_i^{(0)}\}_{i \in \mathcal{I}_{test}}$ for all $i \in \mathcal{I}_{test}$. Similarly to Schapire (2013), one obtains the following algorithm:

Algorithm 1: Pseudo AdaBoost Classifier

Input: Test set $\{[\mathbf{X}_i^\top, Y_i]_{i\in\mathcal{I}_{test}}^\top\}$; Set of classifiers $\{\hat{\theta}_{n,t}\}_{t=1}^T$; Initial weights $w_i^{(0)} = 0$ for all $i \in \mathcal{I}_{test}$

Result: Boosted Classifier $\hat{\theta}_{n,Boost}^{(T)}$

Set $w_i^{(0)} = \frac{1}{|\mathcal{I}_{test}|}$ and $\hat{\theta}_{n,Boost}^{(0)} = 0$;

while $1 \leq t \leq T$ **do**

  Compute $\epsilon_t = \sum_{i\in\mathcal{I}_{test}} w_i^{(t-1)} \mathbb{1}\{\hat{\theta}_{n,t}(\mathbf{X}_i) \neq Y_i\} \cdot \left(\sum_{i\in\mathcal{I}_{test}} w_i^{(t-1)}\right)^{-1}$;

  Compute $\alpha_t = \frac{1}{2} \cdot \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$;

  Update $\hat{\theta}_{n,Boost}^{(t)} = \hat{\theta}_{n,Boost}^{(t-1)} + \alpha_t \cdot \hat{\theta}_{n,t}$;

  Update $w_i^{(t)} = w_i^{(t-1)} \cdot \exp\{-Y_i \cdot \alpha_t \cdot \hat{\theta}_{n,t}(\mathbf{X}_i)\}$;

  Set $t \leftarrow t + 1$;

end

Algorithm 1 refers to a greedy approximation of the minimization problem given in (1.12), where weak learners $\{\hat{\theta}_{n,t}\}_{t=1}^T$ in form of, e.g. decision trees, are already given. In practice, however, a preliminary access to estimators $\{\hat{\theta}_n\}_{t=1}^T$ is not given such that the minimization procedure in (1.12) is usually extended to model-specific parameters, in order to obtain simultaneously the votes $\{\alpha_t\}_{t=1}^T$ and the set of weak learners $\{\hat{\theta}_{n,t}\}_{t=1}^T$. Model-specific parameters in decision trees, for example, are splitting variables, the split location and terminal node values, such that the minimization task is extended to these parameters as well, while focusing on decision trees of simple structure, such as decision stumps as the default setting in Ridgeway (2004). An important parameter to be chosen prior to the start of the algorithm is the number of weak learners T. In Algorithm 1, this was an integrated component of the set of weak learners $\{\hat{\theta}_{n,t}\}_{t=1}^T$, but is usually chosen priorly. As mentioned in Bühlmann and Hothorn (2007), a too large choice of boosting iterations T can lead to a slow overfitting problem. Therefore, a careful choice is required in practice. The extension of the AdaBoost algorithm to multi-class problems, i.e. $K > 2$, with a multi-class exponential loss was conducted in Hastie et al. (2009a). In Breiman (1996a) and Breiman (1999), the AdaBoost algorithm was shown to be equivalent to a steepest descent algorithm in function space, whereas Friedman et al. (2000) and Friedman (2001) extended it to more general settings involving potentially different loss functions than the exponential loss and referred to the boosting principle as a *stagewise additive modeling approach*. It was motivated by the initial minimization problem

$$F^* = \arg\min_{F\in\mathcal{A}} \mathbb{E}[\psi_f(Y, F(\mathbf{X}))|\mathbf{X} = \mathbf{x}], \tag{1.14}$$

where the underlying function class $\mathcal{A}$ is assumed to be of additive nature, that is, $\mathcal{A} = \left\{\sum_{t=1}^T \beta_t \cdot h(\mathbf{x}; \mathbf{a}_t) : h(\mathbf{x}; \mathbf{a}_t) \text{ is a measurable function characterized by } \mathbf{a}_t, T \in \mathbb{N}, \beta_t \in \mathbb{R}\right\}$, which turns the minimization problem (1.14) for a finite data set $\mathcal{D}_n$ into

$$\{\beta_t, \boldsymbol{a}_t\}_{t=1}^T = \arg\min_{\{\beta_t', \boldsymbol{a}_t'\}} \frac{1}{n}\sum_{i=1}^n \psi_f\left(Y_i, \sum_{t=1}^T \beta_t' h(\mathbf{X}_i; \boldsymbol{a}_t')\right). \tag{1.15}$$

Note that the parameters $\{\boldsymbol{a}_t\}_{t=1}^T$ refer to the parametrization of decision trees, if h is chosen among CART learners. Therefore, its domain depends on the nature of the chosen base

learner. Denoting with $F_t(\mathbf{x}) = \sum_{\ell=1}^{t} \beta_\ell \cdot h(\mathbf{x}; \boldsymbol{a}_\ell)$, the negative gradient based on the data set required for conducting the steepest descent method is then given by

$$-g_t(\mathbf{x}_i) = -\left[\frac{\partial \psi_f(Y_i, F(\mathbf{X}_i))}{\partial F(\mathbf{X}_i)}\right]_{F(\mathbf{x})=F_{t-1}(\mathbf{x})}. \tag{1.16}$$

As mentioned in Friedman (2001), the whole analytical structure of the negative gradient $g_t(\mathbf{x})$ is not extractable, but is available at the data points given in $\mathcal{D}_n$. Therefore, a constrained version of the negative gradient is used in place of the unconstrained one $g_t(\mathbf{x})$ leading to an approximation of the negative gradient $-g_t(\mathbf{x})$ by $h(\mathbf{x}; \boldsymbol{a}_t)$, where

$$\{\beta_t, \boldsymbol{a}_t\} = \arg\min_{\boldsymbol{a},\beta} \sum_{i=1}^{n} [-g_t(\mathbf{X}_i) - \beta \cdot h(\mathbf{X}_i; \boldsymbol{a})]^2. \tag{1.17}$$

The step-size ρ in the gradient descent algorithm is then conducted through a line search of the form

$$\rho_t = \arg\min_{\rho} \sum_{i=1}^{n} \psi_f(Y_i, F_{t-1}(\mathbf{X}_i) + \rho h(\mathbf{X}_i; \boldsymbol{a}_t)), \tag{1.18}$$

leading to an update of the form $F_t(\mathbf{x}) = F_{t-1}(\mathbf{x}) + \rho_t \cdot h(\mathbf{x}; \boldsymbol{a}_t)$. The consideration of the boosting method in form of a minimization problem over a functional space of the form $\mathcal{A}$ enabled the extension of this method to regression problems as well. The requirements for conducting boosting for regression is the differentiability of the loss function $\psi = \psi_f$, which also coincides with the loss function used within the AdaBoost algorithm. For the latter, the exponential loss was considered as a continuous monotone and differentiable approximation of the $0-1$ loss. Finally, the **gradient boosting method** for arbitrary supervised learning problems can then be stated as follows:

Algorithm 2: Gradient Boosting

Input: Training set $\mathcal{D}_n$, stopping iteration T, loss function ψ
Result: Gradient Boosting Approximation $F_T(\mathbf{x})$

1 Set $F_0(\mathbf{x}) = \arg\min_{\rho} \sum_{i=1}^{n} \psi(Y_i, \rho)$;
2 **while** $1 \leq t \leq T$ **do**
3 Compute $\tilde{Y}_i = -\left[\frac{\partial \psi(Y_i, F(\mathbf{X}_i))}{\partial F(\mathbf{x})}\right]_{F(\mathbf{x})=F_{t-1}(\mathbf{X}_i)}$ for all $i = 1, \ldots, n$;
4 Set $(\beta, \boldsymbol{a}_t) = \arg\min_{\boldsymbol{a},\beta} \sum_{i=1}^{n} (\tilde{Y}_i - \beta h(\mathbf{X}_i; \boldsymbol{a}))^2$;
5 Set $\rho_t = \arg\min_{\rho} \sum_{i=1}^{n} \psi(Y_i, F_{t-1}(\mathbf{X}_i) + \rho \cdot h(\mathbf{X}_i; \boldsymbol{a}_t))$;
6 Update $F_t(\mathbf{x}) = F_{t-1}(\mathbf{x}) + \rho_t \cdot h_t(\mathbf{x}; \boldsymbol{a}_t)$;
7 Set $t \leftarrow t + 1$;
8 **end**

Note that line 6 in Algorithm 2 is often substituted by $F_t(\mathbf{x}) = F_{t-1}(\mathbf{x}) + \nu \cdot \rho_t \cdot h_t(\mathbf{x}; \boldsymbol{a}_t)$, where the parameter $\nu \in (0, 1]$ is a regularization parameter and corresponds to the *learning rate*. Acording to Friedman (2001), pages $12 - 13$, it aims to reduce over-fitting problems obtained from deriving an approximation beeing too close to the data.

For example, taking into account the squared error loss $\psi(y, F) = (y - F)^2$ in regression problems, i.e. $\mathcal{M} = \mathbb{R}$, and further restricting the function class to decision trees, then the gradient boosting method turns out to fit pseudo-residuals to the subsequent tree (see Friedman, 2001 on page 1194). Deviations to Algorithm 2 do exist such as the **stochastic gradient boosting method**, where randomness is introduced by conducting the computations within the while-loop of Algorithm 2 on subsampled data $\mathcal{D}_{a_n}^*$, where $a_n \leq n$ is the number of points randomly selected from $\mathcal{D}_n$. This way, one aims to combine boosting and bagging methodologies in one algorithm.

1.5 Classification and Regression Trees

Classification and regression trees (CART) are a class of algorithms, that enable the treatment of both, classification and regression learning problems and count towards the class of weak learners, if they are applied singly. That is, the prediction at $\mathbf{X}$ can change drastically, if the set $\mathcal{D}_n$ is changed. As introduced in the previous section, this effect can often be overwhelmed, if the process of bagging or boosting is applied to decision trees. In this subsection, we aim to shortly introduce CART as a weak learner itself. Therefore, let $supp(\mathbf{X}) \subseteq \mathbb{R}^p$ be the support of the random vector $\mathbf{X}$, i.e. the feature vector. The general principle of CART is to separate the support of $\mathbf{X}$ into disjoint sets $\{R_{n,s}\}_{s=1}^J$ such that $supp(\mathbf{X}) = \dot{\bigcup}_{s=1}^J A_{n,s}$. In general, if the feature vector $\mathbf{X}$ contains only continuous variables, then usually, $A_{n,s} = \bigotimes_{j=1}^p [a_{n,s,j}, b_{n,s,j})$. If some of the features are ordinal or nominal, we let $\mathcal{S}_{cat} \subseteq \{1, \ldots, p\}$ be the set of all indices, such that X_j is either nominal or ordinal for all $j \in \mathcal{S}_{cat}$. Regarding nominal scales of measurements, however, most of the CART algorithms distinguish whether they are finitely countable or infinitely countable. From a computational perspective this means that nominal variables are treated as metric, if a specific number of possible attributes is exceeded. In the R-package `randomForest`, for example, not more than 32 factor levels can be treated as nominal or ordinal (Liaw and Wiener, 2002). The region $A_{n,s}$ takes the form $A_{n,s} = \left(\bigotimes_{j \in \{1,\ldots,p\} \setminus \mathcal{S}_{cat}} [a_{n,s,j}, b_{n,s,j}) \right) \otimes \left(\bigotimes_{j \in \mathcal{S}_{cat}} \xi_{n,s,j} \right)$, where $\xi_{n,s,j} \in supp(X_j)$ for all $j \in \mathcal{S}_{cat}$. For every region $A_{n,s}$, $s = 1, \ldots, J$, a *constant* value $\{\hat{c}_{n,s}\}_{s=1}^J$ is then assigned. Constant in this sense does not exclude the possibility that $\hat{c}_{n,\ell}$ is random, but rather refers to the function class CART belongs to, namely to the family of piecewise constant functions, i.e. $\mathcal{A} = \left\{ \sum_{s=1}^J c_s \mathbb{1}\{x \in A_s\} : c_s \in \mathcal{M}, \quad \dot{\bigcup}_{s=1}^J A_s = supp(\mathbf{X}) \right\}$. Therefore, parameters of interest in CART are the constant values $\{c_s\}_{s=1}^J$, the regions $\{A_s\}_{s=1}^J$, and if not directly obtainable, the number of regions J. We emphasized the dependence of these parameters towards the data set $\mathcal{D}_n$ by denoting $\hat{c}_{n,s}$ and $A_{n,s}$ as kind of estimators for c and A_s, for all $s = 1, \ldots, J$. It is worth to notice that in practice, the number of regions J, also called *leaves*, do depend on $\mathcal{D}_n$ as well, such that actually, $J = J_n$.

Now, define a cut within a tree as the pair $[j, z_j]^\top \in \{1, \ldots, p\} \times supp(X_j)$, where $\mathbf{X} = [X_1, \ldots, X_p]^\top \in \mathbb{R}^p$. The determination of a cut is then conducted based on some optimality criterion, which depends on the underlying learning problem. In case of a regression problem, the cut is conducted by maximizing the decrease in empirical variance, that is, for a cut

$j \in \{1, \ldots, p\} \setminus \mathcal{S}_{cat}$ at tree level $1 \leq k \leq \lceil \log_2(J) \rceil + 1$, one has

$$[j_n^{(k)}, z_{j_n}^{(k)}] = \arg\max_{[j,z_j]} \frac{1}{N_n\left(A_{n,s}^{(k)}\right)} \left(\sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{n,s}^{(k)}} \right)^2 \cdot \mathbb{1}\{\mathbf{X}_i \in A_{n,s}^{(k)}\} - \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{n,s,L}^{(k)}(j,z_j)} \mathbb{1}\{X_{i,j} < z_j\} - \bar{Y}_{A_{n,s,R}^{(k)}(j,z_j)} \mathbb{1}\{X_{i,j} \geq z_j\} \right)^2 \cdot \mathbb{1}\{\mathbf{X}_i \in A_{n,s}^{(k)}(j, z_j)\} \right), \tag{1.19}$$

where $A_{n,s,L}^{(k)}(j, z_j) = \{\mathbf{x} \in A_{n,s}^{(k)} : x_j < z_j\}$ is the left part of the hyper-rectangular cell $A_{n,s}^{(k)}$ separated along the cut $[j, z_j]^\top$ and $A_{n,s,R}^{(k)}(j, z_j)$ is the right part of $A_{n,s}^{(k)}$ along the cut $[j, z_j]^\top$, i.e. $A_{n,s,R}^{(k)}(j, z_j) = \{\mathbf{x} \in A_{n,s}^{(k)} : x_j \geq z_j\}$. $N_n(A_{n,s}^{(k)})$ is the cardinality of $A_{n,s}^{(k)}$, i.e. the number of observations in $\mathcal{D}_n$, that fall in $R_{n,s}^{(k)}$. $\bar{Y}_{A_{n,s}^{(k)}}$, $\bar{Y}_{A_{n,s,L}^{(k)}}$ and $\bar{Y}_{A_{n,s,R}^{(k)}}$ denote the corresponding arithmetic mean of $\{Y_i\}_{i:\mathbf{X}_i \in A_{n,s}^{(k)}}$, $\{Y_i\}_{i:\mathbf{X}_i \in A_{n,s,L}^{(k)}}$ and $\{Y_i\}_{i:\mathbf{X}_i \in A_{n,s,R}^{(k)}}$. The specific choice of $\lceil \log_2(J) \rceil + 1$ as an upper bound for the tree level originates from the fact that J refers to the tree-leaves while our trees are binary. Note that the previous cut-representation is only valid for at least ordinal features j. The index $1 \leq s \leq 2^{k-1}$ denotes the corresponding region at level k, which increases exponentially, as k increases. The determination of the cut $[j, z_j]^\top$ according to (1.19) can be executed for the first cut-dimension $j \in \{1, \ldots, p\}$, but requires an approximation for the values of z_j, since the search of the whole region $supp(X_j)$ can be computationally infeasible. In practice, due to the finite amount of data points $n \in \mathbb{N}$, the CART algorithm selects potential cut values z_j among the set $\mathcal{Z}_{n,j} = \{(X_{i,j} + X_{i+1,j})/2 : i = 1, \ldots, n-1\}$, if X_j is interval or ratio scaled (Loh and Shih, 1997). The minimization is then solved by computing the variance reduction as given in (1.19) for every element in $\mathcal{Z}_{n,j}$, given a fixed value of $j \in \{1, \ldots, p\}$. In case that X_j is nominal or ordinal scaled, such that the number of potential attributes does not exceed a pre-defined fixed number, the empirical cut criterion given in (1.19) slightly changes by substituting the indicators to $\mathbb{1}\{X_{i,j} = \xi_j\}$, where $\xi_j \in supp(X_j)$. The notation of the regions in the latter case changes then to $A_{n,s,L}^{(k)}(j, \xi_j) = \{\mathbf{x} \in A_{n,s}^{(k)} : x_j \neq \xi_j\}$ and $A_{n,s,R}^{(k)}(j, \xi_j) = \{\mathbf{x} \in A_{n,s}^{(k)} : x_j = \xi_j\}$ and is motivated by Breiman et al. (1984).

For classification purposes, the usage of the cut criterion as given in (1.19) is not suitable, since neither means nor differences can be computed for nominal features. However, the general idea in CART for classification is to obtain regions $A_{n,s}$, with $1 \leq s \leq J$, such that each region contains response observations $\{Y_i\}_{i=1}^n$ that allow a *clear* assignment of classes according to the majority vote logic within a region. Therefore, one requires a measure, that is able to reflect the *purity* of each region accordingly. Following the same definition of node impurity as in Breiman et al. (1984) on page 24, we state a class of measures, so called *impurity measures*, that are used in CART algorithms.

Definition 1.6 (Impurity Measures)**.** Let $\mathcal{P}_K$ be a set of K-discrete probability measures, i.e. probability measures on $\{1, \ldots, K\}$, which can be expressed in vector notation on the K-dimensional simplex. An impurity measure $\phi_K : \mathcal{P}_K \longrightarrow \mathbb{R}_+$ is a function which satisfies:

- ϕ_K takes its maximum for the discrete uniform distribution, i.e. $\arg\max_{\mathbb{P} \in \mathcal{P}_K} \phi_K(\mathbb{P}) = Dunif\{1, \ldots, K\}$.
- ϕ_K attains its minimum for a K-discrete distribution of the form $[p_1, \ldots, p_K]^\top \in \mathcal{P}_K$ such that $p_s = 1$ and $p_t = 0$ for exactly one s and for all $t \neq s$ with $s, t \in \{1, \ldots, K\}$.

- ϕ_K is symmetric in its input arguments, that is, for all $[p_1, \ldots, p_K]^\top \in \mathcal{P}_K$ and a permutation $\pi : \{1, \ldots, K\} \to \{1, \ldots, K\}$ one has $\phi_K(p_1, \ldots, p_K) = \phi(p_{\pi(1)}, \ldots, p_{\pi(K)})$.

Now, define

$$p_t(A_{n,s}^{(k)}) = \frac{1}{N_n(A_{n,s}^{(k)})} \sum_{i:\mathbf{X}_i \in A_{n,s}^{(k)}} \mathbb{1}\{Y_i = \ell_t\}$$

as the proportion of observations falling in $A_{n,s}^{(k)}$, with class label ℓ_t, $1 \leq t \leq K$ and $\boldsymbol{p}(A_{n,s}^{(k)}) := [p_1(A_{n,s}^{(k)}), \ldots, p_K(A_{n,s}^{(k)})]^\top \in [0,1]^K$. Then, most of the CART algorithms make use of the following impurity measures:

1. The *Misclassification error*, i.e. $\phi_K(\boldsymbol{p}(A_{n,s}^{(k)})) = 1 - \max\limits_{1 \leq t \leq K} p_t(A_{n,s}^{(k)})$,

2. The *Gini-index* $\phi_K(\boldsymbol{p}(A_{n,s}^{(k)})) = \sum\limits_{t=1}^{K} p_t(A_{n,s}^{(k)}) \cdot (1 - p_t(A_{n,s}^{(k)}))$,

3. The *Cross-entropy* or *deviance* $\phi_K(\boldsymbol{p}(A_{n,s}^{(k)})) = -\sum\limits_{t=1}^{K} p_t(A_{n,s}^{(k)}) \cdot \log(p_t(A_{n,s}^{(k)}))$.

Conducting a cut in classification problems is then executed by maximizing the decrease in node impurity, that is

$$\begin{aligned}[j_n^{(k)}, z_{j_n}^{(k)}] = \arg\max_{[j,z_j]} \Bigg\{ & \phi_K(\boldsymbol{p}(A_{n,s}^{(k)})) - \frac{N_n(A_{n,s,L}^{(k)}(j,z_j))}{N_n(A_{n,s}^{(k)})} \cdot \phi_K(\boldsymbol{p}(A_{n,s,L}^{(k)}(j,z_j))) \\ & - \frac{N_n(A_{n,s,R}^{(k)}(j,z_j))}{N_n(A_{n,s}^{(k)})} \cdot \phi_K(\boldsymbol{p}(A_{n,s,R}^{(k)}(j,z_j))) \Bigg\}. \end{aligned} \tag{1.20}$$

The cut criterion in (1.20) is also referred to as the *information gain criterion*. In the Random Forest method, for example, the Gini-index is used as an impurity measure in every tree within the ensemble, but most of the CART algorithms allow variations by explicitly stating the impurity measure in function calls.

After conducting the cuts until one reaches $J = J_n$ regions $A_{n,s}^{(\lceil \log_2(J) \rceil + 1)} = A_{n,s}$ for $1 \leq s \leq J$, the aim is to estimate the *constant* values $\{c_s\}_{s=1}^J$ appropriately. For the latter, one usually distinguishes again between regression and classification learning problems:

$$\hat{c}_{n,s} = \begin{cases} \sum\limits_{i:\mathbf{X}_i \in A_{n,s}} Y_i & \text{for regression,} \\ Mode\{Y_i : \mathbf{X}_i \in A_{n,s}\} & \text{for classification.} \end{cases} \tag{1.21}$$

Growing a CART tree too deep, that is, choosing the number of terminal nodes J very large, can lead to overfitting problems, see e.g. Schaffer (1993) and Hastie et al. (2009b) on page 307. This can be seen by considering the cut criterions given in (1.19) resp. (1.20). Since choosing J sufficiently large such that in each region $A_{n,s} = A_{n,s}^{(\lceil \log_2(J) \rceil + 1)}$, there is at most one observation, one can have a perfect fit of the decision tree to the learning set $\mathcal{D}_n$. However, potentially different data sets can lead to completely wrong predictions in terms of class assignment or mean squared error for regression problems. Therefore, one possibility to reduce the effect of overfitting for deeply grown trees is *pruning*. Following the explanation

given in Hastie et al. (2009b) on page 308, pruning is the process of successively collapsing a grown tree until the single node tree is reached. Then, denoting with Γ_n an already grown tree with regions $\{A_{n,s}\}_{s=1}^{J}$ and *constant* values $\{\hat{c}_{n,s}\}_{s=1}^{J}$, pruning is conducted by collapsing the CART tree successively among the internal nodes, starting from the internal nodes at level $k = \lceil \log_2(J) \rceil$. During the collapsing steps, one obtaines a sequence of subtrees $\{\Gamma_{n,\ell}\}_\ell$, where the aim is to find the smallest subtree Γ_{n,ℓ^*}, such that

$$\ell^* = \arg\min_{\ell} C_\alpha(\Gamma_{n,\ell}) = \arg\min_{\ell} \sum_{s=1}^{J(\Gamma_{n,\ell})} N_n(A_{n,s}(\Gamma_{n,\ell})) \cdot Q_s(\Gamma_{n,\ell}) + \alpha \cdot J(\Gamma_{n,\ell}). \tag{1.22}$$

$J(\Gamma_{n,\ell})$ refers to the number of terminal nodes, or final regions, of the collapsed tree $\Gamma_{n,\ell}$ and analogously, $N_n(A_{n,s}(\Gamma_{n,\ell}))$ is the number of observations falling in $A_{n,s}(\Gamma_{n,\ell})$. The tuning parameter $\alpha \geq 0$ models the trade-off between tree depth and goodness of fit $Q_s(\Gamma_{n,\ell})$ to the training data $\mathcal{D}_n$. For regression problems, one has

$$Q_s(\Gamma_{n,\ell}) = \frac{1}{N_n(A_{n,s}(\Gamma_{n,\ell}))} \sum_{i:\mathbf{X}_i \in A_{n,s}(\Gamma_{n,\ell})} (Y_i - \hat{c}_{n,s})^2,$$

whereas for classification problems, the goodness of fit measure is nothing else than the choosen impurity measure applied on the regions $\{A_{n,s}\}_{s=1}^{J(\Gamma_{n,\ell})}$. As mentioned in Hastie et al. (2009a), for every fixed α, there exists a unique subtree $\Gamma_{n,\ell^*} = \Gamma_{n,\ell(\alpha^*)}$ that actually solves (1.22) and is contained in the sequence of CART trees $\{\Gamma_{n,\ell}\}_\ell$.

The process of pruning is usually conducted after obtaining a grown tree Γ_n. Due to the enormous amount of possible subtrees from Γ_n, the solution to (1.22) can be very time consuming. A possibility to bypass this effect is to use bagging and to drop the cost-complexity pruning. This is usually conducted in Random Forest models, making the latter a faster regression and classification algorithm, see e.g. the R-packages `ranger` and `randomForest` (Liaw and Wiener, 2002; Wright and Ziegler, 2017).

1.6 Proofs of the Chapter

In this subsection, we formally prove that the classifier in Theorem 1.1 indeed solves the multi-class optimization problem given in (1.6).

Proof of Theorem 1.1. Let $\mathbf{x} \in \mathbb{R}^p$ be fixed, $\mathcal{M} = \{\ell_1, \dots, \ell_K\}$ and denote with $\eta(k|\mathbf{x}) = \mathbb{P}[Y = \ell_k|\mathbf{X} = \mathbf{x}]$. Furthermore, let $g \in \mathcal{C}$ be any fixed, but arbitrary classifier and let $g^*(\mathbf{x}) = \arg\max_{k\in\mathcal{M}} \mathbb{P}[Y = k|\mathbf{X} = \mathbf{x}]$. Then one can conduct the following computations:

$$\begin{aligned}
\mathbb{P}[g(\mathbf{X}) \neq Y|\mathbf{X} = \mathbf{x}] &= 1 - \mathbb{P}[g(\mathbf{X}) = Y|\mathbf{X} = \mathbf{x}] \\
&= 1 - \sum_{k=1}^{K} \mathbb{P}[Y = \ell_k, g(\mathbf{X}) = \ell_k|\mathbf{X} = \mathbf{x}] \\
&= 1 - \sum_{k=1}^{K} \mathbb{1}\{g(\mathbf{x}) = \ell_k\} \cdot \mathbb{P}[Y = \ell_k|\mathbf{X} = \mathbf{x}] \\
&= 1 - \sum_{k=1}^{K} \mathbb{1}\{g(\mathbf{x}) = \ell_k\} \cdot \eta(k|\mathbf{x}).
\end{aligned}$$

Applying the same decomposition to $\mathbb{P}[g^*(\mathbf{X}) \neq Y|\mathbf{X} = \mathbf{x}]$ leads to

$$\begin{aligned}
\mathbb{P}[g(\mathbf{X}) \neq Y|\mathbf{X} = \mathbf{x}] - \mathbb{P}[g^*(\mathbf{X}) \neq Y|\mathbf{X} = \mathbf{x}] &= \sum_{k=1}^{K} \eta(k|\mathbf{x}) \left(\mathbb{1}\{g^*(\mathbf{x}) = \ell_k\} - \mathbb{1}\{g(\mathbf{x}) = \ell_k\}\right) \\
&\geq 0. \qquad (1.23)
\end{aligned}$$

The inequality follows by considering the following: Suppose that $g^*(\mathbf{x}) = \ell_{k_0}$ for some fixed $k_0 \in \{1, \dots, K\}$. Then, due to the definition of g^*, one can deduce that $\eta(k_0|\mathbf{x}) \geq \eta(k|\mathbf{x})$ for all $k \neq k_0$ with $k \in \{1, \dots, K\}$. Now distinguish between two cases

- Suppose that g takes the same decision as g^*, i.e. $g(\mathbf{x}) = \ell_{k_0}$. Then clearly,

$$\sum_{k=1}^{K} \eta(k|\mathbf{x}) \left(\mathbb{1}\{g^*(\mathbf{x}) = \ell_k\} - \mathbb{1}\{g(\mathbf{x}) = \ell_k\}\right) = \eta(k_0|\mathbf{x})(1-1) + \sum_{k:k\neq k_0}^{K} \eta(k|\mathbf{x})(0-0) = 0.$$

- Suppose that $g(\mathbf{x}) = \ell_{k_1}$ for some fixed $k_1 \in \{1, \dots, K\}$, such that $k_1 \neq k_0$, i.e. g takes another decision. Then, one obtains:

$$\begin{aligned}
\sum_{k=1}^{K} \eta(k|\mathbf{x}) \left(\mathbb{1}\{g^*(\mathbf{x}) = \ell_k\} - \mathbb{1}\{g(\mathbf{x}) = \ell_k\}\right) &= \eta(k_0|\mathbf{x})(1-0) + \\
&\quad \sum_{k:k\neq k_0}^{K} \eta(k|\mathbf{x})(0 - \mathbb{1}\{g(\mathbf{x}) = \ell_k\}) \\
&= \eta(k_0|\mathbf{x}) - \eta(k_1|\mathbf{x}) \geq 0.
\end{aligned}$$

From inequality (1.23), we can deduce

$$\begin{aligned}
\mathbb{P}[g(\mathbf{X}) \neq Y] - \mathbb{P}[g^*(\mathbf{X}) \neq Y] &= \int \left\{\mathbb{P}[g(\mathbf{X}) \neq Y|\mathbf{X} = \mathbf{x}] - \mathbb{P}[g^*(\mathbf{X}) \neq Y|\mathbf{X} = \mathbf{x}]\right\} d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) \\
&\geq 0,
\end{aligned}$$

which yields to $g_{opt} = g^*$.

■

Chapter 2

Random Forest Models

This chapter is solely devoted to the Random Forest model, since most of this thesis is based on the latter algorithm. The Random Forest can be considered as an ensemble of decision trees, which are combined using the principles of bagging. Beside the random selection of observations in $\mathcal{D}_n$ for the construction of each tree, the Random Forest has the characterizing property to select a random subset $\mathcal{M}_{try} \subseteq \{1, \ldots, p\}$ for possible cuts. That is, the Random Forest randomly selects $m_{try} = |\mathcal{M}_{try}|$ feature indices without replacement from $\{1, \ldots, p\}$, and conducts generally a cut according to the cut criterion given in Chapter 1 equation (1.19) resp. (1.20) using the Gini-index. During the construction of decision trees according to the general principle of the CART logic, pruning of the trees is not conducted. Instead, either a maximal predefined number of leaves is specified, or a maximal number of observations falling in a terminal node is set. This algorithm was initially developed in Breiman (2001) making the Random Forest suitable for different classification and regression problems. It is worth to notice that several modifications of the Random Forest do exist, such as modifications in the random feature selection step, modifications in the bagging procedure or even completely randomized trees (Geurts et al., 2006). One example has been invented and analyzed in Ramosaj and Pauly (2019c) within the issue of missing value imputations and is part of this thesis. Therefore, when we want to address different Random Forests as the one described in Breiman (2001), even with slight modifications regarding the choice of hyper-parameters, we simply say Random Forest models. In the sequel, however, we will describe the traditional Random Forest as given in Breiman (2001), which is widely implemented in different statistical software packages such as R, python or SAS.
In the sequel, we will mainly focus on the mathematical notations given in Scornet et al. (2015) for describing hyper-parameters and mathematical forces involved in the Random Forest and extended them appropriately. Therefore, let us first denote some important hyper-parameters priorly chosen to the tree construction process:

(i) The number of decision trees $M \in \mathbb{N}$ in the ensemble,

(ii) The sampling strategy $\mathcal{S}^*$,

(iii) $m_{try} \in \{1, \ldots, p\}$ the number of pre-selected features for conducting splits,

(iv) a_n the number of selected observations for constructing each tree,

(v) t_n the number of leaves in each tree.

Denote with $\boldsymbol{\Theta}_t = [\boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}]^\top \in \{0,1\}^n \times \{0,1\}^{p_n \times p}$ for $t = 1, \ldots, M$ the generic random vector, where $\boldsymbol{\Theta}_t^{(1)}$ is responsible for the sampling mechanism prior to tree construction and

$\boldsymbol{\Theta}_t^{(2)}$ the random vector modeling feature subspacing. Note that the above notation is novel and is introduced to clarify different mathematical forces of Random Forest models in this thesis. It was especially required for the development of various proofs presented in this work. In Scornet et al. (2015), however, the generic random vector is simply denoted as a sequence of random variables $\{\Theta_t\}_{t=1}^M$. The dimension p_n has to be understood as a potential function of the hyper-parameter t_n, which reflects the tree-depth and therefore, the row-wise independent replications in the random matrix $\boldsymbol{\Theta}_t^{(2)}$. Note that the sampling mechanism and feature subspacing happens independently of each other for every tree $t = 1, \ldots, M$. It therefore immediately follows that $\{\boldsymbol{\Theta}_t\}_{t=1}^M$ is a sequence of iid random vectors. In addition $\boldsymbol{\Theta}_t^{(1)}$ can be separated into $[\Theta_{1,t}^{(1)}, \ldots, \Theta_{n,t}^{(1)}]^\top$ for all $t = 1, \ldots, M$, where $\Theta_{i,t}^{(1)} \in \{0, 1\}$ indicates whether observation $i \in \{1, \ldots, n\}$ has been selected or not. This directly implies that $\Theta_{i,t}^{(1)} \sim Bernoulli(p_i)$ for some probability p_i, which depends on the sampling mechanism and potentially on i, whereas $\sum_{i=1}^{n} \Theta_{i,t}^{(1)} = a_n$. Similarly, one can define $\boldsymbol{\Theta}_t^{(2)} = [\boldsymbol{\Theta}_{1,t}^{(2)}, \ldots, \boldsymbol{\Theta}_{p_n,t}^{(2)}]^\top \in \{0, 1\}^{p_n \times p}$ with $\boldsymbol{\Theta}_{k,t}^{(2)} = [\Theta_{1,k,t}^{(2)}, \ldots, \Theta_{p,k,t}^{(2)}]^\top$, where $\Theta_{j,k,t}^{(2)} \in \{0, 1\}$ indicates whether feature $j \in \{1, \ldots, p\}$ has been selected or not during the feature subspacing step at tree-level-cut k within tree t. That is $\Theta_{j,k,t}^{(2)} \sim Bernoulli(p_j^{(2)})$ with $\sum_{j=1}^{p} \Theta_{j,k,t}^{(2)} = m_{try}$ for all $t = 1, \ldots, M$ and $k = 1, \ldots, p_n$. We have to emphasize that $\{\boldsymbol{\Theta}_{k,t}^{(2)}\}_{k,t}$ remains a sequence of iid random vectors. In the last chapter, we denoted a decision tree as Γ_n, however, this notation was not consistently used throughout the four different articles considered in the thesis, which were prepared for different outlets. To be more precise, we denote with $m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_t, \mathcal{D}_n)$ in this thesis a single decision tree according to the Random Forest algorithm build with random element $\boldsymbol{\Theta}_t$ on the training set $\mathcal{D}_n$ predicting the outcome at $\mathbf{x} \in \mathbb{R}^p$. Then, depending whether the learning problem is a regression or classification problem, the aggregation of the estimators $m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_t, \mathcal{D}_n)$ is denoted by $m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) = m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n)$ and is referred to as the *finite forest estimate*. Hence, we have

$$m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) = \begin{cases} \frac{1}{M} \sum_{t=1}^{M} m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_t, \mathcal{D}_n) & \text{for regression,} \\ Mode\{m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_t, \mathcal{D}_n) : t = 1, \ldots, M\} & \text{for classification.} \end{cases} \tag{2.1}$$

Due to the strong law of large numbers, one can deduce for regression problems $\mathbb{P}_{\boldsymbol{\Theta}}$ - almost surely as $M \to \infty$, (see also Scornet et al., 2015 on page 1719), that

$$m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) \longrightarrow \mathbb{E}_{\boldsymbol{\Theta}}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n)] =: m_n(\mathbf{x}; \mathcal{D}_n) = m_{n,\infty}(\mathbf{x}; \mathcal{D}_n), \tag{2.2}$$

where $\mathbb{E}_{\boldsymbol{\Theta}}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n)] = \mathbb{E}_{\boldsymbol{\Theta}}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n)|\mathcal{D}_n]$ denotes the expectation w.r.t. the random vector $\boldsymbol{\Theta}$, conditioned on the data $\mathcal{D}_n$. $\boldsymbol{\Theta}$ is considered as an independent copy of $\boldsymbol{\Theta}_1$. The quantity $m_n(\mathbf{x}; \mathcal{D}_n)$ is referred to as the *infinite forest estimate*. Obtaining the same result for classification problems is not directly possible. Regarding the latter, suppose that one has a binary classification problem such that w.l.o.g. $\mathcal{M} = \{0, 1\}$. Then, the finite forest estimate for this problem can be rewritten into $m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) = \mathbb{1}\left\{\sum_{t=1}^{M} \mathbb{1}\{m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_t, \mathcal{D}_n) = 1\} \geq M/2\right\}$. Although the strong law of large numbers can be applied here to the sequence $\{\mathbb{1}\{m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_t, \mathcal{D}_n) = 1\}\}_{t=1}^M$ leading to a limit of the form $\mathbb{P}_{\boldsymbol{\Theta}}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_t, \mathcal{D}_n) = 1]$, however, this does not automatically guarantee that $m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n)$ converges almost surely to $\mathbb{1}\{\mathbb{P}_{\boldsymbol{\Theta}}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n)] \geq 1/2\}$.

This is possible, if $\mathbb{P}[\mathbb{P}_{\boldsymbol{\Theta}} m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n)] = 1/2] = 0$, i.e. vanishing the likelihood of discontinuity points, which leads to the possibility of applying the continuous mapping theorem. In this sense, the Random Forest classifier has to be understood as a plugged-in Bayes classifier.

Now returning to the present cut criterion in the Random Forest, it is nothing else than the cut criterion mentioned in Chapter 1 in (1.19) for regression problems, and (1.20) using the Gini index with slight modifications of the corresponding domain the maximization is taken over. Denoting with $L_{n,s}^{(k)}$ the cut criterion at level $1 \leq k \leq \lceil \log_2(t_n) \leq \rceil + 1$ using the cell in $1 \leq s \leq 2^{k-1}$, where $\mathcal{C}_{A_{n,s}^{(k)}}$ denotes all possible cuts in the cell region $A_{n,s}^{(k)}$, we can state the cut criterion as follows:

$$[j_n^*, z_{j_n^*}] = \arg \max_{\substack{j \in \mathcal{M}_{try,s}^{(k)} \\ [j,z] \in \mathcal{C}_{A_{n,s}^{(k)}}}} L_n^{(k)}[j, z], \tag{2.3}$$

where $\mathcal{M}_{try,s}^{(k)} \subseteq \{1, \ldots, p\}$ is the random subset with cardinality m_{try} for cell s at level k. In most of the theoretical work for Random Forests, it is assumed that the feature domain is restricted to the p-dimensional unit cube, i.e. $supp(\mathbf{X}) = [0, 1]^p$, see e.g. Biau (2012); Scornet (2016); Scornet et al. (2015); Wager and Athey (2018). This was usually preceded by the initial, yet unproved idea that Random Forests are invariant under monotone transformations. To formally close this gap, we will state this as a proposition proved at the end of this chapter. It will give the theoretical basis that restricting the support to the p-dimensional unit cube does not have sever generalization effects.

Proposition 2.1. *Let $\mathcal{D}_n = \{[\mathbf{X}_i^\top, Y_i]^\top\}_{i=1}^n$ be a sequence of iid random vectors such that $\mathbf{X}_i = [X_{1,i}, \ldots, X_{p,i}]^\top$. Suppose that $X_{j,1}$ has either a continuous density function or it has a finite support $supp(X_{j,1})$ for all $j \in \{1, \ldots, p\}$. Then there exists a sequence of transformations $\{F_j\}_{j=1}^p$, $F_j : supp(X_j) \to [0, 1]$ such that given the data set $\mathcal{D}_n$ and the generic random variables $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M$, the finite Random Forest is invariant whether it is trained on $\mathcal{D}_n$ or on $\widetilde{\mathcal{D}}_n$, where $\widetilde{\mathcal{D}}_n = \{[\widetilde{\mathbf{X}}_i^\top, Y_i]^\top\}_{i=1}^n$ with $\widetilde{\mathbf{X}}_i = [F_1(X_{1,i}), \ldots, F_p(X_{p,i})]^\top$. In case of metric data, $\{F_j\}_{j=1}^p$ is monotone. The results also hold for the infinite Random Forest, conditioned on $\mathcal{D}_n$.*

It is worth to notice that a Random Forest tree does not directly require that the tree is symmetric, i.e. separating the tree at the root, the left tree part does not need to have the same tree depth as the right part. This might be misleading, since we denoted with $1 \leq s \leq 2^{k-1}$ the corresponding cell or region we want to address at level $1 \leq k \leq \lceil \log_2(t_n) \rceil + 1$. In this case, 2^{k-1} has to be understood as an upper bound, such that for some levels within the tree structure, one might have less than 2^{k-1} regions. This depends on the hyper-parameter t_n, or on the maximal number of terminal node observations one alternatively can fix, such as in R, python or SAS. Algorithm 3 on page 27 summarizes the Random Forest using the above notations.

For regression problems, the Random Forest has the suitable mathematical property that it can be rewritten into the weighted summation of the response $\{Y_i\}_{i=1}^n$, see e.g. Biau (2012) and Biau and Scornet (2016):

$$m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) = \sum_{i=1}^n W_{n,i}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M) \cdot Y_i, \tag{2.4}$$

$$m_n(\mathbf{x}; \mathcal{D}_n) := \mathbb{E}_{\boldsymbol{\Theta}_1}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_1, \mathcal{D}_n)] = m_{n,\infty}(\mathbf{x}; \mathcal{D}_n) = \sum_{i=1}^n W_{n,i}(\mathbf{x}) \cdot Y_i, \tag{2.5}$$

where $W_{n,i}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M) = \frac{1}{M} \sum_{t=1}^{M} \left\{ \frac{\mathbb{1}\{\mathbf{X}_i \in A_n(\mathbf{x}; \boldsymbol{\Theta}_t)\}}{N_n(A_n(\mathbf{x}; \boldsymbol{\Theta}_t))} \right\}$. Here, $A_n(\mathbf{x}; \boldsymbol{\Theta}_t)$ denotes the hyper-rectangular cell containing the fixed point $\mathbf{x}$ of the Random Forest tree designed by $\boldsymbol{\Theta}_t$. To be in line with the previous notations, the cell $A_n(\mathbf{x}; \boldsymbol{\Theta}_t)$ is nothing else than $A_{n,s(\mathbf{x})}^{(\lceil \log_2(t_n) \rceil + 1)}$ for some specific region $s = s(\mathbf{x})$, that depends on $\mathbf{x}$ and is constructed using $\boldsymbol{\Theta}_t$. For the infinite Random Forest, the weights are given by $W_{n,i}(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\Theta}} \left[\frac{\mathbb{1}\{X_i \in A_n(\mathbf{x}; \boldsymbol{\Theta}_1)\}}{N_n(A_n(\mathbf{x}; \boldsymbol{\Theta}_1))} \right]$. It is worth to notice that the weights $\{W_{n,i}(\mathbf{x}, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M)\}_{i=1}^n$ form a sequence of identically distributed random variables, but possess complex dependency structures not only with the features $\mathbf{X}$, but also with the response Y making the Random Forest model complex in mathematical analysis. Therefore, applying theoretical results similar to Barbe and Bertail (1995); Hall (1992); Præstgaard and Wellner (1993) or Efron and Tibshirani (1994) from (weighted) bootstrapping is not straightforward and face difficulties. The weights for both, the finite and infinite Random Forest fulfill the property that $\mathbb{P}$- almost surely $W_{n,i}(\mathbf{x}, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_m), W_{n,i}(\mathbf{x}) \geq 0$ for all $i = 1, \ldots, n$ with $\sum_{i=1}^{n} W_{n,i}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M) = 1$ respectively $\sum_{i=1}^{n} W_{n,i}(\mathbf{x}) = 1$.

The complexity arises due to the dependency of the weights towards the response $\{Y_i\}_{i=1}^n$ originating from the usage of the data $\mathcal{D}_n$ in the cut criterion $L_{n,s}^{(k)}$ for all k and s. The representation of the Random Forest in terms of equations (2.4) and (2.5) enables the consideration of the Random Forest as a nearest neighbor method, such as the authors in Biau and Devroye (2010) have discovered.

Beside the complexity of the Random Forest construction procedure, the prediction of observations $\{\mathbf{X}\}_{i=1}^n$ potentially being part of the training data $\mathcal{D}_n$ is a non-trivial undertaking. This is usually required for evaluating the Random Forest method with regard to its performance, as described in Chapter 1 equation (1.5) resp. (1.8). There, it was assumed that the pair $\mathbf{Z} := [\mathbf{X}^\top, Y]^\top$ is an independent copy of the training data $\mathcal{D}_n$, such that $\mathbf{Z}$ is unseen for $m_{n,M}(\cdot; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n)$ resp. $m_n(\cdot; \mathcal{D}_n)$. In order to evaluate prediction performance of the Random Forest, in general, one would require to split the data set $\mathcal{D}_n$ into a test and training set. This comes with the cost of potentially poor performance, especially when the sample size is small, since the algorithm is trained on a reduced sample size. The Random Forest, or in general any CART-like algorithm combined with the bagging procedure has the nice property that an initial separation into a test and training set is not required. Instead, one can make use of the **Out-Of-Bag** principle. That is, for predicting the outcome Y_i using a feature $\mathbf{X}_i$, use only those trees among the $t = 1, \ldots, M$, that did not use $\mathbf{X}_i$ for constructing the tree using $\boldsymbol{\Theta}_t$. Considering the two possible sampling strategies in Breiman's Random Forest, namely sampling without replacement of $a_n < n$ points or sampling $a_n \leq n$ points with replacement. Then, with positive probability, one can find at least one tree that did not used for a fixed $i \in \{1, \ldots, n\}$ the feature $\mathbf{X}_i$ and Y_i for training. Out-of-Bag Random Forest predictions have therefore been a key part of theoretical analysis in this thesis, enabling a more realistic viewpoint in Random Forest predictions. Therefore, we denote with $m_{n,M}^{OOB}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_m, \mathcal{D}_n)$ and $m_n^{OOB}(\mathbf{x}; \mathcal{D}_n) = m_{n,\infty}^{OOB}(\mathbf{x}; \mathcal{D}_n)$ the Out-Of-Bag prediction of the finite and infinite Random Forest. For the infinite Random Forest, the Out-Of-Bag prediction requires a more thorough analysis, which is shifted to the upcoming subsections emphasizing the theoretical work conducted within this thesis.

Algorithm 3: Random Forest for regression or classification.

Input: Training set $\mathcal{D}_n$, number of decision trees M, $m_{try} \in \{1, \ldots, p\}$, $a_n \in \{1, \ldots, n\}$, $t_n \in \{1, \ldots, a_n\}$ and sampling strategy $\mathcal{S}^*$

Output: Random forest estimate $m_{n,M}$

for $j = 1, \ldots, M$ **do**

  Select a_n data points according to the resampling strategy $\mathcal{S}^*$ from $\mathcal{D}_n$;

  Set $n_{nodes} = 1$;

  while $n_{nodes} \leq t_n$ **do**

    Select without replacement a subset $\mathcal{M}_{try} \subseteq \{1, \ldots, p\}$ with $|\mathcal{M}_{try}| = m_{try}$;

    for $s = 1, \ldots, k = \lceil \log_2(n_{nodes}) \rceil + 1$ **do**

      Find $(j_s^*, z_s^*) = \arg \min_{\substack{j \in \mathcal{M}_{try} \\ (j,z) \in \mathcal{C}_{A_{n,s}^{(k)}}}} L_n^{(k)}(j, z)$, where $\mathcal{C}_{A_{n,s}^{(k)}}$ is the set of all possible cuts in $A_{n,s}^{(k)}$. Moreover, for *regression problems*, we have

$$L_n^{(k)}(j,z) = \frac{1}{N_n(A_{n,s}^{(k)})} \sum_{i=1}^{n} (Y_i - \bar{Y}_{A_{n,s}^{(k)}})^2 \mathbb{1}\{\mathbf{X}_i \in A_{n,s}^{(k)}\} - \frac{1}{N_n(A_{n,s}^{(k)})} \cdot \sum_{i=1}^{n} (Y_i - \bar{Y}_{A_{n,s,L}^{(k)}} \mathbb{1}\{X_{ji} < z\} - \bar{Y}_{A_{n,s,R}^{(k)}} \mathbb{1}\{X_{ji} \geq z\})^2 \mathbb{1}\{\mathbf{X}_i \in A_{n,s}^{(k)}\},$$

      with $A_{n,s,L}^{(k)} = \{\mathbf{x} \in A_{n,s}^{(k)} | x_j < z\}$, $A_{n,s,R}^{(k)} = \{\mathbf{x} \in A_{n,s}^{(k)} | x_j \geq z\}$ and $\bar{Y}_{A_{n,\ell}^{(k)}}$ denotes the mean of the Y_i's over $A_{n,s}^{(k)}$. For *classification problems*, we have

$$L_n^{(k)}(j,z) = \left\{ \phi_K(\boldsymbol{p}(A_{n,s}^{(k)})) - \frac{N_n(A_{n,s,L}^{(k)})}{N_n(A_{n,s}^{(k)})} \cdot \phi_K(\boldsymbol{p}(A_{n,s,L}^{(k)})) - \frac{N_n(A_{n,s,R}^{(k)})}{N_n(A_{n,s}^{(k)})} \cdot \phi_K(\boldsymbol{p}(A_{n,s,R}^{(k)})) \right\},$$

      where $A_{n,s,L}^{(k)} = \{\mathbf{x} \in A_{n,s}^{(k)} | x_j = z\}$, $A_{n,s,R}^{(k)} = \{\mathbf{x} \in A_{n,s}^{(k)} | x_j \neq z\}$ and $\bar{Y}_{A_{n,\ell}^{(k)}}$ denotes the mode of the Y_i's over $A_{n,s}^{(k)}$ with $\phi_K(\boldsymbol{p}(A)) = \sum_{t=1}^{K} p_t(A)(1 - p_t(A))$ and $p_t(A)$ is the relative fraction of observations with class label ℓ_t, $t \in \{1, \ldots, K\}$ in A;

      Cut the cell $A_{n,s}^{(k)}$ at (j_s^*, z_s^*) resulting into $A_{n,s,L}^{(k)}$ and $A_{n,s,R}^{(k)}$;

      Update: $A_{n,2s-1}^{(k+1)} \longleftarrow A_{n,s,L}^{(k)}$ and $A_{n,2s}^{(k+1)} \longleftarrow A_{n,s,R}^{(k)}$;

      $n_{nodes} = 2^{n_{nodes}}$;

    end

  end

  Set $m_n(\cdot; \Theta_j, \mathcal{D}_n)$ as the j-th constructed tree.

end

Result: Collection of M decision trees $\{m_{n,1}(\cdot; \boldsymbol{\Theta}_j, \mathcal{D}_n)\}_{j=1}^{M}$ used to obtain the aggregate regression estimate $m_{n,M}$ in

Beside point predictions, the Random Forest method can also be considered as a tool for *variable selection*, see e.g. Biau and Scornet (2016); Breiman (2001); Gregorutti et al. (2017); Ishwaran (2007); Louppe et al. (2013). The latter refers to the procedure of extracting an *informative* feature subset $\mathcal{S} \subseteq \{1, \dots, p\}$ such that the regression function $\widetilde{m}(\mathbf{x})$ resp. the classifier $\widetilde{g}(\mathbf{x})$ can be reduced to a smaller domain, say $\mathbf{x}_\mathcal{S} \in \mathbb{R}^s$, with $s = |\mathcal{S}|$ and $\widetilde{m}(\mathbf{x}) = m(\mathbf{x}_\mathcal{S})$ resp. $\widetilde{g}(\mathbf{x}) = g(\mathbf{x}_\mathcal{S})$ for all $\mathbf{x}_\mathcal{S} \in \mathbb{R}^s$, with $\mathbf{x}_\mathcal{S} = (x_i)_{i \in \mathcal{S}}$, $\mathbf{x} = (x_i)_{i=1}^p$. Hence, the term *informative* arises from the fact that features in $\{1, \dots, p\} \setminus \mathcal{S}$ are not required for predicting the outcome for every $\mathbf{x} \in \mathbb{R}^p$, such that they can be left out. Following the description given in Biau and Scornet (2016), the Random Forest delivers actually two measures, that can be used for variable selection procedures in both, regression and classification procedures:

1. The *Mean Decrease in Impurity* (MDI) measures for every variable $j \in \{1, \dots, p\}$ its total decrease in node impurity averaged over the forest. That is, for all $j \in \{1, \dots, p\}$ with $L_{n,s}^{(k)} = L_{n,s,\mathbf{\Theta}_t}^{(k)}$ being the cut criterion constructed with $\mathbf{\Theta}_t$ (see Algorithm 3 for its definition), we have

$$MDI_{n,M}(j) = \frac{1}{M} \sum_{t=1}^{M} \sum_{k,s} \frac{|N_n(A_{n,s}^{(k)}(\cdot; \mathbf{\Theta}_t))|}{n} \cdot L_{n,s}^{(k)}(j, z_j) \cdot \mathbb{1}\left\{ L_{n,s}^{(k)}[j, z_j] \geq L_{n,s}^{(k)}[\ell, z_\ell], \forall \ell \neq j, z_j, z_\ell \right\}. \quad (2.6)$$

2. The *Mean Decrease Accuracy* or *permutation importance* (I) measures the decrease in accuracy for each tree $t = 1, \dots, M$ after permuting the values of a feature $j \in \{1, \dots, p\}$ among the set of Out-of-Bag samples and averages the result over the forest. Denoting with $\mathcal{D}_n^{-(t)} = \mathcal{D}_n^{-(t)}(\mathbf{\Theta}_t)$ the index set of Out-of-Bag samples in tree t with cardinality γ_n such that $\pi_{j,t}$ is a permutation along the j-th feature in tree t, then the permutation importance for the variable $j \in \{1, \dots, p\}$ is given by

$$I_{n,M}^{OOB}(j) = \frac{1}{\gamma_n M} \sum_{t=1}^{M} \sum_{i \in \mathcal{D}_n^{(-t)}} \left\{ \psi\left(Y_i, m_{n,M}^{OOB}(\mathbf{X}_i^{\pi_{j,t}}; \mathbf{\Theta}_t)\right) - \psi\left(Y_i, m_{n,M}^{OOB}(\mathbf{X}_i; \mathbf{\Theta}_t)\right) \right\}, \quad (2.7)$$

   where $\psi(x, y) = (x - y)^2$ is used for regression , i.e. the mean squared error. For classification problems, $\psi(x, y) = \mathbb{1}\{x \neq y\}$ as the misclassification error is used.

For practical problems, lower values of MDI resp. I indicate *less informative* features. However, a well-founded statistical inference procedure in terms of hypothesis tests for selecting variables $j \in \{1, \dots, p\}$ based on these measures are rather sparse resp. do not exist so far. One part of this thesis aims to close this gap by proving statistical properties of such measures that are necessary for constructing suitable test statistics based on $MDI_{n,M}$ resp. $I_{n,M}^{OOB}$ in the future.

2.1 Overview of Theoretical Results

This section aims to shortly summarize the theoretical work conducted so far for Random Forest models. They have mainly focused on central limit theorems and consistency properties such as that described in Chapter 1. First, we will recapture consistency results for both, regression and classification Random Forest models and then state central limit theorems developed for Random Forest models, too.

A. Consistency:

For classification purposes, a formal attend to prove consistency of Random Forest models has been conducted in Biau et al. (2008). Therein, universal weak or strong consistency according to Definition 1.4 in Chapter 1 could not be established. Instead, weak consistency (not in L_1-sense, but in probability) for a smaller class of distributions could be shown, such as features having support on $[0,1]^p$ or having non-atomic marginals. As shown in Proposition 2.1, the assumption that features are supported on $[0,1]^p$ is not severe, such that this restriction does not have a great impact. However, the drawbacks of their results are simplifying assumptions on the tree constructing process of the Random Forest. Basically, they aim to destroy the original dependencies in the cut criterion by introducing random split-point selection procedures being independent of $\mathcal{D}_n$. To be more precise, Biau et al. (2008) could prove weak consistency for certain distributions for the following modifications of the Random Forest:

- For the *purely Random Forest classifier*: During the tree construction process, the split direction j and the split value $z = z_j$ are chosen uniformly at random from $\{1, \ldots, p\}$ resp. from the range of the current cell at consideration. Starting from the root cell, the procedure for cutting a cell and conducting the next cut is also random by uniformly selecting the corresponding leave. The procedure is stopped after $\kappa \geq 1$ iterations, where κ is a fixed and pre-defined parameter. The aggregation to the final prediction is conducted similarly to the Random Forest.

- For the *scale-invariant Random Forest classifier*: The tree construction procedure of the scale-invariant Random Forest classifier is similar to the purely Random Forest classifier, except the procedure for selecting the cut value $z = z_j$. Therein, a more data-dependent cutting strategy is thought. At the current leave, select at random an index i among, lets say $N \leq n$ observations in that leave, and determine the cutting value $z = z_j$ such that the i smallest values of these N observations fall in one cell, and the rest in the other cell (see Biau et al., 2008, page 2020).

Note that the above simplifications of the Random Forest are again different to the extremely randomized tree algorithm *Extra-Trees* in Geurts et al. (2006), where subsampling or bootstrapping during bagging is dropped and additional randomization is introduced within each node of a tree in the ensemble. In Extra-Trees, the split-points are chose uniformly at random from each range of $(\mathbf{X}_j)_{j \in \mathcal{E}}$ and $\mathcal{E}$ is randomly chosen subset of $\{1, \ldots, p\}$. The optimal cut is then conducted based on the same impurity measure as in the Random Forest. In the work of Biau et al. (2008), the authors could show that Breiman's Random Forest classifier is not weakly universally consistent. For both, fully and not fully grown trees (i.e. $t_n < a_n$ resp. $t_n = a_n$), they could find a distribution, such that Breiman's Random Forest classifier is not consistent.

For regression problems, a consistency attempt has been made in Biau (2012) again making simplifying assumptions on the tree construction process. Therein, Biau could prove that the *centered Random Forest* is weakly consistent for distributions with finite second response moment and feature support on $[0,1]^p$. The type of Random Forest considered in the latter work assumes that the split direction $j \in \{1, \ldots, p\}$ is chosen with a potentially data dependent probability $p_{n,j} \in (0,1)$. The cut value $z = z_j$ is then chosen as the midpoint of the current node at consideration. This type of Random Forest will exactly result into 2^k leaves, where k is some initial hyper-parameter. However, in Scornet et al. (2015), the consistency result could be extended to Breiman's original Random Forest. Therein, the authors could prove

for fully and not fully grown trees weak consistency for a certain class of distribution. To be more precise, the results in Scornet et al. (2015) can be summarized as follows: Assuming that

$$Y = \sum_{j=1}^{p} m_j(X_j) + \epsilon, \tag{2.8}$$

where $m_j : [0,1] \longrightarrow \mathbb{R}$ is supposed to be continuous, $\mathbf{X} \sim Unif([0,1]^p)$ and that $\epsilon \sim N(0, \sigma^2)$ with $\sigma^2 \in (0, \infty)$, the infinite Random Forest $m_n(\cdot; \mathcal{D}_n) = m_{n,\infty}(\cdot; \mathcal{D}_n)$ is weakly consistent for distributions resp. data generating processes originating from (2.8), provided that $a_n \to \infty$, $t_n \to \infty$ and $t_n(\log(a_n))^9/a_n \to \infty$. For fully grown trees, i.e. $t_n = a_n$, two additional assumptions were required to establish the weak consistency. However, they are rather difficult to verify in practice and can be recalled in Scornet et al. (2015) on page 1723.

