Essays in Predictive and Causal Machine Learning

D I S S E R T A T I O N
of the University of St.Gallen,
School of Management,
Economics, Law, Social Sciences,
International Affairs and Computer Science,
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Gabriel Okasa

from

Slovakia

Approved on the application of

Prof. Dr. Michael Lechner

and

Prof. Dr. Martin Biewen

Dissertation no. 5206

Difo-Druck GmbH, Untersiemau 2022

Essays in Predictive and Causal Machine Learning

D I S S E R T A T I O N
of the University of St.Gallen,
School of Management,
Economics, Law, Social Sciences,
International Affairs and Computer Science,
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Gabriel Okasa

from

Slovakia

Approved on the application of

Prof. Dr. Michael Lechner

and

Prof. Dr. Martin Biewen
Prof. Francesco Audrino, PhD

Dissertation no. 5206

Difo-Druck GmbH, Untersiemau 2022

The University of St.Gallen, School of Management, Economics, Law, Social Sciences, International Affairs and Computer Science, hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St.Gallen, November 8, 2021

The President:

Prof. Dr. Bernhard Ehrenzeller

Dedicated To My Parents.

Acknowledgements

Doing a PhD is a long and challenging journey. I would not have been able to accomplish it without the great support of my supervisor, colleagues, friends and family, to whom I am deeply thankful.

First and foremost, I would like to thank my supervisor Michael Lechner for his extraordinary supervision. I am grateful to him for introducing me into the world of empirical research and causal inference. His careful guidance and valuable advice throughout my doctoral studies are very much appreciated. I have greatly benefited from his helpful feedback that substantially improved this thesis. Besides the thesis supervision, I would like to thank him for the exceptional working conditions at the institute and his permanent availability. It has been an honour to have the opportunity to do a PhD under his supervision.

Furthermore, I would like to thank my colleagues at the institute, who were always there to help me when I struggled. A special thanks goes to Daniel Goller and Michael Zimmert, who accompanied my PhD journey from the very beginning until the end. I am indebted to them for their time, help and feedback on my research ideas, projects and many more. I would also like to thank Michael Knaus and Jana Mareckova for their support and helpful feedback anytime I needed it. Furthermore, I thank Hugo Bodory, Daniel Boller, Petyo Bonev, Conny David, Sandro Heiniger, Alex Krumer, Georgi Lautliev, Carina Steckenleiter and Anthony Strittmatter for their professional as well as personal support throghout my PhD studies.

Moreover, I would like to thank my PhD fellows, who accompanied me both in the course phase as well as in the research phase and made this PhD journey overall an amazing experience. I would like to particularly thank Daniele Ballinari for all the coffee breaks and discussions about research, programming, and beyond. I also thank Mirjam Bächli, Janosch Brenzel-Weiss, Jonathan Chassot, Patrick Chuard, Immanuel Lampe, Benedikt Lennartz, Edouard Mattille, Andrin Pelican, Adam Pigon and Marc-Antoine Ramelet for inspiring discussions and extracurricular activities.

Lastly, I would like to thank my family and friends for their unconditional support. I feel deeply indebted to my parents Zuzana and Milan who always encouraged me to follow my dreams and supported me in all of my endeavours. Without them I would not be here today. Finally, I would like to thank my girlfriend Janice, who supported me throughout the whole PhD journey and always believed in me. Thanks to her, the PhD years have been amongst the most joyful of my life.

St.Gallen, November 9, 2021

Gabriel Okasa

Contents

Summary

This dissertation consists of three chapters devoted to topics in predictive and causal machine learning. Common to all chapters is the synthesis of classical econometric methods and novel machine learning algorithms. Hence this doctoral thesis provides new insights into applications of machine learning for predictive tasks and for causal inference.

The first chapter investigates the estimation of heterogeneous causal effects using machine learning. We focus on the meta-learning framework where the estimation of the causal parameter is decomposed into separate prediction tasks. Using synthetic and empirical simulations we study the finite sample performance of meta-learners based on the Random Forest algorithm under different implementations using sample-splitting and cross-fitting procedures. The results imply that sample-splitting is beneficial in large samples for bias reduction but leads to an increase in variance, whereas cross-fitting keeps the bias low and successfully restores the full sample size efficiency. In contrast, the full-sample estimation is preferable in small samples when using machine learning. Additionally, we provide guidelines for applications of meta-learners in empirical studies depending on particular data characteristics such as treatment shares and sample size.

The second chapter considers the estimation of ordered choice models using machine learning. Similarly, as in the first chapter, we focus on the Random Forest algorithm and develop a new machine learning estimator for models with ordered categorical outcome variable. The proposed Ordered Forest flexibly estimates the conditional ordered choice probabilities while taking the ordering information explicitly into account. In contrast to common machine learning estimators, it is not only suited for prediction tasks, but it also enables the estimation of marginal effects and conducting statistical inference, which provides additional interpretability as in classical econometric estimators. We conduct an extensive simulation study and find a good predictive performance, particularly in settings with nonlinearities and multicollinearity. Furthermore, we demonstrate the estimation of marginal effects and their standard errors in an empirical application.

The third chapter presents an empirical application based on the estimation of causal effects using machine learning. As in the previous two chapters, we rely on the Random Forest method and consider its causal variant, the Modified Causal Forest. Following the rise of online dating, we study the effect of sport activity on partner choice by exploiting a unique dataset from an online dating platform. In particular, we estimate the causal effect of sport frequency on the contact chances, controlling for a large set of observable user characteristics. We find that for male users, doing sport on a weekly basis increases the probability to receive a first message by more than 50%, in comparison to no sport activity. In contrast, we do not find such an evidence for female users. Moreover, the results indicate heterogeneity as for male users the effect increases with higher income.

Zusammenfassung

Die vorliegende Dissertation besteht aus drei Kapitel, welche sich dem prädiktiven und kausalen maschinellen Lernen widmen. Die Synthese von klassischen Methoden der Ökonometrie und neuartigen Algorithmen des maschinellen Lernens ist allen Kapiteln gemeinsam. Somit bietet diese Doktorarbeit neue Erkenntnisse für die Anwendung von maschinellem Lernen für die Prädiktion und kausale Inferenz.

Das erste Kapitel untersucht die Schätzung heterogener kausaler Effekte mittels maschinellen Lernens. Wir fokussieren uns auf das Konzept des Meta-Learning, wobei die Schätzung kausaler Parameter in einzelne Prädiktionsmodelle aufgeteilt wird. Anhand von synthetischen und empirischen Simulationen analysieren wir die Eigenschaften der Meta-Learner in Stichproben mit begrenzter Anzahl an Beobachtungen, basierend auf dem Random Forest Algorithmus. Den Schwerpunkt legen wir hierbei auf verschiedene Implementierungen anhand von Sample-Splitting und Cross-Fitting Prozeduren. Die Ergebnisse belegen, dass Sample-Splitting in grossen Stichproben für die Verzerrungsreduktion hilfreich ist, jedoch führt dies gleichzeitig zu einer Varianzerhöhung. Ebenso reduziert Cross-Fitting in grossen Stichproben die Verzerrung, währenddessen die Effizienz erfolgreich wiederhergestellt wird. Demgegenüber ist bei Anwendung des maschinellen Lernens in kleinen Stichproben die Schätzung basierend auf der ganzen Stichprobe zu bevorzugen. Des Weiteren leiten wir Anwendungsempfehlungen für die Meta-Learner in empirischen Studien ab, die auf bestimmten Datenmerkmalen, wie Treatmentanteile und Stichprobengrösse, beruhen.

Das zweite Kapitel befasst sich mit der Schätzung von Ordered Choice Modellen mittels maschinellen Lernens. Ähnlich wie im ersten Kapitel betrachten wir den Random Forest Algorithmus und entwickeln einen neuen Schätzer für Modelle mit einer geordneten kategorischen abhängigen Variable. Die vorgeschlagene Ordered Forest Methode schätzt flexibel die bedingten geordneten Wahlwahrscheinlichkeiten, worin die Ordnungsinformation ausdrücklich berücksichtigt wird. Im Vergleich mit den gängigen Methoden des maschinellen Lernens, ist die Methode nicht nur für die Prädiktion geeignet, sondern ermöglicht auch die Schätzung marginaler Effekte sowie eine statistische Inferenzanalyse. Somit bietet sie eine zusätzliche Interpretierbarkeit, die auf den klassischen ökonometrischen Methoden beruht. Wir führen eine umfangreiche Simulationsstudie durch und stellen eine gute Vorhersagekraft fest, insbesondere in Szenarien mit Nichtlinearitäten und Multikollinearität. Ferner demonstrieren wir die Schätzung von marginalen Effekten und deren Standardfehler in einer empirischen Anwendung.

Das dritte Kapitel präsentiert eine empirische Anwendung gestützt auf die Schätzung von kausalen Effekten mittels maschinellen Lernens. Sowie in den vorherigen beiden Kapiteln, beziehen wir uns auf die Random Forest Methode und betrachten ihre Kausalversion, den Modified Causal Forest. Nach dem Aufkommen des Online-Datings analysieren wir den Effekt der Sportaktivität auf die Partnerwahl durch die Nutzung eines einzigartigen Datensatzes einer Online-Dating Plattform. Insbesondere schätzen wir den kausalen Effekt der Häufigkeit des Sporttreibens auf die Kontaktchancen, unter Berücksichtigung einer grossen Anzahl von beobachtbaren Benutzercharakteristiken. Wir stellen fest, dass für männliche Benutzer das wöchentliche Sporttreiben die Wahrscheinlichkeit eine erste Nachricht zu erhalten um mehr als 50% erhöht, verglichen zu keiner sportlichen Aktivität. Andererseits finden wir keine solche Evidenz für weibliche Benutzer. Zugleich weisen die Ergebnisse eine Heterogenität auf, indem der Effekt für Männer mit einem höheren Einkommen steigt.

Chapter 1

Meta-Learners for Estimation of Causal Effects: Finite Sample Cross-Fit Performance

Abstract

Estimation of causal effects using machine learning methods has become an active research field in econometrics. In this respect the meta-learning algorithms have gained considerable attention for estimation of heterogeneous causal effects. We study the finite sample performance of various meta-learners for estimation of heterogeneous treatment effects, while explicitly focusing on the usage of sample-splitting and cross-fitting to reduce the overfitting bias. In both synthetic and empirical simulations we find that the performance of the meta-learners in finite samples greatly depends on the estimation procedure. The results imply that sample-splitting and cross-fitting are beneficial in large samples for bias reduction and efficiency of the meta-learners, respectively, whereas full-sample estimation is preferable in small samples. Furthermore, we derive practical recommendations for usage of specific meta-learners in empirical studies depending on particular data characteristics such as treatment shares and sample size.

Keywords: Meta-learners, causal machine learning, heterogeneous treatment effects, Monte Carlo simulation, sample-splitting, cross-fitting.

JEL classification: C15, C18, C31.

1.1 Introduction

In recent years there has been a growing interest in the estimation of causal effects using machine learning algorithms, particularly in the field of economics (Athey, 2018). The newly emerging synthesis of machine learning methods with causal inference has a large potential for a more comprehensive estimation of causal effects (Lechner, 2018). On the one hand, it enables a more flexible estimation of average effects which are of main interest in microeconometrics (Imbens & Wooldridge, 2009). On the other hand, it advances the estimation beyond the average effects and allows for a systematic analysis of effect heterogeneity (Athey & Imbens, 2017). Both of these aspects contribute to a better description of the causal mechanisms and thus to a possibly more efficient treatment allocation (Zhao, Zeng, Rush, & Kosorok, 2012; Kitagawa & Tetenov, 2018; Athey & Wager, 2021; Nie, Brunskill, & Wager, 2021). Hence, applied researchers can greatly benefit from the usage of machine learning methods ranging from evaluation of public policies and business decisions to designing personalized interventions (Andini, Ciani, de Blasio, D'Ignazio, & Salvestrini, 2018; Bansak et al., 2018).

Machine learning estimators as such are, however, primarily designed to tackle prediction problems and thus cannot be used off-the-shelf for causal inference. Therefore, new approaches for the estimation of causal parameters using machine learning emerged. Within the fast developing causal machine learning literature, one strand focused on direct modifications of the existing machine learning algorithms that adjust the objective function for the estimation of causal effects. Such approach has led for example to the developments of Causal Trees (Athey & Imbens, 2016) and Causal Forests (Wager & Athey, 2018; Lechner, 2018; Athey, Tibshirani, & Wager, 2019). While these methods have well-established theoretical properties, they restrict the researcher in the choice of the machine learning method. Another strand of the causal machine learning literature thus proposed general procedures to decompose the causal problem into separate prediction problems that can be solved by standard machine learning algorithms and subsequently combined to estimate the causal parameters of interest. This approach has led to the development of meta-learners for the estimation of causal effects (see e.g. Künzel, Sekhon, Bickel, & Yu, 2019; Kennedy, 2020; or Nie & Wager, 2021).

The meta-learners have received considerable attention for several reasons. First, the meta-learners do not modify the objective function of the machine learning methods but rather combine their predictions in order to estimate the causal effect (Künzel et al., 2019). This enables to directly leverage the superior prediction power of machine learning estimators. Second, the meta-learners are generic algorithms refraining from a specific usage of any particular machine learning method. This allows to apply any suitable supervised learning method for the particular prediction problem at hand. Third, the meta-learners are attractive due to the ease of implementation using standard statistical software. This permits researchers to apply the meta-learners without any potential restrictions due to limited availability in software packages and enables tailored implementation for particular types of data. Despite the attractive features of the meta-learners, there is little guidance for applied researchers on how to choose from a variety of the meta-learners proposed in the literature, with lack of unifying simulation evidence for an assessment of the performance of the meta-learners in applied settings.

The complexity of the meta-learners varies widely and often hinges on the estimation of the nuisance functions such as the conditional mean of the outcome and the treatment, respectively (Chernozhukov et al., 2018). The basic meta-learning algorithms include the S-learner and T-learner which besides the treatment effect function do not require estimation of any additional nuisance functions. However, the most prominent and widely used meta-learners in the literature consist of the X-learner (Künzel et al., 2019), the DR-learner (Kennedy, 2020), and the R-learner (Nie & Wager, 2021), which all require

estimation and combination of several nuisance functions to estimate the causal effect.[1] Due to the machine learning estimation of such nuisance functions the meta-learners are prone to the overfitting bias, i.e. own observation bias. Therefore, sample-splitting has been proposed in the literature to reduce the overfitting bias by using one part of the sample for estimation of the nuisance functions and the other part of the sample for estimation of the causal effect. In order to regain the full sample size efficiency of the estimator cross-fitting repeats the estimation by swapping the samples and averaging the estimated causal effects (Chernozhukov et al., 2018). However, the usage of sample-splitting and cross-fitting is not well understood in practice and the specific definitions of meta-learners differ substantially in their implementation of these procedures. Despite the ambiguous definitions, there is a lack of simulation evidence concerned with the usage of sample-splitting and cross-fitting within the meta-learning framework and thus limited guidance for or against specific implementations. Moreover, there appears to be limited knowledge about how the asymptotic arguments translate into finite sample properties of the meta-learners.

In this paper, we address both of the above issues and study the finite sample properties of the machine learning based meta-learners for estimation of causal effects based on the specific implementations using the full-sample, sample-splitting and cross-fitting procedures for varying sample sizes. We focus on evaluating the estimation of heterogeneous treatment effects as these provide the most detailed description of the underlying causal mechanisms and thus allow for a better assessment of the individualized impacts of an intervention. For this purpose, we review the most widely used meta-learning algorithms together with their theoretical estimation requirements with respect to sample-splitting and cross-fitting and identify their strengths and weaknesses. We conduct both synthetic and empirical simulations comparing the performance of the meta-learners in various settings featuring unequal treatment shares, non-linear functional forms and large-dimensional feature sets. Importantly, within the simulations we explicitly study the convergence performance of the meta-learners based on growing sample sizes up to $32'000$ observations. Furthermore, we derive practical recommendations on the choice of specific meta-learners and the respective estimation procedures for applied empirical work.

The results of our simulation experiments reveal that the choice of the estimation procedure has a large impact on the performance of the machine learning based meta-learners in finite samples. For sufficiently large samples we provide evidence for the theoretical arguments of bias reduction via sample-splitting and cross-fitting, while for smaller samples we observe adverse effects of these procedures when using machine learning. The results show that, if computation time is not a constraint, cross-fitting is always preferable to sample-splitting as it keeps the bias low, while successfully reducing the variance of the estimators even in small samples. Additionally, the results imply heterogeneous impacts of the estimation procedures on the performance of the meta-learners. The X-learner's performance is quite stable regardless of the estimation procedure, whereas the performance of the R-learner and DR-learner is more sensitive to the choice of the estimation procedure. Assessing the performance of the particular meta-learners reveals a clear pattern. In empirical settings with highly imbalanced treatment shares, the X-learner performs best, irrespective of the sample size, while the DR-learner becomes unstable due to extreme propensity scores. For less imbalanced settings the X-learner's performance is still superior in smaller samples, however, it gets outperformed by the DR-learner in larger samples which exhibits the fastest convergence rate in its sample-splitting and cross-fitting version. In empirical settings with balanced treatment shares, the performance of the DR-learner or the R-learner is superior for any sample size considered. The results imply that the usage of less sophisticated S-learner for estimation of causal effects should be avoided, while the T-learner might be a reasonable choice in small samples.

[1] Further examples of some meta-learners proposed in the literature consist of the U-learner and Y-learner (Stadie, Kunzel, Vemuri, & Sekhon, 2018), or the IF-learner (Curth, Alaa, & van der Schaar, 2020) and RA-learner (Curth & van der Schaar, 2021).

This paper contributes to the causal machine learning literature in several ways. First, we provide a unifying simulation evidence of meta-learning algorithms for the estimation of heterogeneous causal effects in large-dimensional and highly non-linear settings based on synthetic and empirical simulations. Second, we explicitly study the meta-learners under the full-sample, sample-splitting and cross-fitting implementations, respectively and thereby provide evidence on the contrast between the asymptotic arguments and finite sample properties. Third, we empirically investigate the convergence performance of the meta-learners by repeating the simulation experiments with growing sample sizes. Finally, we derive relevant practical recommendations for applied empirical work which are based on the particular observable data characteristics.

This paper is organized as follows. We briefly discuss the related literature in Section 1.1.1. Section 1.2 introduces the notation, the parameters of interest and their identification. Section 1.3 reviews the considered meta-learners and the estimation procedures. Section 1.4 describes the synthetic as well as empirical simulations and presents the corresponding results. The main findings of the study are discussed in Section 1.5. Section 1.6 concludes. Further details including descriptive statistics, an exhaustive summary of the main and supplementary results as well as a computation time analysis are provided in Appendices 1.A, 1.B and 1.C, respectively.

1.1.1 Literature

In general the literature on the finite sample properties of causal machine learning estimators under a unified framework seems to be rather scarce. An exception in the econometric literature[2] is Knaus, Lechner, and Strittmatter (2021) who study a wide range of estimators for heterogeneous as well as (group) average treatment effects, including direct estimators as well as some meta-learners in an Empirical Monte Carlo Study as developed in Huber, Lechner, and Wunsch (2013) and Lechner and Wunsch (2013). Knaus et al. (2021) find no estimator to perform uniformly best, but notice that estimators which model both the outcome as well as the treatment process are substantially more robust throughout all data generating processes considered, which can be observed in our simulations as well. Among the meta-learners considered, the DR-learner and the R-learner perform especially well in terms of the root mean squared error. Moreover, using the Random Forest as a base learner turns out to be more stable with better statistical properties in contrast to using the Lasso, particularly in smaller samples, which also motivates the usage of the Random Forest in our simulations. However, although both meta-learners are implemented with cross-fitting, an explicit consideration of different sample-splitting or cross-fitting schemes is missing. Curth and van der Schaar (2021) focus directly on meta-learning algorithms for estimation of heterogeneous treatment effects, but refrain from studying sample-splitting and cross-fitting procedures and rely fully on the full-sample estimation. In this regard, Zivich and Breskin (2021) study the performance of treatment effect estimators based on cross-fitting, including some meta-learners as well. Similarly to Knaus et al. (2021) they find the DR-learner with an ensemble machine learning base learners together with cross-fitting to perform the best among all considered estimators, both in comparison to cases without cross-fitting and to parametric base learners. However, Zivich and Breskin (2021) study exclusively the estimation of average effects without examining convergence performance of the estimators, considering only a single sample size of $3'000$ observations. Recently, Jacob (2020) focuses on the estimation of heterogeneous treatment effects under various cross-fitting schemes for selected meta-learning algorithms. Also, in this simulation study the DR-learner together with the R-learner achieve consistently the best results. Nonetheless, Jacob (2020) stresses the heterogeneous impacts of the particular sample-splitting and cross-fitting procedures on each meta-learner, which is documented in our

[2]Wendling et al. (2018) conduct similar empirical simulation study in medical context.

simulations as well. Nevertheless, even though considering varying sample sizes within the simulation experiments, the considered sample sizes are limited to $2'000$ observations. Overall, none of the above studies focuses directly on the convergence performance of the meta-learners under various estimation procedures which still remains an open question. To the best of our knowledge, this is the first paper that empirically studies the convergence properties of the meta-learners under full-sample, sample-splitting and cross-fitting implementations with growing sample sizes up to several thousands of observations, reaching $32'000$ in our simulations.

Besides the meta-learning framework, there has been also a substantial development of specific causal estimators based on direct modifications of particular machine learning algorithms. Especially, the tree-based estimators have been studied extensively in this respect. These include the above-mentioned Causal Trees (Athey & Imbens, 2016) as well as Causal Boosting (Powers et al., 2018) and Causal Forests (Wager & Athey, 2018) with the extensions of the Modified Causal Forests (Lechner, 2018) and the Generalized Random Forests (Athey et al., 2019). These methods are based on the underlying predictive algorithms of Regression Trees (Breiman, Friedman, Olshen, & Stone, 1984), Boosted Trees (Friedman, 2001) and Random Forests (Breiman, 2001), respectively. Furthemore, Bayesian versions of Regression Trees (Chipman, George, & McCulloch, 1998) have been adapted for estimation of causal effects as well (Hill, 2011; Taddy, Gardner, Chen, & Draper, 2016; Hahn, Murray, & Carvalho, 2020). Besides the estimators based on recursive partitioning, important causal adjustments have been applied in respect to regularization based estimators such as the Lasso (Qian & Murphy, 2011; Belloni, Chernozhukov, & Hansen, 2013; Tian, Alizadeh, Gentles, & Tibshirani, 2014) or Lasso-augmented Support Vector Machines (Imai & Ratkovic, 2013). Additionally, further machine learning algorithms such as the Nearest Neighbours (Fan, Lv, & Wang, 2018) or Neural Networks (Johansson, Shalit, & Sontag, 2016; Shalit, Johansson, & Sontag, 2017; Schwab, Linhardt, & Karlen, 2018; Shi, Blei, & Veitch, 2019) have been transformed towards causal inference as well. For a comprehensive overview of many of these estimators, we refer the interested reader to Athey and Imbens (2019) or Jacob (2021). In this paper, although we focus on the machine learning estimation of causal effects, we refrain from an analysis of these methods due to major conceptual differences to the meta-learning framework and the lack of comparability in terms of the usage of sample-splitting and cross-fitting procedures.

1.2 Framework and Identification

In order to describe the effects of interest and their corresponding identification assumptions we rely on the potential outcome framework (Rubin, 1974). We assume a population $\mathcal{P}$ from which a realization of N *i.i.d.* random variables is given consisting of a random sample $\{Y_i(1), Y_i(0), W_i, X_i\} \sim \mathcal{P}$. Here, we consider a binary treatment variable W_i that is equal to 1 for the treated group and equal to 0 for the control group, respectively. According to the treatment status we define the potential outcome $Y_i(1)$ under treatment for the case when $W_i = 1$ and correspondingly the potential outcome $Y_i(0)$ under control for $W_i = 0$. Additionally, we define a p-dimensional vector of exogenous pre-treatment covariates such that $X_i \in \mathbb{R}^p$. Given this definition we can characterize the *Individual Treatment Effect* (ITE) as follows:

$$\xi_i = Y_i(1) - Y_i(0).$$

However, the fundamental problem of causal inference is that we never observe both potential outcomes at the same time (Holland, 1986). Hence, the observed outcomes are defined according to the observational rule as $Y_i = Y_i(W_i)$. The observed data then consists of the triple $\{Y_i, W_i, X_i\}_{1 \leq i \leq N}$. Nevertheless, it is still possible to identify the expectation of ξ_i under additional assumptions (compare Rubin, 1974; or

Imbens & Rubin, 2015). Thus, we shift the effect of interest towards the *Conditional Average Treatment Effect* (CATE) which takes the expectation of ξ_i, conditional on covariates X_i and is given as:

$$\tau(x) = \mathbb{E}\big[\xi_i \mid X_i = x\big] = \mathbb{E}\big[Y_i(1) - Y_i(0) \mid X_i = x\big] = \mu_1(x) - \mu_0(x)$$

where $\mu_1(x) = \mathbb{E}[Y_i(1) \mid X_i = x]$ and $\mu_0(x) = \mathbb{E}[Y_i(0) \mid X_i = x]$ are the response functions for potential outcomes under treatment and under control, respectively. In this paper we always refer to the CATE with conditioning on all observed exogeneous covariates and thus focusing on the finest level of heterogeneity (see e.g. Knaus et al., 2021).[3] Künzel et al. (2019) point out that the best estimator for $\tau(x)$ is also the best estimator for ξ_i in terms of the mean squared error (MSE).

In order to identify the effects of interest, we need a set of identification assumptions. We operate under the selection-on-observables strategy[4] (see e.g. Imbens & Rubin, 2015) and assume that we observe all relevant confounders, i.e. all covariates X_i that *jointly* influence both the treatment W_i and the potential outcomes, $Y_i(0)$ and $Y_i(1)$. We state the following identification assumptions:

Assumption 1 (Conditional Independence) $\big(Y_i(0), Y_i(1)\big) \perp\!\!\!\perp W_i \mid X_i = x, \forall x \in \text{supp}(X_i)$.

Assumption 2 (Common Support) $0 < \mathbb{P}\big[W_i = 1 \mid X_i = x\big] < 1, \forall x \in \text{supp}(X_i)$.

Assumption 3 (SUTVA) $Y_i = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0)$.

Assumption 4 (Exogeneity) $X_i(0) = X_i(1)$.

According to Assumption 1, also referred to as the conditional ignorability or unconfoudedness assumption, we assume that the potential outcomes are independent of the treatment assignment once conditioned on the covariates, i.e. we assume that there are no hidden confounders. Assumption 2, also known as the overlap assumption, states that the conditional treatment probability is bounded away from 0 and 1 and thus it is possible to observe treated as well as control units for each realization of $X_i = x$. Assumption 3 is known as the stable unit treatment value assumption and indicates that the observed treatment value for a unit is independent of the treatment exposure for other units, which rules out any general equilibrium or spillover effects between treated and controls. Lastly, Assumption 4 specifies that the covariates are not influenced by the treatment.[5] Under these assumptions it follows that

$$\begin{aligned}
\tau(x) &= \mathbb{E}\big[Y_i(1) - Y_i(0) \mid X_i = x\big] && (1.2.1)\\
&= \mathbb{E}\big[Y_i(1) \mid X_i = x\big] - \mathbb{E}\big[Y_i(0) \mid X_i = x\big] && (1.2.2)\\
&= \mathbb{E}\big[Y_i(1) \mid X_i = x, W_i = 1\big] - \mathbb{E}\big[Y_i(0) \mid X_i = x, W_i = 0\big] && (1.2.3)\\
&= \mathbb{E}\big[Y_i \mid X_i = x, W_i = 1\big] - \mathbb{E}\big[Y_i \mid X_i = x, W_i = 0\big] && (1.2.4)
\end{aligned}$$

and thus the CATE can be nonparametrically identified from observable data (Hurwicz, 1950).

[3] In general, the term CATE describes conditional average treatment effects on various aggregation levels. In our case, the CATE corresponds to the *Individualized Average Treatment Effect* (IATE). Additionally, researchers and especially policy makers might be interested in a low-dimensional heterogeneity level based on some pre-specified heterogeneity covariates of interest, which are referred to as the *Group Average Treatment Effects* (GATEs). Such effects are, however, beyond the scope of our study and the interested reader is referred to Zimmert and Lechner (2019), Jacob, Härdle, and Lessmann (2019) and Semenova and Chernozhukov (2021) for a theoretical analysis and to Knaus et al. (2021) for simulation based results or to Cockx, Lechner, and Bollens (2019), Knaus, Lechner, and Strittmatter (2020), Hodler, Lechner, and Raschky (2020) and Goller, Harrer, Lechner, and Wolff (2021) for empirical applications estimating policy relevant GATEs.

[4] For estimation of heterogeneous effects under different identification strategies see e.g. Athey et al. (2019), Bargagli Stoffi and Gnecco (2020) and Biewen and Kugler (2021) for the case of instrumental variables and Gulyas and Pytka (2020) and Zimmert and Zimmert (2020) for the case of difference-in-differences.

[5] Analogously to the definition of potential outcomes, we denote potential covariates under control and under treatment as $X_i(0)$ and $X_i(1)$, respectively.

1.3 Meta-Learning Algorithms and Estimation Procedures

In the machine learning literature meta-learning represents algorithms that exploit knowledge about learning to improve the algorithm's performance, as generally defined by Vilalta and Drissi (2002). These include various algorithms that learn to solve new task from prior learning experience, i.e. *learning to learn* (Schmidhuber, 1987; Thrun & Pratt, 1998), algorithms that learn from multiple related tasks, i.e. *multi-task learning* (Caruana, 1997), or algorithms that learn from multiple models solving identical task, i.e. *ensemble learning* (Dietterich, 2000).[6] Recently, the meta-learning framework has been adopted within the causal machine learning literature for learning causal effects from multiple prediction models (see for example Künzel et al., 2019), which could be termed accordingly as *causal learning*.

At a high level the meta-learners for estimation of heterogeneous causal effects are two-step algorithms. In the first step they define regression functions, in the causal machine learning literature often denoted as the nuisance functions (Chernozhukov et al., 2018; Kennedy, 2020), which can be estimated by any suitable supervised learning method, i.e. the base learner. In the second step they use the estimated nuisance functions to construct an estimator for the causal effect, i.e. the meta-learner. Various meta-learners then differ in the definitions of the nuisance functions and their subsequent usage to obtain the final estimator for the causal effects. Depending on the algorithm complexity, some meta-learners require estimation of only one single model whereas other require estimation of multiple models. This raises the question of data usage within the estimation procedure and thus the possible need for sample-splitting and cross-fitting, respectively.[7]

In general, the nuisance functions are defined as conditional expectations of various types. The most common types are the propensity score function and the response function. First, the propensity score is defined as the conditional probability of a binary treatment W_i given the covariates X_i as follows:

$$e(x) = \mathbb{P}\big[W_i = 1 \mid X_i = x\big].$$

In the causal inference literature the propensity score plays a central role (Rosenbaum & Rubin, 1983) in many matching and reweighting methods to balance the distributions of treated and controls (see Hahn, 1998; and Huber et al., 2013, among others). Second, the response function is broadly defined as the conditional expectation of an outcome variable Y_i given a conditioning set of explanatory variables. The particular definitions of the response function then differ in the specification of the conditioning set and the subset of the data used. For the meta-learners studied in this paper, the following definitions of the response function are of interest:

$$\mu(x, w) = \mathbb{E}[Y_i \mid X_i = x, W_i = w] \tag{1.3.1}$$

$$\mu(x) = \mathbb{E}[Y_i \mid X_i = x] \tag{1.3.2}$$

where Equation 1.3.1 defines the full response function with conditioning on both the covariates X_i as well as the treatment indicator W_i, while $\mu(x, 1)$ and $\mu(x, 0)$ describe the response functions with conditioning on the covariates X_i in the subpopulation under treatment $W_i = 1$ and under control $W_i = 0$, accordingly. Similarly, Equation 1.3.2 defines the full response function with conditioning only on covariates. The meta-learners then use selected nuisance functions together with the available data as

[6] For a recent survey on meta-learning, see Vanschoren (2019).

[7] Recently, related issue of data usage of the meta-learning algorithms with respect to splitting into training and validation set for the learning to learn domain has been discussed by Bai et al. (2020) and Saunshi, Gupta, and Hu (2021).

inputs for the estimation of the CATE function which can be generally denoted as follows:

$$\tau(x) = \zeta\big(W_i, X_i, Y_i, e(x), \mu(x,w), \mu(x)\big)$$

where $\zeta(\cdot)$ is a function of the respective inputs, which is detailed for each particular meta-learning algorithm in Section 1.3.2. The problem arises when estimating the nuisance functions using flexible machine learning methods as these are prone to the overfitting bias, i.e. the 'own observation bias'. The overfitting bias emerges when the in-sample data is fitted too well such that the out-of-sample performance is compromised (see e.g. Hastie, Tibshirani, & Friedman, 2009, for a general discussion of the overfitting issue in machine learning). Hence, a single observation i can have a large influence on the predictions for covariates X_i as pointed out by Athey and Imbens (2019). Chernozhukov et al. (2018) and Newey and Robins (2018) thus propose sample-splitting procedures that allow for elimination of such overfitting biases.[8]

1.3.1 Sample-Splitting and Cross-Fitting

Theoretical arguments express the need for sample-splitting when the causal estimator involves several estimation steps such as the estimation of nuisance functions. Within the meta-learning framework the nuisance functions are typically highly complex and potentially high-dimensional functions estimated by supervised machine learning methods such as penalized regression, tree-based methods, neural networks, etc. Using the same data sample for machine learning estimation of the nuisance function as well as for estimation of the causal effect leads to overfitting which induces a bias in the CATE estimator. On a high level, the bias of the CATE estimator can generally be decomposed into an estimation error of learning the CATE function itself, and the estimation error in learning the nuisance functions, encompassing the overfitting bias (see e.g. Kennedy, 2020). Chernozhukov et al. (2018) show that for the ATE estimation the overfitting bias can be controlled by using sample-splitting, while Kennedy (2020) and Nie and Wager (2021) extend this concept for the CATE estimation. In that case one part of the sample is used to estimate the nuisance functions and the other part is used to estimate the causal effect.[9] As a result, the bias term stemming from overfitting can be shown to be bounded and to converge to zero. Building upon this result, Newey and Robins (2018) propose a different sample-splitting scheme called *double* sample-splitting. In this case, not only the nuisance functions are estimated together on a separate part of the sample but each single nuisance function is estimated on an own separate part of the sample. In practice, the training data is split into $M + 1$ equally sized parts, with M being the number of nuisance functions to estimate and the remaining part of the data serves for estimation of the causal effect. Newey and Robins (2018) show that under the double sample-splitting the bias term converges to zero at a faster rate compared to standard sample-splitting where all nuisances are estimated on the same sample.[10] The double sample-splitting procedure has also been recently implemented by Kennedy (2020) in the context of the DR-learner.

In general, the overfitting bias could also be controlled for by restricting the complexity of the nuisance functions which would, however, prevent high-dimensional settings as well as usage of a variety

[8] Original ideas of using sample-splitting procedures to eliminate own observation bias stem from the literature on density estimation going back to Bickel (1982), Bickel and Ritov (1988) and Powell, Stock, and Stoker (1989) among others.

[9] Sample-splitting procedures are frequently used in causal machine learning literature including Double Machine Learning (Chernozhukov et al., 2018), Causal Forests (Wager & Athey, 2018; Lechner, 2018) or the here-discussed meta-learners (Kennedy, 2020; Nie & Wager, 2021).

[10] The intuition for this result comes from the observation that for estimators using multiple nuisance functions, such as the doubly robust estimators as e.g. the herein discussed DR-learner, the estimation error involves a product of the biases from the estimation of the M nuisance functions. This induces additional nonlinearity bias if all M nuisance functions are estimated using the same data, which gets effectively removed by using separate samples for estimation of each of the M functions. For more details see Newey and Robins (2018) and Kennedy (2020).

of machine learning estimators or ensembles of those.[11] Hence, the advantage of using sample-splitting is to allow for a high degree of complexity of the nuisance functions estimated by a wide class of machine learning estimators (Kennedy, 2020).

It follows that, theoretically, sample-splitting prevents overfitting and thus reduces the bias in the final causal estimator (Chernozhukov et al., 2018; Wager & Athey, 2018). At the same time, however, the variance of the estimator increases as less data is effectively used for estimation. Cross-fitting (Chernozhukov et al., 2018) and respectively *double* cross-fitting (Newey & Robins, 2018) have been proposed in the literature in order to reduce the variance loss induced by sample-splitting. In this procedure, the roles of the data parts get switched such that each part has been used for both the estimation of nuisances as well as the causal effect estimation. The final CATE estimator is then an average of the separate effect estimators produced. This method can be further extended to use more than $M + 1$ splits denoted as K-fold cross-fitting (Chernozhukov et al., 2018) with the final CATE estimator given as:

$$\hat{\tau}(x) = \frac{1}{K} \sum_{k=1}^{K} \hat{\tau}_k(x)$$

where $\hat{\tau}_k(x)$ is the CATE estimator based on the k-th fold.[12]

The above theoretical arguments have a direct impact on the implementation of various meta-learning algorithms. Under the double sample-splitting the more models have to be estimated within the meta-learning algorithm, the more data splits are being implicitly induced, while the impact thereof in finite samples is not clear *a priori* as pointed out by Newey and Robins (2018). As such, the researcher faces a typical bias-variance trade-off with respect to sample-splitting. In order to illustrate the issue it is instructive to decompose the mean squared error (MSE) of a CATE estimator $\hat{\tau}(x)$:

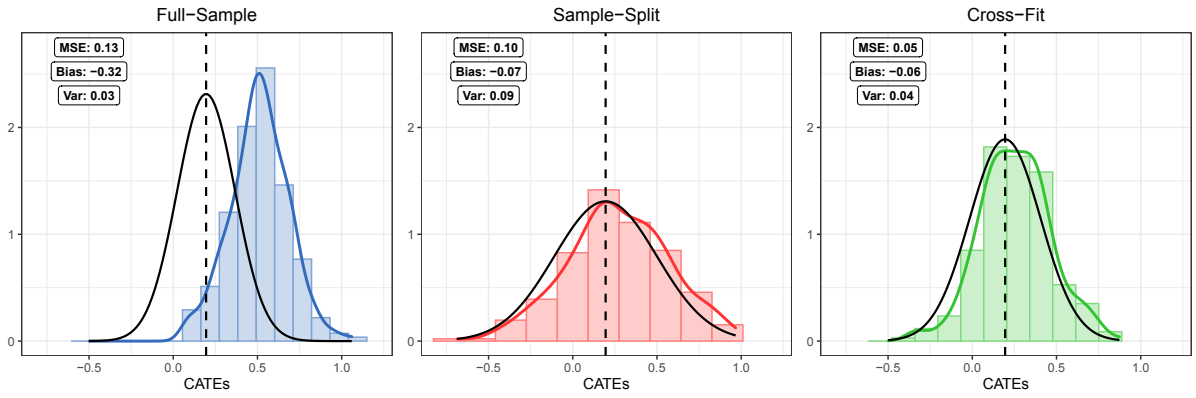
$$MSE\Big(\hat{\tau}(x)\Big) = Var\Big(\hat{\tau}(x)\Big) + \Big(Bias\big(\hat{\tau}(x)\big)\Big)^2.$$

Naively using the full data sample for estimation of both the nuisance functions as well as the CATE function leads to a higher bias due to overfitting but at the same time to lower variance as all available data is used for estimation. Using sample-splitting eliminates the overfitting bias but results in higher variance due to less data being used for estimation. In contrast, cross-fitting both removes the overfitting bias and reduces the variance by effectively using all the available information from the data for estimation. Figure 1.3.1 illustrates this theoretical argument by contrasting the distributions of the CATE parameter under full-sample estimation, double sample-splitting and double cross-fitting, resulting from a Monte Carlo simulation based on a large training sample of $32'000$ observations (further details on the meta-learner and the simulation design are provided in Sections 1.3.2 and 1.4, respectively). We observe that the theoretical arguments can be documented in finite samples too. As such, the full sample version exhibits substantial bias due to overfitting as its distribution is shifted away from the true value of the CATE parameter, but with a rather low variance. On the contrary, the double sample-splitting version successfully eliminates the overfitting bias as the simulated distribution is centered around the true value of the CATE, however with much larger variance. Finally, the double cross-fitting version keeps the reduction in bias whilst having a much lower variance in comparison to the double sample-splitting version as the spread of the CATE distribution comes close to the full sample version, indicating the gain in efficiency of this procedure.

[11] For results in the context of the Lasso estimation under sparsity see Belloni, Chernozhukov, Fernandez-Val, and Hansen (2017).

[12] Increasing the efficiency of a sample-splitting based estimator by swapping the roles of the data samples and averaging the resulting estimates goes back to Schick (1986) in the context of estimation of semi-parametric models.

Figure 1.3.1: CATE distributions under full-sample, sample-splitting and cross-fitting estimation.

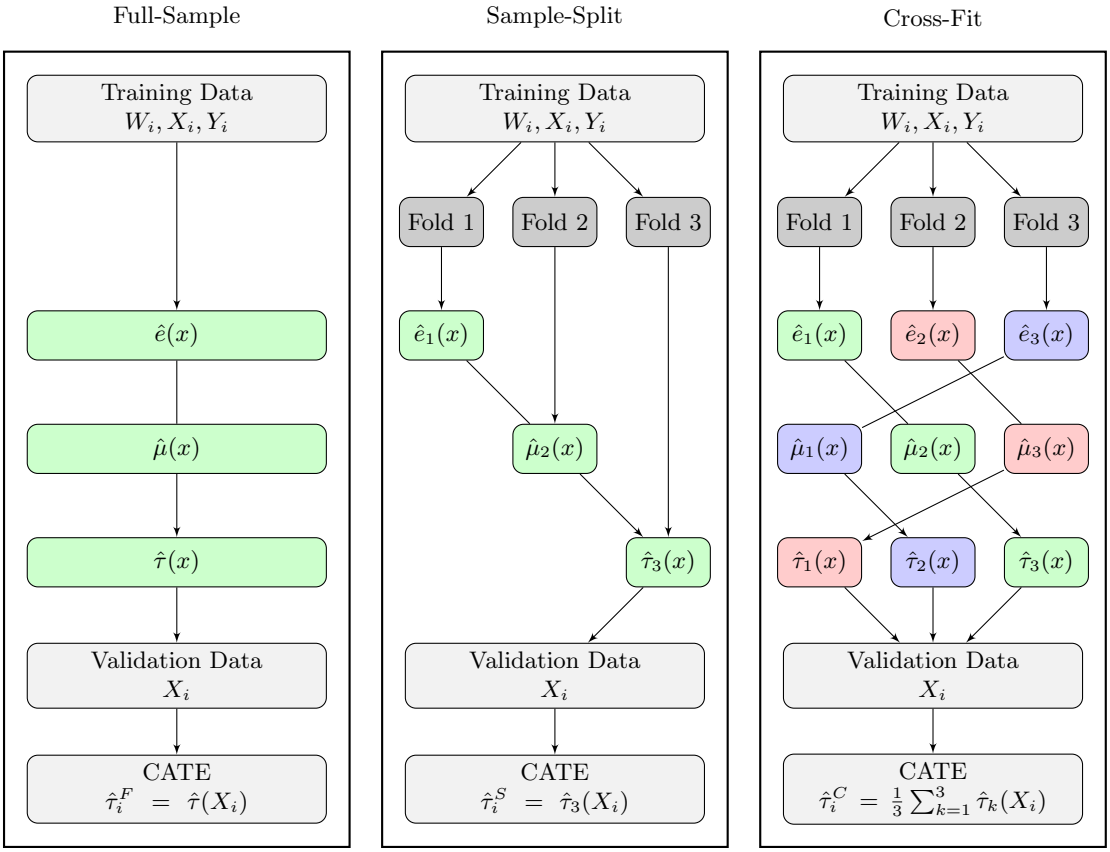


Note: Distributions of the CATE parameter under full-sample estimation (blue), double sample-splitting (red) and double cross-fitting (green) as a result of a Monte Carlo simulation. The CATE distributions are smoothed with the Gaussian kernel using the Silverman's bandwidth. The dashed black line defines the true value of the CATE while the solid black line plots the normal distribution around the true parameter with variance of the estimated CATE distribution. The CATEs are estimated by the DR-learner based on a training sample of $N^T = 32'000$ observations with 250 simulation replications and predicted out-of-sample. Detailed description of the simulation design is given in Section 1.4, while a detailed description of the DR-learner is given in Section 1.3.2.

Apart from the illustrative example above, the empirical question remains the precise quantification of this bias-variance trade-off for various meta-learners and to what degree this might vary with different sample sizes. Different meta-learners use different nuisance functions in different ways which might have an influence on the performance under the particular estimation procedures. Even though sample-splitting and cross-fitting help to eliminate the overfitting bias, in finite samples less data available for estimation might even lead to higher bias due to errors in learning the CATE function itself, especially for small sample sizes. In this paper we address this open question via Monte Carlo simulations and compare the performance of various meta-learners under full-sample, double sample-splitting and double cross-fitting procedure for several different sample sizes to shed more light onto the finite sample properties. We follow Newey and Robins (2018) and choose the double sample-splitting, respectively double cross-fitting procedure due to its theoretically faster convergence rates. Furthermore, we opt for the setting with equally sized $K = M + 1$ folds as suggested by Kennedy (2020). Additionally, we always distinguish between the training and validation data. We use the training data for learning the nuisance function and the CATE function, including the double sample-splitting and double cross-fitting procedure, while we evaluate the CATEs on a set of new validation data. An illustration of the data usage under full-sample estimation, double sample-splitting and double cross-fitting procedure is provided in Figure 1.3.2.

Further motivation for the usage of sample-splitting and cross-fitting stems from the theoretical arguments for conducting statistical inference about the causal parameters of interest. As such, sample-splitting plays a crucial role in obtaining estimators that are not only approximately unbiased but also normally distributed which in turn allows for a valid construction of confidence intervals. In this vein, Chernozhukov et al. (2018) provide results for the estimators of average treatment effect (ATE) that rely on sample-splitting and cross-fitting procedures. Semenova and Chernozhukov (2021) and Zimmert and Lechner (2019) extend this analysis for parametric and nonparametric estimators of group average treatment effects (GATEs), respectively. In the context of Causal Forests, Wager and Athey (2018), Lechner (2018), and Athey et al. (2019) also rely on sample-splitting procedures termed 'honesty' to provide inference for causal effects on various levels of aggregation. Nonetheless, in the context of meta-learning estimation of causal effects, there appears to be lack of unifying model-free theory for conducting statistical inference so far. One exception is the study by Künzel et al. (2019) that analyses various

Figure 1.3.2: Illustration of the full-sample, sample-splitting and cross-fitting procedure.



Note: Illustration of the full-sample (left), double sample-splitting (middle) and double cross-fitting (right) procedures with $K = 3$ folds. The propensity score function is defined by $e(x)$, the response functions in general are denoted by $\mu(x)$ and the CATE function is characterized by $\tau(x)$. Subscripts for the nuisance functions and the CATE function correspond to the fold used for estimation, while the colors indicate the combination of the estimated functions across different folds.

versions of bootstrapping for estimation of standard errors for the CATEs. Recently, Jacob (2021) makes use of such bootstrapping procedures to construct confidence intervals in an empirical application. Besides the computational burden, however, none of the bootstrapping procedures studied by Künzel et al. (2019) seems to reliably provide accurate coverage rates. However, the meta-learners analyzed in Künzel et al. (2019) do not make use of sample-splitting, which could potentially improve the performance of the bootstrapping for estimation of standard errors, given the insights from the related literature. While we do study the properties of the distribution of the CATEs within the simulation experiments in Section 1.4, we do not further analyse the estimation of standard errors mainly due to computational reasons and focus primarily on the point estimators. However, apart from the computational aspects, we note that combining sample-splitting and cross-fitting with bootstrapping for statistical inference about causal effects within the meta-learning framework might be a promising avenue for future research.

1.3.2 Meta-Learners

In the following, we review the meta-learning algorithms for estimation of heterogeneous treatment effects and discuss their advantages and disadvantages in particular empirical settings.

1.3.2.1 S-learner

The first meta-learning algorithm we investigate is the S-learner as denoted by Künzel et al. (2019). According to their naming convention, *S-* stands for *Single* as this meta-learner involves only one single model, namely the full response function, $\mu(x, w)$, that needs to be estimated. In the epidemiology literature the S-learner is also sometimes referred to as g-computation (Robins, 1986; Snowden, Rose, & Mortimer, 2011). The final causal effect is, in this case, obtained as a difference between predictions of the response function with setting the treatment indicator to, $W_i = 1$, and $W_i = 0$, respectively. The algorithm can be described as follows:[13]

Algorithm 1: S-LEARNER

Input: Training Data: $\{(X_i, Y_i, W_i)\}^T$, Validation Data: $\{(X_i)\}^V$

Output: CATE: $\hat{\tau}_S(x) = \hat{E}[Y_i(1) - Y_i(0) \mid X_i = x]$

begin

RESPONSE FUNCTION;

estimate: $\mu(x, w) = E[Y_i \mid X_i = x, W_i = w]$ in $\{(X_i, Y_i, W_i)\}^T$;

CATE FUNCTION;

define: $\hat{\tau}_S(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$;

predict: $\hat{\tau}_S(X_i) = \hat{\mu}(X_i, 1) - \hat{\mu}(X_i, 0)$ in $\{(X_i)\}^V$

end

As can be seen from Algorithm 1, the S-learner does not assign any special role to the treatment indicator W_i within the estimation procedure and uses it only *post hoc* in the computation of the causal effect. Thus, if the treatment indicator is not strongly predictive for the outcome the S-learner will tend to estimate a zero treatment effect.[14] Nevertheless, the S-learner will perform particularly well if the true CATE function is indeed zero, i.e. if $\tau(x) = 0$, which has also been documented in the simulation experiments of Künzel et al. (2019). For the forest based S-learner, Künzel (2019) proposes a modification of the algorithm such that it shrinks towards the ATE instead of zero by performing a Ridge regression in the final leaves of the trees within the forest.[15] In our simulations, we study a simpler modification of the forest based S-learner by always including the treatment indicator in the random subset of covariates when determining the splits. By doing so, we always give the S-learner the chance to split on the treatment indicator which might potentially alleviate the zero-bias issue. We will henceforth denote such learner as the SW-learner, where the W reflects the enforcement of the treatment indicator into the splitting set of covariates. We discuss the behaviour of the SW-learner more closely throughout the simulation results in Section 1.4.3. Furthermore, notice that the Algorithm 1 consists of only one nuisance function that needs to be estimated and thus does not require any sample-splitting or cross-fitting within the training sample induced by multiple nuisance functions, hence it always has access to the full sample of the training data.[16]

[13] As a matter of notation, we refer to the training data used for model estimation with superscript ^{T} as $\{(X_i, Y_i, W_i)\}^T$ and the validation data used for effect prediction with superscript ^{V} as $\{(X_i)\}^V$.

[14] Künzel et al. (2019) argue that the S-learner is actually biased towards zero.

[15] For a detailed explanation of this procedure see Künzel (2019).

[16] Nevertheless, an optional additional sample-splitting or cross-fitting could potentially improve the performance of the S-learner by reducing the possible overfitting of the base learner as such. This is, however, beyond the scope of our analysis and is left for future research.

1.3.2.2 T-learner

The T-learner is another common and widely used meta-learner that we investigate in our study. In the literature it is sometimes also called as the *basic* (Lechner, 2018), *plug-in* (Kennedy, 2020) or *naive* (Nie & Wager, 2021) CATE estimator. According to Künzel et al. (2019), *T-* stands for *Two* as this meta-learner involves two models that need to be estimated, defined by the treatment indicator W_i. These are namely the response function in the treated sample, $\mu(x, 1)$, and the response function in the control sample, $\mu(x, 0)$. This is in contrast to the above S-learner which pools the two response functions into a single one. However, similarly to the S-learner the causal effect is computed as a difference in predictions of the two response functions, which is motivated by the identification result as in Equation (1.2.4). The algorithm can be summarized as follows:[17]

Algorithm 2: T-LEARNER

Input: Training Data: $\{(X_i, Y_i, W_i)\}^T$, Validation Data: $\{(X_i)\}^V$

Output: CATE: $\hat{\tau}_T(X_i) = \hat{E}[Y_i(1) - Y_i(0) \mid X_i = x]$

begin

RESPONSE FUNCTIONS;

estimate: $\mu(x, 1) = E[Y_i \mid X_i = x, W_i = 1]$ in $\{(X_i, Y_i)\}^T_{W_i=1}$;

estimate: $\mu(x, 0) = E[Y_i \mid X_i = x, W_i = 0]$ in $\{(X_i, Y_i)\}^T_{W_i=0}$;

CATE FUNCTION;

define: $\hat{\tau}_T(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$;

predict: $\hat{\tau}_T(X_i) = \hat{\mu}(X_i, 1) - \hat{\mu}(X_i, 0)$ in $\{(X_i)\}^V$

end

Hence the T-learner uses the treatment indicator to split the estimation of the response function into two parts. This procedure is expected to work particularly well if the CATE function is complicated and there are no common trends in the response functions. This phenomenon finds supportive evidence in several simulation studies (see for example Künzel et al., 2019; Jacob, 2020; Curth & van der Schaar, 2021; or Nie & Wager, 2021). Nonetheless, it is expected to work rather poorly if the CATE function is simple, as the response functions are not trained jointly and thus their difference might be unstable (Lechner, 2018; Kennedy, 2020; Nie & Wager, 2021). In terms of the estimation of the nuisance functions, the T-learner behaves similarly to the S-learner, as only the response functions need to be estimated to compute the CATE. As such no additional sample-splitting induced by multiple nuisance functions is required as the response functions are themselves estimated on separate samples defined by treated and control.[18]

1.3.2.3 X-learner

The above mentioned problems of the T-learner are aggravated if the treatment assignment is highly unbalanced, meaning that the vast majority of observations in the sample belongs to only one treatment status. Künzel et al. (2019) therefore propose the X-learner which addresses this issue. The X-learner builds on the T-learner and, as such, first estimates the two response functions $\mu(x, 1)$ and $\mu(x, 0)$. It then uses these estimates to impute the unobserved individual treatment effects for the treated, $\tilde{\xi}^1_i$, and the control, $\tilde{\xi}^0_i$. The imputed effects are in turn used as pseudo-outcomes to estimate the treatment effects in

[17] Notationwise, we refer to a subset of the data defined by a specific value of the variable as for example $W_i = 1$ by a subscript as $\{(X_i, Y_i)\}_{W_i=1}$.

[18] Again, this does not preclude that an optional sample-splitting or cross-fitting might be beneficial for the same reason as in the case of the S-learner (Jacob, 2020, provides some results on this issue for the T-learner). Using an honest forest as a base learner would also add an implicit sample-splitting procedure, however, this is not analysed herein.

the treated sample, $\tau(x,1)$, and the control sample, $\tau(x,0)$, respectively. The final CATE estimate $\tau(x)$ is then a weighted average of these treatment effect estimates weighted by the propensity score, $e(x)$.[19] Thus the X-learner additionally uses the information from the treated to learn about the controls and vice-versa in a *Cross* regression style, hence the X term in its naming label. The learning algorithm can be detailed as follows:

Algorithm 3: X-LEARNER

Input: Training Data: $\{(X_i, Y_i, W_i)\}^T$, Validation Data: $\{(X_i)\}^V$

Output: CATE: $\hat{\tau}_X(X_i) = \hat{E}[Y_i(1) - Y_i(0) \mid X_i = x]$

begin

RESPONSE FUNCTIONS;

estimate: $\mu(x,1) = E[Y_i \mid X_i = x, W_i = 1]$ in $\{(X_i, Y_i)\}^T_{W_i=1}$;

estimate: $\mu(x,0) = E[Y_i \mid X_i = x, W_i = 0]$ in $\{(X_i, Y_i)\}^T_{W_i=0}$;

IMPUTED EFFECTS;

predict: $\tilde{\xi}^1_i = Y_i - \hat{\mu}(X_i, 0)$ in $\{(X_i, Y_i)\}^T_{W_i=1}$;

predict: $\tilde{\xi}^0_i = Y_i - \hat{\mu}(X_i, 1)$ in $\{(X_i, Y_i)\}^T_{W_i=0}$;

TREATMENT EFFECTS;

estimate: $\tau(x,1) = E[\tilde{\xi}^1_i \mid X_i = x, W_i = 1]$ in $\{(X_i, Y_i)\}^T_{W_i=1}$;

estimate: $\tau(x,0) = E[\tilde{\xi}^0_i \mid X_i = x, W_i = 0]$ in $\{(X_i, Y_i)\}^T_{W_i=0}$;

PROPENSITY SCORE;

estimate: $e(x) = P[W_i = 1 \mid X_i = x]$ in $\{(X_i, W_i)\}^T$;

CATE FUNCTION;

define: $\hat{\tau}_X(x) = \hat{e}(x) \cdot \hat{\tau}(x,0) + \big(1 - \hat{e}(x)\big) \cdot \hat{\tau}(x,1)$;

predict: $\hat{\tau}_X(X_i) = \hat{e}(X_i) \cdot \hat{\tau}(X_i,0) + \big(1 - \hat{e}(X_i)\big) \cdot \hat{\tau}(X_i,1)$ in $\{(X_i)\}^V$

end

According to Algorithm 3, the X-learner, in contrast to the T-learner, firstly uses the response functions for imputing the unobserved individual treatment effects instead of directly estimating the CATE. Secondly, these imputed individual treatment effects are used for estimating the CATE and reweighted by the propensity scores. The reweighting helps to put more weight on the treatment effects which have been estimated more precisely, i.e. the ones coming from the larger treated or control sample, respectively. For this reason, the X-learner is expected to work particularly well in unbalanced settings, which is often the case in practice as the share of treated might be restricted financially or otherwise (see Gerber, Green, & Larimer, 2008; Broockman & Kalla, 2016; or Goller, Lechner, Moczall, & Wolff, 2020, for such unbalanced empirical settings). Furthermore, by directly estimating the treatment effects in the second step it enables the estimator to learn structural properties of the CATE function from the data and is thus expected to work well if the CATE function is approximately linear or sparse (Künzel et al., 2019). In simulations of Künzel et al. (2019) the X-learner performs reasonably well even in other non-favourable settings. Notice further that Algorithm 3 requires more estimation steps than the previous two meta-learners. Additionally to the estimation of the response functions, the X-learner requires the estimation of the treatment effect functions as well as the propensity score function. This raises the question of possible overfitting and hence the need for sample-splitting and cross-fitting, respectively. However, there is theoretically no explicit requirement for sample-splitting in the case of the X-learner

[19] In the original definition of the X-learner, the estimation of the propensity score is not exactly specified as it could be any weighting function in general. However, in practice the estimation of the propensity score is recommended (Künzel et al., 2019).

when estimating the nuisance functions, apart from training and validation data split (Künzel et al., 2019). Yet, it might well be that the sample-splitting and further cross-fitting have a non-negligible influence on the performance of the learner in finite samples. We address this issue by implementing the double sample-splitting and double cross-fitting version of the X-learner in the simulation study. For the case of the full-sample estimation we use the out-of-bag predictions of the underlying forest as estimates of the nuisance functions. The out-of-bag predictions are based on the observations that have been left 'out of the bag' when drawing bootstrap samples to estimate the trees of the forest (Hastie et al., 2009). Such observations, however, randomly appear both as training as well as validation observations and thus such out-of-bag predictions are neither the classical in-sample fitted values nor proper out-of-sample predictions.[20]

1.3.2.4 DR-learner

Although the X-learner makes use of the estimation of multiple nuisance functions, it does not provide the double robustness property which exploits the fact that the estimator remains consistent if either the response function or the propensity score function is misspecified (Kennedy, Ma, McHugh, & Small, 2017; Lee, Okui, & Whang, 2017). Recently, Kennedy (2020) proposed the DR-learner where *DR* symbolizes the *Double Robustness* property of the learner. The DR-learner constructs a doubly robust score in the first estimation stage and estimates the CATE in the second stage. There have been many other versions of the DR-learner proposed in the literature, but these were restricted to a particular estimator used in the second stage and are thus not part of the meta-learning framework. For example, Semenova and Chernozhukov (2021) propose a linear estimation of the CATE function, whereas a local-constant estimation is proposed by Zimmert and Lechner (2019) and Fan, Hsu, Lieli, and Zhang (2020), which works well for the estimation of GATEs, i.e. for low-dimensional conditioning set. The main advantage of the DR-learner in comparison to the other versions lies in the general model-free second stage with sharper error bounds and weaker conditions for oracle efficiency (see Kennedy, 2020, for details). However, common to all versions in the literature is the estimation of the doubly robust score[21] by machine learning methods in the first stage also known as Double Machine Learning (Chernozhukov et al., 2018). For a comprehensive overview of the CATE estimators building on the doubly robust score see Knaus (2020). The specific algorithm for the DR-learner is then defined as follows:

[20] We use the out-of-bag predictions for all meta-learners within our analysis.
[21] Also called efficient score or efficient influence function in the literature (Robins & Rotnitzky, 1995; Hahn, 1998).