B. Central Limit Theorems:

Central limit theorems based on Random Forest models have mainly been established for regression learning problems. Considering the work of Mentch and Hooker (2016), the finite Random Forest estimate at a fixed prediction point $\mathbf{x} \in \mathbb{R}^p$ can be considered as an infinite order U-statistics with random kernel, since

$$\begin{aligned} m_{n,M_n}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{M_n}, \mathcal{D}_n) &= \frac{1}{M_n} \sum_{t=1}^{M_n} m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_t, \mathcal{D}_n) \\ &= \frac{1}{M_n} \sum_{t=1}^{M_n} m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_t^{(2)}, \mathcal{D}^*_{a_n}(\boldsymbol{\Theta}_t^{(1)})), \end{aligned}$$

where $\mathcal{D}^*_{a_n}(\boldsymbol{\Theta}_t^{(1)})$ is the resampled data set depending on $\boldsymbol{\Theta}_t^{(1)}$. They make the number of decision trees $M = M_n$ depending on the sample size n. Assuming a Lindeberg-type condition together with the side conditions that $n/M_n \to 0$ and $a_n/\sqrt{n} \to 0$ as $n \to \infty$, a pointwise convergence to the standard normal distribution of the following sequence of statistics could be obtained:

$$\sqrt{M_n} \cdot \frac{m_{n,M_n}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{M_n}, \mathcal{D}_n) - \mathbb{E}[m_{n,M_n}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{M_n}, \mathcal{D}_n)]}{\sqrt{a_n^2 \zeta_{1,a_n}}} \xrightarrow{d} N(0,1), \tag{2.9}$$

where $\zeta_{1,a_n} = Cov(m_{n,1}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{a_n}; \boldsymbol{\Theta}_1), m_{n,1}(\mathbf{X}_1, \mathbf{X}_2', \ldots, \mathbf{X}_{a_n}', \boldsymbol{\Theta}_1'))$, with $\mathbf{X}_i'$ and $\boldsymbol{\Theta}_1'$ being independent copies of $\mathbf{X}_i$ and $\boldsymbol{\Theta}_1$ for $i = 2, 3, \ldots, a_n$.
In Wager and Athey (2018) a similar result was proven, but for different underlying assumptions. There, the authors assumed that the data generating process is of the following form:

1. $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ and $\mathbb{E}[Y^2|\mathbf{X} = \mathbf{x}]$ is assumed to be Lipschitz-continuous with $[\mathbf{X}_i^\top, Y_i]^\top \in [0,1]^p \times \mathbb{R}$,
2. $Var(Y|\mathbf{X} = \mathbf{x}) = \sigma^2(\mathbf{x}) \in (0, \infty)$,
3. $\mathbb{E}[|Y - m(\mathbf{x})|^{2+\delta}|\mathbf{X} = \mathbf{x}] \leq C$ for some constants $\delta, C > 0$, uniformly over all $\mathbf{x} \in [0,1]^p$.

Moreover, additional assumptions on the decision trees used in the Random Forest are set such as:

- The decision trees are *honest*, that is, for every sample point $i = 1, \dots, n$ it uses the response variable Y_i only for the computation of leave values or for the decision where to place splits, but not for both.
- The decision trees are α-*regular* with $\alpha \geq 0.2$, that is, each split leaves at least a fraction α of the accessible training examples on each sides of the splitting node and trees are grown to depth $k \in \mathbb{N}$ for some priorly chosen k such that each leaves contains between k and $2k - 1$ observations.
- The decision trees are *symmetric*, that is, the prediction does not depend on the index ordering of the training examples.
- The decision trees are *random-split trees*, that is, the probability that the j-th feature with $j \in \{1, \dots, p\}$ characterizes the next split is bounded below by ν/p for some $0 < \nu \leq 1$.

Given these assumptions and that $a_n/n^\beta \to 1$ for some β with $1 - \left(1 + \frac{p}{\nu} \cdot \frac{\log(\alpha^{-1})}{\log((1-\alpha)^{-1})}\right)^{-1} < \beta < 1$ as $n \to \infty$, Wager and Athey (2018) show asymptotic normality of the Random Forest and that its variance can be consistently estimated using the infinitesimal jackknife estimator proposed in Wager et al. (2014). To be more precise, there exists a sequence $\sigma_n(\mathbf{x})$ of estimators such that the following asymptotic holds under the above assumptions:

$$\frac{m_{n,\infty}(\mathbf{x}; \mathcal{D}_n) - m(\mathbf{x})}{\sigma_n(\mathbf{x})} \xrightarrow{d} N(0, 1), \tag{2.10}$$

as $n \to \infty$, where $\sigma_n(\mathbf{x})$ can be consistently estimated. However, their result is slightly different to the one given in Mentch and Hooker (2016). The latter delivers a central limit theorem for the quantity $\mathbb{E}[m_{n,M_n}(\mathbf{x}; \boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_{M_n}, \mathcal{D}_n)]$ using the finite Random Forest $m_{n,M_n}(\mathbf{x}; \boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_{M_n}, \mathcal{D}_n)$. These results enable the construction of asymptotic prediction intervals for fixed points $\mathbf{x}$. The latter is an important research question for Machine Learning algorithms, since it combines the procedure of prediction with statistical properties such as uncertainty quantification. This, because having a prediction interval at hand, one can deliver statements within what range the prediction at $\mathbf{x}$ might lie given a certain level, say $1 - \alpha$, $\alpha \in (0, 1)$ of coverage. However, the results given in (2.9) and (2.10) lead to different asymptotic prediction intervals. Denoting with $m_{n,M_n}(\mathbf{x}) = m_{n,M_n}(\mathbf{x}; \boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_{M_n}, \mathcal{D}_n)$ such that

$$\mathcal{C}^{(1)}_{n,1-\alpha}(\mathbf{x}) = \left[m_{n,M_n}(\mathbf{x}) - \frac{\sqrt{a_n^2 \zeta_{1,a_n}}}{\sqrt{M_n}} z_{1-\alpha/2}, \quad m_{n,M_n}(\mathbf{x}) + \frac{\sqrt{a_n^2 \zeta_{1,a_n}}}{\sqrt{M_n}} z_{1-\alpha/2} \right]$$

is an interval with $z_{1-\alpha/2}$ as the $1 - \alpha/2$ quantile of the standard normal distribution for $\alpha \in (0, 1)$, we can deduce with the proven results in Mentch and Hooker (2016) that

$$\mathbb{P}[\mathbb{E}[m_{n,M_n}(\mathbf{x}; \boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_{M_n})] \in \mathcal{C}^{(1)}_{n,1-\alpha}(\mathbf{x})] \longrightarrow 1 - \alpha, \quad \text{as } n \to \infty. \tag{2.11}$$

Hence, the set $\mathcal{C}^{(1)}_{n,1-\alpha}(\mathbf{x})$ can be considered as an asymptotic prediction interval for the quantity $\mathbb{E}[m_{n,M_n}(\mathbf{x}; \boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_{M_n}, \mathcal{D}_n)]$, which might be potentially different to $m(\mathbf{x})$. In order to extend the result of Mentch and Hooker (2016) for prediction intervals covering the interesting quantity $m(\mathbf{x})$ with certainty $1 - \alpha$, one needs to establish pointwise consistency of the Random Forest predictor $m_{n,M_n}(\mathbf{x}; \boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_{M_n}, \mathcal{D}_n)$, which is in general not directly given. Therefore, given the assumption for establishing the result in (2.9), one cannot directly

deduce that $\mathcal{C}^{(1)}_{n,1-\alpha}(\mathbf{x})$ is an asymptotic prediction interval for $m(\mathbf{x})$, one might be interested in. This is different to the result given in (2.10). Therein, the interval

$$\mathcal{C}^{(2)}_{n,1-\alpha} = [m_{n,\infty}(\mathbf{x};\mathcal{D}_n) - \sigma_n(\mathbf{x}) \cdot z_{1-\alpha}, \quad m_{n,\infty}(\mathbf{x};\mathcal{D}_n) + \sigma_n(\mathbf{x}) \cdot z_{1-\alpha}]$$

can be considered as an asymptotic prediction interval for the interesting quantity $m(\mathbf{x})$, i.e.

$$\mathbb{P}[m(\mathbf{x}) \in \mathcal{C}^{(2)}_{n,1-\alpha}(\mathbf{x})] \longrightarrow 1-\alpha, \quad \text{as } n \to \infty. \tag{2.12}$$

This, however, comes with the costs of much stronger assumptions affecting also the tree construction process of the Random Forest, which might neither be verifiable nor valid. Especially in practical problems, for which Breiman's original Random Forest is used, simplifying assumptions on the tree construction process such as those stated above are rarely met.

In Scornet (2016) a central limit theorem for the infinite Random Forest could be established. Therein, Scornet shows that for $M \to \infty$ and conditioned on $\mathcal{D}_n$, one has

$$\sqrt{M} \cdot (m_{n,M}(\mathbf{x};\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_M,\mathcal{D}_n) - m_{n,\infty}(\mathbf{x};\mathcal{D}_n)) \xrightarrow{d} N(0,\widetilde{\sigma}^2(\mathbf{x})), \tag{2.13}$$

where $\widetilde{\sigma}^2(\mathbf{x}) = Var_{\boldsymbol{\Theta}}[m_{n,1}(\mathbf{x};\boldsymbol{\Theta},\mathcal{D}_n)] = Var[m_{n,1}(\mathbf{x};\boldsymbol{\Theta},\mathcal{D}_n)|\mathcal{D}_n] \leq 4 \cdot \max_{1\leq i\leq n} Y_i^2$ while assuming some side conditions regarding the *kernel function* of the Random Forest. Scornet also extended the result of the pointwise convergence as given in (2.13) to the whole functional $m_{n,M}(\cdot;\boldsymbol{\Theta}_1,\cdots,\boldsymbol{\Theta}_M,\mathcal{D}_n)$ resulting into convergence to a Gaussian process. Constructing pointwise prediction intervals based on the result given in (2.13) is not straight forward, since convergence holds conditioned on $\mathcal{D}_n$ for an increasing number of decision trees M in the ensemble. For the purpose of completion, we state the following corollary, which extends the result in (2.13) to prediction intervals:

Corollary 2.1. *Consider Breiman's original Random Forest for regression problems on the data set $\mathcal{D}_n$ with $\mathbb{P}[|Y| < \infty] = 1$ and let $\mathbf{x} \in [0,1]^p$ be fixed. Then, the random set $\mathcal{C}^{(3)}_{n,M,1-\alpha}(\mathbf{x})$ given by*

$$[m_{n,M}(\mathbf{x};\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_M,\mathcal{D}_n) - z_{1-\alpha/2} \cdot \widetilde{\sigma}^2(\mathbf{x}), \quad m_{n,M}(\mathbf{x};\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_M,\mathcal{D}_n) + z_{1-\alpha/2} \cdot \widetilde{\sigma}^2(\mathbf{x})]$$

is a valid $1-\alpha$ pointwise and asymptotic prediction interval for the quantity $m_{n,\infty}(\mathbf{x};\mathcal{D}_n)$, as $M \to \infty$ for any finite n and given $\mathcal{D}_n$.

Note that equation (2.13) was not primarily invented for its usage within prediction intervals as proven in Corollary 2.1, but can be considered as a theoretical tool for illuminating the interaction between the finite and infinite Random Forest. However, our result shows that it can be extended to prediction intervals for the quantity $m_{n,\infty}(\mathbf{x})$ given the data $\mathcal{D}_n$. The pleasant property that the sequence $\{\boldsymbol{\Theta}_t\}_{t=1}^M$ is iid given the data $\mathcal{D}_n$ can lead to the estimation of a consistent estimator $\widehat{\sigma}^2_{n,M}$ for $\widetilde{\sigma}^2(\mathbf{x})$ given the data $\mathcal{D}_n$ using the strong law of large numbers, i.e. we have for $M \to \infty$ in almost sure sense given $\mathcal{D}_n$

$$\widehat{\sigma}^2_{n,M}(\mathbf{x}) = \frac{1}{M-1}\sum_{t=1}^{M}(m_{n,1}(\mathbf{x};\boldsymbol{\Theta}_t,\mathcal{D}_n) - m_{n,M}(\mathbf{x};\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_M,\mathcal{D}_n))^2 \longrightarrow \tilde{\sigma}^2(\mathbf{x}). \tag{2.14}$$

This also allows to account for heteroscedasticity in the residual variance. A potential problem of the above result is the conditioning of the prediction interval on the training set $\mathcal{D}_n$. For

a different training set, say $\mathcal{D}_n'$, one can potentially obtain a completely different prediction interval $\mathcal{C}^{(3)}_{n,M,1-\alpha}(\mathbf{x})$ such that the latter is only valid for the training set at hand, namely $\mathcal{D}_n$. Nevertheless, an interesting result in Scornet (2016) under the assumption that $Y = m(\mathbf{x})+\epsilon$, where ϵ is a centered Gaussian noise with variance $\sigma^2 \in (0,\infty)$ and $\|m\|_\infty := \sup_{\mathbf{x}} |m(\mathbf{x})| < \infty$ is the inequality

$$0 \leq R(m_{n,M}) - R(m_{n,\infty}) \leq \frac{8}{M} \cdot \left(\|m\|_\infty^2 + \sigma^2(1+\log(n))\right), \tag{2.15}$$

with $R(m_{n,M}) = \mathbb{E}[(m_{n,M}(\mathbf{X};\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_M,\mathcal{D}_n)-m(\mathbf{X}))^2]$ and $R(m_n) = \mathbb{E}[(m_n(\mathbf{X})-m(\mathbf{X}))^2]$. Equation (2.15) can be used to estimate or bound the error made for choosing a finite number of decision trees. In addition, it can also be used for choosing an appropriate number of decision trees M, while controlling for accuracy. For example, bounding the error by a fixed number, say $\texttt{err} > 0$, one can chose M, such that

$$M \geq \frac{8}{\texttt{err}} \cdot \left(\|m\|_\infty^2 + \sigma^2(1+\log(n))\right).$$

Regarding the above inequality, the quantities $\|m\|_\infty$ and σ^2 are unknown in practice. However, $\|m\|_\infty$ can be estimated through $\max_{1\leq i\leq n} |m^{OOB}_{n,M}(\mathbf{X}_i)|$, while a consistent estimator for the residual variance can be proposed through our work in $(P3)$ resp. the additional work in Section 2.2.
In addition, we will use equation (2.15) in Section 2.2 for the estimation of potential bias resulting from the choice of finite M in uncertainty quantification.

C. Other interesting theoretical results:

So far, we have recalled theoretical properties of Random Forest models that relate to consistency and central limit theorems. Closely connected to the issue of constructing prediction intervals for regression learning problems, the work of Meinshausen (2006) enables the construction of point-wise prediction intervals without explicitly stating (asymptotic) limiting distributions of test statistics depending on the Random Forest predictor. Instead, the quantity of interest in the work of Meinshausen (2006) is the α-Quantile function

$$Q_\alpha(\mathbf{x}) = \inf\{y \quad | \quad F(y|\mathbf{X}=\mathbf{x}) \geq \alpha\}, \tag{2.16}$$

where $F(y|\mathbf{X}=\mathbf{x}) = \mathbb{P}[Y \leq y|\mathbf{X}=\mathbf{x}]$ is the conditional distribution function of Y given the prediction point $\mathbf{x}$ and $\alpha \in (0,1)$. Estimating or extracting knowledge of Q_α requires knowledge of the conditional distribution function $F(y|\mathbf{X}=\mathbf{x})$. However, Meinshausen used the fact that $F(y|\mathbf{X}=\mathbf{x}) = \mathbb{E}[\mathbb{1}\{Y \leq y\}|\mathbf{X}=\mathbf{x}]$, which is the conditional expectation of a modified random variable $\widetilde{Y}(y) = \mathbb{1}\{Y \leq y\}$. Since Random Forest models for regression learning problems aim to approximate the conditional mean $\mathbb{E}[Y|\mathbf{X}=\mathbf{x}]$ by a weighted sum of the form as given in (2.4) resp. (2.5), his idea was to approximate $F(y|\mathbf{X}=\mathbf{x})$ also by a weighted sum. That is, $F(y|\mathbf{X}=\mathbf{x})$ is approximated by

$$\hat{F}_{n,M}(y|\mathbf{x}) = \sum_{i=1}^{n} W_{n_i}(\mathbf{x};\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_M) \cdot \mathbb{1}\{Y_i \leq y\}, \tag{2.17}$$

where the weights are extracted from a Random Forest trained on the initial data set $\mathcal{D}_n$ and afterwards, the response values are substituted by the random variables $\widetilde{Y}_i(y) = \mathbb{1}\{Y_i \leq y\}$, for all $i = 1,\ldots,n$. In order to establish consistency of the form

$$\sup_{y\in\mathbb{R}} |\hat{F}_{n,M}(y|\mathbf{X}=\mathbf{x}) - F(y|\mathbf{X}=\mathbf{x})| \xrightarrow{\mathbb{P}} 0, \quad \text{as } n\to\infty, \tag{2.18}$$

Meinshausen required Lipschitz-continuity and strict monotonicity in y of the conditional distribution function $F(y|\mathbf{X} = \mathbf{x})$, while the features $\mathbf{X}$ are uniformly distributed on $[0, 1]^p$. In addition, assumptions similar to the α-regularity of the underlying decision trees with $\alpha \in (0, 1/2]$ together with the random-split property are set while the node-size behavior is controlled. Empirical results regarding the usage of Meinshausen's *quantile regression forest* has been conducted in Vaysse and Lagacherie (2017), for example. The results in Meinshausen (2006) have the impact that asymptotic $1 - \alpha$ prediction intervals at $\mathbf{x}$ can be computed, i.e.

$$\mathcal{C}^{(4)}_{n,1-\alpha}(\mathbf{x}) := [\widehat{Q}_{n,\alpha/2}(\mathbf{x}), \quad \widehat{Q}_{n,1-\alpha/2}(\mathbf{x})],$$

where $\hat{Q}_{n,\alpha} = \inf\{y \in \mathbb{R} | \hat{F}_{n,M}(y|\mathbf{X} = \mathbf{x}) \geq \alpha\}$ is the plug-in estimate of the quantile function $Q_\alpha(\mathbf{x})$ using $\hat{F}_{n,M}(y|\mathbf{X} = \mathbf{x})$. Due to the result given in (2.18), one has $Q_{n,\alpha}(\mathbf{x}) \longrightarrow Q_\alpha(\mathbf{x})$ in probability, as $n \to \infty$, for arbitrary but fixed $\mathbf{x} \in \mathbb{R}^p$, and $\alpha \in (0, 1)$. Hence, one has $\mathbb{P}[Y(\mathbf{x}) \in [\widehat{Q}_{n,\alpha/2}(\mathbf{x}), \widehat{Q}_{n,1-\alpha/2}(\mathbf{x})]] \longrightarrow 1 - \alpha$, as $n \to \infty$, assuming $Y(\mathbf{x}) = m(\mathbf{x}) + \epsilon$. Again, $\mathcal{C}^{(4)}_{n,1-\alpha}(\mathbf{x})$ is different to $\mathcal{C}^{(1)}_{n,1-\alpha}(\mathbf{x})$, $\mathcal{C}^{(2)}_{n,1-\alpha}(\mathbf{x})$ and $\mathcal{C}^{(3)}_{n,M,1-\alpha}(\mathbf{x})$ for not being a prediction interval of $m(\mathbf{x})$, $\mathbb{E}[m_{n,M_n}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{M_n})]$ or $m_{n,\infty}(\mathbf{x}; \mathcal{D}_n)$ (the latter for fixed $\mathcal{D}_n$), but rather for the true value $Y(\mathbf{x})$. The latter is actually of primary interest when constructing prediction intervals.

2.2 Overview and Implications of Theoretical Results in this Thesis

Our research on theoretical properties of the Random Forest method has mainly focused on two aspects: First, we aimed to quantify uncertainty for Breiman's original Random Forest for the purpose of prediction. This can also cover the construction of prediction intervals. Secondly, we were interested in theoretical properties of the variable selection mechanism, for which Breiman's original Random Forest method is often used as a tool for feature extraction and thus interpretability in potentially high-dimensional settings. Note that this subsection is not directly devoted to the summary of research articles written during the doctoral study program, but rather emphasizes theoretical implications of the developed results including additional findings after their publication and the motivational background that led us to these research fields. Regarding the latter, Chapter 4 will give more insights on the two articles

- Ramosaj, B. and Pauly M., *Consistent estimation of residual variance with random forest Out-Of-Bag errors.*, Statistics and Probability Letters, 151: 49 − 57, 2019,
- Ramosaj B. and Pauly M., *Asymptotic Unbiasedness of Permutation Importance in Random Forest Models.* arXiv preprint arXiv:1912.03306,

closely connected to the theoretical findings in the upcoming sections.

Uncertainty Quantification

The short summary of theoretical results obtained for Random Forest models have revealed that most of the scientific work has been focused on consistency and uncertainty quantification in Random Forest models. Especially for the latter, rather strict assumptions on the tree construction process have been imposed in order to establish central limit theorems for regression problems, that might either be not verifiable, resp. not valid for the original Random Forest considered in Breiman (2001). Also for the consistency results obtained in

Biau et al. (2008) and Biau (2012), rather simplifying assumptions on the tree construction process have been set. An exception was the work of Scornet (2016), which established weak consistency by means of (1.2) for the infinite Random Forest $m_n(\cdot; \mathcal{D}_n)$. Given these results one research question aimed to be tackled in this thesis is

(H1) Is it possible to construct (asymptotic) prediction intervals for Breiman's original Random Forest without imposing stringent assumptions on the tree construction process ?

We focused on regression learning problems, for which we assume that the relationship between the response Y and the features $\mathbf{X}$ are of the form

$$Y = m(\mathbf{X}) + \epsilon, \tag{2.19}$$

where $\mathbf{X}$ is independent of ϵ such that $\mathbb{E}[\epsilon] = 0$, $Var(\epsilon) = \sigma^2 \in (0, \infty)$ and $supp(\mathbf{X}) = [0,1]^p$. For this case, we considered general prediction intervals for $Y(\mathbf{x})$, $\mathbf{x} \in [0,1]^p$ of the form

$$\mathcal{C}(\mathbf{x}) = [m(\mathbf{x}) + q_{\alpha/2} \cdot \sigma, \quad m(\mathbf{x}) + q_{1-\alpha/2} \cdot \sigma], \tag{2.20}$$

where q_α is the α- quantile of ϵ/σ with $\alpha \in (0,1)$. Since the quantities $m(\mathbf{x})$, q_α and σ are in general not known, one requires estimators of them, say $\widehat{m}_n(\mathbf{x})$ and $\widehat{\sigma}_n^2$, fulfilling

(i) $\widehat{m}_n(\mathbf{x}) \longrightarrow m(\mathbf{x})$ in probability, as $n \to \infty$,

(ii) $\widehat{\sigma}_n^2 \longrightarrow \sigma^2$ in probability, as $n \to \infty$.

(iii) In addition, knowledge on the quantiles q_α for $\alpha \in (0,1)$ is required.

Under these assumptions, one might construct a valid and asymptotic $(1-\alpha)$ prediction interval for the response $Y(\mathbf{x})$ for fixed $\mathbf{x}$ the following way:

$$\mathcal{C}_{n,1-\alpha}(\mathbf{x}) = [\widehat{m}_n(\mathbf{x}) + q_{\alpha/2} \cdot \widehat{\sigma}_n^2, \quad \widehat{m}_n(\mathbf{x}) + q_{1-\alpha/2} \cdot \widehat{\sigma}_n^2]. \tag{2.21}$$

For the sake of completion, we state the asymptotic validity of the prediction interval $\mathcal{C}_{n,1-\alpha}(\mathbf{x})$ as a proposition and a formal proof is given at the end of this chapter.

Proposition 2.2. *Let $\alpha \in (0,1)$ and $\mathbf{x} \in [0,1]^p$ be fixed but arbitrary and assume the regression model given in* (2.19) *together with the assumptions* (i) *and* (ii)*. Denoting with q_α the α-quantile of the scaled residuals ϵ/σ. Then, $\mathcal{C}_{n,1-\alpha}(\mathbf{x})$ is an asymptotic $(1-\alpha)$ prediction interval for $Y(\mathbf{x})$.*

Proposition 2.2 enables the construction of valid, asymptotic prediction intervals, if beside our current framework, the assumptions given in (i) and (ii) are given together with Scornet's result of weak consistency such as the one proven in Scornet et al. (2015). The aim was to construct prediction intervals of the form $\mathcal{C}_{n,1-\alpha}(\mathbf{x})$ using Breiman's original Random Forest method. Regarding (i), the work of Scornet et al. (2015) can be considered as an initial step for fulfilling this assumption. It is worth to notice that Scornet et al. (2015) did not establish pointwise consistency as requested in (i), but establishes an *averaged* version of consistency, i.e. $\widehat{m}_{n,\infty}(\mathbf{X}) \to m(\mathbf{X})$ in L_2-sense and thus in probability under some side conditions mentioned one pages $29-30$ of this thesis. Note that the generalization of condition (i) to an averaged convergence, i.e. to $\widetilde{m}_n(\mathbf{X}) \longrightarrow m(\mathbf{X})$ in probability does not prohibit us in constructing prediction intervals. Although it does not allow us to construct pointwise intervals as $\mathcal{C}_{n,1-\alpha}(\mathbf{x})$, one can still construct a prediction interval for the averaged

version of $\mathcal{C}_{n,1-\alpha}(\mathbf{x})$. Denoting with $a(\mathbf{x})$ and $b(\mathbf{x})$ the left and right boundaries of the prediction interval as given in (2.21), an averaged version of the latter is given by $\mathcal{C}_{n,1-\alpha} = [a, b]$, where $a = \int\limits_{supp(\mathbf{X})} a(\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x})$ and $b = \int\limits_{supp(\mathbf{X})} b(\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x})$. Regarding assumption (ii), no well-founded theoretical work based on Breiman's Random Forest method could be found. Therefore, our first theoretical work was focused on deriving an estimator $\hat{\sigma}_n^2$ for σ^2, which makes use of the Random Forest method and is consistent in the sense of (ii). A preliminary work for different residual variance estimators using the Random Forest method was given in Mendez and Lohr (2011) applying those to a bunch of simulation studies and different empirical data. However, no theoretical guarantees regarding its validity could be delivered.

In the article titled *Consistent estimation of residual variance with random forest Out-Of-Bag errors*, which has been published in Statistics and Probability Letters (see Ramosaj and Pauly, 2019b), we proposed three residual variance estimators and prove their consistency in L_1-sense, which implies consistency in the sense of (ii). They are given by the following estimators:

1. $\widehat{\sigma}^2_{RF} := \frac{1}{n}\sum\limits_{i=1}^{n}(\widehat{\epsilon}_i - \bar{\epsilon}_\cdot)^2$, where $\widehat{\epsilon}_i = Y_i - m^{OOB}_{n,\infty}(\mathbf{X}_i; \mathcal{D}_n)$ and $\bar{\epsilon}_\cdot = \frac{1}{n}\sum\limits_{i=1}^{n}\widehat{\epsilon}_i$,
2. $\widehat{\sigma}^2_{RFboot} := \widehat{\sigma}^2_{RF} - \widehat{R}_B(m^{OOB}_{n,\infty})$, where $\widehat{R}_B(m^{OOB}_{n,\infty})$ is a bootstrapped estimator of the Bias of σ^2_{RF} using the Random Forest. For more details on this, we refer to the summary chapter in this thesis resp. to Ramosaj and Pauly (2019b).
3. $\widehat{\sigma}^2_{RFfast} = \widehat{\sigma}^2_{RF} \cdot \left(1 - \frac{1}{a_n^2}\right)$.

All three estimators can be considered as a preliminary step in answering $(H1)$. We assumed weak consistency in the sense of Definition 1.2 such as proven in Scornet et al. (2015) for the infinite Random Forest $m_n(\cdot; \mathcal{D}_n)$, that is,

$$\lim_{n\to\infty} \mathbb{E}[(m_n(\mathbf{X}) - m(\mathbf{X}))^2] = 0. \tag{2.22}$$

Then, together with some side conditions regarding the sampling strategy and the subsampling rate, we have shown that all three estimators are L_1 - consistent. Note that the result does not yet allow to model heteroscedastic variance in the sense that the residual variance σ^2 is not admitted to depend on $\mathbf{x} \in \mathbb{R}^p$ with $\mathbf{x} \in supp(\mathbf{X})$. A byproduct of our theoretical results regarding the consistency of $\widehat{\sigma}^2_{RF}$, $\widehat{\sigma}^2_{RFfast}$ and $\widehat{\sigma}^2_{RFboot}$ are the consistency results for the Out-of-Bag Random Forest estimator $m^{OOB}_{n,M}$ resp. $m^{OOB}_{n,\infty}$. We could show that under the assumption (2.22), the Out-of-Bag estimator $m^{OOB}_{n,\infty}$ is consistent in the sense of Definition 1.2. A more detailed description can be found in Chapter 4, resp. in Ramosaj and Pauly (2019b). However, digging a little deeper than in $(P3)$ and $(P4)$, new information regarding the interpretation of the limiting quantity $m^{OOB}_{n,\infty}$ can be established. In order to unify the frameworks of both articles $(P3)$ and $(P4)$, we set the following proposition and prove it in the last section:

Proposition 2.3. *For any regression problem with training set $\mathcal{D}_n$ on which the Random Forest according to Algorithm 3 is applied, it holds $\mathbb{P}_{\boldsymbol{\Theta}}$ - almost surely that*

$$\lim_{M\to\infty} m^{OOB}_{n,M}(\mathbf{X}_i; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) \longrightarrow \mathbb{E}_{\boldsymbol{\Theta}_1}[m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_1, \mathcal{D}_n)|\Theta^{(1)}_{i,1} = 0]. \tag{2.23}$$

Furthermore, if $\|m\|_\infty < \infty$ and $\mathbb{P}[|Y_1| < \infty] = 1$ holds, then for any independent copy $\mathbf{X}$ of $\mathbf{X}_1$, condition (2.22) implies the consistency of the finite Random Forest in the sense that

$$\mathbb{E}[(m_{n,M}(\mathbf{X}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) - m(\mathbf{X}))^2] \longrightarrow 0, \quad as\ (M, n) \xrightarrow{seq} (\infty, \infty). \tag{2.24}$$

Proposition 2.3 reveals that the infinite OOB Random Forest is nothing else than a sort of conditional expectation over the generic random element $\boldsymbol{\Theta}$. In $(P3)$, this has been denoted as the infinite Random Forest estimate on the reduced training set $\mathcal{D}_{n-1} := \mathcal{D}_n \setminus \{[\mathbf{X}_i^\top, Y_i]^\top\}$ for fixed $i \in \{1, \ldots, n\}$ (see Ramosaj and Pauly (2019b), Lemma 1). In $(P4)$, we denoted the infinite Random Forest as $\mathbb{E}_{\boldsymbol{\Theta}_{[i]}}[m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_{[i]}, \mathcal{D}_n)]$, where $\boldsymbol{\Theta}_{[i]}$ refers to all those random elements $\boldsymbol{\Theta}$, that did not select the i-th observation. Having in mind the first part of Proposition 2.3, the two quantities in $(P3)$ and $(P4)$ referring to the infinite Random Forest actually indicate the infinite OOB Random Forest as the conditional expectation. The second part of Proposition 2.3 reveals that condition (i) is also given for the finite Random Forest, if condition (2.22) together with the two side-conditions on m and Y are true, while the limit acts sequentially. Therefore, constructing (asymptotically) valid prediction intervals also for the finite Random Forest should be possible with Proposition 2.3 and the upcoming work regarding condition (ii). Based on our consistency results of $\widehat{\sigma}^2_{RF}$, $\widehat{\sigma}^2_{RFfast}$ and $\widehat{\sigma}^2_{RFboot}$, while assuming that the residuals follow a Gaussian distribution with mean 0 and residual variance $\sigma^2 \in (0, \infty)$, we can now obtain for every fixed $i = 1, \ldots, n$ the following random sets:

$$\begin{aligned}
\mathcal{C}_{n,1-\alpha,RF} &= [m^{OOB}_{n,\infty}(\mathbf{X}_i) - z_{1-\alpha/2} \cdot \widehat{\sigma}^2_{RF}, \quad m^{OOB}_{n,\infty}(\mathbf{X}_i) + z_{1-\alpha/2} \cdot \widehat{\sigma}^2_{RF}], \\
\mathcal{C}_{n,1-\alpha,RFboot} &= [m^{OOB}_{n,\infty}(\mathbf{X}_i) - z_{1-\alpha/2} \cdot \widehat{\sigma}^2_{RFboot}, \quad m^{OOB}_{n,\infty}(\mathbf{X}_i) + z_{1-\alpha/2} \cdot \widehat{\sigma}^2_{RFboot}], \\
\mathcal{C}_{n,1-\alpha,RFfast} &= [m^{OOB}_{n,\infty}(\mathbf{X}_i) - z_{1-\alpha/2} \cdot \widehat{\sigma}^2_{RFfast}, \quad m^{OOB}_{n,\infty}(\mathbf{X}_i) + z_{1-\alpha/2} \cdot \widehat{\sigma}^2_{RFfast}].
\end{aligned}$$

In addition, we could also show that $\widehat{\sigma}^2_{RF} \geq \widehat{\sigma}^2_{RFfast} \geq \widehat{\sigma}^2_{RFboot}$ holds $\mathbb{P}$-almost surely which then yields to

$$\mathbb{P}[\mathcal{C}_{n,1-\alpha,RFboot} \subseteq \mathcal{C}_{n,1-\alpha,RFfast} \subseteq \mathcal{C}_{n,1-\alpha,RF}] = 1. \tag{2.25}$$

We are now ready to state a theorem, which enables a possible solution for the research question given in $(H1)$.

Theorem 2.1. *Let $\alpha \in (0,1)$ be fixed and assume the regression model given in* (2.19) *such that the residual variance is Gaussian with zero mean and variance $\sigma^2 \in (0, \infty)$. Then we have:*

1. $\lim_{n\to\infty} \mathbb{P}[Y_i \in \mathcal{C}_{n,1-\alpha,RF}] \geq 1 - \alpha$ *for all* $i \in \{1, \ldots, n\}$, *provided that* (2.22) *is valid.*
2. *Additionally assume that the Random Forest is trained using the sampling without replacement scheme of $a_n < n$ observations such that $a_n^2/n \to \infty$ and assume the validity of* (2.22), *then* $\lim_{n\to\infty} \mathbb{P}[Y_i \in \mathcal{C}_{n,1-\alpha,RFfast}] \geq 1-\alpha$ *and* $\lim_{n\to\infty} \mathbb{P}[Y_i \in \mathcal{C}_{n,1-\alpha,RFboot}] \geq 1-\alpha$.

Note that Theorem 2.1 together with equation (2.25) enables a formal ranking of the random sets $\mathcal{C}_{n,1-\alpha,RF}$, $\mathcal{C}_{n,1-\alpha,RFboot}$ and $\mathcal{C}_{n,1-\alpha,RF}$ leading to an optimal choice when $\mathcal{C}_{n,1-\alpha,RFboot}$ is chosen. However, this comes with the cost of additional computational time for the estimation of the bias $\widehat{R}_B(m_n^{OOB})$ using a Random Forest - based bootstrapping scheme as described in Chapter 4, resp. Mendez and Lohr (2011) and Ramosaj and Pauly (2019b). Regarding the interpretability of the intervals $\mathcal{C}_{n,1-\alpha,RF}$, $\mathcal{C}_{n,1-\alpha,RFboot}$ resp. $\mathcal{C}_{n,1-\alpha,RFfast}$, we cannot conclude that given a fixed observational point, say $\mathbf{X} = \mathbf{x} \in supp(\mathbf{X})$, it holds that $\mathbb{P}[Y(\mathbf{x}) \in \mathcal{C}_{n,1-\alpha,RF}], \mathbb{P}[Y(\mathbf{x}) \in \mathcal{C}_{n,1-\alpha,RFboot}], \mathbb{P}[Y(\mathbf{x}) \in \mathcal{C}_{n,1-\alpha,RFfast}] \geq 1 - \alpha$. This would require the construction of pointwise prediction intervals such as $\mathcal{C}^{(1)}_{n,1-\alpha}(\mathbf{x})$, $\mathcal{C}^{(2)}_{n,1-\alpha}(\mathbf{x})$, $\mathcal{C}^{(3)}_{n,M,1-\alpha}$ resp. $\mathcal{C}^{(4)}_{n,1-\alpha}(\mathbf{x})$ which are closely related to pointwise consistency of Breiman's original Random Forest. Our derived intervals, however, can be considered as an *average* prediction interval for $Y(\mathbf{X}) = m(\mathbf{X}) + \epsilon$, since the probability measure $\mathbb{P}$ is also taken over

the feature $\mathbf{X}$. This is different to the pointwise case, where $\mathbb{P}$ is taken over all random components for fixed $\mathbf{X} = \mathbf{x} \in supp(\mathbf{X})$. Regarding the interpretability of the derived intervals $\mathcal{C}_{n,1-\alpha,RF}$, $\mathcal{C}_{n,1-\alpha,RFboot}$ and $\mathcal{C}_{n,1-\alpha,RFfast}$, one can say that on average, future predictions using Breiman's originial Random Forest might lie within the corresponding range with a probability of at least $1-\alpha$. Hence, *average* prediction intervals for Breiman's original Random Forest can be constructed for regression learning problems, since the assumptions for the validity of 2.22 do not exclude the assumptions considered in our work resp. Theorem 2.1. The coverage strength of the derived prediction intervals for practical problems will be considered in a future work. We plan to conduct an extensive simulation study for potential publication.

Additional Results for Residual Variance Estimators

The proposed estimators $\widehat{\sigma}^2_{RF}$ and $\widehat{\sigma}^2_{RFboot}$ have been analyzed regarding their performance within an extensive simulation study in Mendez and Lohr (2011). However, their simulation study is focused on feature dimensions being not larger than $p = 6$. Nonetheless, the Random Forest method is well-known for its capability of treating high-dimensional data, such as $p \approx n$ and $p > n$ problems as well. Therefore, an interesting practical question regarding the performance of the three estimators $\widehat{\sigma}^2_{RF}$, $\widehat{\sigma}^2_{RFboot}$ and $\widehat{\sigma}^2_{RFfast}$ for features being larger than 6 resp. approaching the number of observations n might be of interest. We could find out that the three estimators tend to be positively biased, as the number of feature dimensions increases. Although this seems to be in conflict with the derived theoretical results such as their L_1-consistency, the estimators should be asymptotically unbiased. Now, recalling the theoretical results for the consistency of the residual variance estimators of the last section, there is one potential source of bias, that is controlled in the theoretical part, but not when it comes to the implementation of them in statistical software-packages such as R, SAS or python: The finiteness of the number of decision trees M. Within the construction of the estimators $\widehat{\sigma}^2_{RF}$, $\widehat{\sigma}^2_{RFboot}$ and $\widehat{\sigma}^2_{RFfast}$, it has always been assumed that the number of decision trees tends to infinity, i.e. for constructing each of them, we took the infinite Random Forest estimator $m_{n,\infty}$. In the sequel, we call the effect of biased introduced by a finite choice of M for bagged learners as the *finite-M-bias* theoretically given by $m^{OOB}_{n,M}(\mathbf{X}_1; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) - m^{OOB}_{n,\infty}(\mathbf{X}_1; \mathcal{D}_n)$. Analytically simplifying the finite-M-bias is not directly possible, but we aim to find an upper bound for it and use it as an estimator for the finite-M-bias. Note that the finite-M-bias has also been identified in Wager et al. (2014) as a serious source of inflation for estimators of $Var(m^{OOB}_{n,\infty}(\mathbf{X}_1))$ resp. $Var(m^{OOB}_{n,M}(\mathbf{X}_1))$. Considering $\widehat{\sigma}^2_{\widetilde{RF},M} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\epsilon}^2_{i,M}$ such that $\widehat{\epsilon}_{i,M} = Y_i - m^{OOB}_{n,M}(\mathbf{X}_i; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n)$ we can set up the following proposition:

Proposition 2.4. *Assume the regression model* (2.19) *while* $\|m\|_\infty =: K < \infty$ *together with* (2.22). *Then the estimator* $\widehat{\sigma}^2_{\widetilde{RF},M}$ *is a consistent estimator of* σ^2 *in the following sense*

$$\lim_{n\to\infty} \lim_{M\to\infty} \mathbb{E}[\widehat{\sigma}^2_{\widetilde{RF},M}] = \sigma^2.$$

If the residuals are additionally centered Gaussian with variance σ^2*, then we have*

$$\mathbb{E}[\widehat{\sigma}^2_{\widetilde{RF},M} - \widehat{\sigma}^2_{\widetilde{RF},\infty}] \leq \frac{8}{M} \cdot \left(\|m\|^2_\infty + \sigma^2(1 + \log(n))\right).$$

The second part of Proposition 2.4 uses the result given in Scornet (2016), inequality (2.15) in order to upper-bound the expected value of the finite-M-bias of the residual variance

estimators. It enables the construction of a finite-M-bias corrected estimator given by

$$\widehat{\sigma}^2_{Mcorrect,M} = \widehat{\sigma}^2_{\widetilde{RF},M} - \frac{8}{M} \cdot \left(\left(\max_{i=1,\ldots,n} |m_{n,M}(\mathbf{X}_i; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n)| \right)^2 + \widehat{\sigma}^2_{\widetilde{RF},M}(1 + \log(n)) \right).$$

In addition to the finite-M-bias-corrected estimator $\widehat{\sigma}^2_{Mcorrect}$, we derive another residual variance estimator aimed to control the finite-M-bias in the estimation $\widehat{\sigma}^2_{\widetilde{RF},M}$. This, because $\widehat{\sigma}^2_{Mcorrect}$ uses an upper bound for the *averaged* finite-M-bias, which might lead to an overestimation of the finite-M-bias. Our idea is based on the following heuristics, which we call the *one-step-estimation procedure*:

1. Fit Breiman's Random Forest method with finite $M \in \mathbb{N}$ to the training data $\mathcal{D}_n$ and use the obtained residuals $\{\widehat{\epsilon}_{i,M}\}_{i=1}^n$ to compute $\widehat{\sigma}_{RF,M}$ as an estimator of σ^2, where $\widehat{\sigma}^2_{RF,M}$ is the same residual variance estimator as $\widehat{\sigma}^2_{RF}$, but only with a finite choice of M within the Random Forest.

2. Modify the training data by setting $\widetilde{Y}_i = \widehat{\epsilon}_{i,M}$ for all $i = 1, \ldots, n$ and denote with $\mathcal{D}^{(1)}_{n,M} = \{[\mathbf{X}_i^\top, \widetilde{Y}_i]\}_{i=1}^n$ the modified training data.

3. Fit the same Random Forest model to the modified training data $\mathcal{D}^{(1)}_{n,M}$ and compute its residual variance estimator according to the same logic as for $\widehat{\sigma}^2_{RF,M}$. Denote the obtained residual variance estimator at this stage with $\widetilde{\sigma}^2_{RFstep,M}$.

Step 3 in the one-step-estimation procedure aims to capture potential bias in the estimation of $\widehat{\sigma}^2_{\widetilde{RF},M}$ by refitting Breiman's Random Forest using the same hyper-parameter choice as in the initial fitting step. If the sample size and the number of trees is sufficiently large while assuming the validity of (2.22) and the finiteness of the supremum norm of the regression function m, we can guarantee that there exists a sequence of random vectors, that converges in probability to the random vectors contained in $\mathcal{D}^{(1)}_{\infty,\infty} = \{[\mathbf{X}_i^\top, \epsilon_i]^\top\}_{i \in \mathbb{N}}$. The latter consists of iid random vectors due to model assumptions and has to be understood as a theoretical set, which is rarely accessible in practice. This has the severe effect that conducting a second Random Forest on $\mathcal{D}^{(1)}_{\infty,\infty}$ while assuming the regression model given in (2.19), we can automatically deduce that a Random Forest based residual variance estimator will be at σ^2. Any deviations to this result can be modeled as potential bias covering not only the finite-M-bias, but also finite-n-bias. We will shortly summarize this observation in a proposition.

Proposition 2.5. *Assume the regression model* (2.19) *together with* (2.22) *and* $\|m\|_\infty =: K < \infty$. *Denoting with* $\mathbf{Z}_{i,n,M} = [\mathbf{X}_i^\top, \hat{\epsilon}_{i,n,M}]^\top$, *there exist a sequence of iid random vectors* $\{\mathbf{Z}_{i,\infty,\infty}\}_{i \in \mathbb{N}}$ *such that for every fixed* $i \in \mathbb{N}$,

$$\mathbf{Z}_{i,n,M} \xrightarrow{\mathbb{P}} \mathbf{Z}_{i,\infty,\infty}, \qquad \text{as } (M,n) \xrightarrow{seq} \infty. \tag{2.26}$$

The one-step-estimation procedure enables us to construct a potentially finite-M-bias estimator, then given by

$$\widehat{\sigma}^2_{RFiter,M} = \widehat{\sigma}^2_{RF,M} - \left| \widehat{\sigma}^2_{RF,M} - \widehat{\sigma}^2_{RFstep,M} \right|. \tag{2.27}$$

Establishing consistency for $\widehat{\sigma}^2_{RFiter}$ requires the consistency of the residual variance estimator $\widehat{\sigma}_{RFstep,M}$ in the one-step-estimation procedure. However, since the underlying data set $\mathcal{D}^{(1)}_{n,M}$ is not independent for finite choices of M resp. n, the establishment of consistency result for

this type of estimator is a delicate issue. If the sample size n and the number of decision trees M is large enough, however, we can assume that $\mathbb{P}[\mathbf{Z}_{i,n,M} \in A, \mathbf{Z}_{j,n,M} \in B] \approx \mathbb{P}[\mathbf{Z}_{i,n,M} \in A] \cdot \mathbb{P}[\mathbf{Z}_{j,n,M} \in B]$ for $A, B \in \mathcal{F}$ such that $\mathbb{P}[\mathbf{Z}_{i,n,M} \in \partial A], \mathbb{P}[\mathbf{Z}_{j,n,M} \in \partial B] = 0$ for all $i \neq j \in \{1, \ldots, n\}$. Note that the sequence $\{\mathbf{Z}_{i,n,M}\}_{i=1}^{n}$ is identically distributed for all choices of n, M, since the random vectors in $\mathcal{D}_n$ share the same property of being identically distributed. Considering the data set $\mathcal{D}_{\infty,\infty} = \{\mathbf{Z}_{i,\infty,\infty}\}_{i\in\mathbb{N}}$, Proposition 2.5 indicates that the latter consists of independent and identically distributed random vectors, which is a requirement for the validity of a regression model of similar type as in (2.19). Having a closer look at the proof of Proposition 2.5, the random vectors in $\mathcal{D}_{\infty,\infty}$ are nothing else than $[\mathbf{X}_i^\top, \epsilon_i]^\top$ such that a corresponding regression model would be of the simple form $\epsilon_i = g(\mathbf{X}_i) + \epsilon_i$, where $g \equiv 0$. By the validity of the regression model (2.19) in the initial step, we can also conclude that $\mathbf{X}_i$ is independent of ϵ_i for the regression problem $\epsilon_i = g(\mathbf{X}_i) + \epsilon_i$. Therefore, a Random Forest trained on $\mathcal{D}_{\infty,\infty}$ would theoretically lead to a residual variance estimator of σ^2. In addition to the estimators $\widehat{\sigma}^2_{RFiter,M}$ and $\widehat{\sigma}^2_{Mcorrect,M}$, we propose a weighted residual variance estimator of the form

$$\widehat{\sigma}^2_{RFmiddle,M} = \frac{1}{2}\left(\widehat{\sigma}^2_{RFiter,M} + \widehat{\sigma}^2_{Mcorrect,M}\right), \tag{2.28}$$

which aims to smooth potential negative bias of the estimator $\widehat{\sigma}^2_{Mcorrect,M}$. The negative bias can especially occur in $\widehat{\sigma}^2_{Mcorrect,M}$ due to the averaged upper-bound for the finite-M-bias, which has been used for its estimation. However, it is yet unknown whether $\widehat{\sigma}^2_{RFiter,M}$ will show positive or negative bias such that the choice of $\widehat{\sigma}^2_{RFmiddle,M}$ is more of heuristic nature. In order to evaluate the performance of the three estimators $\widehat{\sigma}^2_{Mcorrect,M}$, $\widehat{\sigma}^2_{RFiter,M}$ and $\widehat{\sigma}^2_{RFmiddle,M}$, we conducted a simulation study covering the regression problem as given in (2.19) while the residuals are assumed to be centered Gaussian with finite variance $\sigma^2 \in (0, \infty)$. As we could show in the article Ramosaj and Pauly (2019a), potential drivers for the accuracy of the estimation quality using the Random Forest method is the *signal-to-noise* ratio formally given by

$$SN = \frac{Var(m(\mathbf{X}))}{\sigma^2}. \tag{2.29}$$

It is important to note that the feature dimension p plays an influential role on the signal-to-noise ratio SN. This happens through the regression function $m : \mathbb{R}^p \to \mathbb{R}$, where an additive structure of the latter yields potentially larger signal-to-noise ratios, when the feature dimension p increases. This is the case for informative features and features that are independent or show non-negative correlation, i.e. $Cov(m_j(X_j), m_i(X_i)) \geq 0$ for all $i, j \in \{1, \ldots, p\}$, where $m(\mathbf{x}) = \sum_{j=1}^{p} m_j(x_j)$. Therefore, we have $SN = SN(p, \sigma^2)$ in fact. Setting the signal-to-noise ratio to $SN \in \{0.5, 1, 2\}$, we determined the scale of the residual variance by considering the following regression functions with $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top$ and $\mathbf{x} = [x_1, \ldots, x_p] \in [0, 1]^p$:

- The linear case, i.e. $m(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$,
- The polynomial case, i.e. $m(\mathbf{x}) = \sum\limits_{j=1}^{p} \beta_j x_j^j$,
- The trigonometric case, i.e. $m(\mathbf{x}) = 2 \cdot \sin(\boldsymbol{\beta}^\top \mathbf{x} + 2)$,
- The non-continuous case, i.e. $m(\mathbf{x}) = \begin{cases} \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, & \text{if } x_3 > 0.5, \\ \sum\limits_{j=4}^{p} \beta_j x_j + 3, & \text{if } x_3 \leq 0.5, \end{cases}$

while using equation (2.29) with $\mathbf{X} \sim Unif([0,1]^p)$. We have chosen the same class of regression functions as in our article ($P4$). Regarding the feature dimension, we assumed that $p \in \{10, 20, 50, 100\}$ whereby $\boldsymbol{\beta}_{p=10} = \boldsymbol{\beta}_0 \in \mathbb{R}^{10}$, $\boldsymbol{\beta}_{p=20} = [\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_0^\top + 2]^\top \in \mathbb{R}^{20}$, $\boldsymbol{\beta}_{p=50} = [\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_0^\top + 2, \boldsymbol{\beta}_0^\top - 4, \boldsymbol{\beta}_0^\top + 3, \boldsymbol{\beta}_0^\top + 4]^\top$ and $\boldsymbol{\beta}_{p=100} = [\boldsymbol{\beta}_{p=50}^\top, 2\boldsymbol{\beta}_{p=50}^\top]^\top$ with $\boldsymbol{\beta}_0 = [2, 4, 2, -3, 1, 3, 4, 1, 5, -5]^\top$. Under this data generating process, different sample sizes $n \in \{50, 500, 1000\}$ have been created using $MC = 1,000$ Monte-Carlo iterations. Regarding the number of base learners in the ensemble, we have chosen to use $M = 1,000$ decision trees in the Random Forest model throughout every simulation set up while restricting the sampling strategy to sampling without replacement of $\lceil 0.632 \cdot n \rceil$ observations. Although the default number of decision trees in the R-function `randomForest` is 500 (Liaw and Wiener, 2002), we doubled its number due to the reasonable increase in computational time. According to the derived results, it is desirable to include more trees in the ensemble rather than less. The latter is only contrary to additional computational time costs (Scornet, 2016). Therefore, $M = 1,000$ should be a reasonable choice. Furthermore, the choice of $M = 1,000$ decision trees was motivated by the interesting question whether the finite-M-bias corrected estimators are capable in smoothening the bias effect while saving computational time for a smaller choice of decision trees. In addition to that, the residual variance estimator based on the sampling variance of the residuals using the least-square estimate for both, the linear case and the polynomial case have been considered for comparison issues. Thus, a correct model specification for the estimation of the residual variance in the linear and polynomial case was assumed. It will be used as a benchmark estimator for the corresponding regression model. Note that the scale of the y-axis of the boxplots (i.e. its range) have been adapted correspondingly, throughout our simulation results.

The results for the **linear model** (see Figure 2.1) indicate that the weighted combination $\widehat{\sigma}^2_{RFmiddle,1000}$ of $\widehat{\sigma}^2_{RFiter,1000}$ and $\widehat{\sigma}^2_{Mcorrect,1000}$ is on average very close to the true residual variance, even for small sample sizes and increasing feature dimensions. It even outperforms the benchmark estimator under correct model specification, if the sample size is small and the feature dimension increases (see e.g. $n = 50$ for $p \in \{10, 20, 50, 100\}$). For cases where $p \geq n$, the benchmark estimator is not even available, while Random Forest based residual variance estimator deliver results that are, on average, close enough to the true residual variance, depending on the SN as well. The residual variance estimators based on Random Forest residuals $\widehat{\sigma}^2_{RF,1000}$ and $\widehat{\sigma}^2_{RFfast,1000}$ tend to overestimate the true residual variance. The boxplots also indicate that the speed of convergence of $\widehat{\sigma}^2_{RF,1000}$ resp. $\widehat{\sigma}^2_{RFfast,1000}$ is rather slow. The positive bias is a result of equation (2.15) using a finite choice of decision trees in the estimation of $\widehat{\sigma}^2_{RF,1000}$ during the simulation procedure, as the boxplots show. Even for larger sample sizes, the estimators $\widehat{\sigma}^2_{RF,1000}$ and $\widehat{\sigma}^2_{RFfast,1000}$ showed positive bias. For larger signal-to-noise ratios, i.e. $SN \geq 1$, the weighted residual variance estimator $\widehat{\sigma}^2_{RFmiddle,1000}$ tend to slightly overestimate the true residual variance. However, the effect vanishes independent of the feature dimension, as the sample size increases (see Appendix-Section 2.4, Figures 2.5 and 2.9).

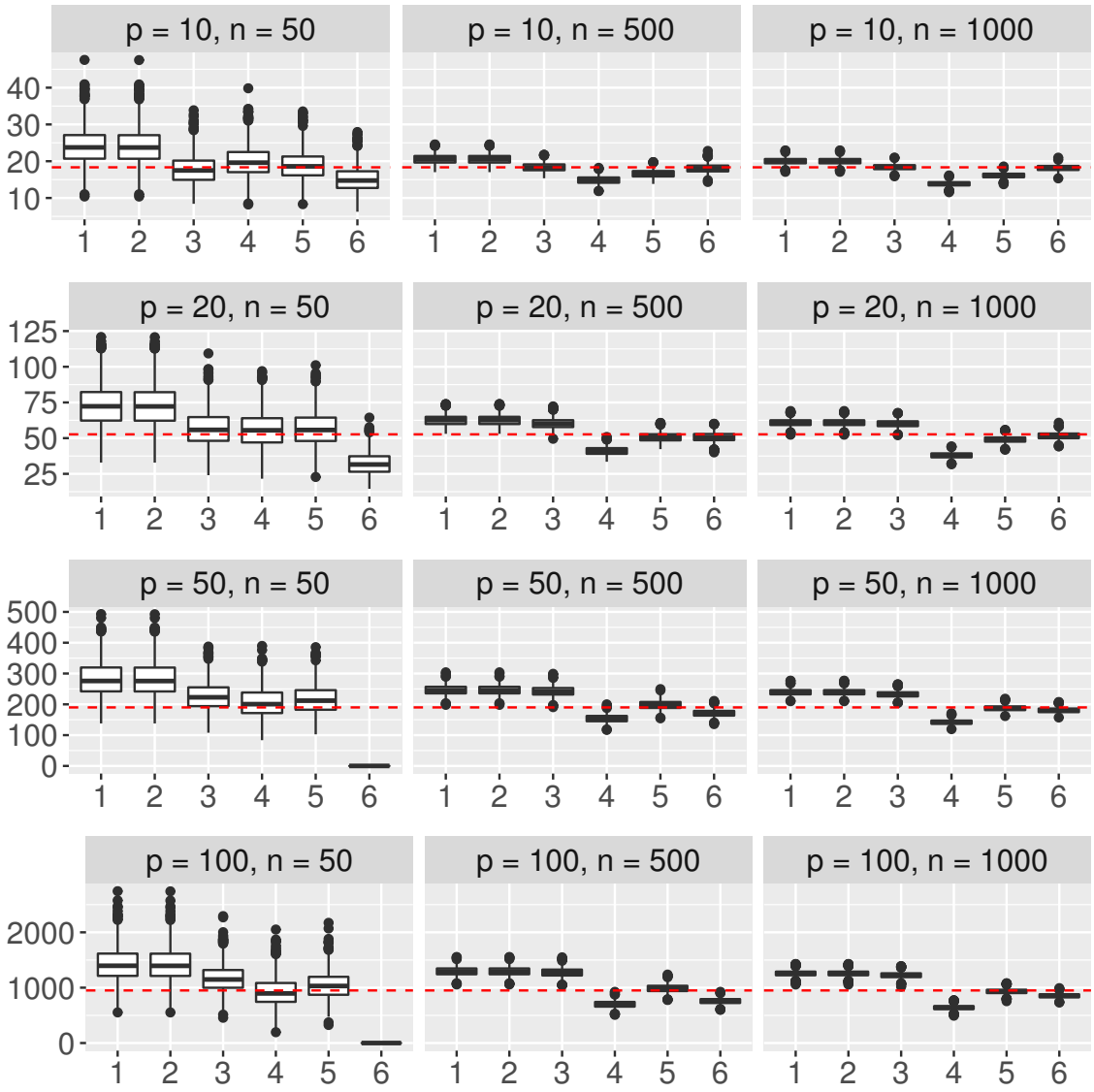


Figure 2.1: Simulation results for the **linear model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN = 0.5}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$ 6 : Benchmark. The red dotted line inidcates the true residual variance.

Regarding the **polynomial model** under a signal-to-noise ratio of $SN = 0.5$ (see Figure 2.2), the weighted residual variance estimator $\widehat{\sigma}^2_{RFmiddle,1000}$ and $\widehat{\sigma}^2_{Mcorrect,1000}$ reveal preferable results being on average close to the true residual variance for small to moderate sample sizes, i.e. $n \in \{50, 500\}$, but revealed a slight inflation when the feature dimension p increased. For larger feature dimensions ($p \geq 50$), the weighted residual variance estimator $\widehat{\sigma}^2_{RFmiddle,1000}$ was slightly overestimating the residual variance. The residual variance estimators $\widehat{\sigma}^2_{Rfiter,1000}$, $\widehat{\sigma}^2_{Mcorrect,1000}$ and $\widehat{\sigma}^2_{RFmiddle,1000}$ beat the benchmark estimator under correct model specification for small and larger sample sizes. In the case of larger sample sizes, the estimator $\widehat{\sigma}^2_{Mcorrect,1000}$ was slightly downsized, but this effect vanished with an increasing feature dimension p. For larger signal-to-noise ratios, i.e. $SN \geq 1$ (see Appendix-Section 2.4, Figures 2.6 and 2.10), the estimators $\widehat{\sigma}^2_{RFiter,1000}$, $\widehat{\sigma}^2_{Mcorrect,1000}$ and $\widehat{\sigma}^2_{RFmiddle,1000}$

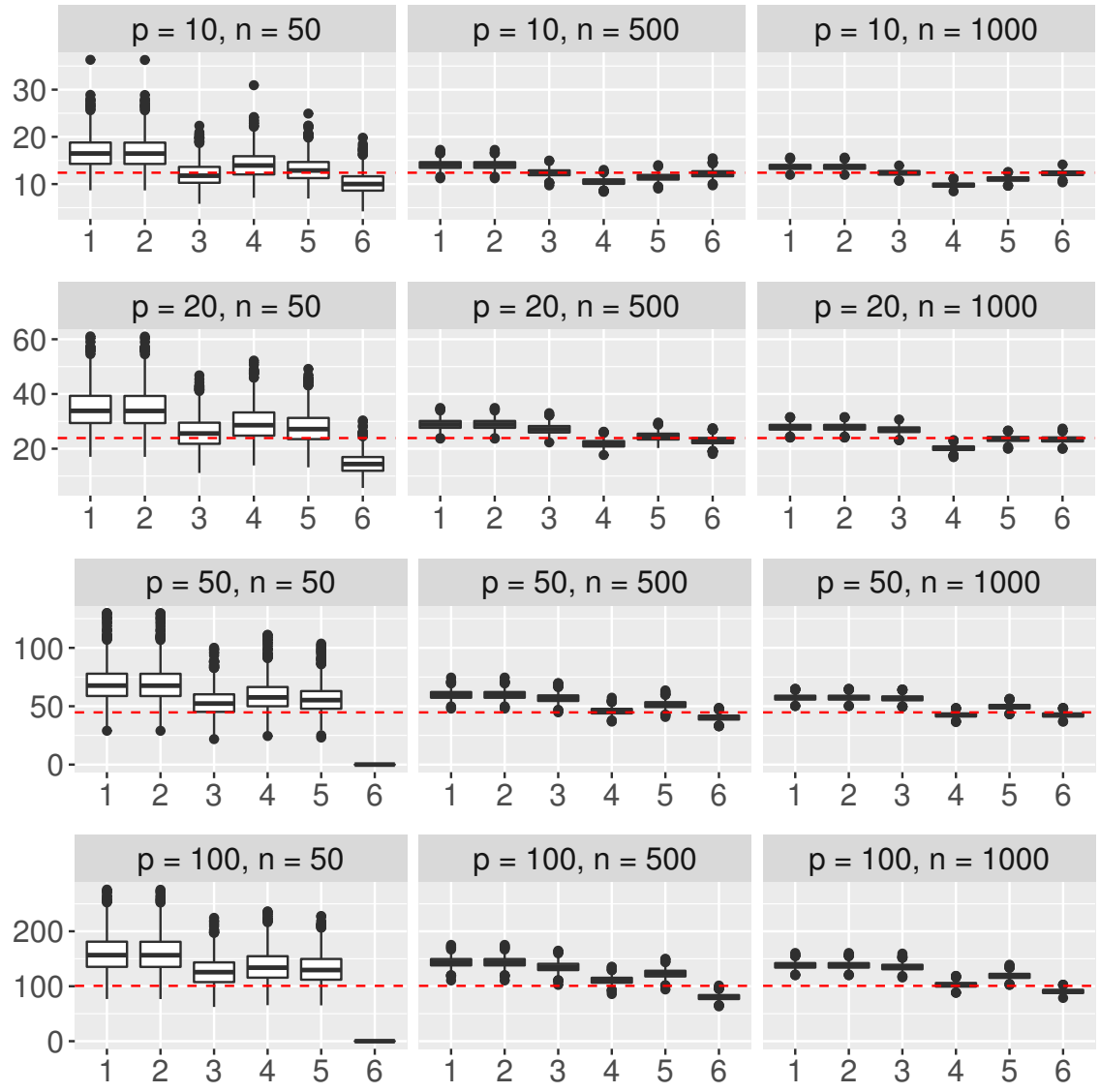


Figure 2.2: Simulation results for the **polynomial model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN = 0.5}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$ 6 : Benchmark. The red dotted line inidcates the true residual variance.

were slightly upsized, when the feature dimension was larger, i.e. $p \geq 50$. However, the effect seemed to vanish, if the sample size was sufficiently large. Under all scenarios, the Random Forest residual variance estimator $\widehat{\sigma}^2_{RF,1000}$ was positively biased indicating the finite-M-bias, even for larger sample sizes.

Turning to the **trigonometric model**, all Random Forest based estimators overestimated the true residual variance for all considered signal-to-noise ratios (see Figure 2.3 and in the Appendix-Section 2.4, Figures 2.7 and 2.11). However, for smaller sample sizes, i.e. $n = 50$, the estimator $\widehat{\sigma}^2_{RFiter,1000}$ resulted into the most preferable results being on average close to the true residual variance, even for larger dimensions $p \geq 50$. For larger sample sizes, the finite-M-bias corrected residual variance estimator $\widehat{\sigma}_{Mcorrect,1000}$ was close to the true resid-

ual variance. Moreover, the weighted residual variance estimator $\widehat{\sigma}^2_{RFmiddle,1000}$ can be still considered as a favorable choice. Similar to the previous results, the Random Forest residual variance estimators $\widehat{\sigma}^2_{RF,1000}$ and $\widehat{\sigma}^2_{RFfast,1000}$ overestimated the residual variance stronger than the other residual variance estimators. This is due to the finite choice of the number of decision trees resulting into a finite-M-bias.

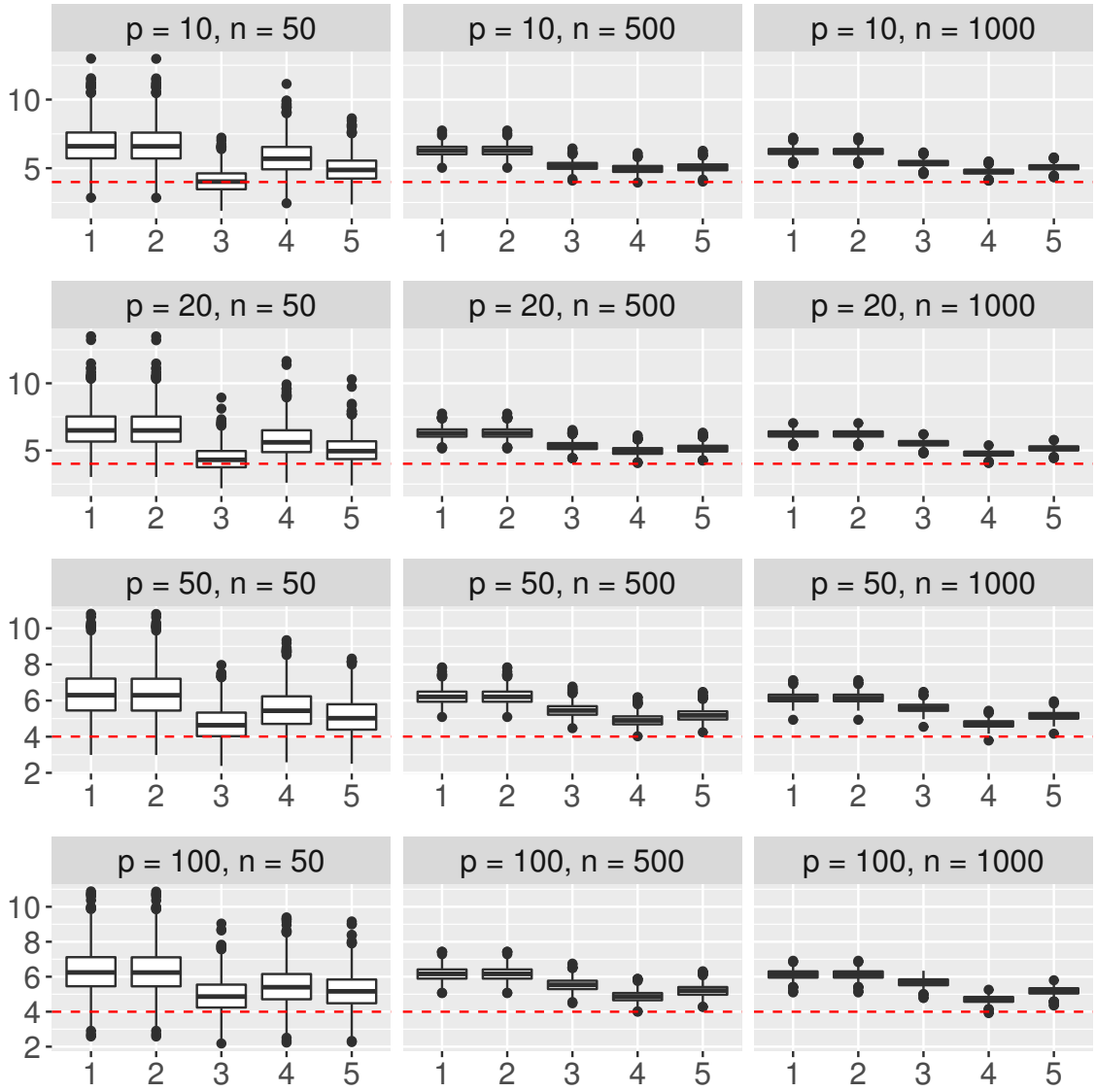


Figure 2.3: Simulation results for the **trigonometric model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN = 0.5}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$. The red dotted line inidcates the true residual variance.

Regarding the **non-continuous model**, the weighted residual variance estimator $\widehat{\sigma}^2_{RFmiddle,1000}$ resulted into the most preferable results across all signal-to-noise ratios (see Figure 2.4 and Figures 2.8 and 2.12 in the Appendix Section 2.4). Nonetheless, $\widehat{\sigma}^2_{RFiter}$ can also be considered as a potential competitor showing similar strong results. Under this model, however, the Random Forest estimators had difficulties in cases where the feature dimension was turning

larger, while the sample size was small (see the cases $p \geq 20$ and $n = 50$ for larger signal-to-noise ratios, e.g. Figure 2.12 in the Appendix Section 2.4). The estimator $\widehat{\sigma}^2_{Mcorrect,1000}$ underestimated the residual variance for larger sample sizes leading to a more suitable choice when the averaged estimator $\widehat{\sigma}^2_{RFmiddle,1000}$ was used. Similar to the previous results, the Random Forest residual variance estimators $\widehat{\sigma}^2_{RF,1000}$ and $\widehat{\sigma}^2_{RFfast,1000}$ overestimated the true residual variance for all considered scenarios, indicating biasedness due to a finite choice of M.

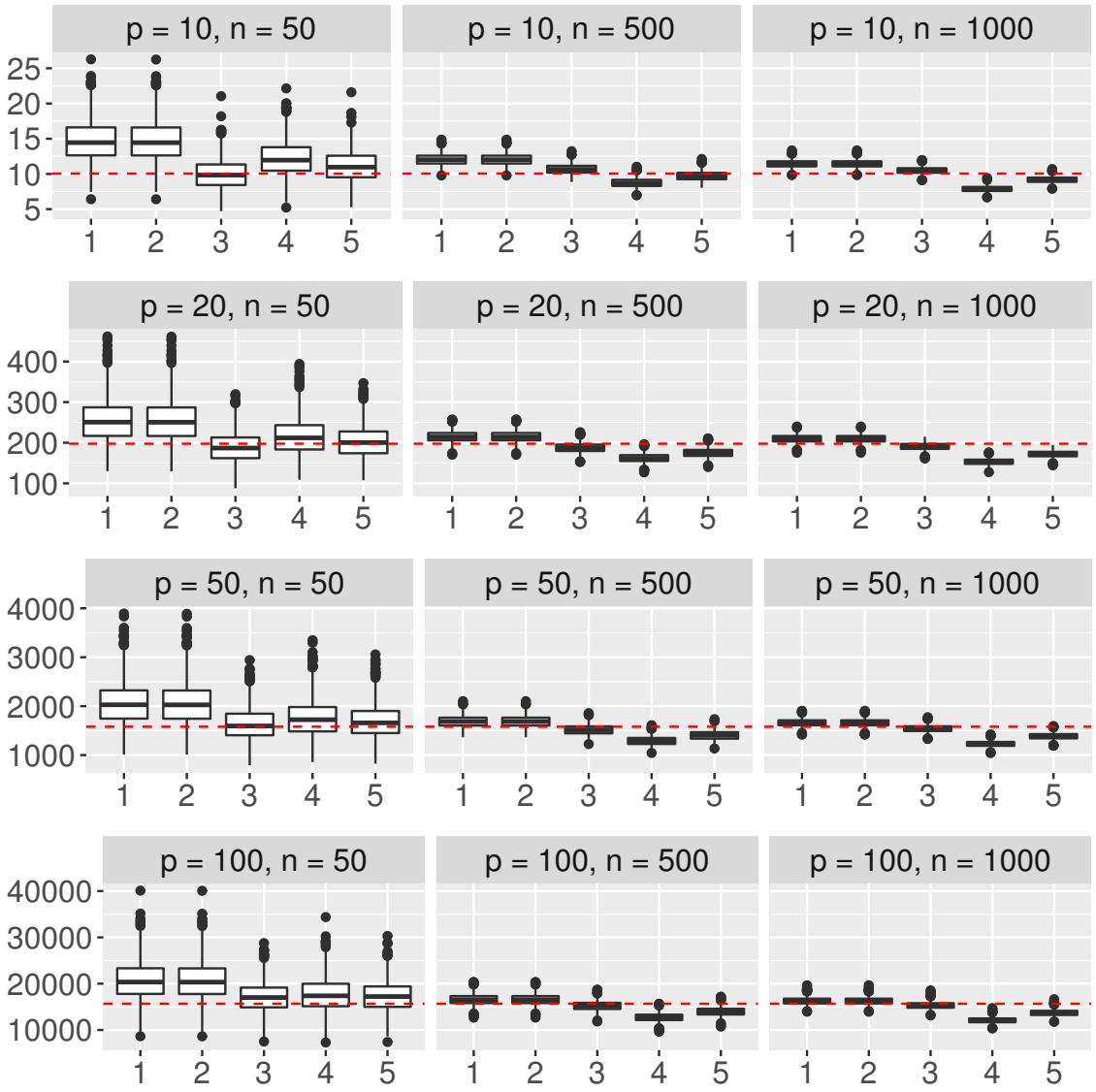


Figure 2.4: Simulation results for the **non-continuous model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN = 0.5}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$. The red dotted line inidcates the true residual variance.

In summary, the finite choice of decision trees in the estimation of the residual variance has to be considered as a serious source of bias. Controlling this effect is therefore important, in order to deliver more accurate results. We derived three residual variance estimators, that aim to control for finite-M-bias, namely $\widehat{\sigma}^2_{RFiter,M}$, $\widehat{\sigma}^2_{Mcorrect,M}$ and $\widehat{\sigma}^2_{RFmiddle,M}$. Our simulation results revealed that the estimator $\widehat{\sigma}^2_{Mcorrect,M}$ tend to underestimate the true residual variance, while the estimator $\widehat{\sigma}^2_{RFiter,M}$ tend to slightly overestimate the residual variance. Therefore, the equal weighting of both estimators leading to $\widehat{\sigma}^2_{RFmiddle,M}$ led to favorable results under various regression functions.

Variable Selection

As mentioned earlier, the Random Forest models are not only used for delivering point predictions. It is also used as a variable selection tool in regression and classification learning problems in both, low and high-dimensional settings. Variable selection plays a key role in many applications, such as the extraction of disease-specific gene expressions in biomedical research or the extraction of main drivers for economic issues such as the interaction of educational level and wealth. For such issues, traditional statistics make use of significance tests for formally testing whether features are related to the response Y by testing whether the corresponding regression coefficient is vanishing. However, such approaches are rather sparse resp. not possible for Machine Learning algorithms such as the Random Forest method, since a proper modeling of the underlying learning problem has not happened prior to fitting. Instead, other measures such as the *mean decrease impurity* or the *permutation measure* play the role of a (vanishing) regression coefficient. Reviewing the literature on the feature extraction mechanism using the measures *mean decrease impurity* or the *permutation measure* of the Random Forest, we found out that most of the work focused on practical applications of these measures. In Strobl et al. (2007) for example, potential bias could be detected in the permutation measure during an extensive simulation study, especially for classification problems. The authors in Gregorutti et al. (2017) considered a formalized version of the permutation importance measure used in Breiman's original Random Forest and prove different identities for the latter. Focusing again on regression learning problems of the form given in (2.19), we were interested in the following research question:

(H2) Beside the criticism on variable importance measures in Random Forest models, can we set up circumstances such that the variable importance measures based on Breiman's Random Forest *correctly* select informative features ? If so, can we deliver theoretical guarantees ?

Regarding research question ($H2$), we found out in Ramosaj and Pauly (2019a) that the permutation importance measure based on Out-of-Bag samples, which is the default measures in statistical software packages such as in R, SAS or python, is asymptotically unbiased, given the following assumptions:

(A1) There is at least *one* informative feature,

(A2) The permutation class is restricted to $\mathcal{V} = \{\pi \in \mathcal{S}_{\gamma_n} : \pi(i) \neq i\}$,

(A3) The features are mutual independent,

(A4) The regression function is bounded,

(A5) The infinite Random Forest is L_2 consistent in the sense of Definition 1.2.

To be more precise, we have found out that there exists a measure $\boldsymbol{I} = [I(1), \ldots, I(p)]^\top \in \mathbb{R}^p$, for which it holds

$$I(j) = \begin{cases} c_j, & \text{if } j \text{ is informative,} \\ 0, & \text{if } j \text{ is not informative,} \end{cases} \tag{2.30}$$

for some $c_j > 0$ defined later such that $\mathbb{E}[I^{OOB}_{n,M}(j)] = 0 = I(j)$ for *non-informative* features j, and for *informative* features j it holds $\lim\limits_{M\to\infty} \mathbb{E}[I^{OOB}_{n,M}(j)] \longrightarrow I(j)$ as $n \to \infty$. The theoretical findings also indicate that simply permuting the observations allowing all kind of permutations might not lead to correct variable selection procedures. The reason to this was the fact that by positive chance, there are observations in the Out-of-Bag set, that will not be permuted, such that the decrease in model accuracy as computed in (2.7) results into vanishing terms, but is then divided by the full cardinality of the Out-of-Bag set. If this effect is sufficiently large, then the convergence to the constant c_j for informative features might not be guaranteed such that instead, it will converge to 0. This theoretical result can be considered as a preliminary step for the following research question, that is not going to be tackled in this thesis, but will be considered for later research:

(H3) Is it possible to use Breiman's originial Random Forest method together with a modified permutation importance measure to conduct statistical inference tests regarding feature significance ?

The research question $(H3)$ requires the finding of a test statistic, that involves both, the theoretical quantity $\boldsymbol{I}$ and the Random Forest permutation importance measure (RFPIM) $I^{OOB}_{n,M}$ in order to be as close as possible to Breiman's original Random Forest. To be more precise, we set for every $j \in \{1, \ldots, p\}$ the following null-hypothesis

$$H_0 : I(j) = 0 \qquad vs. \qquad H_1 : I(j) \neq 0 \tag{2.31}$$

and wonder whether there exist a sequence $\{b_n\}_{n\in\mathbb{N}}$ such that $b_n \nearrow \infty$ as $n \to \infty$ and under the validity of the null-hypothesis, it holds

$$\lim_{n,M\to\infty} \sqrt{b_n} \cdot (I^{OOB}_{n,M}(j) - I(j)) \xrightarrow{d} Z, \tag{2.32}$$

where $\mathbb{E}[Z] = 0$ and $Var(Z) = \zeta \in (0, \infty)$. Such a sequence is likely to be found using central limit theorems. Regarding the latter, we can make use of our findings that $\mathbb{E}[I^{OOB}_{n,M}(j)] = I(j)$ under the null-hypothesis and we are currently working on deriving such results.

2.3 Proofs of the Chapter

Proof of Proposition 2.1. Let $j \in \{1, \dots, p\}$ be fixed but arbitrary. In order to establish the proof, it is necessary to formally define the sequence of transformations $\{F_j\}_{j=1}^{p}$, which depends on the scale of the corresponding random variable. Therefore, we distinguish between the following three cases:

(i) Suppose that X_j is continuous, i.e. it has a continuous density function f_{X_j}. Then define $F_j(x) = F_j|_{supp(X_j)}(x)$ as the distribution function of X_j restricted to its support $supp(X_j)$. Hence, $F_j : supp(X_j) \to [0, 1]$ is strictly increasing.

(ii) Suppose that X_j is metric, but not continuous and has finite support, i.e. $|supp(X_j)| < \infty$. In this case, we assume that X_j is ordinal such that a natural ordering of the elements in $supp(X_j)$ is allowed. Then, we can uniquely define $a_j < \infty$ and $b_j < \infty$ such that $a_j \leq x$ and $b_j \geq x$ for all elements $x \in supp(X_j)$. Hence, we define the function $F_j(x) = \frac{x - a_j}{b_j - a_j}$, which is strictly increasing as well.

(iii) Assume that X_j is not continuous and not metric, e.g. ordinal or nominal, and it has finite support. Without loss of generality, assume that $supp(X_{j,1}) = \{a_1, \dots, a_{K_j}\}$. Then, we can define a mapping $F_j : supp(X_{j,1}) \to [0, 1]$ where $F_j(a_\ell) = 1 - \ell/K_j$ for $\ell \in \{1, \dots, K_j\}$ such that F^{-1} is the unique function mapping the values $\{1 - 1/K_j, 1 - 2/K_j, \dots, 0\} \subset [0, 1]$ to $supp(X_{j,1})$.

We first consider the case for finite M. Therefore, let $t \in \{1, \dots, M\}$ be fixed but arbitrary. Since we condition on the random vector $\boldsymbol{\Theta}_t$ and the set $\mathcal{D}_n$, the re-sampled data $\mathcal{D}_n^*$ and the set for feature subspacing $\mathcal{M}_{try}^{(k)}$ at every level $1 \leq k \leq \lceil \log_2(t_n) \leq \rceil + 1$ are already known. Therefore, the maximization procedure as given in (2.3) turns out to be deterministic. Since the Random Forest constructs hyper-rectangular cells, we obtain for features $j \in \mathcal{M}_{try,s}^{(k)}$ of type (i) and (ii) a region of the form $A_{n,s,j}^{(k)} = [a_{n,s,j}^{(k)}, b_{n,s,j}^{(k)})$. Transforming the features either of type (i) or (ii) using F_j will lead to cells of the form $\widetilde{A}_{n,s,j}^{(k)} = [F_j(a_{n,s,j}^{(k)}), F_j(b_{n,s,j}^{(k)}))$. At every tree level $1 \leq k \leq \lceil \log_2(t_n) \rceil + 1$, the Random Forest algorithms applies its cut criterion on the set $\{Y_i : \mathbf{X}_i \in A_{n,s}^{(k)} = \bigotimes_{r \in \mathcal{M}_{try}^{(k)}} A_{n,s,r}^{(k)}\}$. Now, consider its corresponding index set, i.e. $\mathcal{I}_{n,s}^{(k)} := \{i : \mathbf{X}_i \in A_{n,s}^{(k)}\}$ and the index set of the transformed features of the type (i) or (ii) $\widetilde{I}_{n,s}^{(k)} = \{i : [F_1(X_{1,i}), \dots, F_p(X_{p,i})]^\top \in \widetilde{A}_{n,s}^{(k)}\}$, where $\widetilde{A}_{n,s}^{(k)}$ is defined analogously for the transformed features. We need to show that $\mathcal{I}_{n,s}^{(k)} = \widetilde{I}_{n,s}^{(k)}$ for every tree level k and cell s. Therefore, let $i \in \mathcal{I}_{n,s}^{(k)}$. Since F_j is strictly increasing for both type of features (i) and (ii) given the data $\mathcal{D}_n$, it immediately follows that $F_j(a_{n,s,j}^{(k)}) \leq F_j(X_{j,i}) < F_j(b_{n,s,j}^{(k)})$, hence $i \in \widetilde{\mathcal{I}}_{n,s}^{(k)}$. On the other hand, if $i \in \widetilde{\mathcal{I}}_{n,s}^{(k)}$, then we know that $F_j(a_{n,s,j}^{(k)}) \leq F_j(X_{j,i}) < F_j(b_{n,s,j}^{(k)})$. Applying F^{-1} on the inequalities, it follows that $a_{n,s,j} \leq X_{j,i} < b_{n,s,j}^{(k)}$, since F^{-1} exists and is strictly increasing given the data $\mathcal{D}_n$ for both type of features (i) and (ii).

In case that $j \in \mathcal{M}_{try,s}^{(k)}$ is the feature index for features of type (iii) for some tree level k and region s, the cells in the Random Forest are then of the form $A_{n,s,j}^{(k)} = \{\xi_{n,s,j}^{(k)}\}$ for $\xi_{n,s,j}^{(k)} \in supp(X_{j,1})$. The transformed cells have the form $\widetilde{A}_{n,s,j}^{(k)} = \{F_j(\xi_{n,s,j}^{(k)})\}$. The mapping F_j as given in (iii) can therefore be considered as a re-labeling procedure shrinking the domain to the interval $[0, 1]$. Therefore, $\mathcal{I}_{n,s}^{(k)} = \widetilde{\mathcal{I}}_{n,s}^{(k)}$.

In order to complete the proof for the finite choice of M, we need to show that the terminal node values for trees trained on $\mathcal{D}_n$ or $\widetilde{\mathcal{D}}_n$ are the same, given $\{\boldsymbol{\Theta}_t\}_{t=1}^M$ and $\mathcal{D}_n$. Therefore, consider a terminal node, i.e. $k^* = \lceil \log_2(t_n) \rceil + 1$ and for some $s^* \in \{1, \ldots, t_n\}$, we can immediately follow that $\mathcal{I}_{n,s^*}^{(k^*)} = \widetilde{\mathcal{I}}_{n,s^*}^{(k^*)}$ according to the above results. Hence, the predicted value c_{n,s^*} according to (1.21) in Chapter 1 remains for both, $\mathcal{D}_n$ and $\widetilde{\mathcal{D}}_n$ the same, given $\mathcal{D}_n$ and $\{\boldsymbol{\Theta}_t\}_{t=1}^M$. This leads to $m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) = m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \widetilde{\mathcal{D}}_n)$ given $\mathcal{D}_n$ and $\{\boldsymbol{\Theta}_t\}_{t=1}^M$ for all $\mathbf{x} \in \mathbb{R}^p$, i.e. finite Random Forests are invariant under the monotone transformations F_j, for all $j \in \{1, \ldots, p\}$, given $\mathcal{D}_n$ and $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M$.

For an infinite choice of M, we will receive the infinite Random Forest estimator $m_{n,\infty}(\cdot; \mathcal{D}_n)$. Since the above result is not only valid for $j \in \mathcal{M}_{try,s}^{(k)}$, but for all $j \in \{1, \ldots, p\}$, we can obtain for fixed $\mathcal{D}_n$ the following computations for every $\mathbf{x} \in \mathbb{R}^p$:

$$\begin{aligned} m_{n,\infty}(\mathbf{x}; \mathcal{D}_n) = \mathbb{E}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n) | \mathcal{D}_n] &= \mathbb{E}[\mathbb{E}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n) | \boldsymbol{\Theta}, \mathcal{D}_n] | \mathcal{D}_n] \\ &= \mathbb{E}[\mathbb{E}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \widetilde{\mathcal{D}}_n) | \boldsymbol{\Theta}, \mathcal{D}_n] | \mathcal{D}_n] \\ &= \mathbb{E}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \widetilde{\mathcal{D}}_n) | \mathcal{D}_n] \\ &= \mathbb{E}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \widetilde{\mathcal{D}}_n) | \widetilde{\mathcal{D}}_n] \\ &= m_{n,\infty}(\mathbf{x}; \widetilde{\mathcal{D}}_n). \end{aligned} \tag{2.33}$$

The first equality follows from the definition of the infinite Random Forest. The second equality follows from the measurability of $m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n)$ towards the sigma field generated by $\boldsymbol{\Theta}$ and $\mathcal{D}_n$ and the computation rules of the conditional expectation. The third equality follows from the above results for finite M showing that $m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n) = m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \widetilde{\mathcal{D}}_n)$ holds, given $\mathcal{D}_n$ and $\boldsymbol{\Theta}$. $m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \widetilde{\mathcal{D}}_n)$ remains measurable w.r.t. the sigma field generated by $\boldsymbol{\Theta}$ and $\mathcal{D}_n$, since $\widetilde{\mathcal{D}}_n$ and its elements are transformed values of elements in $\mathcal{D}_n$ using measurable transforms $\{F_j\}_{j=1}^p$, i.e. $\widetilde{\mathcal{D}_n} = \{[F_1(X_{1,i}), \ldots, F_p(X_{p,i}), Y_i]^\top\}_{i=1}^n$. Hence, the second last equality holds.

■

Proof of Corollary 2.1. Consider Breiman's original Random Forest as prescribed in Algorithm 3 and let $\alpha \in (0, 1)$, $\mathbf{x} \in [0, 1]^p$ be fixed. Denote with $z_{1-\alpha/2}$ the $1 - \alpha/2$ quantile of the standard normal distribution. Note that $\{\boldsymbol{\Theta}_t\}_{t=1}^M$ form a sequence of iid random vectors. In addition, we have almost surely

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\Theta}}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n)^2] &= \mathbb{E}_{\boldsymbol{\Theta}} \left[\left(\sum_{i=1}^n W_{n,i}(\mathbf{x}; \boldsymbol{\Theta}) \cdot Y_i \right)^2 \right] \\ &= \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\Theta}}[W_{n,i}(\mathbf{x}; \boldsymbol{\Theta})^2] \cdot Y_i^2 + \sum_{i \neq j} \mathbb{E}_{\boldsymbol{\Theta}}[W_{n,i}(\mathbf{x}; \boldsymbol{\Theta}) W_{n,j}(\mathbf{x}; \boldsymbol{\Theta})] \cdot Y_i Y_j \\ &\leq n \cdot \max_{1 \leq i \leq n} Y_i^2 + n(n-1) \left(\max_{1 \leq i \leq n} |Y_i| \right)^2 \\ &= n^2 \max_{1 \leq i \leq n} |Y_i|^2 < \infty, \end{aligned} \tag{2.34}$$

where the last inequality follows from the assumption that $\mathbb{P}[|Y| < \infty] = 1$. Using the iid property of $\{\boldsymbol{\Theta}\}_{t=1}^M$ and equation (2.34), we can make use of the classical central limit

theorem, while conditioning on $\mathcal{D}_n$ in order to get for $M \to \infty$

$$\begin{aligned}&\sqrt{M} \cdot (m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) - m_{n,\infty}(\mathbf{x}; \mathcal{D}_n)) \\ &= \sqrt{M} \cdot \left(\sum_{t=1}^{M} m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_t, \mathcal{D}_n) - \mathbb{E}_{\boldsymbol{\Theta}}[m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n)] \right) \xrightarrow{d} N(0, \widetilde{\sigma}(\mathbf{x})^2), \end{aligned} \tag{2.35}$$

where $\widetilde{\sigma}(\mathbf{x})^2 = Var_{\boldsymbol{\Theta}}(m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n))$. Therefore, setting $\mathcal{C}^{(3)}_{n,M,1-\alpha}(\mathbf{x})$

$$[m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) - z_{1-\alpha/2} \cdot \widetilde{\sigma}^2(\mathbf{x}), \quad m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) + z_{1-\alpha/2} \cdot \widetilde{\sigma}^2(\mathbf{x})]$$

while conditioning on $\mathcal{D}_n$, we can deduce that

$$\lim_{M\to\infty} \mathbb{P}_{\boldsymbol{\Theta}}[m_{n,\infty}(\mathbf{x}; \mathcal{D}_n) \in \mathcal{C}^{(3)}_{n,M,1-\alpha}(\mathbf{x})] := \lim_{M\to\infty} \mathbb{E}_{\boldsymbol{\Theta}}[\mathbb{1}\{m_{n,\infty}(\mathbf{x}; \mathcal{D}_n) \in \mathcal{C}^{(3)}_{n,M,1-\alpha}(\mathbf{x})\}] = 1 - \alpha.$$

This leads for any finite $n \in \mathbb{N}$ given $\mathcal{D}_n$ to

$$\begin{aligned}\lim_{M\to\infty} \mathbb{P}[m_{n,\infty}(\mathbf{x}; \mathcal{D}_n) \in \mathcal{C}^{(3)}_{n,M,1-\alpha}(\mathbf{x})] &= \lim_{M\to\infty} \mathbb{E}[\mathbb{E}[\mathbb{1}\{m_{n,\infty}(\mathbf{x}; \mathcal{D}_n) \in \widetilde{\mathcal{C}}_{n,M,1-\alpha}(\mathbf{x})\}|\mathcal{D}_n]] \\ &= \mathbb{E}[\lim_{M\to\infty} \mathbb{E}_{\boldsymbol{\Theta}}[\mathbb{1}\{m_{n,\infty}(\mathbf{x}; \mathcal{D}_n) \in \mathcal{C}^{(3)}_{n,M,1-\alpha}(\mathbf{x})\}]] \\ &= \mathbb{E}[1 - \alpha] = 1 - \alpha.\end{aligned}$$

The second equality follows from applying Lebesgue's dominated convergence theorem, since $\mathbb{E}_{\boldsymbol{\Theta}}[\mathbb{1}\{m_{n,\infty}(\mathbf{x}; \mathcal{D}_n) \in \widetilde{\mathcal{C}}_{n,M,1-\alpha}(\mathbf{x})\}|\mathcal{D}_n]$ is bounded by 1. ■

Proof of Proposition 2.2. Let $\mathbf{x} \in \mathbb{R}^p$ and $\alpha \in (0, 1)$ be fixed but arbitrary such that $Y(\mathbf{x}) = m(\mathbf{x}) + \epsilon$. Consider the estimators $\widehat{m}_n(\mathbf{x})$ and $\widehat{\sigma}^2_n$ such that $\widehat{m}_n(\mathbf{x}) \to m(\mathbf{x})$ and $\widehat{\sigma}^2_n \to \sigma^2 > 0$, each in probability, as $n \to \infty$. Hence, we can deduce that $Z_n := \frac{m(\mathbf{x}) - \widehat{m}_n(\mathbf{x})}{\sigma} \longrightarrow 0$ in probability, as $n \to \infty$ by applying the continuous mapping theorem. Therefore, for every sub-sequence $\{n_i\}$, there exist a further common sub-sequence $\{n_{i_j}\}$ such that $Z_{n_{i_j}} \longrightarrow 0$, $\mathbb{P}$-almost surely and $\widehat{\sigma}^2_{n_{i_j}} \longrightarrow \sigma^2$, $\mathbb{P}$-almost surely. This yields to $\sigma/\widehat{\sigma}_{n_{i_j}} \longrightarrow 1$, $\mathbb{P}$- almost surely by the continuous mapping theorem. Hence,

$$T_{n_{i_j}} = \frac{m(\mathbf{x}) - \widehat{m}_{n_{i_j}}(\mathbf{x})}{\sigma} \cdot \frac{\sigma}{\widehat{\sigma}_{n_{i_j}}} \longrightarrow 0, \tag{2.36}$$

$\mathbb{P}$-almost surely. Similarly, we can deduce that $\epsilon/\widehat{\sigma}_{n_{i_j}} \longrightarrow \epsilon/\sigma$, $\mathbb{P}$-almost surely, as $n \to \infty$, which yields $T_{n_{i_j}} + \epsilon/\widehat{\sigma}_n \longrightarrow \epsilon/\sigma$, $\mathbb{P}$ - almost surely. Hence, $T_n + \epsilon/\widehat{\sigma}_n \longrightarrow \epsilon/\sigma$ in probability, as $n \to \infty$. Now, fix $\delta_0 > 0$ and set $W_n = T_n + \epsilon/\widehat{\sigma}_n$. Then we have for $n \to \infty$

$$\begin{aligned}\mathbb{P}[Y(\mathbf{x}) \in \mathcal{C}_{n,1-\alpha}] &= \mathbb{P}\left[q_{\alpha/2} \leq \frac{m(\mathbf{x}) - \widehat{m}_n(\mathbf{x}) + \epsilon}{\widehat{\sigma}_n} \leq q_{1-\alpha/2}\right] \\ &= \mathbb{P}\left[q_{\alpha/2} \leq T_n + \epsilon/\widehat{\sigma}_n \leq q_{1-\alpha/2}\right] \\ &= \mathbb{P}[\{q_{\alpha/2} \leq W_n \leq q_{1-\alpha/2}\} \cap \{|W_n - \epsilon/\sigma| \leq \delta_0\}] + \\ &\quad \mathbb{P}[\{q_{\alpha/2} \leq W_n \leq q_{1-\alpha/2}\} \cap \{|W_n - \epsilon/\sigma| > \delta_0\}] \\ &= \mathbb{P}[\{q_{\alpha/2} \leq (W_n - \epsilon/\sigma) + \epsilon/\sigma \leq q_{1-\alpha/2}\} \cap \{|W_n - \epsilon/\sigma| \leq \delta_0\}] + \\ &\quad \mathbb{P}[\{q_{\alpha/2} \leq W_n \leq q_{1-\alpha/2}\} \cap \{|W_n - \epsilon/\sigma| > \delta_0\}] \\ &= \mathbb{P}[\{q_{\alpha/2} - \delta_0 \leq \epsilon/\sigma \leq q_{1-\alpha/2} + \delta_0\} \cap \{|W_n - \epsilon/\sigma| \leq \delta_0\}] + \\ &\quad \mathbb{P}[\{q_{\alpha/2} \leq W_n \leq q_{1-\alpha/2}\} \cap \{|W_n - \epsilon/\sigma| > \delta_0\}] \\ &\longrightarrow \mathbb{P}[q_{\alpha/2} - \delta_0 \leq \epsilon/\sigma \leq q_{1-\alpha/2} + \delta_0].\end{aligned}$$

The convergence follows, since $\mathbb{P}[|W_n - \epsilon/\sigma| > \delta_0] \to 0$ and therefore, $\mathbb{P}[|W_n - \epsilon/\sigma| \leq \delta_0] \to 1$. Letting $\delta_0 \downarrow 0$ and using the right-continuity property and the monotonicity of the distribution function F of ϵ/σ, we have for $n \to \infty$

$$\begin{aligned}\mathbb{P}[q_{\alpha/2} - \delta_0 \leq \epsilon/\sigma \leq q_{1-\alpha/2} + \delta_0] \overset{n\to\infty}{\longrightarrow} F(q_{1-\alpha/2} + \delta_0) - F(q_{\alpha/2} - \delta_0) &\geq F(q_{1-\alpha/2} + \delta_0) - F(q_{\alpha/2}) \\ &= F(q_{1-\alpha/2}) - \alpha/2 \\ &\overset{\delta_0 \downarrow 0}{\longrightarrow} F(q_{1-\alpha/2}) - \alpha/2 \\ &= 1 - \alpha.\end{aligned}$$

Hence, we have

$$\lim_{n\to\infty} \mathbb{P}[Y(\mathbf{x}) \in \mathcal{C}_{n,1-\alpha}(\mathbf{x})] \geq 1 - \alpha.$$

■

Proof of Proposition 2.3 . Consider an arbitrary regression problem given by a training set $\mathcal{D}_n$ consisting of iid observations $[\mathbf{X}_i^\top, Y_i]$ for $i = 1, \ldots, n$. Fix $i \in \{1, \ldots, n\}$ and let $Z_i = Z_i(M)$ be the number of M regression trees not containing the i-th observation. Then, we can conclude that $Z_i(M) = \sum_{t=1}^{M} \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\}$, where $\{\Theta_{i,t}^{(1)} = 0\}$ refers to the event that the i-th observation has not been selected. Since $\{\Theta_{i,t}^{(1)}\}_{t=1}^{M}$ is a sequence of iid random variables, we can conclude by the strong law of large numbers that $Z_i(M)/M = \frac{1}{M}\sum_{t=1}^{M} \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\} \longrightarrow \mathbb{E}_{\Theta_{i,1}^{(1)}}[\mathbb{1}\{\Theta_{i,1}^{(1)} = 0\}] = \mathbb{P}[\Theta_{i,1}^{(1)} = 0] =: c_n$, whereas

$$c_n = \begin{cases} 1 - a_n/n & \text{for subsampling,} \\ (1 - 1/n)^n & \text{for bootstrapping with replacement.} \end{cases}$$

Assume w.l.o.g. that the first $\{Z_i\}_{i=1}^{M}$ trees do not contain the i-th observation. Then, for the Random Forest Out-of-Bag predictions at $\mathbf{X}_i$ it holds for $M \to \infty$, $\mathbb{P}_{\boldsymbol{\Theta}}$ - almost surely that

$$\begin{aligned} m_{n,M}^{OOB}(\mathbf{X}_i; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) &= \frac{1}{Z_i(M)} \sum_{t=1}^{Z_i(M)} m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_t, \mathcal{D}_n) \\ &= \frac{1}{Z_i(M)/M} \cdot \frac{1}{M} \sum_{t=1}^{M} m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_t, \mathcal{D}_n) \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\} \\ &\longrightarrow \frac{1}{c_n} \cdot \mathbb{E}_{\boldsymbol{\Theta}}[m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}, \mathcal{D}_n) \cdot \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\}] \\ &= \frac{1}{c_n} \cdot \mathbb{P}[\Theta_{i,1}^{(1)} = 0] \cdot \mathbb{E}_{\boldsymbol{\Theta}}[m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}, \mathcal{D}_n) | \Theta_{i,1}^{(1)} = 0] \\ &= \mathbb{E}_{\boldsymbol{\Theta}}[m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}, \mathcal{D}_n) | \Theta_{i,1}^{(1)} = 0]. \end{aligned} \tag{2.37}$$

The first equality is nothing else than the definition of the finite OOB Random Forest. The convergence follows from the strong law of large numbers, since the sequence $\{m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_t, \mathcal{D}_n) \cdot \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\}\}_{t=1}^{M}$ is iid. The second-last equality follows from appplying the law of total expectation.

Regarding the second part of Proposition 2.3, it is assumed that $\sup_{\mathbf{x}\in[0,1]^p} |m(\mathbf{x})| =: K < \infty$. Therefore, we can conclude for any independent copy $\mathbf{X}$ of $\mathbf{X}_1$ almost surely that

$$\begin{aligned}(m(\mathbf{X}) - m_{n,M}(\mathbf{X}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n))^2 &\leq m^2(\mathbf{X}) + 2|m(\mathbf{X})| \cdot |m_{n,M}(\mathbf{X}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n)| \\ &\quad + |m_{n,M}(\mathbf{X}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n)|^2 \\ &\leq K^2 + 2K \max_{1\leq i\leq n} |Y_i| + \left(\max_{1\leq i\leq n} |Y_i|\right)^2 =: C < \infty.\end{aligned}$$

The inequality follows from using the alternative representation of the finite Random Forest as the weighted sum of the response, i.e.

$$\begin{aligned}|m_{n,M}(\mathbf{X}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n)| &= \left|\sum_{i=1}^n W_{n,i}(\mathbf{X}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M) \cdot Y_i\right| \\ &\leq \sum_{i=1}^n W_{n,i}(\mathbf{X}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M)|Y_i| \\ &= \max_{1\leq i\leq n} |Y_i| \cdot \sum_{i=1}^n W_{n,i}(\mathbf{X}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_m) = \max_{1\leq i\leq n} |Y_i|.\end{aligned}$$

Note that the quantity C is independent of M and $\boldsymbol{\Theta}$ such that given $\mathcal{D}_n$ and $\mathbf{X}$, C is constant and finite. Therefore, applying Lebesgue's dominated convergence theorem yields to

$$\begin{aligned}&\lim_{n\to\infty} \lim_{M\to\infty} \mathbb{E}[(m(\mathbf{X}) - m_{n,M}(\mathbf{X}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n))^2] \\ &= \lim_{n\to\infty} \mathbb{E}\left[\lim_{M\to\infty} (m(\mathbf{X}) - m_{n,M}(\mathbf{X}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n))^2\right] \\ &= \lim_{n\to\infty} \mathbb{E}[(m(\mathbf{X}) - m_n(\mathbf{X}))^2] = 0.\end{aligned}$$

The second last equality follows from the continuity of the quadratic function and the last equality from condition (2.22) proved in Scornet et al. (2015). ■

Proof of Theorem 2.1. Consider the first case where the residual variance estimator $\widehat{\sigma}^2_{RF}$ is considered. We could show in Ramosaj and Pauly (2019b) that given the assumption that Random Forests are consistent in the sense of (2.22), such as Breiman's original Random Forest as proven in Scornet et al. (2015), while assuming a regression model of the form (2.19), the residual variance estimator $\widehat{\sigma}^2_{RF}$ is L_1-consistent, which implies consistency in probability. Hence, assumption (ii) is valid for constructing prediction intervals of the form (2.20). Assumption (i) is implied by (2.22), which yields to $m_n(\mathbf{X}) \longrightarrow m(\mathbf{X})$ in probability, as $n \to \infty$. Assuming that the residuals are Gaussian with mean 0 and residual variance $\sigma^2 \in (0, \infty)$ fulfills the assumption given in (iii) such that $q_{\alpha/2} = -z_{1-\alpha/2}$ due to the symmetry of the Gaussian distribution and $q_{1-\alpha/2} = z_{1-\alpha/2}$, where z_α is the α-quantile of the standard normal distribution. The validity of Proposition 2.2 is not restricted to fixed features $\mathbf{X} = \mathbf{x}$, but can be extended to the unconditional version as well. Assuming that (i) holds also for the unconditional version $\widehat{m}_n(\mathbf{X})$, Proposition 2.2 should still be valid. Since this is the case, we can immediately follow that $\mathbb{P}[Y(\mathbf{X}_1) \in \mathcal{C}_{n,1-\alpha,RF}] \geq 1 - \alpha$. Therefore, $\mathcal{C}_{n,1-\alpha,RF}$ can be considered as an average prediction interval for $Y(\mathbf{X}_1)$.
Now considering the second case, i.e. 2., we could show under these assumptions that the residual variance estimators $\widehat{\sigma}^2_{RFboot}$ and $\widehat{\sigma}^2_{RFfast}$ are both L_1- consistent and therefore implying consistency in probability such that (ii) is valid. Since (2.22) is also assumed, we can immediately deduce that (i) is valid leading to $\mathbb{P}[Y(\mathbf{X}) \in \mathcal{C}_{n,1-\alpha,RFboot}], \mathbb{P}[Y(\mathbf{X}) \in \mathcal{C}_{n,1-\alpha,RFfast}] \geq 1 - \alpha$ by the same argumentation as in the first case. ■

Proof of Proposition 2.4. Let $Y = m(\mathbf{X}) + \epsilon$ such that $\|m\|_\infty = K < \infty$. Consider the finite Random Forest estimate using OOB samples, i.e.

$$m_{n,M}^{OOB}(\mathbf{X}_1; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) = \sum_{i=1}^{n} W_{n,i}^{OOB}(\mathbf{X}_1; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M; \mathcal{D}_n) Y_i, \tag{2.38}$$

where $W_{n,i}^{OOB}(\mathbf{X}_1; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) = \frac{1}{Z_1(M)} \sum_{t=1}^{Z_1(M)} W_{n,i}(\mathbf{X}_1; \boldsymbol{\Theta}_t, \mathcal{D}_n)$ and $W_{n,i}^{OOB}(\mathbf{X}_1; \boldsymbol{\Theta}_t, \mathcal{D}_n) = \frac{\mathbb{1}\{\mathbf{X}_i \in A_n^{OOB}(\mathbf{X}_1; \boldsymbol{\Theta}_t)\}}{N_n(A_n^{OOB}(\mathbf{X}_1; \boldsymbol{\Theta}_t))}$. In this representation, we used w.l.o.g. that the first $Z_1(M)$ trees have not used observation $\{[\mathbf{X}_1^\top, Y_1]^\top\}$. This representation enables us to find an upper bound of $m_{n,M}^{OOB}$, that is independent of M, i.e.

$$\begin{aligned} \left| m_{n,M}^{OOB}(\mathbf{X}_1; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) \right| &\leq \sum_{i=1}^{n} W_{n,i}^{OOB}(\mathbf{X}_1; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n) \cdot |Y_i| \\ &\leq \max_{1 \leq i \leq n} |Y_i| =: f_n < \infty. \end{aligned} \tag{2.39}$$

Setting $m_{n,M}(\mathbf{X}_1) = m_{n,M}(\mathbf{X}_1; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n)$ and using (2.39) will lead to

$$|m(\mathbf{X}_1) - m_{n,M}(\mathbf{X}_1)| \leq |m(\mathbf{X}_1)| + |m_{n,M}^{OOB}| < K + f_n < \infty. \tag{2.40}$$

It follows that $|m(\mathbf{X}_1) - m_{n,M}(\mathbf{X}_1)|^2 < (K + f_n)^2$ due to the monotonicity of the quadratic function on the non-negative real line. The obtained upper bound is independent of M, such that Lebesgue's dominated convergence theorem can be used to obtain

$$\begin{aligned} \lim_{M \to \infty} \mathbb{E}[\widehat{\sigma}^2_{\widetilde{RF},M}] &= \lim_{M \to \infty} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(m(\mathbf{X}_i) + \epsilon_i - m_{n,M}^{OOB}(\mathbf{X}_i; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_n))^2] \right\} \\ &= \lim_{M \to \infty} \left\{ \mathbb{E}[(m(\mathbf{X}_1) - m_{n,M}^{OOB}(\mathbf{X}_1))^2] + 2\mathbb{E}[\epsilon_1 \cdot (m(\mathbf{X}_1) - m_{n,M}(\mathbf{X}_1))] + \sigma^2 \right\} \\ &= \lim_{M \to \infty} \mathbb{E}[(m(\mathbf{X}_1) - m_{n,M}^{OOB}(\mathbf{X}_1))^2] + \sigma^2 \\ &= \mathbb{E}[(m(\mathbf{X}_1) - m_n^{OOB}(\mathbf{X}_1))^2] + \sigma^2. \end{aligned} \tag{2.41}$$

The second equality follows from the identical distribution of the sequence $\{\widehat{\epsilon}^2_{i,M}\}_{i=1}^n$, while the third equality from the independence of ϵ_1 and $m_{n,M}^{OOB}(\mathbf{X}_1)$ together with $\mathbb{E}[\epsilon_1] = 0$ and the finiteness of $\mathbb{E}[m(\mathbf{X}_1) - m_{n,M}^{OOB}(\mathbf{X}_1)]$. Using the result given in Ramosaj and Pauly (2019b) which implies the weak convergence of the OOB Random Forest estimate in the sense of Definition 1.2, when condition (2.22) is fulfilled, will lead to the convergence of the estimator $\widehat{\sigma}^2_{\widetilde{RF},M}$ to σ^2, i.e.

$$\lim_{n \to \infty} \mathbb{E}[\widehat{\sigma}^2_{\widetilde{RF},M}] = \sigma^2 + \lim_{n \to \infty} \mathbb{E}[(m(\mathbf{X}_1) - m_n^{OOB}(\mathbf{X}_1))^2] = \sigma^2. \tag{2.42}$$

If the residuals are Gaussian with finite variance $\sigma^2 \in (0, \infty)$, then we can make use of Theorem 3.3 in Scornet (2016), in particular inequality (2.15) which leads to

$$\begin{aligned} \mathbb{E}[\widehat{\sigma}^2_{\widetilde{RF},M} - \widehat{\sigma}^2_{\widetilde{RF},\infty}] &= \mathbb{E}[(Y_1 - m_{n,M}^{OOB}(\mathbf{X}_1))^2 - (Y_1 - m_n^{OOB}(\mathbf{X}_1))^2] \\ &= \mathbb{E}[(m(\mathbf{X}_1) - m_{n,M}^{OOB}(\mathbf{X}_1))^2] - \mathbb{E}[(m(\mathbf{X}_1) - m_n^{OOB}(\mathbf{X}_1))^2] \\ &\leq \frac{8}{M} \cdot \left(\|m\|_\infty^2 + \sigma^2(1 + \log(n)) \right). \end{aligned} \tag{2.43}$$

The second equality follows from applying $Y_1 = m(\mathbf{X}_1) + \epsilon_1$ together with the independence of ϵ_1 and $m_{n,M}^{OOB}(\mathbf{X}_1)$ resp. ϵ_1 and $m_n^{OOB}(\mathbf{X}_1)$. ■

Proof of Proposition 2.5. Let $\mathbf{Z}_{i,n,M} = [\mathbf{X}_i^\top, \widehat{\epsilon}_{i,n,M}]^\top$ and define $\mathbf{Z}_{i,\infty,\infty} = [\mathbf{X}_i^\top, \epsilon_i]^\top$, where $\widehat{\epsilon}_{i,n,M} = Y_i - m_{n,M}^{OOB}(\mathbf{X}_i)$. Recall from the proof construction in Proposition 2.4 , especially equations (2.41) and (2.42) that

$$\lim_{n\to\infty} \lim_{M\to\infty} \mathbb{E}[(m(\mathbf{X}_1) - m_{n,M}^{OOB}(\mathbf{X}_1))^2] = 0. \tag{2.44}$$

The identity applies here as well, since regression model (2.19), condition (2.22) and $\|m\|_\infty < \infty$ holds. Therefore, take $\delta > 0$ and $i \in \{1, \ldots, n\}$. Then we have

$$\begin{aligned}
\lim_{n\to\infty} \lim_{M\to\infty} \mathbb{P}[\|Z_{i,n,M} - Z_{i,\infty,\infty}\| > \delta] &= \lim_{n\to\infty} \lim_{M\to\infty} \mathbb{P}[|\widehat{\epsilon}_{i,M} - \epsilon_i| > \delta] \\
&= \lim_{n\to\infty} \lim_{M\to\infty} \mathbb{P}[|Y_i - m_{n,M}^{OOB}(\mathbf{X}_i) - Y_i + m(\mathbf{X}_i)| > \delta] \\
&= \lim_{n\to\infty} \lim_{M\to\infty} \mathbb{P}[|m(\mathbf{X}_i) - m_{n,M}^{OOB}(\mathbf{X}_i)| > \delta] \\
&\leq \lim_{n\to\infty} \lim_{M\to\infty} \frac{\mathbb{E}[|m(\mathbf{X}_i) - m_{n,M}^{OOB}(\mathbf{X}_i)|^2]}{\delta^2} \\
&= 0.
\end{aligned} \tag{2.45}$$

The inequality follows by applying Markov's inequality and the last step from (2.44). Hence, $\mathbf{Z}_{i,n,M} \xrightarrow{\mathbb{P}} \mathbf{Z}_{i,\infty,\infty}$ as $(M,n) \xrightarrow{seq} (\infty,\infty)$. Since $\{\mathbf{X}_i\}_{i\in\mathbb{N}}$ and $\{\epsilon_i\}_{i\in\mathbb{N}}$ are both independent and identically distributed, the data set $\mathcal{D}_{\infty,\infty} = \{\mathbf{Z}_{i,\infty,\infty}\}_{i\in\mathbb{N}}$ consists of independent and identically distributed random vectors, too. ■

2.4 Appendix of the Chapter

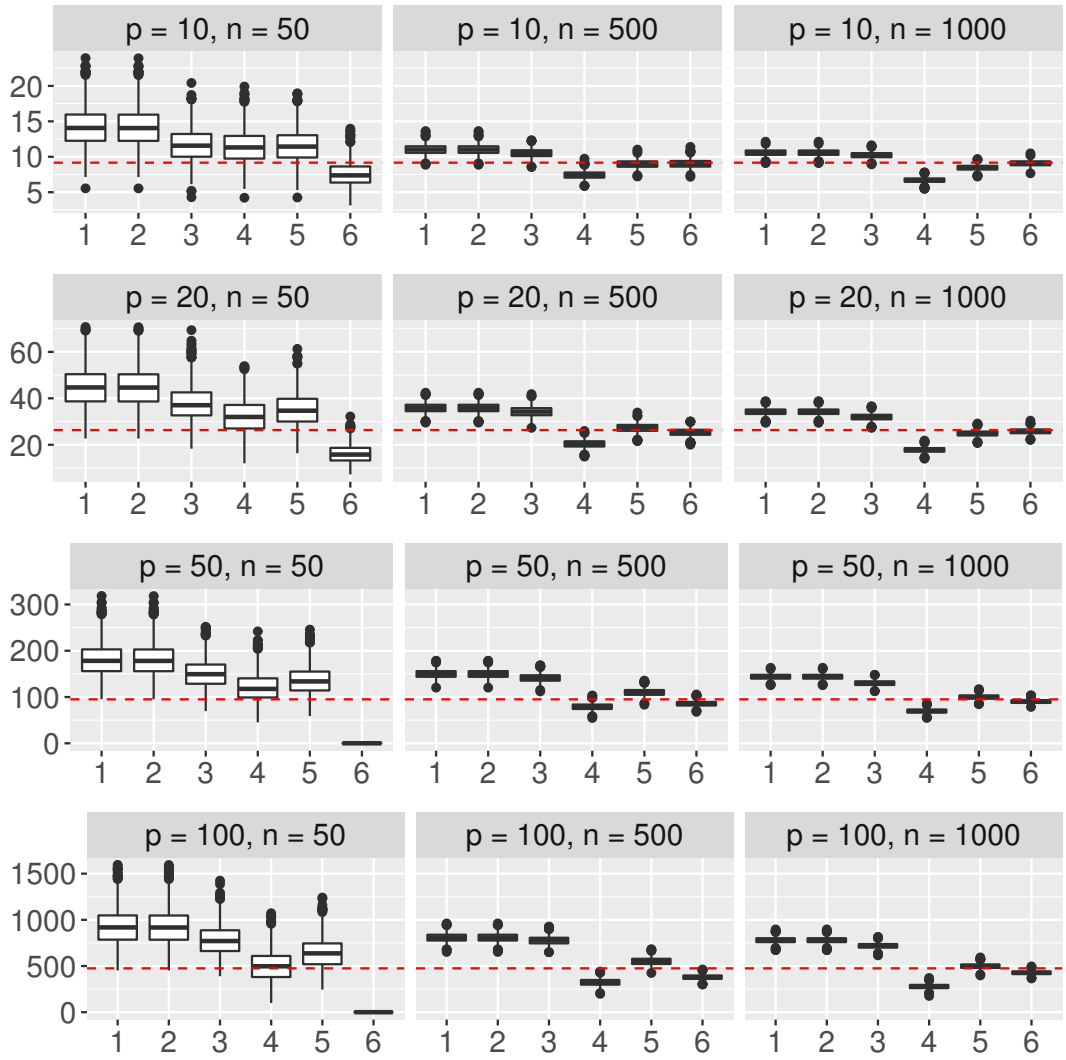


Figure 2.5: Simulation results for the **linear model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN} = \mathbf{1}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$ 6 : Benchmark. The red dotted line inidcates the true residual variance.