Algorithm 4: DR-learner

Input: Training Data: $\{(X_i, Y_i, W_i)\}^T$, Validation Data: $\{(X_i)\}^V$

Output: CATE: $\hat{\tau}_{DR}(x) = \hat{E}[Y_i(1) - Y_i(0) \mid X_i = x]$

begin

Response Functions;

estimate: $\mu(x,1) = E[Y_i \mid X_i = x, W_i = 1]$ in $\{(X_i, Y_i)\}^T_{W_i=1}$;

estimate: $\mu(x,0) = E[Y_i \mid X_i = x, W_i = 0]$ in $\{(X_i, Y_i)\}^T_{W_i=0}$;

Propensity Score;

estimate: $e(x) = P[W_i = 1 \mid X_i = x]$ in $\{(X_i, W_i)\}^T$;

Pseudo Outcome;

predict: $\hat{\psi}_i = \frac{W_i\left(Y_i - \hat{\mu}(X_i,1)\right)}{\hat{e}(X_i)} - \frac{(1-W_i)\left(Y_i - \hat{\mu}(X_i,0)\right)}{1-\hat{e}(X_i)} + \hat{\mu}(X_i,1) - \hat{\mu}(X_i,0)$ in $\{(X_i, Y_i, W_i)\}^T$;

CATE Function;

estimate: $\tau_{DR}(x) = E[\hat{\psi}_i \mid X_i = x]$ in $\{(X_i, Y_i, W_i)\}^T$;

predict: $\hat{\tau}_{DR}(X_i) = \hat{E}[\hat{\psi}_i \mid X_i = x]$ in $\{(X_i)\}^V$

end

As can be seen in Algorithm 4 above, the DR-learner estimates the very same nuisance functions, $\mu(x,0), \mu(x,1)$ and $e(x)$, as the X-learner but uses them in a completely different manner. It combines the nuisance functions as well as the outcome and treatment data in a doubly robust way to construct the pseudo-outcome ψ_i, i.e. the doubly robust score. The score is then regressed on the covariates to estimate the final CATE function. Therefore, the DR-learner can also adapt to structural properties of the CATE such as smoothness or sparsity. For this reason the DR-learner is expected to work well in similar situations as the X-learner with a more balanced treatment assignment, as too extreme propensity scores might possibly yield the estimator unstable (Huber et al., 2013; Powers et al., 2018), especially in high dimensions (D'Amour, Ding, Feller, Lei, & Sekhon, 2021). Moreover, it should have an additional advantage over the X-learner thanks to its double robustness property. The simulations of Kennedy (2020) also suggest a faster convergence rate of the DR-learner in comparison to the X- and T-learner. In order to achieve the optimal rates the DR-learner explicitly requires the double sample-splitting as defined by Newey and Robins (2018), while the double cross-fitting procedure remains optional. Theoretically it is not clear how important the role of the optional cross-fitting is for the DR-learner in finite samples and how much of the efficiency loss due to sample-splitting can be thereby regained. In order to shed light on this issue we investigate the implementations of the DR-learner with double sample-splitting, double cross-fitting, as well as a version with full-sample estimation.

1.3.2.5 R-learner

Yet another approach of first estimating nuisance functions and then using them to learn the treatment effects stems from the literature on partially linear model originally developed by Robinson (1988). Nie and Wager (2021) build on these ideas to flexibly estimate heterogeneous treatment effects and develop the R-learner, where the R stands for the recognition of the contribution of Robinson (1988) as well as for the *Residualization* approach. In the first step, the R-learner estimates the full response function, $\mu(x)$, similarly to the S-learner but without conditioning on the treatment indicator, as well as the propensity score function $e(x)$. It then residualizes the outcome and the treatment by the predictions of the response and the propensity score function, respectively, to construct a modified outcome. In the second step, the R-learner regresses the modified outcome on the covariates, weighted by the squared

residualized treatment[22], i.e. $\big(W_i - \hat{e}(X_i)\big)^2$, to estimate the CATE function (Schuler, Baiocchi, Tibshirani, & Shah, 2018). Analogous transformation of the outcome is also used by the Causal Forest of Athey et al. (2019) termed local centering, or in the G-estimation for sequential trials by Robins (2004). The full estimation procedure of the R-learner can be summarized as follows:

Algorithm 5: R-LEARNER

Input: Training Data: $\{(X_i, Y_i, W_i)\}^T$, Validation Data: $\{(X_i)\}^V$

Output: CATE: $\hat{\tau}_R(x) = \hat{E}[Y_i(1) - Y_i(0) \mid X_i = x]$

begin

RESPONSE FUNCTION;

estimate: $\mu(x) = E[Y_i \mid X_i = x]$ in $\{(X_i, Y_i)\}^T$;

PROPENSITY SCORE;

estimate: $e(x) = P[W_i = 1 \mid X_i = x]$ in $\{(X_i, W_i)\}^T$;

MODIFIED OUTCOME;

predict: $\hat{\phi}_i = \frac{\big(Y_i - \hat{\mu}(X_i)\big)}{\big(W_i - \hat{e}(X_i)\big)}$ in $\{(X_i, Y_i, W_i)\}^T$;

CATE FUNCTION;

estimate: $\tau_R(x) = E[\hat{\phi}_i \mid X_i = x]$ weighted by $\big(W_i - \hat{e}(X_i)\big)^2$ in $\{(X_i, Y_i, W_i)\}^T$;

predict: $\hat{\tau}_R(X_i) = \hat{E}[\hat{\phi}_i \mid X_i = x]$ in $\{(X_i)\}^V$

end

As follows from Algorithm 5, the R-learner separates the estimation into two steps. First, it eliminates the spurious correlations between the response function $\mu(x)$ and the propensity score function $e(x)$ and second, it optimizes the CATE function $\tau_R(x)$. From this standpoint the R-learner follows a related estimation scheme as the DR-learner and is expected to work well in similar settings where the nuisance functions as well as the CATE function might have a high degree of complexity. A possible advantage of the R-learner over the DR-learner might stem from the additional weighting which reduces the impact of extreme propensity scores as pointed out by Jacob (2021). In their simulation experiments, Nie and Wager (2021) show good performance of the R-learner in settings with complicated nuisance functions and rather simple CATE function. Furthermore, for the theoretical results Nie and Wager (2021) explicitly require sample-splitting and cross-fitting, respectively. In particular, they advocate for a 5- or 10-fold cross-fitting procedure as defined by Chernozhukov et al. (2018). In order to examine the importance of the cross-fitting in finite samples we compare the performance of the R-learner as in the above cases with full-sample estimation, double sample-splitting and double cross-fitting, respectively.

1.4 Simulation Study

We study the finite sample performance of meta-learners for estimation of heterogeneous treatment effects based on Random Forests (Breiman, 2001; see also Biau & Scornet, 2016, for a comprehensive introduction). The focus of the Monte Carlo study lies in an assessment of the influence of sample-splitting and cross-fitting in the causal effect estimation. For this purpose we compare the above discussed meta-learners estimated with full-sample, double sample-splitting, and double cross-fitting. We rely on the Random Forest as the base learner for all meta-learners for several reasons. First, different meta-learners

[22] An estimation procedure without the weighting step is in literature referred to as the U-learner (Stadie et al., 2018; Künzel et al., 2019; Nie & Wager, 2021). However, such estimator turned out to be quite unstable in the simulation experiments in Nie and Wager (2021) as well as in those of Künzel et al. (2019) and will thus not be considered further in our analysis.

require estimation of different nuisance functions which involve different types of outcome variables. As such, the response functions mostly involve a continuous outcome variable whereas the propensity score function includes a binary outcome. Hence, when using Random Forests no additional adjustments need to be done in terms of estimation as it automatically estimates probabilities in case of binary outcome and expected values in case of continuous outcomes, respectively. This is in contrast to linear learners such as the Lasso (Tibshirani, 1996), Ridge (Hoerl & Kennard, 1970) or Elastic Net (Zou & Hastie, 2005) where the estimator needs to be modified using appropriate link function for proper probability estimation (see for example Hastie et al., 2009, for the Logit-Lasso). Second, Random Forest is a local nonparametric method which does not need any data pre-processing to flexibly learn the underlying functional form from the data (Hastie et al., 2009). Thus, Random Forest is able to approximate any function with different degrees of complexity which is often the case in treatment effect estimation where the nuisance functions tend to be rather difficult complex functions while the CATE function itself is often simple and sparse (Künzel et al., 2019; Kennedy, 2020). This is again an advantage in comparison to the linear learners mentioned above which become more flexible once an augmented covariate set consisting of polynomials and interactions is created and thus can be regarded as global nonparametric methods (Hastie et al., 2009). Third, in contrast to other flexible state-of-the-art machine learners such as Neural Networks the theoretical properties of Random Forest are better understood which makes it less of a black-box method and thus more amenable to conduct statistical inference (see Meinshausen, 2006; Biau, 2012; Wager, Hastie, & Efron, 2014; Wager, 2014; Scornet, Biau, & Vert, 2015; Mentch & Hooker, 2016; Wager & Athey, 2018; Athey et al., 2019, for a discussion of statistical properties of Random Forests). Additionally, another reason why we do not use the Lasso and the linear learners as such is due to a substantial increase in variance as they are prone to outliers as documented in the simulation studies of Jacob (2020) as well as Knaus et al. (2021). Lastly, from the practical standpoint there is a vast variety of fast and reliable software implementations of Random Forests which makes it easy to use for practitioners.[23]

In order to objectively evaluate the performance and the robustness of different meta-learners in estimating heterogeneous treatment effects with regard to the double sample-splitting and double cross-fitting, we design several simulation scenarios. On the one hand, for each meta-learner we construct such a data generating process (DGP) that suits the particular advantages of the respective meta-learner, i.e. we design a simulation scenario where each meta-learner is expected to work best. Hence, we are able to check if the particular meta-learner outperforms the others and how big the performance discrepancies are for the other meta-learners in comparison to the expected best performing meta-learner. On the other hand, we design a challenging scenario where none of the meta-learners has *a priori* an explicit advantage, which serves as our main simulation design of interest. Thus we can compare the performance of the meta-learners in an objective manner and quantify the deviations to their respective best performance cases. Furthermore, common to all DGPs is the observational study design, i.e. there is always selection into treatment and thus all considered meta-learners have to deal with confounding and not only with modelling the treatment effect itself. Moreover, in contrast to many simulation studies where the nuisances are simple low-dimensional functions (Wager & Athey, 2018; Künzel et al., 2019; Kennedy, 2020), we model all nuisance functions as highly non-linear but sparse functions with large-dimensional covariate space to both challenge the potential of the machine learning methods, though still largely obeying the theory induced limitations. For other challenging simulation designs see also Jacob (2020) or Zivich and Breskin (2021) as well as Lechner (2018) and Knaus et al. (2021) for the Empiri-

[23] In our simulations we use the R-package ranger which provides a fast C++ implementation of Random Forests, particularly suited for high-dimensional data (Wright & Ziegler, 2017). Further options include the grf package written by Tibshirani et al. (2018), the forestry package by Künzel, Liu, Saarinen, Tang, and Sekhon (2020) or the randomForest package by Liaw and Wiener (2002).

cal Monte Carlo Simulations. Importantly, in order to study the approximate convergence rates of the meta-learners we repeat each simulation scenario several times with increasing training sample sizes using $N^T = \{500, 2'000, 8'000, 32'000\}$. We emphasize that the considered sample sizes substantially exceed the ones from previous simulation studies devoted to the analysis of sample-splitting methods, which were limited to $2'000$ (Jacob, 2020) and $3'000$ (Zivich & Breskin, 2021) observations, respectively. Furthermore, the large samples enable us to study the performance of the meta-learners in settings in which the application of machine learning methods is arguably more relevant. We choose to always quadruple the sample size, which allows us to easily benchmark the results with the parametric $\sqrt{N}$ rate, in which case the estimation error is expected to halve with each increase of the sample size. We then evaluate the performance measures on a validation set with sample size of $N^V = 10'000$ to reduce the prediction noise as is usual in many Monte Carlo studies (Janitza, Tutz, & Boulesteix, 2016; Hornung, 2019; Lechner & Okasa, 2019; Jacob, 2020; Knaus et al., 2021). Lastly, in terms of the tuning parameters for the Random Forest base-learner we stick to the default, in the literature commonly used settings, corresponding to $1'000$ trees, the number of randomly chosen split variables set to the square root of number of features, and the minimum leaf size equal to 5.[24] Finally, for each DGP we simulate the training data $R = \{2'000, 1'000, 500, 250\}$ times in total, where we use $2'000$ replications for the smallest sample size and decrease the number of replications down to 250 for the largest sample size, due to computational reasons.[25]

1.4.1 Performance Measures

For the evaluation of the performance of the considered meta-learners with regard to the sample-splitting and cross-fitting in detail, we employ several evaluation measures. First, to assess the overall estimator performance we compute the root mean squared error for each observation i from the validation sample over the R simulation replications:[26]

$$RMSE\big(\hat{\tau}(X_i)\big) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\big(\tau(X_i) - \hat{\tau}^r(X_i)\big)^2}.$$

Next, we decompose the root mean squared error and evaluate the bias and variance component separately to contrast the theoretically expected asymptotic behaviour of sample-splitting and cross-fitting with the finite sample properties. Hence, we additionally compute the mean absolute bias:

$$\mid BIAS\big(\hat{\tau}(X_i)\big) \mid = \frac{1}{R}\sum_{r=1}^{R} \mid \tau(X_i) - \hat{\tau}^r(X_i) \mid$$

as well as the standard deviation of the treatment effects:

$$SD\big(\hat{\tau}(X_i)\big) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\tau}^r(X_i) - \frac{1}{R}\sum_{r=1}^{R}\hat{\tau}^r(X_i)\right)^2}.$$

[24] We refrain from cross-validation or other tuning parameter optimization procedures due to computational constraints. We recommend such optimization in the applied work as it might considerably improve the performance of the estimator (see Curth & van der Schaar, 2021, for an evidence based on Neural Networks), however, for the purposes of the simulation study it would not change the relative ranking of the meta-learners as each of them uses the very same base learner.

[25] Notice, however, that we only halve the number of replications while quadrupling the sample size and as such we may limit a possible deterioration of the performance in terms of the simulation error. A similar strategy for balancing the precision and the computational burden has been used in the simulations by Lechner (2018) or Knaus et al. (2021). Detailed results on the simulation error are provided in Appendix 1.B.2.

[26] We take the square root of the MSE to have the same scale as for the other performance measures, i.e. the absolute bias and the standard deviation.

Furthermore, inspired by the simulation study of Knaus et al. (2021) we also compute the Jarque-Bera statistic (Jarque & Bera, 1980; Bera & Jarque, 1981) to test for the normality of the treatment effect predictions:[27]

$$JB\big(\hat{\tau}(X_i)\big) = \frac{R}{6}\left(S\big(\hat{\tau}(X_i)\big)^2 + \frac{1}{4}(K\big(\hat{\tau}(X_i)\big) - 3)^2\right)$$

where $S\big(\hat{\tau}(X_i)\big)$ and $K\big(\hat{\tau}(X_i)\big)$ is the skewness and the kurtosis of the R treatment effect predictions for observation i, respectively. As a matter of presentation for CATEs, we report the mean values of the RMSE, absolute bias, standard deviation and the Jarque-Bera statistic over N^V validation observations.[28] Additionally, we provide evaluation of further performance measures in Appendix 1.B.2.

1.4.2 Simulation Design

In the general simulation design we follow Künzel et al. (2019) and specify the response functions for potential outcomes under treatment, $\mu_1(x)$, and control, $\mu_0(x)$, the propensity score, $e(x)$, and the treatment effect function, $\tau(x)$, respectively. First, we simulate a p-dimensional matrix of covariates $X_i \in \mathbb{R}^p$ drawing from the uniform distribution, as previously used in simulations of Wager and Athey (2018), Künzel et al. (2019) or Nie and Wager (2021) among others, such that:

$$X_i \sim \mathcal{U}\big([0,1]^{n\times p}\big)$$

and defining the correlation structure according to Falk (1999) using a random correlation matrix Σ_p generated by the method of Joe (2006).[29] Second, we specify the response functions and simulate the potential outcomes as:

$$Y_i(0) = \mu_0(X_i) + \epsilon_i(0)$$

$$Y_i(1) = \mu_1(X_i) + \epsilon_i(1)$$

with errors $\epsilon_i(0), \epsilon_i(1) \overset{iid}{\sim} \mathcal{N}(0,1)$ that are independent of the covariates X_i. Third, we define the propensity score function and simulate the treatment assignment according to:

$$W_i \sim Bern\big(e(X_i)\big)$$

and use the observational rule to set the observed outcomes such that:

$$Y_i = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0)$$

to complete the observable triple $\{(X_i, Y_i, W_i)\}$. The subsequent simulation designs then differ only with respect to how the corresponding functions, namely $\mu_0(x), \mu_1(x), e(x)$ and $\tau(x)$ are specified. For all of our simulations we define the response function under non-treatment according to the well-known Friedman function (1991) to create a difficult yet standardized setting, which has been used also in the simulations of Nie and Wager (2021), as follows:

$$\mu_0(x) = sin\big(\pi \cdot x_1 \cdot x_2\big) + 2 \cdot \big(x_3 - \frac{1}{2}\big)^2 + x_4 + \frac{1}{2} \cdot x_5 \tag{1.4.1}$$

[27] See Thadewald and Büning (2007) for a discussion of the Jarque-Bera test and its comparison to other tests for normality.

[28] As such, we define the average RMSE as $\overline{RMSE} = \frac{1}{N^V}\sum_{i=1}^{N^V} RMSE\big(\hat{\tau}(X_i)\big)$ and analogously for the remaining performance measures. Additionally, for the Jarque-Bera statistic we report also the share of CATEs from the validation sample for which the normality gets rejected at the 5% confidence level. For details, see Appendix 1.B.2.

[29] For a detailed correlation heat map of the covariates and further descriptive statistics of the simulated datasets see Appendix 1.A.

hence effectively creating a highly non-linear but sparse response function which is challenging to estimate on its own.[30] The response function under treatment is then defined simply as:

$$\mu_1(x) = \mu_0(x) + \tau(x)$$

while we vary the specification of the treatment effect function $\tau(x)$ throughout our simulation designs. Lastly, we model the propensity score function similarly to Wager and Athey (2018) and Künzel et al. (2019) using the β distribution with parameters 2 and 4 such that:

$$e(x) = \alpha\Big(1 + \beta_{2,4}\big(f(x)\big)\Big) \tag{1.4.2}$$

while the scale parameter α controls the share of treated in the sample and at the same time helps to bound the resulting probabilities away from 0 and 1 and thus to avoid extreme propensity scores which might yield some meta-learners using such propensities for reweighting unstable (Huber et al., 2013; Powers et al., 2018). We additionally make the propensity score dependent on features X_i of dimension p^e in a non-linear fashion using the functional form specification of Nie and Wager (2021) and set:

$$f(x) = sin(\pi \cdot x_1 \cdot x_2 \cdot x_3 \cdot x_4)$$

which creates a non-linear setting that is challenging to model as opposed to, e.g. polynomial transformations alone. Similarly, such non-linear transformations for the propensity scores using the sine function have been used also in simulations by Lechner (2018) and Knaus et al. (2021).

Table 1.4.1: Overview of the Simulation Study

General Settings	
Number of DGPs	6
Number of Replications R	$\{2'000, 1'000, 500, 250\}$
Training Sample N^T	$\{500, 2'000, 8'000, 32'000\}$
Validation Sample N^V	$10'000$
DGP Settings	
Covariate Space Dimension p	100
Signal Covariates in Response Function p^μ	5
Signal Covariates in Propensity Score Function p^e	4
Signal Covariates in Treatment Function p^τ	$\{0, 1, 2, 3\}$
Forest Settings	
Number of Trees	$1'000$
Random Subset of Split Covariates	$\sqrt{p}$
Minimum Leaf Size	5

As a matter of notation we refer to p as the dimension of the covariate space, p^μ, p^e and p^τ as the dimension of the signal covariates in the response function, the propensity score function, and the CATE function, respectively. We set the aforementioned dimensions as follows: $p = 100$, $p^\mu = 5$, $p^e = 4$ and p^τ varies with forthcoming simulation designs. We note that such large-dimensional covariate set is quite unique as the majority of simulation studies relies on low-dimensional covariate sets (see e.g. Künzel et al., 2019; Jacob, 2020; or Nie & Wager, 2021).[31] We further define the sets of covariates such that $X^{p^\tau} \subset X^{p^e} \subset X^{p^\mu} \subset X^p$. By doing so we make it difficult for the meta-learners to accurately fit the

[30] Note that π refers to the mathematical constant, i.e. $\pi \approx 3.14$.

[31] An exception is the simulation study of Powers et al. (2018) who explicitly study the estimation of heterogeneous treatment effects in high-dimensions.

functions and eliminate the spurious correlations between the response and propensity score functions. Moreover, it also becomes challenging to disentangle the confounding effects from the actual treatment effect heterogeneity which the herein discussed meta-learners are specifically designed for. Finally, a general overview of the simulation study is provided in Table 1.4.1.

1.4.2.1 Simulation 1: balanced treatment and constant zero CATE

The first simulation design features our complicated sparse non-linear nuisance functions as defined above in Equations (1.4.1) and (1.4.2) in contrast to a very simple CATE function. In fact, the treatment effect here is defined as being constant and equal to zero:

$$\tau(x) = 0$$

with a balanced treatment assignment with the scaling factor $\alpha = \frac{1}{4}$ which results in approximately 50% treated and 50% of control units. Such DGP with zero CATE serves as a benchmark and should implicitly suit the S-learner as the treatment indicator is not predictive for the outcome. Nevertheless the other meta-learners with the exception of the T-learner should be also capable of capturing the true zero effect as this is often a showcase example when motivating the particular meta-learners as well as simulating their performance (see Künzel et al., 2019; Kennedy, 2020; and Nie & Wager, 2021, for details).

1.4.2.2 Simulation 2: balanced treatment and complex nonlinear CATE

In the second simulation design we keep the balanced treatment allocation but feature a highly complex and non-linear CATE function resulting from a completely disjoint DGPs for the response function under treatment and under control. As such the response function under control is defined according to Equation (1.4.1), while the response function under treatment is defined as a non-zero constant, i.e. $\mu_1(x) = 1$. The CATE is then defined as:

$$\tau(x) = \mu_1(x) - \mu_0(x).$$

Such simulation setups have been previously used also in Künzel et al. (2019) or in Nie and Wager (2021). In this case the response functions, $\mu_0(x)$ and $\mu_1(x)$, are uncorrelated and thus there is no advantage in pooling those two together. Rather, estimating these two functions separately is the best strategy as there is nothing additional to learn from the other treatment group. For this reason, the T-learner should perform best here, however the meta-learners which also estimate the response functions separately such as the X- and DR-learner are expected to perform well too. Clearly, other meta-learners such as the S- and R-learner which estimate the pooled response function have a disadvantage as they first need to learn the disjoint structure.

1.4.2.3 Simulation 3: highly unbalanced treatment and constant non-zero CATE

In our third simulation design we change the scaling factor in the propensity score function to $\alpha = \frac{1}{12}$ such that we generate approximately 15% treated units.[32] We then model the treatment effect

[32] In contrast to Künzel et al. (2019) we do not specify the treatment imbalance as extreme as 1% treated mostly for computational reasons. Due to our smallest sample size of $N = 500$ used in the simulations and the double sample-splitting procedure, it might occasionally happen that the estimated propensity scores would be exactly zero which would prevent estimation of the DR-learner as well as the R-learner due to the division by zero when constructing the pseudo-outcomes. In our specification, even with the share of the treated being 16.77% in expectation, the aforementioned issue with zero propensity scores still might occur. In such cases, we redraw the sample to ensure at least 15% of treated. However, this

as a constant as for example in Kennedy (2020) or Nie and Wager (2021) and thus create a scenario with highly complicated nuisance functions and very simple CATE function given as:

$$\tau(x) = 1.$$

Accordingly, the X-learner should perform best in this scenario given the high imbalance in the treatment assignment and the sparse CATE function at the same time. In contrast, other meta-learners using the propensity score for reweighting such as the DR- and R-learner might perform worse due to potentially extreme propensity scores close to the $\{0, 1\}$ bounds. Furthermore, the T-learner is clearly disadvantaged in this scenario due to the high treatment imbalance as well as due to the simple CATE function, whereas the S-learner is not expected to work particularly well either due to the relatively bigger effect size bounded away from zero.

1.4.2.4 Simulation 4: unbalanced treatment and simple CATE

In our fourth simulation design we model the CATE function similarly to the above design as a simple non-zero constant and combine it with an indicator function as also used by Künzel et al. (2019) to add more structure to the CATE. As such we define the treatment effect as:

$$\tau(x) = 1 + 1 \cdot \mathbb{1}(x_1 > 0.5)$$

and otherwise keep the DGP same as in the third design while only increasing the share of treated to roughly 25% as is the case in the simulations of Nie and Wager (2021) by setting $\alpha = \frac{1}{8}$. By doing so we should theoretically shift the advantage from the X-learner more onto the DR-learner as both meta-learners are motivated by complex nuisance functions and a simple CATE function with the difference of the X-learner being designed particularly for highly unbalanced treatment allocation. Also the R-learner is expected to perform relatively well in this scenario due to the less unbalanced treatment shares, whereas the S- and T-learner are not expected to perform well for the same reasons as in the above situations.

1.4.2.5 Simulation 5: unbalanced treatment and linear CATE

The fifth simulation design features the same nuisance functions and treatment share as the fourth design, however, here instead of the indicator function we model the treatment effect as a low-dimensional linear function as:

$$\tau(x) = 1 + \frac{1}{2}x_1 + \frac{1}{2}x_2$$

as used in one of the simulation designs of Nie and Wager (2021) where the R-learner performed best and as such it is also expected to have an advantage here. Yet again the X- and DR-learner should perform comparatively well in this setting while the S- and T-learner not so much for the very same reasons as stated above.

1.4.2.6 Main Simulation: unbalanced treatment and nonlinear CATE

In the last simulation design we create arguably the most challenging scenario in which none of the meta-learners has an *a priori* advantage and thus presents our main simulation design of interest. In this case not only the nuisances but also the CATE itself is modelled as a smooth non-linear function of a

occurs only a handful of times out of 2000 draws in total and only for the smallest sample size considered. Nie and Wager (2021) also use similar restrictions on the propensity scores in their simulations due to the very same issue.

slightly larger dimension than in the previous settings, i.e. $p^\tau = 3$. Following Wager and Athey (2018) we specify the CATE function as follows:

$$\tau(x) = 1 + \frac{4}{p^\tau} \sum_{j=1}^{p^\tau} \left(\frac{1}{1 + e^{-12(x_j - 0.5)}} - \frac{1}{2} \right).$$

We further keep the treatment share equal to about 25% and the nuisance functions as previously specified as well. Hence, the meta-learners need to first account for the moderately imbalanced treatment shares, second accurately estimate the complex nuisance functions and disentangle their correlation, and third separate the treatment effect heterogeneity from the selection effects by precisely estimating the non-linear CATE function.

1.4.3 Simulation Results

For the analysis of the simulation results we focus on the Main Simulation design with unbalanced treatment assignment and nonlinear CATE function as this is arguably the most challenging simulation design which does not *a priori* create conditions that would be advantageous for any of the considered meta-learners. We then summarize the results for the rest of the simulation designs for which we provide the detailed results in Appendix 1.B.1. Supplementary results providing additional measures, including the simulation error, bias, skewness, kurtosis, share of CATEs for which the normality has been rejected, as well as the correlation and variance ratio of the estimated and the true CATEs are presented in Appendix 1.B.2.

1.4.3.1 Results of Main Simulation: unbalanced treatment and nonlinear CATE

The performance of the meta-learners in the Main Simulation is depicted in Table 1.4.2. We report the results for the average values of the RMSE, absolute bias, standard deviation and the Jarque-Bera test statistic over the $N^V = 10'000$ predicted CATEs from the validation sample. Figure 1.4.1 details the performance of the meta-learners implemented in the full-sample, double sample-splitting and double cross-fitting versions.

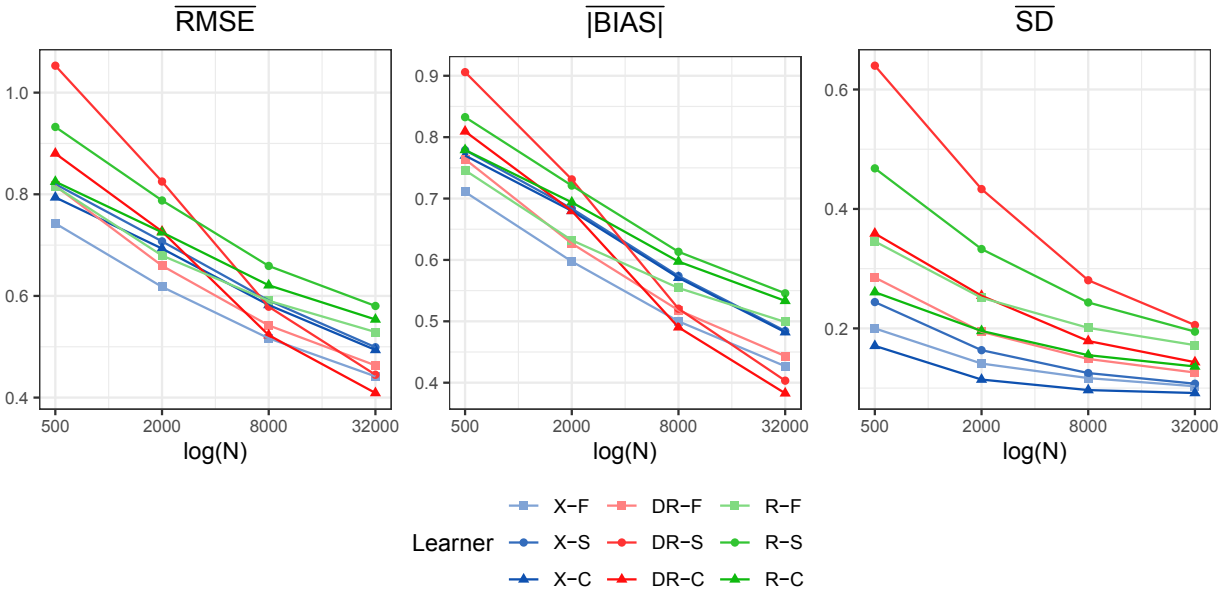
Table 1.4.2: CATE Results for Main Simulation

	$\overline{RMSE}$				$\overline{\lvert BIAS \rvert}$				$\overline{SD}$				$\overline{JB}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.878	0.749	0.651	0.570	0.867	0.739	0.641	0.560	**0.108**	**0.096**	**0.091**	**0.088**	7.140	2.888	2.173	1.936
S-W	0.765	0.634	0.533	0.462	0.717	0.602	0.508	0.443	0.261	0.190	0.149	0.125	**2.086**	2.106	2.019	1.931
T	0.766	0.634	0.533	0.462	0.719	0.602	0.509	0.442	0.260	0.190	0.149	0.125	2.603	**2.085**	2.016	1.924
X-F	**0.743**	**0.618**	**0.517**	0.442	**0.711**	**0.597**	0.500	0.427	0.200	0.141	0.117	0.103	3.490	2.230	2.034	1.857
X-S	0.820	0.707	0.591	0.499	0.779	0.684	0.574	0.484	0.244	0.164	0.125	0.107	5.146	2.680	2.157	1.929
X-C	0.794	0.693	0.582	0.494	0.770	0.680	0.571	0.482	0.171	0.114	0.097	0.092	3.984	2.322	**1.964**	**1.827**
DR-F	0.817	0.659	0.542	0.463	0.764	0.627	0.518	0.443	0.285	0.194	0.149	0.126	141.106	40.528	5.490	2.172
DR-S	1.053	0.825	0.579	0.445	0.906	0.731	0.521	0.403	0.640	0.433	0.281	0.206	567.501	458.729	159.041	36.504
DR-C	0.880	0.727	0.523	**0.409**	0.809	0.680	**0.490**	**0.383**	0.359	0.255	0.179	0.143	52.224	38.216	12.644	3.162
R-F	0.815	0.679	0.590	0.529	0.746	0.632	0.554	0.499	0.346	0.251	0.201	0.172	4.583	3.499	2.225	1.983
R-S	0.932	0.788	0.659	0.580	0.833	0.721	0.613	0.546	0.468	0.333	0.243	0.195	3.959	3.365	2.666	2.028
R-C	0.825	0.725	0.621	0.554	0.779	0.694	0.597	0.533	0.261	0.196	0.155	0.136	2.416	2.184	2.036	1.959

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{SD}$ and $\overline{JB}$ show the mean values of the root mean squared error, absolute bias, standard deviation and the Jarque-Bera test statistic of all 10′000 CATE estimates from the validation sample. The critical values for the JB test statistic are 5.991 and 9.210 at the 5% and 1% level, respectively. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively. Bold numbers indicate the best performing meta-learner for given measure and sample size.

Starting with the most simple S-learner, we observe a competitive performance in terms of the average RMSE for the smaller sample sizes which, however, disappears for larger sample sizes. Taking

Figure 1.4.1: CATE Results for Main Simulation



Note: The results for $\overline{RMSE}$, $\overline{|BIAS|}$, and $\overline{SD}$ show the mean values of the root mean squared error, absolute bias, and standard deviation of all 10′000 CATE estimates from the validation sample. The figure shows the results based on the increasing training samples of $\{500, 2'000, 8'000, 32'000\}$ observations displayed on the log scale. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

a closer look at the results reveals that the competitive performance of the S-learner stems mainly from the very low standard deviation while being substantially biased. Indeed, the variance of the S-learner is the smallest among all meta-learners for all sample sizes. This is mainly due to its tendency to predict effects close to zero if the treatment indicator is not strongly predictive for the outcome as pointed out by Künzel et al. (2019). This explains also the high bias of this estimator as the CATEs vary between -1 and 3 with only a small proportions of the CATEs being equal to zero (see Figure 1.A.6 in Appendix 1.A for details). Nevertheless, the Jarque-Bera test does not indicate evidence against the normality of the predicted CATE distribution, on average.

Considering the modified version of the S-learner with enforcement of the treatment indicator into the forest splitting set, i.e. the SW-learner, we notice almost identical performance to the one of the T-learner. This result can be explained by an observation that once the SW-learner finds the split based on the treatment indicator early within the trees it mimics the disjoint structure of the T-learner. The rest of the recursive partitioning is then very similar to the one of the T-learner which has been also documented for the case of the S-learner in the simulation experiments conducted by Künzel et al. (2019). As a result, it seems that enforcing the treatment indicator into the splitting set helps to alleviate the high bias of the S-learner to some degree, however, it increases the variance of the estimator at the same time. Nevertheless, the bias-variance trade-off in this case results in lower average RMSE in comparison to the S-learner and the SW-learner might thus be preferred over the simple S-learner, when using Random Forest as a base learner. Overall, the SW- and T-learner are very competitive in the smaller sample sizes both in terms of the average RMSE as well as the average absolute bias. However, with access to more training data these two learners do not improve fast enough and are outperformed by the more sophisticated learners in the largest sample size consisting of 32′000 observations. Concerning the distribution of the predicted CATEs there seems to be on average no statistical evidence against the normality, neither for the SW-learner nor for the T-learner.

In contrast to the above mentioned meta-learners the X-learner makes use of the additional estimation of nuisance functions. In its full-sample version the X-learner performs best in terms of the average RMSE for all sample sizes, except the largest one. The good RMSE performance stems partly from the relatively low bias and partly from the relatively low variance of the estimator as the X-learner exhibits the smallest average absolute bias for the smaller sample sizes (500 and 2′000), while having one of the lowest average standard deviations throughout all sample sizes. Interestingly, we only partly document the theoretical properties regarding the sample-splitting and cross-fitting procedures. As such, the full-sample version is the best performing one in terms of the average RMSE as well as in terms of the average absolute bias across all sample sizes, which is a pattern observed in the simulation experiments of Jacob (2020) as well. Accordingly, the sample-splitting version exhibits not only higher values of the average standard deviation but also higher values of the average absolute bias. Nevertheless, we observe that the cross-fitting version successfully regains the efficiency lost due to sample-splitting as it exhibits steadily lower variance than the sample-splitting version and even lower variance than the full-sample version, while having a bias of roughly the same magnitude as the sample-splitting version. As for the distribution of the predicted CATEs, on average, we do not observe evidence for deviations from normality for any of the versions of the X-learner. Additionally, we do not observe any major differences in the speed of convergence between the different versions as can be seen in Figure 1.4.1. Moreover, the absolute differences in the performance measures among the different versions are small in comparison to other meta-learners using nuisance functions. Albeit rather surprising at the first sight, the explanation for this phenomenon comes presumably from the different usage of the propensity score by the X-learner in comparison to the R- and DR-learner. As such, the R- and DR-learner use the propensity score together with the response functions to construct a new pseudo-outcome which is subsequently used to estimate the CATEs. In contrast, the X-learner uses merely the response functions to create the pseudo-outcome, while the propensity score is used only to reweight the final CATE estimates and thus it does not enter into any additional estimation step. Therefore, the X-learner might be less prone to overfitting bias which would partly justify the full-sample estimation as described in Künzel et al. (2019).[33]

Assessing the performance of the DR-learner reveals some interesting insights. The first observation is that the cross-fitting version performs best of all meta-learners in terms of the average RMSE for the largest sample size of 32′000 observations. This comes mainly from the low bias of this estimator as the average absolute bias is the lowest among all learners for the two largest sample sizes, while the average standard deviation is relatively high. However, looking at the average value of the Jarque-Bera statistic suggests evidence against the normality of the predicted CATEs for all but the largest sample size. Inspecting the results more closely reveals that the issue stems from heavy tails of the CATE distributions. The extreme values of the predicted CATEs are mainly caused by the propensity scores which are close to the $\{0, 1\}$ bounds. Similar issues of the DR-learner due to extreme propensity scores have also been documented in the simulation experiments of Knaus et al. (2021) as well as in the empirical application of Knaus (2020). The second observation is that for the DR-learner we clearly see how the theoretical arguments of sample-splitting and cross-fitting translate into the finite sample properties of the estimator. However, these can be documented only for large sample sizes. As such, the bias of the sample-splitting version is smaller than the one of the full-sample version in the largest sample size, while the bias of the cross-fitted version is even slightly lower than the sample-splitting version and is lower than the bias in the full-sample version for both the largest (32′000) and the second largest (8′000) sample considered. For the smaller sample sizes (500 and 2′000) we see that the reduction in the overfitting bias is not large enough in comparison to the bias stemming from the estimation of the CATE function. As such, for small sample sizes the additional splitting of the sample does not leave enough

[33] Nonetheless, this insight might still substantiate the need for sample-splitting, although only with two folds instead of three as used here.

observations to learn the non-linear structure of the CATE. Considering the variance of the estimator, we also observe the theoretically expected pattern. The full-sample version of the DR-learner exhibits the smallest average standard deviation throughout all sample sizes, while the standard deviation for the sample-splitting version is roughly twice as high. Nevertheless, the cross-fitting version successfully reduces the standard deviation for all sample sizes and comes close to the full-sample version, effectively regaining the lost efficiency of the estimator due to sample-splitting. Overall, in terms of the average RMSE this bias-variance trade-off results in favourable performance of the sample-splitting version in the largest and of the cross-fitting version in the two largest samples in comparison to the full-sample version. Considering the distribution of the predicted CATEs we see that the heavy tails problem due to extreme propensity scores is the worst for the sample-splitting version, where even in the second largest sample size of 8′000 observations, the normality is rejected for more than 50% of the predicted CATEs from the validation sample (compare the supplementary results in Appendix 1.B.2). This stems from the smaller samples used for estimation of the propensity scores which are more likely to yield extreme values under imbalanced treatment assignment. We also observe that this issue is less pronounced for the full-sample version. The third and the last observation is yet the probably most noticeable pattern across all performance measures, namely the fast convergence of the sample-splitting and cross-fitting version of the DR-learner which is substantially faster in comparison to all other meta-learners as can be seen in Figure 1.4.1. As such the DR-learner performs almost worst of all, both in terms of the average RMSE and average absolute bias for the smallest sample size of 500 observations, but almost best of all for the largest sample size of 32′000 observations. This provides evidence that the DR-learner is able to learn a highly complex CATE function once enough data becomes available and additionally highlights the need for sample-splitting and cross-fitting in order to achieve the theoretically described optimal performance (Kennedy, 2020).

The performance of the R-learner is competitive with the other meta-learners especially in smaller samples, particularly for the full-sample version. In the smallest sample size of 500 observations the R-learner outperforms the DR-learner in terms of the average RMSE, irrespective of the estimation procedure. However, with growing sample sizes the performance evens out and eventually for the largest sample size of 32′000 observations the R-learner lags behind the majority of the estimators. This is in contrast to previous results from simulations of Jacob (2020) and Knaus et al. (2021) where the R-learner exhibits rather good performance, albeit studied only in smaller samples. The decomposition of the RMSE shows that while the full-sample version of the R-learner exhibits rather low bias, it suffers from a higher variance as can be seen in Figure 1.4.1. Nonetheless, the distributions of the predicted CATEs do not show on average deviations from the normal distribution. This is contrary to the DR-learner and illustrates the advantage of the additional weighting step. As such, even though the R-learner uses the propensity scores for reweighting to construct the modified outcome, it successfully manages to downweight the modified outcomes based on extreme propensity scores to alleviate the heavy tails issues observed in the case of the DR-learner. In particular, even for the sample-splitting version of the R-learner the share of predicted CATEs for which the normality is rejected is an order of magnitude lower in comparison to the DR-learner (see Appendix 1.B.2 for details). In terms of the estimation procedure, we observe a similar pattern as for the X-learner in a sense that the full-sample version performs better with respect to the average RMSE and absolute bias, while the cross-fitting version helps to reduce the variance of the estimator not only in comparison to the sample-splitting version but even in comparison to the full-sample version. The sample-splitting version exhibits higher values of the average absolute bias and standard deviation for all sample sizes considered, while the convergence rates are approximately same as for the full-sample and the cross-fitting version. Hence, there is a lack of indication that the overfitting type of bias reduction could become relevant in bigger samples. Similarly to the DR-learner,

also for the R-learner the differences between the different estimation procedures seem to be higher than those for the X-learner which is again presumably due to the different usage of the propensity scores.

Inspecting the results for the rest of the simulation designs reveals further insights and helps to generalize the findings from the main and most challenging simulation design discussed above.

1.4.3.2 Results of Simulation 1: balanced treatment and constant zero CATE

Within the benchmark Simulation 1 with zero constant CATE the S-learner, as expected, performs best with respect to all performance measures across all sample sizes (see Table 1.B.1 in Appendix 1.B.1). However, the results reveal poor statistical properties of the S-learner as it appears to be substantially biased and inconsistent as the absolute bias as well as standard deviation increase with growing sample size.[34] In general, the performance of the S-learner is, in all simulation designs, plagued by the substantially higher bias than all the other meta-learners, partially accompanied by the consistency issues. The SW-learner is affected by the same issues as the S-learner in Simulation 1 but manages to substantially reduce the bias for the rest of the simulation designs and is generally close to the performance of the T-learner as seen in the Main Simulation.

1.4.3.3 Results of Simulation 2: balanced treatment and complex nonlinear CATE

In Simulation 2 with balanced treatment and complex nonlinear CATE we also observe, as expected, a very good RMSE performance of the T-learner throughout all sample sizes (see Table 1.B.2 in Appendix 1.B.1). However, it exhibits quite high variance which is mostly due to the fact that it estimates two completely disjoint response functions for estimating the CATE. Furthermore, in this design the R-learner in its full-sample version performs particularly well, which comes rather as a surprise as it pools the two disjoint response functions within the estimation procedure. Nevertheless, the R-learner achieves even lower bias than the T-learner for large samples, but with rather high variance which is a pattern observed across all simulation designs.

1.4.3.4 Results of Simulation 3: highly unbalanced treatment and constant non-zero CATE

Simulation 3 features a highly unbalanced treatment assignment and a constant CATE for which the X-learner performs best as expected, throughout all sample sizes and irrespective of the estimation procedure (see Table 1.B.3 in Appendix 1.B.1). Indeed, the differences between the particular versions, i.e. full-sample, sample-splitting and cross-fitting, are quite small which is in contrast to the R- and DR-learner confirming the insights from the Main Simulation. Within this highly unbalanced design the estimation of the propensity score function plays a key role as in this case the estimated propensity scores can get quite often very close to zero. This, however, does not affect the performance of the X-learner as it uses the propensity scores in a fundamentally different way and even the most extreme $\{0, 1\}$ scores would be indeed admissible as pointed out by Künzel et al. (2019). On the contrary, the results show that such extreme propensity scores make now both the R-learner and the DR-learner unstable, with the instability being particularly pronounced in the latter meta-learner. In the case of the DR-learner the

[34] A closer look on the estimation results reveals the reason for this phenomenon. With small sample sizes, the underlying trees of the S-learner's forest are quite shallow and barely split on the treatment indicator resulting in quite homogeneous CATE predictions which are very close to the actual zero effect. However, as the sample size increases, the chance of splits based on the treatment indicator increases which results in more heterogeneous effect predictions spread around zero. Accordingly, the bias as well as the standard deviation increase. Similar consistency issues of the forest-based S-learner seem to appear also in the simulations of Künzel et al. (2019) where the MSE rises with growing sample size for some designs and only stabilizes with very big sample sizes.

heavy tail problem of the CATE distribution is aggravated by more unbalanced treatment assignment as can be seen based on the Jarque-Bera statistic and also on the higher variance of the estimator. While the R-learner manages to avoid this issue by downweighting the observations with extreme propensity scores in less unbalanced settings, it is not fully able to do so when the imbalance is very high and there is potentially a large proportion of propensity scores close to 1. This translates into the higher values of the Jarque-Bera statistic as well as to higher variance and higher bias, too. These issues lead ultimately to bad performance in terms of the average RMSE for both the R- and DR-learner.

1.4.3.5 Results of Simulation 4: unbalanced treatment and simple CATE

In Simulation 4 the imbalance in the treatment assignment is less pronounced which should partly reduce the propensity score issues for the R- and DR-learner. Within this simulation design we observe similar patterns as for the Main Simulation. For the small and medium sized samples the X-learner in the full-sample version performs best in terms of the average RMSE, while it gets outperformed by the DR-learner in its cross-fitting version in the largest sample-size. While the R-learner's performance is quite competitive in smaller samples, it lags behind in larger samples as observed in other simulation designs. As a general pattern, the X-learner remains quite stable with respect to the estimation procedure whereas the DR-learner in its sample-splitting and cross-fitting version exhibits substantially faster convergence than the competing estimators. Nonetheless, based on the Jarque-Bera statistic, the heavy tail issue is less pronounced but still present as can be seen in Table 1.B.4 in Appendix 1.B.1.

1.4.3.6 Results of Simulation 5: unbalanced treatment and linear CATE

Lastly, in Simulation 5 the CATE function gets more involved, while the treatment assignment remains unchanged. The results once again resemble the general pattern (for details see Table 1.B.5 in Appendix 1.B.1). As such the R-learner is competitive mainly in the smaller sample sizes, in this case best performing in the cross-fitting version. The DR-learner in the sample-splitting and cross-fitting version exhibits faster convergence rates, however, in this case the considered sample sizes are not large enough to outperform the X-learner. The X-learner exhibits again little differences regarding the estimation procedure and outperforms the other meta-learners in all performance measures across all sample-sizes.

1.4.4 Empirical Simulation

In order to compare the performance of the meta-learners outside a completely synthetic design of the above simulations we apply the estimators in an arguably more realistic setting using an augmented real dataset. For this purpose we use the data from the data challenge of the 2018 Atlantic Causal Inference Conference (2018 ACIC henceforth). This data is particularly suitable for a comparison of the meta-learners for two reasons. First, the data is based on a randomized control trial in education, namely the National Study of Learning Mindsets (NSLM) by Yeager et al. (2019), and thus provides us with a real data example. Second, the dataset has been augmented to an observational setting with measured confounding and known treatment effects (Carvalho, Feller, Murray, Woody, & Yeager, 2019) which enables us to evaluate the performance of the meta-learners for the estimation of CATEs.

The dataset includes a total of 10′391 observations with 10 covariates, a simulated continuous outcome and a binary treatment, while the share of treated is approximately 25%.[35] The variables are

[35]The dataset can be retrieved online from GitHub. We neglect here the information about the additional school ID for simplicity and comparability reasons.

described in Table 1.A.1 in Appendix 1.A.2. Additionally, to create a more challenging large-dimensional setting, similar to the synthetic simulations, we augment the dataset further with $p = 90$ uniformly distributed covariates, i.e. $X_{11,\dots,100} \sim \mathcal{U}\big([0,1]^{n\times p}\big)$ with the same correlation structure as used within the synthetic simulations.[36] At a high level, we are interested in estimating the treatment effects of an intervention to foster a belief to develop intelligence in students on a measure of student achievement, conditional on observed covariates. The CATEs were generated according to the following specification:

$$\tau(x) = 0.228 + 0.05 \cdot \mathbb{1}(x_1 < 0.07) - 0.05 \cdot \mathbb{1}(x_2 < -0.69) - 0.08 \cdot \mathbb{1}(c_1 \in 1, 13, 14)$$

while the conditional independence assumption holds by construction, the confounding has a complicated functional form. For a detailed description of the data generating process used for the augmentation see Carvalho et al. (2019).

Similarly as in the synthetic simulations we estimate the heterogeneous treatment effects with all meta-learners and evaluate their performance with regard to the point estimates. For this purpose we perform an empirical simulation study inspired, among others, by Lechner (2018) and Künzel et al. (2019) where we first, set apart a validation set of size $N = 1'000$ observations, and second, sample $R = \{2'000, 1'000, 500\}$ training sets each of sizes $N = \{500, 2'000, 8'000\}$ observations from the remaining data. We omit the biggest sample of $N = 32'000$ observations due to the size restrictions of the dataset. We report mean performance measures in a similar fashion as in the previous simulation experiments.

1.4.4.1 Results of Empirical Simulation

The CATE results of the Empirical Simulation for all meta-learners are summarized in Table 1.4.3, while Figure 1.4.2 provides details on the meta-learners in the full-sample, double sample-splitting and double cross-fitting versions.

The results reveal a similar picture to the synthetic simulations in general, with the largest similarities to Simulation 3 and 5 in particular. Accordingly, the X-learner achieves the best performance in terms of the average RMSE as well as average absolute bias in all considered sample-sizes, regardless of the estimation procedure. This emphasizes the good performance of the X-learner in settings with unbalanced treatment assignment and sparse CATE function with structural properties. In the largest sample size of $8'000$ observations, also the DR- and R-learner come close to the performance of the X-learner in terms of the average RMSE, while the simpler SW- and T-learner are competitive mainly in the smaller sample sizes. We also observe a slightly faster convergence of the sample-splitting and cross-fitting version of the DR-learner as in the synthetic simulations, however, the limited sample size in this case does not allow for a sufficiently large improvement to outperform the X-learner. Given the smaller sample sizes in the empirical simulation, we are not able to detect the bias-variance trade-off and the sample-splitting versions always exhibit higher values of the average RMSE, average absolute bias and average standard deviation. This is particularly noticeable for the smallest sample size of 500 observations as there is essentially not enough data left after splitting to learn the correct CATE function. For all meta-learners the cross-fitting versions then always perform better in terms of the variance reduction and even lead to a lower bias in comparison to the sample-splitting versions. These results accentuate the fact that the benefits of sample-splitting in removing the overfitting bias become apparent only for sufficiently large samples. Additionally, we see larger discrepancies between the estimation versions of the DR- and R-learner in comparison to very stable performance of the X-learner, similarly as in the synthetic simulations. Lastly, the results on the distribution of the predicted CATEs resemble those of

[36] For more detailed descriptive statistics of the augmented empirical dataset including correlation heat map of the covariates see Appendix 1.A.2.

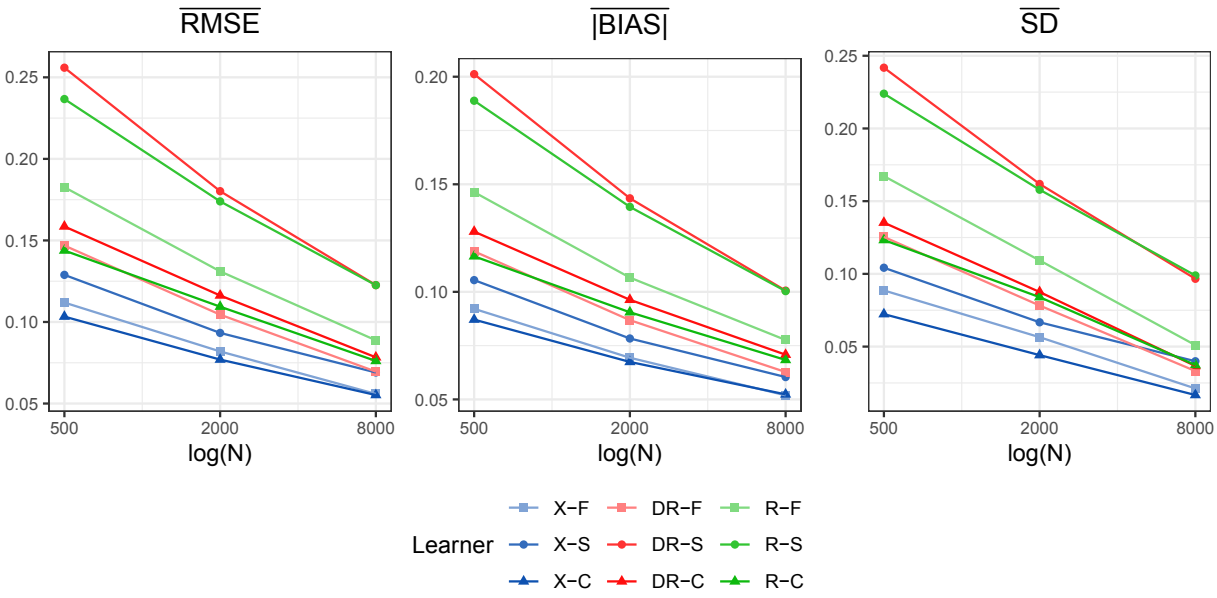
the synthetic simulations with the heavy tail problem of the DR-learner in its sample-splitting version as well as deviations from normality of the S- and SW-learner.

Table 1.4.3: CATE Results for Empirical Simulation

	$\overline{RMSE}$			$\overline{\lvert BIAS \rvert}$			$\overline{SD}$			$\overline{JB}$		
	500	*2000*	*8000*	*500*	*2000*	*8000*	*500*	*2000*	*8000*	*500*	*2000*	*8000*
S	0.175	0.127	0.093	0.171	0.121	0.090	**0.035**	**0.035**	0.023	165.803	7.372	2.054
S-W	0.131	0.109	0.078	0.106	0.090	0.070	0.121	0.084	0.037	57.438	2.065	**1.903**
T	0.150	0.111	0.079	0.122	0.092	0.071	0.127	0.084	0.037	2.050	2.047	2.082
X-F	0.112	0.082	0.056	0.092	0.069	**0.052**	0.089	0.056	0.021	2.043	1.941	2.078
X-S	0.129	0.093	0.069	0.105	0.078	0.060	0.104	0.067	0.040	2.129	2.594	2.004
X-C	**0.103**	**0.077**	**0.055**	**0.087**	**0.067**	**0.052**	0.072	0.044	**0.017**	**1.652**	**1.794**	1.935
DR-F	0.147	0.105	0.070	0.119	0.087	0.063	0.125	0.078	0.033	7.134	3.377	2.001
DR-S	0.256	0.180	0.123	0.201	0.143	0.101	0.242	0.162	0.097	68.837	46.173	6.196
DR-C	0.159	0.116	0.078	0.128	0.096	0.071	0.135	0.088	0.037	6.334	5.329	2.969
R-F	0.183	0.131	0.089	0.146	0.107	0.078	0.167	0.109	0.051	3.819	3.862	2.027
R-S	0.237	0.174	0.123	0.189	0.140	0.100	0.224	0.158	0.099	3.490	3.494	3.117
R-C	0.144	0.109	0.076	0.117	0.091	0.068	0.123	0.084	0.037	2.060	2.222	2.225

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{SD}$ and $\overline{JB}$ show the mean values of the root mean squared error, absolute bias, standard deviation and the Jarque-Bera test statistic of all 1′000 CATE estimates from the validation sample. The critical values for the JB test statistic are 5.991 and 9.210 at the 5% and 1% level, respectively. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively. Bold numbers indicate the best performing meta-learner for given measure and sample size.