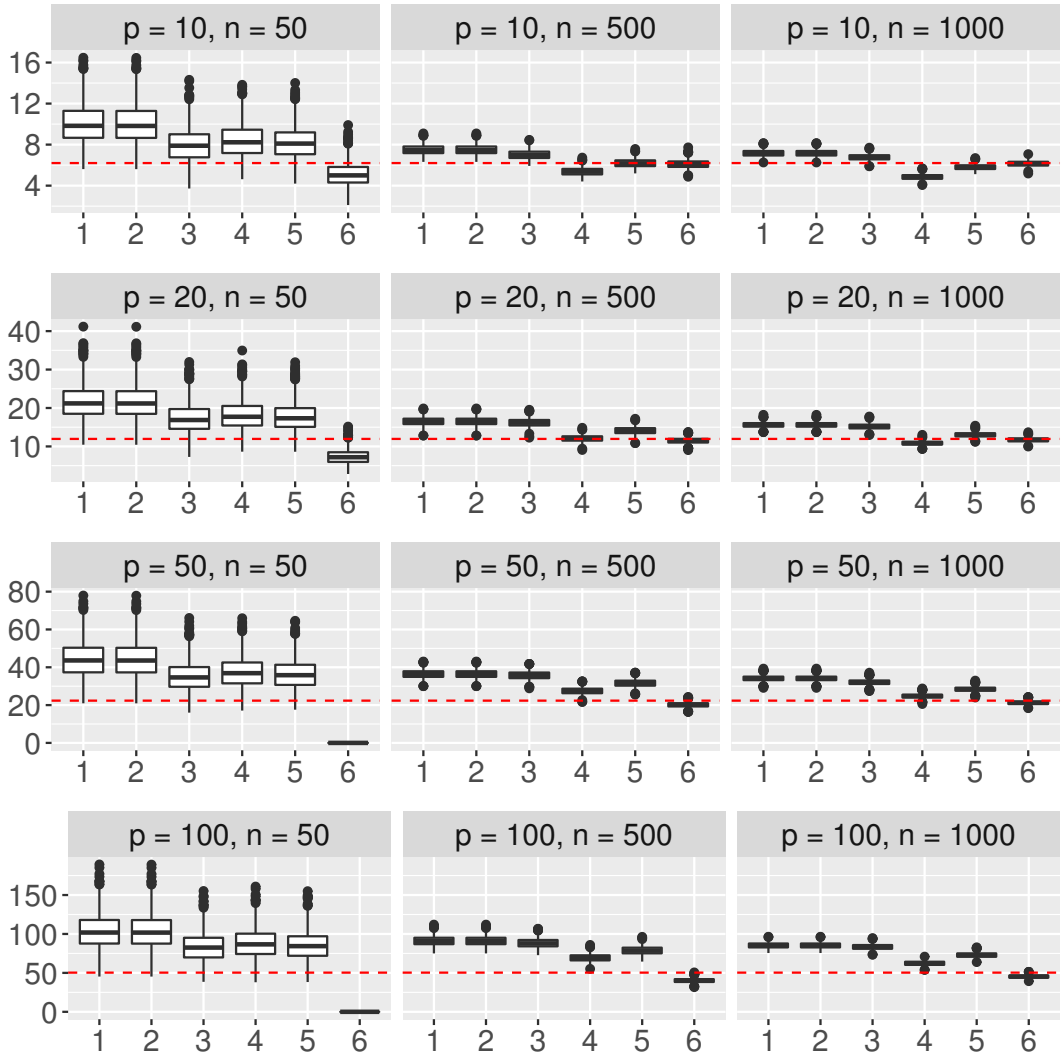


Figure 2.6: Simulation results for the **polynomial model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN} = \mathbf{1}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$ 6 : Benchmark. The red dotted line inidcates the true residual variance.

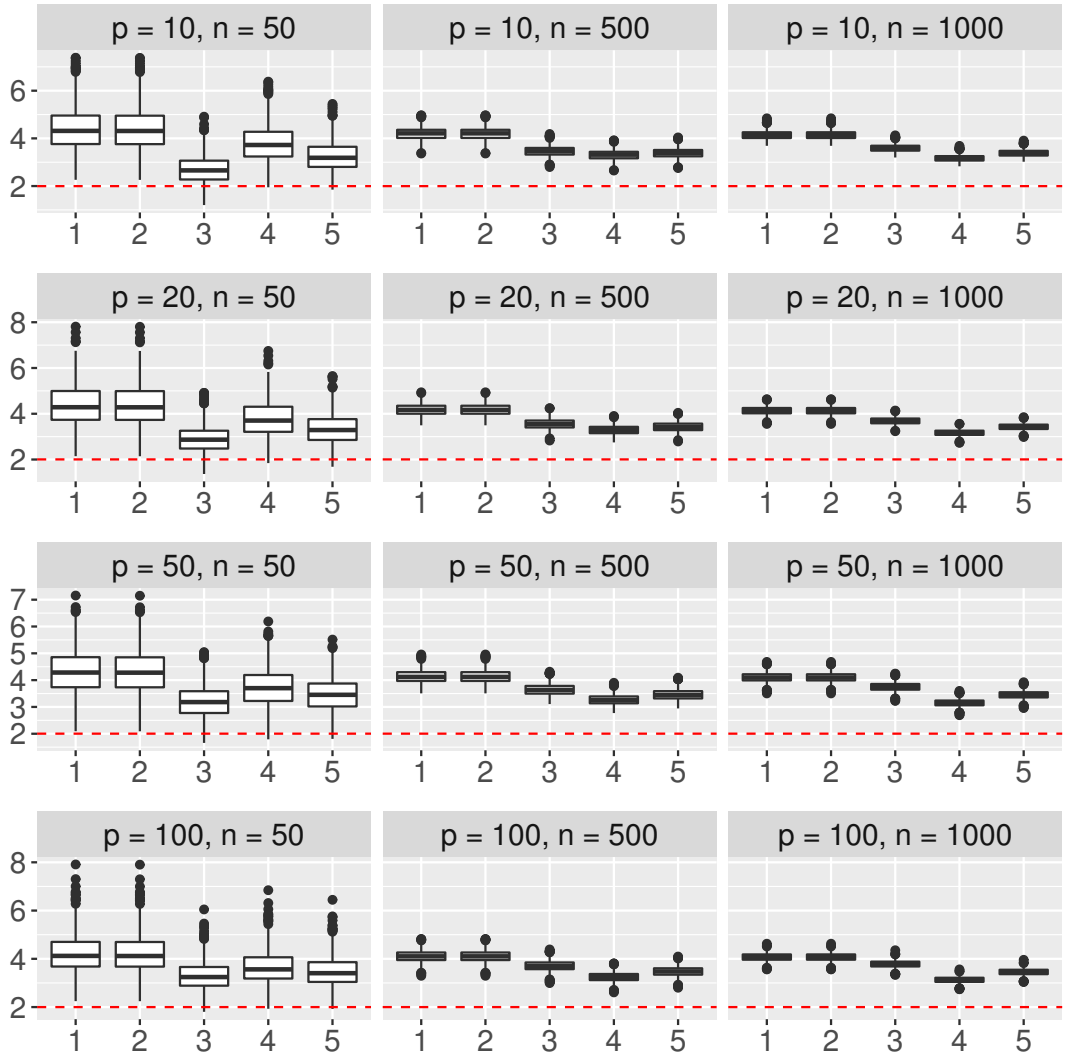


Figure 2.7: Simulation results for the **trigonometric model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN} = \mathbf{1}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$. The red dotted line inidcates the true residual variance.

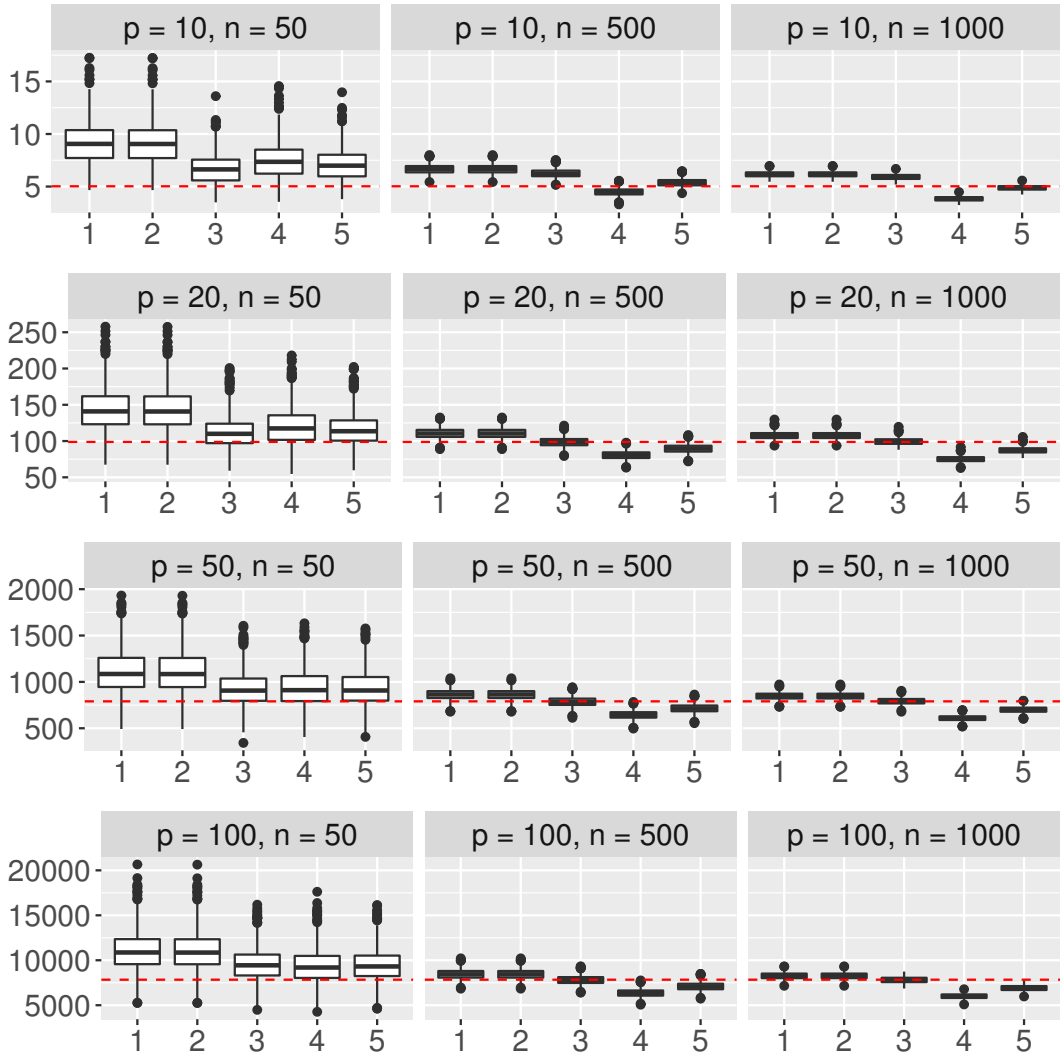


Figure 2.8: Simulation results for the **non-continuous model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN} = \mathbf{1}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$. The red dotted line inidcates the true residual variance.

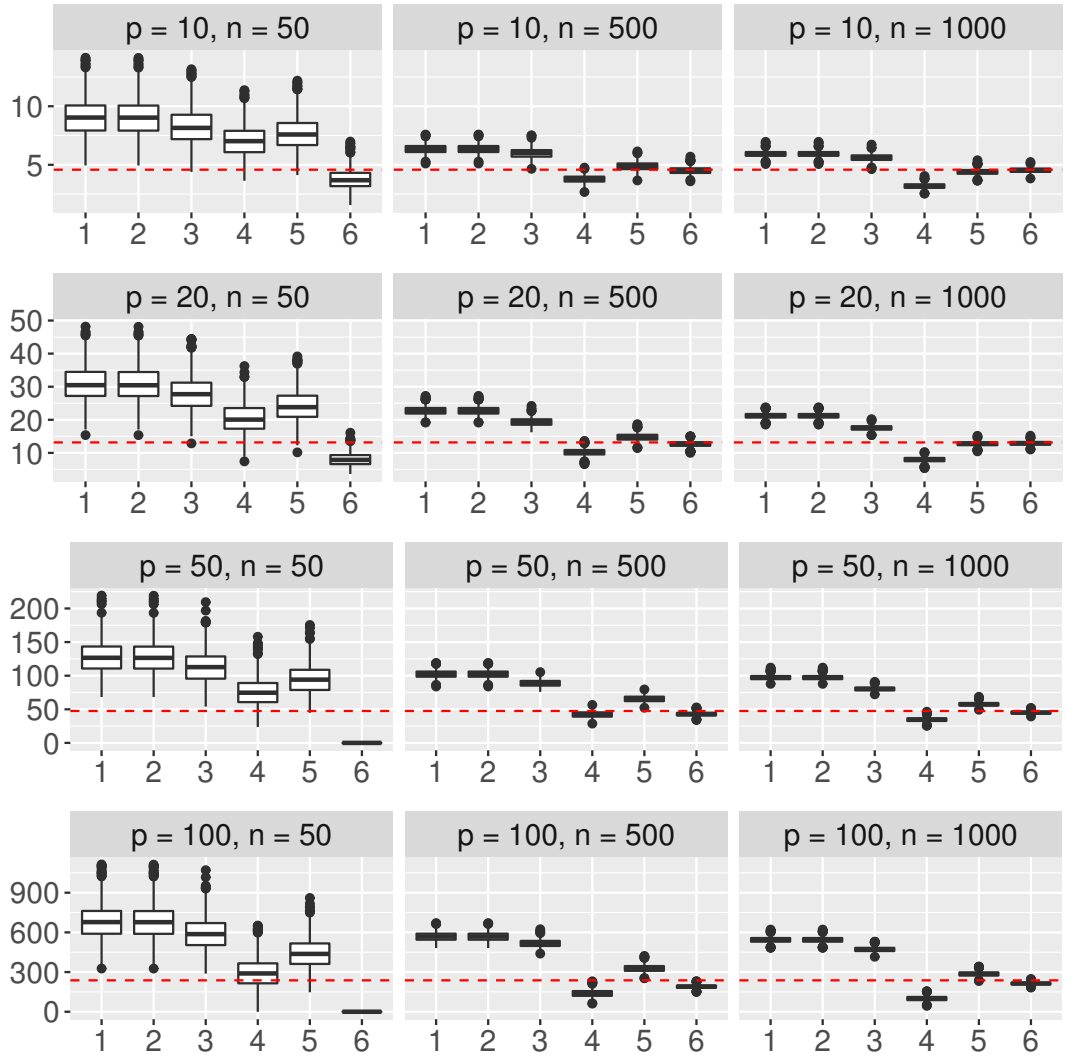


Figure 2.9: Simulation results for the **linear model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN} = \mathbf{2}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$ 6 : Benchmark. The red dotted line inidcates the true residual variance.

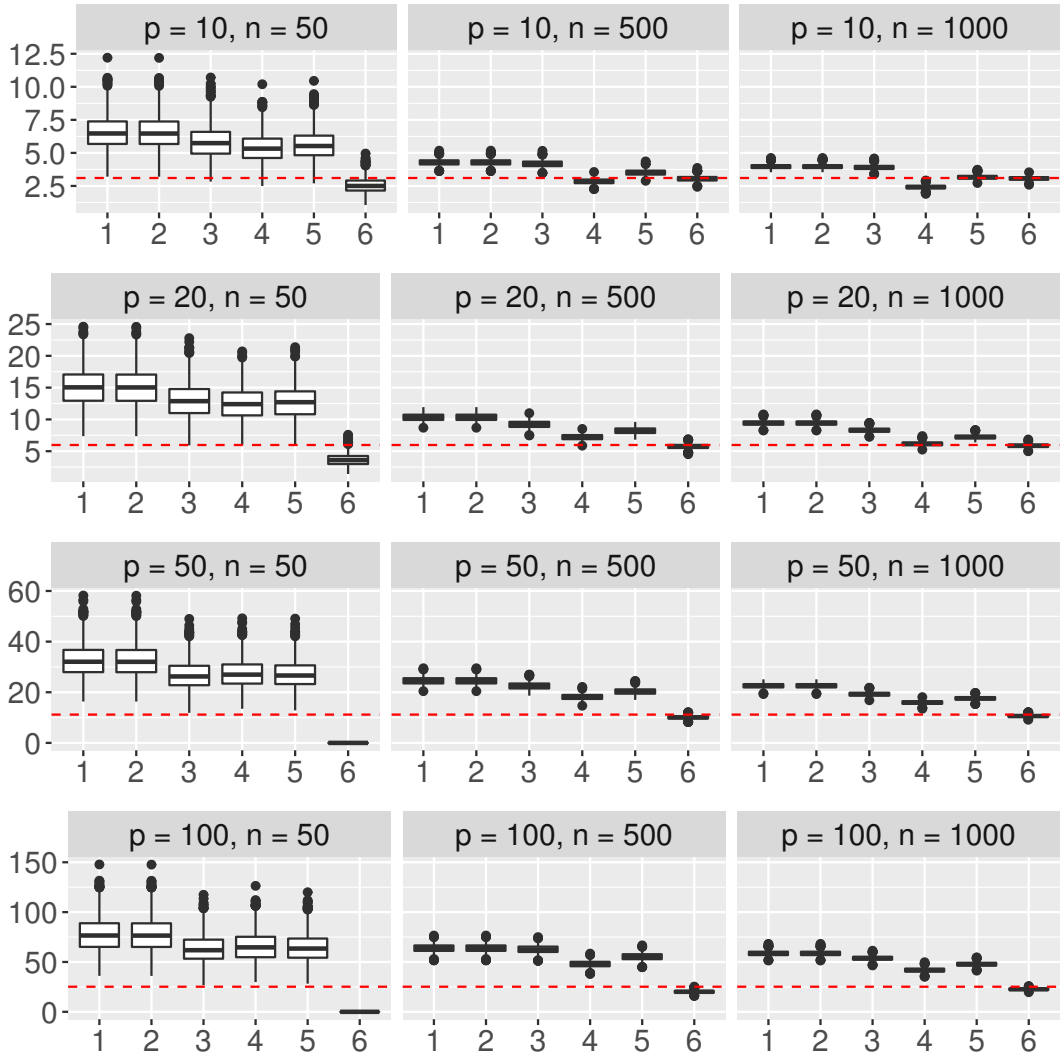


Figure 2.10: Simulation results for the **polynomial model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN} = \mathbf{2}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$ 6 : Benchmark. The red dotted line inidcates the true residual variance.

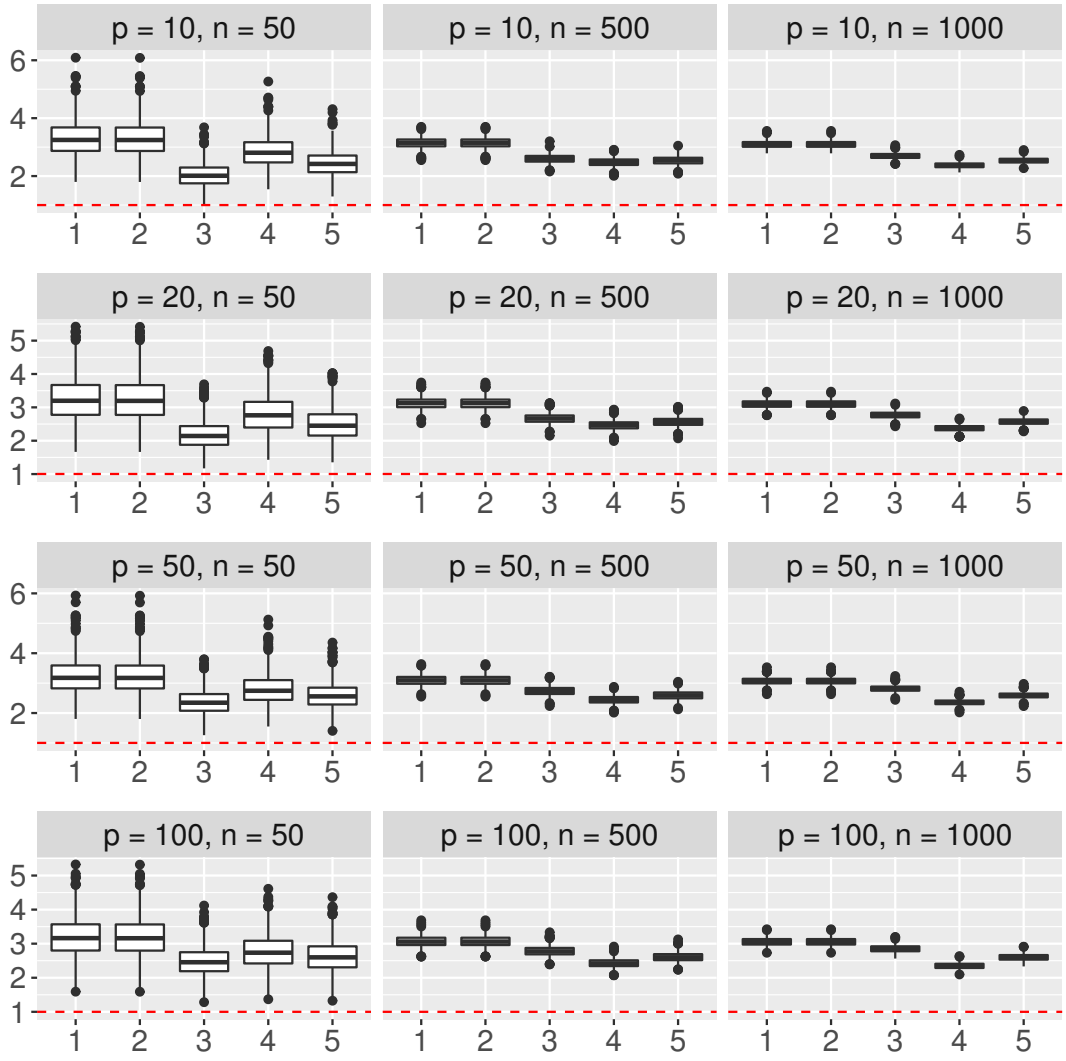


Figure 2.11: Simulation results for the **trigonometric model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN = 2}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$. The red dotted line inidcates the true residual variance.

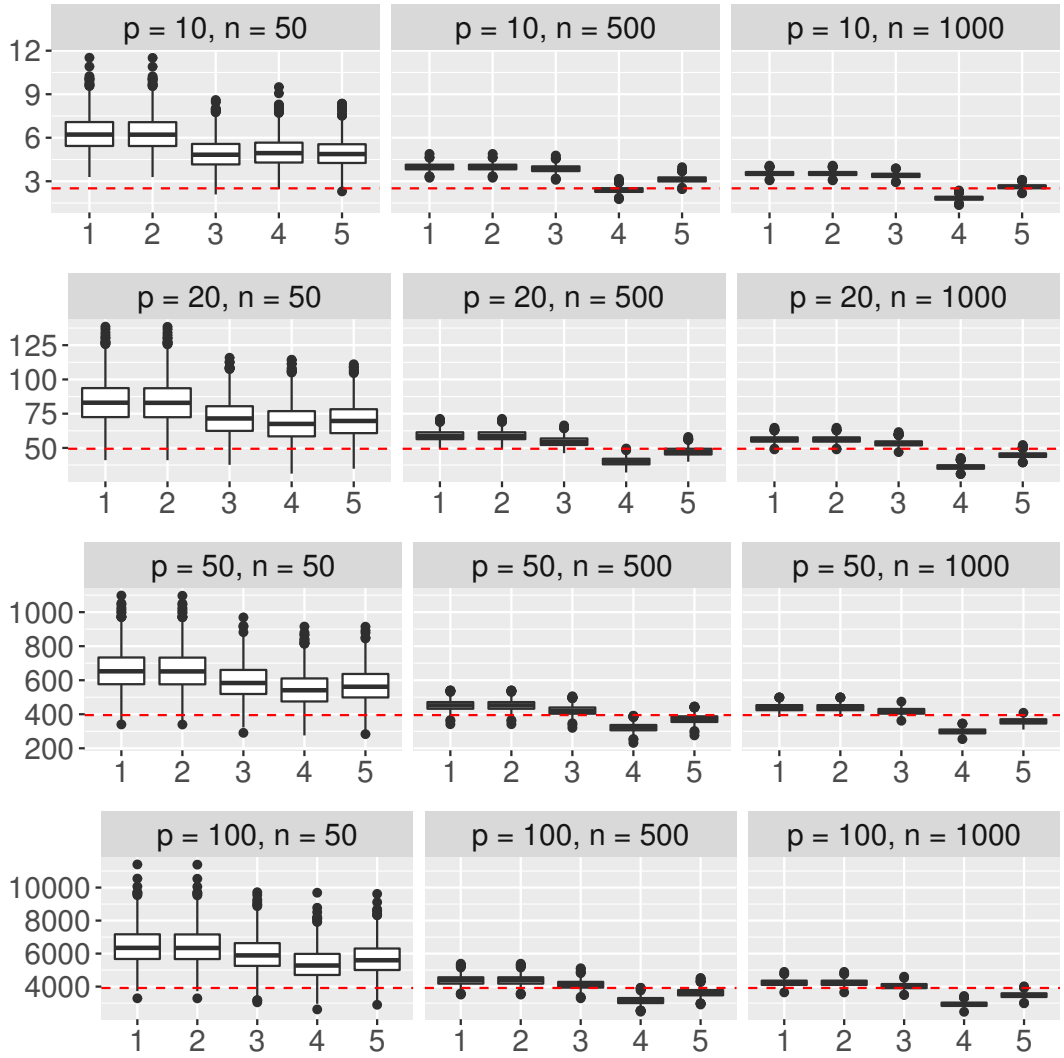


Figure 2.12: Simulation results for the **non-continuous model** with various sample sizes and feature dimensions under a signal-to-noise ratio of $\mathbf{SN} = \mathbf{2}$ using $1,000$ Monte-Carlo iterations. The residual variance estimators are encoded as $1 : \widehat{\sigma}^2_{RF}$ $2 : \widehat{\sigma}^2_{RFfast}$ $3 : \widehat{\sigma}^2_{RFiter}$ $4 : \widehat{\sigma}^2_{Mcorrect,1000}$ $5 : \widehat{\sigma}^2_{RFmiddle,1000}$. The red dotted line inidcates the true residual variance.

Chapter 3

Missing Values and Multiple Imputation

The issue of partially observed data can have an irritating effect on the statistical analysis one aims to conduct. An intuitive idea of overcoming this issue is to simply delete those observations containing missing values (complete case analysis). However, this comes with the cost of loosing important (partial) information, especially when the sample size is small and missing rates are relatively high (White and Carlin, 2010). Another approach is to adopt statistical methods being capable of treating missing values. For example, if the aim is to deliver ML-estimators for regression coefficients, then the corresponding likelihood function should be adopted in order to cover missing values (Enders, 2001; Stubbendick and Ibrahim, 2006; Zhang and Rockette, 2005). We refer to the latter class of methods as *data adjusted methods*. During the adoption of statistical analysis models, it is important to model the mechanism that is responsible for the generation of missing values, called the *missing mechanism*. According to Rubin (1976), the missing mechanism has been modeled from a probabilistic, Bayesian viewpoint. That is, assuming that the data set $\mathcal{D}_n$ consists of iid observations $\mathbf{Z}_i = [Z_{i1}, \dots Z_{ip+1}]^\top = [Y_i, \mathbf{X}_i^\top]^\top \in \mathbb{R}^{p+1}$, we denote with $\mathbf{Y}$ the corresponding data matrix containing the observations in $\mathcal{D}_n = \{\mathbf{Z}_i\}_{i=1}^n$ in a row-wise fashion. That is, $\mathbf{Y} = [\mathbf{Z}_1, \dots, \mathbf{Z}_n]^\top \equiv [\mathbf{K}_1, \dots, \mathbf{K}_{p+1}]^\top \in \mathbb{R}^{n \times p+1}$ is the corresponding data matrix. Note that $\mathbf{K}_1 = [Y_1, \dots, Y_n]^\top$ and $\mathbf{K}_j = [X_{j1}, \dots, X_{jn}]^\top$, $2 \leq j \leq p$, represents the column-wise notation of the data matrix $\mathbf{Y}$. When missing values do occur, the data matrix $\mathbf{Y}$ can usually be separated into $\mathbf{Y} = [\mathbf{Y}_{obs}, \mathbf{Y}_{mis}]$, whereas one does not have access to $\mathbf{Y}_{mis}$, since this part is missing, but one has full access to the observed part $\mathbf{Y}_{obs}$. In order to build a bridge towards the column- and row-wise notation of the data matrix $\mathbf{Y}$, $\mathbf{Y}_{mis}$ has to be understood as all those entries in $\mathbf{Y}$, that are missing such that $\mathbf{Y}_{mis}$ may contain entries of $\mathbf{Z}_i$ resp. $\mathbf{K}_j$, $1 \leq i \leq n$, $1 \leq j \leq p+1$. The process of generating missing values is then modeled by the matrix $\mathbf{R} = [\mathbf{R}_1, \dots, \mathbf{R}_n]^\top \in \mathbb{R}^{n \times p+1}$, where each entry $\mathbf{R}_i = [R_{i,1}, \dots, R_{i,p+1}]^\top$ is a random vector indicating whether observation $i \in \{1, \dots, n\}$ in variable $j \in \{1, \dots, p+1\}$ is observed ($R_{ij} = 1$) or not observed, i.e. missing ($R_{ij} = 0$). Regarding the probability measure $\mathbb{P}$, it requires to emphasize that in fact $\mathbb{P} \in \{\mathbb{P}_\theta : \theta \in \Xi\}$, i.e. the probability distribution $\mathbb{P}$ can be of parametric and non-parametric nature. Supposing that the generation of the missing data matrix $\mathbf{R}$ is conducted from a probability distribution with unknown parameter ξ on the same probability space as $\mathbf{Z}_1$, then the missing mechanism can be separated in the following three cases according to Rubin (1976):

1. The missing mechanism is called *missing completely at random* (MCAR), if $\mathbb{P}[\mathbf{R}|\mathbf{Y}, \xi] = \mathbb{P}[\mathbf{R}|\xi]$ for all ξ and $\mathbf{Y}_{mis}$.

2. The missing mechanism is called *missing at random* (MAR), if $\mathbb{P}[\mathbf{R}|\mathbf{Y}, \xi] = \mathbb{P}[\mathbf{R}|\mathbf{Y}_{obs}, \xi]$ for all ξ and $\mathbf{Y}_{mis}$.

3. The missing mechanism is called *missing not at random* (MNAR), if $\mathbb{P}[\mathbf{R}|\mathbf{Y}, \xi] = \mathbb{P}[\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \xi] \neq \mathbb{P}[\mathbf{R}|\mathbf{Y}_{obs}, \xi]$ for all ξ and $\mathbf{Y}_{mis}$.

Deriving data adjusted methods requires therefore assumptions under which mechanism missing values do occur. From a practical perspective, all three types can have consequences on the later statistical analysis. Suppose for example, that during the electronic collection process of survey data, the machine has lost some of the observations due to a malfunctioning wireless connection. This can be considered as an MCAR example and has less effect on the model complexity during the theoretical development of data adjusted methods. Suppose that the study participants only deliver information regarding their income, if they are male and additionally, the gender is known for all study participants. Then, the missing mechanism in this case is said to be MAR. More delicate is the issue when missing data is generated under the MNAR mechanism, that is, if survey participant reply on an income-related study question only, if their income is either not too low or not too high, for example. In biomedical research, for example, MNAR mechanisms are often present, especially when the patient's information is not available, due to the information itself, i.e. one aims to measure the cholesterol level, but the patient passes away due to a heart attack resulting from a relatively high cholesterol level. Note that the MCAR and MAR mechanism can be sorted in an hierarchical order, that is, the MCAR mechanism implies the MAR mechanism. However, the MNAR mechanism never implies the MAR or MCAR mechanism and vice-versa. Another, yet important part beside the missing mechanism is the *missing pattern*. The latter distinguishes from the missing mechanism by describing the location and structural distribution of missing values in the data matrix $\mathbf{Y}$ rather than the description of the occurrence of missing values as a probabilistic model. According to Enders (2010), missing patterns can be categorized into six different types:

1. The missing pattern is called *monotone*, if whenever $R_{ij} = 0$, then $R_{ik} = 0$ for all $k > j$ with $1 \leq i \leq n$ and $1 \leq j, k \leq p + 1$. That is, after a suitable, column-wise rearrangement of the variables, once a missing value occurs for a unit in a certain variable, then the upcoming features are missing, too.

2. The missing pattern is called *univariate*, if there exists only a single variable $\mathbf{K}_j \in \mathbb{R}^n$ and $1 < k_0 < n$, such that $R_{ij} = 0$ for all $i > k_0$ and $R_{ij} = 1$ for all $i \leq k_0$.

3. The missing pattern is *unit nonresponsively*, if for a sequence of variables, say $\mathbf{K}_{s_1}, \ldots, \mathbf{K}_{s_k}$ with $1 < s_1, \ldots, s_k < p$, it holds $R_{is_\ell} = 0$ for all $i > i_0$ and $1 \leq \ell \leq k$ with $1 < i_0 < n$. That is, for a certain unit group, the same variables are missing.

4. The missing pattern is called *latent*, if there exists a variable $1 \leq j_0 \leq p$, such that $R_{ij_0} = 0$ for all $1 \leq i \leq n$.

5. The missing pattern is called *planned*, if missing values occur due to a planned mechanism.

6. The missing pattern is called *general*, if missing values are scattered throughout the data such that the pattern does not fall into one of the groups mentioned before.

Beside row-wise deletion and data adjustment methods, another, yet intuitive idea, is to impute missing values and conduct later statistical analysis as if missing values have not occurred so far. Regarding the development of suitable imputation methods, the missing

pattern usually plays a more important role than the missing mechanism. The latter is rather important for the later statistical validity of the obtained results. For example, having a missing data matrix $\mathbf{Y}$ that contains a monotone missing pattern, this can result into the sequential application of regression and classification models trained on the fully observed part of $\mathbf{Y}$, while missing parts in $\mathbf{Y}$ are then predicted using the trained model. However, simply imputing missing values (once) and conducting the statistical analysis on the imputed data set can have severe consequences in terms of the validity of later statistical methods. This, because imputation assigns values to missing cases, that will be treated as known and fixed for later statistical analysis without involving potential uncertainty originating from the fact that missing values are present. That is, once missing values are singly imputed, it is assumed that the data analyst afterwards treats the imputed value as certainly known, although the imputation value itself is a *guess* with uncertainty, too. The lack of uncertainty quantification in the later statistical analysis will therefore be not suitable in terms of a valid statistical inference procedure, see e.g. Rubin (2004), pages $11-15$. Therefore, Donald Rubin came across the idea to multiply impute data sets and conduct the later statistical analysis on each obtained and imputed data set. In order to introduce the procedure of multiple imputation, suppose that one is interested in a scientific quantity, say $\boldsymbol{Q} \in \mathbb{R}^k$, $k \in \mathbb{N}$, which would be given, if the entire population would be present. As an example, think of the multivariate normal distribution, i.e. $\mathbb{P} = N_s(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^s \times \mathbb{R}^{s \times s} = \Theta$. In the latter case, $\boldsymbol{Q}$ might either refer to $\boldsymbol{\mu}$ or to $\boldsymbol{\Sigma}$, depending on the viewpoint of the data analyst. In case of $\boldsymbol{Q} = \boldsymbol{\Sigma}$, the latter has then to be understood in vectorized notation instead of a matrix, i.e. $\mathbf{Q} \in \mathbb{R}^{(s^2-s)/2}$ such that $k = (s^2 - s)/2$. The quantity $\boldsymbol{Q}$ is written in vector notation, in order to emphasize the possibility of capturing multivariate quantities during multiple imputation procedures. However, $\boldsymbol{Q}$ can also represent scalar quantities. It is worth to notice that for multivariate $\boldsymbol{Q}$, $Var(\boldsymbol{Q})$ has to be understood as the covariance matrix of $\boldsymbol{Q}$ within the Bayesian viewpoint. In order to motivate Rubin's idea of multiple imputation, we introduce a *stage* model to clarify the theoretical thoughts involved in multiple imputation. The *stage-model* has rather to be understood as a heuristic or philosophic motivation rather than a technical one.

1. *Stage 1* is theoretically the ideal situation, where everything is known and uncertainty is completely deleted. This is the case, if one would have access to the whole *population* leading to certain knowledge about $\boldsymbol{Q}$. At this stage, statistics is not required at all, since $\boldsymbol{Q}$ is always known. In practice, however, this is almost never the case, since sample sizes are *finite* and do not reveal knowledge on the whole population, but rather a part of it. Hence, missing value issues can be considered as a *common* problem for the next stage.

2. At *Stage 2*, one just has access to a *finite* sample size of the whole population, such that at this stage, the quantity $\boldsymbol{Q}$ is not known and is *modeled* by $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}]$ resp. $\mathbb{P}[\boldsymbol{Q}]$ (Bayesian perspective) or estimated by $\widehat{\boldsymbol{Q}}_n$ (frequentist's perspective). Uncertainty regarding the estimation of $\boldsymbol{Q}$ via $\widehat{\boldsymbol{Q}}_n$ from a frequentist's perspective is then expressed in terms of a variance, say $\boldsymbol{U} = Var(\widehat{\boldsymbol{Q}}_n)$. Uncertainty has been implicitly accounted in $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}]$ resp. $\mathbb{P}[\boldsymbol{Q}]$ under the Bayesian perspective. Missing values considered in this thesis are not of *stage-2-type*. This stage is commonly present in statistics, where *missingness* is defined here mainly through the lack of knowledge for $\boldsymbol{Q}$ because of, e.g., a finite sample size. The randomness generated at this stage arises from the *sampling mechanism*.

3. *Stage 3* describes the case where missing values are present in the sense that even the finite sample as prescribed in Stage 2 is lacking partial information. Therefore,

the distribution $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}]$ resp. $\mathbb{P}[\boldsymbol{Q}]$ (Bayesian viewpoint) or $\widehat{\boldsymbol{Q}}_n$ and $\boldsymbol{U}$ are unknown and require to be either *estimated* (frequentist's perspective) or *modeled* (Bayesian perspective). At this stage, beside the sampling mechanism as a source of randomness, the missing mechanism plays a crucial role introducing additional randomness into the observations. Note that the parameter to be estimated increases at this stage to the estimands $\widehat{\boldsymbol{Q}}_n$ and $\boldsymbol{U}$.

4. *Stage 4* is *imputation specific*, and occurs only, if the imputer chooses a finite number of imputations $m \in \mathbb{N}$. The choice of a finite number of imputation results into *missing values* of *Stage-2-type* within this stage.

Introducing the idea of multiple imputation from a mathematical perspective, we follow the same motivation as in Rubin (2004) and Van Buuren (2018) and focus first on the imputation procedure itself, i.e. making Stage 3 complete in the sense of Stage 2. Doing this is mainly motivated by the Bayesian perspective assuming a Bayesian Model regarding $\boldsymbol{Q}$. Hence, extracting knowledge for the estimand $\boldsymbol{Q}$ requires knowledge from $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}]$, the posterior of the hypothetical complete data, which is not directly accessible. Instead, one can have at most knowledge from $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}]$, which itself can be decomposed in the following way

$$\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}] = \int \mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}] \cdot \mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}] d\mathbf{Y}_{mis}. \tag{3.1}$$

Equation (3.1) enables the explanation of multiple imputation procedures from a mathematical perspective. Suppose that one can draw imputed values, say $\mathbf{Y}^*_{mis}$ from the predictive posterior distribution $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]$. Then, one can use these values to compute the quantity of interest $\boldsymbol{Q}$ from the imputed data set $\mathbf{Y}_{imp} = [\mathbf{Y}_{obs}, \mathbf{Y}^*_{mis}]$ and consider the latter as a draw from the posterior $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}]$. In practice, obtaining knowledge for $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}]$ is rather hard. Instead, one could restrict the attention to moments of the posterior distribution. Recalling that the posterior mean of $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}]$ can be rewritten into

$$\mathbb{E}[\boldsymbol{Q}|\mathbf{Y}_{obs}] = \mathbb{E}[\mathbb{E}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}]|\mathbf{Y}_{obs}], \tag{3.2}$$

this will give final practical advice in how to conduct multiple imputation: Generate $m \in \mathbb{N}$ imputed values $\{\mathbf{Y}^*_{mis,t}\}_{t=1}^m$ in order to obtain m data sets $\mathcal{D}_n^{(1)}, \ldots, \mathcal{D}_n^{(m)}$, each consisting of $\mathbf{Y}_{imp,t} = [\mathbf{Y}_{obs}, \mathbf{Y}^*_{mis,t}]$. Based on equation (3.1), compute the quantity $\boldsymbol{Q}$ on each data set as if missing values never occurred and denote them with $\widehat{\boldsymbol{Q}}_{n,1}, \ldots, \widehat{\boldsymbol{Q}}_{n,m}$. The combining rule for obtaining a final estimate for $\boldsymbol{Q}$ is given by equation (3.2), which can be approximated by the strong law of large numbers, if the imputed values are independent draws from $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]$, i.e.

$$\bar{\boldsymbol{Q}}_{n,m} = \frac{1}{m} \sum_{t=1}^{m} \widehat{\boldsymbol{Q}}_{n,t}. \tag{3.3}$$

Quantifying uncertainty in terms of finding appropriate estimators for the posterior variance of $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}]$ is not straight forward. However, making use of the law of total variance (Weiss, 2006), one can decompose the posterior variance into

$$Var(\boldsymbol{Q}|\mathbf{Y}_{obs}) = \mathbb{E}[Var(\boldsymbol{Q}|\mathbf{Y}_{obs}, \boldsymbol{Y}_{mis})|\mathbf{Y}_{obs}] + Var(\mathbb{E}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}]|\mathbf{Y}_{obs}). \tag{3.4}$$

According to Rubin (2004) pages 35, 76 and $84 - 85$, the first component is the average of the variance estimator based on the complete data set $\mathbf{Y} = [\mathbf{Y}_{obs}, \mathbf{Y}_{mis}]$ and is referred to as the *within imputation variance*. The second component indicates the variance between the

posterior means of the completed data. This decomposition enables the consideration of the following total variance estimator for $Var(\boldsymbol{Q}|\mathbf{Y}_{obs})$ within the multiple imputation procedure:

$$\begin{aligned}\widehat{Var}(\boldsymbol{Q}|\mathbf{Y}_{obs}) = \boldsymbol{T}_{n,m} &:= \frac{1}{m}\sum_{t=1}^{m}\boldsymbol{U}_{n,t} + \left(1+\frac{1}{m}\right)\cdot\frac{1}{m-1}\sum_{t=1}^{m}(\widehat{\boldsymbol{Q}}_{n,t}-\bar{\boldsymbol{Q}}_{n,m})(\widehat{\boldsymbol{Q}}_{n,t}-\bar{\boldsymbol{Q}}_{n,m})^{\top} \\ &= \bar{\boldsymbol{U}}_{n,m} + \left(1+\frac{1}{m}\right)\cdot\boldsymbol{B}_{n,m}, \end{aligned} \tag{3.5}$$

where $\boldsymbol{U}_{n,t}$ is an estimate of the variance of $\widehat{\boldsymbol{Q}}_{n,t}$ and the factor $(1+1/m)$ adjusts for the usage of a finite number of multiple imputations $m \in \mathbb{N}$ (Stage 4 in the stage-model), instead of an infinite one (Stage 3). Equation (3.3) and (3.5) are the combination rules in the multiple imputation procedure and are referred to as *Rubin's rule*. These rules together with the description of conducting multiple imputations as draws from the predictive posterior distribution enables the implementation of statistical inference procedures under partially observed data. For example, suppose that in case of no missing values, i.e. $\boldsymbol{Y} = \boldsymbol{Y}_{obs}$, one can obtain an estimator $\widehat{\boldsymbol{Q}}_n$, such that $(\boldsymbol{Q}-\widehat{\boldsymbol{Q}}_n) \sim N_k(\mathbf{0}, \boldsymbol{U})$ holds. When aiming to test for, say $H_0 : \boldsymbol{Q} = \boldsymbol{Q}_0$, as shown in Rubin (2004), the test statistic $D_{n,m} = (\bar{\boldsymbol{Q}}_{n,m}-\boldsymbol{Q}_0)^{\top}\boldsymbol{T}_{n,m}^{-1}(\bar{\boldsymbol{Q}}_{n,m}-\boldsymbol{Q}_0)/k$ follows under H_0 using a *distribution-preserving* multiple imputation procedure an F-distribution with (k,ν) degrees of freedom, where $\nu = (m-1)(1+1/r_m)^2$ and $r_m = (1+1/m)\cdot tr(\boldsymbol{B}_{n,m}\cdot\bar{\boldsymbol{U}}_{n,m}^{-1})/k$ is the *relative increase in variance due to non-response*, see e.g. Rubin (2004) on pages $77-78$. In this context, we refer to a *distribution-preserving* imputation scheme as a method, which generates multiple imputations independently under the same Bayesian model as the data analyst does during the analysis phase. In order to understand Rubin's terminology of *relative increase in variance due to non-response*, think of a scalar $\boldsymbol{Q} = Q$. Then, the total variance would consist only of the first part in the decomposition given in (3.4) and therefore, only of $\bar{\boldsymbol{U}}_{n,m} = \bar{U}_{n,m}$. Thus, the between imputation correction $(1+1/m)\boldsymbol{B}_{n,m} = (1+1/m)B_{n,m}$ accounts for the additional uncertainty originating from the presence of missing entries. Therefore, the fraction $r = (1+1/m)\cdot tr(\boldsymbol{B}_{n,m}\cdot\bar{\boldsymbol{U}}_{n,m}^{-1})/k = \frac{(1+1/m)B_{n,m}}{\bar{U}_{n,m}}$ for the scalar case, i.e. $k=1$, refers to the ratio of both quantities capturing the additional variance increase due to the presence of missing values. Hence, the term *relative increase in variance due to non-response*.

3.1 Validity of Multiple Imputation Procedures

In the previous section, we have introduced different methods when dealing with missing values during the analysis of incomplete data, while a special focus has been set on multiple imputation procedures based on Rubin (1996), Rubin (2004) and Van Buuren (2018). As mentioned in the previous section, multiple imputation procedures have mainly been motivated by equation (3.1), where the complete-case analysis is conducted based on imputed draws from the predictive posterior distribution $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]$. In practice, however, draws from the predictive posterior distribution are rarely possible. Instead one assumes a model, the so called *imputation model*, where imputation draws are generated from the latter. An important research question arising from the idea of multiple imputation is then the following:

(R1) *When does multiple imputation deliver valid statistical inference procedures ?*

Regarding research question ($R1$), general answers have been delivered in Rubin (2004) and Van Buuren (2018). However, there is some ambiguity involved in the question stated, that seems not be clear at first hand. When referring to statistical inference, one usually

thinks of different quantities involved during traditional hypothesis testing. The latter is usually conducted from the frequentist's perspective, where the quantity of interest, here $\boldsymbol{Q}$, is fixed and statistical hypothesis testing is mainly based on the unknown, complete-case tuple $(\widehat{\boldsymbol{Q}}_n, \boldsymbol{U})$, where $\widehat{\boldsymbol{Q}}_n$ is an estimate of $\boldsymbol{Q}$, if no missing values were present and $\boldsymbol{U}$ its corresponding variance-(covariance) matrix. If statistical inference is conducted from a Bayesian perspective, than the interesting quantity is the unknown complete-case posterior $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}]$. The different viewpoints have already been introduced within the stage-model (see Stage 1 - Stage 4). Note that in practice, when missing values do occur, the posterior distribution of $\boldsymbol{Q}$ is actually given by $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{R}]$, which is slightly different to the result obtained in equation (3.1). The latter does not involve the missing mechanism $\mathbf{R}$ being responsible for the generation of the missing data and therefore ignores the missing mechanism as one would be at Stage 2, although Stage 3 resp. 4 should be addressed. In order to clarify circumstances, under which one can indeed consider $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}]$ instead of $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{R}]$ within Bayesian inferential statistics, Rubin (2004) has introduced an *ignorable missing mechanism*, which we will state as a definition.

Definition 3.1 (Ignorability)**.** The missing mechanism is said to be ignorable, if at least MAR holds and the joint prior of θ and ξ decomposes into the product of marginal priors, i.e. $\mathbb{P}[\theta, \xi] = \mathbb{P}[\theta] \cdot \mathbb{P}[\xi]$.

Returning to equation (3.1), it has been mentioned that imputations should be draws from the predictive posterior distribution $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]$. Similarly to the problem of having $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{R}]$ instead of $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}]$, the predictive posterior distribution under partially observed data does also depend on the missing mechanism $\mathbf{R}$. Hence, one actually has $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}]$. Rubin (2004) and Van Buuren (2018) have mentioned that under the MAR mechanism, the predictive posterior distribution does not depend on the missing mechanism, without delivering a formal proof. Therefore, we prove the result and state this as a proposition for completion.

Proposition 3.1. *Assume at least the MAR condition. Then the missing mechanism in the predictive posterior distribution can be ignored, that is* $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}] = \mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]$.

Note that the quantity $\boldsymbol{Q}$ can be considered as a potential function of the data generating parameter θ, i.e. $\boldsymbol{Q} = f(\theta)$ for a measurable function f. Therefore, drawing Bayesian inference for the quantity $\boldsymbol{Q}$ involves drawing Bayesian inference for the data generating parameter θ based on $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{R}]$. There are several factors that can distort the validity of multiple imputation during Bayesian and frequentist's inference procedures. Note that statistical inference in Bayesian analysis is conducted based on $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{R}]$, wheres in frequentist's analysis, statistical inference is mostly based on the tuple $(\widehat{\boldsymbol{Q}}_n, \boldsymbol{U})$ and treats the estimand $\boldsymbol{Q}$ as fixed. Following the general considerations as in Rubin (2004), one can list the following factors:

1. The missing mechanism, if not correctly specified, can distort statistical inference. Since modeling the latter can be challenging and is in general not possible, a preferable situation is the condition of ignorability according to Definition 3.1.

2. Since imputations are constructed based on some model, one has to guarantee that the imputed values are independent draws from the predictive posterior distribution under the posited response mechanism. Therefore, they have to be draws from $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}]$. Hence, a possible source of distortion is a stationary distribution of imputed values, that is different to $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}]$. The same holds for non-stationary distributions, i.e. if (multiply) imputed draws do not share the same distribution yielding to a mixture of distributions after aggregation according to (3.3).

3. In case of incorrect complete-case statistical inference, multiple imputation cannot overcome the distortion by making multiple imputation inference correct again.

4. Multiple imputation intends to conduct complete-case analysis on artificially completed data sets through imputation. Therefore, two models are actually involved during statistical inference procedures: The imputation model, that is responsible for drawing imputations from the predictive posterior distribution $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}]$ under the posited response mechanism, and the analysis model, that accounts under Bayesian inference, for the posterior $\mathbb{P}[\mathbf{Q}|\mathbf{Y}_{obs}, \mathbf{R}]$ resp. the distribution of $\widehat{\boldsymbol{Q}}_n - \boldsymbol{Q}$ under the frequentist's perspective. If both models differ, potential distortion regarding statistical inference might be present.

Having summarized potential sources of distortion for valid statistical inference procedures, we first aim to clarifiy circumstances, under which Bayesian inference procedures based on the posterior $\mathbb{P}[\mathbf{Q}|\mathbf{Y}_{obs}, \mathbf{R}]$ are indeed valid. Although it has been mentioned by several authors such as Brand (1999), Rubin (2004) and Van Buuren (2018) that multiple imputation delivers valid Bayesian inference procedures, however, clear circumstances and a formal proof were not given. Therefore, we will state the latter as a theorem and proof it in the last section of the chapter.

Theorem 3.1 (Bayesian Validity of Multiple Imputations)**.** *Assume that the missing mechanism is ignorable and imputations are independent draws from the predictive posterior distribution* $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]$*. Suppose that the imputer's model and the analyst model coincide. Then the posterior distribution* $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{R}]$ *using the infinite-m multiple imputation procedure can be fully recovered such that Bayesian inference based on* $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{R}]$ *are valid.*

According to Rubin (1996), achieving Bayesian validity is *far more difficult* such that the application of Theorem 3.1 might be limited in practice. However, statistical valid inference is often conducted from the frequentist's perspective. Therein, the interesting quantity $\boldsymbol{Q}$ is treated as fixed and statistical inference is conducted by means of statistical hypothesis testing, i.e. having

$$H_0 : \boldsymbol{Q} = \boldsymbol{Q}_0 \quad vs. \quad H_1 : \boldsymbol{Q} \neq \boldsymbol{Q}_0, \tag{3.6}$$

for a fixed $\boldsymbol{Q}_0 \in \mathbb{R}^k$ in mind, for example. The aim in complete-case analysis is to find a test statistic τ_n that follows a certain distribution under the null-hypothesis either exactly or asymptotically, as $n \to \infty$ such that their quantiles are known. Suppose that $\widehat{\boldsymbol{Q}}_n$ is a complete-case estimator for $\boldsymbol{Q}$ such that, similarly to Rubin (2004), the normality assumption holds, i.e.

$$(\widehat{\boldsymbol{Q}}_n - \boldsymbol{Q}) \sim N_k(\mathbf{0}, \boldsymbol{U}). \tag{3.7}$$

Assumption (3.7) is placed, in order to simplify statistical inference procedures by focusing on the tuple $(\widehat{\boldsymbol{Q}}_n, \boldsymbol{U})$ while making the theoretical verifications in Rubin (2004) regarding the validity of multiple imputation from a frequentist's perspective applicable. Deviations towards the normality assumption as given in (3.7) are shortly discussed in Brand (1999), page 76, where (3.7) can also be seen as fulfilled, if the sample size n is sufficiently large to approximate the sampling distribution of $\widehat{\boldsymbol{Q}}_n$. Regarding the variance $\boldsymbol{U}$, it is assumed that under the absence of missing values (i.e. Stage 2), one would have access to a consistent estimator $\widehat{\boldsymbol{U}}_n$, i.e. for $n \to \infty$ we have

$$\widehat{\boldsymbol{U}}_n \xrightarrow{\mathbb{P}} \boldsymbol{U}. \tag{3.8}$$

Statistical validity of multiple imputation procedures from the frequentist's perspective, using an infinite number of imputations, then concerns with the question whether

$$(\bar{\boldsymbol{Q}}_{n,\infty} - \boldsymbol{Q}) \sim N_k(\mathbf{0}, \boldsymbol{T}_{n,\infty}) \tag{3.9}$$

holds, since the data analysts after imputing missing values by the imputer conducts statistical inference based on (3.9). Potential distortions to (3.9) under (3.7) will yield to invalid statistical inference procedures caused by the usage of the multiple imputation logic. Assuming that the imputations are independent repetitions from a Bayesian model $\mathbb{P}[\boldsymbol{Y}^*_{mis}|\mathbf{Y}_{obs}, \mathbf{R}]$ under which posterior means and variances exists, one has by the strong law of large numbers

$$\begin{aligned}
\bar{\boldsymbol{Q}}_{n,\infty} &:= \lim_{m\to\infty} \boldsymbol{Q}_{n,m} = \mathbb{E}[\boldsymbol{Q}_{n,1}(\mathbf{Y}_{obs}, \mathbf{Y}^*_{mis}, \mathbf{R})|\mathbf{Y}_{obs}, \mathbf{R}], \\
\bar{\boldsymbol{U}}_{n,\infty} &:= \lim_{m\to\infty} \bar{\boldsymbol{U}}_{n,m} = \mathbb{E}[\boldsymbol{U}_{n,1}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}^*_{mis}, \boldsymbol{R})|\mathbf{Y}_{obs}, \mathbf{R}], \\
\bar{\boldsymbol{B}}_{n,\infty} &:= \lim_{m\to\infty} \boldsymbol{B}_{n,m} = \mathbb{E}[\boldsymbol{B}_{n,1}(\mathbf{Y}_{obs}, \mathbf{Y}^*_{mis}, \mathbf{R})|\mathbf{Y}_{obs}, \mathbf{R}].
\end{aligned}$$

Note that the normal distribution in (3.9) has to be understood under both, the response mechanism and the sampling mechanism. Establishing theoretical guarantees for the validity of (3.9) has been done in Rubin (2004) in Chapter 4, page 119 (therein, see Result 4.1). Beside the assumptions that inference procedures are valid under the absence of missing values (i.e. condition (3.7)), a crucial requirement was *proper* imputations. The formal definition of a proper imputation is given in Rubin (2004) on pages 118-119. However, we will use the simplified version of the latter given in Rubin (1996) resp. Van Buuren (2018).

Definition 3.2. A multiple imputation procedure is called proper for a set of complete data statistics $(\widehat{\boldsymbol{Q}}_n, \widehat{\boldsymbol{U}}_n)$ and posited response mechanism, if the following three conditions hold:

1. The multiple imputation point estimate is unbiased for $\widehat{\boldsymbol{Q}}_n$. That is, $\mathbb{E}[\bar{\boldsymbol{Q}}_{n,\infty}|\boldsymbol{Y}] = \mathbb{E}[\widehat{\boldsymbol{Q}}_n(\mathbf{Y}, \mathbf{R})|\mathbf{Y}] = \widehat{\boldsymbol{Q}}_n$ holds $\mathbb{P}$-almost surely. $\mathbb{E}[\bar{\boldsymbol{Q}}_{n,\infty}|\mathbf{Y}]$ has to be understood as the expectation over the response mechanism, i.e. over the conditional distribution $\mathbb{P}[\mathbf{R}|\mathbf{Y}]$.

2. The within-variance estimator of the multiple imputation variance $Var(\bar{\boldsymbol{Q}}_{n,\infty})$ is unbiased for $\widehat{\boldsymbol{U}}_n$. That is, $\mathbb{E}[\bar{\boldsymbol{U}}_{n,\infty}|\mathbf{Y}] = \widehat{\boldsymbol{U}}_n$ holds $\mathbb{P}$ - almost surely.

3. The between imputation variance of the infinite-m multiple imputation procedure is for $Var(\bar{\boldsymbol{Q}}_{n,\infty}|\mathbf{Y})$ unbiased. That is, $\mathbb{E}[\boldsymbol{B}_{n,\infty}|\mathbf{Y}] = Var(\bar{\boldsymbol{Q}}_{n,\infty}|\mathbf{Y})$ holds $\mathbb{P}$-almost surely.

Based on Definition 3.2 together with some other assumptions that aim to bring the analyst's frequentist perspective in virtual agreement with the imputer's Bayesian perspective, Rubin could show in Chapter 4 the validity of statistical inference procedures even under the frequentist's perspective using the multiple imputation logic (therein, Result 4.1). In order to verify the application of Rubin's theoretical results in this thesis, we emphasize that a crucial property for the validity of multiple imputation procedures under the frequentist's perspective requires among others proper imputations and the inferential validity under complete-data (validity at Stage 2). In the next section, we will show that this is not necessarily the case for some Random Forest models used as imputation procedures.

Returning to the process of generating missing values, it is practically difficult to guarantee that imputations are draws from $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]$. However, under a monotone missing pattern, imputation methods can be designed to actually draw from the predictive posterior distribution $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]$ by sequentially applying regression resp. classification tasks on each

of the variables, see e.g Brand (1999), page 48, Raghunathan et al. (2001) or Van Buuren (2018), pages 103 - 105. Regarding the latter missing pattern, we refer to the *Montone Data Imputation* strategy described in Van Buuren (2018) on pages 102 - 104. For general missing patterns, one usually categorizes imputation methods into two cases, see e.g. Van Buuren (2018):

(i) *Joint modeling* (JM), which assumes a pre-specified multivariate distribution for the data matrix $\mathbf{Y}$. That is, the data matrix $\mathbf{Y}$ is row-wise rearranged so that sub-groups of missing patterns are obtained. Then, for every missing pattern, a joint distribution is assumed in order to be able to draw from it. For example, assume that after sorting the rows of $\mathbf{Y}$, a specific group $\mathcal{G} \subseteq \{1, \ldots, n\}$ of missing pattern in $\mathbf{Y}$ has the form $R_i = [0, 0, 0, 1, \ldots, 1]^\top$ for all $i \in \mathcal{G}$. Then, a distribution model for $\mathbb{P}[Y_i, X_{i1}, X_{i2} | X_{i3}, \ldots, X_{ip}, \theta_{1,2,3}]$ using the observations prescribed by $\mathcal{G}$ is assumed, where $\theta_{1,2,3}$ is a parameter for the trivariate conditional distribution. A famous example of a joint modelling strategy is the *data augmentation algorithm* developed by Tanner and Wong (1987).

(ii) *Fully conditional specification* (FCS) is different to the joint modeling strategy by assuming for every variable in the data matrix $\mathbf{Y}$ a model such that imputation actually happens on a variable-by-variable basis. That is, for every variable $1 \leq j \leq p+1$, a distributional model for $\mathbb{P}[Z_{1j} | \mathbf{Z}_{-j}, \theta_j]$ is assumed, where $\mathbf{Z}_{-j} = [Z_{11}, \ldots, Z_{1j-1}, Z_{1j+1}, \ldots, Z_{1p+1}]^\top \in \mathbb{R}^p$ and θ_j is a model parameter. Missing values are then drawn from $\mathbb{P}[Z_{1j} | \mathbf{Z}_{-j}, \widehat{\theta}_j]$, where $\widehat{\theta}_j$ is estimated from the observed data. The process is then iterated until all variables have been treated. A famous example of an FCS approach is the multiple imputation using chained equations (MICE) algorithm implemented in R under the package `mice` (Van Buuren and Groothius-Oudshoorn, 2009).

Practical disadvantages for JM approaches are given, when the joint conditional distribution involves the treatment of mixed-type data. That is, if missing values in specific missing patterns capture both, continuous and non-continous (e.g. nominal) features. This makes the specification of a joint conditional distribution rather hard, where usual assumptions such as multivariate normality are clearly not met. FCS overcomes this issue by specifying a model for every variable such that missing values are imputed in an iterative fashion until all variables have been treated. Then, just the single scale of one feature determines the nature of the conditional distribution $\mathbb{P}[Z_{1j} | \mathbf{Z}_{-j}, \theta_j]$, which makes a suitable choice easier.

In practice, it is often the case that more information is actually available then initially needed for the underlying statistical analysis. Considering for example the breast cancer gene study used as an empirical data example in article ($P2$), various gene expressions beside the ones identified as potential markers for breast cancer are available. Simply ignoring them during the imputation can lead to potential issues:

1. In case that the data analyst conducts additional statistical analyses including variables not considered during imputation, the imputed values will not reflect potential dependencies towards these variables. Therefore, the (stationary) distribution from which imputations are draws can be misspecified.

2. Not including additional information during imputation can lead to *information loss*. The latter has to be understood as the barrier to conduct valid statistical analyses after imputation, because of the involvement of these variables during additional statistical analyses. Since imputation itself is derived through the Bayesian perspective, but the

analyst conducts statistical inference from a frequentist's perspective, the two models can differ, especially when additional variables are missing within the imputation model. In Meng (1994), this has been termed as *inferential uncongeniality.*

Additional variables considered in the imputation model but not directly in the analysis model are called *auxiliary variables.* Formally speaking, the data set $\mathbf{Y}$, based on which the analyst will conduct its statistical procedures, will be extended by auxiliary variables, say $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_s] \in \mathbb{R}^{n \times s}$ such that $\mathbf{W}_\ell \in \mathbb{R}^n$ for all $1 \leq \ell \leq s$ and $s \in \mathbb{N}$. Then, imputations are drawn from the extended imputation model $\mathbb{P}_I[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{W}]$. Conditions, under which the latter model agrees with the predictive posterior distribution $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]$ are given in Meng (1994). Following Meng's thoughts, the aim of including auxiliary variables is to achieve an imputation model, that is more general and covers potential (sub-) models used during the analysis phase. This is especially in line with Theorem 3.1, which requires the conformity of the imputation model and the analyst model. Recent developments in the Machine Learning community, especially the possibility of the Random Forest method to treat mixed-type data in various prediction and variable screening tasks simultaneously, have made its application as a general tool for missing value imputation popular (see e.g. Lunetta et al. 2004, Hudak et al. 2008, Stekhoven and Bühlmann 2012, Hapfelmeier et al. 2012, Shah et al. 2014, Doove et al. 2014 and Deng et al. 2016). The next section aims to cover the involvement of Random Forest models in missing value imputation.

3.2 Multiple Imputation and the Random Forest

So far, important results of multiple imputation procedures have been summarized and clarified. A special focus has been set on the validity of the multiple imputation procedure as a general tool for statistical inference. Since machine learning algorithms are currently used in different industrial and practical applications, it is interesting to know the benefits of them within partially observed data and the multiple imputation procedure. Recalling the advantages of the Random Forest method, this section will summarize the potential usage of the Random Forest method as an imputation algorithm within the multiple imputation framework. Since Random Forest models are mainly based on univariate responses, JM as a general imputation class can be ignored. Therefore, imputing missing values using the Random Forest method is usually conducted within the FCS framework. The Random Forest method itself has first been implemented as an imputation model within the statistical software R under the function `missForest` and is based on the work of Stekhoven and Bühlmann, see e.g. Stekhoven and Bühlmann (2012). Therein, the imputation task is considered as a prediction task, where a Random Forest is trained on observed values of the data matrix $\mathbf{Y}$ and missing values are predicted using the trained Random Forest model. In order to precise the strategy of missing value imputation using the Random Forest method, we will state the algorithm defined in Stekhoven and Bühlmann (2012). The latter has been used in a variety of missing value issues, due to its flexible usage within mixed-type data formats. Similarly to the notation used in our article $(P1)$, let $\pi : \{1, \ldots, p+1\} \to \{1, \ldots, p+1\}$ be a permutation and denote with $\mathbf{K}_j^{obs}$ for $j \in \{1, \ldots, p+1\}$ the sub-vector of $\mathbf{K}_j$, that do not contain any missing values, while $\mathbf{i}_j^{obs} = \{i \in \{1, \ldots, n\} : R_{ij} = 1\}$ is the index-set of observed values in $\mathbf{K}_j$ such that $\mathbf{K}_j^{obs} \in \mathbb{R}^{|\mathbf{i}_j^{obs}|}$. Analogously, $\mathbf{K}_j^{mis}$ and $\mathbf{i}_j^{mis}$ is defined. The rest of the matrix is given by $\mathbf{Y}_{-j}^{obs}$ resp. $\mathbf{Y}_{-j}^{mis}$. That is $\mathbf{Y}_{-j}^{obs}$ is the sub-matrix of $\mathbf{Y}$ that contains all observations given in $\mathbf{i}_j^{obs}$, while excluding variable $\mathbf{K}_j$. Similarly, $\mathbf{Y}_{-j}^{mis}$ is the sub-matrix of $\mathbf{Y}$, that contains all observations given in $\mathbf{i}_j^{mis}$, with the variable $\mathbf{K}_j$ excluded.
Regarding the performance of Algorithm 4 as an imputation method, the authors in Stekhoven

Algorithm 4: Missing Value Imputation according to `missForest`.

Input: Data matrix $\mathbf{Y}$ with missing values, `maxIter`.

Result: Imputed data matrix $\mathbf{Y}^{imp}$.

1 Sort the p columns of the data matrix $\mathbf{Y}$ based on the missing rate in ascending order resulting into a permutation $\pi(1), ..., \pi(p+1)$ of the $p+1$ outcome variables;

2 Initially impute mean resp. mode values for missing continuous resp. categorical realizations of random variables;

3 Starting with the first permuted column $X_{\pi(1)}$, separate the permuted data matrix $\mathbf{Y}_\pi$ into four parts using the previous notation:

$$\left[\begin{array}{c|c} \mathbf{K}^{obs}_{\pi(1)} & \mathbf{Y}^{obs}_{-\pi(1)} \\ \hline \mathbf{K}^{mis}_{\pi(1)} & \mathbf{Y}^{mis}_{-\pi(1)} \end{array}\right] \in \mathbb{R}^{(|\mathbf{i}^{obs}_{\pi(1)}|+|\mathbf{i}^{mis}_{\pi(1)}|)\times(1+(p))}.$$

Train a Random Forest method using the sub-matrix $\mathbf{Y}^{obs}_{-\pi(j)}$ as covariates and $\mathbf{K}^{obs}_{\pi(j)}$ as response variable;

4 Impute the missing values of $\mathbf{K}_{\pi(j)}$ using the trained Random Forest and predict them with $\mathbf{Y}^{mis}_{-\pi(1)}$ as covariate values. Then move to the next variable and repeat steps **3** and **4** until all variables have been treated;

5 As long as the L_2-error of the newly and previously imputed data set has not increased for the first time or the number of iterations is less than `maxIter`, return to step 3;

and Bühlmann (2012) have considered the normalized root means squared error (NRMSE) resp. the proportion of false classification (PFC) of the imputed missing values as potential measures for evaluation. Suppose that the set of features can be decomposed into sets of continuous and non-continuous variables, i.e. $\{1, \dots p\} = \mathcal{C}_c \quad \dot\cup \quad \mathcal{C}_{nc}$. Denoting with $N^{(c)}_{mis}$ the set of all missing instances for continuous outcomes, i.e. $N^{(c)}_{mis} = \{(i,j) \in \{1,\dots,n\} \times \mathcal{C}_c : R_{ij} = 0\}$ while $N^{(nc)}_{mis} = \{(i,j) \in \{1,\dots,n\} \times \mathcal{C}_{nc} : R_{ij} = 0\}$ is the set of all missing instances for non-continuous outcomes. Furthermore, denote with Y^{imp}_{ij} the imputed values and Y^{mis}_{ij} the missing instances for $(i,j) \in N^{(c)}_{mis} \cup N^{(nc)}_{mis}$. Then, the NRMSE is usally used as an evaluation criterion for continuous outcomes, while the PFC is the evaluation criterion for non-continuous outcomes, which are both defined by

$$NRMSE = \frac{\sqrt{\sum_{(i,j)\in N^{(c)}_{mis}} \left(Y^{imp}_{ij} - Y^{mis}_{ij}\right)^2}}{\sqrt{\sum_{(i,j)\in N^{(c)}_{mis}} (Y^{mis}_{ij} - \bar{Y}^{mis}_{\cdot\cdot})^2}}, \tag{3.10}$$

$$PFC = \frac{1}{|N^{(nc)}_{mis}|} \sum_{(i,j)\in N^{(nc)}_{mis}} \mathbb{1}\{Y^{imp}_{ij} \neq Y^{mis}_{ij}\}. \tag{3.11}$$

It has been reported in Stekhoven and Bühlmann (2012) that imputation performance using Algorithm 4 and the Random Forest method showed preferable results compared to the nearest neighbor method or `mice` in terms of low NRMSE resp. PFC, for both, homogeneous (i.e. $\mathcal{C}_c = \emptyset$ or $\mathcal{C}_{nc} = \emptyset$) and mixed-type data. Note that Algorithm 4 was initially invented

within the single-imputation framework. Extending it to the multiple imputation procedure can be conducted through the following two approaches:

1. The intuitive idea is to apply the imputation strategy using the `missForest` method repeatedly m-times, in order to achieve m complete data sets $\mathcal{D}_n^{(1)}, \ldots, \mathcal{D}_n^{(m)}$ and apply Rubin's multiple imputation procedure for aggregation. This is also motivated by the recommendation of Rubin to use independent draws for imputed values. This strategy will guarantee that given the data $\mathcal{D}_n$, imputed values should be independent. In this thesis, we refer to this method as the **RF MI** procedure.

2. A more general idea of the expansion of the Random Forest imputation procedure is given in Doove et al. (2014) and is implemented within the R- function `mice`. The idea is similar to Algorithm 4, but differs in three essential points: Step 2 in Algorithm 4 is replaced by random draws among elements in $\mathbf{K}_{\pi(j)}^{obs}$ for all $j \in \{1, \ldots, p\}$. Step 4 in Algorithm 4 is also replaced by not predicting imputation values as averages over the ensemble of trees, but rather as random draws from observations falling in leaves of the whole ensemble. Step 5 in Algorithm 4 is finally replaced by a fixed, pre-defined number of iterations as a stopping criterion, instead of a mixed version. Throughout the thesis, we will refer to this method as the **RF MICE** procedure.

A certain tendency towards using modern machine learning tools for statistical analysis due to the increased complexity of data generating processes have led us to the following two research questions, which will be tackled in this thesis:

(H4) Is it possible to extend missing value imputation procedures not only to the Random Forest method, but also to other learning algorithms such as boosting procedures ? Do we receive better results in terms of NRMSE or PFC ?

(H5) Using Random Forest models such as those implemented in **RF MI** and **RF MICE** as potential imputation schemes, does this always lead to correct statistical inference procedures ?

Regarding research question ($H4$), answers could be delivered in the first research paper ($P1$) by introducing different bagging and boosting methods. Therein, a detailed prescription of the working principles of the considered algorithms have been delivered and enriched with simulation studies and empirical examples, also outside the framework of Algorithm 4. Regarding the latter, an *incremental imputation algorithm* scheme proposed by Conversano and Siciliano (2009) has been appropriately modified. However, no additional statistical theory has been developed in article ($P1$). The summary of the research article can be found in Chapter 4 of this thesis. Regarding research question ($H5$), answers could be found within the research article ($P2$) under special designs such as paired data and repeated measurements ANOVA. The research article is summarized in Chapter 4 of this thesis and puts its focus more on the methodological and practical aspects of different imputation methods including **RF MI** and **RF MICE**.
Regarding article ($P2$), other imputation methods than Random Forest based models have been used for reasons of comparisons. In order to make them understandable, we will shortly introduce the `norm` and the `pmm` approaches within the `mice` package in R. They can also be found in Van Buuren (2018) on pages 58, 73 and 110. We will refer to these imputation methods as **NORM** and **PMM**. Here, `mice` is the R implementation of the MICE-procedure (multivariate imputation by chained equations) and belongs to the class of FCS imputation

models. It conducts imputations similarly to Algorithm 4, but by randomly sampling initial values for all missing cases from their corresponding set of observed values. Then, for every variable, an univariate response imputation model is assumed in accordance with the description given in 2. The model is trained based on the observed block similarly defined as $\mathbf{Y}_{-j}^{obs}$, $j \in \{1, \ldots, p\}$ and missing values are then imputed by using that model as a prediction model. One iteration then consists of one cycle through all variables such that, by default 5 iterations are then conducted in the R-function `mice` for imputation. The iteration process mimics the burn-in period of a Markovian chain for achieving a stabilizing posterior distribution (cf. Schafer, 1997, pages 91 and 119). `mice` has the advantage that it runs parallel in order to construct m imputation draws. Regarding the model, we additionally focused in this thesis on two non Random Forest based imputation schemes:

1. **NORM** is a Bayesian regression model and assumes for the model fitting tasks that the relation between response (i.e. the variable to impute; similarly to $\mathbf{K}_j^{obs}$) and the covariates (the variables similarly to $\mathbf{Y}_{-j}^{obs}$) is linear yielding the usual least-square estimate $\widehat{\boldsymbol{\beta}}_{LS}$ for the regression coefficient $\boldsymbol{\beta}$ in a linear model. Then, for the residual variance σ^2, a suitable prior is assumed such as the χ^2 distribution, while for the regression coefficients, a (multivariate) normal distribution with parameters as given in Van Buuren (2018) on page 58 is assumed. Imputations are then not draws from the directly trained model, but from a linear model with regression coefficient and residual variance drawn from the chosen prior distribution. Algorithm 3.1 in Van Buuren (2018) on page 58 gives more details on this.

2. **PMM** is a predictive mean matching imputation scheme, which imputes missing values as potential draws from a set of *donors*. The latter are usually defined as $d \in \{1, \ldots, n\}$ observed values, which come *closest* to the predicted values using the same modeling strategy as in **NORM**. In `mice`, the default number of donors is set to $d = 3$ while using a form of stochastic matching distance for evaluating closeness. Then, imputations are random draws from the set of donors. We refer to Van Buuren (2018), on page 73, Algorithm 3.3 for a detailed description.

The next section will extend the work given in $(P2)$ by theoretically deriving potential sources of incorrect statistical inference procedures when using Random Forest based imputation schemes.

3.3 Validity of Random Forest Multiple Imputation Procedures

This section extends the work in $(P2)$ theoretically, by identifying and verifying potential sources of statistical invalidity when using the **RF MI** procedure. Suppose that one has a collection of iid bivariate random vectors $\mathbf{X}_i = [X_{1i}, X_{2i}]^\top \in \mathbb{R}^2$ with $\mathbb{E}[\mathbf{X}_1] = \boldsymbol{\mu} = [\mu_1, \mu_2]^\top \in \mathbb{R}^2$ and $Cov(\mathbf{X}_1) = \boldsymbol{\Sigma} > 0$. Our aim is to test for no time effect in the means, i.e. we consider the two-sided alternative given by

$$H_0: \quad \mu_1 = \mu_2 \qquad \text{vs.} \qquad H_1: \quad \mu_1 \neq \mu_2. \tag{3.12}$$

A formal testing procedure for the null-hypothesis is given by the paired t-test, which is formalized by $\varphi_{n,complete} = \mathbb{1}\{|\tau_n| \geq t_{n-1,1-\alpha/2}\}$, where $t_{n-1,1-\alpha/2}$ is the $1-\alpha/2$-quantile of the t-distribution with $n-1$ degrees of freedom and $\alpha \in (0,1)$. The test statistic τ_n is given by $\tau_n = \sqrt{n} \cdot \bar{d}. / \{\widehat{\sigma}_n\}$ with $d_i = X_{1i} - X_{2i}$ for $i \in \{1, \ldots, n\}$, $\bar{d}.$ the mean of the d_i's and

$\widehat{\sigma}_n^2 = 1/(n-1)\sum_{i=1}^{n}(d_i - \bar{d}_\cdot)^2$ the estimated variance of the d_i's. In order to bring the notation in line with the ones from the previous sections, we have $\widehat{\boldsymbol{Q}}_n = \widehat{Q}_n = \bar{d}_\cdot$ and $\widehat{\boldsymbol{U}}_n = \widehat{\sigma}_n^2$ at this stage (Stage 2). The statistical test $\varphi_{n,complete}$ follows under H_0 a t_{n-1} distribution, if the bivariate vector $\mathbf{X}_1$ is multivariate normal distributed. The latter can also be dropped for the cost of obtaining an asymptotically exact test $\varphi_{n,complete}$, such that the assumption of a normal distribution is not of utmost importance. Within the framework of incomplete paired data, we consider the case where only the first component is partly missing, i.e. we have

$$\underbrace{\begin{bmatrix} X_{11}^{(c)} \\ X_{21}^{(c)} \end{bmatrix}, \ldots, \begin{bmatrix} X_{1n_1}^{(c)} \\ X_{2n_1}^{(c)} \end{bmatrix}}_{\mathbf{X}^{(c)}}, \underbrace{\begin{bmatrix} \text{NA} \\ X_{21}^{(i)} \end{bmatrix}, \ldots, \begin{bmatrix} \text{NA} \\ X_{2n_2}^{(i)} \end{bmatrix}}_{\mathbf{X}^{(i)}}, \tag{3.13}$$

where $n = n_1 + n_2$. Throughout the section, we assume a constant missing rate of $r \in (0,1)$, i.e. we have $n_1 = \lceil (1-r)\cdot n \rceil$ complete-pair observations and $n_2 = n - n_1$ observations with missing values in the first component. We additionally assume that the imputer has access to additional information, beside the one given in (3.13). Thus, we assume that auxiliary variables are available such that the extended data matrix has the form

$$\mathbf{M} = \begin{bmatrix} X_{11} & X_{21} & W_{11} & \ldots & W_{s1} \\ X_{12} & X_{22} & W_{12} & \ldots & W_{s2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{1n_1} & X_{2n_1} & W_{1n_1} & \ldots & W_{sn_1} \\ NA & X_{2n_1+1} & W_{1n_1+1} & \ldots & W_{sn_1+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ NA & X_{2n} & W_{1n} & \ldots & W_{sn} \end{bmatrix} \in \mathbb{R}^{n\times(s+2)}. \tag{3.14}$$

Note that Algorithm 4 makes use of a similar structure as model assumption (3.13) resp. (3.14) at every iteration step. Hence, during the imputation of missing values within Algorithm 4, we will be faced with data structures similar to (3.13). In case of higher feature dimensions, Algorithm 4 will still be using a data structure similar to (3.14), but without the initial iid assumption due to the mean imputation in Step 2. Hence, assuming model (3.13) resp. (3.14) will make the theoretical analysis more accessible while keeping the initial data structure of (3.14). Denoting with $\mathbf{V}_i = [X_{2i}, W_{1i}, \ldots, W_{si}]^\top \in \mathbb{R}^{s+1}$ for $i = 1, \ldots, n$ and $\widetilde{Y}_i = X_{1i}$ for $i = 1, \ldots, n_1$, missing values according to Algorithm 4 are imputed by training a Random Forest on $\mathcal{D}_{n_1} = \{[\mathbf{V}_i^\top, \widetilde{Y}_i]^\top : i = 1, \ldots, n_1\}$ and predicting missing cases by using $\{\mathbf{V}_i : i = n_1 + 1, \ldots, n\}$ and the trained Random Forest model. Imputing missing instances multiple times while obtaining complete-case estimators for the mean difference and its variance for plugging them into the test statistic τ_n can cause potential sources of invalidity of the test $\varphi_{n,complete}$. We mention two main potential sources:

(i) The distribution of τ_n under the null-hypothesis H_0 requires that the samples are iid. This is clearly not met, since imputations are predicted using a Random Forest $m_{n_1,M}(\cdot; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_{n_1})$, which itself is trained on the samples in $\mathcal{D}_{n_1}$. Hence, Algorithm 4 resp. **RF MI** imposes additional, unknown dependencies among the observations. Overcoming this source theoretically is a delicate issue, since central limit theorems for triangular arrays cannot be directly applied here. This, because row-wise independence for imputed values $\widehat{X}_{1,i} = m_{n_1,M}(\mathbf{V}_i; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_{n_1})$ for $i = n_1 + 1, \ldots, n$ is not directly given.

(ii) Incorrectly reflecting the uncertainty, i.e. the variance of the difference $X_{1i} - X_{2i}$ can lead to type-I-error inflation resp. deflation. Within the framework of missing values, the uncertainty involved in the imputation has also to be reflected correctly. This was emphasized by Definition 3.2, which especially required that the between imputation variance is approximately $Var(\bar{\boldsymbol{Q}}_{n,\infty}|\mathbf{Y})$. This was summarized under the term properness. Within the framework of model (3.13), $\boldsymbol{Q}$ is scalar and is nothing else than $\mathbb{E}[X_{11} - X_{21}] = \mu_1 - \mu_2$. Estimating $\boldsymbol{Q}$ using **RF MI** will lead to a scalar estimator that depends on the sample size n, the number of imputations m and the number of decision trees M. It will be denoted by $Q_{n,m,M}$ and refers to the aggregated, completed statistics $\bar{d}_{\cdot}$ using **RF MI** as an imputation scheme within the multiple imputation aggregation logic. Note that $\bar{\boldsymbol{Q}}_{n,m} = \bar{Q}_{n,m} = \bar{Q}_{n,m,M}$ in this case, while $\bar{\boldsymbol{U}}_{n,m} = U_{n,m} = U_{n,m,M}$ is the aggregated, completed variance estimator similarly to $\widehat{\sigma}_n^2$, but estimated on the imputed data set using **RF MI** with M decision trees.

We will not go into more detail regarding source (i). Instead, we will focus on source (ii) and state an example for potential implications regarding (ii).

Example 3.1. *Suppose that under a specific null-hypothesis H_0, T_n fulfills the following central-limit type convergence in distribution: $\sqrt{n}T_n \xrightarrow{H_0} \sigma Z \sim N(0, \sigma^2)$ as $n \to \infty$ with $\sigma^2 > 0$ and $Z \sim N(0,1)$. Then, an asymptotically exact testing procedure under a two-sided alternative is given by $\varphi_n = \mathbb{1}\{|\sqrt{n}T_n| > \sigma z_{1-\alpha/2}\}$, where z_α denotes the α-quantile of the standard normal distribution, with $\alpha \in (0,1)$. Suppose that $\hat{\gamma}_n$ is an inconsistent estimator of σ such that $\hat{\gamma}_n \xrightarrow{\mathbb{P}} \gamma \in (0, \sigma)$ with ratio $c := \sigma/\gamma > 1$. Assume furthermore that there exists a consistent estimator $\widehat{\sigma}_n$ for σ, such that $\widehat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma$, as $n \to \infty$. Then, for the following two test statistics, it follows under the null-hypothesis and from Slutzsky's theorem that*

$$T_n' := \sqrt{n}\frac{T_n}{\hat{\gamma}_n} \longrightarrow cZ, \qquad \widetilde{T}_n := \sqrt{n}\frac{T_n}{\hat{\sigma}_n} \longrightarrow Z \quad \text{as } n \to \infty.$$

Since $c > 1$, the limiting distribution of T_n' is stochastically greater than the limiting distribution of $\tilde{T}_n$ on the non-negative real-line, i.e. $F_{cZ}(x) \leq F_Z(x)$ for all $x \in \mathbb{R}_+$ and $F_{cZ}(x) < F_Z(x)$ for at least one $x \in \mathbb{R}_+$. Here, F denotes the corresponding distribution function under the null-hypothesis. Suppose that the imputer is unaware about the inconsistent estimation of the sampling variance σ^2 by $\hat{\gamma}_n$, whereas $\hat{\gamma}_n$ is a sampling variance estimator after imputation. Therefore, the imputer conducts the statistical analysis as if there has not been any missing values by believing that T_n' is standard normally distributed. However, the imputer's p-value p^{imp} experience a deflation under the null-hypothesis due to the following inequality using the property that T_n' is stochastically greater than $\tilde{T}_n$ on the non-negative real-line and the symmetry of the normal distribution:

$$p^{imp} = 2 \cdot \mathbb{P}[Z > |T_n'|] = 2 \cdot (1 - F_Z(|T_n'|)) \leq 2 \cdot (1 - F_{cZ}(|T_n'|)) = 2 \cdot \mathbb{P}[cZ > |T_n'|] =: p^{true},$$

where p^{true} is the correct p-value after imputation. But this will lead to an inflated type-I error, because we then obtain the following inequality:

$$T_{error}^{imp} := \mathbb{E}_{H_0}[\mathbb{1}\{p^{imp} < \alpha\}] \geq \mathbb{E}_{H_0}[\mathbb{1}\{p^{true} < \alpha\}] =: T_{error}^{true}.$$

Hence, inconsistent sample variance estimators can lead to inflated type-I errors.

Example 3.1 shows that underestimating the true variance of a $\sqrt{n}$-consistent estimator can lead to an inflated type-I-error rate. The upcoming theorem shows that this actually happens, if one uses **RF MI** as an imputation method.

Theorem 3.2. *Assume model* (3.13) *together with potential auxiliary variables as prescribed in* (3.14)*. Then for any finite choice of multiple imputations* $m \in \mathbb{N}$ *and a constant missing rate* $r \in (0,1)$*, the between variance estimator of* ***RF MI*** *tends to zero almost surely, as the number of decision trees increases. That is*

$$\mathbb{P}[\lim_{M\to\infty} B_{n,m,M} = 0] = 1,$$

where $B_{n,m,M}$ *is the between variance estimator as prescribed in* (3.5) *for* $\bar{Q}_{n,m,M}$*.*

Theorem 3.2 reveals that if the number of decision trees is sufficiently large, then the between imputation variance will vanish, if one makes use of the imputation procedure **RF MI**. If one denotes with $B_{n,\infty,\infty}$ the between variance estimate of an infinite multiple imputation procedure using **RF MI** in the sense that $B_{n,\infty,\infty} = \lim_{m\to\infty}\lim_{M\to\infty} B_{n,m,M}$, then this will lead to $\mathbb{E}[B_{n,\infty,\infty}|\mathbf{Y}] = 0$. The above result enables us to simplify the properness assumption as given in Definition 3.2, which was an essential presumption for the validity of inference procedures in frequentist's perspective. Note that under model (3.13) together with the hypothesis (3.12), but with complete cases instead, one would have an estimator $\widehat{Q}_n$ for $Q = \mu_1 - \mu_2$, that simply takes the average over the n complete paired differences $\{X_{1i} - X_{2i}\}_{i=1}^n$. Recalling that the **RF MI** leads to imputation estimators potentially depending on both, the number of iterations m and the number of decision trees M, we will denote its completed estimator as $\bar{Q}_{n,m,M}$, instead of $\bar{Q}_{n,m}$ as in the previous section. We can formulate the following corollary:

Corollary 3.1. *Under model* (3.13) *together with potential auxiliary variables as prescribed in* (3.14)*, the* ***RF MI*** *procedure fulfills condition* 1. *and* 3. *of Definition 3.2, if and only if* $\bar{Q}_{n,\infty,\infty} = \widehat{Q}_n$*, where* $\lim_{m\to\infty}\lim_{M\to\infty} Q_{n,m,M} = \lim_{M\to\infty}\lim_{m\to\infty} Q_{n,m,M} =: Q_{n,\infty,\infty}$ *exists.*

Corollary 3.1 guarantees that **RF MI** partly fulfills the condition of properness (Definition 3.2) under the design of (3.13), if at least $\bar{Q}_{n,\infty,\infty} = \widehat{Q}_n$ holds. Small deviations to this property will automatically lead to *improper* imputations and therefore, its correctness in terms of statistical inference from the frequentist's perspective cannot be guaranteed. This can easily happen, as our simulation examples have shown in $(P2)$. Think of the following example: Suppose that the first pair in model (3.13) is independent of the second pair and independent of the auxiliary variables. We denote with $\{X_{1i}\}_{i=n_1+1}^n$ the true values of the missing cases in $\mathbf{M}$. Then, properness of **RF MI** for the complete estimators $(\widehat{Q}_n, \widehat{U}_n)$ would yield the validity of condition 1 and 3 in Definition 3.2 and therefore, under Corollary 3.1 the equality of both estimators, $\bar{Q}_{n,\infty,\infty}$ and $\widehat{Q}_n$. This will result into the following implications, where we use equation (3.30) in the proof of Corollary 3.1:

$$\begin{aligned}
\bar{Q}_{n,\infty,\infty} &= \widehat{Q}_n \\
\frac{1}{n}\left\{\sum_{i\le n_1}(X_{1i} - X_{2i}) + \sum_{i>n_1}(m_{n,\infty}(\mathbf{V}_i;\mathcal{D}_{n_1}) - X_{2i})\right\} &= \frac{1}{n}\sum_{i=1}^n (X_{1i} - X_{2i}) \\
\frac{1}{n}\left\{\sum_{i\le n_1}(X_{1i} - X_{2i}) + \sum_{i>n_1}(m_{n,\infty}(\mathbf{V}_i;\mathcal{D}_{n_1}) - X_{2i})\right\} &= \frac{1}{n}\left\{\sum_{i\le n_1}(X_{1i} - X_{2i}) + \sum_{i>n_1}(X_{1i} - X_{2i})\right\} \\
\sum_{i>n_1} X_{1i} &= \sum_{i>n_1} m_{n_1,\infty}(\mathbf{V}_i;\mathcal{D}_{n_1}) \\
\sum_{i>n_1} X_{1i} &= \sum_{i>n_1}\sum_{\ell=1}^{n_1} W_{n_1,\ell}(\mathbf{V}_i)\cdot X_{2\ell}.
\end{aligned}$$

The above result imposes dependence structures between $\{X_{1i}\}_{i>n_1}$ and $\{\mathbf{V}_i\}_{i>n_1}$. This is a violation towards the initial independence assumption. Therefore, $\widehat{Q}_n = \bar{Q}_{n,\infty,\infty}$ cannot hold such that **RF MI** is not proper, at least for this scenario. Other scenarios have been discovered in $(P2)$ as well through simulation.

In summary, we can make the following conclusions regarding **RF MI** based on Theorem 3.2 and Corollary 3.1:

1. As the short example for uncorrelated pairs above has shown, **RF MI** cannot guarantee properness according to Definition 3.2. Therefore, **RF MI** does not guarantee the correct validity of statistical inference from the frequentist's perspective, since properness is an unavoidable assumption for statistical validity after imputation (see Result 4.1 in Rubin, 2004).

2. The result that the between variance estimator vanishes almost surely as the number of decision trees in Breiman's Random Forest increases yields that **RF MI** behaves in its limit ($M \to \infty$) as a single imputation scheme. This will ignore Rubin's initial idea of accounting for uncertainty due to the presence of missing values, and therefore acts like as one would be at Stage 2, although one is *living* at Stage 3 resp. Stage 4. Hence, according to Rubin (2004), pages 12 - 13, the lack of considering the *extra variability due to missing cases* can cause variance underestimation as given in Example 3.1 (see also the variance decomposition (3.4) valid at Stage 3). This is different to the general principle in Machine Learning prediction, where additional decision trees in Breiman's Random Forest are recommended and are only contrary to computational time (cf. Wager et al., 2014, Scornet, 2016). However, in case of statistical inference, it can also be contrary to statistical validity as an imputation scheme, which has been revealed by our theoretical and practical work. Therefore, adding additional trees in **RF MI** for the purpose of statistical inference in missing frameworks is not recommended.

3.4 Proofs of the Chapter

Proof of Proposition 3.1. Let the missing mechanism be at least MAR, that is $\mathbb{P}[\mathbf{R}|\mathbf{Y},\xi] = \mathbb{P}[\mathbf{R}|\mathbf{Y}_{obs},\xi]$. Then we have

$$\begin{aligned}
\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs},\mathbf{R}] &= \int \mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs},\mathbf{R},\xi]\cdot\mathbb{P}[\xi]d\xi \\
&= \int \frac{\mathbb{P}[\mathbf{R}|\mathbf{Y},\xi]}{\mathbb{P}[\mathbf{Y}_{obs},\mathbf{R},\xi]}\cdot\mathbb{P}[\mathbf{Y},\xi]\cdot\mathbb{P}[\xi]d\xi \\
&= \int \frac{\mathbb{P}[\mathbf{R}|\mathbf{Y}_{obs},\xi]}{\mathbb{P}[\mathbf{Y}_{obs},\mathbf{R},\xi]}\cdot\mathbb{P}[\mathbf{Y},\xi]\cdot\mathbb{P}[\xi]d\xi \\
&= \int \frac{\mathbb{P}[\mathbf{Y},\xi]}{\mathbb{P}[\mathbf{Y}_{obs},\xi]}\cdot\mathbb{P}[\xi]d\xi \\
&= \int \mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs},\xi]\cdot\mathbb{P}[\xi]d\xi = \mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]. \qquad (3.15)
\end{aligned}$$

The third equality follows from applying the definition of MAR. ■

Proof of Theorem 3.1. We first show that under the ignorability assumption, the missing mechanism can be neglected when conducting Bayesian statistical inference in hypothesis testing. Then we will show that infinite imputation draws enables us to reconstruct the posterior $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}]$, on which Bayesian inference is then conducted. Therefore, let us start with the Bayes Theorem and the assumption of ignorability, which implies the independence of $\boldsymbol{Q} = f(\theta)$ towards ξ. Then we can deduce that

$$\mathbb{P}[f(\theta),\xi|\mathbf{Y}_{obs},\mathbf{R}] = c^{-1}\cdot\mathbb{P}[\mathbf{R},\mathbf{Y}_{obs}|f(\theta),\xi]\cdot\mathbb{P}[f(\theta)]\cdot\mathbb{P}[\xi], \qquad (3.16)$$

where c is a normalizing constant independent of θ and ξ. On other words, $\mathbb{P}[f(\theta),\xi|\mathbf{Y}_{obs},\mathbf{R}] \propto \mathbb{P}[\mathbf{R},\mathbf{Y}_{obs}|f(\theta),\xi]\cdot\mathbb{P}[\theta]\cdot\mathbb{P}[\xi]$. Recalling equation (12) in Schafer (1997) on page 12, one has

$$\mathbb{P}[\mathbf{R},\mathbf{Y}_{obs}|\xi,f(\theta)] = \int \mathbb{P}[\mathbf{R}|\mathbf{Y},\xi]\cdot\mathbb{P}[\mathbf{Y}|f(\theta)]d\mathbf{Y}_{mis}, \qquad (3.17)$$

which yields under the MAR assumption to

$$\begin{aligned}
\mathbb{P}[\mathbf{R},\mathbf{Y}_{obs}|\xi,f(\theta)] &= \mathbb{P}[\mathbf{R}|\mathbf{Y}_{obs},\xi]\cdot\int\mathbb{P}[\mathbf{Y}|f(\theta)]d\mathbf{Y}_{mis} \\
&= \mathbb{P}[\mathbf{R}|\mathbf{Y}_{obs},\xi]\cdot\mathbb{P}[\mathbf{Y}_{obs}|f(\theta)]. \qquad (3.18)
\end{aligned}$$

Since the quantity $\boldsymbol{Q}$ can be considered as a function of the data generating parameter θ, we have $\boldsymbol{Q} = f(\theta)$ for some measurable function f. Finally, we can deduce that

$$\begin{aligned}
\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs},\mathbf{R}] = \mathbb{P}[f(\theta)|\mathbf{Y}_{obs},\mathbf{R}] &= \int\mathbb{P}[f(\theta),\xi|\mathbf{Y}_{obs},\mathbf{R}]d\xi \\
&= c^{-1}\int\mathbb{P}[\mathbf{R},\mathbf{Y}_{obs}|f(\theta),\xi]\cdot\mathbb{P}[f(\theta)]\cdot\mathbb{P}[\xi]d\xi \\
&= \mathbb{P}[\mathbf{Y}_{obs}|f(\theta)]\cdot\mathbb{P}[f(\theta)]\cdot c^{-1}\int\mathbb{P}[\mathbf{R}|\mathbf{Y}_{obs},\xi]\cdot\mathbb{P}[\xi]d\xi \\
&= c^{*}\cdot\mathbb{P}[\mathbf{Y}_{obs}|f(\theta)]\cdot\mathbb{P}[f(\theta)] \\
&= c^{*}\cdot\mathbb{P}[\mathbf{Y}_{obs}|\boldsymbol{Q}]\cdot\mathbb{P}[\boldsymbol{Q}] \\
&= c^{*}\cdot\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}]\cdot\mathbb{P}[\mathbf{Y}_{obs}], \qquad (3.19)
\end{aligned}$$

where $c^* = c^{-1} \cdot \int \mathbb{P}[\mathbf{R}|\mathbf{Y}_{obs}, \xi] \cdot \mathbb{P}[\xi] d\xi$ is independent of θ resp. $\boldsymbol{Q}$. The third equality follows from applying equation (3.16) and the fourth equality from (3.18) as given in Schafer (1997) on page 12. Therefore, we can deduce that $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{R}] \propto \mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}]$. Hence, the missing mechanism can be ignored when considering the posterior distribution $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{R}]$, i.e. considering $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}]$ suffices for Bayesian inference.

Now, return to the predictive posterior distribution $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}]$. Since the missing mechanism is ignorable and thus at least MAR, we know from Proposition 3.1 that $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}]$ $= \mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]$. Therefore, the imputed draws $\{\mathbf{Y}_{mis}^{(t)}\}_{t=1}^m$ from $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}]$ are actually independent draws from $\mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}]$. If we denote with $\mathbb{P}_A$ the Bayesian model of the analyst and $\mathbb{P}_I$ the Bayesian model of the imputer, then we have $\mathbb{P}_A = \mathbb{P}_I =: \mathbb{P}$ due to model assumptions. The imputation procedure will then lead to the preservation of the posterior distribution, since

$$\begin{aligned}\int \mathbb{P}_A[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}] \cdot \mathbb{P}_I[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}] d\mathbf{Y}_{mis} &= \int \mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}] \cdot \mathbb{P}[\mathbf{Y}_{mis}|\mathbf{Y}_{obs}] d\mathbf{Y}_{mis} \\ &= \mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}]. \end{aligned} \tag{3.20}$$

Hence, Bayesian inference should not be distorted and is therefore valid. This result can now be applied to simulate the actual posterior distribution of $\boldsymbol{Q}$ using the independent draws $\{\mathbf{Y}_{mis}^{(t)}\}_{t=}^m$, $m \in \mathbb{N}$ for missing cases. Similarly to Rubin (2004), page 82, let C be some (measurable) region. Then, we have

$$\begin{aligned}\mathbb{P}_A[\boldsymbol{Q} \in C|\mathbf{Y}_{obs}] &= \lim_{m\to\infty} \frac{1}{m} \sum_{t=1}^m \mathbb{P}_A[\boldsymbol{Q} \in C|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}^{(t)}] \\ &= \lim_{m\to\infty} \frac{1}{m} \sum_{t=1}^m \mathbb{P}[\boldsymbol{Q} \in C|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}^{(t)}] \\ &= \mathbb{P}[\boldsymbol{Q} \in C|\mathbf{Y}_{obs}], \end{aligned} \tag{3.21}$$

where the last equality follows from (3.20). Hence, $\mathbb{P}[\boldsymbol{Q}|\mathbf{Y}_{obs}]$ can be recovered within the multiple imputation strategy using an infinite number of imputations. ■

Proof of Theorem 3.2. Let $m \in \mathbb{N}$ be the number of imputations, which is assumed to be fixed. Denote with $\widehat{X}_{1i,M}$ for $i = n_1 + 1, \ldots, n$ the Random Forest prediction using data set $\mathcal{D}_{n_1}$ and a finite choice of decision trees $M \in \mathbb{N}$. That is, $\widehat{X}_{1i,M} = m_{n_1,M}(\mathbf{V}_i; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M, \mathcal{D}_{n_1})$, where $\mathbf{V}_i = [X_{2i}, W_{1i}, \ldots, W_{si}]^\top$. In fact, if one replicates the Random Forest method in order to obtain multiple imputations according to Algorithm 4, this will result into m Random Forest estimates $\{m_{n_1,M}^{(\ell)}\}_{\ell=1}^m$. Therefore, we will have m imputations each with exactly n_2 point predictions $\{\widehat{X}_{1i,M}^{(\ell)}\}_{i=n_1+1}^n$. Since all of them are trained on the data set $\mathcal{D}_{n_1}$, they will only differ w.r.t. the realizations of the vectors $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M$, which are responsible for the sampling mechanism and the feature sub-spacing procedure. Therefore, we have to write $m_{n_1,M}^{(\ell)}(\cdot) = m_{n_1,M}(\cdot; \boldsymbol{\Theta}_1^{(\ell)}, \cdots, \boldsymbol{\Theta}_M^{(\ell)}, \mathcal{D}_{n_1})$ for every $\ell = 1, \ldots, m$. Since imputations are conducted independent of each other, we can conclude that the sequence $\{\boldsymbol{\Theta}_t^{(\ell)}\}_{t,\ell}$ is a sequence of independent and identically distributed random vectors, given the training set $\mathcal{D}_{n_1}$. Now, denote with

$$d_i^{(obs)} = X_{1i} - X_{2i}, \quad i = 1, \ldots, n_1, \tag{3.22}$$

$$d_{i,\ell,M}^{(imp)} = \widehat{X}_{1i,M}^{(\ell)} - X_{2i}, \quad i = n_1 + 1, \ldots, n, \quad \ell = 1, \ldots, m, \tag{3.23}$$

the complete and imputed paired differences. The complete-case estimator based on each computed data set is then given by

$$\begin{aligned}\widehat{Q}_{n,\ell,M} &= \frac{1}{n}\left(\sum_{i=1}^{n_1} d_i^{(obs)} + \sum_{i=n_1+1}^{n} d_{i,\ell,M}^{(imp)}\right) \\ &= \alpha \bar{d}_{\cdot}^{(obs)} + (1-\alpha)\bar{d}_{\cdot,\ell,M}^{(imp)}, \end{aligned} \tag{3.24}$$

where $\alpha = n_1/n = \alpha(r) = \lceil (1-r) \rceil$ is constant, since the missing rate is assumed to be constant. $\bar{d}_{\cdot}^{(obs)}$ and $\bar{d}_{\cdot,\ell,M}^{(imp)}$ denote the corresponding means of $\{d_i^{(obs)}\}_{i=1}^{n_1}$ and $\{d_{i,\ell,M}^{(imp)}\}_{i=n_1+1}^{n}$. If we consider the average of $\frac{1}{n_2}\sum\limits_{i=n_1+1}^{n} \widehat{X}_{1i,M}^{(\ell)}$ over the m-repeated imputations while setting $M' = mM$, we can observe for every $i = n_1+1, \ldots, n$ the following:

$$\begin{aligned}\frac{1}{m}\sum_{\ell=1}^{m} \widehat{X}_{1i,M}^{(\ell)} &= \frac{1}{m}\sum_{\ell=1}^{m} m_{n_1,M}(\mathbf{V}_i; \boldsymbol{\Theta}_1^{(\ell)}, \ldots, \boldsymbol{\Theta}_M^{(\ell)}, \mathcal{D}_{n_1}) \\ &= \frac{1}{m}\sum_{\ell=1}^{m}\frac{1}{M}\sum_{t=1}^{M} m_{n_1,1}(\mathbf{V}_i; \boldsymbol{\Theta}_t^{(\ell)}, \mathcal{D}_{n_1}) \\ &= \frac{1}{M'}\sum_{\ell=1}^{m}\sum_{t=1}^{M} m_{n_1,1}(\mathbf{V}_i; \boldsymbol{\Theta}_t^{(\ell)}, \mathcal{D}_{n_1}) \\ &= \frac{1}{M'}\sum_{t=1}^{M'} m_{n_1,1}(\mathbf{V}_i; \boldsymbol{\Theta}_t, \mathcal{D}_{n_1}) \\ &= m_{n_1,M'}(\mathbf{V}_i; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{M'}, \mathcal{D}_{n_1}), \end{aligned} \tag{3.25}$$

where $\{\boldsymbol{\Theta}_t\}_{t=1}^{M'}$ is obtained after renumbering $\{\boldsymbol{\Theta}_t^{(\ell)}\}_{t,\ell}$. $\{\boldsymbol{\Theta}_t\}_{t=1}^{M'}$ remains a sequence of iid random vectors given the data set $\mathcal{D}_{n_1}$, as explained in the beginning of the proof. Because of the strong law of large numbers and the fact that $\lim\limits_{M\to\infty} M'/M = m$ is constant, we can deduce $\mathbb{P}_{\boldsymbol{\Theta}}$ -almost surely that

$$\begin{aligned}\lim_{M\to\infty} m_{n_1,M'}(\mathbf{V}_i; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{M'}, \mathcal{D}_{n_1}) &= \lim_{M'\to\infty} m_{n_1,M'}(\mathbf{V}_i; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{M'}, \mathcal{D}_{n_1}) \\ &= \mathbb{E}_{\boldsymbol{\Theta}}[m_{n_1,1}(\mathbf{V}_i; \boldsymbol{\Theta}, \mathcal{D}_{n_1})] = m_{n_1,\infty}(\mathbf{V}_i; \mathcal{D}_{n_1}). \end{aligned} \tag{3.26}$$

Returning now to the between imputation variance, we can deduce that the latter depends on the sample size n, the number of imputations m and the number of decision trees M

according to Algorithm 4. Therefore, we obtain the following computations

$$
\begin{aligned}
B_{n,m,M} &= \frac{1}{m-1}\sum_{\ell=1}^{m}\left(\widehat{Q}_{n,\ell,M} - \overline{\widehat{Q}}_{n,\cdot,M}\right)^2 \\
&= \frac{1}{m-1}\sum_{\ell=1}^{m}\left(\alpha\bar{d}_{\cdot}^{(obs)} + (1-\alpha)\bar{d}_{\cdot,\ell,M}^{(imp)} - \left\{\alpha\bar{d}_{\cdot}^{(obs)} + (1-\alpha)\bar{d}_{\cdot,\cdot,M}^{(imp)}\right\}\right)^2 \\
&= \frac{1}{m-1}\sum_{\ell=1}^{m}\left((1-\alpha)\bar{d}_{\cdot,\ell,M}^{(imp)} - (1-\alpha)\bar{d}_{\cdot,\cdot,M}^{(imp)}\right)^2 \\
&= \frac{(1-\alpha)^2}{m-1}\sum_{\ell=1}^{m}\left(\frac{1}{n_2}\sum_{i=n_1+1}^{n}\left\{\left(\widehat{X}_{1i,M}^{(\ell)} - X_{2i}\right) - \frac{1}{m}\sum_{\ell=1}^{m}\left(\widehat{X}_{1i,M}^{(\ell)} - X_{2i}\right)\right\}\right)^2 \\
&= \frac{(1-\alpha)^2}{m-1}\sum_{\ell=1}^{m}\left(\frac{1}{n_2}\sum_{i=n_1+1}^{n}\left\{\widehat{X}_{1i,M}^{(\ell)} - \frac{1}{m}\sum_{\ell=1}^{m}\widehat{X}_{1i,M}^{(\ell)}\right\}\right)^2 \\
&= \frac{(1-\alpha)^2}{m-1}\sum_{\ell=1}^{m}\left(\frac{1}{n_2}\sum_{i=n_1+1}^{n}\Big\{m_{n_1,M}(\mathbf{V}_i;\boldsymbol{\Theta}_1^{(\ell)},\ldots,\boldsymbol{\Theta}_M^{(\ell)},\mathcal{D}_{n_1}) - \right. \\
&\qquad\qquad \left. m_{n_1,M'}(\mathbf{V}_i;\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_{M'},\mathcal{D}_{n_1})\Big\}\right)^2, \qquad (3.27)
\end{aligned}
$$

where the last equality follows from applying equation (3.25) and plugging in the alternative representation of the imputed values $\widehat{X}_{1i,M}^{(\ell)}$ as a Random Forest prediction. Using the result given in (3.26) and plugging it into (3.27) we can finally obtain $\mathbb{P}_{\boldsymbol{\Theta}}$ - almost surely that

$$
\begin{aligned}
\lim_{M\to\infty} B_{n,m,M} &= \frac{(1-\alpha)^2}{m-1}\sum_{\ell=1}^{m}\left(\frac{1}{n_2}\sum_{i=n_1+1}^{n}\lim_{M\to\infty}\Delta_{n_1,M,M'}(\mathbf{V}_i)\right)^2 \\
&= \frac{(1-\alpha)^2}{m-1}\sum_{\ell=1}^{m}\left(\frac{1}{n_2}\sum_{i=n_1+1}^{n} m_{n_1,\infty}(\mathbf{V}_i;\mathcal{D}_{n_1}) - m_{n_1,\infty}(\mathbf{V}_i;\mathcal{D}_{n_1})\right)^2 = 0, \qquad (3.28)
\end{aligned}
$$

where $\Delta_{n_1,M,M'}(\mathbf{V}_i) = m_{n_1,M}(\mathbf{V}_i;\boldsymbol{\Theta}_1^{(\ell)},\ldots,\boldsymbol{\Theta}_M^{(\ell)},\mathcal{D}_{n_1}) - m_{n_1,M'}(\mathbf{V}_i;\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_{M'},\mathcal{D}_{n_1})$. Since $1 = \mathbb{P}_{\boldsymbol{\Theta}}[\lim_{M\to\infty} B_{n,m,M} = 0] = \mathbb{E}[\mathbb{1}\{\lim_{M\to\infty} B_{n,m,M} = 0\}|\mathcal{D}_{n_1}]$, we can extend the almost sure convergence to the whole measure $\mathbb{P}$, by conducting the following computations:

$$
\begin{aligned}
\mathbb{P}[\lim_{M\to\infty} B_{n,m,M} = 0] &= \mathbb{E}[\mathbb{1}\{\lim_{M\to\infty} B_{n,m,M} = 0\}] = \mathbb{E}[\mathbb{E}[\mathbb{1}\{\lim_{M\to\infty} B_{n,m,M} = 0\}|\mathcal{D}_{n_1}]] \\
&= \mathbb{E}[1] = 1.
\end{aligned}
$$

■

Proof of Corollary 3.1. Denote with $\mathbf{Y}$ the truly completed data matrix of $\mathbf{M}$. From Theorem 3.2, we know that the between imputation variance for infinite-m multiple imputation procedure using `missForest` and an infinite number of decision trees vanishes almost surely, i.e. $B_{n,\infty,\infty} := \lim_{m\to\infty}\lim_{M\to\infty} B_{n,m,M} = 0$, $\mathbb{P}$-almost surely. First, we show the existence of $\bar{Q}_{n,\infty,\infty}$. Following the same logic and notation as in the proof of Theorem 3.2, we know that

$\{\boldsymbol{\Theta}_t^{(\ell)}\}_{t,\ell}$ is iid given the data $\mathcal{D}_n$. Then, we have

$$\begin{aligned}\bar{Q}_{n,m,M} &= \frac{1}{m}\sum_{\ell=1}^{m}\widehat{Q}_{n,\ell,M} = \frac{1}{m}\frac{1}{n}\sum_{\ell=1}^{m}\left\{\sum_{i=1}^{n_1} d_i^{(obs)} + d_{i,\ell,M}^{(imp)}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} d_i^{(obs)} + \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{m}\sum_{\ell=1}^{m} m_{n_1,M}(\mathbf{V}_i;\boldsymbol{\Theta}_1^{(\ell)},\ldots,\boldsymbol{\Theta}_M^{(\ell)},\mathcal{D}_{n_1}) - X_{2i}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} d_i^{(obs)} + \frac{1}{n}\sum_{i=1}^{n}\left\{m_{n_1,M'}(\mathbf{V}_i;\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_{M'},\mathcal{D}_{n_1}) - X_{2i}\right\}, \end{aligned} \tag{3.29}$$

where the last equality follows from equation (3.25). Since

$$\begin{aligned}\lim_{m\to\infty}\lim_{M\to\infty} m_{n,M'}(\mathbf{V}_i;\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_{M'},\mathcal{D}_{n_1}) &= \mathbb{E}_{\boldsymbol{\Theta}}[m_{n_1,1}(\mathbf{V}_i;\boldsymbol{\Theta},\mathcal{D}_{n_1})] = m_{n_1,\infty}(\mathbf{V}_i;\mathcal{D}_{n_1}) \\ &= \lim_{M\to\infty}\lim_{m\to\infty} m_{n,M'}(\mathbf{V}_i;\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_{M'},\mathcal{D}_{n_1}),\end{aligned}$$

we can deduce from (3.29) that $\lim_{m\to\infty}\lim_{M\to\infty}\bar{Q}_{n,m,M} = \lim_{M\to\infty}\lim_{m\to\infty}\bar{Q}_{n,m,M}$ holds such that

$$\bar{Q}_{n,\infty,\infty} = \frac{1}{n}\sum_{i=1}^{n_1}(X_{1i} - X_{2i}) + \frac{1}{n}\sum_{i=n_1+1}^{n}(m_{n_1,\infty}(\mathbf{V}_i;\mathcal{D}_{n_1}) - X_{2i}) \tag{3.30}$$

in almost sure sense.
Suppose now that **RF MI** fulfills condition 1. and 3. of Definition 3.2. Then, by the third characteristic of the latter definition, we have that $0 = \mathbb{E}[B_{n,\infty,\infty}|\mathbf{Y}] = Var(\bar{Q}_{n,\infty,\infty}|\mathbf{Y})$. Hence, we have

$$0 = Var(\bar{Q}_{n,\infty,\infty}|\mathbf{Y}) = \mathbb{E}[(\bar{Q}_{n,\infty,\infty} - \mathbb{E}[\bar{Q}_{n,\infty,\infty}|\mathbf{Y}])^2|\mathbf{Y}],$$

which implies that $\mathbb{E}[\bar{Q}_{n,\infty,\infty}|\mathbf{Y}] = \bar{Q}_{n,\infty,\infty}$. However, according to the first characteristic of Definition 3.2, this will yield to $\widehat{Q}_n = \mathbb{E}[\bar{Q}_{n,\infty,\infty}|\mathbf{Y}] = \bar{Q}_{n,\infty,\infty}$. For the other direction, assume that $\widehat{Q}_n = \bar{Q}_{n,\infty,\infty}$. We consider condition 1. and 3. from Definition 3.2.

1. Regarding the first condition of Definition 3.2, we can deduce that

$$\mathbb{E}[\bar{Q}_{n,\infty,\infty}|\mathbf{Y}] = \mathbb{E}[\widehat{Q}_n|\mathbf{Y}] = \widehat{Q}_n. \tag{3.31}$$

   The first equality follows from $\bar{Q}_{n,\infty,\infty} = \widehat{Q}_n$. The second equality from the measurability of $\widehat{Q}_n$ towards the sigma field generated by $\mathbf{Y}$.

2. The last property of Definition 3.2 requires that $\mathbb{E}[B_{n,\infty,\infty}|\mathbf{Y}] = Var(\bar{Q}_{n,\infty,\infty}|\mathbf{Y})$. Since the assumptions of Theorem 3.2 are met, we can conclude that $B_{n,\infty,\infty} = 0$ holds almost surely. Therefore, $\mathbb{E}[B_{n,\infty,\infty}|\mathbf{Y}] = 0$. Similarly, we can obtain in almost sure sense that

$$Var(\bar{Q}_{n,\infty,\infty}|\mathbf{Y}) = Var(\widehat{Q}_n|\mathbf{Y}) = \mathbb{E}[\widehat{Q}_n^2|\mathbf{Y}] - \mathbb{E}[\widehat{Q}_n|\mathbf{Y}]^2 = \widehat{Q}_n^2 - \widehat{Q}_n^2 = 0.$$

   Therefore, $\mathbb{E}[B_{n,\infty,\infty}|\mathbf{Y}] = 0 = Var(\bar{Q}_{n,\infty,\infty}|\mathbf{Y})$.

■

Chapter 4

Summary of the Scientific Articles

4.1 Article 1: *Prediciting Missing Values: A comparative study on non-parametric approaches for imputation.*

In this article, we have focused on research question ($H4$) that has been stated and motivated in Chapter 3. The aim was to use Algorithm 4 as a starting point for implementing other imputation schemes than the Random Forest model within the statistical software R. The key idea was to use CART-based algorithms for imputation, since the latter can be easily adopted to mixed-type data and are easy to tune during choices of potentially suitable hyperparameters. Furthermore, the restriction to CART-based algorithms are comparably faster than other complex Machine Learning algorithms such as neural nets. The latter, however, can still be regarded as a potential field for future research as novel imputation schemes have been mainly focused on the bagging and boosting principles together with nearest neighbor methods (cf. Conversano and Siciliano, 2009, Wang and Feng, 2010, Stekhoven and Bühlmann, 2012). We considered the following algorithms within the scheme of Algorithm 4:

1. The stochastic gradient tree boosting as explained in Chapter 1, Section 1.4 for both types of data, categorical and continuous outcomes. The algorithm is also capable of treating both types simultaneously, which has been done during an empirical data analysis. The package `gbm` within the statistical software R has been used.

2. The C5.0 classifier implemented in the R package `C50` has been used to impute missing cases for categorical outcomes only. The latter is not able to treat continuous outcomes such that its usage is restricted to classification problems. The working principle follows the explanation given in Chapter 1, Section 1.5 while using the cross-entropy as an impurity measure (see Definition 1.6). Differently to the Random Forest, post-pruning will take place. That is, the decision tree is fully grown and then collapsed according to the criterion given in equation (1.22) without using bagging principles and feature sub-spacing.