Figure 1.4.2: CATE Results for Empirical Simulation



Note: The results for $\overline{RMSE}$, $\overline{|BIAS|}$, and $\overline{SD}$ show the mean values of the root mean squared error, absolute bias, and standard deviation of all 1′000 CATE estimates from the validation sample. The figure shows the results based on the increasing training samples of $\{500, 2'000, 8'000\}$ observations displayed on the log scale. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

1.5 Discussion

Given the results of our synthetic and empirical simulations there are several important findings for the estimation of heterogeneous causal effects by the meta-learners and the associated usage of sample-splitting and cross-fitting which are relevant for applied empirical work.

1.5.1 Meta-Learners

In general, the results suggest that meta-learners that directly model both the outcome equations and the selection process perform better, especially in larger samples, which is in line with the insights from the previous literature (see e.g. Knaus et al., 2021). Meta-learners modelling only the outcome equations are competitive only in smaller samples and tend to perform poorly in larger samples as they fail to properly account for the selection into treatment.

In particular, we do not recommend the usage of the S-learner for estimation of heterogeneous causal effects due to empirically documented undesirable statistical properties such as high bias and consistency issues. The herein studied modification of the S-learner, the SW-learner, alleviates the high bias of the S-learner, however, it does not solve the consistency issues. Hence, enforcing the treatment variable into the splitting set of the forest does not constitute an attractive option for estimation of causal effects. In contrast, the T-learner does not suffer from high bias or any consistency issues and has a stable performance as it uses the full data sample without the need of sample-splitting due to potential overfitting. Hence, the T-learner might be an interesting option, if a large sample is not available for the empirical analysis. Related simulation studies (Jacob, 2020; Knaus et al., 2021) find also relatively competitive performance of the T-learner, especially with the Random Forest as a base learner.

Among the meta-learners based on the estimation of nuisance functions, the X-learner performs very well not only in settings with highly unbalanced but also in less unbalanced treatment shares with simple CATEs and demonstrates the theoretically argued capability to learn such CATE structures (Künzel et al., 2019). Moreover, the X-learner exhibits a quite stable performance across all simulation designs with low bias and very low variance, even in small samples. Additionally, due to its particular usage of the propensity scores, the X-learner is not too sensitive to the choice of the estimation procedure. As such, both the full-sample version and the cross-fitting version of the X-learner are viable options, regardless of the sample size. For these reasons, we recommend to use the X-learner for CATE estimation if the researcher is facing a situation with very low number of treated units as well as in less unbalanced settings with potentially limited sample size. In contrast to the X-learner, the DR-learner performs particularly well in settings with nonlinear and complex CATEs if large enough samples are available. However, it tends to be unstable in small samples with unbalanced treatment assignment due to extreme propensity scores, which relates to the results of Jacob (2020) and Knaus et al. (2021). Additionally, for the DR-learner the choice of the estimation procedure is crucial as its sample-splitting and cross-fitting version exhibits the fastest convergence rates of all meta-learners which highlights the theoretical arguments provided in Kennedy (2020). According to the simulation evidence, we advice to employ the cross-fitting version of the DR-learner for the CATE estimation in settings with rather balanced treatment assignment and when large sample is available. Recently, Knaus (2020) proposed the *normalized* DR-learner, that addresses the problem of unstable CATE predictions due to extreme propensity scores which might be a viable option for smaller sample sizes and settings with unbalanced treatment shares. Lastly, the simulation evidence suggests that the R-learner is in comparison to the DR-learner less prone to unstable performance due to extreme propensity scores. However, its performance is competitive only in smaller samples, while the empirically approximated speed of convergence is slower than the one of the DR-learner and seems to depend on the CATE complexity as theoretically argued by Nie and Wager (2021). With respect to the estimation procedure we do not find a clear-cut evidence in favour of a particular version as both the full-sample as well as the cross-fitting version exhibit comparably good performance. Based on this evidence, the R-learner might be an attractive option for estimation of CATEs if the treatment is not too imbalanced and if only a small sample is available. For comparable sample sizes, Knaus et al. (2021) also find the R-learner to have good performance in a variety of settings.

Overall, we point out that based on the simulation evidence, for all meta-learners the approximate convergence rates appear to be substantially slower than the parametric rate of $\sqrt{N}$. This is expected given the insights from previous literature that the estimation of more granular heterogeneous effects is a more difficult task in comparison to the estimation of average effects (compare e.g. Lechner, 2018; or Knaus et al., 2021). However, we note that the approximate convergence rates differ considerably among the meta-learners and their specific implementations as documented in our simulation experiments.

1.5.2 Estimation Procedures

Our simulation evidence suggests that using the full sample for estimation of both the nuisance functions as well as the CATE function leads to a remarkably good performance in terms of both bias and variance in finite samples. Recently, Curth and van der Schaar (2021) also point out that the full-sample estimation seems to work better in practice, especially for small samples. In theory, we would expect lower variance yet higher bias due to overfitting (Chernozhukov et al., 2018). The possible reason for this phenomenon might in our case be due to the out-of-bag predictions of the forest that we use throughout the simulation experiments. Even though these predictions are not out-of-sample *per se* they are not directly based on the observations used for estimation and as such might help to alleviate the overfitting problem when using full sample (compare Athey & Imbens, 2019, for a discussion of out-of-bag predictions in Random Forests). In the causal machine learning literature, such out-of-bag predictions are for example also used in the case of the Generalized Random Forest for the residualization (Athey et al., 2019), similar to the one used in the R-learner. In contrast to the full-sample estimation, using the double sample-splitting for the estimation of the nuisance functions, we effectively use only one third of the available data. Theoretically, we should observe a smaller bias but higher variance of the estimators. However, in almost all cases we observe both higher bias as well as higher variance, particularly for the small sample sizes. Nonetheless, we document the expected bias-variance trade-off for the largest sample sizes. This stems mainly from the fact that using only a third of the smaller samples does not allow a sensible machine learning estimation of the highly non-linear nuisance functions featured in our simulations. However, especially for the DR-learner we do observe faster convergence rates for the sample-splitting version which is compatible with the theoretical convergence arguments (Newey & Robins, 2018; Kennedy, 2020). Hence, it seems to be the case that in order to benefit from the double sample-splitting the training sample must be of sufficient size, otherwise the full sample estimation achieves a better performance. Lastly, the double cross-fitting for estimation of the nuisance components effectively uses all the available information from the data and substantially reduces the variance of the estimators, while keeping the bias low at the same time. This comes at the price of longer computation time in comparison to the sample-splitting procedure as the estimation is repeated several times. Nevertheless, the computation time of the cross-fitting procedure is on average comparable with the full-sample estimation (see Appendix 1.C for details).

Based on the above simulation evidence, it seems reasonable to always use the full-sample estimation together with out-of-bag predictions (if available) when a relatively small sample is available to the applied researcher, whereas to use the double cross-fitting procedure when a relatively large data is accessible, regardless of the choice of a meta-learner. On the contrary, the simulations do not provide any evidence for an advantageous usage of the double sample-splitting over the double cross-fitting, apart from the computational aspects.

1.6 Conclusion

We investigate the finite sample performance of the meta-learners for the estimation of heterogeneous causal effects with focus on the specific estimation implementations related to data usage. In particular, we examine the properties of double sample-splitting and double cross-fitting as defined by Newey and Robins (2018) in contrast to using full sample for estimation. For this purpose, we review several meta-learning algorithms for estimation of causal effects and discuss their advantages and disadvantages in particular empirical settings. We conduct an extensive simulation study with data generating processes involving highly non-linear functional forms and large-dimensional feature space to challenge the machine learning algorithms, while keeping the treatment effect specifications well-structured. Furthermore, we perform an empirical simulation based on an augmented real dataset to reflect an actual empirical setting. Moreover, we repeat the simulation experiments for increasing sample sizes to empirically study the convergence properties of the meta-learners. Based on our simulation evidence, we provide a guideline for empirical researchers and practitioners to better inform the decisions of applying certain method and estimation procedure for their particular research objectives.

The results of our simulation study show that the choice of the estimation procedure can indeed largely impact the performance of the meta-learners in finite samples. On the one hand, we provide an empirical evidence for the theoretical arguments of the bias-variance trade-off related to sample-splitting and cross-fitting which, however, become apparent only if sufficiently large samples are used. On the other hand, we document the adverse effects of these procedures in small samples, when using machine learning. Therefore, we argue that in empirical studies based on small samples, applied researchers should use the full sample for machine learning estimation of both the nuisance functions as well as the treatment effect function as the overfitting bias is in such cases of secondary importance. However, for empirical analyses with access to large data samples, we advocate for the usage of the double cross-fitting for the estimation of treatment effects as the overfitting bias here becomes of primary importance. The double cross-fitting procedure then effectively reduces this overfitting bias and successfully preserves the full sample size efficiency of the estimator. Moreover, if computation time is not a constraint, we discourage applied researchers to use the double sample-splitting procedure due to substantial increase in variance, while having no benefit over the double cross-fitting in terms of bias reduction.

In contrast to the typical drawbacks of simulation studies, the particular design of our simulation experiments with varying sample size and varying treatment shares allows us to draw relevant conclusions that are not solely dependent on the particular specification of the data generating processes, but rely on the data characteristics that an applied researcher can observe without imposing arbitrary assumptions. In particular, the simulation evidence implies a clear advantage for the X-learner, when a researcher is confronted with highly unbalanced treatment shares. This finding holds irrespective of the sample size at hand and as such we recommend the usage of X-learner for estimation of heterogeneous treatment effects whenever the share of treated or controls is around 15% or less. With less unbalanced treatment shares at around 25% of treated or controls, the size of the available sample becomes decisive. For smaller samples with only few hundred observations (500 and 2′000), the simulation evidence again favours the usage of the X-learner. However, for bigger samples with several thousand observations (8′000 and 32′000), our findings favour the DR-learner as it can successfully learn highly complex treatment effect function if enough data is available. Finally, with perfectly balanced treatment shares, the sample size matters less. In such cases, the DR-learner as well as the R-learner are both the preferred estimators. However, we advise against the usage of these two methods in highly imbalanced settings as their performance becomes unstable due to extreme propensity scores. Finally, concerning the simpler meta-learners, we explicitly argue against the usage of the S-learner by applied researchers for estimation of heterogeneous

treatment effects due to the herein empirically documented undesirable statistical properties, while the T-learner might be a reasonable choice in small samples with balanced treatment shares.

Even though we shed light on certain finite sample issues of applying different estimation procedures when using meta-learners for estimation of heterogeneous causal effects, our findings raise new relevant questions. Most importantly, the question of conducting statistical inference about the estimated heterogeneous treatment effects is worth further investigations. Based on the insights in this paper it would be of interest to investigate the performance of the bootstrapping for estimation of standard errors as studied by Künzel et al. (2019) for meta-learners based on the double sample-splitting and double cross-fitting procedures. Moreover, a comparison of such bootstrapping inference procedure for meta-learners and the approaches used in the Causal Forest literature such as the bootstrap of little bags in the Generalized Random Forest (Athey et al., 2019) or the weight-based inference as in the Modified Causal Forest (Lechner, 2018) would be desirable. Furthermore, the performance difference in the point estimation using the out-of-bag vs. in-sample predictions could provide additional insights on the benefits of sample-splitting and cross-fitting procedures and hence to assess the robustness of our results to different types of base learners. Finally, a further simulation comparison between the X-learner, the DR-learner and its normalized version as proposed by Knaus (2020) for highly imbalanced settings would be of interest.

Acknowledgements

A previous version of this paper was presented at research seminars of the University of St.Gallen, EPFL and the GESIS Spring Seminar in Cologne. We thank participants, in particular Michael Lechner, for helpful comments and suggestions. We also thank Francesco Audrino, Daniele Ballinari, Jonathan Chassot, Daniel Goller, Sandro Heiniger, Daniel Jacob, Michael Knaus, Jana Mareckova, Matthias Roesti, Kenneth Younge and Michael Zimmert for their useful feedback. The usual disclaimer applies.

Bibliography

Andini, M., Ciani, E., de Blasio, G., D'Ignazio, A., & Salvestrini, V. (2018). Targeting with machine learning: An application to a tax rebate program in Italy. *Journal of Economic Behavior and Organization*, *156*, 86–102.

Athey, S. (2018). The Impact of Machine Learning on Economics. In *The economics of artificial intelligence: An agenda* (pp. 507–547). University of Chicago Press.

Athey, S. & Imbens, G. W. (2016). Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S. & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Athey, S. & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, *11*(1), 685–725.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148–1178.

Athey, S. & Wager, S. (2021). Policy Learning With Observational Data. *Econometrica*, *89*(1), 133–161.

Bai, Y., Chen, M., Zhou, P., Zhao, T., Lee, J. D., Kakade, S., ... Xiong, C. (2020). How Important is the Train-Validation Split in Meta-Learning? *arXiv preprint arXiv:2010.05843*.

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., & Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science*, *359*(6373), 325–329.

Bargagli Stoffi, F. J. & Gnecco, G. (2020). Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms. *International Journal of Data Science and Analytics*, *9*(3), 315–337.

Belloni, A., Chernozhukov, V., Fernandez-Val, I., & Hansen, C. (2017). Program Evaluation and Causal Inference With High-Dimensional Data. *Econometrica*, *85*(1), 233–298.

Belloni, A., Chernozhukov, V., & Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, *81*(2), 608–650.

Bera, A. K. & Jarque, C. M. (1981). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Monte Carlo Evidence. *Economics Letters*, *7*(4), 313–318.

Biau, G. (2012). Analysis of a Random Forests Model. *Journal of Machine Learning Research*, *13*(1), 1063–1095.

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

Bickel, P. J. & Ritov, Y. (1988). Estimating Integrated Squared Density Derivatives: Sharp Best Order of Convergence Estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, *50*(3), 381–393.

Bickel, P. J. (1982). On Adaptive Estimation. *The Annals of Statistics*, 647–671.

Biewen, M. & Kugler, P. (2021). Two-stage least squares random forests with an application to Angrist and Evans (1998). *Economics Letters*, *204*, 109893.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. CRC Press.

Broockman, D. & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, *352*(6282), 220–224.

Caruana, R. (1997). Multitask Learning. *Machine Learning*, *28*(1), 41–75.

Carvalho, C., Feller, A., Murray, J., Woody, S., & Yeager, D. (2019). Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge. *Observational Studies*, *5*, 21–35.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, *21*(1), 1–68.

Chipman, H. A., George, E. I., & McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, *93*(443), 935–948.

Cockx, B., Lechner, M., & Bollens, J. (2019). Priority to unemployed immigrants?A causal machine learning evaluation of training in Belgium. *arXiv preprint arXiv: 1912.12864*.

Curth, A., Alaa, A. M., & van der Schaar, M. (2020). Semiparametric Estimation and Inference on Structural Target Functions using Machine Learning and Influence Functions. *arXiv preprint arXiv: 2008.06461*.

Curth, A. & van der Schaar, M. (2021). Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms. In *International conference on artificial intelligence and statistics* (pp. 1810–1818). PMLR.

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, *221*(2), 644–654.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, *1857*, 1–15.

Falk, M. (1999). A simple approach to the generation of uniformly distributed random variables with prescribed correlations. *Communications in Statistics Part B: Simulation and Computation*, *28*(3), 785–791.

Fan, Q., Hsu, Y. C., Lieli, R. P., & Zhang, Y. (2020). Estimation of Conditional Average Treatment Effects With High-Dimensional Data. *Journal of Business and Economic Statistics*, 1–15.

Fan, Y., Lv, J., & Wang, J. (2018). DNN: A Two-Scale Distributional Tale of Heterogeneous Treatment Effect Inference. *SSRN Electronic Journal*.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, *19*(1), 1–67.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, *102*(1), 33–48.

Goller, D., Harrer, T., Lechner, M., & Wolff, J. (2021). Active labour market policies for the long-term unemployed: New evidence from causal machine learning. *arXiv preprint arXiv:2106.10141*.

Goller, D., Lechner, M., Moczall, A., & Wolff, J. (2020). Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed. *Labour Economics*, *65*, 101855.

Gulyas, A. & Pytka, K. (2020). Understanding the Sources of Earnings Losses After Job Displacement : A Machine-Learning Approach. *Discussion PaperSeries–CRC TR 224*, (131), 1–70.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, *66*(2), 315–331.

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, *15*(3), 965–1056.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer Science & Business Media.

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

Hodler, R., Lechner, M., & Raschky, P. (2020). Reassessing the Resource Curse using Causal Machine Learning. *CEPR Discussion Paper No. DP15272*.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Hornung, R. (2019). Ordinal Forests. *Journal of Classification*, 1–14.

Huber, M., Lechner, M., & Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, *175*(1), 1–21.

Hurwicz, L. (1950). Generalization of the Concept of Identification. *Statistical inference in dynamic economic models*, *10*, 245–257.

Imai, K. & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, *7*(1), 443–470.

Imbens, G. W. & Rubin, D. B. (2015). *Causal inference: For Statistics, Social, and Biomedical Sciences: An Introduction*.

Imbens, G. W. & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Jacob, D. (2020). Cross-Fitting and Averaging for Machine Learning Estimation of Heterogeneous Treatment Effects. *arXiv preprint arXiv:2007.02852*.

Jacob, D. (2021). CATE meets ML - The Conditional Average Treatment Effect and Machine Learning. *Digital Finance*.

Jacob, D., Härdle, W. K., & Lessmann, S. (2019). Group Average Treatment Effects for Observational Studies. *arXiv preprint arXiv:1911.02688*.

Janitza, S., Tutz, G., & Boulesteix, A. L. (2016). Random forest for ordinal responses: Prediction and variable selection. *Computational Statistics and Data Analysis*, *96*, 57–73.

Jarque, C. M. & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*(3), 255–259.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, *97*(10), 2177–2189.

Johansson, F. D., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference. In *33rd international conference on machine learning, icml 2016* (Vol. 6, pp. 4407–4418).

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.

Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *79*(4), 1229–1245.

Kitagawa, T. & Tetenov, A. (2018). Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, *86*(2), 591–616.

Knaus, M. C. (2020). Double Machine Learning based Program Evaluation under Unconfoundedness. *arXiv preprint arXiv:2003.03191*.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2020). Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach. *Journal of Human Resources*, 0718–9615R1.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, *24*(1), 134–161.

Künzel, S. R. (2019). *Heterogeneous Treatment Effect Estimation Using Machine Learning* (Doctoral dissertation, UC Berkeley).

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165.

Künzel, S., Liu, E., Saarinen, T., Tang, A., & Sekhon, J. (2020). forestry: Forestry. GitHub.

Lechner, M. (2018). Modified Causal Forests for Estimating Heterogeneous Causal Effects. *arXiv preprint arXiv: 1812.09487v2*.

Lechner, M. & Okasa, G. (2019). Random Forest Estimation of the Ordered Choice Model. *arXiv preprint arXiv:1907.02436*.

Lechner, M. & Wunsch, C. (2013). Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics*, *21*, 111–121.

Lee, S., Okui, R., & Whang, Y. J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, *32*(7), 1207–1225.

Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Meinshausen, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research*, *7*(Jun), 983–999.

Mentch, L. & Hooker, G. (2016). Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. *Journal of Machine Learning Research*, *17*(1), 841–881.

Newey, W. K. & Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Nie, X. & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319.

Nie, X., Brunskill, E., & Wager, S. (2021). Learning When-to-Treat Policies. *Journal of the American Statistical Association*, *116*(533), 392–409.

Powell, J. L., Stock, J. H., & Stoker, T. M. (1989). Semiparametric Estimation of Index Coefficients. *Econometrica*, *57*(6), 1403–1430.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., & Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, *37*(11), 1767–1787.

Qian, M. & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics*, *39*(2), 1180–1210.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*(9-12), 1393–1512.

Robins, J. M. (2004). Optimal Structural Nested Models for Optimal Sequential Decisions. (pp. 189–326).

Robins, J. M. & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, *56*(4), 931.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Saunshi, N., Gupta, A., & Hu, W. (2021). A Representation Learning Perspective on the Importance of Train-Validation Splitting in Meta-Learning. *arXiv preprint arXiv:2106.15615*.

Schick, A. (1986). On Asymptotically Efficient Estimation in Semiparametric Models. *The Annals of Statistics*, *14*(3), 1139–1151.

Schmidhuber, J. (1987). *Evolutionary principles in self-referential learning, or on learning how to learn* (Doctoral dissertation, Technische Universität München).

Schuler, A., Baiocchi, M., Tibshirani, R., & Shah, N. (2018). A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*.

Schwab, P., Linhardt, L., & Karlen, W. (2018). Perfect Match: A Simple Method for Learning Representations for Counterfactual Inference with Neural Networks. *arXiv preprint arXiv: 1810.00656*.

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

Semenova, V. & Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, *24*(2), 264–289.

Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In *34th international conference on machine learning, icml 2017* (Vol. 6, pp. 4709–4718).

Shi, C., Blei, D. M., & Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. In *Advances in neural information processing systems* (Vol. 32).

Snowden, J. M., Rose, S., & Mortimer, K. M. (2011). Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, *173*(7), 731–738.

Stadie, B. C., Kunzel, S. R., Vemuri, N., & Sekhon, J. S. (2018). Estimating heterogeneous treatment effects using neural networks with the Y-Learner.

Taddy, M., Gardner, M., Chen, L., & Draper, D. (2016). A Nonparametric Bayesian Analysis of Heterogenous Treatment Effects in Digital Experimentation. *Journal of Business and Economic Statistics*, *34*(4), 661–672.

Thadewald, T. & Büning, H. (2007). Jarque-Bera test and its competitors for testing normality - A power comparison. *Journal of Applied Statistics*, *34*(1), 87–105.

Thrun, S. & Pratt, L. (1998). *Learning to Learn*. Springer Science \& Business Media.

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532.

Tibshirani, J., Athey, S., Wager, S., Friedberg, R., Miner, L., & Wright, M. (2018). grf: Generalized Random Forests. R package version 0.10.2.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Vanschoren, J. (2019). Meta-Learning. In *Automated machine learning* (pp. 35–61). Springer, Cham.

Vilalta, R. & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*, *18*(2), 77–95.

Wager, S. (2014). Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*.

Wager, S. & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wager, S., Hastie, T., & Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, *15*(1), 1625–1651.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., & Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, *37*(23), 3309–3324.

Wright, M. N. & Ziegler, A. (2017). ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., ... Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369.

Zhao, Y., Zeng, D., Rush, A. J., & Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, *107*(499), 1106–1118.

Zimmert, F. & Zimmert, M. (2020). Paid parental leave and maternal reemployment: Do part-time subsidies help or harm? *Economics Working Paper Series*, (No. 2002).

Zimmert, M. & Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high- dimensional confounding. *arXiv preprint arXiv:1908.08779*.

Zivich, P. N. & Breskin, A. (2021). Machine learning for causal inference: On the use of cross-fit estimators. *Epidemiology*, 393–401.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society*, *67*(2), 301–320.

Appendix

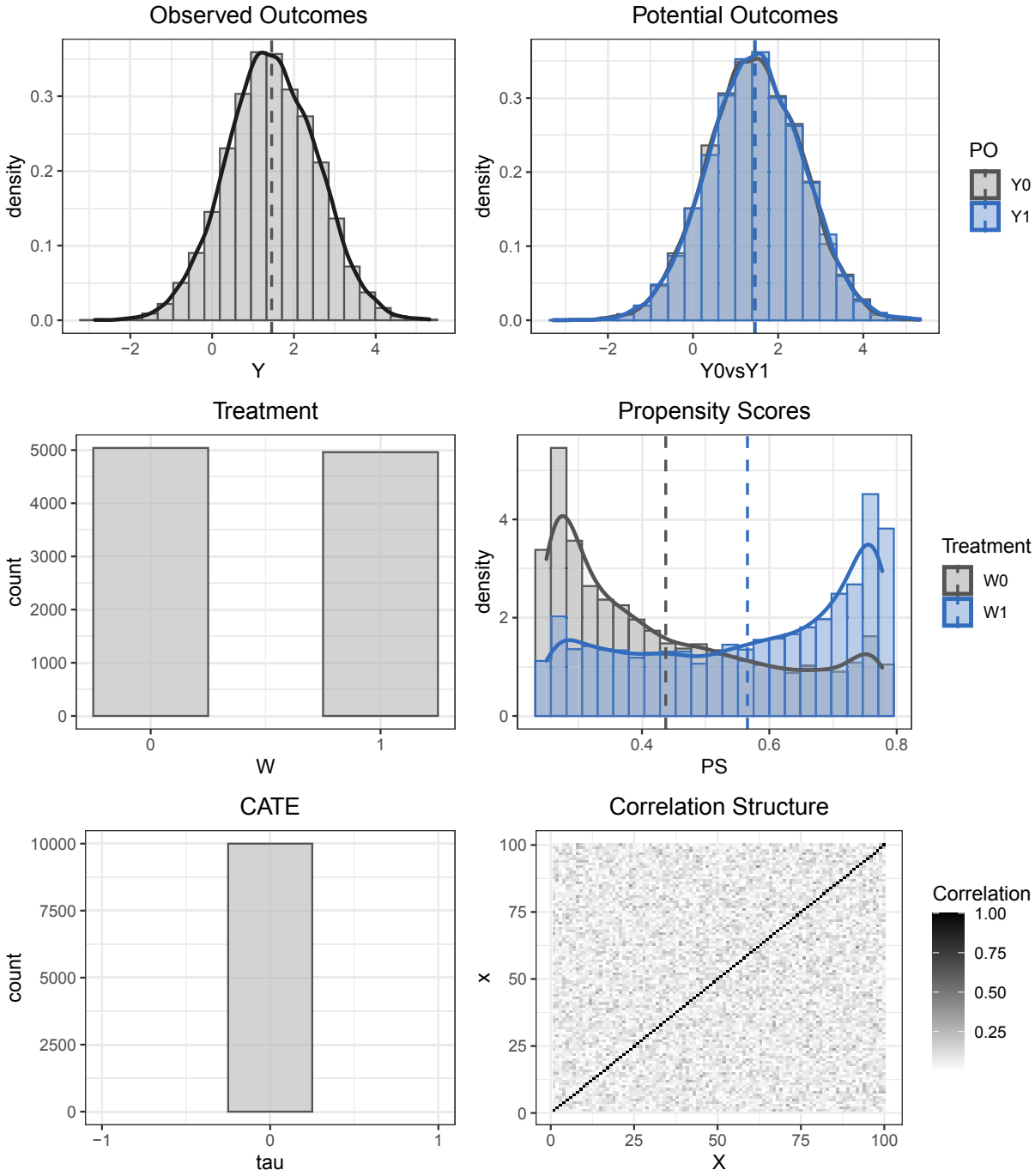
1.A Descriptive Statistics

1.A.1 Synthetic Simulations

This appendix provides the descriptive statistics for the data generated in the six main simulation designs discussed in the main text. For each simulation design we plot the distribution of the observed realized outcomes, Y_i, as well as the potential outcomes, $Y_i(0)$ and $Y_i(1)$. Furthermore, we provide the distribution of the treatment indicator, W_i, together with the propensity score distribution under treatment and under control to visualize the overlap condition. Lastly, we plot the distribution of the true treatment effects, $\tau(X_i)$. Moreover, the plots include a correlation heat map for the covariates X_i. The respective figures for each simulation design are listed below.

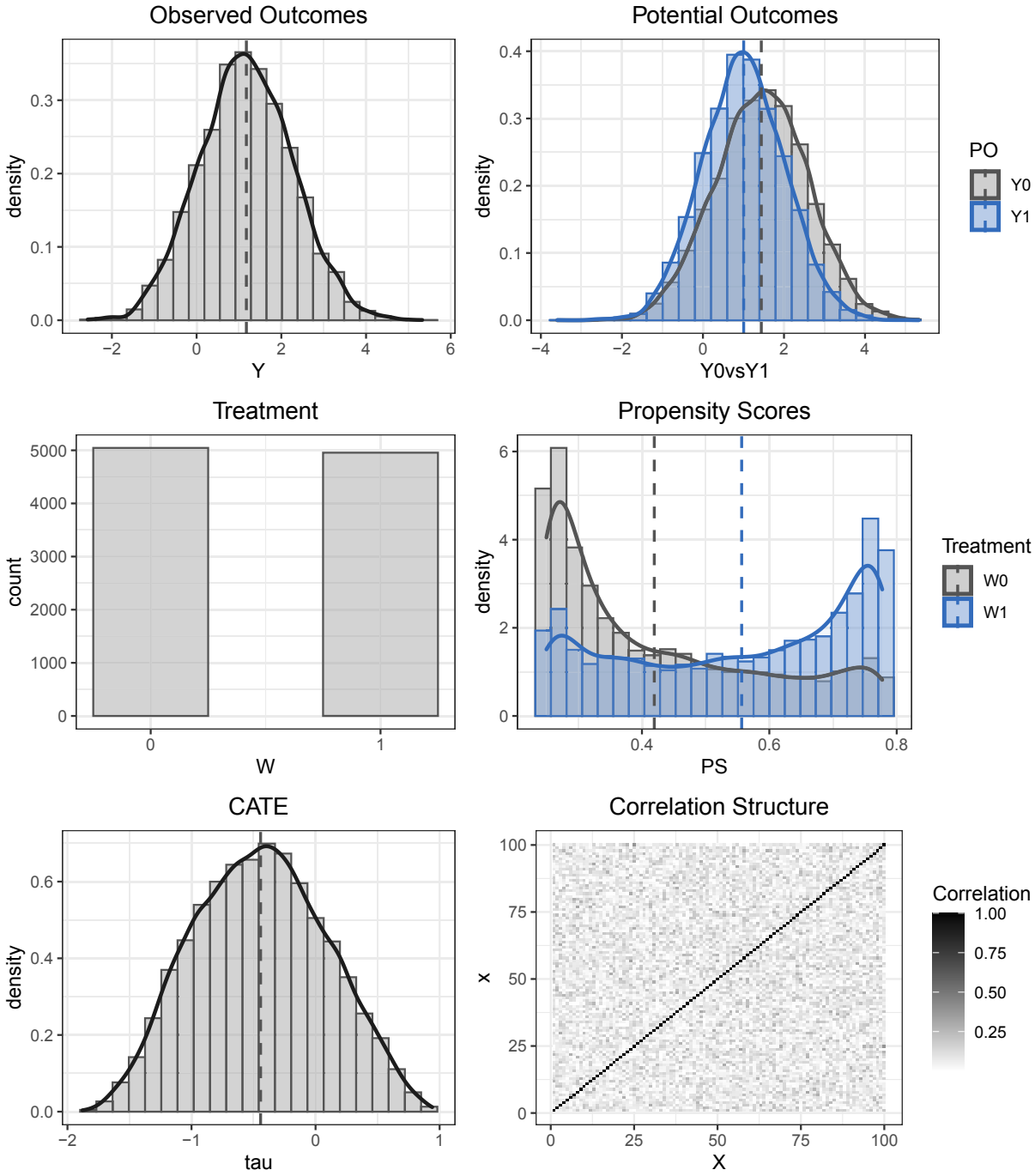
1.A.1.1 Simulation 1: balanced treatment and constant zero CATE

Figure 1.A.1: Descriptive Statistics for the Validation Data in Simulation 1



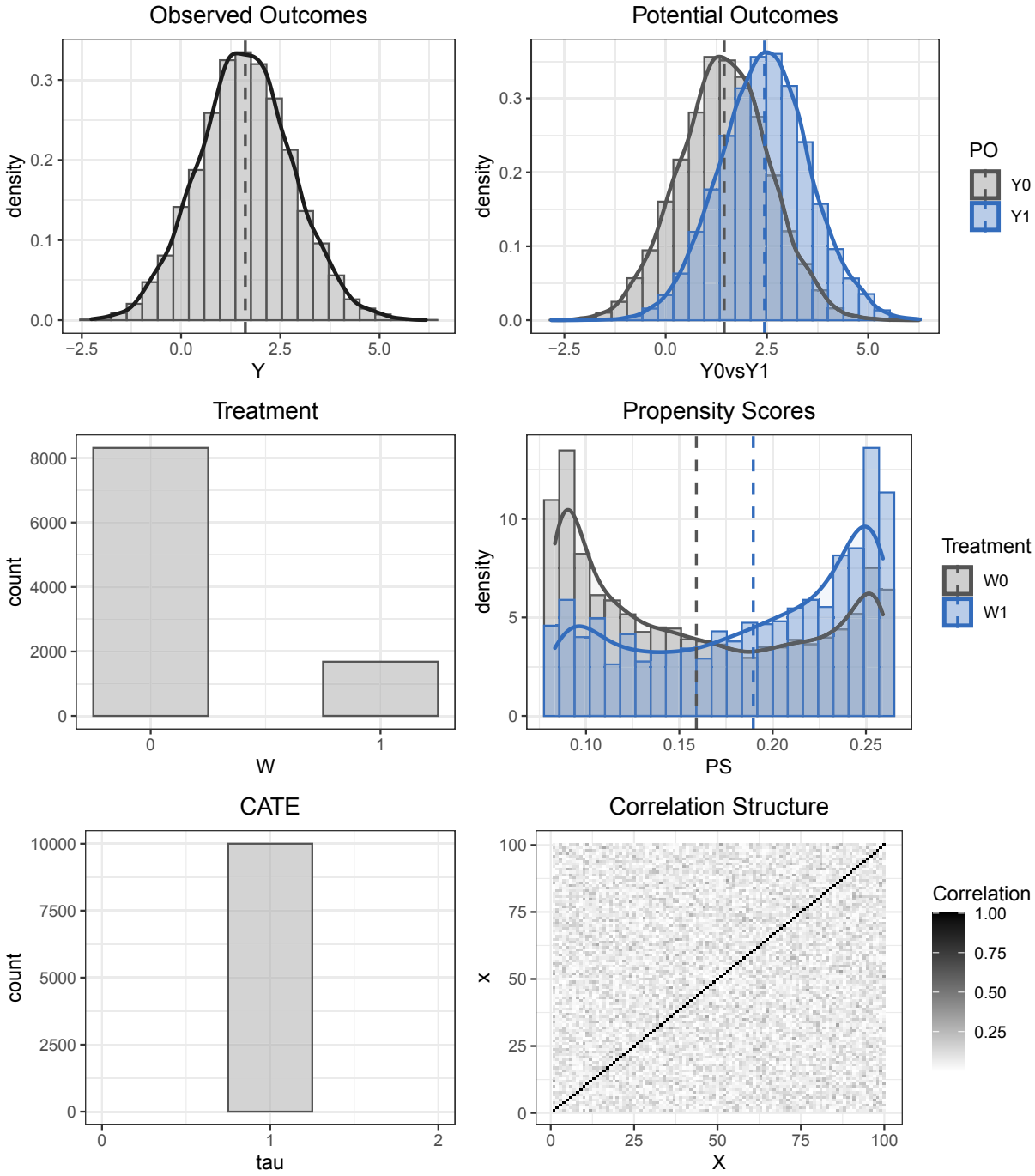
1.A.1.2 Simulation 2: balanced treatment and complex nonlinear CATE

Figure 1.A.2: Descriptive Statistics for the Validation Data in Simulation 2



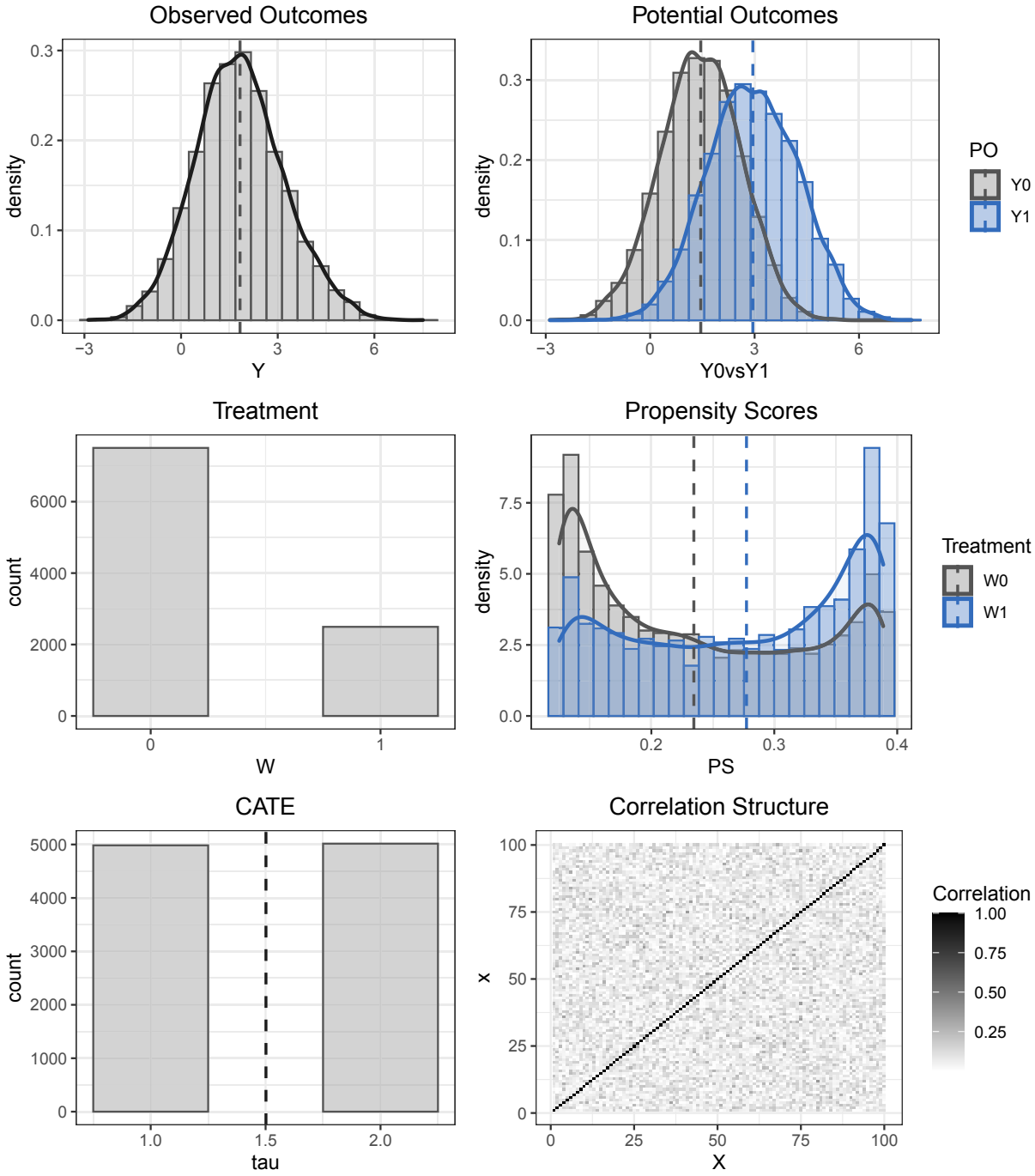
1.A.1.3 Simulation 3: highly unbalanced treatment and constant non-zero CATE

Figure 1.A.3: Descriptive Statistics for the Validation Data in Simulation 3



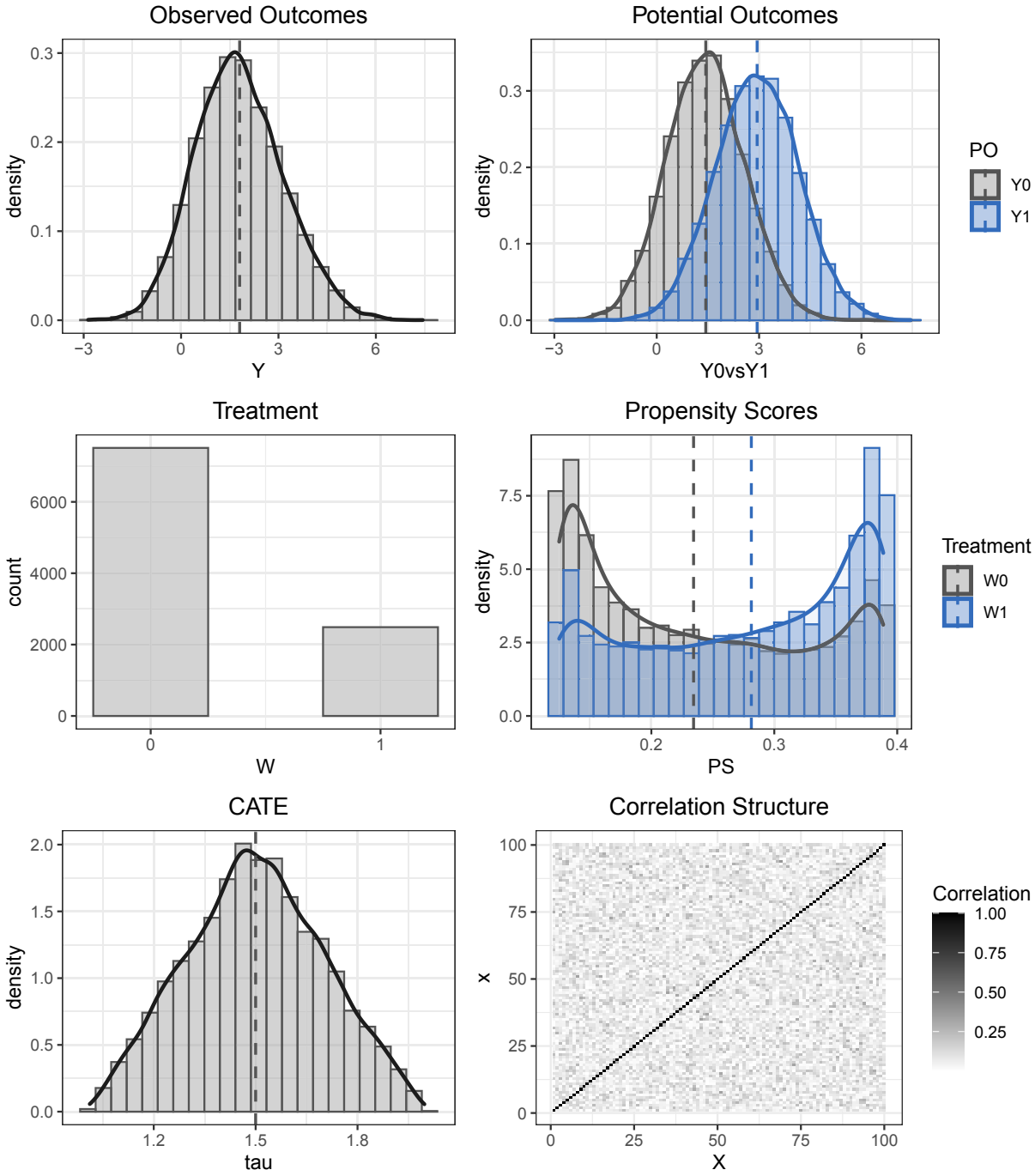
1.A.1.4 Simulation 4: unbalanced treatment and simple CATE

Figure 1.A.4: Descriptive Statistics for the Validation Data in Simulation 4



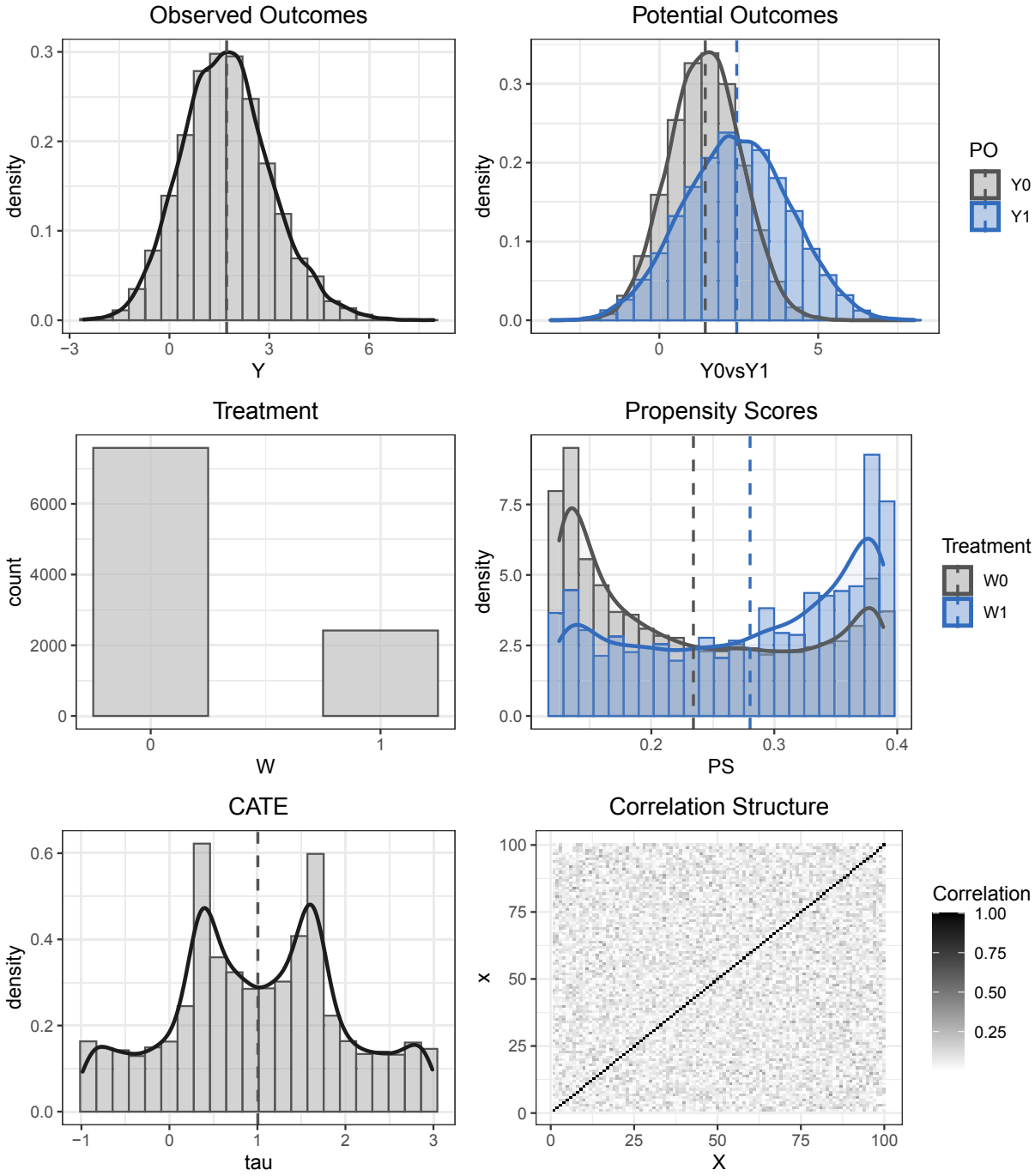
1.A.1.5 Simulation 5: unbalanced treatment and linear CATE

Figure 1.A.5: Descriptive Statistics for the Validation Data in Simulation 5



1.A.1.6 Main Simulation: unbalanced treatment and nonlinear CATE

Figure 1.A.6: Descriptive Statistics for the Validation Data in Main Simulation



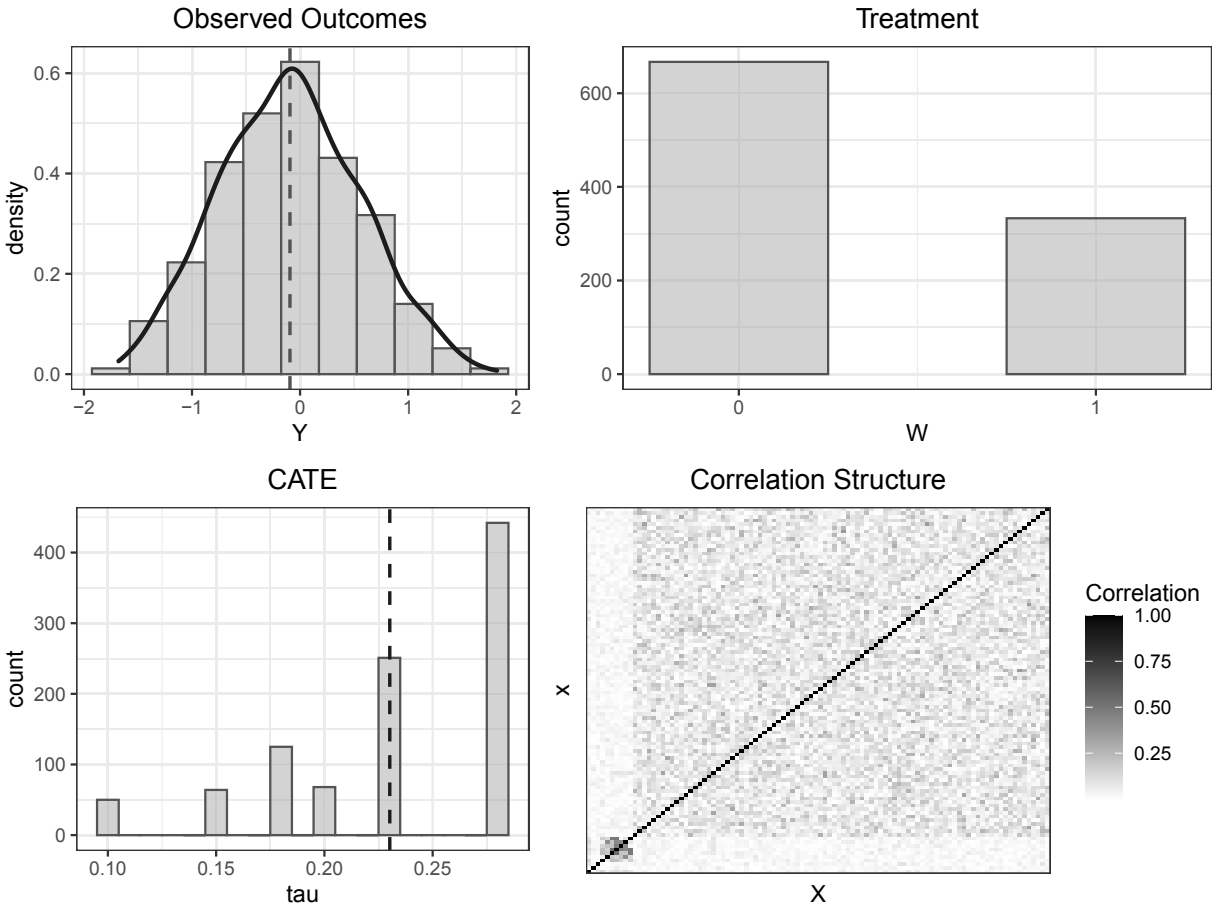
1.A.2 Empirical Simulation

This appendix provides a comprehensive overview of the variables in the augmented real dataset as well as descriptive statistics thereof. Similarly to the results from the main simulation, we plot the distribution of the observed realized outcomes, Y_i, as well as the distribution of the treatment indicator, W_i. Analogously, we plot the distribution of the true treatment effects, $\tau(X_i)$ together with the correlation heat map for the covariates X_i. The respective figures for the distributions of the potential outcomes and the propensity scores under treatment and under control are omitted due to missing data availability for these quantities. The corresponding figures and tables are listed below.

Table 1.A.1: Variable description of the 2018 ACIC dataset. Source: Carvalho, Feller, Murray, Woody, and Yeager (2019).

Variable	Description
Y	outcome measure of achievement recorded post-treatment (continuous variable)
W	treatment indicating receipt of the intervention (binary variable)
S3	student's self-reported expectations for success in the future, a proxy for prior achievement, measured prior to random assignment (ordered categorical variable)
C1	student's race/ethnicity (unordered categorical variable)
C2	student's identified gender (binary variable)
C3	student's first generation status, i.e. first in family to go to college (binary variable)
XC	urbanicity of the school, i.e. rural, suburban, etc. (unordered categorical variable)
X1	school-level mean of students' fixed mindsets, reported prior to random assignment (continuous variable)
X2	school achievement level, measured by test scores and college preparation for the previous 4 cohorts of students (continuous variable)
X3	school racial/ethnic minority composition, i.e. percentage of student body that is Black, Latino, or Native American (continuous variable)
X4	school poverty concentration, i.e. percentage of students who are from families whose incomes fall below the federal poverty line (continuous variable)
X5	School size, i.e. total number of students in all four grade levels in the school (continuous variable)

Figure 1.A.7: Descriptive Statistics for the Validation Data in Empirical Simulation



1.B Simulation Results

1.B.1 Main Results

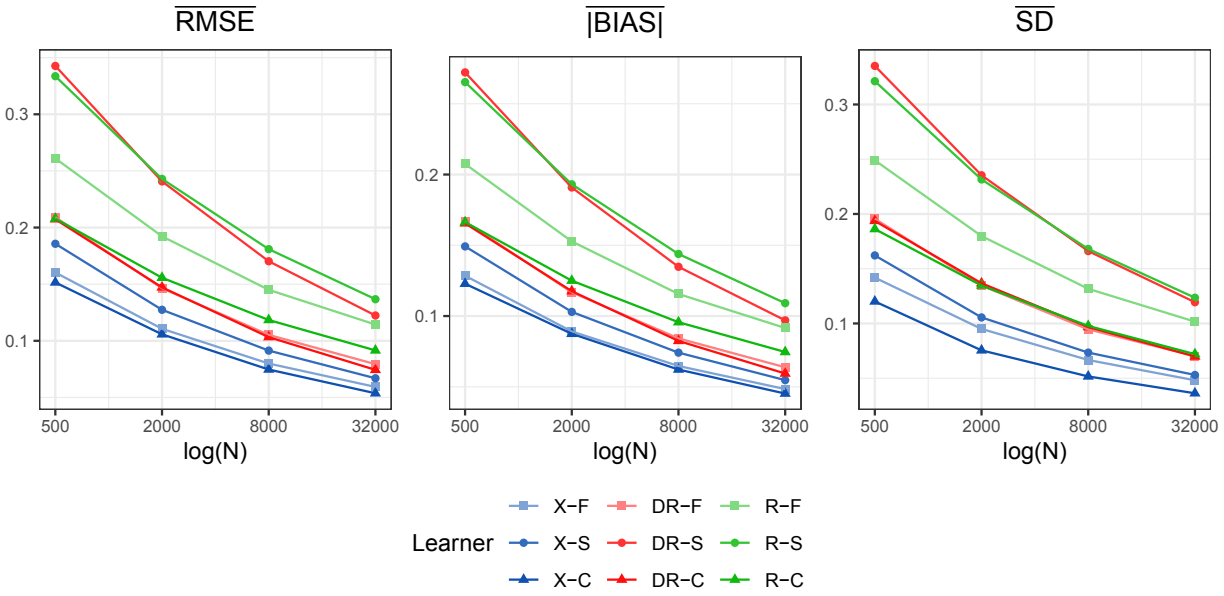
1.B.1.1 Simulation 1: balanced treatment and constant zero CATE

Table 1.B.1: CATE Results for Simulation 1

	$\overline{RMSE}$				$\overline{\lvert BIAS \rvert}$				$\overline{SD}$				$\overline{JB}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	**0.008**	**0.009**	**0.013**	**0.018**	**0.005**	**0.006**	**0.010**	**0.014**	**0.007**	**0.008**	**0.012**	**0.016**	21102.232	3656.421	285.128	12.451
S-W	0.037	0.038	0.049	0.059	0.023	0.025	0.036	0.047	0.033	0.032	0.040	0.047	20410.604	4973.704	364.501	12.494
T	0.225	0.168	0.128	0.101	0.180	0.135	0.103	0.082	0.206	0.149	0.109	0.083	2.071	2.280	**2.000**	**1.912**
X-F	0.160	0.111	0.080	0.059	0.128	0.089	0.065	0.048	0.142	0.095	0.067	0.048	1.689	2.442	2.068	2.002
X-S	0.186	0.127	0.091	0.067	0.149	0.103	0.074	0.055	0.162	0.106	0.073	0.053	1.916	**2.152**	2.180	2.125
X-C	0.152	0.106	0.075	0.054	0.123	0.087	0.062	0.045	0.120	0.075	0.052	0.036	**1.293**	2.393	2.023	1.982
DR-F	0.209	0.146	0.105	0.079	0.167	0.117	0.084	0.064	0.196	0.135	0.095	0.070	4.677	18.496	17.015	7.978
DR-S	0.343	0.241	0.170	0.122	0.272	0.191	0.135	0.097	0.335	0.235	0.166	0.119	5.548	15.737	30.661	41.289
DR-C	0.207	0.147	0.103	0.074	0.166	0.118	0.082	0.059	0.194	0.137	0.097	0.070	2.606	3.632	9.329	13.029
R-F	0.261	0.192	0.145	0.114	0.208	0.153	0.116	0.092	0.249	0.180	0.132	0.102	5.911	23.800	26.134	14.008
R-S	0.334	0.243	0.181	0.137	0.265	0.193	0.144	0.109	0.321	0.232	0.168	0.124	3.192	5.140	15.179	15.980
R-C	0.208	0.156	0.118	0.092	0.166	0.125	0.096	0.075	0.186	0.135	0.098	0.072	2.117	2.586	3.209	3.779

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{SD}$ and $\overline{JB}$ show the mean values of the root mean squared error, absolute bias, standard deviation and the Jarque-Bera test statistic of all 10′000 CATE estimates from the validation sample. The critical values for the JB test statistic are 5.991 and 9.210 at the 5% and 1% level, respectively. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively. Bold numbers indicate the best performing meta-learner for given measure and sample size.