3. The Random Forest method as explained in Algorithm 3, but using stratified sampling during the bagging procedure. The *stratified Random Forest* (Liaw and Wiener, 2002) is used for classification procedures to represent low-frequent outcomes appropriately during the training phase. Therefore, we implemented this method in R only for categorical outcomes.

4. For continuous outcomes, we proposed different re-sampling strategies during the bagging step of the Random Forest method as explained in Algorithm 3. Instead of simply

focusing on subsampling and resampling with replacement, we extended the possibility to use parametric bootstrapped samples from a multivariate normal distribution, where its mean vector and covariance matrix are estimated based on $[\mathbf{K}^{obs}_{\pi(j)}, \mathbf{Y}^{obs}_{-\pi(j)}]$ as given in step 3 of Algorithm 4. This procedure will be abbreviated as the *normal Random Forest*. In addition, we changed the re-sampling step to a multivariate kernel sampling procedure, in which we used a multivariate normal kernel together with a normal-scaled bandwidth-matrix estimator. This procedure is referred to as the *kernel Random Forest*.

5. We included also Bayesian procedures for imputing missing values using CART approaches. The CART construction process as explained in Chapter 1, Section 1.5 is modified by assuming that each characterizing element of decision trees is imposed by a suitable prior distribution. Each tree is then aggregated under an additive modelling scheme similarly to the boosting principle. The method has been developed in Chipman et al. (2010) and implemented in the R-package `BART` (Bayesian Additive Regression Trees).

6. Classical imputation schemes such as the `mice`-package using the predictive mean-matching approach and the `missForest` package have been considered as well.

7. The proposed methods have also been implemented within an alternative Algorithm other than Algorithm 4 and are referred to as the *Incremental Imputation Algorithm* (IIA) developed by Conversano and Siciliano (2009).

Our simulation study can be separated in two parts: First, we artificially generated seven different data sets with dimensions $(n, p) = (250, 15)$ consisting either categorical or continuous variables, where missing values have been artificially inserted under the MCAR or MAR scheme with various missing rates under $MC = 500$ Monte-Carlo iterations. Regarding mixed-type data, we considered the *German Credit Data* from the UCI Machine Learning Repository, which consisted of $n = 1,000$ observations with $p = 20$ variables. In addition, *Facebook data* was taken into consideration from the *myPersonality.com* Facebook application app, which consisted of $n = 463$ participants with $p = 13$ variables. Both data sets have continuous and categorical outcomes. Our results revealed that imputation accuracy in terms of NRMSE and PFC as defined in (3.10) and (3.11) could be improved by the stochastic gradient tree boosting for categorical variables, while the kernel Random Forest performed comparably better for continuous outcomes. This is the case when using the scheme of Algorithm 4, instead of the IIA and the `missForest` as the benchmark model. Superiority towards the benchmark model using the NRMSE and PFC as performance measures have been inferred by the Brunner-Munzel test. Therefore, we proposed the `missBooPF`-Algorithm, that combines the stochastic gradient boosting method and the kernel Random Forest as an imputation scheme, depending on the scale of the variables to be imputed. This comes with the cost of additional computational time loads, which originates from the modified re-sampling scheme among others.

Finally, we could conclude that imputation error in terms of NRMSE and PFC could be reduced when changing Algorithm 4 by using a combination of the kernel Random Forest and the stochastic gradient boosting method, depending on the outcome scale. However, the usage of the NRMSE as a *universal measure* for evaluating imputation schemes can be misleading, as our next article ($P2$) will show, especially under aspects of inferential statistics.

4.2 Article 2: *A cautionary tale on using imputation methods for inference in matched pairs design.*

In this work, we mainly focused on research question ($H5$) by paying special attention to bivariate data with missing cases in both arms. The idea was to compare different multiple imputation procedures together with a recently proposed testing procedure developed by Amro and Pauly (2017) for testing equal means in bivariate data with missing cases. We wanted to know whether test decisions are distorted when data is partially observed compared to the complete-observation case. Therefore, we considered a collection of $n \in \mathbb{N}$ iid bivariate random vectors $\{\mathbf{X}_i\}_{i=1}^n$ such that $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu} = [\mu_1, \mu_2]^\top \in \mathbb{R}^2$ with an arbitrary, positive definite covariance matrix $\boldsymbol{\Sigma} > 0$. Missing cases will appear in both arms such that $n = n_1 + n_2 + n_3$ with

$$\underbrace{\begin{bmatrix} X_{11}^{(c)} \\ X_{21}^{(c)} \end{bmatrix}, \ldots, \begin{bmatrix} X_{1n_1}^{(c)} \\ X_{2n_1}^{(c)} \end{bmatrix}}_{\mathbf{X}^{(c)}}, \underbrace{\begin{bmatrix} X_{11}^{(i)} \\ NA \end{bmatrix}, \ldots, \begin{bmatrix} X_{1n_2}^{(i)} \\ NA \end{bmatrix} \begin{bmatrix} NA \\ X_{21}^{(i)} \end{bmatrix}, \ldots, \begin{bmatrix} NA \\ X_{2n_3}^{(i)} \end{bmatrix}}_{\mathbf{X}^{(i)}}. \tag{4.1}$$

We focused on statistical inference from a frequentist's perspective by considering the null-hypothesis of no mean time effect, i.e.

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_1 : \mu_1 \neq \mu_2.$$

A possible testing procedure for the null-hypothesis is the paired t-test, which should maintain exact or asymptotically exact type-I error results under the complete-case framework of (4.1). However, the procedure faces difficulties under the incomplete framework as given in (4.1). Amro and Pauly (2017) proposed a weighted combination test as a data adjusted method by considering the paired t-statistic for $\mathbf{X}^{(c)}$ and the Welch-type statistic for $\mathbf{X}^{(i)}$, i.e. $T_{ML} = \sqrt{a}T_t(\mathbf{X}^{(c)}) + \sqrt{1-a}T_w(\mathbf{X}^{(i)})$. $T_t(\mathbf{X}^{(c)})$ is in this case the test statistic of the paired t-test based on the samples given in $\mathbf{X}^{(c)}$ and $T_w(\mathbf{X}^{(i)})$ is the test statistic of the Welch's t-test treating the incomplete pairs in $\mathbf{X}^{(i)}$ as two samples. The weighting coefficient a is then given by $a = 2n_1/(n + n_1)$ and critical values for T_{ML} have been computed under a specific permutation scheme permuting complete cases $\mathbf{X}^{(c)}$ and incomplete cases $\mathbf{X}^{(i)}$ separately. From the field of multiple imputation procedures, we considered the following schemes:

1. The FCS approach implemented in R under the package `mice` using the Bayesian linear regression method "norm" as an imputation scheme. The procedure has been explained in detail in Chapter 3 and is referred to as **NORM**.

2. The FCS approach implemented in R under the package `mice` using the predictive mean-matching method "pmm" as an imputation scheme. The procedure has been explained in detail in Chapter 3 and is referred to as **PMM**.

3. The Random Forest imputation scheme introduced in Algorithm 4 of Chapter 3 and implemented in R under the package `missForest` by Stekhoven and Bühlmann (2012) while repeatedly applying this procedure (**RF MI**).

4. The Random Forest imputation scheme implemented in `mice`. This has been explained in detail in Chapter 3 and was referred to as **RF MICE**.

Following the explanations given in Chapter 3, we applied Rubin's combining rule for multiply imputed data sets, where in our case, the quantity of interest $\boldsymbol{Q}$ is the mean difference, i.e.

$\boldsymbol{Q} = \mu_1 - \mu_2$. We conducted an extensive simulation study, where data has been generated under the following data generating process

$$\mathbf{X}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}_i + \boldsymbol{\mu} \tag{4.2}$$

with various covariance matrices and different distributions for $\boldsymbol{\epsilon}_i = [\epsilon_{1i}, \epsilon_{2i}]^\top$ with different sample sizes $n \in \mathbb{N}$. Missing cases have been artificially inserted representing different forms for the triple (n_1, n_2, n_3) under the MCAR framework while repeating the process over $n_{sim} = 10,000$ times.

Our results indicated a heavy inflation of the type-I-error rate when using **RF MI** as a multiple imputation rule under various simulation settings. Similar, but less extreme were the results under the **RF MICE** imputation scheme. The **PMM** approach revealed a slightly more conservative behavior while **NORM** and the data adjusted method T_{ML} were close to the chosen significance level of $\alpha = 0.05$. Regarding the NRMSE as a performance measure for evaluating imputation schemes, we could see that the latter is not suitable when evaluating them in terms of valid statistical inference results. This, because **RF MI** yielded on average the smallest NRMSE values, but inflated type-I-error heavily. We could verbally identify three sources for the inflation of the type-I-error rate such as the partial violation of the independence assumption when using the paired t-test, the (asymptotic) normality when imputing missing cases with the Random Forest as well as the consistency of the complete-case variance estimator after imputation. In the previous Chapter, Section 3.3 we could especially show that the between-variance estimator based on **RF MI** vanishes almost surely (Theorem 3.2) such that Rubin's condition of properness for the validity of multiple imputation procedures is not met. In addition, this result reveals that uncertainty from the imputation itself has not been covered sufficiently well, such that the true variance under a posited response mechanism might be underestimated yielding to inflated type-I-errors. Furthermore, this shows that **RF MI** behave in its limits (as the number of decision trees tends to infinity) like a single imputation scheme. Note that this theoretical result is not covered within $(P2)$, but was established additionally in Chapter 3. In our article, we extended the simulation scenario to repeated measures ANOVA with four endpoints and received similar results while focusing on Wald-Type quadratic forms as a test statistic for testing no mean time effects. This was also the case for the traditionally used *last observation carried forward* approach, where missing instances are imputed by the last available observational point during the four time points.

In addition to the simulation results, we considered gene expressions from a breast cancer study. Therein, the main idea was to identify gene markers for indicating potential breast cancer tissue. For $n = 112$ breast cancer patients, both, normal and tumor tissue samples have been extracted and potential genes isolated. Similarly to the simulation study, the paired t-test has been used for testing equal mean gene expression. We then artificially introduced missing values under the MCAR mechanism for various missing rates. The results indicated slightly different test-decisions when using multiple imputation procedures such as **RF MI**, **RF MICE**, **PMM** or **NORM**, even when using additional information such as auxiliary variables during imputation.

Finally, we could conclude that modern Machine Learning techniques such as the Random Forest procedure can fail to control type-I-error rates in bivariate data with missing cases in both arms. Therefore, these procedures can be considered as unsuitable in its basic implementation form for inference procedures in multiple imputation settings, at least in bivariate data designs such as the one described in (4.1). However, when aiming to solely conduct

prediction during the analysis phase, as our article ($P1$) could illustrate, Random Forest imputation models can still be used for reducing NRMSE. Whether high predictive accuracy and low NRMSE values are closely related to each other remains future research work we are currently focused on.

4.3 Article 3: *Consistent estimation of residual variance with Random Forest Out-of-Bag errors.*

Following the same notation as in Chapter 2 for Random Forests, we considered arbitrary regression problems of the form

$$Y = m(\mathbf{X}) + \epsilon, \tag{4.3}$$

where the support of $\mathbf{X}$ is assumed to be the unit-cube $[0,1]^p$ and the regression function $m : \mathbb{R}^p \to \mathbb{R}$ is a measurable function such that $\mathbb{E}[|m(\mathbf{X})|^2] < \infty$ while $\mathbb{E}[\epsilon] = 0$ and $Var(\epsilon) = \sigma^2$. Furthermore, it is assumed that ϵ is independent of $\mathbf{X}$. Starting from this data generating process, we assume that $\mathcal{D}_n = \{[\mathbf{X}_i^\top, Y_i]_\cdot^\top i = 1, \ldots, n\}$ is a collection of iid random vectors. The main idea of the article was to show that residuals obtained from the Random Forest method using Out-of-Bag samples can be used to construct consistent estimates for the residual variance σ^2. This will smooth the way towards a solution for the research question given in $(H1)$, which has been analyzed in more detail in Chapter 2, under Theorem 2.2. In a first step, we could show that the finite Random Forest estimate based on Out-of-Bag samples $m_{n,M}^{OOB}$ converges almost surely to its infinite Out-of-Bag analogon $m_n^{OOB} = m_{n,\infty}^{OOB}$ (see Lemma 1 in the main article $(P3)$). Assuming that $\mathbf{X}$ is an independent copy of $\mathbf{X}_1$ such that

$$\lim_{n\to\infty} \mathbb{E}[(m(\mathbf{X}) - m_{n,\infty}(\mathbf{X}))^2] = 0 \tag{4.4}$$

holds together with model assumption (4.3), we could formally show that $\widehat{\sigma}_{RF}^2$ introduced in Chapter 2 is L_1-consistent (see Theorem 1 in (P3)). Condition (4.4) has been proven to be valid for the Random Forest method in Scornet et al. (2015) under side conditions, that are stronger than the one we have set. However, they are not exclusive, i.e. substituting condition (4.4) by the ones given in Scornet et al. (2015) and combining them with our additional assumptions will not lead to any mathematical contradictions such that the theoretical results remain valid. Beside the traditional estimation of the residual variance estimator based on the empirical variance of the sequence $\{Y_i - m_n^{OOB}(\mathbf{X}_i)\}_{i=1}^n$, we considered different estimators for σ^2 as well. Simulation studies in Mendez and Lohr (2011) have shown that the residual variance estimator $\widehat{\sigma}_{RF}^2$ can be positively biased. Therefore, we focused on the Random Forest based bootstrap scheme given in Mendez and Lohr (2011) in order to estimate the bias due to finite sample sizes:

1. Given the training set $\mathcal{D}_n$, generate $\epsilon_1^*, \ldots, \epsilon_n^*$ iid (parametric) bootstrapped residuals such that $\mathbb{E}[\epsilon_1|\mathcal{D}_n] = 0$ and $Var(\epsilon_1^*|\mathcal{D}_n) = \widehat{\sigma}_{RF}^2$. This can be done by the normal distribution, for example, i.e. $\epsilon_1^* \sim N(0, \widehat{\sigma}_{RF}^2)$.

2. Use the bootstrapped residuals $\epsilon_1^*, \ldots, \epsilon_n^*$ in order to compute $Y_1^*, \ldots, Y_n^*$ then given by $Y_i^* = m_n^{OOB}(\mathbf{X}_i) + \epsilon_i^*$.

3. Make use of the tree resp. forest structure of m_n^{OOB} and substitute terminal-node values with the bootstrapped samples $Y_1^*, \ldots, Y_n^*$. Since m_n^{OOB} represents the infinite Random Forest, tree resp. forest structure has to be understood in terms of the representation given in (2.5) of Chapter 2 using weights.

4. Repeat steps $1. - 3.$ B-times in order to obtain a Random Forest based bootstrap sequence $\{m_{n,b}^{OOB}\}_{b=1}^B$ as predictions from 3 using the features $\{\mathbf{X}_{i=1}^n\}$.

5. Estimate the bias then given by $\widehat{R}_B(m_n) := \frac{1}{nB}\sum_{b=1}^{B}\sum_{i=1}^{n}(m_{n,b}^{OOB}(\mathbf{X}_i) - m_n^{OOB}(\mathbf{X}_i))^2$ and set the bias-corrected estimator for the residual variance as $\widehat{\sigma}^2_{RFboot} = \widehat{\sigma}^2_{RF} - \widehat{R}_B(m_n)$.

We could additionally show that under the assumption of (4.4) and the regression model (4.3), together with the condition that $a_n^2/n \longrightarrow 0$ as $n \to \infty$, the Random Forest bootstrap-based residual variance estimator $\widehat{\sigma}^2_{RFboot}$ is L_1-consistent, see Theorem 2 in $(P3)$. Since the latter estimator can result into more computational time loadings, we proposed a slight correction given by $\widehat{\sigma}^2_{RFfast} = \widehat{\sigma}^2_{RF}(1 + 1/a_n^2)$. The latter choice has been motivated by Theorem 3 in $(P3)$, which shows that $\widehat{R}_B(m_n) \geq \widehat{\sigma}^2_{RF}/a_n^2$ almost surely conditioned on $\mathcal{D}_n$, as the number of bootstrap replicates B tends to infinity. Note that there is a typo in equation (12) in the main article $(P3)$ where a square root is missing due to the Cauchy-Schwarz inequality. This does not have any further consequences and was caused by the production team of the journal during publication.

4.3.1 Additional Clarifications

In this article, especially under Theorem 1 and Theorem 2, we have shown L_1 - consistency of the residual variance estimators $\widehat{\sigma}^2_{RF}$ respectively $\widehat{\sigma}^2_{RFboot}$. Note that if $\widehat{\sigma}^2_{RF}$ is consistent, then under the considered framework in article $(P3)$, $\widehat{\sigma}^2_{RFfast}$ will be L_1-consistent as well. In this subsection, we want to clarify the interaction between (asymptotic) unbiasedness and L_1-consistency, which has been used throughout the proofs, perhaps without emphasizing it sufficiently enough due to page limitations of the journal.

Considering the proof of Theorem 1, we decomposed the residual variance estimator $\widehat{\sigma}^2_{RF}$ into

$$\widehat{\sigma}^2_{RF} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\epsilon}_{i,n} - \bar{\epsilon}_{\cdot,n})^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}^2_{i,n} - \bar{\epsilon}^2_{\cdot,n}, \tag{4.5}$$

where $\widehat{\epsilon}_{i,n} = Y_i - m_{n,\infty}^{OOB}(\mathbf{X}_i; \mathcal{D}_n)$ and $\bar{\epsilon}_{\cdot,n} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}_{i,n}$. L_1-consistency of $\widehat{\sigma}^2_{RF}$ can be shown, if $\frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}^2_{i,n}$ and $\bar{\epsilon}^2_{\cdot,n}$ are both L_1-consistent, for example. Note that $\bar{\epsilon}^2_{\cdot,n}$ is non-negative such that (asymptotic) unbiasedness implies L_1-consistency, if the limiting quantity is 0. This has been shown in the second part of Theorem 1 in $(P3)$. Hence, $\bar{\epsilon}^2_{\cdot,n}$ is L_1-consistent. Regarding the first quantity $\frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}^2_{i,n}$, we state the following decomposition (see proof of Theorem 1 in $(P3)$):

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}^2_{i,n} &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - m_{n,\infty}(\mathbf{X}_i; \mathcal{D}_n))^2 = \frac{1}{n}\sum_{i=1}^{n}(\Delta_{n,\infty}^{OOB}(\mathbf{X}_i) + \epsilon_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\Delta_{n,\infty}^{OOB}(\mathbf{X}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot \Delta_{n,\infty}^{OOB}(\mathbf{X}_i) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2,\end{aligned} \tag{4.6}$$

where $\Delta_{n,\infty}^{OOB}(\mathbf{X}_i) := m(\mathbf{X}_i) - m_{n,\infty}^{OOB}(\mathbf{X}_i; \mathcal{D}_n)$. Using the decomposition (4.6), L_1-consistency of $\frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}^2_{i,n}$ is then established by showing that $\frac{1}{n}\sum_{i=1}^{n}\Delta_{n,\infty}^{OOB}(\mathbf{X}_i)^2$ and $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot \Delta_{n,\infty}^{OOB}(\mathbf{X}_i)$ converge in L_1-sense, while $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2$ is an L_1-consistent estimator for σ^2. Regarding the first

part, we can deduce that

$$0 \leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \Delta_{n,\infty}^{OOB}(\mathbf{X}_i)^2\right] = \mathbb{E}[\Delta_{n,\infty}^{OOB}(\mathbf{X}_1)^2] \longrightarrow 0, \text{ as } n \to \infty, \tag{4.7}$$

where the first equality follows from the identical distribution of $\{\Delta_{n,\infty}^{OOB}(\mathbf{X}_i)\}_{i=1}^n$, as argued in the proof of Theorem 1 in $(P3)$ and the convergence from condition (4.4). Regarding the second part of the decomposition (4.6), we can conclude that

$$\begin{aligned} 0 \leq \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot \Delta_{n,\infty}^{OOB}(\mathbf{X}_i)\right|\right] &\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left|\epsilon_i \cdot \Delta_{n,\infty}^{OOB}(\mathbf{X}_i)\right|\right] \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\epsilon_i^2]^{1/2} \cdot \mathbb{E}[\Delta_{n,\infty}^{OOB}(\mathbf{X}_i)^2]^{1/2} \\ &= \sigma \cdot \mathbb{E}[\Delta_{n,\infty}^{OOB}(\mathbf{X}_1)^2]^{1/2} \longrightarrow 0. \end{aligned} \tag{4.8}$$

Similarly to the proof of Theorem 1 in $(P3)$, the second inequality results from applying the triangular inequality, while third equality is the application of the Cauchy-Schwarz inequality. The convergence is a consequence of assumption (4.4). Therefore, result (4.8) implies L_1-consistency which itself implies convergence in probability. Establishing L_1-consistency of $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$ can be conducted by using Vitali's convergence theorem (Rudin, 1987, page 166, Klenke, 2008, page 141). To apply the latter, note that $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$ converges almost surely to $\sigma^2 < \infty$ due to the iid structure of $\{\epsilon_i\}_{i=1}^n$. This implies convergence in probability. Together with

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2\right|\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\epsilon_i^2] = \sigma^2 < \infty \tag{4.9}$$

and

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 \cdot \mathbb{1}\{A\}\right] \leq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\epsilon_i^2] = \sigma^2 < \infty, \tag{4.10}$$

for all $A \in \mathcal{F}$, it follows that $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$ is uniformly integrable. Hence, by Vitali's convergence theorem, we can guarantee that

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 - \sigma^2\right|\right] \longrightarrow 0, \text{as } n \to \infty. \tag{4.11}$$

Note that the first inequality in (4.10) follows from $\epsilon_i^2 \mathbb{1}\{A\} \leq \epsilon_i^2$ almost surely, since $\epsilon_i^2 \geq 0$ almost surely for all $i = 1, \ldots, n$. This will make the L_1-consistency of $\widehat{\sigma}_{RF}^2$ clearer. Establishing L_1-consistency in Theorem 2 in $(P3)$ has been conducted by the usage of Theorem 1 in $(P3)$ among others.

Regarding the proof of Theorem 2 in $(P3)$ on page 55, we concluded that

$$\sum_{j\neq\ell} \mathbb{E}[W_{n,j}^{OOB}(\mathbf{X}_i) W_{n,\ell}^{OOB}(\mathbf{X}_i) \epsilon_j \epsilon_\ell] \leq a_{n-1}^2/(n-1)\mathbb{E}[\epsilon_j \epsilon_\ell], \tag{4.12}$$

where $i \in \{1, \ldots, n\}$ is fixed but arbitrary. The inequality, however, is true, if the residuals are non-negative. In order to show the convergence of (4.12), we need to set up two additional conditions in order for Theorem 2 to hold under the model (4.3). Following the notation introduced in Chapter 2, we note that p_n as the row-dimension of $\boldsymbol{\Theta}_t^{(2)}$, $t \in \{1, \ldots, M\}$ is actually a function of the number of terminal nodes t_n of tree t and it holds $p_n \geq t_n - 1$. We denote the coupling probability of $\mathbf{X}_i$ and $\mathbf{X}_j$ lying in the same cell due to the feature subspacing procedure at node k as $\mathbb{P}_{\boldsymbol{\Theta}_{k,t}^{(2)}}[\mathbf{X}_i \overset{\boldsymbol{\Theta}_{k,t}^{(2)}}{\leftrightarrow} \mathbf{X}_j]$ and set the following two conditions:

(E1) Assume that there exists $\gamma \in [0, 1)$ such that $\mathbb{P}_{\boldsymbol{\Theta}_{k,t}^{(2)}}[\mathbf{X}_i \overset{\boldsymbol{\Theta}_{k,t}^{(2)}}{\leftrightarrow} \mathbf{X}_j] \leq \gamma$.

(E2) The sampling rate a_n and the number of leaves in each tree t_n behaves assymptotically in the following sense: $a_n^2 \gamma^{t_n} \to 0$ as $n \to \infty$.

A necessary condition for $(E1)$ to hold is $m_{try} < p$, because otherwise, the coupling probability will take values in $\{0, 1\}$. A sufficient condition for $(E2)$ to hold is to set the number of leaves in each tree in at least $o(n)$-order. In this case, $a_n^2/t_n = a_n^2/n \cdot n/t_n \longrightarrow 0$, as $n \to \infty$, i.e. t_n grows faster to ∞ as a_n^2 does which yields $(E2)$ due to $\gamma \in [0, 1)$. Under the stated assumptions in Theorem 2 of $(P3)$ together with $(E1)$ and $(E2)$, we can now guarantee its validity. This is due to the following conclusions:

$$
\begin{aligned}
W_{n,j}^{OOB}(\mathbf{X}_i) &\leq \max_{1 \leq i \leq n} \mathbb{P}_{\boldsymbol{\Theta}_t}[\mathbf{X}_i \overset{\boldsymbol{\Theta}_t}{\leftrightarrow} \mathbf{X}_j] = \max_{1 \leq i \leq n} \mathbb{P}_{\boldsymbol{\Theta}_t^{(1)}}[\mathbf{X}_i \overset{\boldsymbol{\Theta}_t^{(1)}}{\leftrightarrow} \mathbf{X}_j] \cdot \mathbb{P}_{\boldsymbol{\Theta}_t^{(2)}}[\mathbf{X}_i \overset{\boldsymbol{\Theta}_t^{(2)}}{\leftrightarrow} \mathbf{X}_j] \\
&\leq \frac{a_{n-1}}{n-1} \cdot \max_{1 \leq i \leq n} \mathbb{P}_{\boldsymbol{\Theta}^{(2)}}[\mathbf{X}_i \overset{\boldsymbol{\Theta}^{(2)}}{\leftrightarrow} \mathbf{X}_j] = \frac{a_{n-1}}{n-1} \max_{1 \leq i \leq n} \prod_{k=1}^{p_n} \mathbb{P}_{\boldsymbol{\Theta}_{k,t}^{(2)}}[\mathbf{X}_i \overset{\boldsymbol{\Theta}_{k,t}^{(2)}}{\leftrightarrow} \mathbf{X}_j] \\
&\leq \frac{a_{n-1}}{n-1} \prod_{k=1}^{p_n} \gamma \leq \frac{a_{n-1}}{n-1} \gamma^{t_n - 1}. \qquad (4.13)
\end{aligned}
$$

The first equality follows from the independence of $\boldsymbol{\Theta}_t^{(1)}$ and $\boldsymbol{\Theta}_t^{(2)}$, while the second inequality from (11) on page 55 of $(P3)$. The second equality follows from the fact that $\{\boldsymbol{\Theta}_{k,t}^{(2)}\}_{k=2}^{p_n}$ is a sequence of iid random vectors, whereas the last inequality follows from $p_n \geq t_n - 1$. Using the Cauchy-Schwarz inequality together with the result in (4.13), we can now obtain

$$
\begin{aligned}
\sum_{j \neq \ell} \mathbb{E}[W_{n,j}^{OOB}(\mathbf{X}_i) W_{n,\ell}^{OOB}(\mathbf{X}_i) |\epsilon_j \epsilon_\ell|] &\leq \sum_{j \neq \ell} \mathbb{E}[\epsilon_j^2 \epsilon_\ell^2]^{1/2} \mathbb{E}[W_{n,j}^{OOB}(\mathbf{X}_i)^2 W_{n,\ell}^{OOB}(\mathbf{X}_i)^2]^{1/2} \\
&= \sigma^2 \sum_{j \neq \ell} \mathbb{E}[W_{n,j}^{OOB}(\mathbf{X}_i)^2 W_{n,\ell}^{OOB}(\mathbf{X}_i)^2]^{1/2} \\
&\leq \sigma^2 \cdot n(n-1) \frac{a_{n-1}^2}{(n-1)^2} \gamma^{2(t_n - 1)} \\
&\leq \sigma^2 \frac{n(n-1)}{(n-1)^2} a_n^2 \gamma^{t_n - 1} \longrightarrow 0, \quad \text{as } n \to \infty. \qquad (4.14)
\end{aligned}
$$

The first equality follows from the independence of ϵ_j and ϵ_ℓ, since $j \neq \ell$, whereas the second inequality follows from (4.13). The convergence results obviously from $(E2)$.

4.4 Article 4: *Asymptotic Unbiasedness of the Permutation Importance Measure in Random Forest Models.*

In our last work, we paid special attention to the feature selection process of the Random Forest method as described in Algorithm 3 of Chapter 2 while aiming to solve research question $(H2)$. We assume a regression model of the form

$$Y = \widetilde{m}(\mathbf{X}) + \epsilon, \tag{4.15}$$

where $\mathbf{X}$ is again assumed to have $[0,1]^p$ as support and is independent of ϵ such that $\mathbb{E}[\epsilon] = 0$ and $Var(\epsilon) = \sigma^2 \in (0, \infty)$. Regarding the measurable regression function $\widetilde{m} : \mathbb{R}^p \longrightarrow \mathbb{R}$, we assumed that there exists an informative subset $\mathcal{S} \subseteq \{1, \ldots, p\}$ such that (4.15) can be further simplified into $m(\mathbf{X}_{\mathcal{S}})$, where $\mathbf{X}_{\mathcal{S}} \in \mathbb{R}^s$ with $s = |\mathcal{S}|$ is the reduced vector consisting of only those covariates, that are relevant according to the composition of the index set $\mathcal{S}$. Extracting relevant features according to the set $\mathcal{S}$ is in practice a challenging tasks. Within the Random Forest method, however, there exist several approaches how relevant features can be extracted such as the permutation measure of the Random Forest method, even for $p > n$ problems (see e.g. Jiang et al., 2004,Díaz-Uriarte and De Andres, 2006,Menze et al., 2009 or Qi, 2012 pages 307-323). The idea is to randomly permute variables in a specific feature j for observations not being considered during the training phase of the Random Forest method, the so called Out-of-Bag samples. Then the decrease in empirical mean squared error is computed and used for assessing the importance of a particular feature. Denoting with $\mathcal{D}_n^{-(t)}$ the Out-of-Bag set for decision tree $t \in \{1, \ldots, M\}$, while $\pi_{j,t}$ is the permutation for variable $j \in \{1, \ldots, p\}$ in decision tree t, we can state the Random Forest permutation importance measure (RFPIM) based on Out-of-Bag samples for feature $j \in \{1, \ldots, p\}$ by

$$I_{n,M}^{OOB}(j) = \frac{1}{M\gamma_n} \sum_{t=1}^{M} \sum_{i \in \mathcal{D}_n^{-(t)}} \left\{ (Y_i - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}; \mathbf{\Theta}_t, \mathcal{D}_n))^2 - (Y_i - m_{n,1}(\mathbf{X}_i; \mathbf{\Theta}_t, \mathcal{D}_n))^2 \right\}. \tag{4.16}$$

We could show that under model (4.15) together with assumptions $(A1)$, $(A3)$ and $(A4)$ there exists a limiting quantity $I(j)$ for $j \in \{1, \ldots, p\}$ that enables the discrimination between variables in $\mathcal{S}$ and variables not in $\mathcal{S}$ (see Proposition 2 in $(P4)$). This is given by

$$I(j) = \begin{cases} \mathbb{E}[(m(\mathbf{X}_1) - m(\mathbf{X}_{j,1}))^2], \text{ if } j \in \mathcal{S}, \\ 0, \text{ else } . \end{cases} \tag{4.17}$$

Finally, we then could show that for uninformative variables, i.e. $j \in \mathcal{S}^C$, the RFPIM reveals unbiased results while for informative ones, the RFPIM showed an asymptotic unbiasedness. That is,

$$\mathbb{E}[I_{n,M}^{OOB}(j)] = I(j) = 0, \text{ if } j \in \mathcal{S}^C, \text{ while } \lim_{M \to \infty} \mathbb{E}[I_{n,M}^{OOB}(j)] \longrightarrow I(j) \text{ for } j \in \mathcal{S} \text{ as } n \to \infty$$

has been proven under the assumptions (A1) - (A5) in Theorem 1 in the main article $(P4)$. Especially assumption (A2) reveals some novel insights into the operating mode of the RFPIM. Instead of allowing all kind of permutations, we thereby had to restrict the permutations to a sub-class, that excludes the possibility that $\pi_{j,t} = id$ and $\pi_{j,t}(i) = i$ for $i \in \mathcal{D}_n^{-(t)}$. Without the latter assumptions, it is unclear whether Theorem 1 in $(P4)$ holds or not. Therefore, simply permuting features might not be suitable for assessing variable importance in Random Forest.

In addition, we could identify potential sources that might distort the RFPIM, especially for finite choices of M and n. Beside the large number of decision trees needed and the large samples for obtaining optimal results in terms of Theorem 1, noisy data can affect the convergence rate of $I_{n,M}^{OOB}(j)$ especially for variables $j \in \mathcal{S}$. Hence, we focused on the signal-to-noise ratio given by

$$SN = \frac{Var(\widetilde{m}(\mathbf{X}))}{\sigma^2} \tag{4.18}$$

and could show that if the Random Forest cuts are cut-consistent, than the effect of the noise represented by its variance σ^2 vanishes from the cut-criterion $L_n^{(k)}(j,z)$ given in Algorithm 3. If the systematic signal coming from the regression function $\widetilde{m}$ is strong enough, i.e. larger than the noise, than the signal-to-noise ratio will be correspondingly larger than 1. This effect, together with the results given in $(P3)$ leads to the consistent estimation of SN using various residual variance estimators proposed in this thesis.

In order to support our theoretical findings, we have conducted an extensive simulation study under scenarios where $p < n$ and $p > n$. Considering the same regression functions as in Chapter 2, Section 2.1 under a sparse regression framework, the simulation results supported our findings for all four considered regression functions: the linear case, the polynomial case, the trigonometric case and the non-continuous case. Especially for non-informative features, the RFPIM was on average vanishing, while for informative ones, i.e. $j \in \mathcal{S}$, RFPIM slowly seemed to converge to $I(j)$. In order to emphasize the effect of the signal-to-noise ratio on the RFPIM, we have chosen simulation settings such that $SN \in \{0.5, 1, 3, 5\}$. Larger values of SN indicated a clearer distinction between variables in $\mathcal{S}$ and $\mathcal{S}^C$ on average.

Although it has been reported that the Random Forest for classification purposes reveals importance measures that are biased for categorical variables with larger domain cardinalities, we could theoretically close this gap for regression problems under the assumptions $(A1)$ - $(A5)$. Hence, Random Forest regression reveals an unbiased importance measure and this can be extended to $p > n$ problem types unless these assumptions are not violated. We plan to extend our work by allowing correlation among the features, which can not be covered from our theoretical results yet.

Chapter 5

Conclusion and Outlook

In the first part of this thesis, we considered the application of Random Forest models and Boosting methods as imputation models for partially observed data. In the second part, we focused on some crucial properties such as uncertainty quantification and feature selection in sparse regression problems using Random Forest. In a first simulation study, we could show that the traditionally used Random Forest with sub-sampling or resampling can be enhanced as an imputation method when switching to a kernel-based bootstrapping procedure. This resulted in better imputation performance for continuous outcomes using the normalized root mean squared error as a performance measure. In case of mixed-type data, our results revealed that the inclusion of the stochastic gradient boosting method enhanced the imputation procedure leading to the proposal of a mixed model consisting of the Random Forest and the stochastic gradient boosting (see ($P1$)). In terms of statistically valid inference procedure, we could find out that modern Machine Learning tools such as the Random Forest as an imputation scheme led to an inflation of type-I-errors in subsequent mean comparisons in repeated measure designs. The results suggest that inflation might happen for repeated measures designs in general, when testing mean time effects (see ($P2$)). We extended the findings in article ($P2$) by proving that the **RF MI** as an imputation schemes leads to almost surely vanishing between variance estimators according to Rubin's combining rule, see e.g. Chapter 3, Section 3.3. Hence, the method does not correctly reflect uncertainty originating form the fact that missing cases are present. Furthermore, this revealed that **RF MI** is not *proper* therefore not meeting the assumptions for an inferentially valid imputation method, at least for monotone and univariate missing patterns. We are currently preparting another paper on this result.

Regarding uncertainty quantification, which plays a crucial role in statistical inference procedures for delivering variance estimators in central-limit-type theorems, we could show that the L_2-consistent Random Forest can lead to consistent residual variance estimators for regression problems. The latter requirement of L_2-consistency according to Definition 1.2 has been tackled in Biau et al. (2008) and Scornet et al. (2015) already, where we extended that result by setting this as a theoretically verifiable assumption. In ($P3$), we proposed different estimators and proved its consistency when using Random Forests. We extended the estimators proposed in ($P3$) in this thesis correcting them for potentially finite-M-bias as a source of distortion, see e.g. Chapter 2, Section 2.2. The latter originates from a finite choice of decision trees in the Random Forest ensemble and we also plan to publish this result together with extensive simulations.

Random Forest models are not only used as a prediction tool, but as a variable selection tool

in potentially high-dimensional regression and classification problems. Recent scientific work has been focused on addressing potential failures of Random Forest internal variable selection measures such as the permutation measure. In $(P4)$, we could show that the RFPIM can lead to correct variable screening procedures under specific assumptions for regression-type problems. They reveal that simply permuting features among the Out-of-Bag set might not guarantee the delivery of (asymptotically) correct variable selection measures. Instead, one has to consider certain sub-classes of potential permutations.

The last theoretical result in $(P4)$ enables the consideration of central-limit-type theorems for the RFPIM. We conjecture that it is possible to show that there exists a non-negative sequence $a_n \nearrow \infty$ such that either of the following holds

$$\sqrt{a_n}(I^{OOB}_{n,M}(j) - I(j)) \xrightarrow{d} Z, \text{ as } n \to \infty, \quad j \in \{1, \dots, p\} \tag{5.1}$$

$$\sqrt{a_n}(I^{OOB}_{n,\infty}(j) - I(j)) \xrightarrow{d} Z, \text{ as } n \to \infty, \quad j \in \{1, \dots, p\} \tag{5.2}$$

where Z follows a known distribution, for example the normal distribution. Then one would be equipped to conduct hypothesis tests of the form

$$H_0 : I(j) = 0 \quad \text{vs.} \quad H_1 : I(j) \neq 0 \tag{5.3}$$

using the Random Forest method. This could also be conducted for cases where $p > n$ unless the assumptions in $(P4)$ are not violated. Note that the null-hypothesis that variables are unimportant according to the definition given in $(P4)$ implies the the null-hypothesis H_0 in (5.3). Therefore, rejecting H_0 leads to the rejection of the initial hypothesis that variables are unimportant. In cases where $Z \sim N(0, \sigma^2)$, we then need to find suitable estimators for σ^2, that may be motivated by the work in $(P3)$ and Chapter 2. Thus, based on the results (5.1) or (5.2) one can construct an asymptotically valid test for testing H_0 in (5.3) with modern Machine Learning tools such as the Random Forest. The idea can be considered as a future research work, for which the author is currently preparing a grant proposal that will indeed extend the results in this thesis and make Machine Learning tools more interpretable in terms of Figure 1.

Regarding the results that Random Forest methods in its traditional form might be unsuitable as an imputation scheme in partially observed data led us to the question in which cases the latter should still be used as an imputation scheme. Therefore, we are currently focusing on prediction schemes with missing covariates and analyze prediction accuracy of various bagging and boosting methods under partially observed covariates in another research project.

Appendix A

Original Articles and their Supplementary Materials.

In the sequel, you can partly find the original articles, on which this thesis has been build upon. Due to publishing rights, only the DOI of the articles ($P1$), ($P2$) and ($P3$) are listed here. Accessing those articles together with their supplementary material can be done using the DOI or directly contacting the corresponding journal through their website. Since article ($P4$) is in the process of being published, an un-modified version has been attached to this thesis on the next page.

- Article ($P1$) together with its supplementary material can be found under this DOI:

  `https://doi.org/10.1007/s00180-019-00900-3`

- Article ($P2$) together with its supplementary material can be found under this DOI:

  `https://doi.org/10.1093/bioinformatics/btaa082`

- Article ($P3$) can be found under this DOI:

  `https://doi.org/10.1016/j.spl.2019.03.017`

Asymptotic Unbiasedness of the Permutation Importance Measure in Random Forest Models.

Burim Ramosaj*, Markus Pauly

Faculty of Statistics
Institute of Mathematical Statistics and Applications in Industry
Technical University of Dortmund
44227 Dortmund, Germany

Abstract

Variable selection in sparse regression models is an important task as applications ranging from biomedical research to econometrics have shown. Especially for higher dimensional regression problems, for which the link function between response and covariates cannot be directly detected, the selection of informative variables is challenging. Under these circumstances, the Random Forest method is a helpful tool to predict new outcomes while delivering measures for variable selection. One common approach is the usage of the permutation importance. Due to its intuitive idea and flexible usage, it is important to explore circumstances, for which the permutation importance based on Random Forest correctly indicates informative covariates. Regarding the latter, we deliver theoretical guarantees for the validity of the permutation importance measure under specific assumptions and prove its (asymptotic) unbiasedness. An extensive simulation study verifies our findings.

Keywords: Random Forest, Unbiasedness, Permutation Importance, Out-of-Bag Samples, Statistical Learning

1. Introduction

Random Forest is a non-parametric classification and regression algorithm being known for its good predictive performance and simple applicability under various settings. The method is based on constructing each tree in the forest by bagging procedures, which enables the construction of several estimators based on *Out-of-Bag* principles, such as prediction points or variance estimates. Main advantages of the Random Forest method compared to other Machine Learning tools is its relative ease in hyper-parameter tuning while delivering internal estimates of the mean squared error. Due to its complicated mathematical description, including data-dependent weighting, theoretical results such as consistency or central limit theorems have only been derived recently, see e.g. [1, 2, 3, 4].

Beyond its usage for prediction, Random Forest models can also be used as a tool for variable selection. Especially in high dimensional learning problems, where the number of variables exceeds the number of ob-

* *Corresponding Author:* Burim Ramosaj
Email address: burim.ramosaj@tu-dortmund.de

servations, the extraction of an informative feature subset is beneficial from three perspectives: Firstly, a reduced and simplified model is more accessible and interpretable than models of higher dimensions leading to faster and easier data collection processes. Secondly, model accuracy can sometimes even be enhanced under lower dimensional models bypassing the possibility of overfitting. Thirdly, a reduced model makes well-known statistical inference procedures applicable. As mentioned in [5], the Random Forest model can be considered as an embedded model, where variable selection is an integral part of the tree construction. In selecting variables, the Random Forest method delivers two measures: The permutation importance as well as the mean decrease impurity. For classification, the mean decrease impurity summarizes the decrease of the Gini impurity after conducting a cut over the whole tree structure and averages the result over all trees. For regression problems, this measure turns into the summation of the decrease in variance after conducting a cut at every node of the tree, averaged over the forest. The principle of the permutation importance is slightly different: In order to mimic the effect of a variable on the response, its values from the set of Out-of-Bag samples are randomly permuted and the decrease in model accuracy averaged over all trees is measured.

Although simple to apply and intuitive, both measures have been criticized. In [6], for example, the authors could illustrate that the Gini importance for classification problems tends to prefer variables with larger numbers of categories and scale measurements. Furthermore, different results could be obtained when switching the sampling procedure in the bagging step to sampling with replacement instead of without replacement. In [7], additional criticism was addressed towards the permutation importance, arguing that the permutation of the corresponding feature does not only break the relation with the response variable, but also with other potentially correlated covariates. This effect of correlated features has since been part of several research [6, 8, 9, 10, 11, 12, 13]. Nevertheless, several authors such as [6, 11] claimed that the permutation importance led to more accurate results than the importance measure based on decrease in node impurities. However, theoretical guarantees for the validity of the traditional Random Forest method regarding its importance measures are rather sparse. An exception is given in [14], where a theoretical approach has been conducted within the framework of correlated features in additive regression models. Therein, the authors showed different identities of a formalized version of the Random Forest permutation importance measure (RFPIM).

The contributions of this paper are twofold: First, we aim to clarify the criticism on the RFPIM from a theoretical perspective. Therefore, we state assumptions, for which the permutation importance measure does correctly select informative features and prove its (asymptotic) unbiasedness. This way, we also close the gap between the formalized version of the permutation measure as considered in [14] and the empirical permutation measure computed in a Random Forest model. Secondly, we identify main drivers for the quality of the RFPIM and support our findings by an extensive simulation study covering high-dimensional settings, too.

2. Model Framework and Random Forest

Our framework covers regression models, for which the covariable space is assumed to lie on the p-dimensional unit space, i.e. $\mathbf{X} \in [0,1]^p$. In fact, this assumption does not have sever generalization effects, since Random Forest models are invariant under (strictly) monotone transformations. For discrete distributions of $\mathbf{X}$, one could alternatively assume a finite support, such that a $[0,1]$-standardization exists for every feature $j = 1, \ldots, p$. Furthermore, we will assume that the relationship between the response variable Y and the covariates $\mathbf{X}$ can be modeled through

$$Y = \widetilde{m}(\mathbf{X}) + \epsilon, \tag{1}$$

where $\tilde{m} : [0,1]^p \longrightarrow \mathbb{R}$ is a measurable function and $\mathbf{X}$ is independent of ϵ with $\mathbb{E}[\epsilon] = 0$, $Var(\epsilon) \equiv \sigma^2 \in (0, \infty)$. For sparse learning problems, not all of the given covariates are necessary, that is, there is a subset $\mathcal{S}$ with cardinality s less than p that covers all the information about Y. Assuming without loss of generality that $\mathcal{S} = \{1, \ldots, s\}$, the regression model (1) can then be reduced to

$$Y = m(\mathbf{X}_{\mathcal{S}}) + \epsilon, \tag{2}$$

where $\mathbf{X}_{\mathcal{S}} = [X_1, \ldots, X_s]$ and $m : [0,1]^s \longrightarrow \mathbb{R}$ is another measurable function such that $\widetilde{m}(\mathbf{X}) = m(\mathbf{X}_{\mathcal{S}})$. The specification of $\mathcal{S}$, or also known as *variable selection, feature selection or subset selection*, can be challenging, especially when the relationship is not linear or not deducible at all. Formally speaking, we refer to a variable $j \in \{1, \ldots, p\}$ as *informative* or *important*, if the corresponding regression model given in (1) can be reduced to a regression model of the form (2). This leads to the independence of Y towards X_j given all other covariates for features $j \notin \mathcal{S}$. That is $Y \perp\!\!\!\perp X_j | X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$. For differentiable link-functions $\widetilde{m}$, one can alternatively define a variable as *unimportant* or *uninformative*, if for $\boldsymbol{h}_j = [0, \ldots, 0, h, 0, \ldots, 0]^\top \in \mathbb{R}^p$, with $h \in \mathbb{R}$ lying at the j-th position, it holds

$$\frac{\partial \widetilde{m}(\mathbf{x})}{\partial x_j} := \lim_{\|\boldsymbol{h}_j\| \to 0} \frac{\widetilde{m}(\mathbf{x} + \boldsymbol{h}_j) - \widetilde{m}(\mathbf{x})}{\|\boldsymbol{h}_j\|} = 0. \tag{3}$$

Then a feature $j \in \{1, \ldots, p\}$ is said to be *informative* or *important*, if it is not *uninformative* or *unimportant*. Under the scenario of a differentiable link function $\widetilde{m}$, both definitions given in (2) and (3) for an *informative* or *important* variable can be shown to be equivalent using a Taylor expansion of $\widetilde{m}$.

Although there are several approaches in extracting informative features, difficulties exist if the underlying link function is of complex analytical structure. The Random Forest method enables the extraction of informative features during the training phase of the algorithm. To accept this, let us shortly recall the Random Forest. Given a training set

$$\mathcal{D}_n = \{[\mathbf{X}_i^\top, Y_i]^\top \in [0,1]^p \times \mathbb{R} : i = 1, \ldots, n\}, \tag{4}$$

of iid pairs $[\mathbf{X}_i^\top, Y_i]^\top$, $i = 1, \ldots, n$, the Random Forest method estimates the functional relationship of $\widetilde{m}$ by piecewise constant functions over random partitions of the feature space. To be more precise, the Random Forest model for regression is a collection of $M \in \mathbb{N}$ decision trees, where for each tree, a bootstrap

sample is taken from $\mathcal{D}_n$ using with or without replacement procedures. This is denoted as the resampling strategy $\mathcal{P}$. Furthermore, at each node of the tree, feature sub-spacing is conducted selecting $v_{try} \in \{1, \ldots, p\}$ features for possible split direction. Denote with $\boldsymbol{\Theta}$ the generic random variable responsible for both, the bootstrap sample construction and the feature sub-spacing procedure. Then, $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M$ are assumed to be independent copies of $\boldsymbol{\Theta}$ responsible for this random process in the corresponding tree, independent of $\mathcal{D}_n$. The combination of the trees is then conducted through averaging. i.e.

$$m_{n,M}(\mathbf{x}; \boldsymbol{\Theta}_1, \ldots \boldsymbol{\Theta}_M, \mathcal{D}_n) = \frac{1}{M} \sum_{j=1}^{M} m_{n,1}(\mathbf{x}; \boldsymbol{\Theta}_j, \mathcal{D}_n) \tag{5}$$

and is referred to as the finite forest estimate of $\widetilde{m}$, where $\mathbf{x} \in [0,1]^p$ is a fixed point. Here, $m_{n,1}(\cdot; \boldsymbol{\Theta}_j, \mathcal{D}_n)$ refers to a single tree in the Random Forest build with $\boldsymbol{\Theta}_j$, $j = 1, \ldots, M$. As explained in [3], the strong law of large numbers (for $M \to \infty$) allows to study $\mathbb{E}_{\Theta}[m_n(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n)]$ instead of (5). Hence, we set

$$m_n(\mathbf{x}) = m_n(\mathbf{x}; \mathcal{D}_n) = \mathbb{E}_{\boldsymbol{\Theta}}[m_n(\mathbf{x}; \boldsymbol{\Theta}, \mathcal{D}_n)], \tag{6}$$

where $\mathbb{E}_{\Theta}$ denotes the expectation over $\boldsymbol{\Theta}$ given the training set $\mathcal{D}_n$, i.e. $\mathbb{E}_{\boldsymbol{\Theta}}[m_n(\mathbf{x}; \boldsymbol{\Theta}_j, \mathcal{D}_n)] = \mathbb{E}[m_n(\mathbf{x}; \Theta, \mathcal{D}_n)|\mathcal{D}_n]$.
Similar to [3], we refer to the Random Forest algorithm by identfiying three parameters responsible for the Random Forest tree construction:

- $v_{try} \in \{1, \ldots, p\}$, the number of pre-selected directions for splitting,
- $a_n \in \{1, \ldots, n\}$, the number of sampled points in the bootstrap step and
- $t_n \in \{1, \ldots, a_n\}$, the number of leaves in each tree.

A detailed algorithm is given on page 1720 in [3], for example.
An advantage of the Random Forest method is the delivery of internal measures such as predictions or prediction accuracy without initially separating the training set $\mathcal{D}_n$ such as in cross-validation procedures. This is possible by making use of the bagging principle and Out-Of-Bag (OOB) samples. The latter extracts all random trees that have not used a fixed observation $\mathbf{X}_i$ in the set $\mathcal{D}_n$ during training and averages the prediction results over all those trees. In the sequel, we will denote with $m_{n,M}^{OOB}(\mathbf{X}_i)$ the OOB prediction of $\mathbf{X}_i \in \mathcal{D}_n$ using the finite forest estimate and $m_n^{OOB}(\mathbf{X}_i) = \mathbb{E}_{\boldsymbol{\Theta}_{[i]}}[m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_{[i]}, \mathcal{D}_n)]$ the corresponding infinite forest OOB prediction, where $\boldsymbol{\Theta}_{[i]}$ is the generic random vector, which has not selected observation $i \in \{1, \ldots, n\}$. Note that the authors in [15] could show that even for the OOB finite forest prediction, it holds $\mathbb{P}_{\Theta}$ - almost surely that

$$m_{n,M}^{OOB}(\mathbf{X}_i) \longrightarrow m_n^{OOB}(\mathbf{X}_i), \quad \text{as } M \to \infty.$$

In the sequel, it is required to have a look at a certain averaging step in the random tree ensemble of the Random Forest and its asymptotic behavior in case of $M \to \infty$. For later use, we state this as a Proposition.

Proposition 1. *Assume regression model* (1) *and fix* $i \in \{1, \dots, n\}$. *Then it holds* $\mathbb{P}_\Theta$ *- almost-surely that*

$$\frac{1}{M}\sum_{t=1}^{M} m_{n,1}(\mathbf{X}_i; \mathbf{\Theta}_t, \mathcal{D}_n) \cdot \mathbb{1}\{\mathbf{X}_i \textit{ has not been selected}\} \longrightarrow c_n \cdot m_n^{OOB}(\mathbf{X}_i), \quad \textit{as } M \to \infty,$$

where

$$c_n = \begin{cases} 1 - a_n/n, & \textit{if observations are subsampled (draws without replacement),} \\ (1 - 1/n)^n, & \textit{if observations are bootstrapped with replacement.} \end{cases}$$

3. Permutation Importance of the Random Forest

Returning to the extraction of relevant features, the Random Forest permutation importance makes use of the Out-of-Bag principle. That is, for every tree constructed in the forest, the increase of mean squared error evaluated on the corresponding Out-of-Bag sample after permuting its observations along the j-th variable is measured, with $j \in \{1, \dots, p\}$. Hence, the measure clearly depends on the sampling strategy $\mathcal{P}$ chosen prior to tree construction. This could be seen in [6] for example, where different results were obtained depending on the sampling strategy given in $\mathcal{P}$. Formally speaking, the permutation importance can be defined as

$$I_{n,M}^{OOB}(j) := \frac{1}{M\gamma_n}\sum_{t=1}^{M}\sum_{i \in \mathcal{D}_n^{-(t)}} \left\{(Y_i - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}; \mathbf{\Theta}_t))^2 - (Y_i - m_{n,1}(\mathbf{X}_i; \mathbf{\Theta}_t))^2\right\} \tag{7}$$

for all $j \in \{1, \dots, p\}$, where $\mathcal{D}_n^{-(t)} = \mathcal{D}_n^{-(t)}(\mathbf{\Theta}_t)$ is the Out-of-Bag sample for the t-th tree, i.e. the set of observations not selected for training $m_{n,1}(\cdot; \mathbf{\Theta}_t, \mathcal{D}_n)$. The cardinality γ_n of $\mathcal{D}_n^{-(t)}$ clearly depends on the sampling strategy $\mathcal{P}$. Moreover, $\pi_{j,t}$ is the non-trivial permutation of observations in $\mathcal{D}_n^{-(t)}$ along the j-th variable in decision tree $t \in \{1, \dots, M\}$. In [16] and [14], a *theoretical version* of $I_{n,M}^{OOB}(j)$, $j \in \{1, \dots, p\}$ was given by

$$\begin{aligned} I(j) &:= \mathbb{E}[(Y_1 - \widetilde{m}(\mathbf{X}_{j,1}))^2] - \mathbb{E}[(Y_1 - \widetilde{m}(\mathbf{X}_1))^2] \\ &= \mathbb{E}[(Y_1 - \widetilde{m}(\mathbf{X}_{j,1}))^2] - \sigma^2, \end{aligned} \tag{8}$$

where $\mathbf{X}_{j,1} = [X_{1,1}, \dots, X_{j-1,1}, Z_j, X_{j+1,1}, \dots, X_{p,1}]^\top$ and Z_j is an independent copy of $X_{j,1}$, independent of Y_1. The intuition behind the definition in (8) is that $I(j)$, $j = 1, \dots, p$ measures the increase in variation after eliminating potential dependencies between the j-th variable and the response.
Assuming an additive regression model, i.e. $\widetilde{m}(\mathbf{x}) = \sum_{j=1}^{p} \widetilde{m}_j(x_j)$, [14] proved that

$$I(j) = \begin{cases} 2 \cdot Cov(Y, \widetilde{m}_j(X_j)) - \sum_{k \neq j} Cov(\widetilde{m}_j(X_j), \widetilde{m}_k(X_k)) & \text{if } \mathbb{E}[\widetilde{m}_j(X_j)] = 0, \\ 2 \cdot Var(\widetilde{m}_j(X_j)), & \text{else,} \end{cases} \tag{9}$$

for $j \in \{1, \dots, p\}$, where $I(j)$ can be further simplified in case of a multivariate normal distribution for $[\mathbf{X}^\top, Y]^\top \in \mathbb{R}^{p+1}$, see e.g. Proposition 2 in [14]. So far, however, it is completely unclear in which sense the quantities $\boldsymbol{I}_{n,M}^{OOB} = [I_{n,M}^{OOB}(1), \dots, I_{n,M}^{OOB}(p)]^\top$ and $\boldsymbol{I} = [I(1), \dots, I(p)]^\top$ relate to each other. This is of

important interest, since $\boldsymbol{I}$ can, e.g., be considered as a key quantity for future significance tests during feature extraction. Below, we will study their relation in detail under a more general set-up not requiring the additivity of the link function $\widetilde{m}$. Instead, we set up some more general assumptions, under which we can guarantee asymptotically, that $\boldsymbol{I}_{n,M}$ is an unbiased estimator of $\boldsymbol{I}$. This will open new paths for feature selection tests using Random Forest.

Assumptions.

(A1) There is at least one informative variable, i.e. $|\mathcal{S}| \geq 1$,

(A2) Permutations are restricted to the class $\mathcal{V} = \{\pi \in \mathcal{S}_{\gamma_n} : \pi(i) \neq i\}$, *where* $\mathcal{S}_{\gamma_n}$ *is the symmetric group,*

(A3) The features $\mathbf{X} = [X - 1, \ldots, X_p]^\top$ *are mutual independent.*

(A4) $\sup_{\mathbf{x}} |\widetilde{m}(\mathbf{x})| < \infty$,

(A5) Infinite Random Forests are L_2*-consistent, i.e.* $\lim_{n\to\infty} \mathbb{E}[(\widetilde{m}(\mathbf{X}) - m_n(\mathbf{X}))^2] = 0$, *where* $\mathbf{X}$ *is an independent copy of* $\mathbf{X}_1$.

Condition $(A1)$ ensures that the random forest is not forced to select among non-informative variable. This can happen if $|\mathcal{S}| = 0$, since the tree construction process will continue until either a pre-defined number of leaves t_n is reached or each leave in a tree consists of at most a pre-specified number of observations. Condition (A2) is important from a technical perspective, in order to achieve (asymptotic) unbiasedness. Furthermore, this condition reveals some drawbacks of the traditional permutation approaches: considering arbitrary permutations $\pi \in \mathcal{S}_{\gamma_n}$, we cannot guarantee the (asymptotic) unbiasedness of the RFPIM. Hence, one should carefully consider implementations of RFPIM in statistical software packages such as R or python with regard to this assumption. Condition $(A3)$ is essential in this context. The permutation used in $\boldsymbol{I}^{OOB}_{n,M}$ aims to break the relationship between the response variable and the corresponding covariate. In case of dependency structures among the other covariables, however, this dependency is then also broken clouding the primary effect of dependencies between the response and the covariable of interest. Note that assumption $(A3)$ implies the assumption of no multicolinearity. Condition $(A4)$ is rather technical. Instead, one could replace it with $\widetilde{m}$ being continuous, since the domain of $\mathbf{X}$ is the p-dimensional unit cube $[0, 1]^p$. An important assumption is $(A5)$, which was formally proven to be valid for Random Forest models in [3]. There, the authors proved the L_2 - consistency of the same Random Forest method as considered in our work. Note that their assumptions for the validity of $(A5)$ do not exclude $(A3)$ and $(A4)$. Instead, one could completely overtake the assumptions given in Theorem 1 or Theorem 2 listed in [3] and replace them with $(A3)$ - $(A5)$. Assumptions $(A1)$ and $(A2)$ have then to be considered as additional assumptions in this context. A formal proof of this is given in the Appendix. However, for generality and as we also state non-asymptotic results, we decided to work with ours.