Figure 1.B.1: CATE Results for Simulation 1



Note: The results for $\overline{RMSE}$, $\overline{|BIAS|}$, and $\overline{SD}$ show the mean values of the root mean squared error, absolute bias, and standard deviation of all 10′000 CATE estimates from the validation sample. The figure shows the results based on the increasing training samples of $\{500, 2'000, 8'000, 32'000\}$ observations displayed on the log scale. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

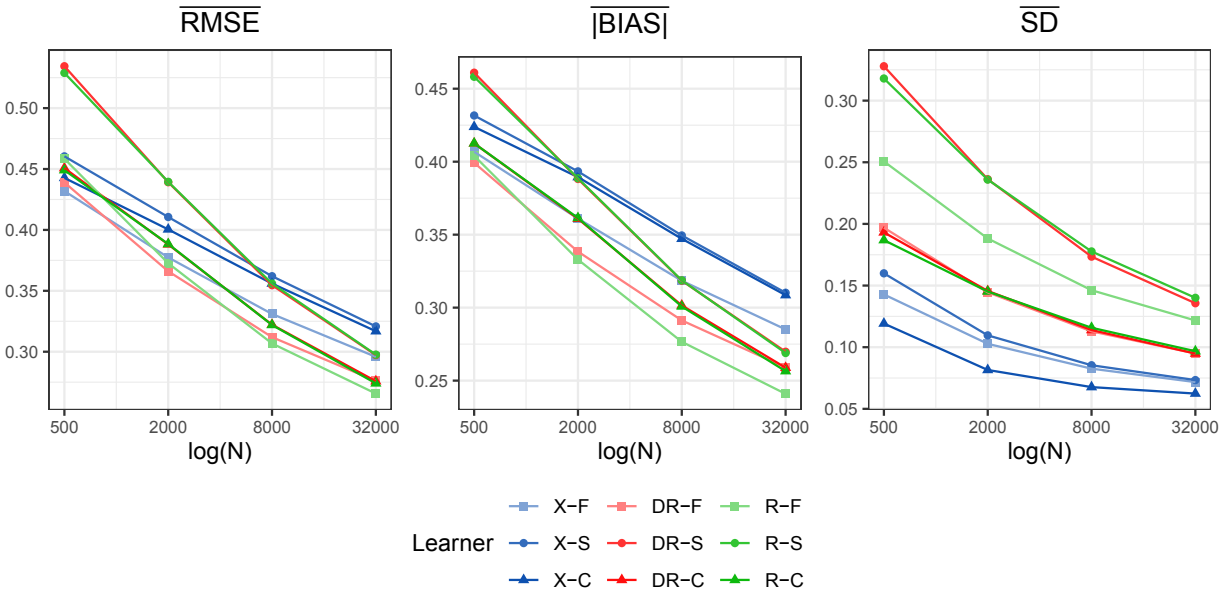
1.B.1.2 Simulation 2: balanced treatment and complex nonlinear CATE

Table 1.B.2: CATE Results for Simulation 2

	$\overline{RMSE}$				$\overline{\lvert BIAS \rvert}$				$\overline{SD}$				$\overline{JB}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.527	0.442	0.374	0.326	0.522	0.434	0.366	0.317	**0.055**	**0.068**	**0.066**	0.064	711.571	17.960	4.186	2.537
S-W	0.463	**0.357**	**0.303**	**0.265**	0.431	**0.328**	0.280	0.246	0.177	0.151	0.120	0.099	268.394	2.900	2.520	2.207
T	0.434	0.358	**0.303**	**0.265**	**0.392**	**0.328**	0.280	0.246	0.204	0.154	0.120	0.099	2.206	2.464	2.466	2.250
X-F	**0.432**	0.377	0.331	0.296	0.407	0.361	0.318	0.285	0.143	0.103	0.083	0.072	1.915	2.167	**1.936**	1.906
X-S	0.460	0.411	0.362	0.321	0.432	0.393	0.349	0.310	0.160	0.110	0.085	0.073	2.048	**2.046**	2.156	1.957
X-C	0.443	0.400	0.356	0.317	0.424	0.389	0.347	0.309	0.119	0.082	0.068	**0.062**	**1.417**	2.139	1.955	**1.900**
DR-F	0.439	0.366	0.312	0.276	0.399	0.338	0.291	0.259	0.197	0.144	0.113	0.095	3.392	2.919	2.104	1.936
DR-S	0.534	0.439	0.355	0.297	0.461	0.388	0.318	0.270	0.328	0.236	0.173	0.136	5.134	8.959	4.011	2.371
DR-C	0.451	0.388	0.322	0.276	0.413	0.361	0.302	0.259	0.193	0.146	0.114	0.095	2.498	2.525	2.158	1.980
R-F	0.458	0.373	0.307	0.266	0.404	0.333	**0.277**	**0.241**	0.251	0.188	0.146	0.122	4.201	4.206	2.437	2.021
R-S	0.529	0.439	0.356	0.298	0.458	0.389	0.319	0.269	0.318	0.236	0.178	0.140	2.989	3.550	4.630	3.400
R-C	0.449	0.388	0.322	0.274	0.413	0.361	0.301	0.256	0.187	0.145	0.116	0.097	2.195	2.280	2.107	1.940

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{SD}$ and $\overline{JB}$ show the mean values of the root mean squared error, absolute bias, standard deviation and the Jarque-Bera test statistic of all 10′000 CATE estimates from the validation sample. The critical values for the JB test statistic are 5.991 and 9.210 at the 5% and 1% level, respectively. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively. Bold numbers indicate the best performing meta-learner for given measure and sample size.

Figure 1.B.2: CATE Results for Simulation 2



Note: The results for $\overline{RMSE}$, $\overline{|BIAS|}$, and $\overline{SD}$ show the mean values of the root mean squared error, absolute bias, and standard deviation of all 10′000 CATE estimates from the validation sample. The figure shows the results based on the increasing training samples of $\{500, 2'000, 8'000, 32'000\}$ observations displayed on the log scale. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

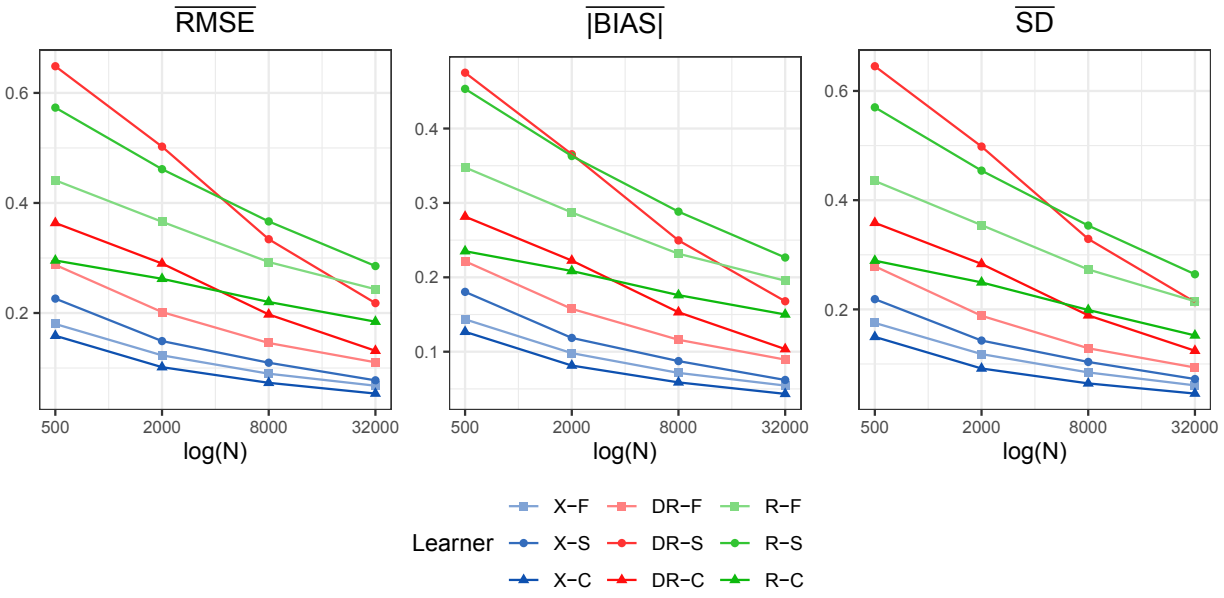
1.B.1.3 Simulation 3: highly unbalanced treatment and constant non-zero CATE

Table 1.B.3: CATE Results for Simulation 3

	$\overline{RMSE}$				$\overline{\lvert BIAS \rvert}$				$\overline{SD}$				$\overline{JB}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.645	0.475	0.359	0.279	0.638	0.468	0.352	0.272	**0.099**	**0.084**	0.072	0.062	2.888	2.667	**2.111**	1.981
S-W	0.246	0.191	0.146	0.111	0.197	0.154	0.119	0.091	0.233	0.163	0.121	0.090	4.611	2.281	2.385	1.993
T	0.244	0.191	0.146	0.111	0.195	0.154	0.119	0.091	0.227	0.164	0.121	0.090	2.806	2.271	2.243	**1.964**
X-F	0.180	0.123	0.090	0.068	0.144	0.098	0.072	0.054	0.175	0.118	0.085	0.061	3.663	2.552	4.441	2.820
X-S	0.226	0.149	0.110	0.078	0.180	0.119	0.087	0.062	0.219	0.143	0.104	0.072	**2.367**	2.678	3.058	3.192
X-C	**0.159**	**0.102**	**0.073**	**0.054**	**0.127**	**0.081**	**0.059**	**0.043**	0.150	0.092	**0.064**	**0.046**	6.541	**1.969**	2.263	1.994
DR-F	0.287	0.202	0.146	0.110	0.222	0.158	0.116	0.089	0.279	0.188	0.129	0.093	3060.536	812.294	244.016	38.545
DR-S	0.649	0.502	0.334	0.218	0.475	0.365	0.250	0.168	0.645	0.498	0.329	0.213	1276.433	1496.545	795.711	258.858
DR-C	0.364	0.290	0.197	0.131	0.282	0.223	0.153	0.104	0.359	0.283	0.189	0.124	112.249	149.500	126.268	43.274
R-F	0.441	0.366	0.293	0.243	0.348	0.287	0.232	0.195	0.435	0.354	0.273	0.215	14.590	27.616	19.045	8.171
R-S	0.573	0.461	0.366	0.285	0.453	0.363	0.288	0.227	0.570	0.454	0.353	0.264	7.887	12.474	18.638	11.149
R-C	0.295	0.262	0.220	0.184	0.235	0.208	0.176	0.150	0.289	0.249	0.199	0.152	2.822	3.570	4.324	3.162

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{SD}$ and $\overline{JB}$ show the mean values of the root mean squared error, absolute bias, standard deviation and the Jarque-Bera test statistic of all 10′000 CATE estimates from the validation sample. The critical values for the JB test statistic are 5.991 and 9.210 at the 5% and 1% level, respectively. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively. Bold numbers indicate the best performing meta-learner for given measure and sample size.

Figure 1.B.3: CATE Results for Simulation 3



Note: The results for $\overline{RMSE}$, $\overline{|BIAS|}$, and $\overline{SD}$ show the mean values of the root mean squared error, absolute bias, and standard deviation of all 10′000 CATE estimates from the validation sample. The figure shows the results based on the increasing training samples of $\{500, 2'000, 8'000, 32'000\}$ observations displayed on the log scale. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

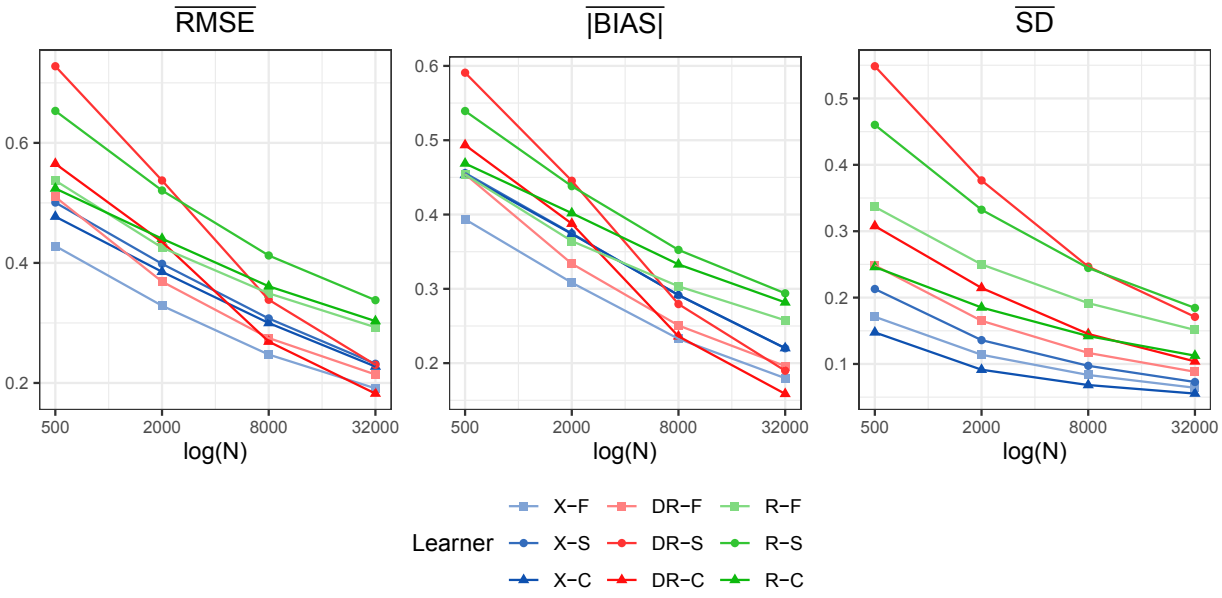
1.B.1.4 Simulation 4: unbalanced treatment and simple CATE

Table 1.B.4: CATE Results for Simulation 4

	$\overline{RMSE}$				$\overline{\lvert BIAS \rvert}$				$\overline{SD}$				$\overline{JB}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.834	0.616	0.472	0.370	0.825	0.606	0.462	0.361	**0.105**	**0.090**	0.078	0.069	**1.935**	2.120	2.075	1.951
S-W	0.443	0.336	0.258	0.206	**0.390**	**0.300**	**0.233**	0.187	0.229	0.162	0.120	0.093	2.575	2.330	2.124	1.957
T	0.443	0.335	0.258	0.206	**0.390**	**0.300**	**0.233**	0.187	0.229	0.163	0.120	0.093	2.552	2.278	2.130	1.938
X-F	**0.428**	**0.329**	**0.247**	0.191	0.394	0.308	**0.233**	0.180	0.171	0.114	0.083	0.064	3.731	2.312	2.210	1.978
X-S	0.501	0.399	0.307	0.232	0.456	0.375	0.291	0.220	0.213	0.136	0.097	0.073	6.602	2.762	2.298	2.113
X-C	0.477	0.385	0.300	0.227	0.453	0.374	0.292	0.220	0.148	0.091	**0.068**	**0.055**	5.617	**2.026**	**2.023**	**1.898**
DR-F	0.510	0.369	0.275	0.214	0.454	0.334	0.251	0.196	0.249	0.165	0.117	0.088	116.949	156.252	41.933	5.158
DR-S	0.728	0.537	0.339	0.230	0.591	0.445	0.279	0.190	0.549	0.377	0.247	0.171	497.136	530.045	407.510	97.233
DR-C	0.565	0.435	0.269	**0.182**	0.493	0.388	0.236	**0.159**	0.308	0.215	0.145	0.104	51.595	50.726	42.424	15.770
R-F	0.537	0.426	0.349	0.293	0.454	0.364	0.303	0.258	0.337	0.250	0.192	0.151	8.764	13.754	7.349	2.839
R-S	0.653	0.521	0.412	0.338	0.539	0.438	0.352	0.294	0.460	0.332	0.245	0.184	7.025	6.592	7.512	5.029
R-C	0.524	0.440	0.361	0.303	0.469	0.402	0.333	0.282	0.246	0.185	0.142	0.113	2.700	2.900	2.732	2.318

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{SD}$ and $\overline{JB}$ show the mean values of the root mean squared error, absolute bias, standard deviation and the Jarque-Bera test statistic of all 10′000 CATE estimates from the validation sample. The critical values for the JB test statistic are 5.991 and 9.210 at the 5% and 1% level, respectively. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively. Bold numbers indicate the best performing meta-learner for given measure and sample size.

Figure 1.B.4: CATE Results for Simulation 4



Note: The results for $\overline{RMSE}$, $\overline{|BIAS|}$, and $\overline{SD}$ show the mean values of the root mean squared error, absolute bias, and standard deviation of all 10′000 CATE estimates from the validation sample. The figure shows the results based on the increasing training samples of $\{500, 2'000, 8'000, 32'000\}$ observations displayed on the log scale. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

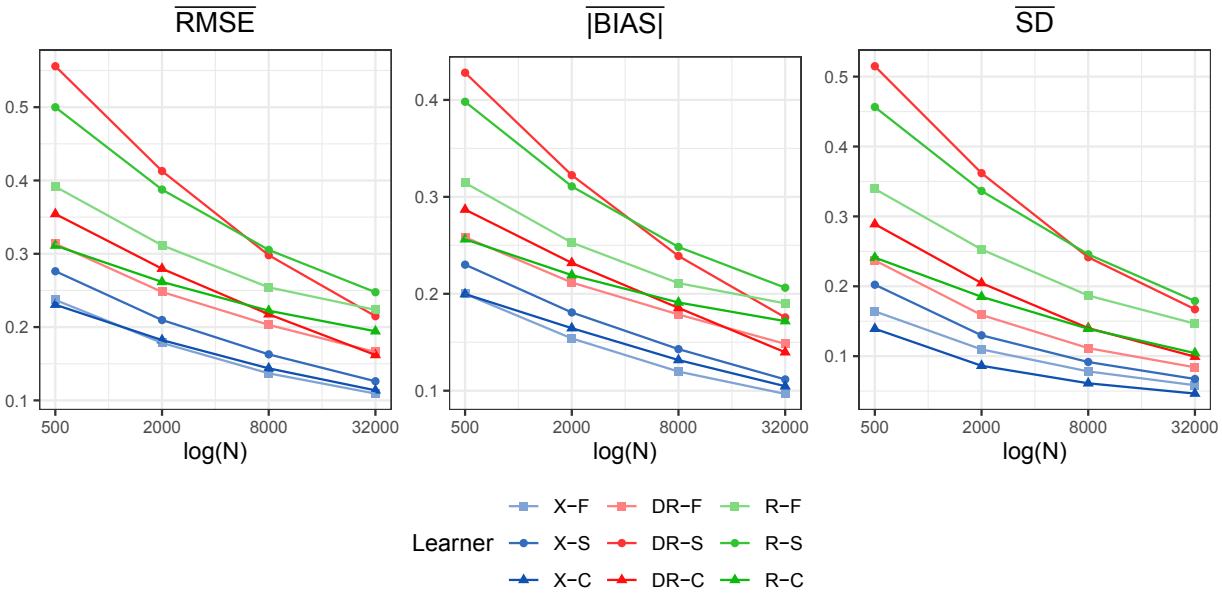
1.B.1.5 Simulation 5: unbalanced treatment and linear CATE

Table 1.B.5: CATE Results for Simulation 5

	$\overline{RMSE}$				$\overline{\lvert BIAS \rvert}$				$\overline{SD}$				$\overline{JB}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.823	0.606	0.461	0.358	0.817	0.599	0.454	0.351	**0.101**	0.087	0.075	0.066	**1.796**	2.150	2.057	1.986
S-W	0.305	0.244	0.196	0.164	0.255	0.209	0.170	0.145	0.222	0.159	0.117	0.089	2.457	2.189	2.054	1.957
T	0.305	0.244	0.196	0.164	0.255	0.209	0.171	0.145	0.222	0.159	0.117	0.089	2.497	2.173	2.026	**1.936**
X-F	0.237	**0.178**	**0.137**	**0.109**	**0.200**	**0.154**	**0.120**	**0.097**	0.164	0.110	0.078	0.058	3.329	2.228	2.102	2.022
X-S	0.276	0.210	0.163	0.126	0.230	0.181	0.143	0.112	0.202	0.130	0.092	0.067	6.639	2.811	2.296	2.421
X-C	**0.231**	0.182	0.144	0.114	**0.200**	0.165	0.132	0.105	0.139	**0.086**	**0.061**	**0.046**	4.474	**2.014**	**2.004**	2.037
DR-F	0.314	0.248	0.203	0.166	0.258	0.212	0.179	0.148	0.237	0.159	0.112	0.084	123.780	364.063	249.515	26.116
DR-S	0.556	0.413	0.298	0.215	0.428	0.322	0.239	0.176	0.515	0.362	0.242	0.167	453.484	685.910	651.725	174.087
DR-C	0.354	0.280	0.217	0.162	0.287	0.232	0.185	0.140	0.289	0.205	0.140	0.099	50.509	61.849	72.770	22.771
R-F	0.392	0.312	0.254	0.223	0.314	0.253	0.211	0.190	0.339	0.253	0.187	0.146	12.888	28.931	17.700	4.568
R-S	0.500	0.388	0.305	0.248	0.398	0.311	0.248	0.206	0.457	0.336	0.246	0.179	9.107	10.173	16.684	12.581
R-C	0.311	0.262	0.222	0.194	0.256	0.219	0.191	0.172	0.241	0.185	0.139	0.104	2.925	3.617	3.874	3.072

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{SD}$ and $\overline{JB}$ show the mean values of the root mean squared error, absolute bias, standard deviation and the Jarque-Bera test statistic of all 10′000 CATE estimates from the validation sample. The critical values for the JB test statistic are 5.991 and 9.210 at the 5% and 1% level, respectively. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively. Bold numbers indicate the best performing meta-learner for given measure and sample size.

Figure 1.B.5: CATE Results for Simulation 5



Note: The results for $\overline{RMSE}$, $\overline{|BIAS|}$, and $\overline{SD}$ show the mean values of the root mean squared error, absolute bias, and standard deviation of all 10′000 CATE estimates from the validation sample. The figure shows the results based on the increasing training samples of $\{500, 2'000, 8'000, 32'000\}$ observations displayed on the log scale. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

1.B.2 Supplementary Results

This appendix provides supplementary results based on additional performance measures, complementing those from Section 1.4.1. To understand the simulation noise and thus the precision the average RMSE is measured with, we compute the standard error of the average RMSE following Knaus et al. (2021) as:

$$SE(\overline{RMSE}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{N^V}\sum_{i=1}^{N^V}\left(\tau(X_i) - \hat{\tau}^r(X_i)\right)^2 - \overline{RMSE}\right)^2}.$$

Additionally, besides the absolute bias, we evaluate also the bias without the absolute value given by:

$$BIAS\big(\hat{\tau}(X_i)\big) = \frac{1}{R}\sum_{r=1}^{R}\left(\tau(X_i) - \hat{\tau}^r(X_i)\right)$$

We further evaluate also the components of the Jarque-Bera statistic separately, namely the skewness, i.e. $S\big(\hat{\tau}(X_i)\big)$ and the kurtosis, i.e. $K\big(\hat{\tau}(X_i)\big)$ defined by:

$$S\big(\hat{\tau}(X_i)\big) = \frac{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\tau}^r(X_i) - \frac{1}{R}\sum_{r=1}^{R}\hat{\tau}^r(X_i)\right)^3}{\left(\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\tau}^r(X_i) - \frac{1}{R}\sum_{r=1}^{R}\hat{\tau}^r(X_i)\right)^2\right)^{3/2}} \quad \text{and} \quad K\big(\hat{\tau}(X_i)\big) = \frac{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\tau}^r(X_i) - \frac{1}{R}\sum_{r=1}^{R}\hat{\tau}^r(X_i)\right)^4}{\left(\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\tau}^r(X_i) - \frac{1}{R}\sum_{r=1}^{R}\hat{\tau}^r(X_i)\right)^2\right)^2}.$$

As in the main simulation results, we report the averages of the above measures over the validation sample N^V. Complementary to the average values of the Jarque-Bera statistic presented in the main text, herein we report the share of CATEs for which the normality gets rejected at the 5% level. In order to further evaluate the performance on the replication level we compute the correlation between the true and the estimated treatment effects given by:

$$CORR = \frac{1}{R}\sum_{r=1}^{R}\left(\rho\big(\boldsymbol{\tau}, \hat{\boldsymbol{\tau}}^r\big)\right)$$

where $\boldsymbol{\tau}$ is a vector of size N^V containing the true treatment effects from the validation sample and $\hat{\boldsymbol{\tau}}^r$ is a vector of size N^V containing the estimated treatment effects for the validation sample at the replication r, while $\rho(\cdot)$ denotes the correlation function. Similarly, we compute also the variance ratio of the true and the estimated treatment effects as follows:

$$VARR = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{Var(\hat{\boldsymbol{\tau}}^r)}{Var(\boldsymbol{\tau})}\right)$$

where $Var(\cdot)$ denotes the variance. The full results including the main and the supplementary performance measures are listed in Tables 1.B.6 - 1.B.12 below.

1.B.2.1 Simulation 1: balanced treatment and constant zero CATE

Table 1.B.6: CATE Results for Simulation 1

	$\overline{RMSE}$				$SE(\overline{RMSE})$				$\overline{\lvert BIAS\rvert}$				$\overline{BIAS}$				$\overline{SD}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.008	0.009	0.013	0.018	0.005	0.005	0.006	0.004	0.005	0.006	0.010	0.014	-0.003	-0.004	-0.006	-0.008	0.007	0.008	0.012	0.016
S-W	0.037	0.038	0.049	0.059	0.028	0.024	0.024	0.015	0.023	0.025	0.036	0.047	-0.017	-0.020	-0.028	-0.033	0.033	0.032	0.040	0.047
T	0.225	0.168	0.128	0.101	0.046	0.024	0.013	0.007	0.180	0.135	0.103	0.082	-0.087	-0.072	-0.060	-0.048	0.206	0.149	0.109	0.083
X-F	0.160	0.111	0.080	0.059	0.054	0.026	0.014	0.006	0.128	0.089	0.065	0.048	-0.074	-0.055	-0.041	-0.029	0.142	0.095	0.067	0.048
X-S	0.186	0.127	0.091	0.067	0.071	0.034	0.018	0.008	0.149	0.103	0.074	0.055	-0.090	-0.071	-0.053	-0.037	0.162	0.106	0.073	0.053
X-C	0.152	0.106	0.075	0.054	0.068	0.034	0.018	0.008	0.123	0.087	0.062	0.045	-0.092	-0.074	-0.052	-0.036	0.120	0.075	0.052	0.036
DR-F	0.209	0.146	0.105	0.079	0.044	0.021	0.011	0.006	0.167	0.117	0.084	0.064	-0.072	-0.055	-0.043	-0.032	0.196	0.135	0.095	0.070
DR-S	0.343	0.241	0.170	0.122	0.071	0.029	0.013	0.006	0.272	0.191	0.135	0.097	-0.069	-0.045	-0.030	-0.017	0.335	0.235	0.166	0.119
DR-C	0.207	0.147	0.103	0.074	0.046	0.020	0.009	0.004	0.166	0.118	0.082	0.059	-0.072	-0.050	-0.029	-0.015	0.194	0.137	0.097	0.070
R-F	0.261	0.192	0.145	0.114	0.039	0.020	0.012	0.006	0.208	0.153	0.116	0.092	-0.077	-0.066	-0.058	-0.048	0.249	0.180	0.132	0.102
R-S	0.334	0.243	0.181	0.137	0.073	0.032	0.018	0.010	0.265	0.193	0.144	0.109	-0.089	-0.072	-0.064	-0.055	0.321	0.232	0.168	0.124
R-C	0.208	0.156	0.118	0.092	0.050	0.026	0.015	0.008	0.166	0.125	0.096	0.075	-0.092	-0.077	-0.065	-0.054	0.186	0.135	0.098	0.072

	$\overline{SKEW}$				$\overline{KURT}$				$JB\%$				$CORR$				$VARR$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	2.638	1.809	0.929	0.307	17.334	10.674	5.654	3.539	1.000	1.000	1.000	0.539								
S-W	2.744	2.171	1.150	0.339	17.212	12.179	6.003	3.493	1.000	1.000	1.000	0.586								
T	0.005	-0.009	-0.004	-0.008	3.016	3.038	3.006	2.983	0.058	0.075	0.052	0.044								
X-F	0.010	-0.030	-0.001	-0.022	2.990	3.058	3.002	2.985	0.028	0.091	0.055	0.051								
X-S	0.017	-0.023	0.003	-0.008	2.951	3.023	3.029	3.003	0.036	0.066	0.063	0.060								
X-C	0.002	-0.040	0.003	-0.018	2.985	3.061	3.006	2.986	0.010	0.086	0.052	0.049								
DR-F	-0.008	-0.031	-0.020	-0.010	3.112	3.235	3.196	3.100	0.211	0.253	0.154	0.086								
DR-S	-0.009	-0.040	-0.060	-0.040	3.150	3.308	3.410	3.359	0.271	0.359	0.274	0.164								
DR-C	-0.013	-0.031	-0.039	-0.019	3.054	3.099	3.158	3.128	0.102	0.158	0.152	0.099								
R-F	0.017	0.015	0.013	0.007	3.145	3.296	3.256	3.147	0.275	0.296	0.181	0.104								
R-S	0.027	0.009	0.014	0.012	3.077	3.161	3.251	3.222	0.150	0.216	0.211	0.132								
R-C	0.003	0.003	0.001	0.008	3.024	3.065	3.081	3.067	0.065	0.092	0.104	0.083								

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{BIAS}$, $\overline{SD}$, $\overline{SKEW}$, and $\overline{KURT}$ show the mean values of the root mean squared error, absolute bias, bias, standard deviation, skewness and kurtosis of all 10′000 CATE estimates from the validation sample. $SE(\overline{RMSE})$ depicts the standard error of the average RMSE and $JB\%$ presents the share of CATEs for which the Jarque-Bera test has been rejected at the 5% level. The results for $CORR$ and $VARR$ show the values of the correlation and variance ratio between the true and the estimated CATEs over all replications. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

1.B.2.2 Simulation 2: balanced treatment and complex nonlinear CATE

Table 1.B.7: CATE Results for Simulation 2

	$\overline{RMSE}$				$SE(\overline{RMSE})$				$\overline{\lvert BIAS\rvert}$				$\overline{BIAS}$				$\overline{SD}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.527	0.442	0.374	0.326	0.114	0.092	0.075	0.064	0.522	0.434	0.366	0.317	-0.369	-0.258	-0.190	-0.143	0.055	0.068	0.066	0.064
S-W	0.463	0.357	0.303	0.265	0.084	0.050	0.045	0.041	0.431	0.328	0.280	0.246	-0.159	-0.036	-0.029	-0.023	0.177	0.151	0.120	0.099
T	0.434	0.358	0.303	0.265	0.056	0.049	0.045	0.041	0.392	0.328	0.280	0.246	-0.040	-0.034	-0.029	-0.023	0.204	0.154	0.120	0.099
X-F	0.432	0.377	0.331	0.296	0.069	0.066	0.061	0.056	0.407	0.361	0.318	0.285	-0.031	-0.022	-0.017	-0.012	0.143	0.103	0.083	0.072
X-S	0.460	0.411	0.362	0.321	0.071	0.071	0.067	0.061	0.432	0.393	0.349	0.310	-0.041	-0.029	-0.021	-0.015	0.160	0.110	0.085	0.073
X-C	0.443	0.400	0.356	0.317	0.076	0.075	0.070	0.063	0.424	0.389	0.347	0.309	-0.042	-0.032	-0.022	-0.014	0.119	0.082	0.068	0.062
DR-F	0.439	0.366	0.312	0.276	0.058	0.053	0.048	0.044	0.399	0.338	0.291	0.259	-0.032	-0.026	-0.021	-0.016	0.197	0.144	0.113	0.095
DR-S	0.534	0.439	0.355	0.297	0.059	0.050	0.044	0.038	0.461	0.388	0.318	0.270	-0.027	-0.017	-0.013	-0.008	0.328	0.236	0.173	0.136
DR-C	0.451	0.388	0.322	0.276	0.060	0.057	0.050	0.044	0.413	0.361	0.302	0.259	-0.030	-0.021	-0.013	-0.006	0.193	0.146	0.114	0.095
R-F	0.458	0.373	0.307	0.266	0.052	0.045	0.039	0.034	0.404	0.333	0.277	0.241	-0.039	-0.034	-0.030	-0.027	0.251	0.188	0.146	0.122
R-S	0.529	0.439	0.356	0.298	0.060	0.050	0.043	0.038	0.458	0.389	0.319	0.269	-0.042	-0.036	-0.033	-0.030	0.318	0.236	0.178	0.140
R-C	0.449	0.388	0.322	0.274	0.061	0.058	0.050	0.044	0.413	0.361	0.301	0.256	-0.045	-0.040	-0.034	-0.028	0.187	0.145	0.116	0.097

	$\overline{SKEW}$				$\overline{KURT}$				$JB\%$				$CORR$				$VARR$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	-1.172	-0.282	-0.138	-0.087	4.608	3.106	3.021	3.003	1.000	0.832	0.206	0.087	0.430	0.737	0.850	0.890	772.103	31.143	11.674	6.871
S-W	-0.820	-0.054	-0.003	0.003	3.494	3.015	3.027	2.999	1.000	0.116	0.075	0.058	0.490	0.728	0.846	0.896	51.036	5.794	4.379	3.494
T	0.000	-0.007	0.002	0.004	3.030	3.043	3.025	3.002	0.070	0.085	0.072	0.057	0.465	0.722	0.846	0.896	6.865	5.692	4.382	3.492
X-F	0.001	-0.012	0.007	0.001	3.010	3.026	2.984	2.971	0.045	0.067	0.045	0.043	0.458	0.723	0.847	0.891	20.826	14.657	8.854	5.962
X-S	0.008	-0.014	0.011	-0.007	2.931	3.022	3.006	2.981	0.043	0.058	0.062	0.050	0.266	0.553	0.785	0.868	25.303	22.204	13.868	8.249
X-C	0.000	-0.022	0.007	-0.006	2.993	3.029	2.983	2.975	0.015	0.065	0.045	0.045	0.391	0.677	0.835	0.886	55.123	33.303	15.678	8.554
DR-F	-0.003	-0.013	-0.000	0.000	3.082	3.054	3.002	2.976	0.151	0.100	0.052	0.046	0.429	0.707	0.842	0.891	7.946	6.991	5.202	4.024
DR-S	-0.002	-0.021	-0.009	-0.010	3.135	3.199	3.090	3.016	0.247	0.249	0.113	0.063	0.224	0.455	0.718	0.841	3.595	4.565	4.468	3.649
DR-C	-0.006	-0.017	-0.005	0.001	3.057	3.047	3.005	2.978	0.099	0.087	0.056	0.051	0.359	0.628	0.828	0.895	9.232	8.713	5.869	4.112
R-F	0.012	0.006	0.015	0.010	3.107	3.086	3.018	2.986	0.200	0.132	0.064	0.050	0.398	0.668	0.824	0.884	4.473	4.449	3.642	3.005
R-S	0.014	0.003	0.017	0.011	3.072	3.108	3.073	3.008	0.130	0.151	0.097	0.059	0.227	0.456	0.716	0.840	3.805	4.558	4.304	3.512
R-C	0.001	0.002	0.013	0.017	3.033	3.033	3.000	2.983	0.069	0.074	0.053	0.048	0.363	0.630	0.828	0.897	9.803	8.713	5.702	3.970

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{BIAS}$, $\overline{SD}$, $\overline{SKEW}$, and $\overline{KURT}$ show the mean values of the root mean squared error, absolute bias, bias, standard deviation, skewness and kurtosis of all 10′000 CATE estimates from the validation sample. $SE(\overline{RMSE})$ depicts the standard error of the average RMSE and $JB\%$ presents the share of CATEs for which the Jarque-Bera test has been rejected at the 5% level. The results for $CORR$ and $VARR$ show the values of the correlation and variance ratio between the true and the estimated CATEs over all replications. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

1.B.2.3 Simulation 3: highly unbalanced treatment and constant non-zero CATE

Table 1.B.8: CATE Results for Simulation 3

	$\overline{RMSE}$				$SE(\overline{RMSE})$				$\overline{\lvert BIAS\rvert}$				$\overline{BIAS}$				$\overline{SD}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.645	0.475	0.359	0.279	0.077	0.043	0.025	0.014	0.638	0.468	0.352	0.272	0.638	0.468	0.352	0.272	0.099	0.084	0.072	0.062
S-W	0.246	0.191	0.146	0.111	0.059	0.025	0.015	0.009	0.197	0.154	0.119	0.091	-0.045	-0.050	-0.042	-0.032	0.233	0.163	0.121	0.090
T	0.244	0.191	0.146	0.111	0.053	0.025	0.015	0.008	0.195	0.154	0.119	0.091	-0.057	-0.050	-0.041	-0.032	0.227	0.164	0.121	0.090
X-F	0.180	0.123	0.090	0.068	0.060	0.024	0.012	0.006	0.144	0.098	0.072	0.054	-0.041	-0.033	-0.025	-0.016	0.175	0.118	0.085	0.061
X-S	0.226	0.149	0.110	0.078	0.095	0.040	0.019	0.007	0.180	0.119	0.087	0.062	-0.058	-0.042	-0.034	-0.021	0.219	0.143	0.104	0.072
X-C	0.159	0.102	0.073	0.054	0.074	0.031	0.015	0.007	0.127	0.081	0.059	0.043	-0.053	-0.043	-0.033	-0.021	0.150	0.092	0.064	0.046
DR-F	0.287	0.202	0.146	0.110	0.058	0.022	0.012	0.007	0.222	0.158	0.116	0.089	-0.049	-0.040	-0.029	-0.017	0.279	0.188	0.129	0.093
DR-S	0.649	0.502	0.334	0.218	0.204	0.093	0.039	0.018	0.475	0.365	0.250	0.168	-0.052	-0.038	-0.025	-0.007	0.645	0.498	0.329	0.213
DR-C	0.364	0.290	0.197	0.131	0.075	0.032	0.016	0.008	0.282	0.223	0.153	0.104	-0.048	-0.036	-0.026	-0.005	0.359	0.283	0.189	0.124
R-F	0.441	0.366	0.293	0.243	0.048	0.028	0.022	0.020	0.348	0.287	0.232	0.195	0.032	0.040	0.043	0.048	0.435	0.354	0.273	0.215
R-S	0.573	0.461	0.366	0.285	0.141	0.057	0.034	0.025	0.453	0.363	0.288	0.227	0.032	0.038	0.042	0.044	0.570	0.454	0.353	0.264
R-C	0.295	0.262	0.220	0.184	0.044	0.022	0.017	0.019	0.235	0.208	0.176	0.150	0.030	0.041	0.043	0.046	0.289	0.249	0.199	0.152

	$\overline{SKEW}$				$\overline{KURT}$				$JB\%$				$CORR$				$VARR$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.060	0.062	0.037	0.027	2.991	3.007	2.999	2.983	0.113	0.101	0.059	0.052								
S-W	-0.070	-0.005	-0.011	-0.003	3.102	3.039	3.028	2.988	0.277	0.075	0.067	0.049								
T	-0.024	-0.007	-0.009	-0.003	3.062	3.042	3.024	2.987	0.121	0.076	0.067	0.046								
X-F	-0.055	-0.015	-0.023	-0.004	3.080	3.069	3.111	3.035	0.186	0.095	0.123	0.068								
X-S	-0.017	0.030	-0.012	-0.010	3.085	3.074	3.095	3.083	0.084	0.107	0.110	0.091								
X-C	-0.101	-0.014	-0.025	-0.001	3.127	3.020	3.021	2.998	0.430	0.053	0.067	0.049								
DR-F	-0.006	-0.035	-0.041	-0.020	5.349	4.944	3.984	3.340	0.949	0.674	0.337	0.147								
DR-S	-0.011	-0.070	-0.096	-0.095	6.502	7.416	6.268	4.523	1.000	0.995	0.817	0.397								
DR-C	-0.033	-0.047	-0.056	-0.055	4.033	4.468	4.258	3.538	0.996	0.928	0.557	0.229								
R-F	-0.026	-0.025	-0.012	0.006	3.325	3.511	3.375	3.159	0.658	0.542	0.282	0.121								
R-S	0.035	-0.003	-0.018	-0.005	3.222	3.382	3.464	3.292	0.468	0.520	0.378	0.183								
R-C	-0.015	-0.011	-0.006	0.009	3.067	3.129	3.159	3.086	0.117	0.173	0.178	0.105								

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{BIAS}$, $\overline{SD}$, $\overline{SKEW}$, and $\overline{KURT}$ show the mean values of the root mean squared error, absolute bias, bias, standard deviation, skewness and kurtosis of all 10′000 CATE estimates from the validation sample. $SE(\overline{RMSE})$ depicts the standard error of the average RMSE and $JB\%$ presents the share of CATEs for which the Jarque-Bera test has been rejected at the 5% level. The results for $CORR$ and $VARR$ show the values of the correlation and variance ratio between the true and the estimated CATEs over all replications. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

1.B.2.4 Simulation 4: unbalanced treatment and simple CATE

Table 1.B.9: CATE Results for Simulation 4

	$\overline{RMSE}$				$SE(\overline{RMSE})$				$\overline{\lvert BIAS\rvert}$				$\overline{BIAS}$				$\overline{SD}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.834	0.616	0.472	0.370	0.121	0.106	0.091	0.076	0.825	0.606	0.462	0.361	0.825	0.605	0.458	0.354	0.105	0.090	0.078	0.069
S-W	0.443	0.336	0.258	0.206	0.049	0.032	0.024	0.020	0.390	0.300	0.233	0.187	-0.077	-0.071	-0.059	-0.048	0.229	0.162	0.120	0.093
T	0.443	0.335	0.258	0.206	0.049	0.033	0.024	0.020	0.390	0.300	0.233	0.187	-0.076	-0.071	-0.059	-0.048	0.229	0.163	0.120	0.093
X-F	0.428	0.329	0.247	0.191	0.040	0.026	0.017	0.011	0.394	0.308	0.233	0.180	-0.061	-0.052	-0.038	-0.027	0.171	0.114	0.083	0.064
X-S	0.501	0.399	0.307	0.232	0.052	0.031	0.020	0.014	0.456	0.375	0.291	0.220	-0.077	-0.064	-0.050	-0.034	0.213	0.136	0.097	0.073
X-C	0.477	0.385	0.300	0.227	0.033	0.021	0.015	0.010	0.453	0.374	0.292	0.220	-0.079	-0.067	-0.048	-0.034	0.148	0.091	0.068	0.055
DR-F	0.510	0.369	0.275	0.214	0.046	0.033	0.020	0.013	0.454	0.334	0.251	0.196	-0.064	-0.055	-0.038	-0.027	0.249	0.165	0.117	0.088
DR-S	0.728	0.537	0.339	0.230	0.122	0.059	0.030	0.016	0.591	0.445	0.279	0.190	-0.055	-0.050	-0.034	-0.014	0.549	0.377	0.247	0.171
DR-C	0.565	0.435	0.269	0.182	0.043	0.035	0.021	0.012	0.493	0.388	0.236	0.159	-0.069	-0.054	-0.030	-0.010	0.308	0.215	0.145	0.104
R-F	0.537	0.426	0.349	0.293	0.046	0.030	0.021	0.015	0.454	0.364	0.303	0.258	-0.029	-0.022	-0.012	-0.005	0.337	0.250	0.192	0.151
R-S	0.653	0.521	0.412	0.338	0.092	0.049	0.032	0.023	0.539	0.438	0.352	0.294	-0.035	-0.032	-0.019	-0.014	0.460	0.332	0.245	0.184
R-C	0.524	0.440	0.361	0.303	0.032	0.026	0.018	0.015	0.469	0.402	0.333	0.282	-0.039	-0.030	-0.018	-0.011	0.246	0.185	0.142	0.113

	$\overline{SKEW}$				$\overline{KURT}$				$JB\%$				$CORR$				$VARR$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.015	0.028	0.030	0.019	3.010	3.000	2.993	2.976	0.050	0.056	0.056	0.047	0.598	0.816	0.904	0.942	24.819	10.142	5.669	3.726
S-W	-0.027	-0.003	-0.001	-0.005	3.047	3.026	3.002	2.982	0.101	0.076	0.058	0.044	0.507	0.763	0.878	0.926	4.538	3.298	2.466	2.002
T	-0.027	-0.005	-0.002	-0.006	3.044	3.025	3.003	2.977	0.094	0.076	0.060	0.044	0.509	0.764	0.878	0.927	4.528	3.295	2.467	1.994
X-F	-0.064	-0.017	-0.010	-0.013	3.065	3.020	3.006	2.966	0.192	0.076	0.068	0.048	0.644	0.877	0.952	0.977	9.886	5.412	3.200	2.330
X-S	-0.071	-0.027	-0.009	-0.014	3.185	3.066	3.038	2.992	0.426	0.112	0.073	0.057	0.337	0.734	0.909	0.963	14.543	9.199	4.845	2.973
X-C	-0.103	-0.029	-0.017	-0.010	3.089	2.962	2.996	2.959	0.374	0.049	0.052	0.042	0.496	0.850	0.945	0.975	31.636	11.894	5.217	3.047
DR-F	-0.058	-0.059	-0.034	-0.011	3.848	3.745	3.319	3.054	0.798	0.443	0.200	0.073	0.242	0.717	0.894	0.949	4.993	4.567	3.159	2.400
DR-S	-0.067	-0.115	-0.101	-0.076	5.197	5.478	4.775	3.768	0.999	0.933	0.587	0.267	0.061	0.324	0.742	0.892	1.351	1.832	1.821	1.578
DR-C	-0.026	-0.054	-0.059	-0.047	3.683	3.762	3.578	3.234	0.958	0.710	0.371	0.146	0.107	0.498	0.870	0.948	3.482	4.144	2.451	1.768
R-F	-0.018	-0.007	-0.002	0.001	3.215	3.280	3.152	3.031	0.443	0.342	0.165	0.067	0.251	0.524	0.712	0.825	2.355	2.844	2.901	2.686
R-S	0.049	0.024	0.010	0.003	3.182	3.217	3.212	3.106	0.412	0.314	0.211	0.110	0.105	0.308	0.564	0.739	1.873	2.403	2.835	2.879
R-C	-0.028	0.000	0.002	-0.004	3.053	3.076	3.060	3.016	0.110	0.118	0.098	0.066	0.174	0.476	0.731	0.846	5.354	5.714	4.761	3.722

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{BIAS}$, $\overline{SD}$, $\overline{SKEW}$, and $\overline{KURT}$ show the mean values of the root mean squared error, absolute bias, bias, standard deviation, skewness and kurtosis of all 10′000 CATE estimates from the validation sample. $SE(\overline{RMSE})$ depicts the standard error of the average RMSE and $JB\%$ presents the share of CATEs for which the Jarque-Bera test has been rejected at the 5% level. The results for $CORR$ and $VARR$ show the values of the correlation and variance ratio between the true and the estimated CATEs over all replications. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

1.B.2.5 Simulation 5: unbalanced treatment and linear CATE

Table 1.B.10: CATE Results for Simulation 5

	$\overline{RMSE}$				$SE(\overline{RMSE})$				$\overline{\lvert BIAS\rvert}$				$\overline{BIAS}$				$\overline{SD}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.823	0.606	0.461	0.358	0.069	0.043	0.031	0.028	0.817	0.599	0.454	0.351	0.817	0.599	0.454	0.351	0.101	0.087	0.075	0.066
S-W	0.305	0.244	0.196	0.164	0.046	0.029	0.021	0.017	0.255	0.209	0.170	0.145	-0.076	-0.067	-0.054	-0.044	0.222	0.159	0.117	0.089
T	0.305	0.244	0.196	0.164	0.046	0.029	0.021	0.017	0.255	0.209	0.171	0.145	-0.076	-0.068	-0.055	-0.044	0.222	0.159	0.117	0.089
X-F	0.237	0.178	0.137	0.109	0.044	0.026	0.017	0.014	0.200	0.154	0.120	0.097	-0.062	-0.052	-0.038	-0.028	0.164	0.110	0.078	0.058
X-S	0.276	0.210	0.163	0.126	0.068	0.032	0.021	0.015	0.230	0.181	0.143	0.112	-0.074	-0.065	-0.050	-0.034	0.202	0.130	0.092	0.067
X-C	0.231	0.182	0.144	0.114	0.049	0.030	0.022	0.017	0.200	0.165	0.132	0.105	-0.078	-0.067	-0.048	-0.034	0.139	0.086	0.061	0.046
DR-F	0.314	0.248	0.203	0.166	0.041	0.025	0.023	0.020	0.258	0.212	0.179	0.148	-0.064	-0.054	-0.038	-0.027	0.237	0.159	0.112	0.084
DR-S	0.556	0.413	0.298	0.215	0.136	0.053	0.024	0.016	0.428	0.322	0.239	0.176	-0.053	-0.051	-0.034	-0.014	0.515	0.362	0.242	0.167
DR-C	0.354	0.280	0.217	0.162	0.054	0.024	0.019	0.016	0.287	0.232	0.185	0.140	-0.068	-0.053	-0.030	-0.010	0.289	0.205	0.140	0.099
R-F	0.392	0.312	0.254	0.223	0.038	0.023	0.020	0.019	0.314	0.253	0.211	0.190	-0.021	-0.016	-0.009	-0.003	0.339	0.253	0.187	0.146
R-S	0.500	0.388	0.305	0.248	0.105	0.042	0.024	0.020	0.398	0.311	0.248	0.206	-0.023	-0.024	-0.014	-0.010	0.457	0.336	0.246	0.179
R-C	0.311	0.262	0.222	0.194	0.037	0.022	0.021	0.023	0.256	0.219	0.191	0.172	-0.028	-0.021	-0.013	-0.007	0.241	0.185	0.139	0.104

	$\overline{SKEW}$				$\overline{KURT}$				$JB\%$				$CORR$				$VARR$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.004	0.030	0.027	0.026	2.981	2.998	2.991	2.984	0.034	0.062	0.054	0.050	0.211	0.415	0.561	0.672	6.435	4.582	4.002	3.502
S-W	-0.024	-0.005	-0.007	-0.012	3.045	3.023	3.000	2.988	0.087	0.067	0.054	0.048	-0.035	0.115	0.332	0.524	1.094	1.571	2.232	2.654
T	-0.024	-0.005	-0.006	-0.008	3.045	3.025	2.996	2.981	0.091	0.065	0.053	0.047	-0.034	0.115	0.331	0.524	1.093	1.573	2.230	2.652
X-F	-0.056	-0.015	-0.008	-0.012	3.061	3.020	3.015	2.987	0.163	0.067	0.059	0.051	0.202	0.481	0.728	0.849	2.966	3.413	3.225	2.767
X-S	-0.063	-0.025	-0.013	-0.016	3.202	3.087	3.032	3.009	0.442	0.120	0.075	0.063	0.093	0.270	0.563	0.780	3.071	3.496	3.862	3.255
X-C	-0.081	-0.026	-0.011	-0.021	3.101	2.960	2.999	2.983	0.260	0.047	0.051	0.054	0.154	0.400	0.713	0.861	7.940	7.788	6.045	3.922
DR-F	-0.064	-0.072	-0.047	-0.026	3.874	4.167	3.685	3.167	0.803	0.500	0.250	0.101	-0.038	0.013	0.159	0.442	0.965	1.702	2.879	3.695
DR-S	-0.073	-0.121	-0.132	-0.113	5.095	5.798	5.324	4.050	0.998	0.933	0.601	0.285	-0.012	0.016	0.101	0.345	0.251	0.361	0.691	1.205
DR-C	-0.029	-0.062	-0.077	-0.061	3.678	3.871	3.805	3.335	0.957	0.750	0.407	0.167	-0.012	0.030	0.167	0.515	0.664	1.013	1.885	2.627
R-F	-0.028	-0.017	-0.002	-0.007	3.282	3.454	3.273	3.069	0.563	0.445	0.203	0.082	0.004	0.055	0.114	0.150	0.400	0.631	1.037	1.510
R-S	0.043	0.020	0.004	-0.010	3.234	3.309	3.373	3.229	0.525	0.438	0.302	0.154	-0.014	0.016	0.063	0.124	0.310	0.419	0.680	1.134
R-C	-0.029	-0.002	-0.002	-0.001	3.064	3.117	3.124	3.055	0.129	0.165	0.146	0.081	-0.027	0.024	0.102	0.193	0.923	1.223	1.840	2.659

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{BIAS}$, $\overline{SD}$, $\overline{SKEW}$, and $\overline{KURT}$ show the mean values of the root mean squared error, absolute bias, bias, standard deviation, skewness and kurtosis of all 10′000 CATE estimates from the validation sample. $SE(\overline{RMSE})$ depicts the standard error of the average RMSE and $JB\%$ presents the share of CATEs for which the Jarque-Bera test has been rejected at the 5% level. The results for $CORR$ and $VARR$ show the values of the correlation and variance ratio between the true and the estimated CATEs over all replications. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

1.B.2.6 Main Simulation: unbalanced treatment and nonlinear CATE

Table 1.B.11: CATE Results for Main Simulation

	$\overline{RMSE}$				$SE(\overline{RMSE})$				$\overline{\lvert BIAS\rvert}$				$\overline{BIAS}$				$\overline{SD}$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.878	0.749	0.651	0.570	0.203	0.169	0.142	0.121	0.867	0.739	0.641	0.560	0.578	0.413	0.305	0.229	0.108	0.096	0.091	0.088
S-W	0.765	0.634	0.533	0.462	0.123	0.107	0.093	0.082	0.717	0.602	0.508	0.443	-0.135	-0.121	-0.099	-0.081	0.261	0.190	0.149	0.125
T	0.766	0.634	0.533	0.462	0.123	0.107	0.093	0.081	0.719	0.602	0.509	0.442	-0.139	-0.121	-0.099	-0.081	0.260	0.190	0.149	0.125
X-F	0.743	0.618	0.517	0.442	0.128	0.111	0.095	0.082	0.711	0.597	0.500	0.427	-0.124	-0.102	-0.077	-0.060	0.200	0.141	0.117	0.103
X-S	0.820	0.707	0.591	0.499	0.137	0.127	0.109	0.093	0.779	0.684	0.574	0.484	-0.147	-0.123	-0.096	-0.073	0.244	0.164	0.125	0.107
X-C	0.794	0.693	0.582	0.494	0.144	0.132	0.112	0.095	0.770	0.680	0.571	0.482	-0.151	-0.126	-0.095	-0.072	0.171	0.114	0.097	0.092
DR-F	0.817	0.659	0.542	0.463	0.126	0.112	0.097	0.085	0.764	0.627	0.518	0.443	-0.116	-0.095	-0.067	-0.049	0.285	0.194	0.149	0.126
DR-S	1.053	0.825	0.579	0.445	0.133	0.097	0.076	0.064	0.906	0.731	0.521	0.403	-0.102	-0.085	-0.053	-0.021	0.640	0.433	0.281	0.206
DR-C	0.880	0.727	0.523	0.409	0.118	0.112	0.088	0.072	0.809	0.680	0.490	0.383	-0.118	-0.088	-0.049	-0.017	0.359	0.255	0.179	0.143
R-F	0.815	0.679	0.590	0.529	0.112	0.101	0.095	0.090	0.746	0.632	0.554	0.499	-0.115	-0.100	-0.081	-0.066	0.346	0.251	0.201	0.172
R-S	0.932	0.788	0.659	0.580	0.120	0.110	0.100	0.095	0.833	0.721	0.613	0.546	-0.126	-0.117	-0.095	-0.081	0.468	0.333	0.243	0.195
R-C	0.825	0.725	0.621	0.554	0.130	0.123	0.110	0.102	0.779	0.694	0.597	0.533	-0.131	-0.115	-0.094	-0.077	0.261	0.196	0.155	0.136
	$\overline{SKEW}$				$\overline{KURT}$				$JB\%$				$CORR$				$VARR$			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	0.115	0.071	0.047	0.024	3.006	2.966	2.990	2.968	0.466	0.115	0.062	0.045	0.624	0.831	0.904	0.934	92.890	27.182	12.760	7.598
S-W	-0.024	-0.016	-0.011	-0.026	3.004	2.999	2.988	2.967	0.055	0.056	0.054	0.044	0.524	0.798	0.891	0.922	13.575	8.180	5.033	3.633
T	-0.035	-0.017	-0.014	-0.027	3.035	2.996	2.986	2.967	0.097	0.055	0.051	0.044	0.514	0.798	0.891	0.922	13.471	8.176	5.034	3.626
X-F	-0.060	-0.030	-0.017	-0.023	3.054	2.984	2.985	2.950	0.171	0.064	0.049	0.040	0.664	0.894	0.950	0.967	24.558	11.070	6.021	4.070
X-S	-0.067	-0.030	-0.028	-0.019	3.143	3.047	3.002	2.964	0.323	0.106	0.060	0.045	0.367	0.754	0.919	0.957	38.577	21.499	9.669	5.561
X-C	-0.079	-0.042	-0.021	-0.021	3.064	2.953	2.987	2.944	0.209	0.066	0.049	0.037	0.530	0.852	0.946	0.966	79.900	27.087	10.181	5.644
DR-F	-0.106	-0.073	-0.034	-0.022	3.915	3.381	3.081	2.982	0.827	0.366	0.119	0.052	0.317	0.770	0.912	0.948	13.371	10.505	6.196	4.273
DR-S	-0.143	-0.216	-0.148	-0.084	5.350	5.320	4.033	3.317	1.000	0.947	0.526	0.189	0.095	0.406	0.812	0.918	3.696	4.807	3.992	2.940
DR-C	-0.080	-0.111	-0.075	-0.044	3.678	3.629	3.243	3.034	0.960	0.668	0.243	0.085	0.162	0.580	0.899	0.950	9.167	9.660	4.855	3.102
R-F	-0.009	-0.006	-0.013	-0.018	3.126	3.077	3.012	2.982	0.233	0.128	0.063	0.049	0.368	0.692	0.832	0.890	7.963	7.751	6.147	4.980
R-S	0.031	0.018	0.002	-0.009	3.107	3.097	3.048	2.992	0.207	0.151	0.082	0.054	0.166	0.449	0.732	0.846	6.605	7.982	7.305	6.036
R-C	-0.021	-0.006	-0.014	-0.020	3.043	3.018	3.003	2.966	0.088	0.063	0.052	0.042	0.271	0.624	0.843	0.902	17.941	15.725	9.647	6.705

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{BIAS}$, $\overline{SD}$, $\overline{SKEW}$, and $\overline{KURT}$ show the mean values of the root mean squared error, absolute bias, bias, standard deviation, skewness and kurtosis of all $10'000$ CATE estimates from the validation sample. $SE(\overline{RMSE})$ depicts the standard error of the average RMSE and $JB\%$ presents the share of CATEs for which the Jarque-Bera test has been rejected at the 5% level. The results for $CORR$ and $VARR$ show the values of the correlation and variance ratio between the true and the estimated CATEs over all replications. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

1.B.2.7 Empirical Simulation

Table 1.B.12: CATE Results for Empirical Simulation

	$\overline{RMSE}$			$SE(\overline{RMSE})$			$\overline{\lvert BIAS\rvert}$			$\overline{BIAS}$			$\overline{SD}$		
	500	*2000*	*8000*	*500*	*2000*	*8000*	*500*	*2000*	*8000*	*500*	*2000*	*8000*	*500*	*2000*	*8000*
S	0.175	0.127	0.093	0.025	0.015	0.009	0.171	0.121	0.090	0.171	0.119	0.085	0.035	0.035	0.023
S-W	0.131	0.109	0.078	0.031	0.014	0.011	0.106	0.090	0.070	-0.011	-0.050	-0.043	0.121	0.084	0.037
T	0.150	0.111	0.079	0.027	0.014	0.011	0.122	0.092	0.071	-0.063	-0.053	-0.044	0.127	0.084	0.037
X-F	0.112	0.082	0.056	0.029	0.014	0.011	0.092	0.069	0.052	-0.056	-0.045	-0.036	0.089	0.056	0.021
X-S	0.129	0.093	0.069	0.041	0.019	0.011	0.105	0.078	0.060	-0.065	-0.054	-0.043	0.104	0.067	0.040
X-C	0.103	0.077	0.055	0.035	0.018	0.013	0.087	0.067	0.052	-0.065	-0.054	-0.043	0.072	0.044	0.017
DR-F	0.147	0.105	0.070	0.026	0.014	0.010	0.119	0.087	0.063	-0.061	-0.051	-0.042	0.125	0.078	0.033
DR-S	0.256	0.180	0.123	0.055	0.023	0.011	0.201	0.143	0.101	-0.066	-0.056	-0.047	0.242	0.162	0.097
DR-C	0.159	0.116	0.078	0.031	0.015	0.011	0.128	0.096	0.071	-0.068	-0.057	-0.046	0.135	0.088	0.037
R-F	0.183	0.131	0.089	0.022	0.011	0.009	0.146	0.107	0.078	-0.051	-0.043	-0.034	0.167	0.109	0.051
R-S	0.237	0.174	0.123	0.046	0.021	0.011	0.189	0.140	0.100	-0.058	-0.049	-0.042	0.224	0.158	0.099
R-C	0.144	0.109	0.076	0.026	0.013	0.010	0.117	0.091	0.068	-0.058	-0.050	-0.040	0.123	0.084	0.037
	$\overline{SKEW}$			$\overline{KURT}$			$JB\%$			$CORR$			$VARR$		
	500	*2000*	*8000*	*500*	*2000*	*8000*	*500*	*2000*	*8000*	*500*	*2000*	*8000*	*500*	*2000*	*8000*
S	0.671	0.178	0.031	3.310	2.998	2.988	1.000	0.511	0.051	0.050	0.135	0.328	10.481	2.315	1.666
S-W	0.394	0.014	0.007	2.838	2.979	2.984	1.000	0.049	0.040	0.061	0.162	0.359	0.562	0.331	0.448
T	0.002	0.007	0.001	3.001	2.986	2.995	0.055	0.056	0.058	0.058	0.158	0.357	0.200	0.323	0.447
X-F	0.007	0.013	0.007	3.006	2.981	2.999	0.053	0.049	0.055	0.103	0.223	0.458	0.542	0.748	0.912
X-S	0.020	0.022	0.007	3.024	3.045	2.993	0.063	0.100	0.045	0.058	0.134	0.297	0.544	0.716	0.981
X-C	0.009	0.011	0.007	2.987	2.964	2.982	0.028	0.027	0.049	0.094	0.197	0.393	1.327	1.527	1.665
DR-F	0.012	0.008	0.001	3.100	3.025	2.986	0.209	0.105	0.054	0.059	0.139	0.372	0.209	0.388	0.645
DR-S	0.013	0.016	0.006	3.678	3.476	3.130	0.906	0.472	0.161	0.031	0.066	0.161	0.065	0.110	0.220
DR-C	0.018	0.014	-0.009	3.158	3.095	3.055	0.318	0.190	0.098	0.053	0.113	0.251	0.185	0.309	0.564
R-F	-0.012	-0.009	-0.006	3.068	3.042	2.990	0.179	0.125	0.047	0.072	0.145	0.306	0.105	0.188	0.320
R-S	-0.002	-0.015	-0.006	3.090	3.083	3.052	0.166	0.154	0.109	0.040	0.080	0.174	0.074	0.116	0.212
R-C	0.002	-0.006	-0.007	3.004	2.987	3.023	0.053	0.060	0.062	0.068	0.139	0.275	0.219	0.333	0.545

Note: The results for the $\overline{RMSE}$, $\overline{|BIAS|}$, $\overline{BIAS}$, $\overline{SD}$, $\overline{SKEW}$, and $\overline{KURT}$ show the mean values of the root mean squared error, absolute bias, bias, standard deviation, skewness and kurtosis of all 1′000 CATE estimates from the validation sample. $SE(\overline{RMSE})$ depicts the standard error of the average RMSE and $JB\%$ presents the share of CATEs for which the Jarque-Bera test has been rejected at the 5% level. The results for $CORR$ and $VARR$ show the values of the correlation and variance ratio between the true and the estimated CATEs over all replications. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

1.C Computation Time

In order to assess the computational trade-offs among different estimation schemes as well as different meta-learners we evaluate the computational time for each meta-learner and each estimation scheme for each sample size over 10 replications of the Main Simulation to illustrate the performance. The results are summarized in Table 1.C.1 and Figure 1.C.1 below.

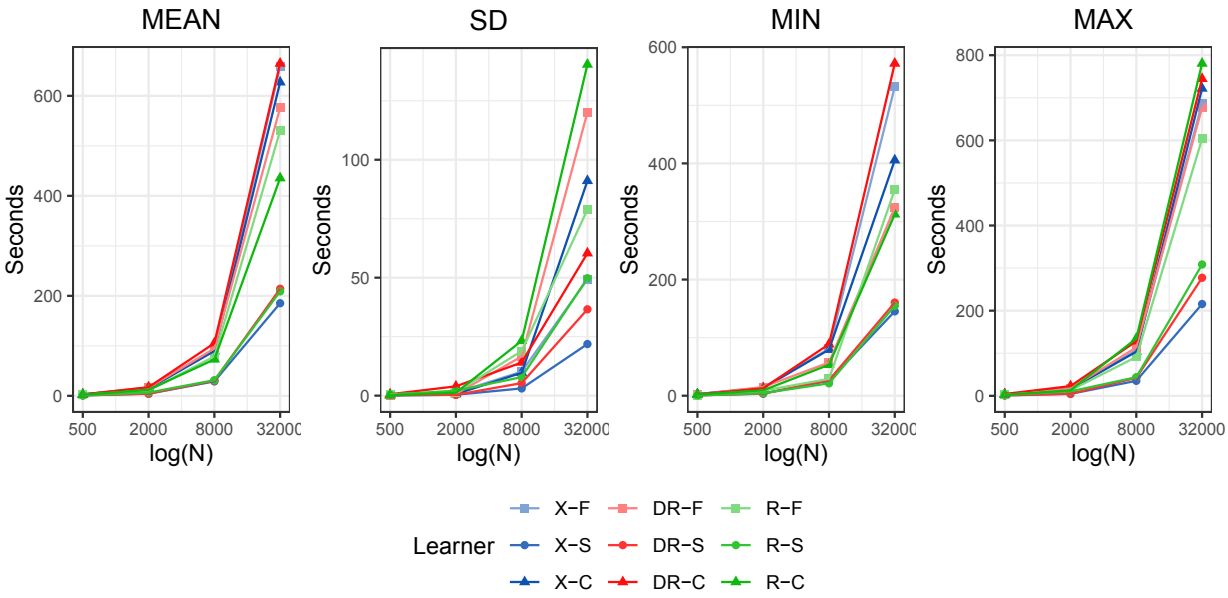
1.C.1 Main Simulation: unbalanced treatment and nonlinear CATE

Table 1.C.1: Computation Time Results for Main Simulation

	MEAN				SD				MIN				MAX			
	500	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*	*500*	*2000*	*8000*	*32000*
S	1.492	8.786	53.385	252.165	0.039	0.991	9.777	0.551	1.440	7.110	43.000	251.050	1.560	9.860	67.790	252.810
S-W	1.416	6.730	39.117	263.195	0.051	1.015	5.804	4.636	1.330	4.890	31.920	250.630	1.500	8.050	49.950	267.140
T	1.203	6.168	38.933	238.932	0.043	0.880	5.982	26.988	1.110	4.460	29.530	162.260	1.260	7.540	47.340	249.560
X-F	2.894	16.512	92.803	658.892	0.081	1.303	10.148	49.316	2.770	14.070	80.950	531.960	2.980	17.590	106.570	687.320
X-S	0.915	4.233	28.863	185.232	0.074	0.380	3.061	21.873	0.760	3.500	25.830	145.250	1.000	4.640	35.250	215.630
X-C	3.027	14.353	89.644	627.236	0.265	0.657	9.687	91.057	2.790	13.300	79.280	405.540	3.470	15.720	103.810	721.920
DR-F	2.262	15.998	94.615	576.230	0.129	0.696	16.618	120.102	2.080	14.920	57.130	323.300	2.490	17.040	113.820	676.650
DR-S	0.836	4.292	30.218	214.072	0.189	0.345	5.321	36.595	0.580	3.780	26.640	160.380	1.200	4.940	42.280	277.320
DR-C	2.684	17.272	105.261	664.728	0.596	3.941	14.058	60.375	2.150	12.770	88.690	572.030	4.300	22.850	128.400	744.670
R-F	2.058	10.588	78.830	530.529	0.594	1.430	18.923	78.864	0.840	8.750	28.900	354.890	2.450	13.020	91.270	603.520
R-S	0.919	6.514	31.234	208.910	0.429	2.084	7.845	49.782	0.530	4.420	21.340	154.770	2.100	10.440	43.750	308.480
R-C	2.177	11.934	72.912	435.450	0.080	1.230	23.239	140.374	2.020	9.240	53.290	312.420	2.250	12.970	134.670	780.560

Note: The results for the MEAN, SD, MIN, and MAX show the values of the mean, standard deviation, minimum and maximum for the computation time in seconds based on 10 simulation replications. The computation time includes both the estimation as well as the prediction task. No multithreading used within the estimation of meta-learners. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

Figure 1.C.1: Computation Time Results for Main Simulation



Note: The results for the MEAN, SD, MIN, and MAX show the values of the mean, standard deviation, minimum and maximum for the computation time in seconds based on 10 simulation replications. The figure shows the results based on the increasing training samples of $\{500, 2'000, 8'000, 32'000\}$ observations displayed on the log scale. Additionally, X-F, DR-F, R-F denote the full-sample versions of the meta-learners, while X-S, DR-S, R-S and X-C, DR-C, R-C denote the sample-splitting and cross-fitting versions, respectively.

Chapter 2

Random Forest Estimation of the Ordered Choice Model

Co-author: Michael Lechner

Abstract

In this paper we develop a new machine learning estimator for ordered choice models based on the random forest. The proposed *Ordered Forest* flexibly estimates the conditional choice probabilities while taking the ordering information explicitly into account. In addition to common machine learning estimators, it enables the estimation of marginal effects as well as conducting inference and thus provides the same output as classical econometric estimators. An extensive simulation study reveals a good predictive performance, particularly in settings with non-linearities and near-multicollinearity. An empirical application contrasts the estimation of marginal effects and their standard errors with an ordered logit model.

Keywords: Ordered choice models, random forests, probabilities, marginal effects, machine learning.

JEL classification: C14, C25, C40.

2.1 Introduction

Many empirical models deal with categorical dependent variables which have an inherent ordering. In such cases the outcome variable is measured on an ordered scale such as level of education defined by primary, secondary and tertiary education or income coded into low, middle and high income level. Further examples include survey outcomes on self-assessed health status (bad, good, very good, see e.g. Case, Lubotsky, & Paxson, 2002; or Murasko, 2008), level of life satisfaction and happiness (Boes, Staub, & Winkelmann, 2010; and Boes & Winkelmann, 2010) or political opinions (do not agree, agree, strongly agree, see e.g. Jackson & Darrow, 2005; or Jackman, 2009) as well as grades, scores and various ratings and valuations (see Butler, Finegan, & Siegfried, 1998; Hamermesh & Parker, 2005; Afonso, Gomes, & Rother, 2009; or Gogas, Papadimitriou, & Agrapetidou, 2014, for some further examples). Moreover, even sports outcomes resulting in loss, draw and win are part of such modelling framework (e.g. Goller, Knaus, Lechner, & Okasa, 2018). So far, the ordered probit or ordered logit model represent workhorse models in such cases. The main advantage of these models is the ease of estimation, usually done by maximum likelihood. However, the major disadvantage are the strong parametric assumptions which are imposed for convenience rather than derived from any substantive knowledge about the application. Unfortunately, the desired marginal effects are sensitive to these assumptions. Although there is a large literature on how to generalize these assumptions in case of binary choice models (Matzkin, 1992; Ichimura, 1993; Klein & Spady, 1993), or multinomial (unordered) choice models (Lee, 1995; Fox, 2007), limited work has been done for ordered choice models (Lewbel, 2000; Klein & Sherman, 2002; also see Stewart, 2005, for an overview).

In this paper, we exploit recent advances in the machine learning literature to develop an estimator for conditional choice probabilities as well as marginal effects together with inference procedures when the outcome variable has an ordered categorical nature. The proposed *Ordered Forest* estimator is based on the regression random forest algorithm as introduced by Breiman (2001) and makes use of cumulative probability predictions based on binary indicators of respective ordered categories to flexibly estimate the single choice probabilities of the particular ordered category, conditional on covariates. Furthermore, to analyze the relationship of the ordered choice probabilities with the covariates, the *Ordered Forest* exploits numerical derivative approximations for estimation of the mean marginal effects and marginal effects at mean as the typical quantities of interest in the field of discrete choice models (see e.g. Greene & Hensher, 2010). Finally, in order to quantify the estimation uncertainty of the above parameters, the *Ordered Forest* adapts the weight-based inference proposed by Lechner (2018) using the asymptotic results of Wager and Athey (2018) for the consistency and normality of random forest predictions for the case of ordered categorical outcomes. Thus *Ordered Forest* estimator provides not only the point estimate for the conditional choice probabilities and the corresponding marginal effects, but also an estimate for the respective standard errors.

We investigate the predictive performance of the estimator by comparing it to classical and other competing methods via a large-scale Monte Carlo simulation study as well as using real datasets. The results from the synthetic simulation reveal good performance of the *Ordered Forest* in finite samples throughout all simulation designs, including high-dimensional settings. In particular, the superior performance of the estimator over the parametric ordered logit becomes apparent when dealing with nonlinear functional forms and near-multicollinearity among covariates. Furthermore, the *Ordered Forest* outperforms the competing forest-based estimators in the most complex simulation designs. Additionally, the results from the empirical evaluation further confirm the good predictive performance of the estimator in real datasets. Lastly, an empirical application demonstrates the estimation of the marginal effects and the associated inference procedure. The empirical results highlight the value of the additional flexibility

in the effect estimation of relevant economic parameters. Moreover, to enable the usage of the method by applied researchers a free software implementation of the *Ordered Forest* estimator has been developed in R (R Core Team, 2018) and is provided in an R-package orf (Lechner & Okasa, 2019) available at the official CRAN repository.[1]

This paper contributes to the econometric as well as machine learning literature in several ways. In terms of econometrics, this paper develops a new estimator of the ordered choice models based on a machine learning algorithm. The proposed *Ordered Forest* estimator improves on the classical parametric models such as ordered logit and ordered probit models by allowing *ex-ante* flexible functional forms as well as allowing for a larger covariate space. The latter is a feature of many machine learning methods, but is typically absent from standard econometrics. In terms of machine learning, this paper develops a new type of random forest estimator adapted to ordered categorical outcomes. As such, the proposed *Ordered Forest* extends the classical regression forests as developed by Breiman (2001) and Wager and Athey (2018) specifically for estimation of ordered choice models and thus expands the forest-based estimators for particular econometric models such as for example the survival forest (Hothorn, Lausen, Benner, & Radespiel-Tröger, 2004) designed for estimation of survival models or the quantile regression forest (Meinshausen, 2006) for estimation of conditional quantiles. Additionally to the above forest-based estimators, the *Ordered Forest* further advances machine learning methods with the estimation of marginal effects and the inference thereof, a feature of many parametric models, but generally missing in the machine learning literature. Hence, our contribution is twofold. First, with respect to the literature on parametric estimation of the ordered choice models, the *Ordered Forest* represents a flexible estimator without any parametric assumptions, while providing essentially the same information as an ordered parametric model. Second, with respect to the machine learning literature, the *Ordered Forest* achieves more precise estimation of ordered choice probabilities, while adding estimation of marginal effects as well as statistical inference thereof.

This paper is organized as follows. Section 2.2 discusses the related literature concerning parametric and machine learning methods for the estimation of ordered choice models. Section 2.3 reviews the random forest algorithm and its theoretical properties. In Section 2.4 the *Ordered Forest* estimator is introduced including the estimation of the conditional choice probabilities, marginal effects and the inference procedure. The Monte Carlo simulation is presented in Section 2.5. Section 2.6 shows an empirical application. Section 2.7 concludes. Further details regarding estimation methods, the simulation study and the empirical application are provided in Appendices 2.A, 2.B and 2.C, respectively.

2.2 Literature

In econometrics, the ordered probit and ordered logit models are widely used when there are ordered response variables (McCullagh, 1980). These models build on the latent regression model assuming an underlying continuous outcome Y_i^* as a linear function of regressors X_i with unknown coefficients β, while assuming that the latent error term u_i follows the standard normal or the logistic distribution. Furthermore, the ordered discrete outcome Y_i represents categories that cover a certain range of the latent continuous Y_i^* and is determined by unknown threshold parameters α_m. Formally, in the case of the ordered logit the latent model is defined as:

$$Y_i^* = X_i'\beta + u_i, \qquad (u_i \mid X_i) \sim Logistic(0, \pi^2/3) \tag{2.2.1}$$

[1] Additionally, an implementation of the estimator in GAUSS is available online and on *ResearchGate*. A Python version of the estimator focused on prediction exercise is available on GitHub.

with unknown threshold parameters $\alpha_0 < \alpha_1 < ... < \alpha_M$ such that:

$$Y_i = m \quad \text{if} \quad \alpha_{m-1} < Y_i^* \leq \alpha_m \quad \text{for} \quad m = 1, ..., M, \tag{2.2.2}$$

where the coefficients and the thresholds are commonly estimated via maximum likelihood with the delta method or bootstrapping used for inference. Notice, that the outer thresholds are $\alpha_0 = -\infty$ and $\alpha_M = \infty$. The above latent model is also often motivated by the quantity of interest, i.e. the conditional choice probabilities which are given by:

$$P[Y_i = m \mid X_i = x] = \Lambda\big(\alpha_m - X_i'\beta\big) - \Lambda\big(\alpha_{m-1} - X_i'\beta\big), \tag{2.2.3}$$

where the link function $\Lambda(\cdot)$ is the logistic cdf mapping the real line onto the unit interval. Thus, the estimated probabilities are bounded between 0 and 1. The marginal effects are further given as partial derivative of the probabilities in (2.2.3) as:

$$\frac{\partial P[Y_i = m \mid X_i = x]}{\partial x^k} = \Big[\lambda\big(\alpha_{m-1} - X_i'\beta\big) - \lambda\big(\alpha_m - X_i'\beta\big)\Big]\beta_k, \tag{2.2.4}$$

where x^k is the k-th element of X_i and β_k is the corresponding coefficient, while $\lambda(\cdot)$ being the logistic pdf.

Although such models are relatively easy to estimate, they impose strong parametric assumptions which hinder the flexibility of these models. Apart from the assumptions about the distribution of the error term, further functional form assumptions are being imposed. As is clear from (2.2.1), the coefficients β are constant across the outcome classes which is often labelled as the parallel regression assumption (Williams, 2016). This inflexibility affects both the estimation of the choice probabilities as well as the estimation of marginal effects. For these reasons, generalizations of these models have been proposed in the literature in order to relax some of the assumptions. An example of such models is the generalized ordered logit model (McCullagh & Nelder, 1989), where the parallel regression assumption is abandoned. Boes and Winkelmann (2006) provide an excellent overview of several other generalized parametric models. However, all of these models retain some of the distributional assumptions which limit their modelling flexibility.

Besides the standard econometric literature on parametric specifications of ordered choice models (for an overview see Agresti, 2002; or Boes & Winkelmann, 2006), a new strand of literature devoted to relaxing the parametric assumptions by using novel machine learning methods is emerging. Particularly, the tree-based methods have gained considerable attention. Although the classical CART algorithms introduced by Breiman, Friedman, Olshen, and Stone (1984) are very powerful in both regression as well as in classification (see Loh, 2011, for a review), there is a need for adjustment when predicting ordered response. In the case of regression, the *discrete* nature of the outcome is not being taken into account and in the case of classification, the *ordered* nature of the outcome is not being taken into account. For these reasons, a strand of the literature focused particularly on adjustments towards ordered classification rather than regression which excludes the estimation of the conditional probabilities as is the case in the parametric ordered choice models. For example, Kramer, Widmer, Pfahringer, and De Groeve (2001) propose a simple procedure for constructing a distance-sensitive classification learner using post-processing classification rules. Another approach suggested in the literature is to modify the splitting criterion directly. In particular, the usage of alternative impurity measures as opposed to the Gini coefficient in case of classification trees have been suggested, namely the generalized Gini criterion (Breiman et al., 1984) or the ordinal impurity function (Piccarreta, 2008). Both of these measures put higher penalty on misclassification the more distant the predicted category is from the true one. It follows that the above

methods focus on estimating ordered classes rather than estimating ordered class probabilities, as is the focus of this paper.

The above ideas, however, have not been much used in practice. The reason might be the well-known drawbacks of single trees which suffer from unstable splits and a lack of smoothness (Hastie, Tibshirani, & Friedman, 2009). A natural extension of the CART algorithms is the random forest first introduced by Breiman (2001). However, the random forest algorithm as well as CART is primarily suitable for either regression or classification exercises. As such, appropriate modifications of the standard random forest algorithm are desired in order to predict conditional probabilities of discrete outcomes while taking the ordering nature into account. Hothorn, Hornik, and Zeileis (2006b) propose a random forest algorithm building on their conditional inference framework for recursive partitioning which can also deal with ordered outcomes. The difference to standard regression forests lies in a different splitting criterion using a test statistic where the conditional distribution at each split is based on permutation tests (for details see Strasser & Weber, 1999; and Hothorn et al., 2006b). Their proposed ordinal forest regression assumes an underlying latent continuous response Y_i^* as is the case in standard ordered choice models. Hothorn et al. (2006b) define a score vector $s(m) \in \mathbb{R}^M$, with $m = 1, ..., M$ observed ordered classes. This scores reflect the distances between the classes. The authors suggest to set the scores as midpoints of the intervals of Y_i^* which define the classes. As the underlying Y_i^* is unobserved, such a suggestion results in $s(m) = m$ and ordinal forest regression collapses to a standard forest regression as pointed out by Janitza, Tutz, and Boulesteix (2016).[2] However, although the tree building step coincides, the prediction step differs as the estimates are the choice probabilities calculated as the proportions of the respective outcome classes falling into the same leaf instead of averages of the outcomes. As such, for each leaf within a tree, the prediction is computed for each value of the ordered categorical outcome as its share within the leaf, resulting in a probability predictions between 0 and 1. This is in contrast to standard prediction procedures, which would compute an average of all values of the ordered categorical outcome. Nevertheless, after computing the single-tree predictions as the relative frequencies of the ordered outcomes, the forest estimates of the conditional choice probabilities $\hat{P}[Y_i = m \mid X_i = x]$ are computed by taking the averages of the choice probabilities produced by each tree, i.e. the same aggregation scheme as in a regression forest. Hornung (2019a) points out that setting $s(m) = m$ implies inherently assuming that the class widths, i.e. the adjacent intervals of the continuous outcome variable Y_i^* determining the descrete outcome Y_i are of the same length. This, however, does not have to hold in general and these intervals might not follow any particular pattern.[3] In order to address this issue, Hornung (2019a) proposes an ordinal forest method, which optimizes these interval widths by maximizing the out-of-bag (OOB) prediction performance of the forests.[4] However, on the contrary to the approach of Hothorn et al. (2006b), the forest algorithm used is based on the forest as developed by Breiman (2001), while the primary target is to predict the ordinal class and the choice probabilities are obtained as relative frequencies of trees predicting the particular class. As such, each tree predicts the most probable value of the ordered categorical outcome. Thereupon, the forest prediction for the conditional choice probability is computed as the share of trees predicting the particular categorical value of the ordered outcome. This is in contrast to the estimation scheme by Hothorn et al. (2006b), where the probability prediction step occurs at the level of trees, instead of at the level of forest as is the case here. Hornung (2019a) shows better prediction performance of such ordinal forests which optimize the class widths of Y_i^* in comparison to the conditional forests. Without the optimization step, the author denotes such forest as the naive

[2] Janitza et al. (2016) perform also a simulation study to test the robustness of the suggested score values by setting $s(m) = m^2$, but do not find any significant differences to simple $s(m) = m$.

[3] Recently, Buri and Hothorn (2020) and Tutz (2021) proposed score-free methods based on random forests that do not rely on the underlying continuous intervals of the observed ordered classes.

[4] This approach could be regarded as semiparametric as it uses the nonparametric structure of the trees and assumes a particular parametric distribution (standard normal) within its optimization procedure.

ordinal forest.[5]

While both of the discussed approaches take the ordering information of the outcomes into account, they focus mainly on prediction and variable importance without considering estimation of the marginal effects or the associated inference for the effects which are a fundamental part of the classical econometric ordered choice models. In addition, although both of these methods demonstrate good predictive performance, none of them provides theoretical guarantees with regards to the distribution of these predictions. Further, it is worth to mention that in practice both methods suffer from considerable computational costs. In case of the conditional forest, the additional permutation tests that need to be performed to evaluate the test statistic at each split result in a considerably longer computation time. For the ordinal forest, the additional optimization step for the class widths requires a prior estimation of a large number of forests (1000 by default) which also leads to a substantially longer computation time (see Tables 2.B.26 and 2.B.27 in Appendix 2.B.4 for further details).

There is also a strand of literature which is concerned with the estimation of ordered outcome models in high-dimensional settings based on regularization methods. Examples of this approach include penalized ordered outcome models by Wurm, Rathouz, and Hanlon (2017) who make use of a standard ordered logit/probit regression while introducing an elastic net penalization term. Harrell (2015) describes a cumulative logit model with a ridge type of penalty. Archer et al. (2014) implement the GMIFS (generalized monotone incremental forward stagewise) algorithm for penalized ordered outcome models which is similar to the Lasso type penalty. However, although the penalized models can deal with high dimensions, when the true model is relatively "sparse", they nevertheless belong to a specific parametric class such as the ordered logit/probit (see Hastie et al., 2009). Such models can only become more flexible, and thus partially relax the parametric assumptions when generating a large number of polynomials and interactions of available covariates prior to estimation. It follows that these models use a global approximation of the functional form and cannot learn it adaptively as tree-based approaches do. In contrast, random forests do not impose parametric assumptions and can learn any arbitrary relationship in a nonparametric way by locally adaptive estimation in small neighbourhoods of the data. It follows that random forests use a local approximation of the functional form, without any need for prior pre-processing of the data. As such, random forests are nonlinear in covariates and although there are no specific statistical tests to find such a random forest structure, essentially random forests can approximate any structure, including a global linear structure, if a sufficient amount of data is provided. For these reasons, the remainder of this paper focuses on the forest-based methods.

2.3 Random Forests

Random forests as introduced by Breiman (2001) became quickly a very popular prediction method thanks to its good prediction accuracy, while being relatively simple to tune. Further advantages of random forests as a nonparametric technique are the high degree of flexibility and ability to deal with large number of predictors, while coping better with the curse of dimensionality problem in comparison to classical nonparametric methods such as kernel or local linear regression (see for example Racine, 2008). Random forests are based on bootstrap aggregation, i.e. the so-called bagging of single regression (or classification) trees where the covariates considered for each next split within a tree are selected at random. More precisely, the random forest algorithm draws a bootstrap sample $Z_i^*(X_i, Y_i)$ of size N from the available training data for $b = 1, ..., B$ bootstrap replications. For each bootstrapped sample, a random-forest tree $\hat{T}_b$ is grown by recursive partitioning until the minimum leaf size is reached. At each

[5] A more detailed description of the conditional as well as the ordinal forest is provided in Appendix 2.A.2 and 2.A.3, respectively.

of the splits, m out of p covariates chosen at random are considered. After all B trees are grown in this fashion, the regression random forest estimate of the conditional mean $E[Y_i \mid X_i = x]$ is the ensemble of the trees:

$$\hat{RF}^B(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{T}_b(x) \quad \text{with} \quad \hat{T}_b(x) = \frac{1}{\mid \{i : X_i \in L_b(x)\} \mid}\sum_{\{i:X_i \in L_b(x)\}} Y_i, \tag{2.3.1}$$

where $L_b(x)$ denotes a leaf containing x. Single trees, if grown sufficiently deep, have a low bias, but fairly high variance. By averaging over many single trees with randomly choosing the set of observations and split covariates, the variance of the estimator is being reduced substantially. First, the variance reduction is achieved through bagging. The higher the number of bootstrap replications, the lower the variance. Second, the variance is further reduced through the random selection of covariates. The lower is the number of considered covariates for a split, the more is the correlation between the trees reduced and consequently, the bigger is the variance reduction of the average (Hastie et al., 2009).

Another attractive feature of random forests is the weighted average representation of the final estimate of the conditional mean $E[Y_i \mid X_i = x]$. As such we can rewrite the random forest prediction as follows:

$$\hat{RF}^B(x) = \sum_{i=1}^{N}\hat{w}_i(x)Y_i, \tag{2.3.2}$$

where the weights are defined as:

$$\hat{w}_{b,i}(x) = \frac{\mathbf{1}(\{X_i \in L_b(x)\})}{\mid \{i : X_i \in L_b(x)\} \mid} \quad \text{with} \quad \hat{w}_i(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{w}_{b,i}(x). \tag{2.3.3}$$

As such the forest weights $\hat{w}_i(x)$ are again an average over all single tree weights. These tree weights capture if the training example X_i falls into the leaf $L_b(x)$ scaled by the size of that leaf. Notice, that the weights are locally adaptive. Intuitively, random forests resemble the classical nonparametric kernel regression with an adaptive, data-driven bandwidth and with limited curse of dimensionality. One can show that in the regression case, the random forest estimate as defined in (2.3.1) is equivalent to the weighting estimate defined in (2.3.2). This weighting perspective of random forests has been firstly suggested by Hothorn et al. (2004) and Meinshausen (2006) in the scope of survival and quantile regression, respectively. Recently, Athey, Tibshirani, and Wager (2019) point out the usefulness of the random forest weights in various estimation tasks. In this spirit, we will later on in Section 2.4.3 use the forest induced weights explicitly for inference as has been recently suggested by Lechner (2018).

Besides the huge popularity of random forests for prediction, the statistical literature focused on establishing asymptotic properties of random forests as well (Meinshausen, 2006; Biau, 2012; Scornet, Biau, & Vert, 2015; Mentch & Hooker, 2016). A major step towards formally valid inference has been done in a recent work by Wager (2014) and Wager and Athey (2018) who prove consistency and asymptotic normality of random forest predictions, under some modifications of the standard random forest algorithm. These modifications concern both the tree-building procedure as well as the tree-aggregation scheme. First, the tree aggregation is now done using subsampling without replacement instead of bootstrapping. Second, the tree building procedure introduces the major and crucial condition of so-called honesty as first suggested by Athey and Imbens (2016). A tree is honest, if it does not use the same responses for both, placing splits and estimating the within-leaf predictions. This can be achieved by the so-called double-sample trees, which split the random subsample of training data $Z_i^*(X_i, Y_i)$ into two disjoint sets of the same size, while the one is used for placing splits and the other one for estimating

the predictions. Furthermore, for the consistency it is essential that the size of the leaves L of the trees becomes small relative to the sample size as N gets large.[6] This is achieved by introducing some randomness in choosing the splitting variables. Particularly, each covariate receives a minimum amount of positive chance of a split. Such constructed tree is then said to be a random-split tree. Additionally, the trees are required to be α-regular, meaning that after each split, both of the child nodes contain at least a fraction α of the training data (specifically, $\alpha \leq 0.2$ is required). Lastly, trees have to be symmetric in a sense that the order of the training data is independent of the predictor output. Overall, apart from subsampling and honesty the above conditions are not particularly binding and do not fundamentally deviate from the standard regression random forest. Lastly, some additional regularity conditions need to be satisfied for the asymptotic arguments to hold. In particular, the data $Z_i(X_i, Y_i) \in [0,1]^p \times \mathbb{R}$ comes from *i.i.d.* sampling, the p-dimensional covariates $X_i \sim \mathcal{U}([0,1]^p)$ are independently and uniformly distributed, the conditional means $E[Y_i \mid X_i = x]$ and $E[Y_i^2 \mid X_i = x]$ are Lipschitz-continuous, the variance is bounded away from zero, $Var[Y_i \mid X_i = x] > 0$, and the number of subsampling replications is large enough to eliminate the Monte Carlo effects, while an appropriate scaling of the subsample size s_N is ensured.[7] Then, under the above assumptions, the random forest predictions can be shown to be (pointwise) asymptotically Gaussian and unbiased. We use this result to provide an inference procedure for the marginal effects of the *Ordered Forest* discussed in Section 2.4.3.

2.4 Ordered Forest Estimator

The general idea of the *Ordered Forest* estimator is to provide a flexible alternative for estimation of ordered choice models that can deal with a large-dimensional covariate space. As such, the main goal is the estimation of conditional ordered choice probabilities, i.e. $P[Y_i = m \mid X_i = x]$ as well as marginal effects, i.e. the changes in the estimated probabilities in association with changes in covariates. Correspondingly, the variability of the estimated effects is of interest and therefore a method for conducting statistical inference is provided as well. The latter two features go beyond the traditional machine learning estimators which focus solely on the prediction exercise, and complement the prediction with the same econometric output as the traditional parametric estimators.

2.4.1 Conditional Choice Probabilities

The main idea of the estimation of the ordered choice probabilities by a random forest algorithm lies in the estimation of cumulative, i.e. nested probabilities based on binary indicators. As such, for an *i.i.d* random sample of size $N (i = 1, ..., N)$, consider an ordered outcome variable $Y_i \in \{1, ..., M\}$ with ordered classes m. Then the binary indicators are given as $Y_{m,i} = \mathbf{1}(Y_i \leq m)$ for outcome classes $m = 1, ..., M-1$. First, the ordered model is transformed into multiple overlapping binary models which are estimated by random forests yielding the predictions for the cumulative probabilities, i.e $\hat{Y}_{m,i} = \hat{P}[Y_{m,i} = 1 \mid X_i = x]$. Second, the estimated cumulative probabilities are differenced to isolate the respective class probabilities $P_{m,i} = P[Y_i = m \mid X_i = x]$. Hence the estimate for the conditional probability of the m-th ordered class

[6] Wager and Athey (2018) point out that the leaves need to be relatively small in all dimensions of the covariate space. This implies that the high-dimensional settings are not considered and hence the theoretical asymptotic results might not hold in such settings.

[7] The condition of uniformity of covariates is due to simplicity and is not particularly binding as the result holds also with a density bounded away from zero and infinity as argued by Wager and Athey (2018). Furthermore, the Lipschitz-continuity of the conditional mean appears not too restrictive as the random forest estimates have in general smooth response surfaces when $B \to \infty$, i.e. the number of bootstrap or subsampling iterations goes to infinity (Bühlmann & Yu, 2002). Lastly, the appropriate scaling of the subsample size s_N does not affect the asymptotic normality, but violations might lead to asymptotic bias as pointed out by Wager and Athey (2018). For a detailed description of the conditions as well as of the proof, see Wager and Athey (2018).

is given by subtracting two adjacent cumulative probabilities as $\hat{P}_{m,i} = \hat{Y}_{m,i} - \hat{Y}_{m-1,i}$. Formally, the proposed estimation procedure can be described as follows:

1. Create $M-1$ binary indicator variables such as

$$Y_{m,i} = \mathbf{1}(Y_i \leq m) \qquad \text{for} \qquad m = 1, ..., M-1, \tag{2.4.1}$$

   where m is known and given by the definition of the dependent variable.
2. Estimate regression random forest for each of the $M-1$ indicators as

$$P[Y_{m,i} = 1 \mid X_i = x] = \sum_{i=1}^{N} w_{m,i}(x) Y_{m,i} \qquad \text{for} \qquad m = 1, ..., M-1, \tag{2.4.2}$$

   where the forest weights are defined as $w_{m,i}(x) = \frac{1}{B} \sum_{b=1}^{B} w_{m,b,i}(x)$ with trees weights given by $w_{m,b,i}(x) = \frac{\mathbf{1}(\{X_i \in L_{b,m}(x)\})}{|\{i: X_i \in L_{b,m}(x)\}|}$ with leaves $L_{b,m}(x)$ for a total of B trees.
3. Obtain forest predictions for each of the $M-1$ indicators as

$$\hat{Y}_{m,i} = \hat{P}[Y_{m,i} = 1 \mid X_i = x] = \sum_{i=1}^{N} \hat{w}_{m,i}(x) Y_{m,i} \qquad \text{for} \qquad m = 1, ..., M-1, \tag{2.4.3}$$

   where $\hat{Y}_{m,i}$ are estimated cumulative probabilities.
4. Compute ordered probabilities for each distinct class as

$$\hat{P}_{m,i} = \hat{Y}_{m,i} - \hat{Y}_{m-1,i} \qquad \text{for} \qquad m = 2, ..., M \tag{2.4.4}$$

   with

$$\hat{Y}_{M,i} = 1 \qquad \text{and} \qquad \hat{P}_{1,i} = \hat{Y}_{1,i} \tag{2.4.5}$$

   and

$$\hat{P}_{m,i} = 0 \qquad \text{if} \qquad \hat{P}_{m,i} < 0 \tag{2.4.6}$$

$$\hat{P}_{m,i} = \frac{\hat{P}_{m,i}}{\sum_{m=1}^{M} \hat{P}_{m,i}} \qquad \text{for} \qquad m = 1, ..., M, \tag{2.4.7}$$

where equation (2.4.4) makes use of the cumulative (nested) probability feature. As such, the predicted values of two subsequent binary indicator variables $Y_{m,i}$ are subtracted from each other to isolate the probability of the higher order class.[8] In equation (2.4.5) the first part is given by construction as follows from the indicator function (2.4.1) that all values of Y_i fullfil the condition for $m = M$ and from the fact that cumulative probabilities must add up to 1. The second part defines the probability of the lowest value of the ordered outcome variable. This follows directly from the random forest estimation as the created indicator variable $Y_{1,i}$ describes the very lowest value of the ordered outcome classes and as such, no modification of its predicted value is necessary to obtain a valid probability prediction. Line (2.4.6) ensures that the computed probabilities from (2.4.4) do not become negative. This might occasionally happen especially if the respective outcome classes comprise of very few observations. This issue is well-known also from the generalized ordered logit model where the parallel regression assumption is relaxed (see McCullagh & Nelder, 1989, p. 155). However, even though it is possible in theory, growing honest

[8] Similar transformations of an ordered model into multiple binary models have been proposed in the classification literature. Kwon, Han, and Lee (1997) introduce the so-called ordinal pairwise partitioning method in the context of neural networks. Yet the closest to our work is the approach by Frank and Hall (2001) who make use of the cumulative model explicitly.

trees seems to largely prevent this from happening in practice. Lastly, in case if negative predictions should occur and thus being set to zero, (2.4.7) defines a normalization step to ensure that all class probabilities sum up to 1. Notice, that such an approach requires estimation of $M-1$ forests in the training data, which might appear to be computationally expensive. However, given that most empirical problems involve a rather limited number of outcome classes (usually not exceeding 10 distinct classes) and the relatively fast estimation of standard regression forest[9] without any additional permutation test nor optimization steps needed as is the case for the conditional or the ordinal forests, respectively, the here proposed procedure shall be computationally advantageous (see Tables 2.B.26 and 2.B.27 in Appendix 2.B.4).

2.4.2 Marginal Effects

After estimating the conditional ordered choice probabilities, it is of interest to investigate how the estimated probabilities are associated with covariates, i.e. how the changes in the covariates translate into changes in the probabilities. Typical measures for such relationships in standard nonlinear econometrics are the marginal, or, partial effects. Thus, for nonlinear models, including ordered choice models, two fundamental measures are of common interest, mean marginal effects and marginal effects at the mean of the covariates.[10] These quantities are feasible also in the case of the *Ordered Forest* estimator. Due to the character of the ordered choice model, the marginal effects on all probabilities of different values of the ordered outcome classes are estimated, i.e. $P[Y_i = m \mid X_i = x]$. In the following, let us define the marginal effect for an element x^k of X_i as follows:

$$ME_i^{k,m}(x) = \frac{\partial P[Y_i = m \mid X_i^k = x^k, X_i^{-k} = x^{-k}]}{\partial x^k}, \tag{2.4.8}$$

with X_i^k and X_i^{-k} denoting the elements of X_i with and without the k-th element, respectively.[11] Next, let us define the marginal effect for categorical variables as a discrete change in the following way:

$$ME_i^{k,m}(x) = P[Y_i = m \mid X_i^k = \lceil x^k \rceil, X_i^{-k} = x^{-k}] - P[Y_i = m \mid X_i^k = \lfloor x^k \rfloor, X_i^{-k} = x^{-k}], \tag{2.4.9}$$

where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote upper and lower integer values, respectively, such that a difference of one unit is respected. Notice, that in the case of a binary variable this leads to the respective probabilities being evaluated at $\lceil x^k \rceil = 1$ and $\lfloor x^k \rfloor = 0$ as is usual for ordered choice models. From the above definitions of marginal effects, we obtain the desired quantity of interest, i.e. the marginal effect at mean by evaluating $ME_i^{k,m}(x)$ at the population mean of X_i, for which the sample mean is a natural proxy. The mean marginal effect is obtained by taking sample averages of $ME_i^{k,m}(x)$, i.e. $\frac{1}{N}\sum_{i=1}^{N} ME_i^{k,m}(x)$.

Having formally defined the desired marginal effects, the next issue is the estimation of these effects. For the case of binary and categorical covariates X^k, this appears straightforward as the estimated *Ordered Forest* model provides predicted values for all probabilities at all values x^k. As such, the estimate $\hat{ME}_i^{k,m}(x)$ of marginal effects defined in equation (2.4.9) remains as a difference of the two conditional probabilities estimated by the *Ordered Forest*. However, it is less obvious for continuous variables, where derivatives are needed. As the estimates of the choice probabilities are averaged leaf means, the marginal effect is not explicit and not differentiable. In the nonparametric literature Stoker (1996) and Powell and Stoker (1996), among others, are directly concerned with estimating average derivatives. However, these

[9] The computational speed of the regression forests depends on many tuning parameters, of which the number of bootstrap replications, i.e. grown trees is the most decisive one.

[10] One can evaluate the marginal effect at any arbitrarily chosen value. The default option is usually the mean or the median.

[11] As a matter of notation, capitals denote random variables, whereas small letters refer to the particular realizations of the random variable.

methods lack convenience of estimation and have thus not been widely adopted by empirical researchers.[12] Therefore, we approximate the derivative by a discrete analogue based on the definition of a derivative as follows:

$$\hat{ME}_i^{k,m}(x) = \frac{\hat{P}[Y_i = m \mid X_i^k = x^{kU}, X_i^{-k} = x^{-k}] - \hat{P}[Y_i = m \mid X_i^k = x^{kL}, X_i^{-k} = x^{-k}]}{x^{kU} - x^{kL}} \tag{2.4.10}$$

$$= \frac{\hat{P}_{m,i}(x^{kU}) - \hat{P}_{m,i}(x^{kL})}{x^{kU} - x^{kL}}, \tag{2.4.11}$$

with x^{kU}, x^{kL} defined as $x^{kU} = x^k + h \cdot \sigma(x^k)$ and $x^{kL} = x^k - h \cdot \sigma(x^k)$, while ensuring that the support of x^k is respected, and where $\sigma(\cdot)$ denotes standard deviation and h controls the window size for evaluating the marginal effect. We recommend to set $h = 0.1$ to achieve accurate evaluation at the margin.[13] Hence, the approximation targets the marginal change in the value of the covariate X_i^k. Notice, that such an estimation of marginal effects is much more demanding exercise than solely predicting the choice probabilities. Therefore, it is expected that considerably more subsampling iterations are needed for a good performance.

2.4.3 Inference

The building block of the *Ordered Forest* are the estimates of conditional probabilities such as $P[Y_{m,i} = 1 \mid X_i = x]$. Particularly, the *Ordered Forest* makes use of linear combinations of such probability estimates made by the random forest for both the conditional ordered choice probabilities as well as for the corresponding marginal effects. Therefore, for conducting inference on these quantities, it is sufficient to ensure that the underlying estimates of conditional probabilities are asymptotically normally distributed. Here, we combine the results of Wager and Athey (2018) and Lechner (2018). First, we use the asymptotic results of Wager and Athey (2018) who show that the consistency and normality of random forest predictions hold also when dealing with binary outcomes, and thus also hold for probability predictions of type $P[Y_{m,i} = 1 \mid X_i = x]$. Hence, the final *Ordered Forest* estimates for the conditional ordered choice probabilities and the marginal effects, based on a forest algorithm respecting the conditions discussed in Section 2.3, inherit the consistency and normality properties. Second, we adapt the inference procedure for random forests as developed by Lechner (2018) to estimate the variance of the conditional ordered choice probabilities and the corresponding marginal effects.

The here proposed method for conducting approximate inference of the estimated marginal effects utilizes the weight-based representation of random forest predictions and adapts the weight-based inference proposed by Lechner (2018) for the case of the *Ordered Forest* estimator.[14] The main condition for conducting weight-based inference is to ensure that the weights and the outcomes are independent. In general, the weights are functions of the covariates for the observation i and the training data. In order to estimate the variance of the marginal effects successfully, the conditioning set of the weights must be reduced. Therefore, if the observation i is not part of the training data and there is *i.i.d.* sampling, then the weights depend only on the observation i and are furthermore independent of the outcomes (for a formal analysis, see Lechner, 2018). This is achieved through sample splitting where one half of the sample is used to build the forest, and thus to determine the weights, and the other half to estimate

[12] The issues range from estimation difficulty, possibly non-standard distribution of the estimator, to ambiguous choices of nuisance parameters.

[13] We have additionally experimented with $h = 0.5$ and $h = 1$ which resulted in incrementally larger effect sizes. Generally, the lower the window size h, the more local the effect and the higher the window size h, the more global the effect becomes. As Burden and Faires (2011) point out, the window size h should not be chosen too small due to the instability of the numerical derivative approximations. In the software implementation in the R package orf, users can control this parameter by changing the argument window. See Lechner and Okasa (2019) for more details.

[14] See also Lechner (2002) and Imbens and Abadie (2006) for related approaches.

the effects using the respective outcomes. Notice that this condition goes beyond honesty as defined in Wager and Athey (2018) as this requires not only estimating honest trees but estimating honest forest as a whole. The reason for this is the fact that the weights are not based on the estimated trees, but on the estimated forest. Therefore, to ensure independence between the weights and outcomes, the honesty condition must be w.r.t. to the forest and it is not sufficient to build honest trees only. This comes, however, at the expense of the efficiency of the estimator as less data are effectively used. Nevertheless, the simulation evidence in Lechner (2018) suggests that this efficiency loss is small, if present at all.[15]

Since the *Ordered Forest* estimator is based on differences of random forest predictions for adjacent outcome categories, also the covariance term enters the variance formula of the final estimator[16] as opposed to the Modified Causal Forests developed in Lechner (2018). Further, the estimation of marginal effects is based on differences of single *Ordered Forest* predictions which also needs to be taken into account.[17] Let us first rewrite the marginal effects in terms of weighted means of the outcomes as follows:

$$
\begin{aligned}
\hat{ME}_i^{k,m}(x) &= \frac{\hat{P}_{m,i}(x^{kU}) - \hat{P}_{m,i}(x^{kL})}{x^{kU} - x^{kL}} \\
&= \frac{1}{x^{kU} - x^{kL}} \cdot \left(\left[\sum_{i=1}^{N} \hat{w}_{i,m}(x^{kU}) Y_{i,m} - \sum_{i=1}^{N} \hat{w}_{i,m-1}(x^{kU}) Y_{i,m-1} \right] - \left[\sum_{i=1}^{N} \hat{w}_{i,m}(x^{kL}) Y_{i,m} - \sum_{i=1}^{N} \hat{w}_{i,m-1}(x^{kL}) Y_{i,m-1} \right] \right) \\
&= \frac{1}{x^{kU} - x^{kL}} \cdot \left(\left[\sum_{i=1}^{N} \hat{w}_{i,m}(x^{kU}) Y_{i,m} - \sum_{i=1}^{N} \hat{w}_{i,m}(x^{kL}) Y_{i,m} \right] - \left[\sum_{i=1}^{N} \hat{w}_{i,m-1}(x^{kU}) Y_{i,m-1} - \sum_{i=1}^{N} \hat{w}_{i,m-1}(x^{kL}) Y_{i,m-1} \right] \right) \\
&= \frac{1}{x^{kU} - x^{kL}} \cdot \left(\sum_{i=1}^{N} \tilde{w}_{i,m}(x^{kU} x^{kL}) Y_{i,m} - \sum_{i=1}^{N} \tilde{w}_{i,m-1}(x^{kU} x^{kL}) Y_{i,m-1} \right),
\end{aligned}
$$

where $\tilde{w}_{i,m}(x^{kU}x^{kL}) = \hat{w}_{i,m}(x^{kU}) - \hat{w}_{i,m}(x^{kL})$, and $\tilde{w}_{i,m-1}(x^{kU}x^{kL}) = \hat{w}_{i,m-1}(x^{kU}) - \hat{w}_{i,m-1}(x^{kL})$ are the new weights defining the marginal effect. As such the quantity of interest for inference becomes the variance of the above expression given as:

$$
\begin{aligned}
Var\left(\hat{ME}_i^{k,m}(x)\right) &= Var\left(\frac{1}{x^{kU} - x^{kL}} \cdot \left(\sum_{i=1}^{N} \tilde{w}_{i,m}(x^{kU} x^{kL}) Y_{i,m} - \sum_{i=1}^{N} \tilde{w}_{i,m-1}(x^{kU} x^{kL}) Y_{i,m-1} \right) \right) \\
&= Var\left(\frac{\sum_{i=1}^{N} \tilde{w}_{i,m}(x^{kU} x^{kL}) Y_{i,m}}{x^{kU} - x^{kL}} \right) + Var\left(\frac{\sum_{i=1}^{N} \tilde{w}_{i,m-1}(x^{kU} x^{kL}) Y_{i,m-1}}{x^{kU} - x^{kL}} \right) \\
&\quad - 2 \cdot Cov\left(\frac{\sum_{i=1}^{N} \tilde{w}_{i,m}(x^{kU} x^{kL}) Y_{i,m}}{x^{kU} - x^{kL}}; \frac{\sum_{i=1}^{N} \tilde{w}_{i,m-1}(x^{kU} x^{kL}) Y_{i,m-1}}{x^{kU} - x^{kL}} \right),
\end{aligned}
$$

which suggests the following estimator for the variance:[18]

$$
\begin{aligned}
&\hat{Var}\left(\hat{ME}_i^{k,m}(x)\right) = \frac{N}{N-1} \cdot \frac{1}{(x^{kU} - x^{kL})^2} \cdot \\
&\cdot \left(\sum_{i=1}^{N} \left(\tilde{w}_{i,m}(x^{kU} x^{kL}) Y_{i,m} - \frac{1}{N} \sum_{i=1}^{N} \tilde{w}_{i,m}(x^{kU} x^{kL}) Y_{i,m} \right)^2 + \sum_{i=1}^{N} \left(\tilde{w}_{i,m-1}(x^{kU} x^{kL}) Y_{i,m-1} - \frac{1}{N} \sum_{i=1}^{N} \tilde{w}_{i,m-1}(x^{kU} x^{kL}) Y_{i,m-1} \right)^2 \right. \\
&\left. - 2 \cdot \sum_{i=1}^{N} \left(\tilde{w}_{i,m}(x^{kU} x^{kL}) Y_{i,m} - \frac{1}{N} \sum_{i=1}^{N} \tilde{w}_{i,m}(x^{kU} x^{kL}) Y_{i,m} \right) \cdot \left(\tilde{w}_{i,m-1}(x^{kU} x^{kL}) Y_{i,m-1} - \frac{1}{N} \sum_{i=1}^{N} \tilde{w}_{i,m-1}(x^{kU} x^{kL}) Y_{i,m-1} \right) \right),
\end{aligned}
$$

[15] The so-called cross-fitting to avoid the efficiency loss as suggested by Chernozhukov et al. (2018) does not appear to be applicable here as the independence of the weights and the outcomes would not be ensured.

[16] One could avoid the covariance term with an additional sample split, which might, however, further lead to a decreased efficiency of the estimator.

[17] Notice, that for outcome classes $m = 1$ and $m = M$, the variance formula simplifies substantially.

[18] Here, we estimate the variance with sample counterparts. An alternative approach, as in Lechner (2018), would be to first apply the law of total variance and, second, estimate the conditional moments by nonparametric methods. However, due to the presence of the covariance term the conditioning set contains 2 variables which causes the convergence rate to decrease and hence such variance estimation might even result in less precise estimates, depending on the sample size.

where for the marginal effects at the mean of the covariates the weights $\tilde{w}_{i,m}(x^{kU}x^{kL})$ and the scaling factor $1/(x^{kU}-x^{kL})^2$ are evaluated at the respective sample means, whereas for the mean marginal effects the average of the weights $\frac{1}{N}\sum_{i=1}^{N}\tilde{w}_{i,m}(x^{kU}x^{kL})$ and of the scaling factor $1/\left(\frac{1}{N}\sum_{i=1}^{N}(x^{kU}-x^{kL})\right)^2$ is used. Notice also the fact that the scaling factor drops out in the case of categorical covariates. According to the simulation study in Lechner (2018) the weight-based inference in case of the Modified Causal Forests tends to be rather conservative for the individual effects and rather accurate for aggregate effects. The results from the here conducted empirical application resemble this pattern where inference for the marginal effects at the mean of the covariates is more conservative in comparison to inference for the mean marginal effects (see also Appendix 2.C.2 for a comparison).

2.5 Monte Carlo Simulation

In order to investigate the finite sample performance of the proposed *Ordered Forest* estimator, we perform a Monte Carlo simulation study comparing competing estimators for ordered choice models based on the random forest algorithm. As a parametric benchmark, we take the ordered logistic regression. The considered models are specifically the following: (i) ordered logit (McCullagh, 1980), (ii) naive ordinal forest (Hornung, 2019a), (iii) ordinal forest (Hornung, 2019a), (iv) conditional forest (Hothorn et al., 2006b), and (v) *Ordered Forest* as developed in Section 2.4. Within the simulation study the *Ordered Forest* estimator is analyzed more closely to study the finite sample performance of the estimator depending on the particular forest building schemes and the way the ordering information is being taken into account. Regarding the former we study the *Ordered Forest* based on the standard random forest as in Breiman (2001), i.e. with boostrapping and *without* honesty as well as based on the adjusted random forest as in Wager and Athey (2018), i.e. with subsampling and *with* honesty. Regarding the latter we study an alternative approach for estimating the conditional choice probabilities which could be labelled as a 'multinomial' forest. In that case, the ordering information is not being taken into account and the probabilities of each category are estimated directly. The details of this approach are provided in Appendix 2.A.1. Given this, the *Ordered Forest* estimator should perform better than the multinomial forest in terms of the prediction accuracy thanks to the incorporation of additional information from the ordering of the outcome classes.

Table 2.5.1: General Settings of the Simulation

Monte Carlo	
observations in training set	200 (800)
observations in testing set	10000
replications	100
covariates with effect	15
trees in a forest	1000
randomly chosen covariates	$\sqrt{p}$
minimum leaf size[19]	5

General settings regarding the sample size, the number of replications, as well as forest-specific tuning parameters for the Monte Carlo simulation are depicted in Table 2.5.1. Furthermore, a detailed description of the software implementation of the respective estimators as well as the software specific tuning parameters are discussed in Appendix 2.B.4.

[19] Due to the conceptual differences of the conditional forests, an alternative stopping rule ensuring growing deep trees is chosen. See details in Appendix 2.B.4.

2.5.1 Data Generating Process

In terms of the data generating process, we built upon an ordered logit model as defined in (2.2.1) and (2.2.2). As such we simulate the underlying continuous latent variable Y_i^* as a linear function of regressors X_i, while drawing the error term u_i from the logistic distribution. Then, the continuous outcome Y_i^* is discretized into an ordered categorical outcome Y_i based on the threshold parameters α_m.[20] Furthermore, the intercept term is fixed to zero, i.e. $\beta_0 = 0$ and thus the thresholds are relative to this value of the intercept. As a result, such DGP captures the probability of the latent variable Y_i^* falling into a particular class given the location defined by the deterministic component of the model together with its stochastic component (Carsey & Harden, 2013).

In simulations of the data generating process, different numbers of possible discrete ordered classes are considered, particularly $M = \{3, 6, 9\}$ which corresponds to the simulation set-up used in Janitza et al. (2016) and Hornung (2019a). Further, both equal class widths, i.e. equally spaced threshold parameters α_m, as well as randomly spaced thresholds, while still preserving the monotonicity of the discrete outcome Y_i, are considered. For the latter, the threshold quantiles are drawn from the uniform distribution, i.e. $\alpha_m^q \sim U(0, 1)$ and ordered afterwards. For the former, the threshold quantiles are equally spaced between 0 and 1 depending on the number of classes. The β coefficients are specified as having fixed coefficient size, namely $\beta_1, ..., \beta_5 = 1$, $\beta_6, ..., \beta_{10} = 0.75$ and $\beta_{11}, ..., \beta_{15} = 0.5$ as is also the case in Janitza et al. (2016) and Hornung (2019a). Moreover, an option for nonlinear effects is introduced, too. As such, the covariates do not enter the functional form linearly, but are given by a sine function $sin(2X_i)$ as for example in Lin, Li, and Sun (2014), which is hard to model as opposed to other nonlinearities such as polynomials or interactions. The set of covariates X_i is drawn from the multivariate normal distribution with zero mean and a pre-specified variance-covariance matrix Σ, i.e. $X_i \sim \mathcal{N}(0, \Sigma)$, where Σ is specified either as an identity matrix and as such implying zero correlation between regressors, or it is specified to have a specific correlation structure between regressors[21] as follows:

$$\rho_{i,j} = \begin{cases} 1 & \text{for} \quad i = j \\ 0.8 & \text{for} \quad i \neq j; i, j \in \{1, 3, 5, 7, 9, 11, 13, 15\} \\ 0 & \text{otherwise} , \end{cases}$$

which is inspired by the correlation structure from the simulations in Janitza et al. (2016) and Hornung (2019a). Further, an option to include additional variables with zero effect is implemented as well. As such, another 15 covariates are added to the covariate space with $\beta_{16} = ... = \beta_{30} = 0$ from which 10 are again drawn from the normal distribution with zero mean and unit variance, i.e. $X_{i,0}^c \sim \mathcal{N}(0, 1)$ and 5 are dummies drawn from the binomial distribution, i.e. $X_{i,0}^d \sim \mathcal{B}(0.5)$. As the performance of the *Ordered Forest* estimator in high-dimensional settings is of particular interest, due to yet not fully understood theoretical properties in such settings, we include an option for additionally enlarging the covariate space with 1000 zero effect covariates $X_{i,0} \sim \mathcal{N}(0, 1)$, effectively creating a setting with $p >> N$. In the high-dimensional case the ordered logit is excluded from the simulations for obvious reasons. Overall, considering all the possible combinations for specifying the DGP, we end up with 72 different DGPs.[22]

[20] The thresholds are determined beforehand according to fixed threshold quantiles α_m^q of a large sample of $N = 1'000'000$ observations of the latent Y_i^* from the very same DGP to reflect the realized outcome distribution and then used afterwards in the simulations as a part of the deterministic component.