Our first result shows an alternative expression of the quantity $\boldsymbol{I}$ defined in (8), which makes variable selection possible for the Random Forest permutation importance.

Proposition 2. *Assume the regression model* (1) *and conditions* $(A1)$, $(A3)$ *and* $(A4)$. *Then for every variable* $j \in \{1, \ldots, p\}$ *it holds*

$$I(j) = \begin{cases} \mathbb{E}[(\widetilde{m}(\mathbf{X}_1) - \widetilde{m}(\mathbf{X}_{j,1}))^2], & \text{if } j \in \mathcal{S}, \\ 0, & \text{else .} \end{cases}$$

This property allows us to define the permutation importance as unbiased or asymptotically unbiased, if $\mathbb{E}[\boldsymbol{I}_{n,M}^{OOB}] = \boldsymbol{I}$ resp. $\lim_{M\to\infty} \mathbb{E}[\boldsymbol{I}_{n,M}^{OOB}] \longrightarrow \boldsymbol{I}$, as $n \to \infty$. Proposition 2 can be considered as an extension of the results given in equation (9), since the assumption for the link-function being additive is dropped. Anyhow, the above considerations finally lead to the main result of the current paper: the (asymptotic) unbiasedness of RFPIM.

Theorem 1. *Under model (1) and conditions* $(A1)$ - $(A5)$ *while sampling is restricted to sampling without replacement, the RFPIM is unbiased for* $j \in \mathcal{S}^C = \{1, \ldots, p\} \setminus \mathcal{S}$ *and asymptotically unbiased for* $j \in \mathcal{S}$. *That is for* $j \in \mathcal{S}^C$ *it holds*

$$\mathbb{E}[I_{n,M}^{OOB}(j)] = 0 = I(j)$$

and for $j \in \mathcal{S}$ *we have*

$$\lim_{M\to\infty} \mathbb{E}[I_{n,M}^{OOB}(j)] \longrightarrow I(j), \quad \text{as } n \to \infty.$$

Theorem 1 and equation (9) under the assumption of an additive link function reveal some important insights about the RFPIM. In case of non-informative variables, i.e. Y is independent of X_j or equivalently, $\partial\widetilde{m}(\mathbf{x})/\partial x_j \equiv 0$, the empirical variable importance does not select on average across non-informative variables. However, if the variable is informative, that is $\partial\widetilde{m}(\mathbf{x})/\partial x_j \neq 0$ and X_j depends on Y, this will lead to $I(j) > 0$, such that on average, their is enough discriminating power between informative and non-informative variables. Furthermore, the theoretical results obtained from Theorem 1 and equation (9) allow the sorting of variables according to their signal strength, if the underlying link-function is assumed to be additive. Hence, under the assumptions $(A1)$ - $(A5)$ together with the assumption that $\widetilde{m}$ decomposes into an additive expansion of measurable functions, the RFPIM does not only detect informative variables, but also delivers an internal ranking across variables in $\mathcal{S}$. In addition, the theoretical results in Theorem 1 also reveal that unimportant variables tends to 0 stronger than important ones, since the unbiasedness is exact in that case for any sample size $n \in \mathbb{N}$ and number of base learners $M \in \mathbb{N}$. The theoretical findings also indicate that the discriminating power of the permutation importance depends on the sample size of the training set $\mathcal{D}_n$ and the number of base learners M. Larger sample sizes with a relatively large number of decision trees in the Random Forest should deliver stronger discriminating power between variables in $\mathcal{S}$ and $\{1, \ldots, p\} \setminus \mathcal{S}$. Note that the theoretical findings do not reveal insights into the rate of convergence of the asymptotic. However, an important factor influencing the discriminating power of the permutation importance measure that cannot be directly extracted from the theoretical findings so far is the random noise arising from the residuals ϵ. These contaminate the data especially depending on the scale of their variance σ^2. Nonetheless, if the systematic

signal arising from the link function $m(\mathbf{x})$ is strong enough, the effect of noise can be appeased. Thus, keeping an eye on the ratio

$$SN = \frac{Var(\widetilde{m}(\mathbf{X}))}{\sigma^2} \tag{10}$$

is an important task during the computation of the RFPIM. We refer to this meaasure as the *signal-to-noise* ratio, which is formally defined in [17]. Although this factor cannot be directly detected based on the results in Theorem 1, a closer look at the specific cut criterion used in the Random Forest will deliver some insights into the interaction of SN and the permutation measure $\boldsymbol{I}_{n,M}^{OOB} = [I_{n,M}^{OOB}(1), \ldots, I_{n,M}^{OOB}(p)]^\top \in \mathbb{R}^p$. Recall that the empirical cut criterion of the Random Forest model within the construction of each tree is given by

$$\begin{aligned} L_{n,t}^{(k)}(j,z) = &\frac{1}{N_n(A_\ell^{(k)})} \sum_{i=1}^{n} (Y_i - \bar{Y}_{A_\ell^{(k)}})^2 \mathbb{1}\{\mathbf{X}_i \in A_\ell^{(k)}\} \\ &- \frac{1}{N_n(A_\ell^{(k)})} \sum_{i=1}^{n} (Y_i - \bar{Y}_{A_{\ell,L}^{(k)}} \mathbb{1}\{X_{ji} < z\} - \bar{Y}_{A_{\ell,R}^{(k)}} \mathbb{1}\{X_{ji} \geq z\})^2 \mathbb{1}\{\mathbf{X}_i \in A_\ell^{(k)}\} \end{aligned} \tag{11}$$

for $t = 1, \ldots, M$. Here $A_\ell^{(k)} = A_\ell^{(k)}(\Theta_t) \subset [0,1]^p$ denotes the hyper-rectangular cell obtained after cutting the tree at level $k \in \{1, \ldots, \lceil \log_2(t_n) \rceil + 1\}$, $A_{\ell,L}^{(k)} = A_{\ell,L}^{(k)}(\Theta_t)$ denotes the left hyper-rectangular cell after cutting $A_\ell^{(k)}$ on variable j in z, i.e. $A_{\ell,L}^{(k)} = \{\mathbf{x} \in A_\ell^{(k)} : x_j < z\}$ and $A_{\ell,R}^{(k)} = A_{\ell,R}^{(k)}(\Theta_t)$ is the corresponding right hyper-rectangular cell $\{\mathbf{x} \in A_\ell^{(k)} : x_j \geq z\}$. Moreover, $\bar{Y}_A$ is the mean of all Y's, that belong to the cell A and $N_n(A)$ refers to the number of observations falling into cell A. As stated in [3], the strong law of large numbers for $n \to \infty$ leads to the consideration of

$$\begin{aligned} L^{(k)}(j,z) = &Var[Y_1 | \mathbf{X}_1 \in A_\ell^{(k)}] - \mathbb{P}[X_{j,1} < z | \mathbf{X}_1 \in A_\ell^{(k)}]_, \cdot Var[Y | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} < z] \\ &- \mathbb{P}[X_{j,1} \geq z | \mathbf{X}_1 \in A_\ell^{(k)}] \cdot Var[Y_1 | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} \geq z] \end{aligned} \tag{12}$$

such that $L_{n,1}^{(k)}(j,z) \longrightarrow L^{(k)}(j,z)$ holds $\mathbb{P}$ - almost surely for all $(j,z) \in \{1, \ldots, p\} \times [0,1]$. If we oppose the cut criterion of the Random Forest to the variance decomposition of the response, we obtain

$$Var(Y_1) = Var(\widetilde{m}(\mathbf{X}_1)) + \sigma^2. \tag{13}$$

Assuming that the Random Forest is cut-consistent, that is

$$(j_n, z_n) := \arg\max_{j,z} L_{n,t}^{(k)}(j,z) \longrightarrow (j,z) = \arg\max_{j,z} L^{(k)}(j,z), \quad \mathbb{P}- \text{ almost surely}, \tag{14}$$

the influence of the signal-to-noise ratio on the cuts (j_n, z_n) reduces immediately, since the residual variance drops out of the theoretical cut criterion which is then given by $L^{(k)}(j,z) = Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}] - \mathbb{P}[X_{j,1} < z | \mathbf{X}_1 \in A_\ell^{(k)}] \cdot Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} < z] - \mathbb{P}[X_{j,1} \geq z | \mathbf{X}_1 \in A_\ell^{(k)}] \cdot Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} \geq z]$. For a formal proof, we refer to the Appendix. However, this clearly depends on the sample size and the assumption that Random Forest cuts are consistent M-estimators in the sense of (14). The proof of the latter should therefore be considered in future research. In case of σ^2 being larger than $Var(\widetilde{m}(\mathbf{X}))$, the cut (j_n, z_n) conducted by the Random Forest might be inflated in terms of potentially selecting non-informative variables. The estimation of SN can therefore be considered as an additional control mechanism

in computing $\boldsymbol{I}_{n,M}^{OOB}$. The authors in [15] proved the consistency of several estimators for σ^2, which are based on the sampling variance of residuals obtained from the Random Forest model using Out-of-Bag samples. These results enables practitioners to consistently estimate the signal-to-noise ratio given by

$$\widehat{SN}_n = \frac{|\hat{\sigma}_Y^2 - \hat{\sigma}_{RF}^2|}{\hat{\sigma}_{RF}^2}, \tag{15}$$

where $\hat{\sigma}_Y^2$ is the sampling variance of the response Y and $\hat{\sigma}_{RF}^2$ an residual variance estimator as given in [15]. In the sequel, we simply restrict our attention to the residual sampling variance estimator $\hat{\sigma}_{RF}^2 = 1/n \sum_{i=1}^{n} (\hat{\epsilon}_i - \bar{\hat{\epsilon}}_n)^2$ for σ^2 as described in [15], where $\hat{\epsilon}_i = Y_i - m_n^{OOB}(\mathbf{X}_i)$ and $\bar{\hat{\epsilon}}_n$ is its corresponding mean.

4. Simulation Study

In order to provide practical evidence for the theoretical results of the previous section, we simulated artificial data and computed the empirical variable importance measure based on Out-of-Bag estimates for every variable. In doing so, several regression functions have been considered that are in line with the assumptions of the previous section. We first consider $p = 10$ covariates whose influence on Y is described by means of a regression coefficient vector $\boldsymbol{\beta}_0 = [2, 4, 2, -3, 1, 0, 0, 0, 0, 0]^\top$. The data is then generated under the following frameworks:

1. For the simplest case, we assume a linear model, i.e. $m(\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}_0$, for $i = 1, \ldots, n$.
2. Here, we assume a polynomial relationship, that is, $m(\mathbf{x}_i) = \sum_{j=1}^{p} \beta_{0,j} x_{i,j}^j$ for $i = 1, \ldots, n$.
3. In order to capture recurrent effects, a trigonometric function is assumed, i.e. $m(\mathbf{x}_i) = 2 \cdot \sin(\mathbf{x}_i^\top \boldsymbol{\beta}_0 + 2)$ for $i = 1, \ldots, n$.
4. Finally, the effect of non-continuous functions is considered, that is

$$m(\mathbf{x}_i) = \begin{cases} \beta_{0,1} x_{i,1} + \beta_{0,2} x_{i,2} + \beta_{0,3} x_{i,3}, & \text{if } x_{i,3} > 0.5 \\ \beta_{0,4} x_{i,4} + \beta_{0,5} x_{i,5} + 3 & \text{if } x_{i,3} \leq 0.5 \end{cases}$$

for $i = 1, \ldots, n$.

We used $MC = 1,000$ Monte-Carlo iterations to approximate the expectation of $\boldsymbol{I}_{n,M}^{OOB}$. That is, for every $m_c \in \{1, \ldots, MC\}$, we generated $\mathcal{D}_n^{m_c} = \{[\mathbf{X}_i^{m_c \top}, Y_i^{m_c}]^\top : i = 1, \ldots, n\}$, where $\mathbf{X}_i^{m_c} \sim Unif([0,1]^p)$ and $Y_i = m(\mathbf{X}_i^{mc}) + \epsilon_i$ for every $i = 1, \ldots, n$ and $m_c = 1, \ldots, MC$. On every generated data set $\mathcal{D}_n^{mc}$, the empirical permutation importance based on Out-of-Bag samples $I_{n,M;m_c}^{OOB}(j)$, $j \in \{1, \ldots, p\}$ is then computed. By the strong law of large number, we can guarantee almost surely that

$$\bar{I}_{n,M;\cdot}^{OOB}(j) := \frac{1}{MC} \sum_{m_c=1}^{MC} I_{n,M;m_c}^{OOB}(j) \longrightarrow \mathbb{E}[I_{n,M}^{OBB}(j)], \tag{16}$$

as $MC \to \infty$, which should give some practical insights into Theorem 1. Different sample sizes of the form $n \in \{50, 100, 500, 1000\}$ should also reflect the behavior of the permutation importance as prescribed in

Theorem 1. Throughout our simulations, we used $M = 1,000$ decision trees in the Random Forest model and trained it using sampling without replacement of $a_n = \lceil 2/3n \rceil < n$ data points.

Regarding the noise ϵ, a centered Gaussian distribution with homoscedastic variance σ^2 is assumed. As explained at the end of Section 3, the discriminative power of the permutation importance measure clearly depends on the signal-to-noise ratio. In order to explore this effect, a signal-to-noise ratio of $SN \in \{0.5, 1, 3, 5\}$ is considered. That is, the residual variance σ^2 is determined by setting $\sigma^2 = Var(m(\mathbf{X}_1)) \cdot SN^{-1}$.

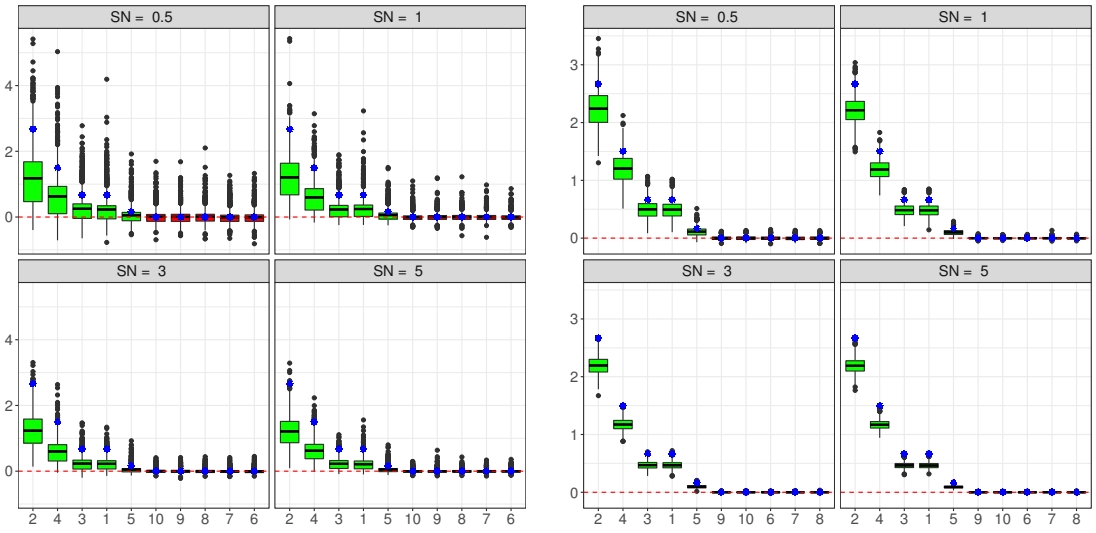
We additionally generated data under high-dimensional settings, i.e. for $\boldsymbol{\beta}_1 = [2, 4, 2, -3, 1, \mathbf{0}^\top]^\top \in \mathbb{R}^{n+5}$ and $n \in \{50, 100, 500, 1000\}$, we generated $\mathcal{D}_n^{m_c}$ and computed the permutation importance for every Monte-Carlo set $\mathcal{D}_n^{m_c}$. This leads to regression problems of the type $p > n$, for which Theorem 1 - unless not any of the given assumptions are violated - should also be valid.

4.1. Simulation Results

In this section, we present the simulation result for all four models $1. - 4.$ with $p = 10$ and a sample size of $n \in \{50, 1000\}$. The results for the other sample sizes are moved to the supplement. Note that the solid black lines in the boxplots represented in Figure 1 to 4, refer not to the median, but to the empirical mean $\bar{I}^{OOB}_{n,M;\cdot}(j)$ as computed in (16). The blue star point $\star$ in the plots refer to the expected value of the permutation importance based on Out-of-Bag samples. For additive models such as the linear and polynomial model, a direct computation of $\boldsymbol{I}$ could be obtained using equation (9). For non-additive link-functions, such as in the trigonometric or non-continuous case, the results given in Proposition 2 are used and approximated with additional $1,000$ Monte-Carlo iterations.

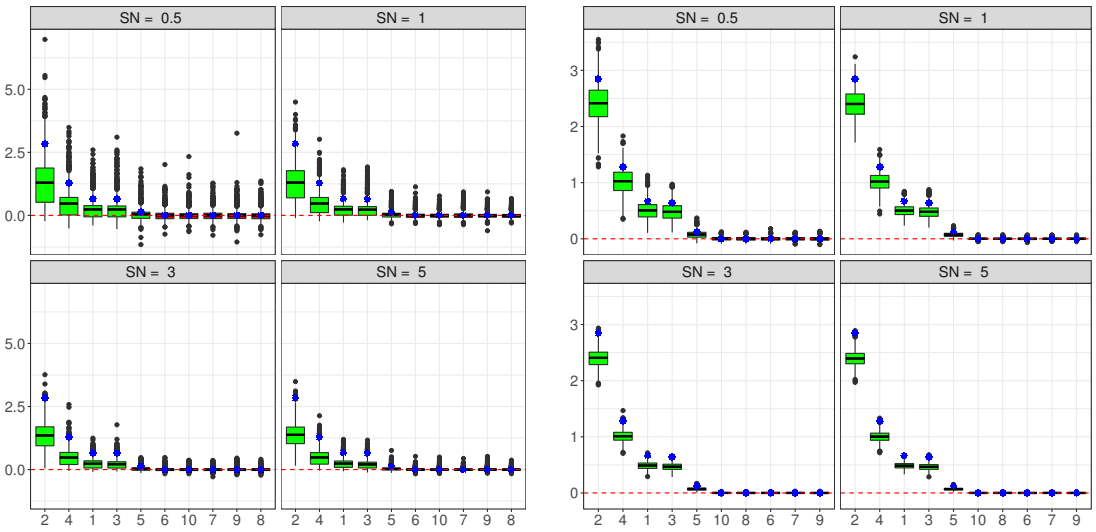
Figure 1 gives boxplots of the permutation importance of all ten variables over all Monte-Carlo iterations for the **linear model**. It is apparent that in case of small sample sizes (left panel), the permutation importance had difficulties in clearly distinguishing informative and non-informative variables. This is in line with the asymptotic results obtained in Theorem 1. The simulation results reveal that this depends on the signal-to-noise ratio and the scale of the regression coefficient, as discussed in Section 3. For a signal-to-noise ratio less than 1, a clear distinction was rather hard. Under the same sample size, with a signal-to-noise ratio larger than 1, the permutation importance could distinguish informative and non-informative variables clearer. Smaller regression coefficients being close to 0 such as $\beta_{0,5}$ resulted into lower permutation importance values. This is in line with equation (9), which results into $I(5) = \beta_{0,5}^2/6 = 1/6 \leq \min_{j \in \{1,\dots,5\}} I(j)$. For larger sample sizes (right panel), the distinction power of the permutation importance is stronger making the dependence towards the signal-to-noise ratio weaker, as shown in Section 3, considering the asymptotic of $I^{OOB}_{n,M}(j)$, $j \in \mathcal{S}$.

Regarding the **polynomial model**, the distinction power of the permutation importance increased, which can be extracted from Figure 2. Under this setting, a sufficiently large signal-to-noise ratio could lead to a stronger distinction even for small sample sizes like $n = 50$ (left panel). Larger sample sizes emphasized the distinction making the selection clearer and more independent towards the signal-to-noise ratio as shown in Section 3 by considering the cut criterion used in the Random Forest. In addition, the empirical mean of the simulated result approached its theoretical, asymptotic counterpart $\boldsymbol{I}$ as proven in Theorem 1.



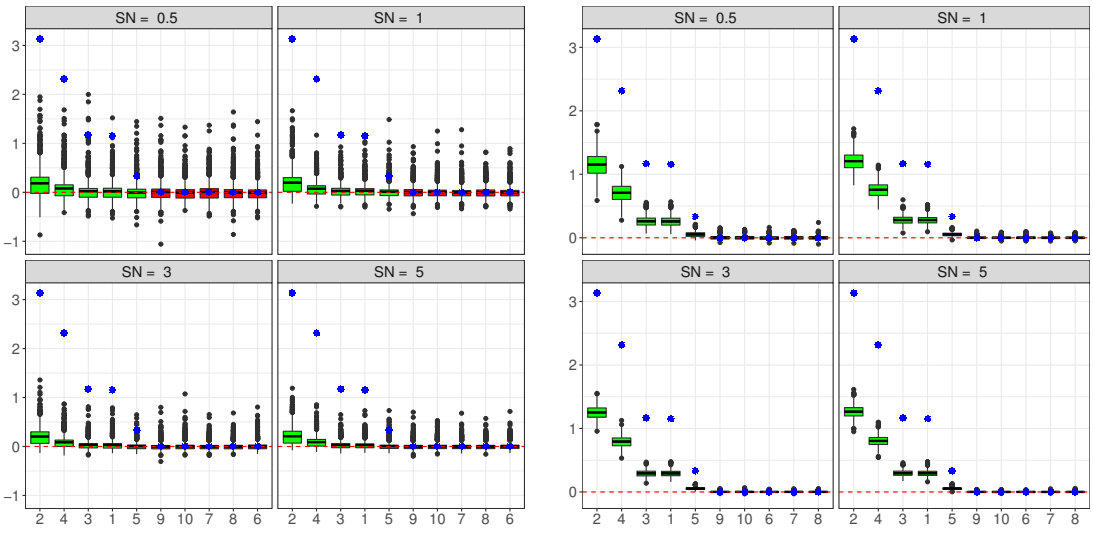
(a) $n = 50$ (b) $n = 1,000$

Figure 1: Permutation importance with various signal-to-noise ratios under a **linear model** as described in 1. using $MC = 1,000$ Monte-Carlo iterations with a sample size of (a) $n = 50$ and (b) $n = 1,000$. The solid line refers to the empirical mean $\bar{I}^{OOB}_{n,M;\cdot}$ and $\star$ to its expectation.



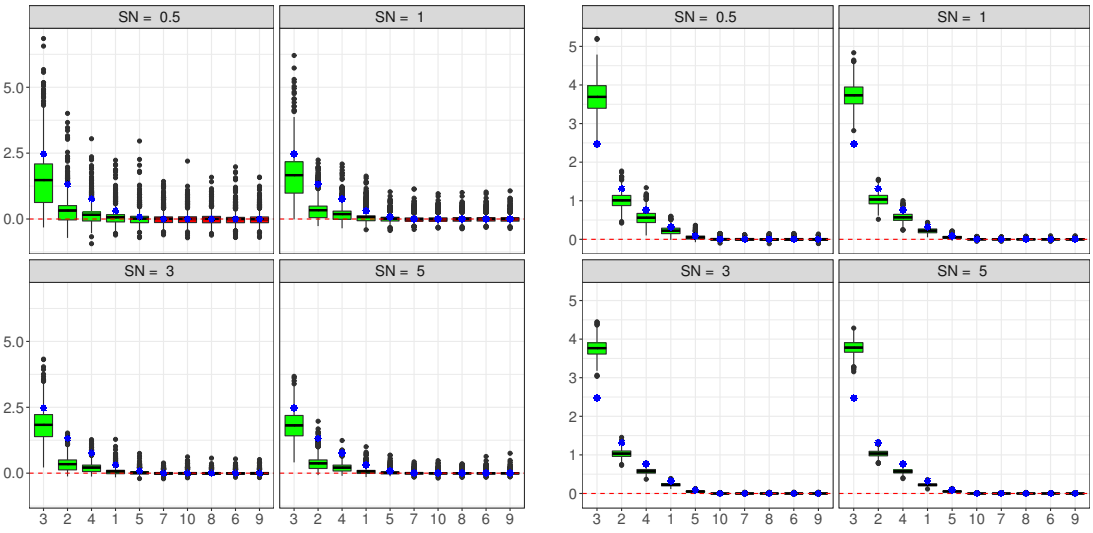
(a) $n = 50$ (b) $n = 1,000$

Figure 2: Permutation importance with various signal-to-noise ratios under a **polynomial model** as described in 2. using $MC = 1,000$ Monte-Carlo iterations with a sample size of (a) $n = 50$ and (b) $n = 1,000$. The solid line refers to the empirical mean $\bar{I}^{OOB}_{n,M;\cdot}$ and $\star$ to its expectation.



(a) $n = 50$ (b) $n = 1,000$

Figure 3: Permutation importance with various signal-to-noise ratios under a **trigonometric model** as described in 3. using $MC = 1,000$ Monte-Carlo iterations with a sample size of (a) $n = 50$ and (b) $n = 1,000$. The solid line refers to the empirical mean $\bar{I}^{OOB}_{n,M;\cdot}$ and $\star$ to a Monte-Carlo approximation of its expectation.



(a) $n = 50$ (b) $n = 1,000$

Figure 4: Simulation results for the permutation importance with various signal-to-noise ratios under a **non-continuous model** as described in 4. using $MC = 1,000$ Monte-Carlo iterations with a sample size of (a) $n = 50$ and (b) $n = 1,000$. The solid line refers to the empirical mean $\bar{I}^{OOB}_{n,M;\cdot}$ and $\star$ to a Monte-Carlo approximation of its expectation.

For the polynomial model, we can also make use of equation (9), which will lead us to $I(j) = 2\beta_j^2 \left(\frac{1}{2j+1} - \frac{1}{(j+1)^2}\right)$ for $j \in \mathcal{S}$. This justifies the relatively small values of $I^{OOB}_{n,M}(5)$, which should lie around $25/198 \approx 0.13$.

Regarding the **trigonometric** link function, the permutation importance measure lost in separating force when the sample size was relatively small. Here, a larger signal-to-noise ration was helpful, but for weak signals such as β_5, a clear distinction was rather hard. The results turned quickly into the right direction, when the sample size increased (right panel), as illustrated in Figure 3. In the latter scenario, the permutation importance was able to distinguish between elements in $\mathcal{S}$ and $\{1, \dots, p\} \setminus \mathcal{S}$ while the empirical mean approached its theoretical counterpart $\boldsymbol{I}$. This was rather independent of the signal-to-noise ratio, as discussed in Section 3. Note that under this model, equation (9) cannot be applied. However, it seems that a stronger or weaker signal resulted into lower or higher permutation importance.

Moving to the **non-continuous** case with linear sub-functions, a stronger distinction power could be obtained compared to the linear link function. This, although equation 9 is not applicable. A detailed result of the permutation importance measure under this setting can be extracted from Figure 4. There, the boxplot indicated a strong discriminative power towards non-informative variables for larger data sets, independent of the signal-to-noise ratio. The empirical mean of the simulated importance measures approached its theoretical counterpart $\boldsymbol{I}$ for an increased sample size. In addition, more importance is put on variable 3 compared to the other frameworks. This arises from the usage of the third variable for both, the localization of the discontinuity point and its contribution to the response through the linear sub-function. However, this effect should vanish asymptotically according to Theorem 1, as long as the assumptions are met.

Under all settings, it is worth to notice that the permutation importance resulted into larger variability, if the variables were informative. For non-informative variables, the Random Forest was *sure* which variables were non-informative, especially when sample size increased. In fact, under all simulation settings, the RFPIM attained values very close to zero. This supports the findings in Theorem 1 for unimportant features, as the permutation importance is exactly unbiased in this case.

The boxplots of the permutation importance for the **high-dimensional** settings are summarized in Figures 5 - 8 given in the supplement. Under this framework, the linear model (see Figure 5 in the supplement) lost in distinction power compared to $p < n$ problems, especially when the sample size was relatively small. Although $p > n$, an increase in n led to an increase in separation force between variables in $\mathcal{S}$ and $\{1, \dots, p\}$ making the results clearer for $n \geq 500$. The empirical mean of the permutation importance moved closer to its theoretical counterpart $I(j)$, $j = 1, \dots, 5$. For $j \in \{6, \dots, 10\}$, they were almost exactly to zero as proven in Theorem 1. There is also an increase in variation under the high-dimensional setting. Regarding the polynomial model (see Figure 6 in the supplement), similar results could be obtained compared to the $p < n$ regression problem. However, the permutation importance was slightly downsized for all variables, but the distinction force was similar. Under the trigonometric function with $p > n$ (see Figure 7 in the supplement), the permutation importance lost in separation force when the sample size was small. Evaluable results could be obtained for $n = 1,000$, but the permutation measure was again downsized for all variables compared to its analogon under $p < n$. The simulation reveals that the convergence of the expectation is slower compared to its $p < n$ analogon. The non-continuous case (see Figure 8 in the supplement) led to similar results than

under the scenario of p being less than n, with the exception that the permutation importance was again slightly downsized for all variables again.

Final Thoughts. Under both settings, i.e. $p < n$ and $p > n$, the permutation importance measure ranked the variables correctly according to the results given in equation (9) for the linear and polynomial model. The ranking remained the same for the trigonometric case, but was slightly changed when the sample size was rather small in high-dimensional settings. The ranking of the variables changed under the non-continuous model, where additional importance was set to variable 3 for playing the role of a discontinuity point and its systematic influence on Y through the sub-function. However, according to our findings, this effect should vanish asymptotically.

5. Conclusion

We proved the (asymptotic) unbiasedness of the permutation importance measure originating from the Random Forest for regression models. Our results are mainly based on assuming that features are independent, and hence uncorrelated while requiring that the Random Forest is L_2-consistent. Furthermore, we identified main drivers for the quality of the variable selection process such as the signal-to-noise ratio by explicitly considering the cut criterion of the Random Forest model. An extensive simulation study has been conducted for low- ($p < n$) and high-dimensional ($p > n$) regression frameworks. The results support our theoretical findings: even under high-dimensional settings, the permutation importance was able to correctly select among informative features, when the sample size was sufficiently large. Our findings also indicate that potential future research is worth to be conducted on (i) the consistency of the involved cut-criterion and (ii) the (asymptotic) distribution of the Random Forest permutation importance as a preliminary step towards the construction of valid statistical testing procedures for feature selection.

Acknowledgement

We are very thankful to Gérard Biau and Erward Scornet for fruitful disucssions on Random Forest related issues during a scholary visit at the Sorbonne Université and the École Polytechnique.

6. Appendix

In this section we state the proofs of Propositions 1 and 2 and Theorem 1. Additional proofs mentioned in the article are shifted at the end of this section.

Proof of Proposition 1. Let $i \in \{1, \ldots, n\}$ be fixed and $\mathbf{X}_i \in \mathcal{D}_n$. Let $\{\mathbf{\Theta}_t\}_{t=1}^M$ be the sequence of iid generic random vectors on the probability space $(\Omega_\Theta, \mathcal{F}_\Theta, \mathbb{P}_\Theta)$ being responsible for the sampling procedure and the feature sub-spacing in the Random Forest algorithm. Note that the generic random vector can then be decomposed into $\mathbf{\Theta}_t = [\mathbf{\Theta}_t^{(1)}, \mathbf{\Theta}_t^{(2)}]^\top$, where $\mathbf{\Theta}_t^{(1)} \in \{0,1\}^n$ indicates whether a certain observation has been selected in tree t and $\mathbf{\Theta}_t^{(2)}$ models feature sub-spacing. Furthermore, denote with $Z_i = Z_i(M)$ the number of the M regression trees not containing the i-th observation. Then we can conclude that

$$Z_i(M) \sim Bin(M, c_n), \quad \text{where} \quad c_n = \begin{cases} 1 - a_n/n & \text{for subsampling,} \\ (1-1/n)^n & \text{for bootstrapping with replacement,} \end{cases}$$

with $c_n > 0$. Since $Z_i(M) = \sum_{\ell=1}^{M} B_\ell$, with $B_\ell \sim Bernoulli(c_n)$ independent and identically distributed under $\mathbb{P}_\Theta$, it follows by the strong law of large numbers that $V_{n,M} := Z_i(M)/M \xrightarrow{a.s.} \mathbb{E}[B_1] = c_n$, as $M \to \infty$. This implies that $Z_i(M) \xrightarrow{a.s.} \infty$, as $M \to \infty$. Assuming without loss of generality that the first $Z_i(M)$ decision trees do not contain the i-th observation, this will yield to

$$R_{n,M} := \frac{1}{Z_i(M)} \sum_{t=1}^{Z_i(M)} m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_t, \mathcal{D}_n) \longrightarrow m_n^{OOB}(\mathbf{X}_i) \quad \mathbb{P}_\Theta - a.s. \text{ as } M \to \infty, \tag{17}$$

where $m_n^{OOB}(\mathbf{X}_i) = \mathbb{E}_{\boldsymbol{\Theta}_{[i]}}[m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_{[i]}, \mathcal{D}_n)]$ with $\boldsymbol{\Theta}_{[i]} = [\boldsymbol{\Theta}^{(1)}, \Theta^{(2)}]$, such that $\Theta_i^{(1)} = 0$. Now, let $\mathbf{K}_{n,M} = [V_{n,M}, R_{n,M}]^\top \in \mathbb{R}^2$ and set $N = N_1 \cup N_2$, where $N_1 = \{\omega \in \Omega_\Theta : V_{n,M}(\omega) \nrightarrow c_n\}$ and $N_2 = \{\omega \in \Omega_\Theta : R_{n,M}(\omega) \nrightarrow m_n^{OOB}(\mathbf{X}_i)\}$. Since $\mathbb{P}_\Theta(N_1) = \mathbb{P}_\Theta(N_2) = 0$, it follows immediately that $0 \leq \mathbb{P}_\Theta(N) = \mathbb{P}_\Theta(N_1) + \mathbb{P}_\Theta(N_2) = 0$, i.e. N is a null-set. Hence,

$$\mathbf{K}_{n,M} \longrightarrow [c_n, m_n^{OOB}(\mathbf{X}_i)]^\top, \quad \mathbb{P}_\Theta - \text{ almost-surely as } M \to \infty. \tag{18}$$

Since $\{\boldsymbol{\Theta}_t\}_{t=1}^M$ is a sequence of iid random variables, we can again assume without loss of generality, that the first $Z_i(M)$ do not contain the i-th observation. Therefore, we can conclude that

$$\begin{aligned} \frac{1}{M} \sum_{t=1}^{M} m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_t) \mathbb{1}\{\mathbf{X}_i \text{ has not been selected}\} &= \frac{Z_i(M)}{M} \frac{1}{Z_i(M)} \sum_{t=1}^{Z_i(M)} m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_t) \\ &\longrightarrow c_n \cdot m_n^{OOB}(\mathbf{X}_i), \end{aligned} \tag{19}$$

$\mathbb{P}_\Theta -$ almost-surely as $M \to \infty$. The convergence follows by applying the continuous mapping theorem on the function $g(x, y) = x \cdot y$ using $\mathbf{K}_{n,M}$ and (18). □

Proof of Proposition 2. Let $\mathbf{X} = [X_1, \ldots, X_p]^\top \in \mathbb{R}^p$ be an independent copy of $\mathbf{X}_1$ such that $Y = \widetilde{m}(\mathbf{X}) + \epsilon$ as in regression model (1). Furthermore, Let $j \in \{1, \ldots, p\} \setminus \mathcal{S}$, i.e. j is non-informative. According to our definition of being non-informative and the assumption that there are no dependencies among the features $\{X_j\}_{j=1}^p$, this will lead us to Y being indepdendent of X_j, while X_j is also independent towards all other features X_ℓ, $\ell \neq j \in \{1, \ldots, p\}$. Denoting with $\mathbf{X}_j = [X_1, \ldots, X_{j-1}, Z_j, X_{j+1}, \ldots, X_p]^\top \in \mathbb{R}^p$, while Z_j is an independent copy of X_j, independent of X_ℓ and Y for all $\ell \neq j$, this will yield to $[\mathbf{X}_j^\top, Y]^\top \overset{d}{=} [\mathbf{X}^\top, Y]^\top$. Hence, we will obtain

$$\begin{aligned} I(j) &= \mathbb{E}[(Y - \widetilde{m}(\mathbf{X}_j))^2] - \mathbb{E}[(Y - \widetilde{m}(\mathbf{X}))^2] \\ &= \mathbb{E}[(Y - \widetilde{m}(\mathbf{X}))^2] - \mathbb{E}[(Y - \widetilde{m}(\mathbf{X}))^2] \\ &= 0. \end{aligned} \tag{20}$$

On the other hand, if $j \in \mathcal{S}$, i.e. j is informative, than we can deduce the following computations, where the third equation follows from the independence of $\mathbf{X}$ and ϵ together with $\mathbb{E}[\epsilon] = 0$. The second last equality follows from assumption $(A3)$ leading to $\mathbf{X}_j \stackrel{d}{=} \mathbf{X}$.

$$\begin{aligned} I(j) &= \mathbb{E}[(Y - \widetilde{m}(\mathbf{X}_j))^2] - \mathbb{E}[(Y - \widetilde{m}(\mathbf{X}))^2] \\ &= \mathbb{E}[(Y - \widetilde{m}(\mathbf{X}) + \widetilde{m}(\mathbf{X}) - \widetilde{m}(\mathbf{X}_j))^2] - \mathbb{E}[(Y - \widetilde{m}(\mathbf{X}))^2] \\ &= \mathbb{E}[(\widetilde{m}(\mathbf{X}) - \widetilde{m}(\mathbf{X}_j))^2] + 2\mathbb{E}[\epsilon(\widetilde{m}(\mathbf{X}) - \widetilde{m}(\mathbf{X}_j)))] \\ &= \mathbb{E}[(\widetilde{m}(\mathbf{X}) - \widetilde{m}(\mathbf{X}_j))^2]. \end{aligned} \tag{21}$$

□

Proof of Theorem 1. Let $j \in \{1, \ldots, p\}$, $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, M\}$ be fixed but arbitrary and assume that the Random Forest sampling mechanism is restricted to sampling $a_n \in \{1, \ldots, n\}$ points without replacement such that $a_n < n$. Denote with $\mathcal{D}_n^{(t)}$ the collection of points selected for tree $t \in 1, \ldots, M$. Then we denote with $\mathcal{D}_n^{-(t)} = \mathcal{D}_n \setminus \mathcal{D}_n^{(t)}$ the subset of $\mathcal{D}_n$ in tree $t \in \{1, \ldots, M\}$ with cardinality γ_n for which its elements have not been selected during the sampling procedure. Note that the cardinality of $\mathcal{D}_n^{-(t)}$ remains fixed for all $t = 1, \ldots, M$ and is given by $\gamma_n = n - a_n$, which is different to sampling with replacement. In addition, we set $\mathcal{D}_{n,\mathbf{X}}^{-(t)} = \{\mathbf{X}_i : [\mathbf{X}_i^\top, Y_i]^\top \in \mathcal{D}_n^{-(t)}\}$ to be the set of all features $\mathbf{X}$ that belong to $\mathcal{D}_n^{-(t)}$, i.e. that have been selected during resampling. Then we recall from (7) that the permutation variable importance based on OOB estimates is given by

$$\begin{aligned} I_{n,M}^{OOB}(j) &= \frac{1}{M\gamma_n} \sum_{t=1}^{M} \sum_{i \in \mathcal{D}_n^{-(t)}} \left\{ (Y_i - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}; \boldsymbol{\Theta}_t))^2 - (Y_i - m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_t))^2 \right\} \\ &= \frac{1}{M\gamma_n} \sum_{t=1}^{M} \sum_{i=1}^{n} \left\{ (Y_i - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}; \boldsymbol{\Theta}_t))^2 - (Y_i - m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_t))^2 \right\} \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_n^{-(t)}\}, \end{aligned} \tag{22}$$

where $\pi_{j,t}$ is a *real* permutation of the j-th covariable in $\mathcal{D}_{n,\mathbf{X}}^{-(t)}$, where we call a permutation as *real*, if $\pi_{j,t} \in \{\pi \in \mathcal{S}_{\gamma_n} : \pi(i) \neq i\} =: \mathcal{V}$ and $\mathcal{S}_{\gamma_n}$ is the symmetric group. Although we did not yet specify the dependence of $\mathcal{D}_n^{(t)}$ and $\mathcal{D}_n^{-(t)}$ towards the generic random vector $\boldsymbol{\Theta}_t$ in the Random Forest mechanism, it is worth to notice that in fact, $\mathcal{D}_n^{(t)} = \mathcal{D}_n^{(t)}(\boldsymbol{\Theta}_t)$ and $\mathcal{D}_n^{-(t)} = \mathcal{D}_n^{-(t)}(\boldsymbol{\Theta}_t)$.
Then, the following results can be obtained:

$$\begin{aligned} \mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i))^2 \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] &= \mathbb{E}[\mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i))^2 \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\} | \mathcal{D}_n]] \\ &= \mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i))^2 \mathbb{P}[\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}(\Theta_t) | \mathcal{D}_n]] \\ &= \mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i))^2 (1 - \mathbb{P}[\mathbf{X}_i \notin \mathcal{D}_{n,\mathbf{X}}^{-(t)}(\Theta_t) | \mathcal{D}_n])] \\ &= \left(1 - \frac{\binom{n-1}{a_n-1}}{\binom{n}{a_n}}\right) \mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i))^2] \\ &= \frac{n - a_n}{n} \mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i))^2] \end{aligned} \tag{23}$$

The second equality follows from the measurability of $(Y_i - m(\mathbf{X}_i))$ and $\mathbb{P}[\mathbf{X}_i \notin \mathcal{D}_{n,\mathbf{X}}^{-(t)}(\Theta_t) | \mathcal{D}_n]$ is the probability of not selecting a fixed observation i among n elements, when resampling is conducted without replacement.

Returning to the sequence of iid generic random vectors $\{\boldsymbol{\Theta}_t\}_{t=1}^M$, we recall that we can separate each generic random vector into $\boldsymbol{\Theta}_t = [\boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}]$, where $\boldsymbol{\Theta}_t^{(1)}$ models the sampling mechanism prior to tree construction and $\boldsymbol{\Theta}_t^{(2)}$ is the random variable modeling feature sub-spacing during the tree construction. Note that in case of $m_{try} = p$, it follows that $\boldsymbol{\Theta}_t = \boldsymbol{\Theta}_t^{(1)}$. Furthermore, $\boldsymbol{\Theta}_t^{(1)}$ can be decomposed into

$$\boldsymbol{\Theta}_t^{(1)} = [\Theta_{1,t}^{(1)}, \ldots, \Theta_{n,t}^{(1)}]^\top \in \{0,1\}^n, \tag{24}$$

where each entry $\Theta_{\ell,t}^{(1)}$, $1 \leq \ell \leq n$ is Bernoulli distributed indicating whether observation ℓ has been selected during the sampling procedure. For sampling without replacement the sequence $\{\Theta_{\ell,t}^{(1)}\}_{\ell=1}^n$ does not consist of independent random variables. However, it holds that $\sum_{\ell=1}^n \Theta_{\ell,t} = a_n$ and that $\boldsymbol{\Theta}_t^{(1)}$ is independent of $(\mathbf{X}_i, Y_i, \boldsymbol{\Theta}_t^{(2)})$ for all $t = 1, \ldots, M$ and all $i = 1, \ldots, n$. Let $\Delta_n(\mathbf{X}_i, Y_i, \boldsymbol{\Theta}_t) = \Delta_n(\mathbf{X}_i, Y_i, \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}) := (\widetilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i, \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}))^2$, declare $\mathbf{X}_i'$ as an independent copy of $\mathbf{X}_i$ independent of $m_{n,1}$ and set $\mathcal{G} = \{[v_1, \ldots, v_n]^\top \in \{0,1\}^n : v_1 + \cdots + v_n = a_n\}$ and $\mathcal{G}_i := \{\mathbf{v} \in \mathcal{G} : v_i = 0\}$. Then we observe the following equality

$$\begin{aligned}
\mathbb{E}[\Delta_n(\mathbf{X}_i, Y_i, \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}) \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\}] &= \sum_{\boldsymbol{\ell} \in \mathcal{G}} \mathbb{E}[\Delta_n(\mathbf{X}_i, Y_i, \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}) \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\} | \boldsymbol{\Theta}_t^{(1)} = \boldsymbol{\ell}] \cdot \mathbb{P}[\boldsymbol{\Theta}_t^{(1)} = \boldsymbol{\ell}] \\
&= \sum_{\boldsymbol{\ell} \in \mathcal{G}_i} \mathbb{E}[\Delta_n(\mathbf{X}_i, Y_i, \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}) | \boldsymbol{\Theta}_t^{(1)} = \boldsymbol{\ell}] \cdot \mathbb{P}[\boldsymbol{\Theta}_t^{(1)} = \boldsymbol{\ell}] \\
&= \sum_{\boldsymbol{\ell} \in \mathcal{G}_i} \mathbb{E}[\Delta_n(\mathbf{X}_i', Y_i', \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}) | \boldsymbol{\Theta}_t^{(1)} = \boldsymbol{\ell}] \cdot \mathbb{P}[\boldsymbol{\Theta}_t^{(1)} = \boldsymbol{\ell}] \\
&= \mathbb{E}[\Delta_n(\mathbf{X}_i', Y_i', \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}) \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\}],
\end{aligned} \tag{25}$$

where the second last equality follows from the independence of $\boldsymbol{\Theta}_t^{(1)}$ and $(\mathbf{X}_i, Y_i, \boldsymbol{\Theta}_t^{(2)})$ and $(\mathbf{X}_i, Y_i, \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}) \stackrel{d}{=} (\mathbf{X}_i', Y_i', \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)})$. Now, using (25), we obtain

$$\begin{aligned}
0 \leq \mathbb{E}[(\widetilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_t))^2 \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] &= \mathbb{E}[\Delta_n(\mathbf{X}_i, Y_i, \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}) \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] \\
&= \mathbb{E}[\Delta_n(\mathbf{X}_i, Y_i, \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}) \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\}] \\
&= \mathbb{E}[\Delta_n(\mathbf{X}_i', Y_i', \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}) \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\}] \\
&= \mathbb{E}[\Delta_n(\mathbf{X}_i', Y_i', \boldsymbol{\Theta}_t^{(1)}, \boldsymbol{\Theta}_t^{(2)}) \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] \\
&= \mathbb{E}[(\widetilde{m}(\mathbf{X}_i') - m_{n,1}(\mathbf{X}_i'; \boldsymbol{\Theta}_t))^2 \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] \\
&=: C_{n,i,t}.
\end{aligned} \tag{26}$$

Note that the random tree estimate $m_{n,1}(\mathbf{X}_i'; \boldsymbol{\Theta}_t)$ can be rewritten into

$$m_{n,1}(\mathbf{X}_i'; \boldsymbol{\Theta}_t) = \sum_{j=1}^n W_{n,j}(\mathbf{X}_i'; \boldsymbol{\Theta}_t) Y_j, \tag{27}$$

where $W_{n,j}(\mathbf{X}_i'; \boldsymbol{\Theta}_t) = \frac{\mathbb{1}\{\mathbf{X}_j \in A_n(\mathbf{X}_i'; \boldsymbol{\Theta}_t)\}}{N_n(A_n(\mathbf{X}_i'; \boldsymbol{\Theta}_t))}$ with $A_n(\mathbf{X}_i'; \boldsymbol{\Theta}_t)$ being the hyper-rectangular cell containing $\mathbf{X}_i'$ under the random tree constructed by $\boldsymbol{\Theta}_t$ and $N_n(A_n(\mathbf{X}_i'; \boldsymbol{\Theta}_t))$ the number of observations falling in that

hyper-rectangular cell. This way, one can deduce that $0 \leq W_{n,j}(\mathbf{X}_i';\boldsymbol{\Theta}_t) \leq 1$ for all $j = 1,\ldots,n$ and $\sum_{j=1}^{n} W_{n,j}(\mathbf{X}_i';\boldsymbol{\Theta}_t) = 1$. Since $K := \sup_{\mathbf{x}} |\widetilde{m}(\mathbf{x})| < \infty$ by $(A4)$ one obtains $\mathbb{E}[Y_1^2] = \mathbb{E}[\widetilde{m}(\mathbf{X}_1)^2] + \sigma^2 < K^2 + \sigma^2 < \infty$ and together with the Cauchy-Schwarz inequality, it holds for all $n \in \mathbb{N}$ that

$$\begin{aligned}
C_{n,i,t} &\leq \mathbb{E}[(|\widetilde{m}(\mathbf{X}_i') - m_{n,1}(\mathbf{X}_i';\boldsymbol{\Theta}_t)|)^2] \leq \mathbb{E}[(|\widetilde{m}(\mathbf{X}_i')| + |m_{n,1}(\mathbf{X}_i';\boldsymbol{\Theta}_t)|)^2] \\
&\leq K^2 + 2K \cdot \mathbb{E}\left[\left|\sum_{j=1}^{n} W_{n,j}(\mathbf{X}_i';\boldsymbol{\Theta}_t) Y_j\right|^2\right]^{1/2} + \mathbb{E}\left[\left|\sum_{j=1}^{n} W_{n,j}(\mathbf{X}_i';\boldsymbol{\Theta}_t) Y_j\right|^2\right] \\
&\leq K^2 + 2K \cdot \mathbb{E}\left[\left(\sum_{j=1}^{n} W_{n,j}(\mathbf{X}_i';\boldsymbol{\Theta}_t)|Y_j|\right)^2\right]^{1/2} + \mathbb{E}\left[\left(\sum_{j=1}^{n} W_{n,j}(\mathbf{X}_i';\boldsymbol{\Theta}_t)|Y_j|\right)^2\right] \\
&\leq K^2 + 2K \cdot \mathbb{E}\left[\left(\sum_{j=1}^{n} |Y_j|\right)^2\right]^{1/2} + \mathbb{E}\left[\left(\sum_{j=1}^{n} |Y_j|\right)^2\right] \\
&\leq K^2 + 2Kn(\mathbb{E}[Y_1^2])^{1/2} + n^2\mathbb{E}[Y_1^2] < \infty. \qquad (28)
\end{aligned}$$

Set $\Delta_{n,i}(\boldsymbol{\Theta}_t) = \widetilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i;\boldsymbol{\Theta}_t)$ and recall that $\epsilon_i = Y_i - \widetilde{m}(\mathbf{X}_i)$ according to model (1). Then it follows from the law of total probability that

$$\begin{aligned}
\mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i))(\widetilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i;\boldsymbol{\Theta}_t))\mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] &= \mathbb{E}[\epsilon_i \cdot \Delta_{n,i}(\boldsymbol{\Theta}_t)\mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] \\
&= \mathbb{P}[\Theta_{i,t}^{(1)} = 0] \cdot \mathbb{E}[\epsilon_i \cdot \Delta_{n,i}(\boldsymbol{\Theta}_t)|\Theta_{i,t}^{(1)} = 0] \\
&= \frac{\gamma_n}{n} \cdot \mathbb{E}[\epsilon_i|\Theta_{i,t}^{(1)} = 0] \cdot \mathbb{E}[\Delta_{n,i}(\Theta_t)|\Theta_{i,t}^{(1)} = 0] \\
&= 0, \qquad (29)
\end{aligned}$$

since given the condition $\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}$, or equivalently, $\Theta_{i,t}^{(1)} = 0$, ϵ_i is independent of $\Delta_{n,i}(\boldsymbol{\Theta}_t)$. Furthermore note that we used the independence of ϵ_i towards $\mathbf{X}_i$ and $\Theta_{i,t}^{(1)}$ leading to $\mathbb{E}[\epsilon_i|\Theta_{i,t}^{(1)} = 0] = \mathbb{E}[\epsilon_i] = 0$.

240

Hence, combining the results from (23), (26) and (29), we obtain

$$\begin{aligned}
\mathbb{E}[(Y_i - m_{n,1}(\mathbf{X}_i;\boldsymbol{\Theta}_t))^2\mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] &= \mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i))^2\mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] \\
&\quad + 2\mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i))(\widetilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i;\Theta_t))\mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] \\
&\quad + \mathbb{E}[(\widetilde{m}(\mathbf{X}_i) - m_{n,1}(\mathbf{X}_i;\boldsymbol{\Theta}_t))^2\mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] \\
&= \frac{n - a_n}{n}\mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i))^2] + C_{n,i,t} \qquad (30)
\end{aligned}$$

Defining $\tilde{\mathbf{X}}_{j,i} = [X_{1,i},\ldots X_{j-1,i}, Z_j, X_{j+1,i},\ldots,X_{p,i}]^\top$ for $i = 1,\ldots,n$, where Z_j is independent of $[X_{1,i},\ldots,X_{j-1,i},X_{j+1,i},\ldots,X_{p,i}]$ and ϵ_i and Y_i, but has the same marginal distribution as $X_j \overset{d}{=} X_{j,i}$, we can deduce that for any arbitrary measurable function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ and $\pi \in \mathcal{V}$, it holds:

$$\begin{aligned}
\mathbb{E}[f(X_{1,i},\ldots,X_{j,\pi(i)},\ldots,X_{p,i})] &= \mathbb{E}[\mathbb{E}[f(X_{1,i},\ldots,X_{j,\pi(i)},\ldots,X_{p,i})|\pi]] \\
&= \mathbb{E}[f(\tilde{\mathbf{X}}_{j,i})], \qquad (31)
\end{aligned}$$

since $\mathbb{E}[f(X_{1,i},\dots,X_{j,\pi(i)},\dots,X_{p,i})|\pi] \stackrel{d}{=} \mathbb{E}[f(\tilde{\mathbf{X}}_{j,i})]$ due to the independence of the samples.
Now, following exactly the same calculation rules as in the derivation of equation (23), while also using (31), we receive

$$\begin{aligned}\mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] &= \frac{n-a_n}{n}\mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2] \\ &= \frac{n-a_n}{n}\mathbb{E}[(Y_i - \widetilde{m}(\tilde{\mathbf{X}}_{j,i}))^2]. \end{aligned} \tag{32}$$

Now denote with $\tilde{\mathbf{X}}'_{j,i}$ an independent copy of $\tilde{\mathbf{X}}_{j,i}$ independent of $m_{n,1}$. Since sampling is restricted to without replacement, the permutation $\pi_{j,t}$ is independent of $\mathbf{\Theta}_t$, $\mathcal{D}_n$ and hence independent of $\mathcal{D}_{n,\mathbf{X}}^{-(t)}$. This would be different if sampling is conducted with replacement, since the cardinality of $\mathcal{D}_{n,\mathbf{X}}^{-(t)}$ would be random leading to the dependence of $\pi_{j,t}$ towards $\mathbf{\Theta}_t$. This independence allows us to conduct the following computations

$$\begin{aligned} 0 &\leq \mathbb{E}[(\widetilde{m}(\mathbf{X}_i^{\pi_{j,t}}) - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}};\mathbf{\Theta}_t))^2 \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] \\ &= \mathbb{E}[(\widetilde{m}(\tilde{\mathbf{X}}_{j,i}) - m_{n,1}(\tilde{\mathbf{X}}_{j,i};\mathbf{\Theta}_t))^2 \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\}] && \text{(33a)} \\ &= \mathbb{E}[(\widetilde{m}(\tilde{\mathbf{X}}'_{j,i}) - m_{n,1}(\tilde{\mathbf{X}}'_{j,i};\mathbf{\Theta}_t))^2 \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\}] && \text{(33b)} \\ &= \mathbb{E}[(\widetilde{m}(\mathbf{X}'_{i}) - m_{n,1}(\mathbf{X}'_{i};\mathbf{\Theta}_t))^2 \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\}] \\ &= \mathbb{E}[(\widetilde{m}(\mathbf{X}'_{i}) - m_{n,1}(\mathbf{X}'_{i};\mathbf{\Theta}_t))^2 \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}\}] = C_{n,i,t}, && \text{(33)} \end{aligned}$$

where equality (33a) follows from applying (31), equality (33b) from the calculation results obtained from equation (25) and (26) and the second last equality from $\mathbf{X}'_i \stackrel{d}{=} \tilde{\mathbf{X}}'_{j,i}$ together with the independence property towards all other random elements, under the event that $\Theta_{i,t}^{(1)} = 0$.