[21] Note that with a too high multicollinearity, the ordered logit model breaks down. With restricting the level of multicollinearity, the logit model can be still reasonably compared to the other competing methods.

[22] For the low-dimensional setting we have $n = 4$ options for the DGP settings, out of which we can choose from none to all of them, whereby the ordering does not matter, we end up with 16 possible combinations as given by the formula $\sum_{r=0}^{n} \binom{n}{r}$, each for 3 possible numbers of outcome classes resulting in 48 different DGPs. For the high-dimensional setting we have $n = 3$ options as the additional noise variables are always considered. This for all 3 distinct numbers of outcome classes yields 24 different DGPs.

For all of them we simulate a training dataset of size $N = 200$ and a testing dataset of size $N = 10'000$ for evaluating the prediction performance of the considered methods. We simulate the large testing set for three main reasons. First, the large testing set enables us to reduce the prediction noise and thus provides a more reliable measure for average out-of-sample performance of the estimators. Second, the large testing set also helps to reduce the simulation noise and thus to obtain more precise estimates for the performance measures. Third, we choose the large testing set to ensure further comparability with the simulation studies performed by Janitza et al. (2016) and Hornung (2019a). Note that such a large testing set is also common choice in many other simulation studies (see e.g. Jacob, 2020; or Knaus, Lechner, & Strittmatter, 2021). Further, we focus more closely on the simulation designs corresponding to the least and the most complex DGPs for which we simulate also a training set of size $N = 800$. The former DGP (labelled as simple DGP henceforth) corresponds exactly to an ordered logit model as in (2.2.1) with equal class widths, uncorrelated covariates with linear effects and without any additional zero effect variables. The latter DGP (labelled as complex DGP henceforth) features random class widths, correlated covariates with nonlinear effects and additional zero effect variables. For each replication, we estimate the model on the training set and evaluate the predictions on the testing set, for all tested methods.

2.5.2 Evaluation Measures

In order to properly evaluate the prediction performance we use two measures of accuracy, namely the mean squared error (MSE) and the ranked probability score (RPS). The former evaluates the error of the estimated conditional choice probabilities as a squared difference from the true values of the conditional choice probabilities. Given our simulation design, we know these true values, which are given as in equation (2.2.3). Hence, we can define the Monte Carlo average MSE as:

$$AMSE = \frac{1}{R}\sum_{j=1}^{R}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{m=1}^{M}\Big(P[Y_{i,j} = m \mid X_{i,j} = x] - \hat{P}[Y_{i,j} = m \mid X_{i,j} = x]\Big)^2,$$

where j refers to the j-th simulation replication, while R being the total number of replications. The second measure, the RPS as developed by Epstein (1969) is arguably the preffered measure for the evaluation of probability forecasts for ordered outcomes as it takes the ordering information into account (see Gneiting & Raftery, 2007; and Constantinou & Fenton, 2012). The Monte Carlo average RPS can be defined as follows:

$$ARPS = \frac{1}{R}\sum_{j=1}^{R}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{M-1}\sum_{m=1}^{M}\Big(P[Y_{i,j} \leq m \mid X_{i,j} = x] - \hat{P}[Y_{i,j} \leq m \mid X_{i,j} = x]\Big)^2,$$

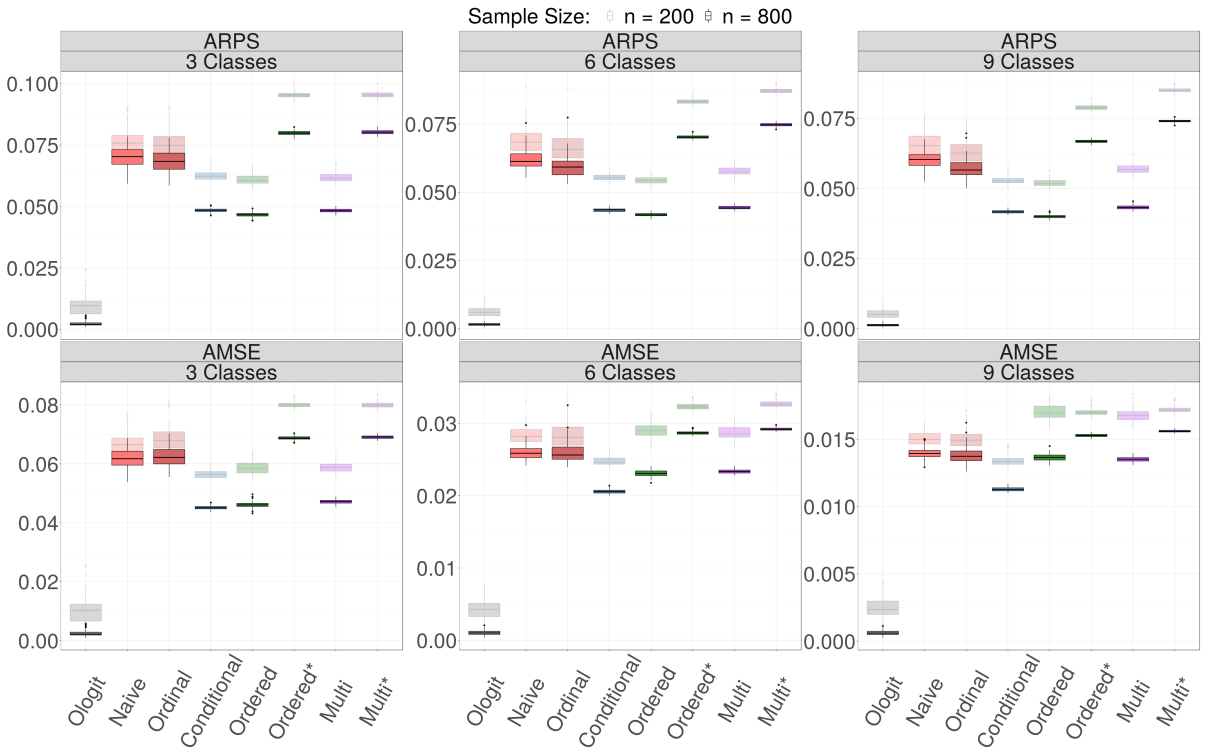
where on the contrary to the MSE, the difference between the cumulative choice probabilities is measured. The RPS can be seen as a generalization of the Brier Score (Brier, 1950) for multiple, ordered outcomes. As such, it measures the discrepancy between the predicted cumulative distribution function and the true one. Nevertheless, although the ordering information is taken into account, the relative distance between the classes is not reflected as pointed out by Janitza et al. (2016).

2.5.3 Simulation Results

For the sake of brevity, here we focus mainly on the simulation results obtained for the simple and for the complex DGP, while the results for all 72 DGPs are provided in Appendix 2.B.2. Figures 2.5.1

and 2.5.2 summarize the results for the low-dimensional setting for the simple and the complex DGP, respectively. Similarly, Figures 2.5.3 and 2.5.4 present the results for the simple and the complex DGP for the high-dimensional setting. The upper panels of the figures show the ARPS, the preferred accuracy measure, whereas the lower panels show the AMSE as a complementary measure. Within the figures the transparent boxplots in the background show the results for the smaller sample size along with the bold boxplots in the foreground showing the results for the bigger sample size. From left to right the figures present the results for 3, 6 and 9 outcome classes, respectively. The figures compare the prediction accuracy of the ordered logit, naive ordinal forest, ordinal forest, conditional forest, *Ordered Forest* and the multinomial forest, where the asterisk (*) denotes the honest version of the last two forests considered. Further tables with more detailed results and statistical tests for mean differences in the prediction errors are listed in Appendix 2.B.1.

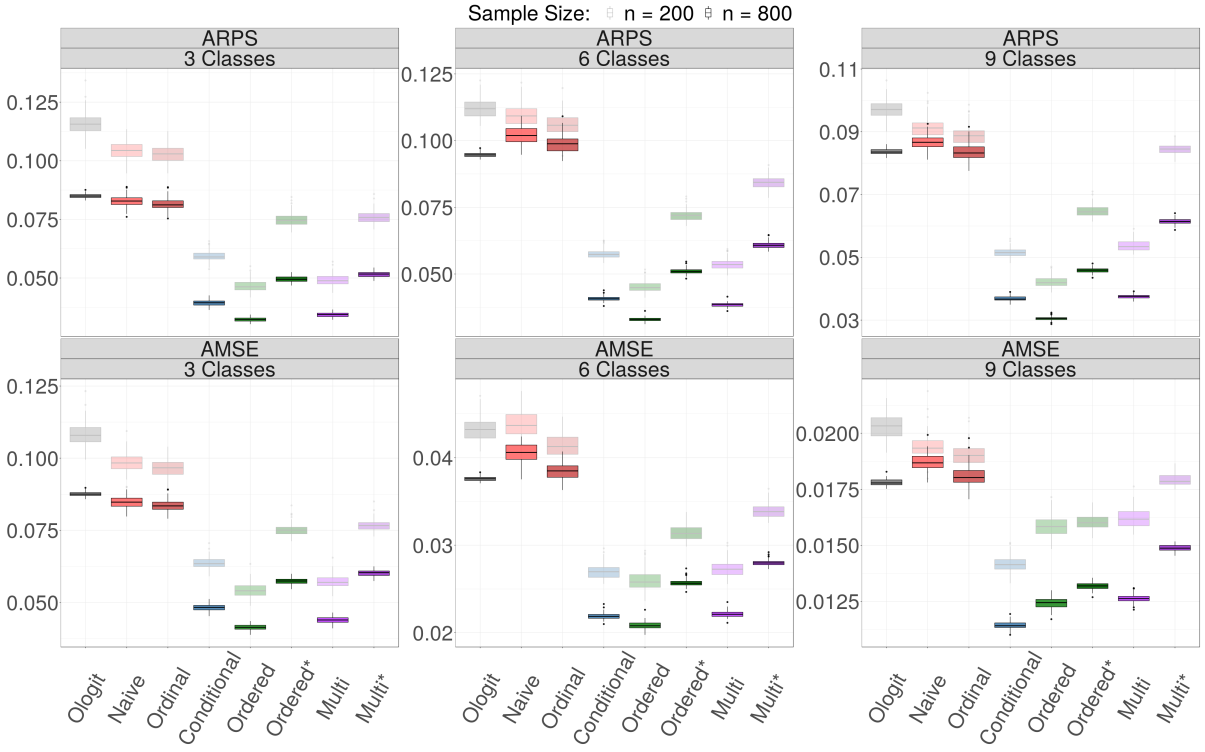
Figure 2.5.1: Simulation Results: Simple DGP & Low Dimension



Note: Figure summarizes the prediction accuracy results based on 100 simulation replications. The upper panel contains the ARPS and the lower panel contains the AMSE. The boxplots show the median and the interquartile range of the respective measure. The transparent boxplots denote the results for the small sample size, while the bold boxplots denote the results for the big sample size. From left to right the results for 3, 6, and 9 outcome classes are displayed.

In the low-dimensional setting with the simple DGP it is expected that the ordered logistic regression should perform best in terms of both the AMSE as well as the ARPS. Indeed, we do observe this results in Figure 2.5.1 as the ordered logit model performs unanimously best out of the considered models, reaching almost zero prediction error. Among the flexible forest-based estimators, the proposed *Ordered Forest* belongs to those better performing methods in terms of both accuracy measures. The honest versions of the forests lag behind what points at the efficiency loss due to the additional sample splitting. Overall, the ranking of the estimators stays stable with regards to the number of outcome categories. Additional pattern common to all estimators is the lower prediction error and increased precision with growing sample size.

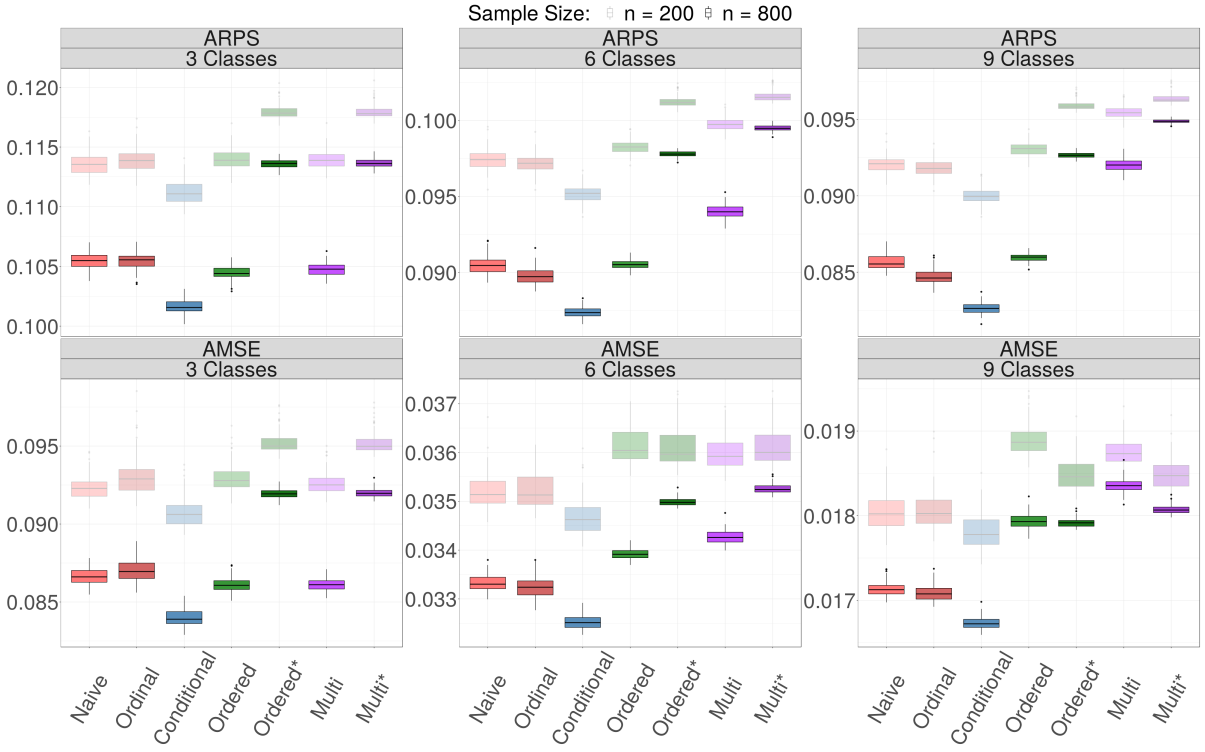
Figure 2.5.2: Simulation Results: Complex DGP & Low Dimension



Note: Figure summarizes the prediction accuracy results based on 100 simulation replications. The upper panel contains the ARPS and the lower panel contains the AMSE. The boxplots show the median and the interquartile range of the respective measure. The transparent boxplots denote the results for the small sample size, while the bold boxplots denote the results for the big sample size. From left to right the results for 3, 6, and 9 outcome classes are displayed.

In the case of the complex DGP, the performance of the flexible forest-based estimators is expected to be better in comparison to the parametric ordered logit. This can be seen in Figure 2.5.2 as the ordered logit lags behind the majority of the flexible methods in both accuracy measures. The somewhat higher prediction errors of the naive and the ordinal forest compared to the other forest-based methods might be due to their different primary target which are the ordered classes instead of the ordered probabilities as is the case for the other methods. In this respect the conditional forest exhibits considerably good prediction performance. The *Ordered Forest* outperforms the competing forest-based estimators in terms of the ARPS throughout all outcome class scenarios and also in terms of the AMSE in two scenarios, being outperformed only by the conditional forest in case of 9 outcome classes. Interestingly, the multinomial forest performs very well across all scenarios. However, it is consistently worse than the *Ordered Forest* with bigger discrepancy between the two the more outcome classes are considered. This points to the value of the ordering information and the ability of the *Ordered Forest* to utilize it in the estimation. With regards to the sample size, we observe the same pattern as in Figure 2.5.1.

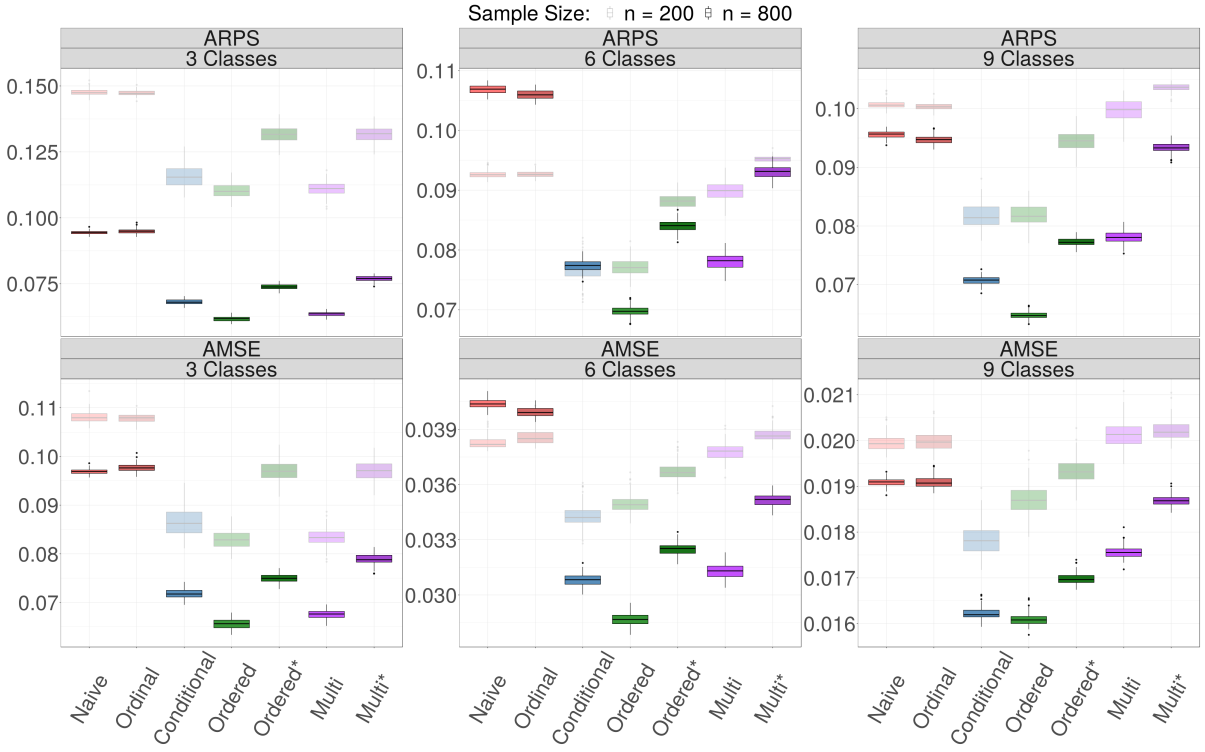
Figure 2.5.3: Simulation Results: Simple DGP & High Dimension



Note: Figure summarizes the prediction accuracy results based on 100 simulation replications. The upper panel contains the ARPS and the lower panel contains the AMSE. The boxplots show the median and the interquartile range of the respective measure. The transparent boxplots denote the results for the small sample size, while the bold boxplots denote the results for the big sample size. From left to right the results for 3, 6, and 9 outcome classes are displayed.

Considering the high-dimensional setting for the case of the simple DGP, we see in Figure 2.5.3 that the *Ordered Forest* slightly lags behind the other methods, except the scenarios with 3 outcome classes. In comparison, the conditional forest performs best in terms of the ARPS as well as in terms of the AMSE. Also the naive and the ordinal forest exhibit better performance compared to the previous simulation designs. However, it should be noted that the overall differences in the magnitude of the prediction errors are much lower within this simulation design as compared to the previous designs. Further, taking a closer look at the ARPS results of the multinomial forest we clearly see that in the simple ordered design the ignorance of the ordering information really harms the predictive performance of the estimator the more outcome classes are considered. Additionally, it is interesting to see that the performance gain due to a bigger sample size seems to be much less for the honest version of the forests in the high-dimensional setting as opposed to the low-dimensional setting.

Figure 2.5.4: Simulation Results: Complex DGP & High Dimension



Note: Figure summarizes the prediction accuracy results based on 100 simulation replications. The upper panel contains the ARPS and the lower panel contains the AMSE. The boxplots show the median and the interquartile range of the respective measure. The transparent boxplots denote the results for the small sample size, while the bold boxplots denote the results for the big sample size. From left to right the results for 3, 6, and 9 outcome classes are displayed.

Lastly, the case of the complex DGP in the high-dimensional setting as in Figure 2.5.4 shows some interesting patterns. In general, all of the methods exhibit good predictive performance as the loss in the prediction accuracy due to the high-dimensional covariate space is small. Additionally, although dealing with the most complex design, no substantial loss in the prediction accuracy can be observed in comparison to the less complex designs. This fact demonstrates the ability of the random forest algorithm as such to effectively cope with highly nonlinear functional forms even in high dimensions. Further, it seems that the role of the sample size is of particular importance in this complex design. On the contrary to the previous designs, where the prediction accuracy increases almost by a constant amount for all estimators and thus does not change their relative ranking, here it does not hold anymore. First, some estimators seem to learn faster than others, i.e. to have a faster rate of convergence. As such in the small sample size the *Ordered Forest* has in some settings higher values of the ARPS as well as the AMSE than the conditional forest, however manages to outperform the conditional forest in the bigger training sample. This becomes most apparent in the case of 9 outcome classes. Here, the median of the ARPS is almost the same for the two methods based on the small training sample, but significantly lower for the *Ordered Forest* based on the larger training sample.[23] Second, for the ordinal forest the prediction accuracy even worsens with the bigger training sample, which might hint on possible convergence issues. This might possibly come from the fact that the estimator comprises multiple distinct optimization and partly nonlinear transformation steps that are tied together, but lack formal asymptotic arguments to analyse the impacts and propagation of the estimation errors into the final point estimator. Overall, the *Ordered Forest* achieves the lowest ARPS as well as AMSE within this design, closely followed by the

[23] See Appendix 2.B.1 for the detailed results of the statistical tests conducted.

conditional and the multinomial forest. However, the generally good performance of the conditional forest might be due to a different type of the stopping criterion, which enables growing very deep trees that are possibly deeper than the classical Breiman (2001) trees with pre-specified minimum leaf size and as such might achieve lower bias which is then reflected in the lower values of ARPS as well as of AMSE.

In addition to the four main simulation designs discussed above, we also inspect all 72 different DGPs to analyze the performance and the sensitivity of the *Ordered Forest* to the particular features of the simulated DGPs (for details see Appendix 2.B.2). In case of both the low-dimensional setting, as well as the high-dimensional setting, the *Ordered Forest* performs particularly well if there are nonlinear effects accompanied by near-multicollinearity of regressors as such as well as together with additional noise variables or randomly spaced thresholds. Furthermore, the honest version of the *Ordered Forest* achieves consistently lower prediction accuracy in both settings. It seems that in small samples the increase in variance due to honesty dominates the reduction in the bias of the estimator. In order to further investigate the impact of the honesty feature in bigger samples as well as the convergence of the *Ordered Forest*, we quadruple the size of the training set once again and repeat the main simulation for the *Ordered Forest* and its honest version with $N = 3'200$ observations (see Appendix 2.B.1 for the full results). Firstly, for both versions we observe that with growing sample size the prediction errors get lower and the precision increases. However, the rate of convergence seems to be slower than the parametric rate of $\sqrt{N}$. Secondly, we observe the same pattern as in the smaller sample sizes, namely slightly lower prediction accuracy for the honest version of the *Ordered Forest* which stays roughly constant across all simulation designs. Hence, even in the biggest sample the additional variance dominates the bias reduction. However, it should be noted that for a prediction exercise honesty is an optional choice, while if inference is of interest, honesty becomes binding.

2.5.4 Empirical Results

Additionally to the above synthetic simulations, we explore the performance of the *Ordered Forest* estimator based on real datasets[24] previously used in Janitza et al. (2016) and Hornung (2019a). Table 2.5.2 summarizes the features of the datasets and the descriptive statistics are provided in Appendix 2.B.3.1. We compare our estimator in terms of the prediction accuracy to all the estimators used in the above Monte Carlo simulation.

Table 2.5.2: Description of the Datasets

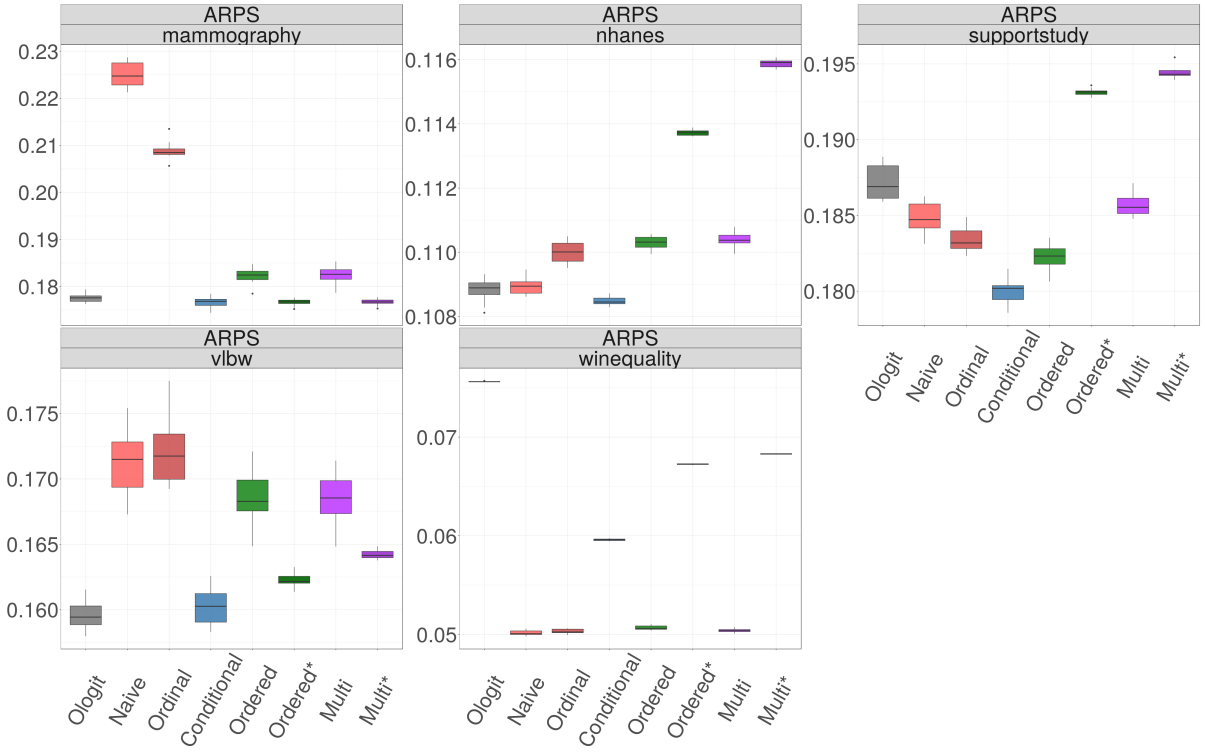
Datasets Summary						
Dataset	Sample Size	Outcome	Class Range			Covariates
Wine Quality	4893	Quality Score	1 (moderate)	-	6 (high)	11
Mammography	412	Visits History	1 (never)	-	3 (over year)	5
Nhanes	1914	Health Status	1 (excellent)	-	5 (poor)	26
Vlbw	218	Physical Condition	1 (threatening)	-	9 (optimal)	10
Support Study	798	Disability Degree	1 (none)	-	5 (fatal)	15

Similarly to Hornung (2019a) we evaluate the prediction accuracy based on a repeated cross-validation in order to reduce the dependency of the results on the particular training and test sample splits. As such we perform a 10-fold cross-validation on each dataset, i.e. we randomly split the dataset in 10 equally sized folds and use 9 folds for training the model and 1 fold for validation. This process is repeated such that each fold serves as a validation set exactly once. Lastly, we repeat this whole procedure 10 times

[24] The here proposed algorithm has been already applied and is in use for predicting match outcomes in football, see Goller et al. (2018) and SEW Soccer Analytics for details.

and report average accuracy measures. The results of the cross-validation exercise for the ARPS as well as the AMSE are summarized in Figures 2.5.5 and 2.5.6, respectively. Similarly as for the simulation results Appendix 2.B.3 contains more detailed statistics.

Figure 2.5.5: Cross-Validation: ARPS

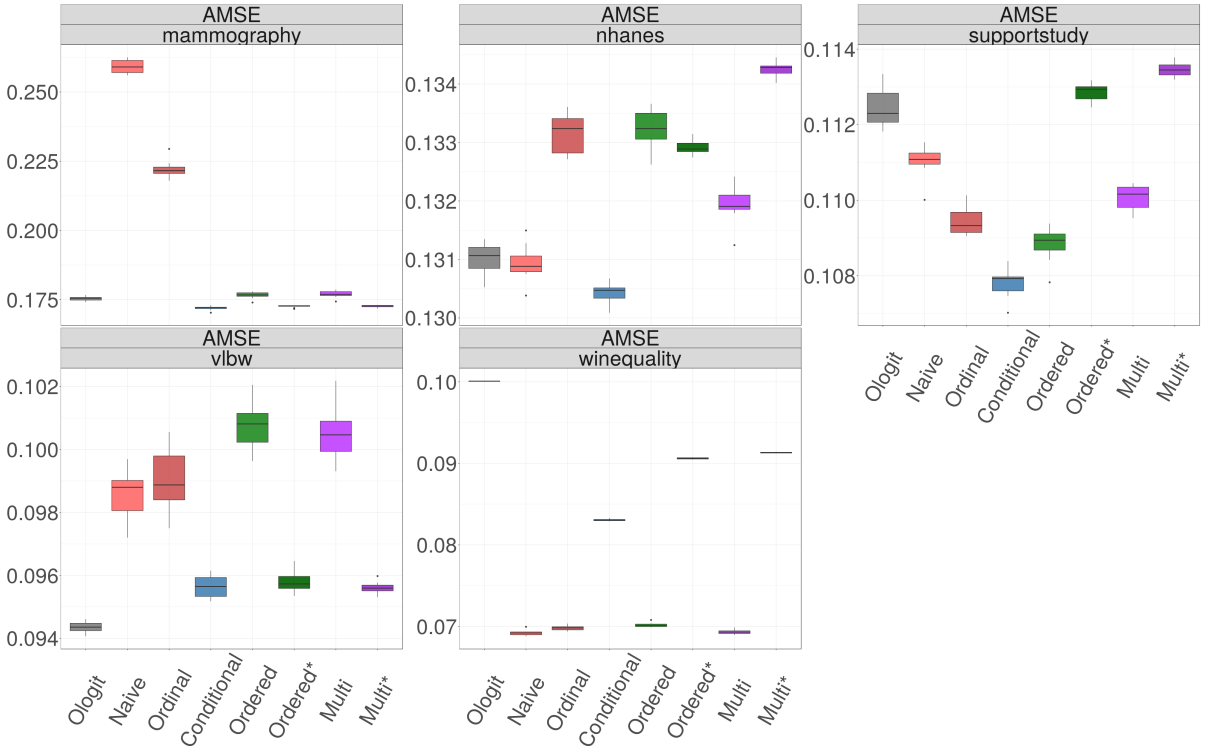


Note: Figure summarizes the prediction accuracy results in terms of the ARPS based on 10 repetitions of 10-fold cross-validation for respective datasets. The boxplots show the median and the interquartile range of the respective measure.

The main difference in evaluating the prediction accuracy in comparison to the simulation study is the fact that we do not observe the underlying ordered class probabilities, but only the realized ordered classes. This affects the computation of the accuracy measures and it can be expected that the prediction errors are somewhat higher in comparison to the simulation data, which is also the case here. Overall, the results imply a substantial heterogeneity in the prediction accuracy across the considered datasets. On the one hand, the parametric ordered logit does well in small samples (*vlbw*) whereas the forest-based methods are somewhat lagging behind. This is not surprising as a lower precision in small samples is the price to pay for the additional flexibility. On the other hand, in the largest sample (*winequality*) the ordered logit is clearly the worst performing method and all forest-based methods perform substantially better. With respect to the *Ordered Forest* estimator we observe relatively high prediction accuracy for three datasets (*mammography*, *supportstudy*, *winequality*) and relatively low prediction accuracy for two datasets (*nhanes*, *vlbw*) in comparison to the competing methods. The good performance in the *winequality* and the *supportstudy* dataset is expected due to the large samples available. In case of the *mammography* dataset, even when smaller in sample size, the *Ordered Forest* maintains the good prediction performance, with its honest version doing even better. The worse performance for the *vlbw* dataset might be due to the small sample size. However, the honest version of the *Ordered Forest* performs rather well. The relatively poor performance in the case of the *nhanes* dataset comes rather at surprise as the sample size is rather large. Nevertheless, here the differences among all estimators are very small in magnitude, in fact the smallest among the considered datasets. Overall, the empirical results provide

an evidence for a good predictive performance of the new *Ordered Forest* estimator based on various real datasets.

Figure 2.5.6: Cross-Validation: AMSE



Note: Figure summarizes the prediction accuracy results in terms of the AMSE based on 10 repetitions of 10-fold cross-validation for respective datasets. The boxplots show the median and the interquartile range of the respective measure.

2.6 Empirical Application

For an analysis of the relationship between the covariates and the predicted ordered choice probabilities we estimate the marginal effects for the *Ordered Forest* and compare these to the marginal effects estimated by the ordered logit. We estimate both common measures for marginal effects, i.e. the mean marginal effects as well as the marginal effects at covariate means. The main difference between the ordered logit and the *Ordered Forest* is the fact that the *Ordered Forest* does not use any parametric link function in the estimation of the marginal effects and as such does not impose any functional form on these estimates. As a result, the *Ordered Forest* does neither fix the sign of the marginal effects estimates nor revert it exactly once within the class range as is the case for the ordered logit (the so-called 'single crossing' feature, see i.e. Boes & Winkelmann, 2006; or Greene & Hensher, 2010) but rather estimates these in a data-driven manner. Nevertheless, the *Ordered Forest*, same as the ordered logit, still ensures that the marginal effects across the class range sum up to zero (being more likely to be in some particular classes must imply being less likely to be in some other classes). As such the *Ordered Forest* not only enables a more flexible estimation of the ordered choice probabilities but also of the marginal effects.

In order to showcase the *Ordered Forest* estimation of marginal effects, we revisit the question of self-assessed health status and its relationship with socio-economic characteristics as for example analyzed previously by Case et al. (2002) and Murasko (2008). In our empirical application we analyze the dataset from the 2009 National Health Interview Survey (NHIS) used in Angrist and Pischke (2014) which includes an ordered categorical outcome indicating a self-assessed health status. The specific survey

question of interest reads as: *'Would you say your health in general is excellent, very good, good, fair, or poor?'* and is coded on an ordered scale ranging from 1 (poor) to 5 (excellent). We examine how the ordered choice probabilities of the self-assessed health status differ for individuals with and without a coverage by private health insurance (see Levy & Meltzer, 2008, for a review of insurance effects on health) as well as how these probabilities vary with further socio-demographic characteristics, namely age, race and family size as well as economic characteristics, namely education, employment status and family income. The considered dataset is well-suited for demonstrating the evaluation of marginal effects for several reasons. First, the dataset features an ordered categorical outcome with 5 distinct ordered categories, which are unevenly distributed and thus challenging for estimating the associated marginal effects. Second, the dataset includes both continuous as well as categorical covariates which enables an exhaustive demonstration of the evaluation of marginal effects for various variable types. Third, the dataset contains more than 18′000 observations which allows for a precise estimation of the marginal effects. The descriptive statistics for the considered dataset are presented in Appendix 2.C.1.[25] We follow the data preparation of Angrist and Pischke (2014) and discard all observations with missing values and retain only individuals from single family households and those of age between 26 and 59 years as those do not yet qualify for the public health insurance program Medicare.

First of all, in order to describe the differences in the health status based on the health insurance we inspect the ordered class probabilities for the self-reported health status for individuals with and without a private health insurance contract. The descriptive results are reported in Table 2.6.1 below, including statistical evidence for the differences between the two groups. The descriptive evidence suggests that individuals with health insurance have a higher probability to be in excellent or very good health condition and at the same time have a lower probability to be in good or fair health condition. This evidence is both statistically precise and economically relevant. Furthermore, individuals with health insurance seem to have also a lower probability to be in poor health condition. However the evidence for that is less pronounced, both in statistical as well as in economic terms.

Table 2.6.1: Differences in Health Status based on Health Insurance: NHIS Dataset

NHIS Dataset					
	Health Insurance				
Health Status	Yes	No	Diff	tValue	pValue
Poor	1.07	1.51	-0.44	-1.84	6.61
Fair	4.81	10.19	-5.38	-9.28	0.00
Good	23.26	35.14	-11.88	-12.66	0.00
Very good	36.31	27.54	8.77	9.70	0.00
Excellent	34.55	25.62	8.93	10.08	0.00
N	15816	2974			

Next, in order to investigate the differences in the health status based on the health insurance we estimate the ordered choice probabilities for the self-reported health status conditional on having a private health insurance contract and further socio-economic characteristics using the *Ordered Forest* and the ordered logit and evaluate the corresponding marginal effects. Table 2.6.2 contains the estimated mean marginal effects for each outcome class for all covariates together with the associated standard errors, t-values, p-values as well as conventional significance levels for both the *Ordered Forest* as well as the ordered logit.[26]

[25] The dataset is freely accessible from the R-package stevedata (Miller, 2021) or in the data appendix of Angrist and Pischke (2014) available online.

[26] The results for the marginal effects at mean are available in Appendix 2.C.2.

In general, we see similar patterns in terms of the effect sizes and effect direction for both the *Ordered Forest* and the ordered logit. However, we do observe more variability in terms of the effect direction in case of the *Ordered Forest* as we would also expect given the flexibility arguments discussed above. In terms of uncertainty of the effects the weight-based inference seems to be slightly more conservative than the delta method used in the ordered logit. Nevertheless, the *Ordered Forest* also detects very precise effects which are not discovered by the ordered logit.

Table 2.6.2: Mean Marginal Effects: NHIS Dataset

Dataset		**Ordered Forest**					**Ordered Logit**				
Variable	Class	Effect	Std.Error	t-Value	p-Value		Effect	Std.Error	t-Value	p-Value	
Health Insurance	1	0.23	0.08	2.89	0.38	***	-0.11	0.05	-2.19	2.85	**
	2	-0.95	0.49	-1.93	5.35	*	-0.49	0.22	-2.22	2.68	**
	3	-4.51	1.99	-2.27	2.33	**	-1.32	0.59	-2.25	2.44	**
	4	4.44	1.80	2.47	1.35	**	0.02	0.03	0.65	51.88	
	5	0.78	2.47	0.32	75.21		1.90	0.83	2.29	2.22	**
Female	1	-0.19	0.12	-1.59	11.16		0.02	0.03	0.68	49.85	
	2	0.05	0.31	0.17	86.56		0.10	0.14	0.68	49.81	
	3	0.52	0.70	0.74	45.99		0.26	0.39	0.68	49.80	
	4	0.44	0.86	0.52	60.63		0.00	0.01	0.59	55.20	
	5	-0.82	1.16	-0.70	48.08		-0.39	0.57	-0.68	49.80	
Non White	1	0.38	0.15	2.57	1.02	**	0.36	0.05	7.02	0.00	***
	2	0.57	0.42	1.36	17.53		1.60	0.20	7.89	0.00	***
	3	5.97	1.12	5.32	0.00	***	4.10	0.48	8.57	0.00	***
	4	-4.23	1.09	-3.87	0.01	***	-0.26	0.08	-3.12	0.18	***
	5	-2.69	1.57	-1.72	8.57	*	-5.81	0.65	-8.87	0.00	***
Age	1	0.04	0.01	4.22	0.00	***	0.04	0.00	12.60	0.00	***
	2	0.15	0.03	5.09	0.00	***	0.20	0.01	19.77	0.00	***
	3	0.45	0.07	6.07	0.00	***	0.54	0.02	23.87	0.00	***
	4	-0.01	0.09	-0.13	89.49		0.01	0.01	1.24	21.54	
	5	-0.62	0.12	-5.10	0.00	***	-0.78	0.03	-24.15	0.00	***
Education	1	0.00	0.00	0.41	68.03		-0.11	0.01	-11.61	0.00	***
	2	-0.01	0.00	-1.73	8.42	*	-0.51	0.03	-16.80	0.00	***
	3	-0.02	0.01	-2.80	0.52	***	-1.39	0.07	-18.94	0.00	***
	4	0.00	0.01	0.71	48.08		-0.02	0.02	-1.23	21.83	
	5	0.02	0.01	2.57	1.03	**	2.04	0.11	18.85	0.00	***
Family Size	1	0.00	0.00	0.32	74.77		-0.01	0.01	-0.81	42.01	
	2	-0.00	0.01	-0.21	83.33		-0.04	0.05	-0.81	41.95	
	3	-0.06	0.02	-3.49	0.05	***	-0.12	0.14	-0.81	41.93	
	4	-0.03	0.02	-1.78	7.51	*	-0.00	0.00	-0.67	50.41	
	5	0.10	0.02	4.97	0.00	***	0.17	0.21	0.81	41.94	
Employed	1	-3.99	0.50	-7.94	0.00	***	-0.42	0.06	-7.30	0.00	***
	2	-3.81	0.73	-5.19	0.00	***	-1.86	0.23	-8.21	0.00	***
	3	2.58	1.15	2.25	2.44	**	-4.77	0.53	-8.98	0.00	***
	4	4.37	1.24	3.51	0.04	***	0.39	0.11	3.55	0.04	***
	5	0.84	1.82	0.46	64.34		6.66	0.71	9.42	0.00	***
Income	1	-0.11	0.04	-3.00	0.27	***	-0.00	0.00	-12.07	0.00	***
	2	-0.46	0.14	-3.27	0.11	***	-0.00	0.00	-17.73	0.00	***
	3	-0.06	0.51	-0.12	90.68		-0.00	0.00	-20.61	0.00	***
	4	0.36	0.37	0.97	33.42		-0.00	0.00	-1.24	21.41	
	5	0.27	0.45	0.61	54.03		0.00	0.00	20.96	0.00	***

Significance levels correspond to: $***. < 0.01$, $**. < 0.05$, $*. < 0.1$.

Notes: Table shows the comparison of the mean marginal effects in % points between the *Ordered Forest* and the ordered logit. The effects are estimated for all classes, together with the corresponding standard errors, t-values and p-values. The standard errors for the *Ordered Forest* are estimated using the weight-based inference and for the ordered logit are obtained via the delta method.

In particular, inspecting the variable of interest, namely the indicator for private health insurance, we immediately see the additional flexibility of the *Ordered Forest*. While both methods estimate positive marginal effects of having a private health insurance on the probability of being in very good or excellent health condition and negative marginal effects for being in good or fair health condition, the *Ordered Forest* estimates also a positive effect for being in poor health condition, whereas the ordered logit is

forced to estimate a negative effect due to its above-mentioned single-crossing property. As such, the *Ordered Forest* estimates a non-monotonic effect of having a private health insurance across the class probabilities. The results suggest that on one hand individuals with health insurance are less likely to be in good or fair health condition by 4.51 or 0.95 % points, respectively. On the other hand, individuals with health insurance are more likely to be in very good or excellent health condition by 4.44 or 0.78 % points, respectively, but they are also more likely to be in poor health condition by 0.23 % points. As the decision to sign up for a private health insurance is not random, i.e. the data comes from a non-experimental setting, it is not possible to uncover the causal effect without strong assumptions. One might, however, argue that based on the partial correlation evidence, due to the regular medical care and prevention the health insurance increases the likelihood of being in rather good health condition, but also that individuals with rather poor health condition are more likely to sign up for a private health insurance to cover up for the expected medical care costs. As can be seen, the *Ordered Forest* enables for such a non-monotonic effects analysis, while the ordered logit does not permit such mechanism to take place at all. Overall, in terms of effect sizes, for both estimators we observe smaller magnitudes in comparison to the unconditional differences presented in Table 2.6.1. However, the effect sizes estimated by the *Ordered Forest* are slightly bigger than those of the ordered logit. With regards to the statistical uncertainty around the estimated marginal effects, both methods exhibit similar level of precision.

Inspecting the effects of the additional conditioning variables, we see that neither the *Ordered Forest* nor the ordered logit find evidence for gender influencing the health class probabilities as the estimated effects are of small magnitude and lack statistical precision. In contrast, both methods estimate a higher probability of being in poor, fair or good health condition and conversely a lower probability of being in very good or excellent health condition for people of color, an effect that is sizeable and statistically precise. In this case, we note the slightly more conservative standard errors of the *Ordered Forest*. Furthermore, both methods estimate a higher likelihood of being in rather bad health condition and a lower likelihood of being in rather good health condition for increasing age with similar effect sizes as well as with similar statistical precision. In terms of education, there seem to be a positive relationship with regard to the probability of being in an excellent health condition. However, this effect is less pronounced for the *Ordered Forest* considering both the effect size and the precision in comparison to the ordered logit. The same positive relationship can be observed also for the family size and although the economic relevance of this effect is rather small, the *Ordered Forest* estimates this effect with high statistical precision, whereas the ordered logit does not find statistical evidence in this respect. Considering the employment status, both methods estimate lower likelihood of being in rather bad health condition and a higher likelihood of being in good health condition with comparable effect sizes as well as statistical precision. Lastly, the *Ordered Forest* and the ordered logit both estimate a positive relationship with regards to the income level. As such, individuals with higher income are less likely to be in rather bad health condition and more likely to be in rather good health condition. In case of the *Ordered Forest*, the effect sizes are slightly larger, but with lower statistical precision, finding relevant evidence only for the negative effects on the fair and poor health status, whereas in case of the ordered logit the statistical precision is higher, however with effectively estimating a zero effect. This might be due to the somewhat higher collinearity between the education and income level (0.45), which suggests a better handling of near-multicollinearity among covariates of the *Ordered Forest* as has been documented in the simulation study. Overall, however, the main advantage of the estimation of the marginal effects by the *Ordered Forest* stems from a more flexible, data-driven approximation of possible nonlinearities in the functional form.

2.7 Conclusion

In this paper, we develop and apply a new machine learning estimator of the econometric ordered choice models based on the random forest algorithm. The *Ordered Forest* estimator is a flexible alternative to parametric ordered choice models such as the ordered logit or ordered probit which does not rely on any distributional assumptions and provides essentially the same output as the parametric models, including the estimation of the marginal effects as well as the associated inference. The proposed estimator utilizes the flexibility of random forests and can thus naturally deal with nonlinearities in the data and with a large-dimensional covariate space, while taking the ordering information of the categorical outcome variable into account. Hence, the estimator flexibly estimates the conditional ordered choice probabilities without restrictive assumptions about the distribution of the error term, or other assumptions such as the single index and constant threshold assumptions as is the case for the parametric ordered choice models (see Boes & Winkelmann, 2006, for a discussion of these assumptions). Further, the estimator allows also the estimation of the marginal effects, i.e. how the estimated conditional ordered choice probabilities vary with changes in covariates. The weighted representation of these effects enables the weight-based inference as suggested by Lechner (2018). The fact that the estimator comprises of linear combinations of random forest predictions ensures that the theoretical guarantees of Wager and Athey (2018) are satisfied. Additionally, a free software implementation of the *Ordered Forest* estimator is available in the R-package orf on the official CRAN repository to enable the usage of the method by applied researchers.

The performance of the *Ordered Forest* estimator is studied and compared to other competing estimators in an extensive Monte Carlo simulation as well as using real datasets. The simulation results suggest good performance of the estimator in finite samples, including also high-dimensional settings. The advantages of the machine learning estimation compared to a parametric method become apparent when dealing with near-multicollinearity and highly nonlinear functional forms. In such cases all of the considered forest-based estimators perform better than the ordered logit in terms of the prediction accuracy. Among the forest-based estimators the *Ordered Forest* proposed in this paper performs well throughout all simulated DGPs and outperforms the competing methods in the most complex simulation designs. The empirical evidence using real datasets supports the findings from the Monte Carlo simulation. Additionally, the estimation of the marginal effects as well as the inference procedure seems to work well in the presented empirical example.

Despite the attractive properties of the *Ordered Forest* estimator, many interesting questions are left open. Particularly, a further extension of the Monte Carlo simulation to study the sensitivity of the *Ordered Forest* in respect to tuning parameters of the underlying random forest as well as in respect to different simulation designs would be of interest. Similarly, the performance of the estimator with and without honesty for larger sample sizes should be further investigated. Also, the optimal choice of the size of the window for evaluating the marginal effects would be worth to explore. Additionally, besides the theoretical guarantees for the point estimator, a rigorous asymptotic analysis of the weight-based inference procedure for the estimation of standard errors would be beneficial to describe the exact theoretical properties. Lastly, it would be of great interest to see more real data applications of the *Ordered Forest* estimator such as for example in Kim, Lym, and Kim (2021), especially for large samples.

Acknowledgements

A previous version of the paper was presented at research seminars of the University of St.Gallen, at the German Statistical Week in Trier and the Statistics of Machine Learning Conference in Prague. We thank participants, in particular Francesco Audrino, Daniel Goller, Michael Knaus, two anonymous referees and an associate editor for helpful comments. The usual disclaimer applies.

Bibliography

Afonso, A., Gomes, P., & Rother, P. (2009). Ordered response models for sovereign debt ratings. *Applied Economics Letters*, *16*(8), 769–773.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley series in probability and statistics.

Angrist, J. D. & Pischke, J. S. (2014). *Mastering 'metrics: The path from cause to effect*.

Archer, K. J., Hou, J., Zhou, Q., Ferber, K., Layne, J. G., & Gentry, A. E. (2014). Ordinalgmifs: An R package for ordinal regression in high-dimensional data settings. *Cancer Informatics*, *13*, 187–195.

Athey, S. & Imbens, G. W. (2016). Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148–1178.

Biau, G. (2012). Analysis of a Random Forests Model. *Journal of Machine Learning Research*, *13*(1), 1063–1095.

Boes, S., Staub, K., & Winkelmann, R. (2010). Relative status and satisfaction. *Economics Letters*, *109*(3), 168–170.

Boes, S. & Winkelmann, R. (2006). Ordered response models. In *Modern econometric analysis: Surveys on recent developments* (pp. 167–181).

Boes, S. & Winkelmann, R. (2010). The Effect of Income on Positive and Negative Subjective Well-Being. *Social Indicators Research*, *95*(2), 111–128.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. CRC Press.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, *78*(1), 1–3.

Bühlmann, P. & Yu, B. (2002). Analyzing bagging.

Burden, R. L. & Faires, J. D. (2011). *Numerical Analysis 9th Edition*.

Buri, M. & Hothorn, T. (2020). Model-based random forests for ordinal regression. *International Journal of Biostatistics*, *16*(2).

Butler, J. S., Finegan, T. A., & Siegfried, J. J. (1998). Does more calculus improve student learning in intermediate micro- and macroeconomic theory? *Journal of Applied Econometrics*, *13*(2), 185–202.

Carroll, N. (2018). oglmx: Estimation of Ordered Generalized Linear Models. R package version 3.0.0.0.

Carsey, T. M. & Harden, J. J. (2013). *Monte Carlo simulation and resampling methods for social science*. Sage Publications.

Case, A., Lubotsky, D., & Paxson, C. (2002). Economic status and health in childhood: The origins of the gradient. *American Economic Review*, *92*(5), 1308–1334.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, *21*(1), 1–68.

Constantinou, A. C. & Fenton, N. E. (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, *8*(1).

Epstein, E. S. (1969). A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology*, *8*(6), 985–987.

Fox, J. T. (2007). Semiparametric estimation of multinomial discrete-choice models using a subset of choices. *RAND Journal of Economics*, *38*(4), 1002–1019.

Frank, E. & Hall, M. (2001). A simple approach to ordinal classification. In *European conference on machine learning* (Vol. 2167, pp. 145–156).

Gneiting, T. & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Gogas, P., Papadimitriou, T., & Agrapetidou, A. (2014). Forecasting bank credit ratings. *Journal of Risk Finance*, *15*(2), 195–209.

Goller, D., Knaus, M. C., Lechner, M., & Okasa, G. (2018). Predicting Match Outcomes in Football by an Ordered Forest Estimator. *Economics Working Paper Series*, *No.1811*.

Greene, W. H. & Hensher, D. A. (2010). *Modeling ordered choices: A primer*. Cambridge University Press.

Hamermesh, D. S. & Parker, A. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review*, *24*(4), 369–376.

Harrell, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Harrell, F. E. (2019). rms: Regression Modeling Strategies. R package version 5.1-3.1.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer Science & Business Media.

Hornung, R. (2019a). Ordinal Forests. *Journal of Classification*, 1–14.

Hornung, R. (2019b). ordinalForest: Ordinal Forests: Prediction and Variable Ranking with Ordinal Target Variables. R package version 2.3-1.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006a). Survival ensembles. *Biostatistics*, *7*(3), 355–373.

Hothorn, T., Hornik, K., & Zeileis, A. (2006b). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Hothorn, T., Lausen, B., Benner, A., & Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in Medicine*, *23*(1), 77–91.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, *58*(1-2), 71–120.

Imbens, G. W. & Abadie, A. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, *74*(1), 235–267.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*.

Jackson, D. J. & Darrow, T. I. (2005). The influence of celebrity endorsements on young adults' political opinions. *Harvard International Journal of Press/Politics*, *10*(3), 80–98.

Jacob, D. (2020). Cross-Fitting and Averaging for Machine Learning Estimation of Heterogeneous Treatment Effects. *arXiv preprint arXiv:2007.02852*.

Janitza, S., Tutz, G., & Boulesteix, A. L. (2016). Random forest for ordinal responses: Prediction and variable selection. *Computational Statistics and Data Analysis*, *96*, 57–73.

Kim, S., Lym, Y., & Kim, K. J. (2021). Developing crash severity model handling class imbalance and implementing ordered nature: Focusing on elderly drivers. *International Journal of Environmental Research and Public Health*, *18*(4), 1–22.

Klein, R. W. & Sherman, R. P. (2002). Shift restrictions and semiparametric estimation in ordered response models. *Econometrica*, *70*(2), 663–691.

Klein, R. W. & Spady, R. H. (1993). An Efficient Semiparametric Estimator for Binary Response Models. *Econometrica*, *61*(2), 387–421.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, *24*(1), 134–161.

Kramer, S., Widmer, G., Pfahringer, B., & De Groeve, M. (2001). Prediction of Ordinal Classes Using Regression Trees. *Fundamenta Informaticae*, *47*, 1–13.

Kwon, Y. S., Han, I., & Lee, K. C. (1997). Ordinal Pairwise Partitioning (OPP) Approach to Neural Networks Training in Bond rating. *International Journal of Intelligent Systems in Accounting, Finance & Management*, *6*(1), 23–40.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics*, *84*(2), 205–220.

Lechner, M. (2018). Modified Causal Forests for Estimating Heterogeneous Causal Effects. *arXiv preprint arXiv: 1812.09487v2*.

Lechner, M. & Okasa, G. (2019). orf: Ordered Random Forests. CRAN R package version 0.1.3.

Lee, L.-F. (1995). Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics*, *65*(2), 381–428.

Levy, H. & Meltzer, D. (2008). The impact of health insurance on health. In *Annual review of public health* (Vol. 29, pp. 399–409).

Lewbel, A. (2000). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics*, *97*(1), 145–177.

Lin, Z., Li, Q., & Sun, Y. (2014). A consistent nonparametric test of parametric regression functional form in fixed effects panel data models. *Journal of Econometrics*, *178*(1), 167–179.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*, 14–23.

Matzkin, R. L. (1992). Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and The Binary Choice Models. *Econometrica*, *60*(2), 239–270.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. New York Chapman & Hall.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B. 42*(2), 109–142.

Meinshausen, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research*, *7*(Jun), 983–999.

Mentch, L. & Hooker, G. (2016). Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. *Journal of Machine Learning Research*, *17*(1), 841–881.

Miller, S. (2021). stevedata: Steve's Toy Data for Teaching About a Variety of Methodological, Social, and Political Topics.

Murasko, J. E. (2008). An evaluation of the age-profile in the relationship between household income and the health of children in the United States. *Journal of Health Economics*, *27*(6), 1489–1502.

Piccarreta, R. (2008). Classification trees for ordinal variables. *Computational Statistics*, *23*(3), 407–427.

Powell, J. L. & Stoker, T. M. (1996). Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics*, *75*(2), 291–316.

R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Racine, J. S. (2008). Nonparametric Econometrics: A Primer. *Foundations and Trends® in Econometrics*, *3*(1), 1–88.

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

Stewart, M. B. (2005). A comparison of semiparametric estimators for the ordered response model. *Computational Statistics and Data Analysis*, *49*(2), 555–573.

Stoker, T. M. (1996). Smoothing bias in the measurement of marginal effects. *Journal of Econometrics*, *72*(1-2), 49–84.

Strasser, H. & Weber, C. (1999). On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, *8*, 220–250.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, *9*(1), 1–11.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 1–21.

Tibshirani, J., Athey, S., Wager, S., Friedberg, R., Miner, L., & Wright, M. (2018). grf: Generalized Random Forests. R package version 0.10.2.

Tutz, G. (2021). Ordinal Trees and Random Forests: Score-Free Recursive Partitioning and Improved Ensembles. *arXiv preprint arXiv: 2102.00415*.

Wager, S. (2014). Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*.

Wager, S. & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Williams, R. (2016). Understanding and interpreting generalized ordered logit models. *Journal of Mathematical Sociology*, *40*(1), 7–20.

Wright, M. N. & Ziegler, A. (2017). ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17.

Wurm, M. J., Rathouz, P. J., & Hanlon, B. M. (2017). Regularized Ordinal Regression and the ordinalNet R Package. *arXiv preprint arXiv:1706.05003*, 1–42.

Appendix

2.A Other Machine Learning Estimators

2.A.1 Multinomial Forest

Considering the *Ordered Forest* estimator a possible modification for models with categorical outcome variable *without* an inherent ordering appears to be straightforward. Instead of estimating cumulative probabilities and afterwards isolating the respective class probabilities, we can estimate the class probabilities $P_{m,i} = P[Y_i = m \mid X_i = x]$ directly. As such the binary outcomes are now constructed to indicate the particular outcome classes separately. Then the random forest predictions for each class yield the conditional choice probabilities which need to be afterwards normalized to sum up to 1. Formally, consider (un)ordered categorical outcome variable $Y_i \in \{1, ..., M\}$ with classes m and sample size $N (i = 1, ..., N)$. Then, the estimation procedure can be described as follows:

1. Create M binary indicator variables such as

$$Y_{m,i} = \mathbf{1}(Y_i = m) \qquad \text{for} \qquad m = 1, ..., M. \tag{2.A.1}$$

   where m is known and given by the definition of the dependent variable.

2. Estimate regression random forest for each of the M indicators as

$$P[Y_{m,i} = 1 \mid X_i = x] = \sum_{i=1}^{N} w_{m,i}(x) Y_{m,i} \qquad \text{for} \qquad m = 1, ..., M, \tag{2.A.2}$$

   where the forest weights are defined as $w_{m,i}(x) = \frac{1}{B} \sum_{b=1}^{B} w_{m,b,i}(x)$ with trees weights given by $w_{m,b,i}(x) = \frac{\mathbf{1}(\{X_i \in L_{b,m}(x)\})}{|\{i: X_i \in L_{b,m}(x)\}|}$ with leaves $L_{b,m}(x)$ for a total of B trees.

3. Obtain forest predictions for each of the M indicators as

$$\hat{Y}_{m,i} = \hat{P}[Y_{m,i} = 1 \mid X_i = x] = \sum_{i=1}^{N} \hat{w}_{m,i}(x) Y_{m,i} \qquad \text{for} \qquad m = 1, ..., M, \tag{2.A.3}$$

   where $\hat{Y}_{m,i}$ are estimated probabilities.

4. Compute probabilities for each class as

$$\hat{P}_{m,i} = \hat{Y}_{m,i} \qquad \text{for} \qquad m = 1, ..., M \tag{2.A.4}$$

$$\hat{P}_{m,i} = \frac{\hat{P}_{m,i}}{\sum_{m=1}^{M} \hat{P}_{m,i}} \qquad \text{for} \qquad m = 1, ..., M, \tag{2.A.5}$$

where the equation (2.A.4) defines the probabilities of all M classes and subsequent equation (2.A.5) ensures that the probabilities sum up to 1 as this might not be the case otherwise. Similarly to the

Ordered Forest estimator, also the multinomial forest is a linear combination of the respective forest predictions and as such also inherits the theoretical properties stemming from random forest estimation as described in Section 2.3 of the main text.

2.A.2 Conditional Forest

The conditional forest as discussed in Section 2.2 of the main text is grown with the so-called conditional inference trees. The main idea is to provide an unbiased way of recursive splitting of the trees using a test statistic based on permutation tests (Strasser & Weber, 1999). To describe the estimation procedure, consider an ordered categorical outcome $Y_i \in (1, ..., M)$ with ordered classes m and sample size $N (i = 1, ..., N)$. Further, define binary case weights $w_i \in \{0, 1\}$ which determine if the observation is part of the current leaf. Then, the algorithm developed by Hothorn et al. (2006b) can be described as follows:

1. Test the global null hypothesis of independence between any of the P covariates and the outcome, for the particular case weights, given a bootstrap sample Z_b. Afterwards, select the p-th covariate $X_{i,p}$ with the strongest association with the outcome Y_i, or stop if the null hypothesis cannot be rejected. The association is measured by a linear statistic T given as:

$$T_p(Z_b, w) = \sum_{i=1}^{N} w_i g_p(X_{i,p}) h(Y_i), \tag{2.A.6}$$

   where $g_p(\cdot)$ and $h(\cdot)$ are specific transformation functions.

2. Split the covariate sample space $\mathcal{X}_p$ into two disjoint sets $\mathcal{I}$ and $\mathcal{J}$ with adapted case weights $w_i \mathbf{1}(X_{i,p} \in \mathcal{I})$ and $w_i \mathbf{1}(X_{i,p} \in \mathcal{J})$ determining the observations falling into the subset $\mathcal{I}$ and $\mathcal{J}$, respectively. Then, the split is chosen by evaluating a two-sample statistic as a special case of 2.A.6:

$$T_p^{\mathcal{I}}(Z_b, w) = \sum_{i=1}^{N} w_i \mathbf{1}(X_{i,p} \in \mathcal{I}) h(Y_i) \tag{2.A.7}$$

   for all possible subsets $\mathcal{I}$ of the covariate sample space $\mathcal{X}_p$.

3. Repeat steps 1 and 2 recursively with modified case weights.

Hence, the above algorithm distinguishes between variable selection (step 1) and splitting rule (step 2), while both relying on the variations of the test statistic $T_p(Z_b, w)$. In practice, however, the distribution of this statistic under the null hypothesis is unknown and depends on the joint distribution of Y_i and $X_{i,p}$. For this reason, the permutation tests are applied to abstract from the dependency by fixing the covariates and conditioning on all possible permutations of the outcomes. Then, the conditional mean and covariance of the test statistic can be derived and the asymptotic distribution can be approximated by Monte Carlo procedures, while Strasser and Weber (1999) proved its normality. Finally, variables and splits are chosen according to the lowest p-value of the test statistic $T_p(Z_b, w)$ and $T_p^{\mathcal{I}}(Z_b, w)$, respectively.

Besides the permutation tests, the choice of the tranformation functions $g_p(\cdot)$ and $h(\cdot)$ is important and depends on the type of the variables. For continuous outcome and covariates, identity transformation is suggested. For the case of an ordinal regression which is of interest here, the transformation function is given through the score function $s(m)$. If the underlying latent Y_i^* is unobserved, it is suggested that $s(m) = m$ and thus $h(Y_i) = Y_i$. Hence, in the tree building the ordered outcome is treated as a continuous one (Janitza et al., 2016). Then, however, the leaf predictions are the choice probabilities computed as proportions of the outcome classes falling within the leaf, instead of fitting a within leaf

constant. The final conditional forest predictions for the choice probabilities are the averaged conditional tree probability predictions. Such obtained choice probabilities are analyzed in the Monte Carlo study in Section 2.5 of the main text.

2.A.3 Ordinal Forest

In the following, the algorithm for the ordinal forest as developed by Hornung (2019a) is described. To begin with, consider an ordered categorical outcome $Y_i \in (1, ..., M)$ with ordered classes m and sample size $N (i = 1, ..., N)$. Then, for a set of optimization forests $b = 1, ..., B_{sets}$:

1. Draw $M-1$ uniformly distributed variables $D_{b,m} \sim U(0, 1)$ and sort them according to their values. Further, set $D_{b,1} = 0$ and $D_{b,M+1} = 1$.
2. Define a score set $S_{b,m} = \{S_{b,1}, ..., S_{b,M}\}$ with scores constructed as $S_{b,m} = \Phi^{-1}\big(\frac{D_{b,m}+D_{b,m+1}}{2}\big)$ for $m = 1, ..., M$, where $\Phi(\cdot)$ is the cdf of the standard normal.
3. Create a new continuous outcome $Z_{b,i} = (Z_{b,1}, ..., Z_{b,N})$ by replacing each class value m of the original ordered categorical Y_i by the m-th value of the score set $S_{b,m}$ for all $m = 1, ..., M$.
4. Use $Z_{b,i}$ as dependent variable and estimate a regression forest $RF_{S_{b,m}}$ with B_{prior} trees.
5. Obtain the out-of-bag (OOB) predictions for the continuous $Z_{b,i}$ and transform them into predictions for Y_i as follows: $\hat{Y}_{b,i} = m$ if $\hat{Z}_{b,i} \in \big]\Phi^{-1}(D_{b,m}, \Phi^{-1}(D_{b,m+1})\big]$ for all $i = 1, ..., N$.
6. Compute a performance measure for the given forest $\hat{RF}_{S_{b,m}}$ based on some performance function of type $f(Y_i, \hat{Y}_{b,i})$.

After estimating B_{sets} of optimization forests, take S_{best} of these which achieved the best performance according to the performance function. Then, construct the final set of uniformly distributed variables $D_1, ..., D_{M+1}$ as an average of those from S_{best} for $m = 1, ..., M+1$. Finally, form the optimized score set $S_m = \{S_1, ..., S_M\}$ with scores constructed as $S_m = \Phi^{-1}\big(\frac{D_m+D_{m+1}}{2}\big)$ for $m = 1, ..., M$. The continuous outcome $Z_i = (Z_1, ..., Z_N)$ is then similarly as in the optimization procedure constructed by replacing each m value of the original outcome Y_i by the m-th value of the optimized score set S_m for all $m = 1, ..., M$. Finally, estimate the regression forest RF_{final} using Z_i as the dependent variable. On one hand, the class prediction of such an ordinal forest is one of the M ordered classes which has been predicted the most by the respective trees of the forest. On the other hand, the probability prediction is obtained as a relative frequency of trees predicting the particular class. Such predicted choice probabilities are analyzed in the conducted Monte Carlo study in Section 2.5 of the main text. Further, the so-called naive forest corresponds to the ordinal forest with omitting the above described optimization procedure.

2.B Simulation Study

2.B.1 Main Simulation Results

In the following Tables 2.B.1, 2.B.2, 2.B.3, 2.B.4 and 2.B.5 are summarized the simulation results presented in Section 2.5.3 of the main text. Each table specifies the particular simulation design as follows: the column *Class* indicates the number of outcome classes, *Dim.* specifies the dimension, *DGP* characterizes the data generating process as defined in the main text and *Statistic* contains summary statistics of the simulation results. In particular, the mean of the respective accuracy measure and its standard deviation. Furthermore, rows *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods. The alternative hypothesis is that the mean of the *Ordered Forest* is less than the mean of the other method to test if the *Ordered Forest* achieves significantly lower prediction error than the other considered methods. Furthermore, Figures 2.B.1, 2.B.2, 2.B.3 and 2.B.4 complement the results presented in Section 2.5.3 of the main text for the simulations with the increased sample size.

2.B.1.1 ARPS: Sample Size = 200

Table 2.B.1: Simulation results: Accuracy Measure = ARPS & Sample Size = 200

Simulation Design				Comparison of Methods							
Class	Dim.	DGP	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
3	Low	Simple	mean	0.0097	0.0765	0.0755	0.0625	0.0609	0.0954	0.0619	0.0954
			st.dev.	0.0042	0.0056	0.0055	0.0018	0.0020	0.0011	0.0019	0.0012
			t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0002	0.0000
			wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
3	Low	Complex	mean	0.1156	0.1044	0.1028	0.0593	0.0466	0.0748	0.0491	0.0760
			st.dev.	0.0047	0.0039	0.0038	0.0023	0.0026	0.0028	0.0024	0.0027
			t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
3	High	Simple	mean		0.1135	0.1139	0.1112	0.1140	0.1180	0.1139	0.1179
			st.dev.		0.0009	0.0010	0.0009	0.0008	0.0006	0.0008	0.0006
			t-test		1.0000	0.7676	1.0000		0.0000	0.7268	0.0000
			wilcox-test		0.9999	0.8438	1.0000		0.0000	0.7191	0.0000
3	High	Complex	mean		0.1476	0.1474	0.1156	0.1102	0.1316	0.1110	0.1317
			st.dev.		0.0013	0.0010	0.0041	0.0029	0.0031	0.0029	0.0031
			t-test		0.0000	0.0000	0.0000		0.0000	0.0287	0.0000
			wilcox-test		0.0000	0.0000	0.0000		0.0000	0.0272	0.0000
6	Low	Simple	mean	0.0062	0.0687	0.0665	0.0554	0.0544	0.0833	0.0577	0.0872
			st.dev.	0.0020	0.0048	0.0050	0.0012	0.0014	0.0009	0.0016	0.0010
			t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
6	Low	Complex	mean	0.1122	0.1093	0.1058	0.0574	0.0452	0.0719	0.0536	0.0842
			st.dev.	0.0040	0.0045	0.0044	0.0017	0.0020	0.0022	0.0021	0.0024
			t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
6	High	Simple	mean		0.0974	0.0972	0.0951	0.0983	0.1012	0.0998	0.1016
			st.dev.		0.0006	0.0006	0.0006	0.0005	0.0004	0.0005	0.0004
			t-test		1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
			wilcox-test		1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
6	High	Complex	mean		0.0927	0.0927	0.0766	0.0772	0.0882	0.0898	0.0952
			st.dev.		0.0006	0.0005	0.0020	0.0016	0.0014	0.0018	0.0006
			t-test		0.0000	0.0000	0.9878		0.0000	0.0000	0.0000
			wilcox-test		0.0000	0.0000	0.9887		0.0000	0.0000	0.0000
9	Low	Simple	mean	0.0054	0.0653	0.0629	0.0528	0.0519	0.0789	0.0569	0.0850
			st.dev.	0.0018	0.0042	0.0042	0.0012	0.0014	0.0009	0.0017	0.0009
			t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
9	Low	Complex	mean	0.0973	0.0912	0.0887	0.0515	0.0421	0.0647	0.0537	0.0845
			st.dev.	0.0031	0.0033	0.0032	0.0015	0.0016	0.0019	0.0018	0.0017
			t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
9	High	Simple	mean		0.0921	0.0918	0.0900	0.0931	0.0959	0.0955	0.0964
			st.dev.		0.0006	0.0006	0.0006	0.0005	0.0003	0.0004	0.0003
			t-test		1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
			wilcox-test		1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
9	High	Complex	mean		0.1007	0.1004	0.0817	0.0819	0.0945	0.0997	0.1036
			st.dev.		0.0007	0.0007	0.0020	0.0017	0.0015	0.0019	0.0006
			t-test		0.0000	0.0000	0.7875		0.0000	0.0000	0.0000
			wilcox-test		0.0000	0.0000	0.8473		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the RPS based on 100 simulation replications for the sample size of 200 observations. The first column denotes the number of outcome classes. Columns 2 and 3 specify the dimension and the DGP, respectively. The fourth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.1.2 AMSE: Sample Size = 200

Table 2.B.2: Simulation results: Accuracy Measure = AMSE & Sample Size = 200

Simulation Design				**Comparison of Methods**							
Class	Dim.	DGP	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
3	Low	Simple	mean	0.0103	0.0669	0.0682	0.0565	0.0587	0.0800	0.0587	0.0800
			st.dev.	0.0044	0.0041	0.0044	0.0015	0.0022	0.0009	0.0016	0.0010
			t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.3900	0.0000
			wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.2614	0.0000
3	Low	Complex	mean	0.1081	0.0985	0.0965	0.0637	0.0543	0.0752	0.0572	0.0768
			st.dev.	0.0039	0.0034	0.0029	0.0020	0.0026	0.0021	0.0022	0.0019
			t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
3	High	Simple	mean		0.0923	0.0931	0.0908	0.0930	0.0952	0.0926	0.0952
			st.dev.		0.0008	0.0013	0.0009	0.0009	0.0007	0.0007	0.0007
			t-test		1.0000	0.2408	1.0000		0.0000	0.9980	0.0000
			wilcox-test		1.0000	0.5433	1.0000		0.0000	0.9977	0.0000
3	High	Complex	mean		0.1081	0.1079	0.0863	0.0828	0.0970	0.0834	0.0971
			st.dev.		0.0012	0.0009	0.0028	0.0019	0.0021	0.0020	0.0021
			t-test		0.0000	0.0000	0.0000		0.0000	0.0264	0.0000
			wilcox-test		0.0000	0.0000	0.0000		0.0000	0.0364	0.0000
6	Low	Simple	mean	0.0043	0.0284	0.0283	0.0248	0.0291	0.0324	0.0287	0.0327
			st.dev.	0.0014	0.0012	0.0018	0.0007	0.0010	0.0005	0.0008	0.0005
			t-test	1.0000	1.0000	0.9998	1.0000		0.0000	0.9958	0.0000
			wilcox-test	1.0000	1.0000	1.0000	1.0000		0.0000	0.9953	0.0000
6	Low	Complex	mean	0.0433	0.0438	0.0413	0.0270	0.0260	0.0314	0.0274	0.0339
			st.dev.	0.0014	0.0017	0.0014	0.0008	0.0011	0.0009	0.0010	0.0008
			t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
6	High	Simple	mean		0.0352	0.0352	0.0347	0.0361	0.0361	0.0360	0.0361
			st.dev.		0.0003	0.0004	0.0004	0.0004	0.0004	0.0003	0.0004
			t-test		1.0000	1.0000	1.0000		0.8112	0.9994	0.6394
			wilcox-test		1.0000	1.0000	1.0000		0.8788	0.9989	0.6579
6	High	Complex	mean		0.0383	0.0386	0.0343	0.0350	0.0367	0.0378	0.0387
			st.dev.		0.0003	0.0004	0.0006	0.0005	0.0005	0.0005	0.0004
			t-test		0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
			wilcox-test		0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
9	Low	Simple	mean	0.0025	0.0150	0.0149	0.0134	0.0170	0.0170	0.0168	0.0172
			st.dev.	0.0008	0.0005	0.0007	0.0004	0.0006	0.0003	0.0005	0.0002
			t-test	1.0000	1.0000	1.0000	1.0000		0.5492	0.9993	0.0040
			wilcox-test	1.0000	1.0000	1.0000	1.0000		0.3269	0.9985	0.0003
9	Low	Complex	mean	0.0203	0.0194	0.0190	0.0142	0.0159	0.0161	0.0162	0.0179
			st.dev.	0.0006	0.0006	0.0005	0.0003	0.0005	0.0003	0.0004	0.0003
			t-test	0.0000	0.0000	0.0000	1.0000		0.0006	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	1.0000		0.0004	0.0000	0.0000
9	High	Simple	mean		0.0180	0.0181	0.0178	0.0189	0.0185	0.0188	0.0185
			st.dev.		0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
			t-test		1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
			wilcox-test		1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
9	High	Complex	mean		0.0200	0.0200	0.0178	0.0187	0.0193	0.0201	0.0202
			st.dev.		0.0002	0.0002	0.0003	0.0003	0.0003	0.0003	0.0002
			t-test		0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
			wilcox-test		0.0000	0.0000	1.0000		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the MSE based on 100 simulation replications for the sample size of 200 observations. The first column denotes the number of outcome classes. Columns 2 and 3 specify the dimension and the DGP, respectively. The fourth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.1.3 ARPS: Sample Size = 800

Table 2.B.3: Simulation results: Accuracy Measure = ARPS & Sample Size = 800

Simulation Design				Comparison of Methods							
Class	Dim.	DGP	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
3	Low	Simple	mean	0.0023	0.0701	0.0685	0.0484	0.0466	0.0799	0.0483	0.0803
			st.dev.	0.0009	0.0043	0.0045	0.0007	0.0009	0.0008	0.0008	0.0008
			t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
3	Low	Complex	mean	0.0849	0.0828	0.0813	0.0394	0.0323	0.0495	0.0344	0.0516
			st.dev.	0.0009	0.0024	0.0026	0.0012	0.0009	0.0013	0.0010	0.0012
			t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
3	High	Simple	mean		0.1055	0.1055	0.1017	0.1044	0.1136	0.1047	0.1136
			st.dev.		0.0007	0.0007	0.0006	0.0005	0.0004	0.0005	0.0003
			t-test		0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
			wilcox-test		0.0000	0.0000	1.0000		0.0000	0.0001	0.0000
3	High	Complex	mean		0.0944	0.0949	0.0681	0.0616	0.0738	0.0635	0.0770
			st.dev.		0.0007	0.0010	0.0010	0.0010	0.0010	0.0009	0.0011
			t-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
6	Low	Simple	mean	0.0015	0.0619	0.0595	0.0435	0.0417	0.0702	0.0443	0.0748
			st.dev.	0.0005	0.0037	0.0039	0.0006	0.0007	0.0007	0.0006	0.0006
			t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
6	Low	Complex	mean	0.0947	0.1020	0.0986	0.0408	0.0330	0.0510	0.0384	0.0608
			st.dev.	0.0009	0.0031	0.0031	0.0009	0.0007	0.0010	0.0008	0.0012
			t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
6	High	Simple	mean		0.0905	0.0898	0.0874	0.0905	0.0978	0.0940	0.0995
			st.dev.		0.0006	0.0005	0.0004	0.0003	0.0002	0.0004	0.0002
			t-test		0.6597	1.0000	1.0000		0.0000	0.0000	0.0000
			wilcox-test		0.8939	1.0000	1.0000		0.0000	0.0000	0.0000
6	High	Complex	mean		0.1069	0.1060	0.0774	0.0698	0.0840	0.0781	0.0931
			st.dev.		0.0007	0.0007	0.0010	0.0009	0.0010	0.0013	0.0011
			t-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
9	Low	Simple	mean	0.0013	0.0603	0.0570	0.0417	0.0400	0.0668	0.0432	0.0741
			st.dev.	0.0004	0.0032	0.0035	0.0006	0.0006	0.0006	0.0007	0.0006
			t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
9	Low	Complex	mean	0.0837	0.0867	0.0836	0.0368	0.0305	0.0459	0.0375	0.0614
			st.dev.	0.0009	0.0023	0.0027	0.0008	0.0006	0.0008	0.0006	0.0009
			t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
9	High	Simple	mean		0.0857	0.0847	0.0826	0.0860	0.0927	0.0920	0.0949
			st.dev.		0.0005	0.0005	0.0004	0.0003	0.0002	0.0004	0.0001
			t-test		1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
			wilcox-test		1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
9	High	Complex	mean		0.0956	0.0947	0.0708	0.0648	0.0773	0.0781	0.0933
			st.dev.		0.0006	0.0007	0.0007	0.0006	0.0007	0.0011	0.0009
			t-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the RPS based on 100 simulation replications for the sample size of 800 observations. The first column denotes the number of outcome classes. Columns 2 and 3 specify the dimension and the DGP, respectively. The fourth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.1.4 AMSE: Sample Size = 800

Table 2.B.4: Simulation results: Accuracy Measure = AMSE & Sample Size = 800

Simulation Design				Comparison of Methods							
Class	Dim.	DGP	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
3	Low	Simple	mean	0.0025	0.0618	0.0624	0.0451	0.0461	0.0688	0.0472	0.0691
			st.dev.	0.0009	0.0032	0.0036	0.0006	0.0010	0.0006	0.0007	0.0006
			t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
			wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
3	Low	Complex	mean	0.0875	0.0848	0.0834	0.0482	0.0414	0.0574	0.0439	0.0602
			st.dev.	0.0008	0.0020	0.0020	0.0011	0.0010	0.0011	0.0011	0.0010
			t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
3	High	Simple	mean		0.0866	0.0870	0.0840	0.0861	0.0920	0.0861	0.0920
			st.dev.		0.0005	0.0007	0.0005	0.0005	0.0003	0.0004	0.0003
			t-test		0.0000	0.0000	1.0000		0.0000	0.5234	0.0000
			wilcox-test		0.0000	0.0000	1.0000		0.0000	0.4713	0.0000
3	High	Complex	mean		0.0969	0.0977	0.0717	0.0656	0.0749	0.0675	0.0789
			st.dev.		0.0006	0.0008	0.0009	0.0010	0.0009	0.0009	0.0011
			t-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
6	Low	Simple	mean	0.0010	0.0260	0.0260	0.0206	0.0231	0.0287	0.0234	0.0292
			st.dev.	0.0003	0.0010	0.0014	0.0003	0.0005	0.0002	0.0003	0.0002
			t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
			wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
6	Low	Complex	mean	0.0376	0.0406	0.0384	0.0219	0.0208	0.0257	0.0221	0.0280
			st.dev.	0.0003	0.0010	0.0009	0.0004	0.0004	0.0004	0.0004	0.0003
			t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
6	High	Simple	mean		0.0333	0.0332	0.0325	0.0339	0.0350	0.0343	0.0353
			st.dev.		0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001
			t-test		1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
			wilcox-test		1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
6	High	Complex	mean		0.0404	0.0399	0.0308	0.0287	0.0325	0.0313	0.0352
			st.dev.		0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0003
			t-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
9	Low	Simple	mean	0.0006	0.0140	0.0138	0.0113	0.0136	0.0153	0.0135	0.0156
			st.dev.	0.0002	0.0004	0.0006	0.0002	0.0003	0.0001	0.0002	0.0001
			t-test	1.0000	0.0000	0.0121	1.0000		0.0000	1.0000	0.0000
			wilcox-test	1.0000	0.0000	0.0241	1.0000		0.0000	1.0000	0.0000
9	Low	Complex	mean	0.0178	0.0187	0.0181	0.0114	0.0124	0.0132	0.0126	0.0149
			st.dev.	0.0001	0.0004	0.0005	0.0002	0.0003	0.0002	0.0002	0.0001
			t-test	0.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
9	High	Simple	mean		0.0171	0.0171	0.0167	0.0179	0.0179	0.0184	0.0181
			st.dev.		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
			t-test		1.0000	1.0000	1.0000		0.9803	0.0000	0.0000
			wilcox-test		1.0000	1.0000	1.0000		0.9670	0.0000	0.0000
9	High	Complex	mean		0.0191	0.0191	0.0162	0.0161	0.0170	0.0176	0.0187
			st.dev.		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
			t-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
			wilcox-test		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the MSE based on 100 simulation replications for the sample size of 800 observations. The first column denotes the number of outcome classes. Columns 2 and 3 specify the dimension and the DGP, respectively. The fourth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

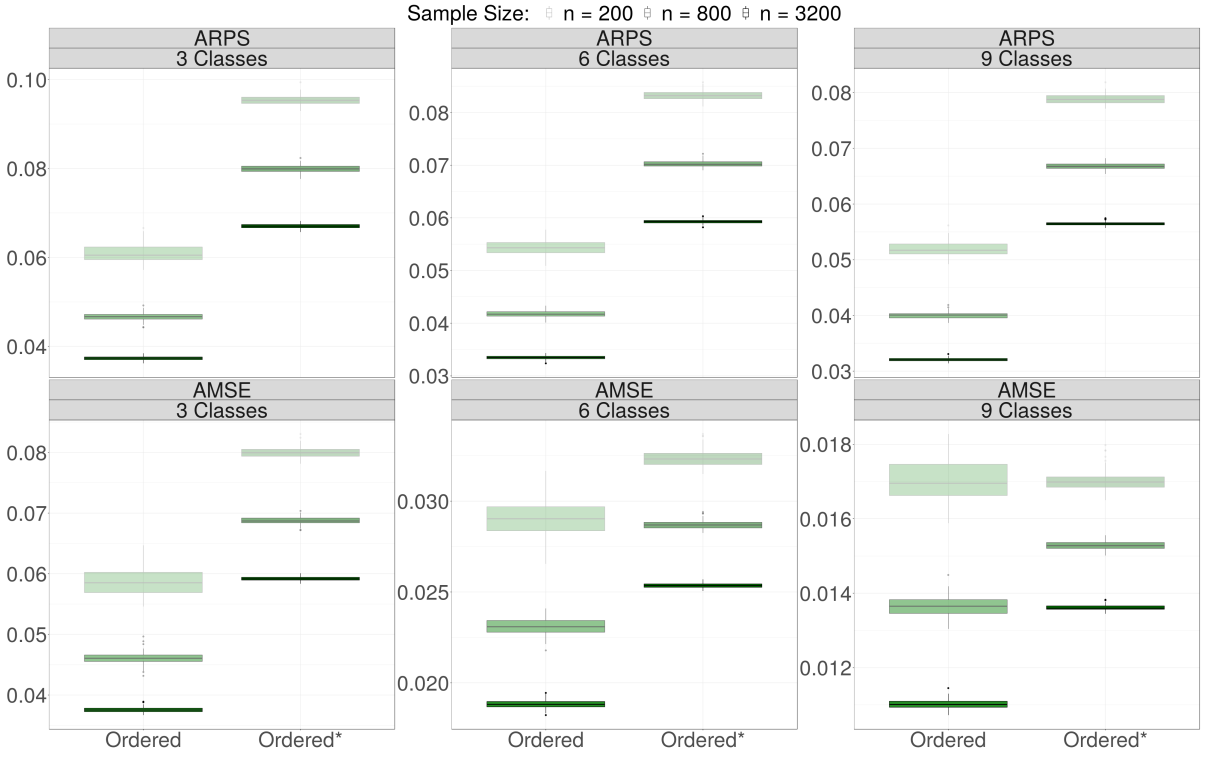
2.B.1.5 ARPS & AMSE: Sample Size = 3200

Table 2.B.5: Simulation results: Accuracy Measure = ARPS/AMSE & Sample Size = 3200

Simulation Design				ARPS		AMSE	
Class	Dim.	DGP	Statistic	Ordered	Ordered*	Ordered	Ordered*
3	Low	Simple	mean	0.0373	0.0670	0.0376	0.0591
			st.dev.	0.0004	0.0005	0.0005	0.0004
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000
3	Low	Complex	mean	0.0285	0.0415	0.0243	0.0336
			st.dev.	0.0004	0.0005	0.0003	0.0004
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000
3	High	Simple	mean	0.0956	0.1069	0.0798	0.0872
			st.dev.	0.0003	0.0002	0.0002	0.0002
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000
3	High	Complex	mean	0.0498	0.0620	0.0557	0.0653
			st.dev.	0.0004	0.0005	0.0005	0.0005
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000
6	Low	Simple	mean	0.0335	0.0593	0.0188	0.0253
			st.dev.	0.0004	0.0004	0.0002	0.0001
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000
6	Low	Complex	mean	0.0255	0.0367	0.0162	0.0197
			st.dev.	0.0003	0.0004	0.0002	0.0002
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000
6	High	Simple	mean	0.0825	0.0923	0.0314	0.0335
			st.dev.	0.0002	0.0002	0.0001	0.0000
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000
6	High	Complex	mean	0.0526	0.0656	0.0264	0.0292
			st.dev.	0.0004	0.0004	0.0002	0.0001
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000
9	Low	Simple	mean	0.0321	0.0565	0.0110	0.0136
			st.dev.	0.0003	0.0003	0.0001	0.0001
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000
9	Low	Complex	mean	0.0244	0.0350	0.0098	0.0110
			st.dev.	0.0002	0.0003	0.0001	0.0001
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000
9	High	Simple	mean	0.0783	0.0875	0.0165	0.0172
			st.dev.	0.0002	0.0002	0.0000	0.0000
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000
9	High	Complex	mean	0.0559	0.0697	0.0145	0.0160
			st.dev.	0.0004	0.0004	0.0001	0.0001
			t-test		0.0000		0.0000
			wilcox-test		0.0000		0.0000

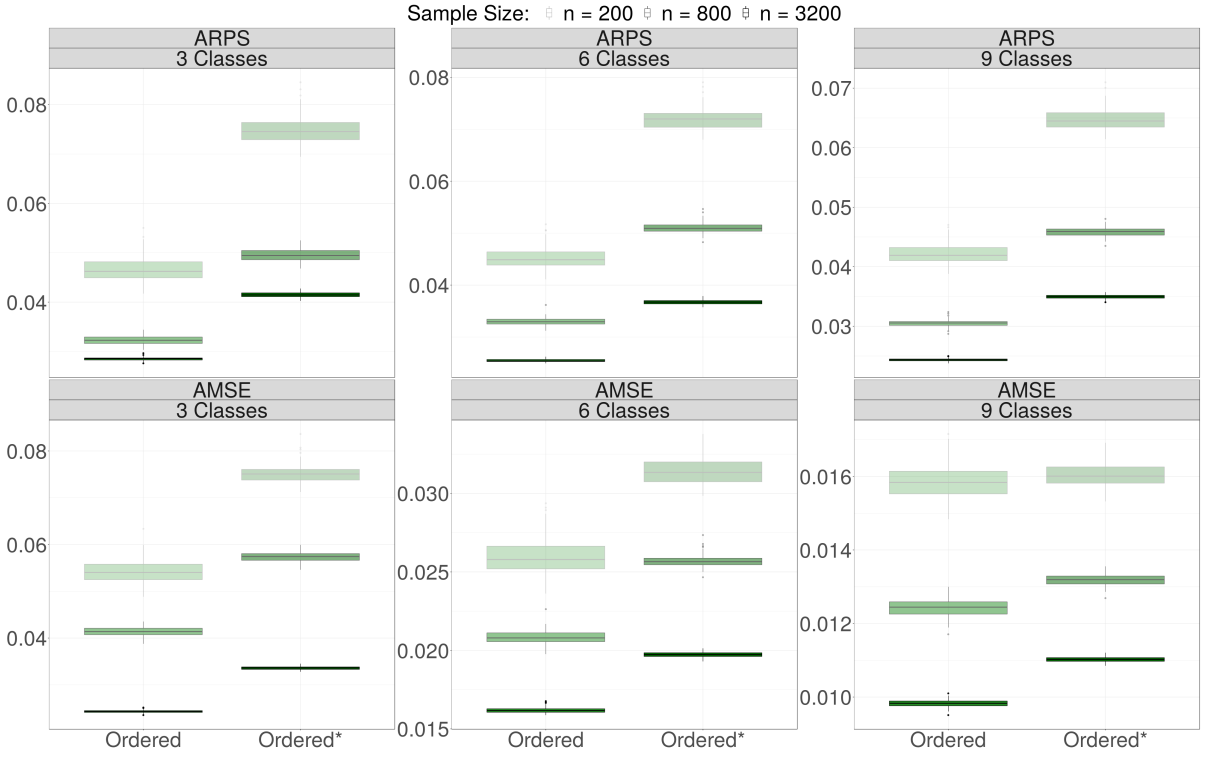
Notes: Table reports the average measures of the RPS and MSE based on 100 simulation replications for the sample size of 3200 observations. The first column denotes the number of outcome classes. Columns 2 and 3 specify the dimension and the DGP, respectively. The fourth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and the honest version of the *Ordered Forest*.

Figure 2.B.1: Ordered Forest Simulation Results: Simple DGP & Low Dimension



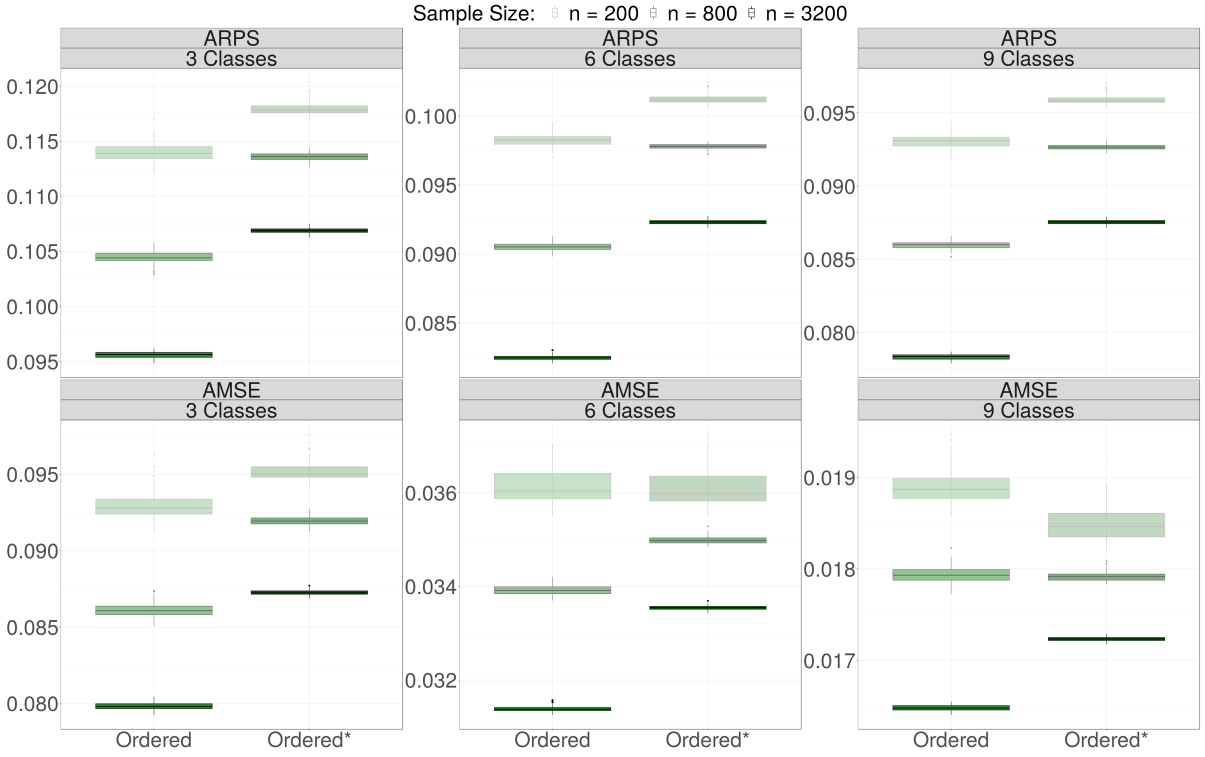
Note: Figure summarizes the prediction accuracy results based on 100 simulation replications. The upper panel contains the ARPS and the lower panel contains the AMSE. The boxplots show the median and the interquartile range of the respective measure. The transparent boxplots denote the results for the small sample size, the semi-transparent ones denote the medium sample size, while the bold boxplots denote the results for the big sample size. From left to right the results for 3, 6, and 9 outcome classes are displayed.

Figure 2.B.2: Ordered Forest Simulation Results: Complex DGP & Low Dimension



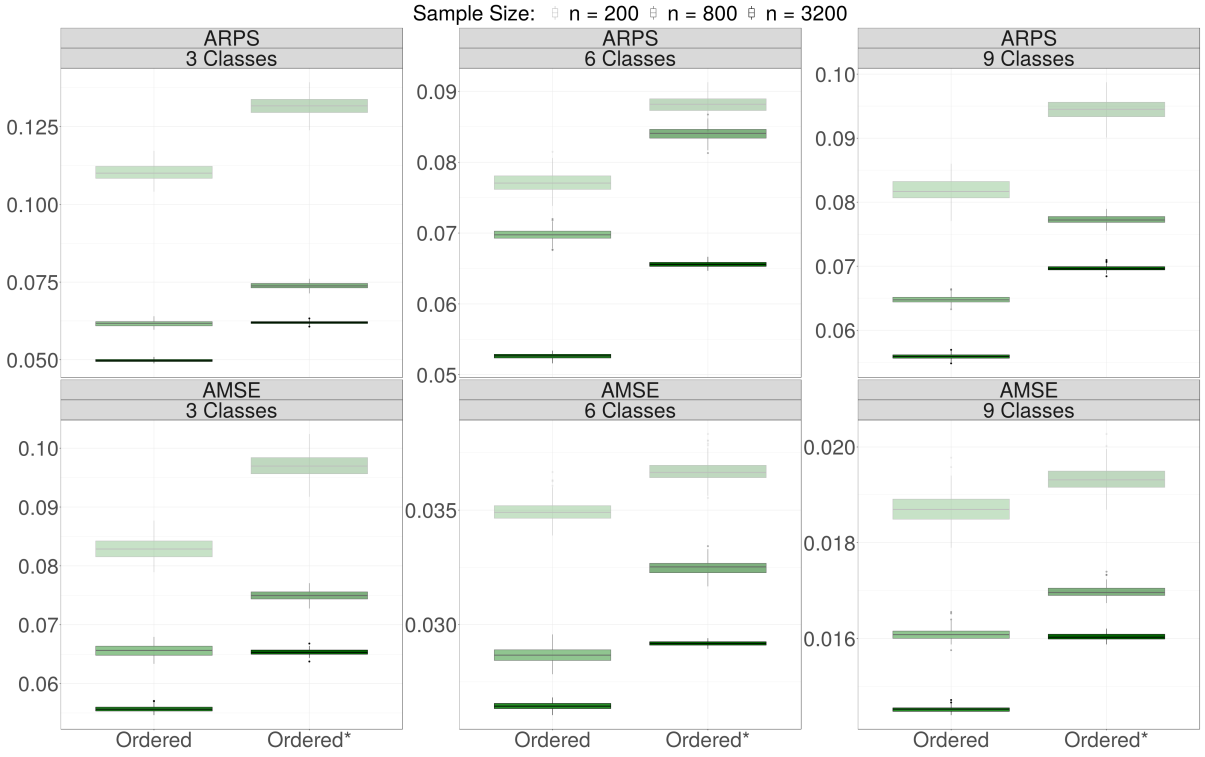
Note: Figure summarizes the prediction accuracy results based on 100 simulation replications. The upper panel contains the ARPS and the lower panel contains the AMSE. The boxplots show the median and the interquartile range of the respective measure. The transparent boxplots denote the results for the small sample size, the semi-transparent ones denote the medium sample size, while the bold boxplots denote the results for the big sample size. From left to right the results for 3, 6, and 9 outcome classes are displayed.

Figure 2.B.3: Ordered Forest Simulation Results: Simple DGP & High Dimension



Note: Figure summarizes the prediction accuracy results based on 100 simulation replications. The upper panel contains the ARPS and the lower panel contains the AMSE. The boxplots show the median and the interquartile range of the respective measure. The transparent boxplots denote the results for the small sample size, the semi-transparent ones denote the medium sample size, while the bold boxplots denote the results for the big sample size. From left to right the results for 3, 6, and 9 outcome classes are displayed.

Figure 2.B.4: Ordered Forest Simulation Results: Complex DGP & High Dimension



Note: Figure summarizes the prediction accuracy results based on 100 simulation replications. The upper panel contains the ARPS and the lower panel contains the AMSE. The boxplots show the median and the interquartile range of the respective measure. The transparent boxplots denote the results for the small sample size, the semi-transparent ones denote the medium sample size, while the bold boxplots denote the results for the big sample size. From left to right the results for 3, 6, and 9 outcome classes are displayed.

2.B.2 Complete Simulation Results

Tables 2.B.6 to 2.B.17 below summarize the simulation results for all 72 different DGPs, complementing the main results presented in Section 2.5.3 of the main text. Each table specifies the particular simulation design as follows: the first column *DGP* provides the identifier for the data generating process. Columns 2 to 5 specify the particular characteristics of the respective DGP, namely if the DGP features additional noise variables (*noise*), 15 in the low-dimensional case and 1000 in the high-dimensional case, nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* contains summary statistics of the simulation results. In particular, the mean of the respective accuracy measure (*mean*) and its standard deviation (*st.dev.*). Furthermore, rows *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods. The alternative hypothesis is that the mean of the *Ordered Forest* is less than the mean of the other method to test if the *Ordered Forest* achieves significantly lower prediction error than the other considered methods.

2.B.2.1 ARPS: Low Dimension with 3 Classes

Table 2.B.6: Simulation Results: Accuracy Measure = ARPS & Low Dimension with 3 Classes

Simulation Design						Comparison of Methods							
DGP	noise	nonlin	multi	random	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
1	✗	✗	✗	✗	mean	0.0097	0.0765	0.0755	0.0625	0.0609	0.0954	0.0619	0.0954
					st.dev.	0.0042	0.0056	0.0055	0.0018	0.0020	0.0011	0.0019	0.0012
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0002	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
2	✓	✗	✗	✗	mean	0.0216	0.0840	0.0832	0.0738	0.0754	0.1041	0.0763	0.1041
					st.dev.	0.0054	0.0046	0.0048	0.0015	0.0016	0.0013	0.0016	0.0013
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0001	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0001	0.0000
3	✗	✓	✗	✗	mean	0.0904	0.0715	0.0726	0.0688	0.0681	0.0824	0.0672	0.0824
					st.dev.	0.0045	0.0031	0.0033	0.0021	0.0022	0.0013	0.0020	0.0013
					t-test	0.0000	0.0000	0.0000	0.0132		0.0000	0.9988	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0070		0.0000	0.9976	0.0000
4	✗	✗	✓	✗	mean	0.0097	0.1236	0.1194	0.0316	0.0297	0.0449	0.0297	0.0493
					st.dev.	0.0031	0.0079	0.0079	0.0015	0.0015	0.0013	0.0014	0.0013
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.4099	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.3721	0.0000
5	✗	✗	✗	✓	mean	0.0104	0.0730	0.0698	0.0611	0.0594	0.0942	0.0607	0.0948
					st.dev.	0.0035	0.0072	0.0066	0.0017	0.0020	0.0015	0.0021	0.0016
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
6	✓	✓	✗	✗	mean	0.1052	0.0772	0.0781	0.0763	0.0768	0.0863	0.0759	0.0862
					st.dev.	0.0066	0.0025	0.0030	0.0021	0.0020	0.0011	0.0019	0.0011
					t-test	0.0000	0.1612	0.0004	0.9717		0.0000	0.9998	0.0000
					wilcox-test	0.0000	0.1979	0.0004	0.9589		0.0000	0.9996	0.0000
7	✓	✗	✓	✗	mean	0.0221	0.1349	0.1321	0.0344	0.0335	0.0502	0.0353	0.0569
					st.dev.	0.0064	0.0060	0.0057	0.0013	0.0011	0.0021	0.0013	0.0022
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
8	✓	✗	✗	✓	mean	0.0196	0.0750	0.0753	0.0669	0.0694	0.0938	0.0699	0.0940
					st.dev.	0.0056	0.0036	0.0040	0.0017	0.0019	0.0010	0.0018	0.0010
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0412	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0339	0.0000
9	✗	✓	✓	✗	mean	0.1116	0.1204	0.1170	0.0486	0.0401	0.0706	0.0422	0.0722
					st.dev.	0.0030	0.0077	0.0075	0.0022	0.0021	0.0025	0.0021	0.0024
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
10	✗	✓	✗	✓	mean	0.0905	0.0703	0.0693	0.0673	0.0668	0.0808	0.0672	0.0809
					st.dev.	0.0047	0.0042	0.0042	0.0023	0.0023	0.0013	0.0023	0.0013
					t-test	0.0000	0.0000	0.0000	0.0650		0.0000	0.0923	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0862		0.0000	0.0921	0.0000
11	✗	✗	✓	✓	mean	0.0111	0.1299	0.1284	0.0312	0.0295	0.0428	0.0298	0.0463
					st.dev.	0.0042	0.0115	0.0121	0.0017	0.0017	0.0014	0.0016	0.0015
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0868	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0901	0.0000
12	✓	✓	✓	✗	mean	0.1297	0.1232	0.1209	0.0639	0.0483	0.0809	0.0512	0.0819
					st.dev.	0.0051	0.0058	0.0055	0.0024	0.0025	0.0028	0.0023	0.0027
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
13	✓	✓	✗	✓	mean	0.0915	0.0682	0.0697	0.0675	0.0689	0.0764	0.0677	0.0764
					st.dev.	0.0063	0.0022	0.0024	0.0020	0.0020	0.0011	0.0019	0.0011
					t-test	0.0000	0.9877	0.0036	1.0000		0.0000	1.0000	0.0000
					wilcox-test	0.0000	0.9813	0.0032	1.0000		0.0000	1.0000	0.0000
14	✓	✗	✓	✓	mean	0.0235	0.1219	0.1194	0.0319	0.0312	0.0468	0.0324	0.0524
					st.dev.	0.0068	0.0052	0.0050	0.0015	0.0014	0.0020	0.0014	0.0021
					t-test	1.0000	0.0000	0.0000	0.0012		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0008		0.0000	0.0000	0.0000
15	✗	✓	✓	✓	mean	0.1118	0.1222	0.1204	0.0482	0.0396	0.0688	0.0411	0.0712
					st.dev.	0.0042	0.0087	0.0092	0.0024	0.0025	0.0026	0.0024	0.0026
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
16	✓	✓	✓	✓	mean	0.1156	0.1044	0.1028	0.0593	0.0466	0.0748	0.0491	0.0760
					st.dev.	0.0047	0.0039	0.0038	0.0023	0.0026	0.0028	0.0024	0.0027
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the RPS based on 100 simulation replications for the sample size of 200 observations with 3 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 15 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.2.2 ARPS: Low Dimension with 6 Classes

Table 2.B.7: Simulation Results: Accuracy Measure = ARPS & Low Dimension with 6 Classes

Simulation Design						Comparison of Methods							
DGP	noise	nonlin	multi	random	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
17	✗	✗	✗	✗	mean	0.0062	0.0687	0.0665	0.0554	0.0544	0.0833	0.0577	0.0872
					st.dev.	0.0020	0.0048	0.0050	0.0012	0.0014	0.0009	0.0016	0.0010
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
18	✓	✗	✗	✗	mean	0.0129	0.0726	0.0708	0.0645	0.0669	0.0901	0.0709	0.0932
					st.dev.	0.0034	0.0026	0.0028	0.0013	0.0012	0.0007	0.0013	0.0007
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
19	✗	✓	✗	✗	mean	0.0749	0.0610	0.0608	0.0585	0.0593	0.0707	0.0597	0.0725
					st.dev.	0.0022	0.0030	0.0027	0.0016	0.0018	0.0010	0.0020	0.0010
					t-test	0.0000	0.0000	0.0000	0.9996		0.0000	0.0947	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.9995		0.0000	0.0966	0.0000
20	✗	✗	✓	✗	mean	0.0059	0.1111	0.1071	0.0285	0.0273	0.0407	0.0292	0.0539
					st.dev.	0.0016	0.0050	0.0061	0.0010	0.0009	0.0011	0.0010	0.0015
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
21	✗	✗	✗	✓	mean	0.0062	0.0670	0.0648	0.0544	0.0537	0.0816	0.0569	0.0853
					st.dev.	0.0022	0.0044	0.0044	0.0013	0.0014	0.0009	0.0015	0.0009
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
22	✓	✓	✗	✗	mean	0.0853	0.0650	0.0651	0.0644	0.0664	0.0735	0.0675	0.0748
					st.dev.	0.0049	0.0022	0.0022	0.0016	0.0014	0.0008	0.0014	0.0006
					t-test	0.0000	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
23	✓	✗	✓	✗	mean	0.0106	0.1177	0.1145	0.0313	0.0307	0.0462	0.0377	0.0640
					st.dev.	0.0028	0.0038	0.0049	0.0010	0.0008	0.0014	0.0011	0.0018
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
24	✓	✗	✗	✓	mean	0.0148	0.0745	0.0722	0.0655	0.0677	0.0919	0.0718	0.0946
					st.dev.	0.0040	0.0032	0.0029	0.0012	0.0013	0.0009	0.0014	0.0008
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
25	✗	✓	✓	✗	mean	0.0952	0.0995	0.0961	0.0439	0.0372	0.0630	0.0418	0.0747
					st.dev.	0.0020	0.0041	0.0043	0.0016	0.0016	0.0017	0.0016	0.0020
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
26	✗	✓	✗	✓	mean	0.0733	0.0590	0.0594	0.0573	0.0582	0.0691	0.0586	0.0707
					st.dev.	0.0024	0.0021	0.0020	0.0015	0.0015	0.0010	0.0015	0.0009
					t-test	0.0000	0.0017	0.0000	1.0000		0.0000	0.0660	0.0000
					wilcox-test	0.0000	0.0041	0.0000	1.0000		0.0000	0.0809	0.0000
27	✗	✗	✓	✓	mean	0.0053	0.1069	0.1046	0.0278	0.0266	0.0401	0.0286	0.0533
					st.dev.	0.0014	0.0048	0.0056	0.0010	0.0009	0.0011	0.0009	0.0015
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
28	✓	✓	✓	✗	mean	0.1090	0.1022	0.1001	0.0564	0.0447	0.0709	0.0527	0.0843
					st.dev.	0.0041	0.0031	0.0030	0.0015	0.0018	0.0020	0.0018	0.0024
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
29	✓	✓	✗	✓	mean	0.0881	0.0666	0.0662	0.0658	0.0676	0.0751	0.0697	0.0764
					st.dev.	0.0051	0.0024	0.0022	0.0016	0.0015	0.0008	0.0015	0.0006
					t-test	0.0000	0.9997	1.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
30	✓	✗	✓	✓	mean	0.0118	0.1214	0.1161	0.0317	0.0309	0.0469	0.0378	0.0642
					st.dev.	0.0032	0.0046	0.0055	0.0009	0.0008	0.0014	0.0012	0.0019
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
31	✗	✓	✓	✓	mean	0.0931	0.0956	0.0925	0.0434	0.0368	0.0619	0.0414	0.0731
					st.dev.	0.0019	0.0044	0.0045	0.0015	0.0014	0.0016	0.0014	0.0020
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
32	✓	✓	✓	✓	mean	0.1122	0.1093	0.1058	0.0574	0.0452	0.0719	0.0536	0.0842
					st.dev.	0.0040	0.0045	0.0044	0.0017	0.0020	0.0022	0.0021	0.0024
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the RPS based on 100 simulation replications for the sample size of 200 observations with 6 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 15 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.2.3 ARPS: Low Dimension with 9 Classes

Table 2.B.8: Simulation Results: Accuracy Measure = ARPS & Low Dimension with 9 Classes

Simulation Design						Comparison of Methods							
DGP	noise	nonlin	multi	random	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
33	✗	✗	✗	✗	mean	0.0054	0.0653	0.0629	0.0528	0.0519	0.0789	0.0569	0.0850
					st.dev.	0.0018	0.0042	0.0042	0.0012	0.0014	0.0009	0.0017	0.0009
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
34	✓	✗	✗	✗	mean	0.0112	0.0693	0.0672	0.0609	0.0638	0.0855	0.0704	0.0901
					st.dev.	0.0028	0.0027	0.0027	0.0012	0.0012	0.0007	0.0013	0.0006
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
35	✗	✓	✗	✗	mean	0.0706	0.0573	0.0572	0.0555	0.0567	0.0669	0.0590	0.0698
					st.dev.	0.0023	0.0026	0.0027	0.0014	0.0015	0.0009	0.0016	0.0007
					t-test	0.0000	0.0220	0.0445	1.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0788	0.2389	1.0000		0.0000	0.0000	0.0000
36	✗	✗	✓	✗	mean	0.0052	0.1057	0.1047	0.0277	0.0263	0.0396	0.0303	0.0601
					st.dev.	0.0014	0.0046	0.0056	0.0009	0.0009	0.0010	0.0010	0.0014
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
37	✗	✗	✗	✓	mean	0.0054	0.0627	0.0608	0.0518	0.0511	0.0774	0.0558	0.0835
					st.dev.	0.0019	0.0036	0.0035	0.0012	0.0014	0.0009	0.0016	0.0010
					t-test	1.0000	0.0000	0.0000	0.0001		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0002		0.0000	0.0000	0.0000
38	✓	✓	✗	✗	mean	0.0806	0.0607	0.0608	0.0606	0.0629	0.0695	0.0661	0.0715
					st.dev.	0.0036	0.0016	0.0018	0.0013	0.0012	0.0008	0.0014	0.0007
					t-test	0.0000	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
39	✓	✗	✓	✗	mean	0.0086	0.1122	0.1102	0.0301	0.0295	0.0443	0.0408	0.0710
					st.dev.	0.0017	0.0036	0.0041	0.0009	0.0008	0.0012	0.0011	0.0017
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
40	✓	✗	✗	✓	mean	0.0106	0.0663	0.0646	0.0586	0.0615	0.0820	0.0679	0.0866
					st.dev.	0.0028	0.0026	0.0026	0.0011	0.0012	0.0008	0.0012	0.0007
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
41	✗	✓	✓	✗	mean	0.0897	0.0929	0.0897	0.0417	0.0356	0.0596	0.0424	0.0776
					st.dev.	0.0017	0.0037	0.0038	0.0014	0.0013	0.0015	0.0014	0.0018
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
42	✗	✓	✗	✓	mean	0.0701	0.0565	0.0564	0.0545	0.0556	0.0657	0.0579	0.0685
					st.dev.	0.0025	0.0024	0.0024	0.0015	0.0014	0.0008	0.0016	0.0007
					t-test	0.0000	0.0006	0.0010	1.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0028	0.0066	1.0000		0.0000	0.0000	0.0000
43	✗	✗	✓	✓	mean	0.0051	0.1034	0.1025	0.0273	0.0258	0.0394	0.0298	0.0593
					st.dev.	0.0015	0.0040	0.0045	0.0008	0.0007	0.0010	0.0009	0.0014
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
44	✓	✓	✓	✗	mean	0.1018	0.0956	0.0933	0.0534	0.0432	0.0673	0.0550	0.0873
					st.dev.	0.0035	0.0031	0.0031	0.0013	0.0016	0.0017	0.0019	0.0021
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
45	✓	✓	✗	✓	mean	0.0763	0.0587	0.0588	0.0582	0.0605	0.0664	0.0638	0.0684
					st.dev.	0.0040	0.0019	0.0018	0.0014	0.0012	0.0007	0.0011	0.0006
					t-test	0.0000	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
46	✓	✗	✓	✓	mean	0.0084	0.1079	0.1066	0.0292	0.0286	0.0432	0.0391	0.0699
					st.dev.	0.0021	0.0034	0.0040	0.0008	0.0007	0.0012	0.0012	0.0017
					t-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
47	✗	✓	✓	✓	mean	0.0881	0.0915	0.0887	0.0411	0.0352	0.0588	0.0414	0.0765
					st.dev.	0.0017	0.0039	0.0041	0.0014	0.0012	0.0014	0.0014	0.0016
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
48	✓	✓	✓	✓	mean	0.0973	0.0912	0.0887	0.0515	0.0421	0.0647	0.0537	0.0845
					st.dev.	0.0031	0.0033	0.0032	0.0015	0.0016	0.0019	0.0018	0.0017
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the RPS based on 100 simulation replications for the sample size of 200 observations with 9 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 15 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.2.4 ARPS: High Dimension with 3 Classes

Table 2.B.9: Simulation Results: Accuracy Measure = ARPS & High Dimension with 3 Classes