Similarly, set $\tilde{\Delta}_{n,i}^{(j)}(\mathbf{\Theta}_t) = \widetilde{m}(\tilde{\mathbf{X}}_{j,i}) - m_{n,1}(\tilde{\mathbf{X}}_{j,i};\mathbf{\Theta}_t)$ and $\tilde{\epsilon}_{j,i} = Y_i - \widetilde{m}(\tilde{\mathbf{X}}_{j,i})$. Then, recall from model (1) that

$$\begin{aligned}\mathbb{E}[\tilde{\epsilon}_{j,i}|\mathbf{X}_i \in \mathcal{D}_{n,\mathbf{X}}^{-(t)}] &= \mathbb{E}[Y_i - \widetilde{m}(\tilde{\mathbf{X}}_{j,i})|\Theta_{i,t}^{(1)} = 0] \\ &= \mathbb{E}[\widetilde{m}(\mathbf{X}_i) + \epsilon_i - \widetilde{m}(\tilde{\mathbf{X}}_{j,i})|\Theta_{i,t}^{(1)} = 0] \\ &= \mathbb{E}[\widetilde{m}(\mathbf{X}_i)|\Theta_{i,t}^{(1)} = 0] + \mathbb{E}[\epsilon_i|\Theta_{i,t}^{(1)} = 0] - \mathbb{E}[\widetilde{m}(\tilde{\mathbf{X}}_{j,i})|\Theta_{i,t}^{(1)} = 0] \\ &= \mathbb{E}[\epsilon_i] + \mathbb{E}[\widetilde{m}(\mathbf{X}_i)] - \mathbb{E}[\widetilde{m}(\tilde{\mathbf{X}}_{j,i})] = \mathbb{E}[\widetilde{m}(\mathbf{X}_i)] - \mathbb{E}[\widetilde{m}(\mathbf{X}_i)] \\ &= 0, \end{aligned} \tag{34}$$

where we explicitly used assumption ($A3$) in the second-last equality equality and the independence of $\mathbf{\Theta}_t^{(1)}$ towards ϵ_i and $\mathbf{X}_i$ in the fourth equality. Now, consider

$$
\begin{aligned}
\mathbb{E}[\tilde{\epsilon}_{j,i} \cdot \tilde{\Delta}^{(j)}_{n,i}(\boldsymbol{\Theta}_t) \cdot \mathbb{1}\{\Theta^{(1)}_{i,t} = 0\}] &= \mathbb{E}[\epsilon_i \cdot \tilde{\Delta}^{(j)}_{n,i}(\boldsymbol{\Theta}_t) \cdot \mathbb{1}\{\Theta^{(1)}_{i,t} = 0\}] + \\
&\quad + \mathbb{E}[(\widetilde{m}(\mathbf{X}_i) - \widetilde{m}(\tilde{\mathbf{X}}_{j,i}))(\widetilde{m}(\tilde{\mathbf{X}}_{j,i}) - m_{n,1}(\tilde{\mathbf{X}}_{j,i}; \boldsymbol{\Theta}_t)) \cdot \mathbb{1}\{\Theta^{(1)}_{i,t} = 0\}] \\
&= \mathbb{P}[\Theta^{(1)}_{i,t} = 0] \cdot \mathbb{E}[\epsilon_i] \cdot \mathbb{E}[\tilde{\Delta}^{(j)}_{n,i}(\boldsymbol{\Theta}_t) | \Theta^{(1)}_{i,t} = 0] \\
&\quad + \mathbb{E}[(\widetilde{m}(\mathbf{X}_i) - \widetilde{m}(\tilde{\mathbf{X}}_{j,i}))(\widetilde{m}(\tilde{\mathbf{X}}_{j,i}) - m_{n,1}(\tilde{\mathbf{X}}_{j,i}; \boldsymbol{\Theta}_t)) \cdot \mathbb{1}\{\Theta^{(1)}_{i,t} = 0\}] \\
&= \mathbb{E}[(\widetilde{m}(\mathbf{X}_i) - \widetilde{m}(\tilde{\mathbf{X}}_{j,i}))(\widetilde{m}(\tilde{\mathbf{X}}_{j,i}) - m_{n,1}(\tilde{\mathbf{X}}_{j,i}; \boldsymbol{\Theta}_t)) \cdot \mathbb{1}\{\Theta^{(1)}_{i,t} = 0\}] \\
&= \mathbb{P}[\Theta^{(1)}_{i,t} = 0] \cdot \mathbb{E}[(\widetilde{m}(\mathbf{X}_i) - \widetilde{m}(\tilde{\mathbf{X}}_{j,i}))(\widetilde{m}(\tilde{\mathbf{X}}_{j,i}) - m_{n,1}(\tilde{\mathbf{X}}_{j,i}; \boldsymbol{\Theta}_t)) | \Theta^{(1)}_{i,t} = 0] \\
&= \frac{\gamma_n}{n} \cdot Cov_{\Theta^{(1)}_{i,t}=0}\left(\{\widetilde{m}(\mathbf{X}_i) - \widetilde{m}(\tilde{\mathbf{X}}_{j,i})\}; \{(\widetilde{m}(\tilde{\mathbf{X}}_{j,i}) - m_{n,1}(\tilde{\mathbf{X}}_{j,i}; \boldsymbol{\Theta}_t))\}\right) \\
&=: \frac{\gamma_n}{n} \cdot \xi^{(j)}_{n,i}(\boldsymbol{\Theta}_{\boldsymbol{t}}) \qquad (35)
\end{aligned}
$$

The second equality follows from the law of total expectation and the independence of ϵ_i and $\tilde{\Delta}^{(j)}_{n,i}$ under the event that $\Theta^{(1)}_{i,t} = 0$, i.e. that the i-th observation has not been selected during training. The third equality follows from equation (34). The second last equality follows from the fact that $\mathbb{E}[\widetilde{m}(\mathbf{X}_i) - \widetilde{m}(\tilde{\mathbf{X}}_{j,i})] = \mathbb{E}[\widetilde{m}(\mathbf{X}_i)] - \mathbb{E}[\widetilde{m}(\tilde{\mathbf{X}}_{j,i})] = 0$, since $\widetilde{m}(\mathbf{X}_i) \overset{d}{=} \widetilde{m}(\tilde{\mathbf{X}}_{j,i})$. Finally, we can now obtain

$$
\begin{aligned}
&\mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i^{\pi_{j,t}}))(\widetilde{m}(\mathbf{X}_i^{\pi_{j,t}}) - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}; \boldsymbol{\Theta}_t))\mathbb{1}\{\mathbf{X}_i \in \mathcal{D}^{-(t)}_{n,\mathbf{X}}\}] \\
&= \mathbb{E}[(Y_i - \widetilde{m}(\tilde{\mathbf{X}}_{j,i}))(\widetilde{m}(\tilde{\mathbf{X}}_{j,i}) - m_{n,1}(\tilde{\mathbf{X}}_{j,i}; \boldsymbol{\Theta}_t))\mathbb{1}\{\Theta^{(1)}_{i,t} = 0\}] \\
&= \mathbb{E}[\tilde{\epsilon}_{i,j} \cdot \tilde{\Delta}^{(j)}_{n,i}(\boldsymbol{\Theta}_t) \cdot \mathbb{1}\{\Theta^{(1)}_{i,t} = 0\}] \\
&= \frac{\gamma_n}{n} \cdot \xi^{(j)}_{n,i}(\boldsymbol{\Theta}_{\boldsymbol{t}}) \qquad (36)
\end{aligned}
$$

In the second equality, we used (31), while the last equality follows from applying equation (35).
Using the results from (32), (33) and (36), one can now obtain:

$$
\begin{aligned}
\mathbb{E}[(Y_i - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}; \Theta_t))^2 \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}^{-(t)}_{n,\mathbf{X}}\}] &= \mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i^{\pi_{j,t}}))^2 \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}^{-(t)}_{n,\mathbf{X}}\}] \\
&\quad + \mathbb{E}[(\widetilde{m}(\mathbf{X}_i^{\pi_{j,t}}) - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}; \Theta_t))^2 \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}^{-(t)}_{n,\mathbf{X}}\}] \\
&\quad + 2\mathbb{E}[\tilde{\epsilon}_{i,j} \cdot \tilde{\Delta}^{(j)}_{n,i}(\Theta_t) \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}^{-(t)}_{n,\mathbf{X}}\}] \\
&= \frac{n - a_n}{n} \mathbb{E}[(Y_i - \widetilde{m}(\tilde{\mathbf{X}}_{j,i}))^2] + C_{n,i,t} + \frac{2\gamma_n}{n} \cdot \xi^{(j)}_{n,i}(\boldsymbol{\Theta}_t) \qquad (37)
\end{aligned}
$$

Finally, using (30) and (37) together with (28), we obtain

$$
\begin{aligned}
\mathbb{E}[I^{(OOB)}_{n,M}(j)] &= \frac{1}{M\gamma_n} \sum_{t=1}^{M} \sum_{i=1}^{n} \mathbb{E}[\{(Y_i - m_{n,1}(\mathbf{X}_i^{\pi_{j,t}}; \boldsymbol{\Theta}_t))^2 - (Y_i - m_{n,1}(\mathbf{X}_i; \boldsymbol{\Theta}_t))^2\} \mathbb{1}\{\mathbf{X}_i \in \mathcal{D}^{-(t)}_{n,\mathbf{X}}\}] \\
&= \frac{1}{M\gamma_n} \sum_{t=1}^{M} \sum_{i=1}^{n} \left\{ \frac{n - a_n}{n} \{\mathbb{E}[(Y_i - \widetilde{m}(\tilde{\mathbf{X}}_{j,i}))^2] - \mathbb{E}[(Y_i - \widetilde{m}(\mathbf{X}_i))^2]\} + C_{n,i,t} - C_{n,i,t} + \frac{2\gamma_n}{n} \cdot \xi^{(j)}_{n,i}(\boldsymbol{\Theta}_t) \right\} \\
&= \frac{n - a_n}{\gamma_n} \left\{ \mathbb{E}[(Y_1 - \widetilde{m}(\tilde{\mathbf{X}}_{j,1}))^2] - \mathbb{E}[(Y_1 - \widetilde{m}(\mathbf{X}_1))^2] \right\} + \frac{2}{\gamma_n} \sum_{i=1}^{n} \left(\frac{1}{M} \sum_{t=1}^{M} \frac{\gamma_n}{n} \xi^{(j)}_{n,i}(\boldsymbol{\Theta}_t) \right) \\
&= \mathbb{E}[(Y_1 - \widetilde{m}(\tilde{\mathbf{X}}_{j,1}))^2] - \mathbb{E}[(Y_1 - \widetilde{m}(\mathbf{X}_1))^2] + 2 \cdot \left(\frac{1}{M} \sum_{t=1}^{M} \xi^{(j)}_{n,1}(\boldsymbol{\Theta}_{\boldsymbol{t}}) \right) \qquad (38)
\end{aligned}
$$

where the second last equality follows from the identical distribution (in i) of the sequence $\{Y_i - m(\tilde{\mathbf{X}}_{j,i})\}_{i=1}^n$, respectively $\{Y_i - m(\mathbf{X}_i)\}_{i=1}^n$. The last equality follows from the identical distribution of the sequence $\{\xi_{n,i}^{(j)}(\boldsymbol{\Theta}_t)\}_{i=1}^n$.

Without loss of generality, assume that the first $1 \leq s \leq p$ features are informative, i.e. $\mathcal{S} = \{1, \ldots, s\}$ and define $\mathbf{X}_{i;\mathcal{S}} = [X_{1,i}, X_{2,i}, \ldots, X_{s,i}]^\top \in \mathbb{R}^s$, the i-th random vector reduced to informative features characterized by $\mathcal{S}$. Similarly, let $\tilde{\mathbf{X}}_{j,i;\mathcal{S}}$ be the reduced random vector of $\tilde{\mathbf{X}}_{j,i}$, in which the j-th position is substituted by Z_j, with $1 \leq j \leq s$.

We distinguish between two cases: First, let $j \in \mathcal{S}^C = \{1, \ldots, p\} \setminus \mathcal{S}$. Under this scenario, we know that $\widetilde{m}(\tilde{\mathbf{X}}_{j,1}) = \widetilde{m}(\mathbf{X}_1) = m(\mathbf{X}_{1;\mathcal{S}})$. Hence, we have

$$\begin{aligned}
\xi_{n,1}^{(j)}(\boldsymbol{\Theta}_t) &= Cov_{\Theta_{1,t}^{(1)}=0}\Big(\{\widetilde{m}(\mathbf{X}_1) - \widetilde{m}(\tilde{\mathbf{X}}_{j,1})\}; \{(\widetilde{m}(\tilde{\mathbf{X}}_{j,1}) - m_{n,1}(\tilde{\mathbf{X}}_{j,1}; \boldsymbol{\Theta}_t)) \cdot \mathbb{1}\{\Theta_{1,t}^{(1)} = 0\}\}\Big) \\
&= Cov_{\Theta_{1,t}^{(1)}=0}\Big(0; \{(\widetilde{m}(\tilde{\mathbf{X}}_{j,1}) - m_{n,1}(\tilde{\mathbf{X}}_{j,1}; \boldsymbol{\Theta}_t)) \cdot \mathbb{1}\{\Theta_{1,t}^{(1)} = 0\}\Big) = 0. \qquad (39)
\end{aligned}$$

Therefore, it immediately follows by applying (38) and (39) that

$$\begin{aligned}
\mathbb{E}[I_{n,M}^{(OOB)}(j)] &= \mathbb{E}[(Y_1 - \widetilde{m}(\tilde{\mathbf{X}}_{j,1}))^2] - \mathbb{E}[(Y_1 - \widetilde{m}(\mathbf{X}_1))^2] \\
&= \mathbb{E}[(Y_1 - m(\mathbf{X}_{1;\mathcal{S}}))^2] - \mathbb{E}[(Y_1 - m(\mathbf{X}_{1;\mathcal{S}}))^2] \\
&= 0 = I(j). \qquad (40)
\end{aligned}$$

Secondly, let $j \in \mathcal{S}$ be informative. Then notice that

$$\begin{aligned}
\frac{\gamma_n}{n}\frac{1}{M}\sum_{t=1}^{M}\xi_{n,1}^{(j)}(\boldsymbol{\Theta}_t) &= \frac{\gamma_n}{n}\frac{1}{M}\sum_{t=1}^{M} Cov_{\Theta_{i,t}^{(1)}=0}\Big(\{\widetilde{m}(\mathbf{X}_1) - \widetilde{m}(\tilde{\mathbf{X}}_{j,1})\}; \{\widetilde{m}(\tilde{\mathbf{X}}_{j,1}) - m_{n,1}(\tilde{\mathbf{X}}_{j,1}; \boldsymbol{\Theta}_t))\}\Big) \\
&= \frac{\gamma_n}{n}\frac{1}{M}\sum_{t=1}^{M}\mathbb{E}[(m(\mathbf{X}_{1;\mathcal{S}}) - m(\tilde{\mathbf{X}}_{j,1;\mathcal{S}})) \cdot (\widetilde{m}(\tilde{\mathbf{X}}_{j,1}) - m_{n,1}(\tilde{\mathbf{X}}_{j,1}; \boldsymbol{\Theta}_t))|\Theta_{1,t}^{(1)} = 0] \\
&= \frac{1}{M}\sum_{t=1}^{M}\mathbb{E}[(m(\mathbf{X}_{1;\mathcal{S}}) - m(\tilde{\mathbf{X}}_{j,1;\mathcal{S}})) \cdot (\widetilde{m}(\tilde{\mathbf{X}}_{j,1}) - m_{n,1}(\tilde{\mathbf{X}}_{j,1}; \boldsymbol{\Theta}_t)) \cdot \mathbb{1}\{\Theta_{1,t}^{(1)} = 0\}] \\
&= \mathbb{E}\left[(m(\mathbf{X}_{1;\mathcal{S}}) - m(\tilde{\mathbf{X}}_{j,1;\mathcal{S}})) \cdot \frac{Z_1(M)}{M} \cdot \Big(m(\tilde{\mathbf{X}}_{j,1}) - m_{n,M}^{OOB}(\tilde{\mathbf{X}}_{j,1})\Big)\right], \qquad (41)
\end{aligned}$$

where $Z_1(M) = \sum_{t=1}^{M}\mathbb{1}\{\Theta_{1,t}^{(1)} = 0\} = \sum_{t=1}^{M}\mathbb{1}\{\mathbf{X}_1 \text{ has not been selected under } \boldsymbol{\Theta}_t\}$ is the number of times the first observation has not been selected during the sampling procedure and $m_{n,M}^{OOB}(\tilde{\mathbf{X}}_{j,1}) = \frac{1}{Z_1(M)}\sum_{t=1}^{M} m_{n,1}(\tilde{\mathbf{X}}_{j,1}; \boldsymbol{\Theta}_t) \cdot \mathbb{1}\{\Theta_{i,t}^{(1)} = 0\}$. Due to assumption $(A4)$, we can deduce that

$$|m(\mathbf{X}_{1;\mathcal{S}}) - m(\tilde{\mathbf{X}}_{1,j;\mathcal{S}})| \leq 2K < \infty \qquad (42)$$

On the other hand, we observe the following bound:

$$\begin{aligned}
\left|\frac{Z_1(M)}{M} \cdot \Big(\widetilde{m}(\tilde{\mathbf{X}}_{1,j}) - m_{n,M}^{OOB}(\tilde{\mathbf{X}}_{1,j})\Big)\right| &\leq K + \sum_{\ell=1}^{n} W_{n,\ell}(\tilde{\mathbf{X}}_{1,j}; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M) \cdot |Y_\ell| \\
&\leq K + \max_{1 \leq \ell \leq n}|Y_\ell| =: K + f_n, \qquad (43)
\end{aligned}$$

where $W_{n,\ell}(\cdot;\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_M) = \frac{1}{M}\sum_{t=1}^{M} W_{n,\ell}(\cdot;\boldsymbol{\Theta}_t)$. Hence, we can deduce by applying (42) and (43) that

$$|f_{n,M}| := \left|(m(\mathbf{X}_{1;\mathcal{S}}) - m(\tilde{\mathbf{X}}_{j,1;\mathcal{S}})) \cdot \frac{Z_1(M)}{M} \cdot \left(\widetilde{m}(\tilde{\mathbf{X}}_{j,1}) - m_{n,M}^{OOB}(\tilde{\mathbf{X}}_{1,j})\right)\right| \leq 2K(K+f_n) =: g_n, \tag{44}$$

i.e. g_n is a finite upper bound for $|f_{n,M}|$, independent of M such that $\mathbb{E}_{\boldsymbol{\Theta}}[|g_n|] = 2K \cdot (K+f_n) < \infty$, where $f_n := \max_{1\leq\ell\leq n} |Y_\ell|$. Applying Lebesgue's dominated convergence theorem while using Proposition 1 under the sampling without replacement scheme with $c_n = 1 - a_n/n = \gamma_n/n$ and using $Z_1(M)/M \longrightarrow c_n$ as $M \to \infty$ due to (19), we obtain

$$\begin{aligned}
&\lim_{M\to\infty} \mathbb{E}\left[(m(\mathbf{X}_{1;\mathcal{S}}) - m(\tilde{\mathbf{X}}_{j,1;\mathcal{S}})) \cdot \frac{Z_1(M)}{M}\left(\widetilde{m}(\tilde{\mathbf{X}}_{j,1}) - m_{n,M}^{OOB}(\tilde{\mathbf{X}}_{j,1})\right)\right] \\
&= \frac{\gamma_n}{n}\mathbb{E}[(m(\mathbf{X}_{1;\mathcal{S}}) - m(\tilde{\mathbf{X}}_{j,1;\mathcal{S}}))(\widetilde{m}(\tilde{\mathbf{X}}_{j,1}) - m_n^{OOB}(\tilde{\mathbf{X}}_{j,1}))] \\
&=: \frac{\gamma_n}{n} J_n
\end{aligned} \tag{45}$$

Note that J_n can be bounded the following way using the Cauchy-Schwarz inequality:

$$J_n \leq |J_n| \leq \sqrt{\mathbb{E}[|m(\mathbf{X}_{1;\mathcal{S}}) - m(\tilde{\mathbf{X}}_{j,1;\mathcal{S}})|^2]}\sqrt{\mathbb{E}[|\widetilde{m}(\tilde{\mathbf{X}}_{j,1}) - m_n^{OOB}(\tilde{\mathbf{X}}_{j,1})|^2]} \tag{46}$$

Since $J_n \geq -|J_n|$ and due to assumption $(A5)$, we can deduce that $\lim_{n\to\infty} J_n = 0$. Note that the L_2 consistency of the Random Forest estimate m_n^{OOB} for Out-of-Bag samples follows by $(A5)$ and a Corollary given in [15]. Finally, we can conclude with (41) and (45) that

$$\lim_{n\to\infty}\lim_{M\to\infty} \frac{1}{M}\sum_{t=1}^{M} \xi_{n,1}^{(j)}(\boldsymbol{\Theta}_t) = \lim_{n\to\infty} \frac{n}{\gamma_n}\frac{\gamma_n}{n} J_n = \lim_{n\to\infty} J_n = 0, \tag{47}$$

which completes the proof.

□

In the sequel, we will shortly deliver proofs for the following claims, that have been mentioned in the main article: (i) We argued that a variable $j \in \{1,\ldots,p\}$ is *important*, if the partial derivate of $\widetilde{m}(\mathbf{x})$ w.r.t. x_j vanishes, i.e. we claimed the equivalence of both definitions (2) and (3) mentioned in the article. (ii) We claimed that the assumptions given in [3] can replace $(A3) - (A5)$. (iii) We claimed that the theoretical cut criterion $L^{(k)}(j,z)$ is independent of the residual noise σ^2.

Proof of (i). Suppose that being important is defined through (2) and assume without loss of generality, that the first $s \leq p$ features are important, i.e. $\widetilde{m}(\mathbf{x}) = m(\mathbf{x}_{\mathcal{S}})$, where $\mathbf{x}_{\mathcal{S}} = [x_1,\ldots,x_s]^\top \in \mathbb{R}^s$. Then it follows immediately that $\frac{\partial\widetilde{m}(\mathbf{x})}{\partial x_j} = 0$ for all $j \in \{1,\ldots,p\}\setminus\mathcal{S}$, since $\widetilde{m}(\mathbf{x}) = m(\mathbf{x}_{\mathcal{S}})$ does not depend on j. Hence, variable $j \in \{1,\ldots,p\}\setminus\mathcal{S}$ is unimportant according to the definition given in (3).

For the other direction, define the set $\mathcal{C} := \{k \in \{1,\ldots,p\} : \frac{\partial\widetilde{m}(\mathbf{x})}{\partial x_k} \neq 0\}$ and suppose that $j \in \{1,\ldots,p\}$ is informative in the sense that $j \in \mathcal{C}$. Then, let $\mathbf{a} \in \mathbb{R}^p$ be fixed but arbitrary. Using the multivariate Taylor expansion of $\widetilde{m}$ at $\mathbf{a}$, one has

$$\begin{aligned}
\widetilde{m}(\mathbf{x}) &\approx \widetilde{m}(\mathbf{a}) + \nabla\widetilde{m}(\mathbf{a})^\top(\mathbf{x}-\mathbf{a}) \\
&= \widetilde{m}(\mathbf{a}) + \sum_{s\in\mathcal{C}} \widetilde{m}_s'(a_s)(x_s - a_s) =: m(\mathbf{x}_{\mathcal{C}})
\end{aligned} \tag{48}$$

which yields to $\mathcal{S} = \mathcal{C}$, i.e. the function $\widetilde{m}$ can be reduced to a function of potentially lower dimension, since $\mathbf{a}$ is chosen arbitrary and (48) holds for any fixed $\mathbf{a}$.

□

Proof of (ii). Recalling some of the assumptions given in [3] in order to establish L_2 consistency, we have

1. $\widetilde{m}(\mathbf{x}) = \sum_{k=1}^{p} \widetilde{m}_k(x_k)$, where $\{m_k(x_k)\}_{k=1}^{p}$ is a sequence of univariate and continuous functions.
2. The feature vector $\mathbf{X} = [X_1, \ldots, X_p]^\top \in \mathbb{R}^p$ is assumed to be uniformly distributed over $[0,1]^p$.
3. The residuals are assumed to be centered Gaussian with variance $\sigma^2 \in (0, \infty)$, independent of $\mathbf{X}$.
4. Sampling is restricted to sampling without replacement such that $a_n \to \infty$, $t_n \to \infty$ and $\frac{t_n \cdot (\log(a_n))^9}{a_n} \to 0$ as $n \to \infty$.

Now, since $\widetilde{m}_k$ is continuous for every $k \in \{1, \ldots, p\}$ according to 1., it immediately follows that $\widetilde{m}$ resp. $|\widetilde{m}(\mathbf{x})|$ is continuous. Hence, since $[0,1]^p$ as the support of $\mathbf{X}$ is compact, so is the set $\{\widetilde{m}(\mathbf{x}) : \mathbf{x} \in [0,1]^p\}$, which then yields to $\sup_{\mathbf{x} \in [0,1]^p} \widetilde{m}(\mathbf{x}) = \max_{\mathbf{x} \in [0,1]^p} \widetilde{m}(\mathbf{x}) = K < \infty$. This is nothing else than assumption $(A4)$. Furthermore, we have from 2. that $\mathbf{X} \sim Unif([0,1]^p)$, which yields to $f_{\mathbf{X}}(x_1, \ldots x_p) = \mathbb{1}\{\mathbf{x} \in [0,1]^p\} = \prod_{j=1}^{p} \mathbb{1}\{X_j \in [0,1]\} = \prod_{j=1}^{p} f_{Unif(0,1)}(x_j)$, i.e. the multivariate density decomposes into the product of univariate densities. Therefore, the sequence of random variables $\{X_j\}_{j=1}^{p}$ is mutual independent. Hence, assumption $(A3)$ follows. Assuming that the residuals are centered Gaussian with finite variance σ^2 as given in 3. is nothing else than the specification of our assumption that $\mathbb{E}[\epsilon] = 0$ and $Var(\epsilon) \in (0, \infty)$ by imposing explicitly the Gaussian distribution. Assumption $(A5)$ then immediately follows by using Theorem 1 in [3] and the assumptions 1 - 4. Assumptions $(A1)$ and $(A2)$ are not required in [3], and hence, they do not prohibit us to use Theorem 1 in [3]. Therefore, they can be taken over additionally.

□

Proof of (iii). Consider the theoretical cut criterion $L^{(k)}(j,z)$ at level $1 \leq k \leq \lceil \log_2(t_n) \rceil + 1$ with $1 \leq \ell \leq 2^{k-1}$. Then we can see that this is independent of σ^2:

$$
\begin{aligned}
L^{(k)}(j,z) &= Var[Y_1 | \mathbf{X}_1 \in A_\ell^{(k)}] - \mathbb{P}[X_{j,1} < z | \mathbf{X}_1 \in A_\ell^{(k)}], \cdot Var[Y_1 | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} < z] \\
&\quad - \mathbb{P}[X_{j,1} \geq z | \mathbf{X}_1 \in A_\ell^{(k)}] \cdot Var[Y_1 | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} \geq z] \\
&= Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}] + Var[\epsilon_1 | \mathbf{X}_1 \in A_\ell^{(k)}] - \mathbb{P}[X_{j,1} < z | \mathbf{X}_1 \in A_\ell^{(k)}] \cdot \Big\{ Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} < z] + \\
&\quad Var[\epsilon_1 | \mathbf{X}_1 \in A_\ell^{(k)} X_{j,1} < z] \Big\} - \mathbb{P}[X_{j,1} \geq z | \mathbf{X}_1 \in A_\ell^{(k)}] \cdot \Big\{ Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} \geq z] + \\
&\quad Var[\epsilon_1 | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} \geq z] \Big\} \\
&= Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}] + \sigma^2 - \mathbb{P}[X_{j,1} < z | \mathbf{X}_1 \in A_\ell^{(k)}] \cdot \Big\{ Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} < z] + \sigma^2 \Big\} - \\
&\quad \mathbb{P}[X_{j,1} \geq z | \mathbf{X}_1 \in A_\ell^{(k)}] \cdot \Big\{ Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} \geq z] + \sigma^2 \Big\} \\
&= Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}] - \mathbb{P}[X_{j,1} < z | \mathbf{X}_1 \in A_\ell^{(k)}] \cdot Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} < z] - \\
&\quad \mathbb{P}[X_{j,1} \geq z | \mathbf{X}_1 \in A_\ell^{(k)}] \cdot Var[\widetilde{m}(\mathbf{X}_1) | \mathbf{X}_1 \in A_\ell^{(k)}, X_{j,1} \geq z],
\end{aligned}
$$

where the third equality follows from the independence of ϵ_1 and $\mathbf{X}_1$. □

References

References

[1] S. Wager, S. Athey, Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, Journal of the American Statistical Association 113 (523) (2018) 1228–1242.

[2] S. Wager, T. Hastie, B. Efron, Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife, Journal of Machine Learning Research 15 (1) (2014) 1625–1651.

[3] E. Scornet, G. Biau, J.-P. Vert, Consistency of Random Forests, The Annals of Statistics 43 (4) (2015) 1716–1741.

[4] L. Mentch, G. Hooker, Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests, The Journal of Machine Learning Research 17 (1) (2016) 841–881.

[5] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3 (Mar) (2003) 1157–1182.

[6] C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: Illustrations, sources and a solution, BMC Bioinformatics 8 (1) (2007) 25.

[7] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, BMC bioinformatics 9 (1) (2008) 307.

[8] K. J. Archer, R. V. Kimes, Empirical characterization of random forest variable importance measures, Computational Statistics & Data Analysis 52 (4) (2008) 2249–2260.

[9] K. K. Nicodemus, J. D. Malley, Predictor correlation impacts machine learning algorithms: implications for genomic studies, Bioinformatics 25 (15) (2009) 1884–1890.

[10] K. K. Nicodemus, J. D. Malley, C. Strobl, A. Ziegler, The behaviour of random forest permutation-based variable importance measures under predictor correlation, BMC Bioinformatics 11 (1) (2010) 110.

[11] K. K. Nicodemus, Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures, Briefings in Bioinformatics 12 (4) (2011) 369–373.

[12] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (10) (2010) 1340–1347.

[13] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Variable selection using Random Forests, Pattern Recognition Letters 31 (14) (2010) 2225–2236.

[14] B. Gregorutti, B. Michel, P. Saint-Pierre, Correlation and variable importance in random forests, Statistics and Computing 27 (3) (2017) 659–678.

[15] B. Ramosaj, M. Pauly, Consistent estimation of residual variance with random forest Out-Of-Bag errors, Statistics & Probability Letters 151 (2019) 49–57.

[16] R. Zhu, D. Zeng, M. R. Kosorok, Reinforcement Learning Trees, Journal of the American Statistical Association 110 (512) (2015) 1770–1784.

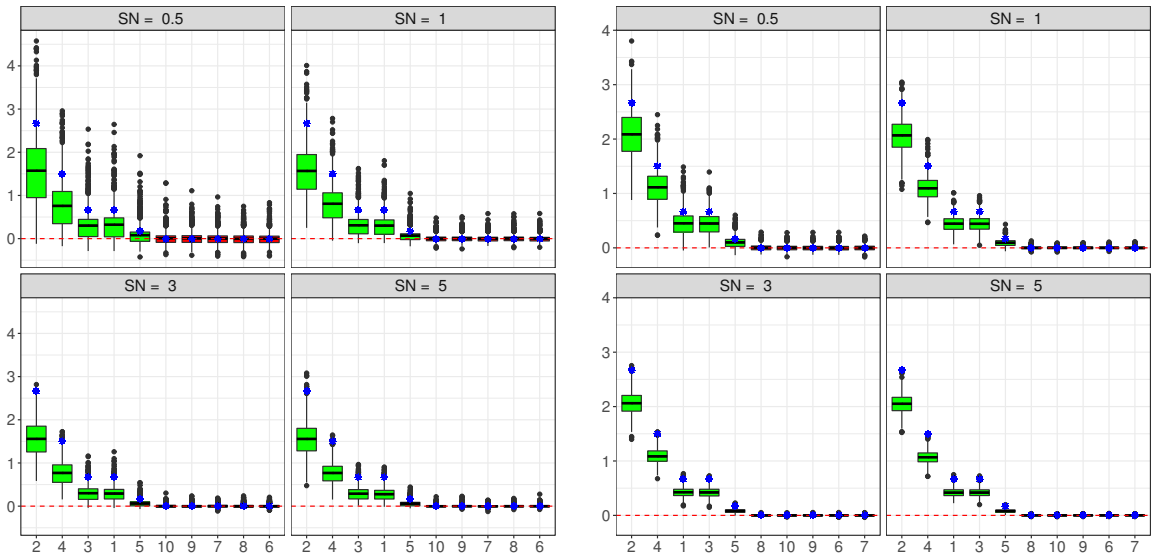
[17] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd Edition, Springer, New York, NY, 2009.

Supplementary Material to: Asymptotic Unbiasedness of Variable Importance Measures in Random Forest Models.

Burim Ramosaj*, Markus Pauly

Faculty of Statistics
Institute of Mathematical Statistics and Applications in Industry
Technical University of Dortmund
44227 Dortmund, Germany

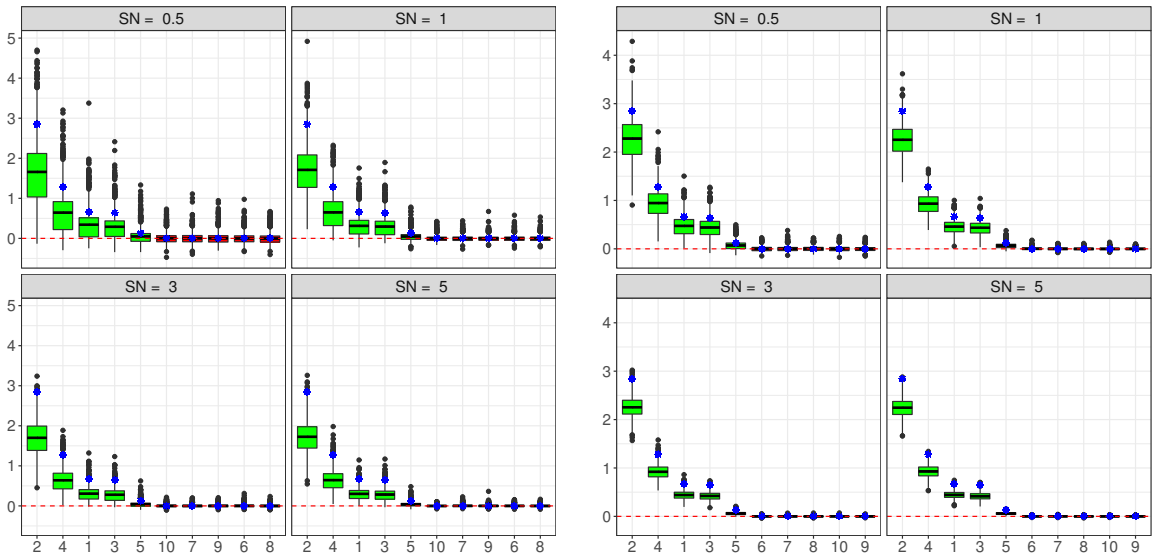
1. Results for $p < n$ Problems



(a) $n = 100$ (b) $n = 500$

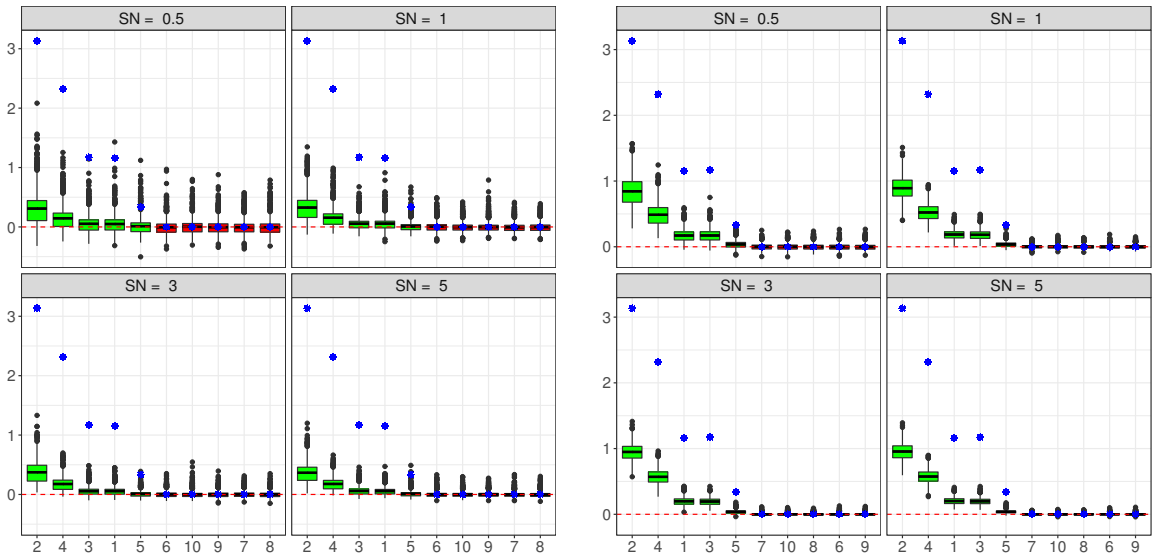
Figure 1: Simulation results for the permutation importance with various signal-to-noise ratios under a **linear model** as described in (1) of the main article using $MC = 1,000$ Monte-Carlo iterations with a sample size of (a) $n = 100$ and (b) $n = 500$. The solid lines refer to the empirical mean and $\star$ to its expectation.

* *Corresponding Author:* Burim Ramosaj
Email address: burim.ramosaj@tu-dortmund.de



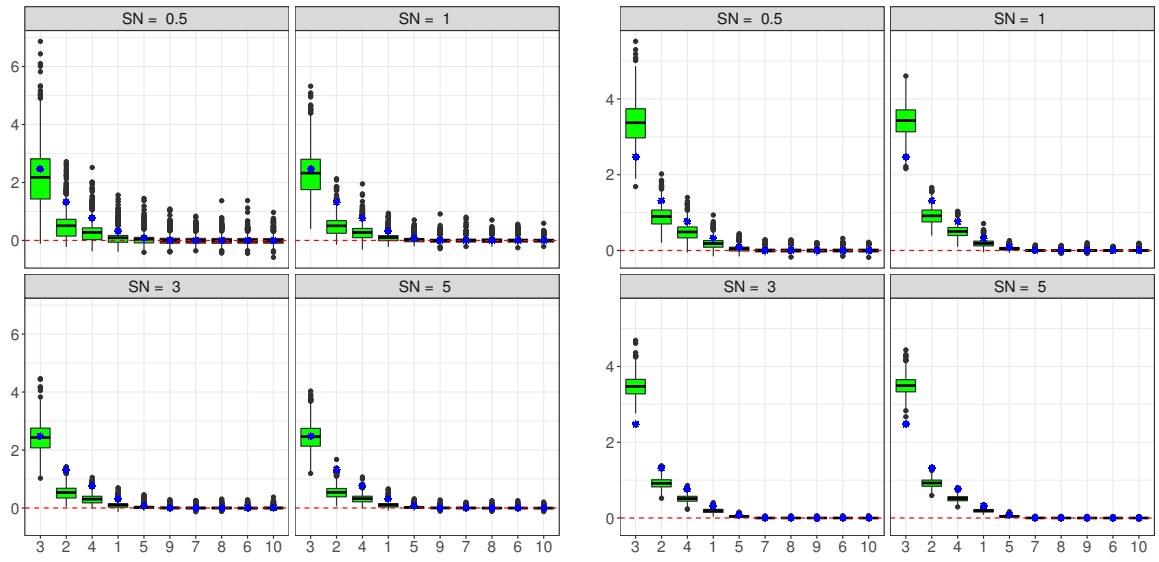
(a) $n = 100$ (b) $n = 500$

Figure 2: Simulation results for the permutation importance with various signal-to-noise ratios under a **polynomial model** as described in (1) of the main article using $MC = 1,000$ Monte-Carlo iterations with a sample size of (a) $n = 100$ and (b) $n = 500$. The solid lines refer to the empirical mean and $\star$ to its expectation.



(a) $n = 100$ (b) $n = 500$

Figure 3: Simulation results for the permutation importance with various signal-to-noise ratios under a **trigonometric model** as described in (1) of the main article using $MC = 1,000$ Monte-Carlo iterations with a sample size of (a) $n = 100$ and (b) $n = 500$. The solid lines refer to the empirical mean and $\star$ to a Monte-Carlo approximation of its expectation.



(a) $n = 100$ (b) $n = 500$

Figure 4: Simulation results for the permutation importance with various signal-to-noise ratios under a **non-continuous model** as described in (1) of the main article using $MC = 1,000$ Monte-Carlo iterations with a sample size of (a) $n = 100$ and (b) $n = 500$. The solid lines refer to the empirical mean and $\star$ to a Monte-Carlo approximation of its expectation.

		$n = 50$				$n = 100$			
	$SN =$	0.5	1	3	5	0.5	1	3	5
Model	linear	0.189	0.364	0.807	1.001	0.248	0.509	1.181	1.528
	polynomial	0.184	0.362	0.807	1.033	0.243	0.5	1.197	1.594
	trigonometric	0.1	0.1	0.1	0.1	0.061	0.064	0.102	0.119
	non-continuous	0.158	0.309	0.726	0.936	0.204	0.473	1.152	1.523
		$n = 500$				$n = 1,000$			
	$SN =$	0.5	1	3	5	0.5	1	3	5
Model	linear	0.365	0.743	1.937	2.781	0.400	0.808	2.178	3.240
	polynomial	0.365	0.754	1.995	2.919	0.395	0.812	2.246	3.400
	trigonometric	0.098	0.215	0.451	0.549	0.170	0.335	0.7	0.862
	non-continuous	0.357	0.759	2.094	3.153	0.395	0.829	2.376	3.714

Table 1: Estimator $\widehat{SN}_n$ as given in equation (14) of the main article under various sample sizes and signal-to-noise ratios using $MC = 1,000$ Monte-Carlo iterates for the $p < n$ regression problem.

Table 1 refers to the estimator $\widehat{SN}_n$ of SN as proposed in the main article under various sample sizes. One can see that $\widehat{SN}_n$ tends to be smaller than SN, but slowly moves to SN for an increased sample size.

2. Results for $p > n$ Problems

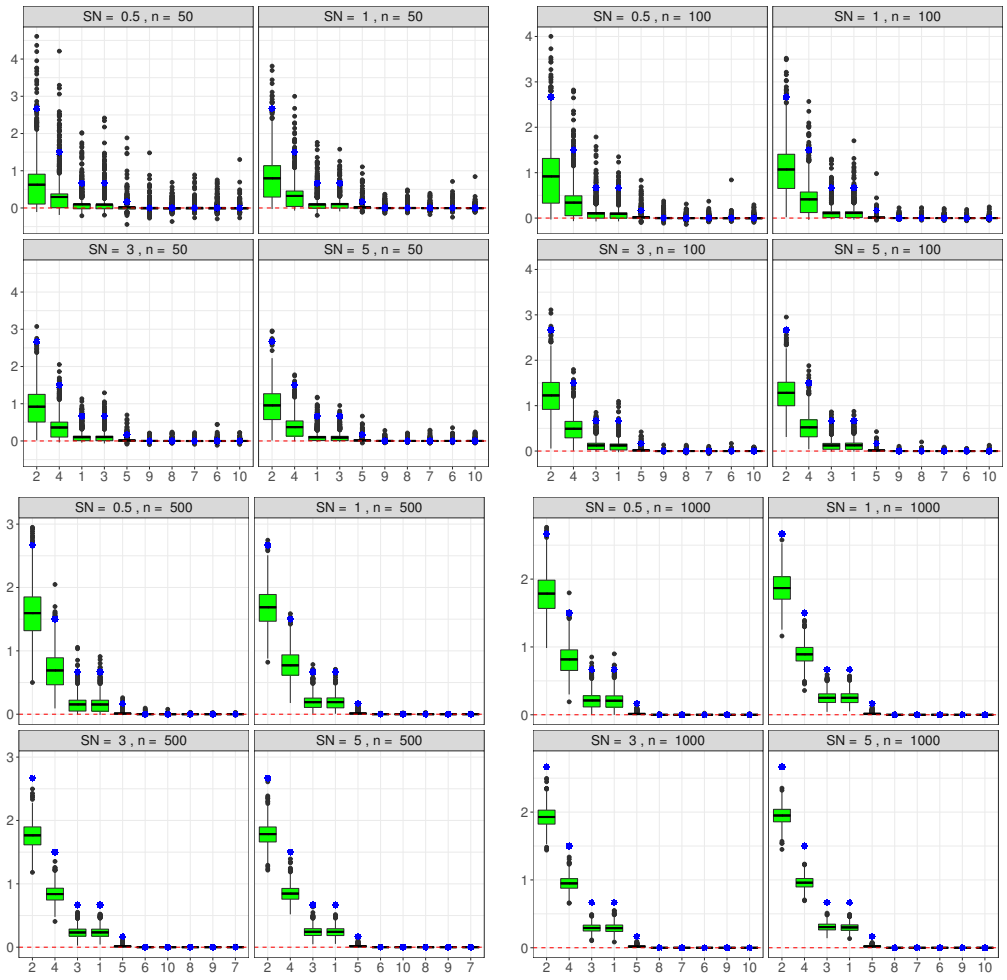


Figure 5: Simulation results for the permutation importance with various signal-to-noise ratios under a **linear model** as described in (1) of the main article using $MC = 1,000$ Monte-Carlo iterations under the **high-dimensional** setting. The solid lines refer to the empirical mean and $\star$ to its expectation.

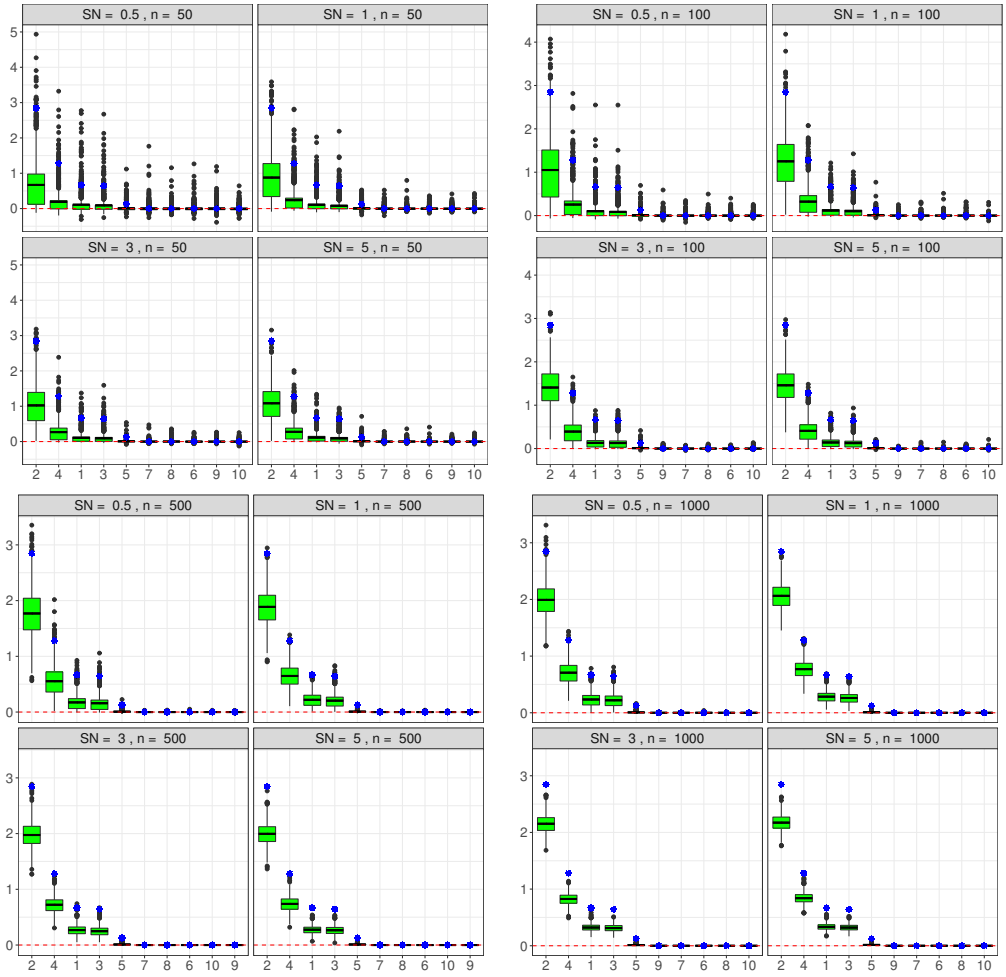


Figure 6: Simulation results for the permutation importance with various signal-to-noise ratios under a **polynomial model** as described in (1) of the main article using $MC = 1,000$ Monte-Carlo iterations under the **high-dimensional** setting. The solid lines refer to the empirical mean and $\star$ to its expectation.

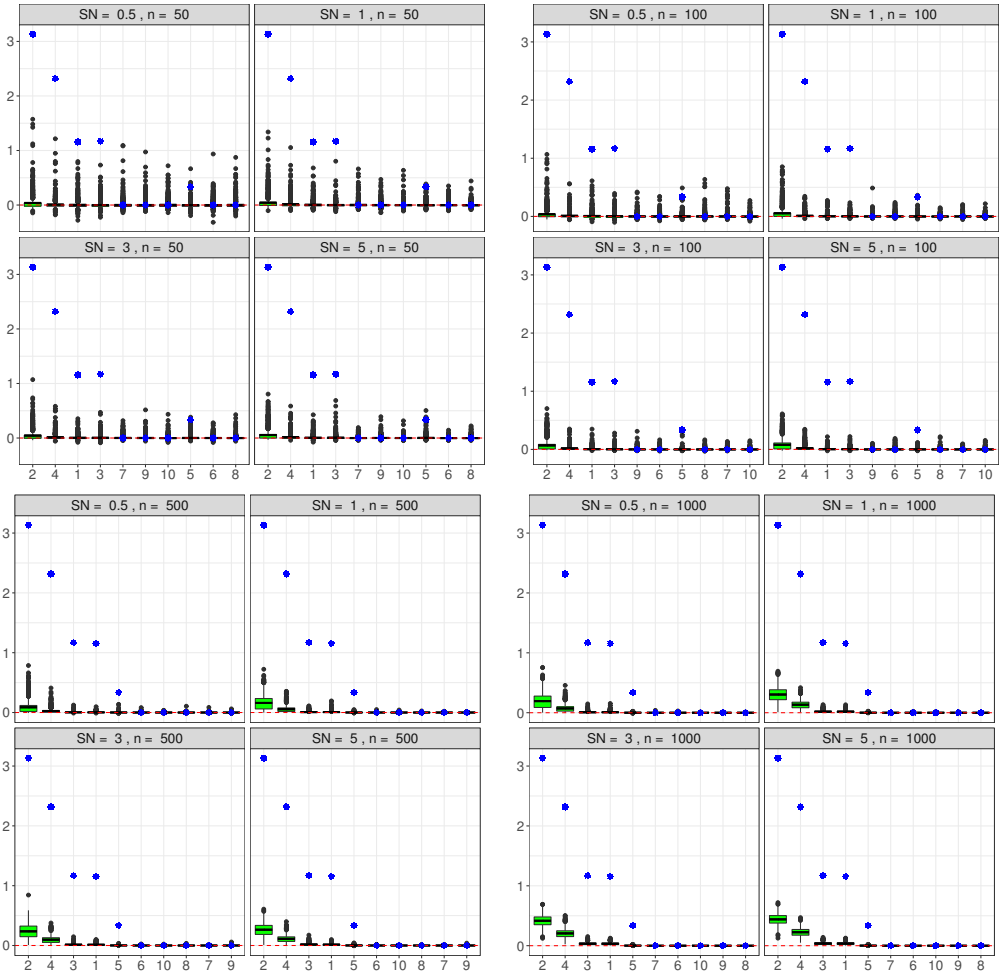


Figure 7: Simulation results for the permutation importance with various signal-to-noise ratios under a **trigonometric model** as described in (1) of the main article using $MC = 1,000$ Monte-Carlo iterations under the **high-dimensional** setting. The solid lines refer to the empirical mean and $\star$ to a Monte-Carlo approximation of its expectation.

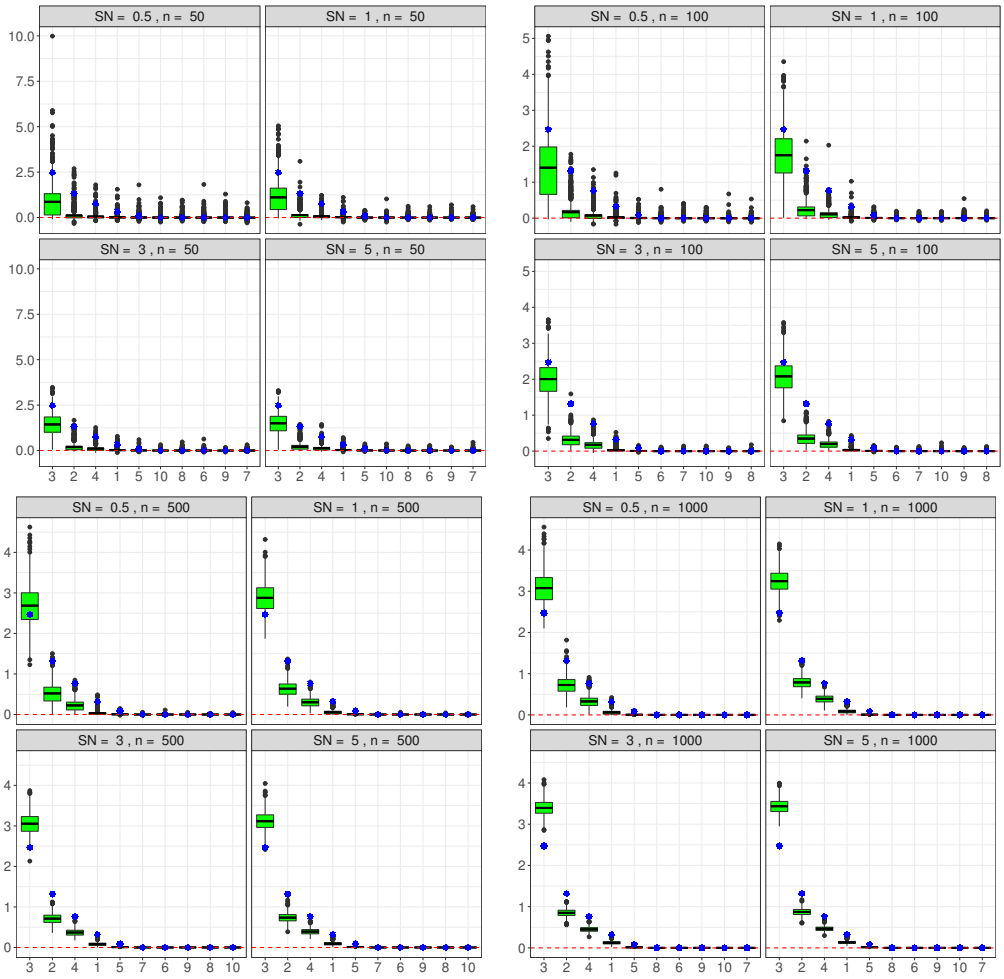


Figure 8: Simulation results for the permutation importance with various signal-to-noise ratios under a **non-continuous model** as described in (1) of the main article using $MC = 1,000$ Monte-Carlo iterations under the **high-dimensional** setting. The solid lines refer to the empirical mean and $\star$ to a Monte-Carlo approximation of its expectation.

List of Figures

Bibliography

L. Amro and M. Pauly. Permuting incomplete paired data: a novel exact and asymptotic correct randomization test. *Journal of Statistical Computation and Simulation*, 87(6): 1148–1159, 2017.

B. Andreas, S. Werner, and S. Yi. *Degrees of Boosting: A Study of Loss Functions for Classification and Class Probability Estimation.* Manuscript, 2005.

E. Angelini, G. di Tollo, and A. Roli. A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance*, 48(4):733–755, 2008.

P. Barbe and P. Bertail. *The Weighted Bootstrap*, volume 98. Springer Science & Business Media, New York, NY, 1995.

G. Biau. Analysis of a Random Forests Model. *Journal of Machine Learning Research*, 13 (Apr):1063–1095, 2012.

G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518, 2010.

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

G. Biau, L. Devroye, and G. Lugosi. Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research*, 9(Sep):2015–2033, 2008.

J. J. Brand. *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets.* PhD-Thesis, 1999.

L. Breiman. Arcing classifiers. *Annals of Statistics*, 26:123–40, 1996a.

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

L. Breiman. Prediction Games and Arcing Algorithms. *Neural Computation*, 11(7):1493–1517, 1999.

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

L. Breiman, J. Friedman, R. Ohlsen, and C. Stone. *Classification and Regression Trees.* Chapman & Hall/CRC, 1984.

P. Bühlmann and T. Hothorn. Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, 22(4):477–505, 2007.

P. Bühlmann and B. Yu. Analyzing Bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

C. Conversano and R. Siciliano. Incremental Tree-Based Missing Data Imputation with Lexicographic Ordering. *Journal of Classification*, 26(3):361–379, 2009.

Y. Deng, C. Chang, M. S. Ido, and Q. Long. Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Scientific Reports*, 6(1):1–10, 2016.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013.

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

L. L. Doove, S. Van Buuren, and E. Dusseldorp. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104, 2014.

B. Efron and R. Tibshirani. Improvements on Cross-Validation: The 632+ Bootstrap Method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, NY, 1994.

C. K. Enders. A Primer on Maximum Likelihood Algorithms available for Use with Missing Data. *Structural Equation Modeling*, 8(1):128–141, 2001.

C. K. Enders. *Applied Missing Data Analysis*. Guilford Press, 2010.

Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

J. Friedman, T. Hastie, and R. Tibshirani. Additive Logistic Regression: A Statistical View of Boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

J. H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, pages 1189–1232, 2001.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63 (1):3–42, 2006.

B. Gregorutti, B. Michel, and P. Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678, 2017.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression* . Springer Science & Business Media, 2006.

P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media, New York, NY, 1992.

A. Hapfelmeier, T. Hothorn, and K. Ulm. Recursive partitioning on incomplete data using surrogate decisions and multiple imputation. *Computational Statistics & Data Analysis*, 56(6):1552–1565, 2012.

T. Hastie, S. Rosset, J. Zhu, and H. Zou. Multi-class AdaBoost. *Statistics and its Interface*, 2(3):349–360, 2009a.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, NY, 2 edition, 2009b.

L. Hoogerheide, J. H. Block, and R. Thurik. Family background variables as instruments for education in income regressions: A Bayesian analysis. *Economics of Education Review*, 31 (5):515–523, 2012.

A. T. Hudak, N. L. Crookston, J. S. Evans, D. E. Hall, and M. J. Falkowski. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment*, 112(5):2232–2245, 2008.

H. Ishwaran. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537, 2007.

H. Jiang, Y. Deng, H.-S. Chen, L. Tao, Q. Sha, J. Chen, C.-J. Tsai, and S. Zhang. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics*, 5(1):81, 2004.

D. Keysers, T. Deselaers, C. Gollan, and H. Ney. Deformation Models for Image Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1422–1435, 2007.

A. Klenke. *Wahrscheinlichkeitstheorie*, volume 2. Springer, 2008.

A. Liaw and M. Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

W.-Y. Loh and Y.-S. Shih. Split Selection Methods for Classification Trees. *Statistica Sinica*, 7(4):815–840, 1997.

X. Long, L. Chen, C. Jiang, L. Zhang, and A. D. N. Initiative. Prediction and classification of Alzheimer disease based on quantification of MRI deformation. *PloS ONE*, 12(3), 2017.

G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts. Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*, pages 431–439, 2013.

K. L. Lunetta, L. B. Hayward, J. Segal, and P. Van Eerdewegh. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*, 5(1):32, 2004.

P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, FL, 2 edition, 1989.

N. Meinshausen. Quantile Regression Forests. *Journal of Machine Learning Research*, 7 (Jun):983–999, 2006.

G. Mendez and S. Lohr. Estimating residual variance in random forest regression. *Computational Statistics & Data Analysis*, 55(11):2937–2950, 2011.

X.-L. Meng. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, pages 538–558, 1994.

L. Mentch and G. Hooker. Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. *The Journal of Machine Learning Research*, 17(1):841–881, 2016.

B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10(1):213, 2009.

J. Præstgaard and J. A. Wellner. Exchangeably Weighted Bootstraps of the General Empirical Process. *The Annals of Probability*, 21(4):2053–2086, 1993.

Y. Qi. *Random forest for Bioinformatics*. Springer, 2012.

J. R. Quinlan. Bagging, Boosting, and C4. 5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, pages 725–730, Portland, OR, 1996.

T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, 27(1):85–96, 2001.

B. Ramosaj and M. Pauly. Asymptotic Unbiasedness of the Permutation Importance Measure in Random Forest Models. *arXiv preprint arXiv:1912.03306*, 2019a.

B. Ramosaj and M. Pauly. Consistent estimation of residual variance with random forest Out-Of-Bag errors. *Statistics & Probability Letters*, 151:49–57, 2019b.

B. Ramosaj and M. Pauly. Predicting missing values: a comparative study on non-parametric approaches for imputation. *Computational Statistics*, 34(4):1741–1764, 2019c.

G. Ridgeway. The gbm Package. *R Foundation for Statistical Computing*, 5(3), 2004.

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

D. B. Rubin. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*, volume 81. John Wiley & Sons, 2004.

W. Rudin. *Real and Complex Analysis*. McGraw-Hill Book Co., 1987.

J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC, London, UK, 1997.

C. Schaffer. Overfitting Avoidance as Bias. *Machine Learning*, 10(2):153–178, 1993.

R. E. Schapire. Explaining AdaBoost. *Empirical Inference*, pages 37–52, 2013.

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

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

A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology*, 179(6):764–774, 2014.

D. J. Stekhoven and P. Bühlmann. Missforest: non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.

A. L. Stubbendick and J. G. Ibrahim. Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica*, pages 1143–1167, 2006.

C. D. Sutton. Classification and Regression Trees, Bagging, and Boosting. *Handbook of Statistics*, 24:303–329, 2005.

M. Tanner and W. H. Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

S. Van Buuren. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, 2018.

S. Van Buuren and K. Groothius-Oudshoorn. mice: Multiple Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 2009.

K. Vaysse and P. Lagacherie. Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, 291:55–64, 2017.

S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

S. Wager, T. Hastie, and B. Efron. Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.

C. Wang and Z. Feng. Boosting with missing predictors. *Biostatistics*, 11(2):195–212, 2010.

N. A. Weiss. *A Course in Probability*. Addison-Wesley, 2006.

I. R. White and J. B. Carlin. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29(28):2920–2931, 2010.

M. N. Wright and A. Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01.

Z. Zhang and H. E. Rockette. On maximum likelihood estimation in parametric regression with missing covariates. *Journal of Statistical Planning and Inference*, 134(1):206–223, 2005.