Simulation Design						Comparison of Methods						
DGP	noise	nonlin	multi	random	Statistic	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
49	✓	✗	✗	✗	mean	0.1135	0.1139	0.1112	0.1140	0.1180	0.1139	0.1179
					st.dev.	0.0009	0.0010	0.0009	0.0008	0.0006	0.0008	0.0006
					t-test	1.0000	0.7676	1.0000		0.0000	0.7268	0.0000
					wilcox-test	0.9999	0.8438	1.0000		0.0000	0.7191	0.0000
50	✓	✓	✗	✗	mean	0.0896	0.0899	0.0901	0.0903	0.0907	0.0901	0.0907
					st.dev.	0.0008	0.0010	0.0008	0.0007	0.0007	0.0007	0.0006
					t-test	1.0000	0.9997	0.9840		0.0002	0.9973	0.0004
					wilcox-test	1.0000	1.0000	0.9929		0.0000	0.9989	0.0000
51	✓	✗	✓	✗	mean	0.1534	0.1529	0.0827	0.0766	0.1082	0.0867	0.1134
					st.dev.	0.0011	0.0012	0.0024	0.0025	0.0029	0.0024	0.0026
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
52	✓	✗	✗	✓	mean	0.1253	0.1252	0.1224	0.1248	0.1296	0.1250	0.1296
					st.dev.	0.0013	0.0013	0.0010	0.0009	0.0007	0.0009	0.0007
					t-test	0.0011	0.0115	1.0000		0.0000	0.1664	0.0000
					wilcox-test	0.0013	0.0140	1.0000		0.0000	0.1515	0.0000
53	✓	✓	✓	✗	mean	0.1299	0.1300	0.1048	0.1016	0.1200	0.1021	0.1202
					st.dev.	0.0011	0.0012	0.0034	0.0027	0.0026	0.0027	0.0025
					t-test	0.0000	0.0000	0.0000		0.0000	0.0674	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0494	0.0000
54	✓	✓	✗	✓	mean	0.0997	0.0996	0.0999	0.0998	0.1004	0.0997	0.1004
					st.dev.	0.0012	0.0013	0.0012	0.0012	0.0011	0.0011	0.0012
					t-test	0.5772	0.8438	0.3065		0.0000	0.6432	0.0000
					wilcox-test	0.6792	0.9705	0.2427		0.0000	0.7183	0.0000
55	✓	✗	✓	✓	mean	0.1678	0.1667	0.0862	0.0836	0.1167	0.0906	0.1195
					st.dev.	0.0015	0.0013	0.0026	0.0030	0.0029	0.0029	0.0029
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
56	✓	✓	✓	✓	mean	0.1476	0.1474	0.1156	0.1102	0.1316	0.1110	0.1317
					st.dev.	0.0013	0.0010	0.0041	0.0029	0.0031	0.0029	0.0031
					t-test	0.0000	0.0000	0.0000		0.0000	0.0287	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0272	0.0000

Notes: Table reports the average measures of the RPS based on 100 simulation replications for the sample size of 200 observations with 3 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 1000 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.2.5 ARPS: High Dimension with 6 Classes

Table 2.B.10: Simulation Results: Accuracy Measure = ARPS & High Dimension with 6 Classes

Simulation Design						Comparison of Methods						
DGP	noise	nonlin	multi	random	Statistic	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
57	✓	✗	✗	✗	mean	0.0974	0.0972	0.0951	0.0983	0.1012	0.0998	0.1016
					st.dev.	0.0006	0.0006	0.0006	0.0005	0.0004	0.0005	0.0004
					t-test	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
58	✓	✓	✗	✗	mean	0.0762	0.0762	0.0765	0.0773	0.0772	0.0776	0.0773
					st.dev.	0.0006	0.0006	0.0006	0.0005	0.0005	0.0004	0.0004
					t-test	1.0000	1.0000	1.0000		0.9803	0.0000	0.7833
					wilcox-test	1.0000	1.0000	1.0000		0.9838	0.0000	0.7449
59	✓	✗	✓	✗	mean	0.1336	0.1327	0.0747	0.0675	0.0968	0.0912	0.1152
					st.dev.	0.0008	0.0010	0.0013	0.0016	0.0015	0.0016	0.0017
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
60	✓	✗	✗	✓	mean	0.0845	0.0845	0.0826	0.0857	0.0880	0.0872	0.0883
					st.dev.	0.0005	0.0005	0.0006	0.0004	0.0003	0.0004	0.0003
					t-test	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
61	✓	✓	✓	✗	mean	0.1091	0.1088	0.0891	0.0885	0.1026	0.1010	0.1105
					st.dev.	0.0009	0.0008	0.0025	0.0021	0.0018	0.0023	0.0010
					t-test	0.0000	0.0000	0.0547		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0626		0.0000	0.0000	0.0000
62	✓	✓	✗	✓	mean	0.0658	0.0659	0.0660	0.0669	0.0665	0.0672	0.0666
					st.dev.	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005
					t-test	1.0000	1.0000	1.0000		1.0000	0.0006	0.9998
					wilcox-test	1.0000	1.0000	1.0000		1.0000	0.0000	1.0000
63	✓	✗	✓	✓	mean	0.1167	0.1163	0.0682	0.0606	0.0872	0.0820	0.1052
					st.dev.	0.0007	0.0008	0.0014	0.0016	0.0015	0.0018	0.0015
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
64	✓	✓	✓	✓	mean	0.0927	0.0927	0.0766	0.0772	0.0882	0.0898	0.0952
					st.dev.	0.0006	0.0005	0.0020	0.0016	0.0014	0.0018	0.0006
					t-test	0.0000	0.0000	0.9878		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.9887		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the RPS based on 100 simulation replications for the sample size of 200 observations with 6 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 1000 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.2.6 ARPS: High Dimension with 9 Classes

Table 2.B.11: Simulation Results: Accuracy Measure = ARPS & High Dimension with 9 Classes

Simulation Design						Comparison of Methods						
DGP	noise	nonlin	multi	random	Statistic	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
65	✓	✗	✗	✗	mean	0.0921	0.0918	0.0900	0.0931	0.0959	0.0955	0.0964
					st.dev.	0.0006	0.0006	0.0006	0.0005	0.0003	0.0004	0.0003
					t-test	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
66	✓	✓	✗	✗	mean	0.0721	0.0720	0.0724	0.0732	0.0730	0.0739	0.0731
					st.dev.	0.0006	0.0005	0.0006	0.0005	0.0004	0.0004	0.0004
					t-test	1.0000	1.0000	1.0000		0.9959	0.0000	0.8717
					wilcox-test	1.0000	1.0000	1.0000		0.9991	0.0000	0.9308
67	✓	✗	✓	✗	mean	0.1268	0.1260	0.0713	0.0648	0.0926	0.0979	0.1175
					st.dev.	0.0008	0.0009	0.0013	0.0013	0.0014	0.0017	0.0015
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
68	✓	✗	✗	✓	mean	0.0904	0.0902	0.0884	0.0915	0.0941	0.0937	0.0946
					st.dev.	0.0006	0.0006	0.0005	0.0005	0.0003	0.0004	0.0003
					t-test	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	1.0000	1.0000		0.0000	0.0000	0.0000
69	✓	✓	✓	✗	mean	0.1031	0.1028	0.0838	0.0838	0.0967	0.1024	0.1061
					st.dev.	0.0007	0.0007	0.0021	0.0017	0.0016	0.0016	0.0005
					t-test	0.0000	0.0000	0.4695		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.5044		0.0000	0.0000	0.0000
70	✓	✓	✗	✓	mean	0.0706	0.0707	0.0710	0.0718	0.0716	0.0724	0.0717
					st.dev.	0.0007	0.0007	0.0006	0.0006	0.0005	0.0005	0.0006
					t-test	1.0000	1.0000	1.0000		0.9903	0.0000	0.8186
					wilcox-test	1.0000	1.0000	1.0000		0.9983	0.0000	0.8723
71	✓	✗	✓	✓	mean	0.1246	0.1238	0.0704	0.0636	0.0911	0.0966	0.1153
					st.dev.	0.0007	0.0008	0.0014	0.0013	0.0014	0.0016	0.0018
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
72	✓	✓	✓	✓	mean	0.1007	0.1004	0.0817	0.0819	0.0945	0.0997	0.1036
					st.dev.	0.0007	0.0007	0.0020	0.0017	0.0015	0.0019	0.0006
					t-test	0.0000	0.0000	0.7875		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.8473		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the RPS based on 100 simulation replications for the sample size of 200 observations with 9 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 1000 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.2.7 AMSE: Low Dimension with 3 Classes

Table 2.B.12: Simulation Results: Accuracy Measure = AMSE & Low Dimension with 3 Classes

Simulation Design						Comparison of Methods							
DGP	noise	nonlin	multi	random	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
1	✗	✗	✗	✗	mean	0.0103	0.0669	0.0682	0.0565	0.0587	0.0800	0.0587	0.0800
					st.dev.	0.0044	0.0041	0.0044	0.0015	0.0022	0.0009	0.0016	0.0010
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.3900	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.2614	0.0000
2	✓	✗	✗	✗	mean	0.0227	0.0723	0.0727	0.0648	0.0682	0.0859	0.0684	0.0859
					st.dev.	0.0056	0.0034	0.0038	0.0013	0.0015	0.0010	0.0014	0.0010
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.1609	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.1287	0.0000
3	✗	✓	✗	✗	mean	0.0700	0.0576	0.0609	0.0552	0.0586	0.0644	0.0565	0.0644
					st.dev.	0.0032	0.0024	0.0032	0.0016	0.0021	0.0011	0.0016	0.0011
					t-test	0.0000	0.9980	0.0000	1.0000		0.0000	1.0000	0.0000
					wilcox-test	0.0000	0.9954	0.0000	1.0000		0.0000	1.0000	0.0000
4	✗	✗	✓	✗	mean	0.0124	0.1217	0.1166	0.0378	0.0370	0.0500	0.0367	0.0554
					st.dev.	0.0040	0.0068	0.0068	0.0017	0.0018	0.0014	0.0017	0.0016
					t-test	1.0000	0.0000	0.0000	0.0005		0.0000	0.8458	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0003		0.0000	0.8530	0.0000
5	✗	✗	✗	✓	mean	0.0096	0.0594	0.0567	0.0495	0.0511	0.0726	0.0517	0.0732
					st.dev.	0.0032	0.0057	0.0047	0.0015	0.0018	0.0011	0.0017	0.0011
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0044	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0024	0.0000
6	✓	✓	✗	✗	mean	0.0809	0.0612	0.0636	0.0604	0.0638	0.0671	0.0617	0.0670
					st.dev.	0.0048	0.0019	0.0030	0.0016	0.0017	0.0010	0.0015	0.0010
					t-test	0.0000	1.0000	0.7436	1.0000		0.0000	1.0000	0.0000
					wilcox-test	0.0000	1.0000	0.9265	1.0000		0.0000	1.0000	0.0000
7	✓	✗	✓	✗	mean	0.0283	0.1297	0.1262	0.0411	0.0407	0.0548	0.0427	0.0634
					st.dev.	0.0083	0.0052	0.0049	0.0015	0.0015	0.0020	0.0017	0.0022
					t-test	1.0000	0.0000	0.0000	0.0734		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0494		0.0000	0.0000	0.0000
8	✓	✗	✗	✓	mean	0.0230	0.0722	0.0746	0.0660	0.0705	0.0855	0.0701	0.0857
					st.dev.	0.0065	0.0028	0.0038	0.0014	0.0018	0.0008	0.0015	0.0008
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.9630	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.9578	0.0000
9	✗	✓	✓	✗	mean	0.0968	0.1066	0.1012	0.0493	0.0443	0.0660	0.0465	0.0680
					st.dev.	0.0024	0.0070	0.0060	0.0018	0.0021	0.0018	0.0019	0.0016
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
10	✗	✓	✗	✓	mean	0.0667	0.0538	0.0533	0.0507	0.0530	0.0599	0.0529	0.0600
					st.dev.	0.0034	0.0033	0.0031	0.0017	0.0019	0.0010	0.0018	0.0010
					t-test	0.0000	0.0119	0.1801	1.0000		0.0000	0.6401	0.0000
					wilcox-test	0.0000	0.0332	0.2314	1.0000		0.0000	0.7041	0.0000
11	✗	✗	✓	✓	mean	0.0132	0.1201	0.1172	0.0328	0.0326	0.0427	0.0327	0.0472
					st.dev.	0.0050	0.0111	0.0113	0.0017	0.0018	0.0013	0.0018	0.0016
					t-test	1.0000	0.0000	0.0000	0.1763		0.0000	0.3026	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.1287		0.0000	0.3376	0.0000
12	✓	✓	✓	✗	mean	0.1104	0.1064	0.1027	0.0616	0.0506	0.0737	0.0540	0.0751
					st.dev.	0.0039	0.0051	0.0043	0.0020	0.0024	0.0022	0.0021	0.0019
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
13	✓	✓	✗	✓	mean	0.0765	0.0595	0.0630	0.0587	0.0632	0.0646	0.0605	0.0646
					st.dev.	0.0050	0.0016	0.0029	0.0016	0.0019	0.0011	0.0016	0.0011
					t-test	0.0000	1.0000	0.7290	1.0000		0.0000	1.0000	0.0000
					wilcox-test	0.0000	1.0000	0.8231	1.0000		0.0000	1.0000	0.0000
14	✓	✗	✓	✓	mean	0.0311	0.1273	0.1244	0.0413	0.0408	0.0553	0.0420	0.0626
					st.dev.	0.0090	0.0044	0.0043	0.0019	0.0019	0.0021	0.0018	0.0023
					t-test	1.0000	0.0000	0.0000	0.0584		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.0428		0.0000	0.0000	0.0000
15	✗	✓	✓	✓	mean	0.0878	0.1016	0.0962	0.0420	0.0374	0.0566	0.0387	0.0593
					st.dev.	0.0031	0.0081	0.0072	0.0019	0.0022	0.0018	0.0020	0.0018
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
16	✓	✓	✓	✓	mean	0.1081	0.0985	0.0965	0.0637	0.0543	0.0752	0.0572	0.0768
					st.dev.	0.0039	0.0034	0.0029	0.0020	0.0026	0.0021	0.0022	0.0019
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the MSE based on 100 simulation replications for the sample size of 200 observations with 3 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 15 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.2.8 AMSE: Low Dimension with 6 Classes

Table 2.B.13: Simulation Results: Accuracy Measure = AMSE & Low Dimension with 6 Classes

Simulation Design						Comparison of Methods							
DGP	noise	nonlin	multi	random	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
17	✗	✗	✗	✗	mean	0.0043	0.0284	0.0283	0.0248	0.0291	0.0324	0.0287	0.0327
					st.dev.	0.0014	0.0012	0.0018	0.0007	0.0010	0.0005	0.0008	0.0005
					t-test	1.0000	1.0000	0.9998	1.0000		0.0000	0.9958	0.0000
					wilcox-test	1.0000	1.0000	1.0000	1.0000		0.0000	0.9953	0.0000
18	✓	✗	✗	✗	mean	0.0083	0.0294	0.0292	0.0272	0.0311	0.0337	0.0310	0.0341
					st.dev.	0.0021	0.0007	0.0010	0.0005	0.0006	0.0003	0.0005	0.0003
					t-test	1.0000	1.0000	1.0000	1.0000		0.0000	0.9791	0.0000
					wilcox-test	1.0000	1.0000	1.0000	1.0000		0.0000	0.9805	0.0000
19	✗	✓	✗	✗	mean	0.0245	0.0216	0.0222	0.0207	0.0257	0.0237	0.0240	0.0237
					st.dev.	0.0007	0.0009	0.0009	0.0006	0.0009	0.0004	0.0008	0.0004
					t-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
20	✗	✗	✓	✗	mean	0.0065	0.0600	0.0568	0.0238	0.0259	0.0299	0.0263	0.0356
					st.dev.	0.0017	0.0019	0.0023	0.0008	0.0009	0.0007	0.0008	0.0007
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0006	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0002	0.0000
21	✗	✗	✗	✓	mean	0.0043	0.0283	0.0282	0.0248	0.0291	0.0324	0.0289	0.0327
					st.dev.	0.0014	0.0013	0.0015	0.0007	0.0009	0.0005	0.0007	0.0005
					t-test	1.0000	1.0000	1.0000	1.0000		0.0000	0.9689	0.0000
					wilcox-test	1.0000	1.0000	1.0000	1.0000		0.0000	0.9661	0.0000
22	✓	✓	✗	✗	mean	0.0273	0.0223	0.0228	0.0220	0.0263	0.0242	0.0249	0.0243
					st.dev.	0.0015	0.0007	0.0008	0.0005	0.0007	0.0004	0.0006	0.0003
					t-test	0.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	0.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
23	✓	✗	✓	✗	mean	0.0114	0.0607	0.0580	0.0258	0.0266	0.0319	0.0305	0.0396
					st.dev.	0.0030	0.0014	0.0017	0.0007	0.0007	0.0008	0.0008	0.0007
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
24	✓	✗	✗	✓	mean	0.0088	0.0306	0.0296	0.0274	0.0306	0.0346	0.0310	0.0350
					st.dev.	0.0023	0.0010	0.0011	0.0005	0.0006	0.0004	0.0006	0.0004
					t-test	1.0000	0.6721	1.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.3992	1.0000	1.0000		0.0000	0.0000	0.0000
25	✗	✓	✓	✗	mean	0.0374	0.0396	0.0377	0.0234	0.0256	0.0292	0.0254	0.0315
					st.dev.	0.0007	0.0016	0.0013	0.0008	0.0011	0.0008	0.0009	0.0007
					t-test	0.0000	0.0000	0.0000	1.0000		0.0000	0.9637	0.0000
					wilcox-test	0.0000	0.0000	0.0000	1.0000		0.0000	0.9567	0.0000
26	✗	✓	✗	✓	mean	0.0245	0.0215	0.0224	0.0207	0.0256	0.0236	0.0241	0.0237
					st.dev.	0.0008	0.0007	0.0009	0.0006	0.0008	0.0004	0.0007	0.0005
					t-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
27	✗	✗	✓	✓	mean	0.0060	0.0587	0.0560	0.0236	0.0254	0.0297	0.0262	0.0355
					st.dev.	0.0015	0.0017	0.0019	0.0007	0.0009	0.0007	0.0009	0.0007
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
28	✓	✓	✓	✗	mean	0.0416	0.0396	0.0384	0.0271	0.0272	0.0312	0.0280	0.0338
					st.dev.	0.0014	0.0012	0.0009	0.0006	0.0009	0.0008	0.0008	0.0006
					t-test	0.0000	0.0000	0.0000	0.8880		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.8212		0.0000	0.0000	0.0000
29	✓	✓	✗	✓	mean	0.0292	0.0239	0.0240	0.0231	0.0268	0.0255	0.0261	0.0256
					st.dev.	0.0016	0.0009	0.0008	0.0005	0.0007	0.0003	0.0005	0.0003
					t-test	0.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	0.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
30	✓	✗	✓	✓	mean	0.0115	0.0618	0.0580	0.0242	0.0246	0.0306	0.0285	0.0375
					st.dev.	0.0029	0.0018	0.0020	0.0006	0.0006	0.0007	0.0008	0.0006
					t-test	1.0000	0.0000	0.0000	0.9999		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	0.9999		0.0000	0.0000	0.0000
31	✗	✓	✓	✓	mean	0.0378	0.0394	0.0375	0.0236	0.0256	0.0295	0.0256	0.0317
					st.dev.	0.0008	0.0014	0.0013	0.0007	0.0009	0.0007	0.0008	0.0006
					t-test	0.0000	0.0000	0.0000	1.0000		0.0000	0.6494	0.0000
					wilcox-test	0.0000	0.0000	0.0000	1.0000		0.0000	0.6416	0.0000
32	✓	✓	✓	✓	mean	0.0433	0.0438	0.0413	0.0270	0.0260	0.0314	0.0274	0.0339
					st.dev.	0.0014	0.0017	0.0014	0.0008	0.0011	0.0009	0.0010	0.0008
					t-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the MSE based on 100 simulation replications for the sample size of 200 observations with 6 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 15 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.2.9 AMSE: Low Dimension with 9 Classes

Table 2.B.14: Simulation Results: Accuracy Measure = AMSE & Low Dimension with 9 Classes

Simulation Design						Comparison of Methods							
DGP	noise	nonlin	multi	random	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
33	✗	✗	✗	✗	mean	0.0025	0.0150	0.0149	0.0134	0.0170	0.0170	0.0168	0.0172
					st.dev.	0.0008	0.0005	0.0007	0.0004	0.0006	0.0003	0.0005	0.0002
					t-test	1.0000	1.0000	1.0000	1.0000		0.5492	0.9993	0.0040
					wilcox-test	1.0000	1.0000	1.0000	1.0000		0.3269	0.9985	0.0003
34	✓	✗	✗	✗	mean	0.0046	0.0155	0.0154	0.0144	0.0176	0.0175	0.0175	0.0178
					st.dev.	0.0011	0.0004	0.0004	0.0003	0.0004	0.0002	0.0003	0.0002
					t-test	1.0000	1.0000	1.0000	1.0000		0.9697	0.9696	0.0011
					wilcox-test	1.0000	1.0000	1.0000	1.0000		0.9359	0.9544	0.0003
35	✗	✓	✗	✗	mean	0.0123	0.0110	0.0114	0.0107	0.0147	0.0121	0.0137	0.0121
					st.dev.	0.0004	0.0004	0.0005	0.0004	0.0006	0.0003	0.0005	0.0003
					t-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
36	✗	✗	✓	✗	mean	0.0044	0.0333	0.0321	0.0148	0.0168	0.0185	0.0175	0.0222
					st.dev.	0.0010	0.0008	0.0009	0.0004	0.0005	0.0005	0.0005	0.0003
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
37	✗	✗	✗	✓	mean	0.0026	0.0152	0.0154	0.0136	0.0173	0.0172	0.0170	0.0175
					st.dev.	0.0009	0.0005	0.0006	0.0003	0.0006	0.0002	0.0004	0.0002
					t-test	1.0000	1.0000	1.0000	1.0000		0.8591	1.0000	0.0034
					wilcox-test	1.0000	1.0000	1.0000	1.0000		0.7952	1.0000	0.0032
38	✓	✓	✗	✗	mean	0.0136	0.0112	0.0115	0.0111	0.0144	0.0122	0.0137	0.0122
					st.dev.	0.0006	0.0003	0.0004	0.0003	0.0004	0.0002	0.0003	0.0002
					t-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
39	✓	✗	✓	✗	mean	0.0066	0.0335	0.0323	0.0159	0.0167	0.0192	0.0200	0.0242
					st.dev.	0.0012	0.0006	0.0007	0.0005	0.0005	0.0006	0.0005	0.0004
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
40	✓	✗	✗	✓	mean	0.0046	0.0152	0.0152	0.0142	0.0175	0.0172	0.0173	0.0174
					st.dev.	0.0011	0.0004	0.0005	0.0003	0.0004	0.0002	0.0003	0.0002
					t-test	1.0000	1.0000	1.0000	1.0000		1.0000	0.9998	0.9655
					wilcox-test	1.0000	1.0000	1.0000	1.0000		1.0000	0.9999	0.9525
41	✗	✓	✓	✗	mean	0.0190	0.0198	0.0192	0.0127	0.0158	0.0156	0.0152	0.0170
					st.dev.	0.0004	0.0007	0.0007	0.0004	0.0006	0.0004	0.0005	0.0003
					t-test	0.0000	0.0000	0.0000	1.0000		0.9652	1.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	1.0000		0.9503	1.0000	0.0000
42	✗	✓	✗	✓	mean	0.0125	0.0112	0.0117	0.0108	0.0147	0.0122	0.0138	0.0122
					st.dev.	0.0004	0.0004	0.0007	0.0003	0.0005	0.0002	0.0005	0.0002
					t-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
43	✗	✗	✓	✓	mean	0.0043	0.0335	0.0324	0.0149	0.0167	0.0187	0.0176	0.0225
					st.dev.	0.0013	0.0007	0.0009	0.0005	0.0005	0.0005	0.0005	0.0004
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
44	✓	✓	✓	✗	mean	0.0208	0.0198	0.0193	0.0143	0.0159	0.0164	0.0162	0.0181
					st.dev.	0.0007	0.0006	0.0005	0.0003	0.0006	0.0004	0.0005	0.0003
					t-test	0.0000	0.0000	0.0000	1.0000		0.0000	0.0001	0.0000
					wilcox-test	0.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
45	✓	✓	✗	✓	mean	0.0130	0.0110	0.0113	0.0108	0.0142	0.0118	0.0134	0.0118
					st.dev.	0.0006	0.0003	0.0003	0.0003	0.0004	0.0003	0.0003	0.0002
					t-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
46	✓	✗	✓	✓	mean	0.0070	0.0335	0.0325	0.0166	0.0173	0.0200	0.0204	0.0250
					st.dev.	0.0016	0.0005	0.0007	0.0004	0.0005	0.0005	0.0005	0.0004
					t-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	1.0000	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
47	✗	✓	✓	✓	mean	0.0192	0.0203	0.0198	0.0130	0.0159	0.0158	0.0153	0.0173
					st.dev.	0.0004	0.0007	0.0007	0.0004	0.0006	0.0004	0.0005	0.0003
					t-test	0.0000	0.0000	0.0000	1.0000		0.6681	1.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	1.0000		0.5336	1.0000	0.0000
48	✓	✓	✓	✓	mean	0.0203	0.0194	0.0190	0.0142	0.0159	0.0161	0.0162	0.0179
					st.dev.	0.0006	0.0006	0.0005	0.0003	0.0005	0.0003	0.0004	0.0003
					t-test	0.0000	0.0000	0.0000	1.0000		0.0006	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000	1.0000		0.0004	0.0000	0.0000

Notes: Table reports the average measures of the MSE based on 100 simulation replications for the sample size of 200 observations with 9 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 15 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.2.10 AMSE: High Dimension with 3 Classes

Table 2.B.15: Simulation Results: Accuracy Measure = AMSE & High Dimension with 3 Classes

Simulation Design						Comparison of Methods						
DGP	noise	nonlin	multi	random	Statistic	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
49	✓	✗	✗	✗	mean	0.0923	0.0931	0.0908	0.0930	0.0952	0.0926	0.0952
					st.dev.	0.0008	0.0013	0.0009	0.0009	0.0007	0.0007	0.0007
					t-test	1.0000	0.2408	1.0000		0.0000	0.9980	0.0000
					wilcox-test	1.0000	0.5433	1.0000		0.0000	0.9977	0.0000
50	✓	✓	✗	✗	mean	0.0692	0.0698	0.0696	0.0702	0.0699	0.0696	0.0699
					st.dev.	0.0009	0.0013	0.0010	0.0009	0.0009	0.0008	0.0008
					t-test	1.0000	0.9907	0.9999		0.9649	1.0000	0.9852
					wilcox-test	1.0000	1.0000	1.0000		0.9887	1.0000	0.9944
51	✓	✗	✓	✗	mean	0.1385	0.1379	0.0864	0.0752	0.1008	0.0881	0.1087
					st.dev.	0.0009	0.0010	0.0019	0.0021	0.0023	0.0019	0.0018
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
52	✓	✗	✗	✓	mean	0.0906	0.0904	0.0884	0.0902	0.0931	0.0902	0.0931
					st.dev.	0.0011	0.0013	0.0008	0.0008	0.0007	0.0008	0.0007
					t-test	0.0006	0.0794	1.0000		0.0000	0.3296	0.0000
					wilcox-test	0.0010	0.1853	1.0000		0.0000	0.2606	0.0000
53	✓	✓	✓	✗	mean	0.1079	0.1083	0.0910	0.0888	0.1010	0.0892	0.1013
					st.dev.	0.0009	0.0011	0.0025	0.0020	0.0019	0.0019	0.0019
					t-test	0.0000	0.0000	0.0000		0.0000	0.0936	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0745	0.0000
54	✓	✓	✗	✓	mean	0.0706	0.0703	0.0703	0.0705	0.0706	0.0704	0.0706
					st.dev.	0.0010	0.0011	0.0010	0.0009	0.0009	0.0008	0.0009
					t-test	0.1479	0.9409	0.8941		0.1495	0.7655	0.1796
					wilcox-test	0.1712	0.9972	0.9496		0.0718	0.8048	0.1178
55	✓	✗	✓	✓	mean	0.1291	0.1276	0.0725	0.0678	0.0914	0.0758	0.0954
					st.dev.	0.0016	0.0010	0.0020	0.0021	0.0021	0.0020	0.0020
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
56	✓	✓	✓	✓	mean	0.1081	0.1079	0.0863	0.0828	0.0970	0.0834	0.0971
					st.dev.	0.0012	0.0009	0.0028	0.0019	0.0021	0.0020	0.0021
					t-test	0.0000	0.0000	0.0000		0.0000	0.0264	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0364	0.0000

Notes: Table reports the average measures of the MSE based on 100 simulation replications for the sample size of 200 observations with 3 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 1000 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.2.11 AMSE: High Dimension with 6 Classes

Table 2.B.16: Simulation Results: Accuracy Measure = AMSE & High Dimension with 6 Classes

Simulation Design						Comparison of Methods						
DGP	noise	nonlin	multi	random	Statistic	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
57	✓	✗	✗	✗	mean	0.0352	0.0352	0.0347	0.0361	0.0361	0.0360	0.0361
					st.dev.	0.0003	0.0004	0.0004	0.0004	0.0004	0.0003	0.0004
					t-test	1.0000	1.0000	1.0000		0.8112	0.9994	0.6394
					wilcox-test	1.0000	1.0000	1.0000		0.8788	0.9989	0.6579
58	✓	✓	✗	✗	mean	0.0246	0.0246	0.0246	0.0257	0.0248	0.0252	0.0248
					st.dev.	0.0003	0.0003	0.0003	0.0004	0.0003	0.0002	0.0003
					t-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
59	✓	✗	✓	✗	mean	0.0622	0.0617	0.0459	0.0383	0.0494	0.0479	0.0553
					st.dev.	0.0003	0.0003	0.0005	0.0007	0.0007	0.0006	0.0006
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
60	✓	✗	✗	✓	mean	0.0339	0.0341	0.0335	0.0350	0.0347	0.0348	0.0348
					st.dev.	0.0003	0.0004	0.0004	0.0004	0.0004	0.0003	0.0004
					t-test	1.0000	1.0000	1.0000		1.0000	0.9993	0.9999
					wilcox-test	1.0000	1.0000	1.0000		1.0000	0.9995	1.0000
61	✓	✓	✓	✗	mean	0.0397	0.0397	0.0351	0.0358	0.0383	0.0380	0.0399
					st.dev.	0.0004	0.0004	0.0007	0.0006	0.0005	0.0006	0.0004
					t-test	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
62	✓	✓	✗	✓	mean	0.0229	0.0231	0.0229	0.0241	0.0231	0.0235	0.0231
					st.dev.	0.0004	0.0005	0.0005	0.0005	0.0005	0.0004	0.0005
					t-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
63	✓	✗	✓	✓	mean	0.0628	0.0629	0.0481	0.0405	0.0512	0.0506	0.0583
					st.dev.	0.0003	0.0004	0.0005	0.0008	0.0007	0.0006	0.0005
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
64	✓	✓	✓	✓	mean	0.0383	0.0386	0.0343	0.0350	0.0367	0.0378	0.0387
					st.dev.	0.0003	0.0004	0.0006	0.0005	0.0005	0.0005	0.0004
					t-test	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the MSE based on 100 simulation replications for the sample size of 200 observations with 6 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 1000 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.2.12 AMSE: High Dimension with 9 Classes

Table 2.B.17: Simulation Results: Accuracy Measure = AMSE & High Dimension with 9 Classes

Simulation Design						Comparison of Methods						
DGP	noise	nonlin	multi	random	Statistic	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
65	✓	✗	✗	✗	mean	0.0180	0.0181	0.0178	0.0189	0.0185	0.0188	0.0185
					st.dev.	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
					t-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
66	✓	✓	✗	✗	mean	0.0123	0.0123	0.0123	0.0133	0.0124	0.0129	0.0124
					st.dev.	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
					t-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
67	✓	✗	✓	✗	mean	0.0339	0.0337	0.0263	0.0224	0.0281	0.0284	0.0316
					st.dev.	0.0002	0.0002	0.0003	0.0005	0.0004	0.0003	0.0003
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
68	✓	✗	✗	✓	mean	0.0181	0.0181	0.0179	0.0190	0.0186	0.0188	0.0186
					st.dev.	0.0002	0.0002	0.0003	0.0003	0.0003	0.0002	0.0003
					t-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
69	✓	✓	✓	✗	mean	0.0198	0.0199	0.0178	0.0187	0.0193	0.0201	0.0201
					st.dev.	0.0002	0.0002	0.0003	0.0003	0.0003	0.0002	0.0002
					t-test	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
70	✓	✓	✗	✓	mean	0.0124	0.0124	0.0124	0.0133	0.0125	0.0130	0.0125
					st.dev.	0.0002	0.0002	0.0002	0.0003	0.0002	0.0002	0.0002
					t-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
					wilcox-test	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
71	✓	✗	✓	✓	mean	0.0338	0.0337	0.0262	0.0225	0.0281	0.0285	0.0315
					st.dev.	0.0002	0.0002	0.0004	0.0005	0.0005	0.0003	0.0004
					t-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
72	✓	✓	✓	✓	mean	0.0200	0.0200	0.0178	0.0187	0.0193	0.0201	0.0202
					st.dev.	0.0002	0.0002	0.0003	0.0003	0.0003	0.0003	0.0002
					t-test	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000
					wilcox-test	0.0000	0.0000	1.0000		0.0000	0.0000	0.0000

Notes: Table reports the average measures of the MSE based on 100 simulation replications for the sample size of 200 observations with 9 outcome classes. Columns 1 to 5 specify the DGP identifier and its features, namely 1000 additional noise variables (*noise*), nonlinear effects (*nonlin*), multicollinearity among covariates (*multi*), and randomly spaced thresholds (*random*). The sixth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.3 Empirical Results

In this section we present more detailed and supplementary results regarding the empirical results (Section 2.5.4) discussed in the main text. In the following the descriptive statistics for the considered datasets and the results for the prediction accuracy are summarized.

2.B.3.1 Descriptive Statistics

Table 2.B.18: Descriptive Statistics: mammography dataset

Mammography Dataset						
variable	type	mean	sd	median	min	max
SYMPT*	Categorical	2.97	0.95	3.00	1.00	4.00
PB	Numeric	7.56	2.10	7.00	5.00	17.00
HIST*	Categorical	1.11	0.31	1.00	1.00	2.00
BSE*	Categorical	1.87	0.34	2.00	1.00	2.00
DECT*	Categorical	2.66	0.56	3.00	1.00	3.00
y*	Categorical	1.61	0.77	1.00	1.00	3.00

Table 2.B.19: Descriptive Statistics: nhanes dataset

Nhanes Dataset						
variable	type	mean	sd	median	min	max
sex*	Categorical	1.51	0.50	2.00	1.00	2.00
race*	Categorical	2.87	1.00	3.00	1.00	5.00
country_of_birth*	Categorical	1.34	0.79	1.00	1.00	4.00
education*	Categorical	3.37	1.24	3.00	1.00	5.00
marital_status*	Categorical	2.31	1.74	1.00	1.00	6.00
waistcircum	Numeric	100.37	16.37	99.40	61.60	176.70
Cholesterol	Numeric	196.89	41.59	193.00	97.00	432.00
WBCcount	Numeric	7.30	2.88	6.90	1.60	83.20
AcuteIllness*	Categorical	1.25	0.43	1.00	1.00	2.00
depression*	Categorical	1.39	0.76	1.00	1.00	4.00
ToothCond*	Categorical	3.05	1.24	3.00	1.00	5.00
sleepTrouble*	Categorical	2.28	1.28	2.00	1.00	5.00
wakeUp*	Categorical	2.41	1.30	2.00	1.00	5.00
cig*	Categorical	1.51	0.50	2.00	1.00	2.00
diabetes*	Categorical	1.14	0.34	1.00	1.00	2.00
asthma*	Categorical	1.15	0.36	1.00	1.00	2.00
heartFailure*	Categorical	1.03	0.16	1.00	1.00	2.00
stroke*	Categorical	1.03	0.18	1.00	1.00	2.00
chronicBronchitis*	Categorical	1.07	0.26	1.00	1.00	2.00
alcohol	Numeric	3.93	20.18	2.00	0.00	365.00
heavyDrinker*	Categorical	1.17	0.37	1.00	1.00	2.00
medicalPlaceToGo*	Categorical	1.92	0.67	2.00	1.00	5.00
BPsys	Numeric	124.44	18.62	122.00	78.00	230.00
BPdias	Numeric	71.18	11.84	72.00	10.00	114.00
age	Numeric	49.96	16.68	50.00	20.00	80.00
BMI	Numeric	29.33	6.66	28.32	14.20	73.43
y*	Categorical	2.77	1.00	3.00	1.00	5.00

Table 2.B.20: Descriptive Statistics: supportstudy dataset

Supportstudy Dataset						
variable	type	mean	sd	median	min	max
age	Numeric	62.80	16.27	65.29	20.30	100.13
sex*	Categorical	1.54	0.50	2.00	1.00	2.00
dzgroup*	Categorical	3.23	2.48	2.00	1.00	8.00
num.co	Numeric	1.90	1.34	2.00	0.00	7.00
scoma	Numeric	12.45	25.29	0.00	0.00	100.00
charges	Numeric	59307.91	86620.70	28416.50	1635.75	740010.00
avtisst	Numeric	23.53	13.60	20.00	1.67	64.00
race*	Categorical	1.36	0.88	1.00	1.00	5.00
meanbp	Numeric	84.52	27.64	77.00	0.00	180.00
wblc	Numeric	12.62	9.31	10.50	0.05	100.00
hrt	Numeric	98.59	32.93	102.50	0.00	300.00
resp	Numeric	23.60	9.54	24.00	0.00	64.00
temp	Numeric	37.08	1.25	36.70	32.50	41.20
crea	Numeric	1.80	1.74	1.20	0.30	11.80
sod	Numeric	137.64	6.34	137.00	118.00	175.00
y*	Categorical	2.90	1.81	2.00	1.00	5.00

Table 2.B.21: Descriptive Statistics: vlbw dataset

Vlbw Dataset						
variable	type	mean	sd	median	min	max
race*	Categorical	1.57	0.50	2.00	1.00	2.00
bwt	Numeric	1094.89	260.44	1140.00	430.00	1500.00
inout*	Categorical	1.03	0.16	1.00	1.00	2.00
twn*	Categorical	1.24	0.43	1.00	1.00	2.00
lol	Numeric	7.73	19.47	3.00	0.00	192.00
magsulf*	Categorical	1.18	0.39	1.00	1.00	2.00
meth*	Categorical	1.44	0.50	1.00	1.00	2.00
toc*	Categorical	1.24	0.43	1.00	1.00	2.00
delivery*	Categorical	1.41	0.49	1.00	1.00	2.00
sex*	Categorical	1.50	0.50	1.00	1.00	2.00
y*	Categorical	5.09	2.58	6.00	1.00	9.00

Table 2.B.22: Descriptive Statistics: winequality dataset

Winequality Dataset						
variable	type	mean	sd	median	min	max
fixed.acidity	Numeric	6.85	0.84	6.80	3.80	14.20
volatile.acidity	Numeric	0.28	0.10	0.26	0.08	1.10
citric.acid	Numeric	0.33	0.12	0.32	0.00	1.66
residual.sugar	Numeric	6.39	5.07	5.20	0.60	65.80
chlorides	Numeric	0.05	0.02	0.04	0.01	0.35
free.sulfur.dioxide	Numeric	35.31	17.01	34.00	2.00	289.00
total.sulfur.dioxide	Numeric	138.38	42.51	134.00	9.00	440.00
density	Numeric	0.99	0.00	0.99	0.99	1.04
pH	Numeric	3.19	0.15	3.18	2.72	3.82
sulphates	Numeric	0.49	0.11	0.47	0.22	1.08
alcohol	Numeric	10.51	1.23	10.40	8.00	14.20
y*	Categorical	3.87	0.88	4.00	1.00	6.00

2.B.3.2 Prediction Accuracy

Tables 2.B.23 and 2.B.24 summarize in detail the results of the prediction accuracy exercise using real datasets for the ARPS and the AMSE, respectively. The first column *Data* specifies the dataset, the second column *Class* defines the number of outcome classes of the dependent variable and the third column *Size* indicates the number of observations. Similarly to the simulation results, the column *Statistic* contains summary statistics and statistical tests results for the equality of means between the results of the *Ordered Forest* and all the other methods.

Table 2.B.23: Empirical Results: Accuracy Measure = ARPS

Dataset Summary				**Comparison of Methods**							
Data	Class	Size	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
mammography	3	412	mean	0.1776	0.2251	0.2089	0.1767	0.1823	0.1766	0.1826	0.1767
			st.dev.	0.0010	0.0027	0.0021	0.0013	0.0018	0.0008	0.0019	0.0007
			t-test	1.0000	0.0000	0.0000	1.0000		1.0000	0.3999	1.0000
			wilcox-test	1.0000	0.0000	0.0000	1.0000		1.0000	0.3153	1.0000
nhanes	5	1914	mean	0.1088	0.1089	0.1100	0.1085	0.1103	0.1137	0.1104	0.1159
			st.dev.	0.0004	0.0003	0.0004	0.0001	0.0002	0.0001	0.0002	0.0001
			t-test	1.0000	1.0000	0.9839	1.0000		0.0000	0.2106	0.0000
			wilcox-test	1.0000	1.0000	0.9738	1.0000		0.0000	0.2179	0.0000
supportstudy	5	798	mean	0.1872	0.1849	0.1834	0.1800	0.1823	0.1931	0.1857	0.1944
			st.dev.	0.0011	0.0010	0.0009	0.0008	0.0008	0.0003	0.0007	0.0004
			t-test	0.0000	0.0000	0.0052	1.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0073	1.0000		0.0000	0.0000	0.0000
vlbw	9	218	mean	0.1595	0.1713	0.1724	0.1603	0.1686	0.1623	0.1685	0.1642
			st.dev.	0.0011	0.0026	0.0030	0.0014	0.0021	0.0005	0.0020	0.0003
			t-test	1.0000	0.0100	0.0023	1.0000		1.0000	0.5143	1.0000
			wilcox-test	1.0000	0.0116	0.0010	1.0000		1.0000	0.5733	1.0000
winequality	6	4893	mean	0.0756	0.0501	0.0503	0.0596	0.0507	0.0673	0.0504	0.0683
			st.dev.	0.0000	0.0003	0.0002	0.0001	0.0002	0.0001	0.0002	0.0000
			t-test	0.0000	1.0000	0.9992	0.0000		0.0000	0.9971	0.0000
			wilcox-test	0.0000	0.9999	0.9986	0.0000		0.0000	0.9966	0.0000

Notes: Table reports the average measures of the RPS based on 10 repetitions of 10-fold cross-validation. The fourth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

Table 2.B.24: Empirical Results: Accuracy Measure = AMSE

Dataset Summary				**Comparison of Methods**							
Data	Class	Size	Statistic	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
mammography	3	412	mean	0.1754	0.2593	0.2222	0.1720	0.1766	0.1726	0.1770	0.1726
			st.dev.	0.0007	0.0025	0.0031	0.0008	0.0012	0.0004	0.0013	0.0004
			t-test	0.9923	0.0000	0.0000	1.0000		1.0000	0.2467	1.0000
			wilcox-test	0.9943	0.0000	0.0000	1.0000		1.0000	0.2179	1.0000
nhanes	5	1914	mean	0.1310	0.1309	0.1332	0.1304	0.1332	0.1329	0.1319	0.1343
			st.dev.	0.0003	0.0003	0.0003	0.0002	0.0003	0.0001	0.0003	0.0001
			t-test	1.0000	1.0000	0.7067	1.0000		0.9936	1.0000	0.0000
			wilcox-test	1.0000	1.0000	0.6579	1.0000		0.9955	1.0000	0.0000
supportstudy	5	798	mean	0.1124	0.1110	0.1094	0.1078	0.1088	0.1129	0.1101	0.1135
			st.dev.	0.0005	0.0004	0.0004	0.0004	0.0004	0.0002	0.0003	0.0002
			t-test	0.0000	0.0000	0.0020	1.0000		0.0000	0.0000	0.0000
			wilcox-test	0.0000	0.0000	0.0008	0.9999		0.0000	0.0000	0.0000
vlbw	9	218	mean	0.0944	0.0986	0.0990	0.0956	0.1008	0.0958	0.1006	0.0956
			st.dev.	0.0002	0.0008	0.0009	0.0004	0.0008	0.0003	0.0009	0.0002
			t-test	1.0000	1.0000	0.9999	1.0000		1.0000	0.7224	1.0000
			wilcox-test	1.0000	1.0000	0.9999	1.0000		1.0000	0.7821	1.0000
winequality	6	4893	mean	0.1001	0.0692	0.0698	0.0831	0.0702	0.0906	0.0693	0.0913
			st.dev.	0.0000	0.0003	0.0003	0.0001	0.0003	0.0001	0.0003	0.0001
			t-test	0.0000	1.0000	0.9960	0.0000		0.0000	1.0000	0.0000
			wilcox-test	0.0000	1.0000	0.9974	0.0000		0.0000	1.0000	0.0000

Notes: Table reports the average measures of the MSE based on 10 repetitions of 10-fold cross-validation. The fourth column *Statistic* shows the mean and the standard deviation of the accuracy measure for all methods. Additionally, *t-test* and *wilcox-test* contain the p-values of the parametric t-test as well as the nonparametric Wilcoxon test for the equality of means between the results of the *Ordered Forest* and all the other methods.

2.B.4 Software Implementation

The Monte Carlo study has been conducted using the R statistical software (R Core Team, 2018) in version 3.5.2 (Eggshell Igloo) and the respective packages implementing the estimators used. With regards to the forest-based estimators the main tuning parameters, namely the number of trees, the number of randomly chosen covariates and the minimum leaf size have been specified according to the values in Table 2.5.1 in the main text.

Table 2.B.25: Overview of Software Packages and Tuning Parameters

Software Implementation and Tuning Parameters								
method	Ologit	Naive	Ordinal	Conditional	Ordered	Ordered*	Multi	Multi*
package	rms	ordinalForest	ordinalForest	party	ranger	grf	ranger	grf
function	lrm	ordfor	ordfor	cforest	ranger	regression_forest	ranger	regression_forest
max. iterations	25	-	-	-	-	-	-	-
trees	-	1000	1000	1000	1000	1000	1000	1000
random subset	-	$\sqrt{p}$	$\sqrt{p}$	$\sqrt{p}$	$\sqrt{p}$	$\sqrt{p}$	$\sqrt{p}$	$\sqrt{p}$
leaf size	-	5	5	0	5	5	5	5
B_{sets}	-	0	1000	-	-	-	-	-
B_{prior}	-	0	100	-	-	-	-	-
performance	-	equal	equal	-	-	-	-	-
S_{best}	-	0	10	-	-	-	-	-

In terms of the particular R packages used the ordered logistic regression has been implemented using the rms package (version 5.1-3) written by Harrell (2019). The respective lrm function for fitting the ordered logit has been used with the default parameters, except setting the maximum number of iterations, maxit=25 as for some of the DGPs the ordered logit has experienced convergence issues. Next, the naive and the ordinal forest have been applied based on the ordinalForest package in version 2.3 (Hornung, 2019b) with the ordfor function. As described in Appendix 2.A.3 the ordinal forest introduces additional tuning parameters for which we use the default parameters as suggested in the package manual. Further, the conditional forest has been estimated with the package party in version 1.3-1 (Hothorn, Bühlmann, Dudoit, Molinaro, & Van Der Laan, 2006a; Strobl, Boulesteix, Zeileis, & Hothorn, 2007; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). Regarding the choice of the tuning parameters, we rely on the default parameters of the cforest function. A particularity of the conditional forest is, due to the conceptual differences to standard regression forest in terms of the splitting criterion, the choice of the stopping rule. This is controlled by the significance level α (see Appendix 2.A.2 for details). However, in order to grow deep trees we follow the suggestion in the package manual to set mincriterion= 0, which has been also used in the simulation study conducted in Janitza et al. (2016). Lastly, the *Ordered Forest* as well as the multinomial forest algorithms are implemented using the package ranger in version 0.11.1 (Wright & Ziegler, 2017) with the default hyperparameters. The honest versions of the above two estimators rely on the grf package in version 0.10.2 (Tibshirani et al., 2018) with the default hyperparameters as well. A detailed overview of packages with the corresponding tuning parameters is provided in Table 2.B.25.

Furthermore, Tables 2.B.26 and 2.B.27 compare the absolute and relative computation time of the respective methods. For comparison purposes, we measure the computation time for the four main DGPs presented in Section 2.5.3 of the main text, namely the simple DGP in the low- and high-dimensional case as well as the complex DGP in the low- and high-dimensional case, for both the small sample size ($N = 200$) and the big sample size ($N = 800$) for all considered number of outcome classes. We estimate the model based on the training set and predict the class probabilities for a test set of size $N = 10'000$ as in the main simulation. We repeat this procedure 10 times and report the average computation time. The tuning parameters and the software implementations are chosen as defined in Table 2.5.1 in the main text and Table 2.B.25 herein, respectively. All simulations are computed on a 64-Bit Windows machine

with 4 cores (1.80GHz) and 16GB RAM storage.

Table 2.B.26: Absolute Computation Time in Seconds

Simulation Design					**Comparison of Methods**						
Class	Dim.	DGP	Size	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
3	Low	Simple	200	0.01	1.22	10.33	46.61	0.62	1.24	0.91	1.86
3	Low	Simple	800	0.02	1.58	40.83	150.84	1.03	1.96	1.61	2.98
3	Low	Complex	200	0.02	1.19	11.93	47.43	0.63	1.26	0.98	1.92
3	Low	Complex	800	0.03	1.71	52.45	150.59	1.08	1.94	1.73	3.06
3	High	Simple	200		3.50	61.89	64.28	4.05	5.08	6.06	7.27
3	High	Simple	800		13.91	332.60	175.76	7.19	7.10	12.19	11.02
3	High	Complex	200		3.46	60.25	59.98	4.02	4.96	6.02	7.10
3	High	Complex	800		13.83	325.65	173.63	6.83	6.61	11.50	10.66
6	Low	Simple	200	0.02	1.88	12.79	46.80	1.47	3.00	1.74	3.52
6	Low	Simple	800	0.03	2.28	48.98	151.58	2.45	4.75	3.10	5.82
6	Low	Complex	200	0.03	1.85	14.75	46.97	1.56	3.12	1.85	3.66
6	Low	Complex	800	0.04	2.54	64.44	151.84	2.68	4.82	3.30	6.02
6	High	Simple	200		4.21	69.80	64.14	10.24	11.74	12.01	13.63
6	High	Simple	800		15.86	386.02	176.27	19.34	17.43	26.24	19.97
6	High	Complex	200		4.11	70.51	60.85	9.98	11.52	11.95	13.61
6	High	Complex	800		15.85	371.69	174.17	18.11	17.18	24.43	19.52
9	Low	Simple	200	0.03	2.32	20.53	46.70	2.27	4.71	2.44	5.03
9	Low	Simple	800	0.04	2.69	57.22	145.21	3.82	7.29	4.61	7.99
9	Low	Complex	200	0.03	2.29	22.86	47.36	2.40	4.83	2.65	5.28
9	Low	Complex	800	0.05	3.07	79.15	151.36	4.27	7.75	5.81	8.68
9	High	Simple	200		4.85	80.76	63.25	16.05	17.84	17.69	19.56
9	High	Simple	800		16.91	413.74	169.91	31.34	26.91	38.95	27.38
9	High	Complex	200		4.62	78.86	57.68	15.79	17.78	17.57	19.59
9	High	Complex	800		18.10	437.04	175.07	31.12	27.33	37.59	28.16

Notes: Table reports the average absolute computation time in seconds based on 10 simulation replications of training and prediction. The first column denotes the number of outcome classes. Columns 2 and 3 specify the dimension and the DGP, respectively. The fourth column contains the number of observations in the training set. The prediction set consists of 10 000 observations.

The results reveal the expected pattern for the *Ordered Forest*. The more outcome classes the longer the computation time as by definition of the algorithm more forests have to be estimated. Furthermore, we also observe a longer computation time if the number of observation and/or the number of considered splitting covariates increases which is also an expected behaviour. However, the computation time is not sensitive to the particular DGP which it should not be either. The latter two patterns are true for all considered methods. In comparison to the other forest-based methods, the computational advantage of the *Ordered Forest* becomes apparent. The *Ordered Forest* outperforms the ordinal and the conditional forest in all cases. In some cases the *Ordered Forest* is even more than 100 times faster and even in the closest cases it is more than 3 times faster than the two. In absolute terms this translates to computation time of around 1 second for the *Ordered Forest* and around 50 seconds for the ordinal and around 150 seconds for the conditional forest in the most extreme case. Contrarily, in the closest case, the computation time for the *Ordered Forest* is around 15 seconds, while for the ordinal forest this is around 80 seconds and around 60 seconds for the conditional forest. This points to the additional computation burden of the ordinal and the conditional forest due to the optimization procedure and the permutation tests, respectively. The only exception is the naive forest which does not include the optimization step. Furthermore, we observe a slightly longer computation time for the multinomial forest in comparison to the *Ordered Forest*, which is due to one extra forest being estimated. The honest versions of the two forests take a bit longer in general, but this seems to reverse once bigger samples are considered (in terms

of both number of observations as well as number of considered covariates).

Table 2.B.27: Relative Computation Time

Simulation Design				**Comparison of Methods**							
Class	Dim.	DGP	Size	Ologit	Naive	Ordinal	Cond.	Ordered	Ordered*	Multi	Multi*
3	Low	Simple	200	0.02	1.98	16.76	75.66	1	2.02	1.48	3.02
3	Low	Simple	800	0.02	1.53	39.68	146.59	1	1.91	1.56	2.90
3	Low	Complex	200	0.03	1.87	18.79	74.70	1	1.99	1.55	3.03
3	Low	Complex	800	0.03	1.59	48.79	140.09	1	1.81	1.61	2.84
3	High	Simple	200		0.86	15.27	15.86	1	1.25	1.50	1.79
3	High	Simple	800		1.94	46.28	24.46	1	0.99	1.70	1.53
3	High	Complex	200		0.86	14.99	14.92	1	1.23	1.50	1.77
3	High	Complex	800		2.02	47.68	25.42	1	0.97	1.68	1.56
6	Low	Simple	200	0.02	1.28	8.73	31.95	1	2.05	1.19	2.40
6	Low	Simple	800	0.01	0.93	19.95	61.74	1	1.94	1.26	2.37
6	Low	Complex	200	0.02	1.18	9.45	30.09	1	2.00	1.19	2.34
6	Low	Complex	800	0.02	0.94	24.02	56.59	1	1.80	1.23	2.24
6	High	Simple	200		0.41	6.81	6.26	1	1.15	1.17	1.33
6	High	Simple	800		0.82	19.96	9.11	1	0.90	1.36	1.03
6	High	Complex	200		0.41	7.07	6.10	1	1.16	1.20	1.36
6	High	Complex	800		0.88	20.52	9.62	1	0.95	1.35	1.08
9	Low	Simple	200	0.01	1.02	9.03	20.54	1	2.07	1.07	2.21
9	Low	Simple	800	0.01	0.70	14.98	38.01	1	1.91	1.21	2.09
9	Low	Complex	200	0.01	0.95	9.51	19.69	1	2.01	1.10	2.19
9	Low	Complex	800	0.01	0.72	18.55	35.48	1	1.82	1.36	2.03
9	High	Simple	200		0.30	5.03	3.94	1	1.11	1.10	1.22
9	High	Simple	800		0.54	13.20	5.42	1	0.86	1.24	0.87
9	High	Complex	200		0.29	5.00	3.65	1	1.13	1.11	1.24
9	High	Complex	800		0.58	14.04	5.63	1	0.88	1.21	0.90

Notes: Table reports the average relative computation time with regards to the *Ordered Forest* estimator based on 10 simulation replications of training and prediction. The first column denotes the number of outcome classes. Columns 2 and 3 specify the dimension and the DGP, respectively. The fourth column contains the number of observations in the training set. The prediction set consists of 10 000 observations.

Generally, the sensitivity with regards to the computation time appears to be very different for the considered methods. For the *Ordered Forest* as well as the multinomial forest, including their honest versions, the most important aspect is clearly the number of outcome classes. For the naive and the ordinal forest the number of observations seems to be most decisive and for the conditional forest paradoxically the size of the prediction set is most relevant. Overall, the above result support the theoretical argument of the *Ordered Forest* being computationally advantageous in comparison to the ordinal and the conditional forest.

2.C Empirical Application

In this appendix we provide the descriptive statistics for the dataset used in the empirical application of the main text as well as supplementary results containing the estimation of marginal effects.

2.C.1 Descriptive Statistics

Table 2.C.1: Descriptive Statistics: NHIS Dataset

NHIS Dataset						
variable	type	mean	sd	median	min	max
Health Status*	Categorical	3.93	0.95	4.00	1.00	5.00
Health Insurance*	Categorical	0.84	0.37	1.00	0.00	1.00
Female*	Categorical	0.50	0.50	0.50	0.00	1.00
Non White*	Categorical	0.20	0.40	0.00	0.00	1.00
Age	Numeric	42.72	8.70	43.00	26.00	59.00
Education	Numeric	13.74	2.99	14.00	0.00	18.00
Family Size	Numeric	3.63	1.37	4.00	2.00	18.00
Employed*	Categorical	0.82	0.39	1.00	0.00	1.00
Income*	Categorical	94178.04	56738.46	85985.78	19282.93	167844.53

Table 2.C.2: Descriptive Statistics by Class: NHIS Dataset

NHIS Dataset					
	Health Status				
variable	poor	fair	good	very good	excellent
Health Status	1.14	5.66	25.14	34.92	33.13
Health Insurance	79.07	71.50	77.88	87.52	87.76
Female	49.77	51.08	49.28	50.43	49.92
Non White	31.63	23.89	22.84	18.18	18.21
Age	47.65	45.37	43.75	42.73	41.30
Education	12.11	12.20	12.89	13.97	14.46
Family Size	3.33	3.68	3.68	3.59	3.64
Employed	28.84	65.57	80.99	84.35	84.21
Income	53409.03	62473.99	78957.11	99685.45	106743.21
N	215	1063	4724	6562	6226
share in %	1.14	5.66	25.14	34.92	33.13

Note: Means of variables for respective outcome class displayed. Shares for dummy variables are indicated in %.

2.C.2 Marginal Effects

In what follows, the results for the marginal effects at mean are presented for the considered NHIS dataset. Similarly as in the main text, the effects are computed for each outcome class of the dependent variable both for the *Ordered Forest* as well as for the ordered logit. The estimations are done in R version 3.6.1 using the orf package (Lechner & Okasa, 2019) in version 0.1.3 for the *Ordered Forest* and the oglmx package (Carroll, 2018) in version 3.0.0.0 for the ordered logit.

Table 2.C.3: Marginal Effects at Mean: NHIS Dataset

Dataset		**Ordered Forest**					**Ordered Logit**				
Variable	Class	Effect	Std.Error	t-Value	p-Value		Effect	Std.Error	t-Value	p-Value	
Age	1	0.01	0.01	0.69	48.80		0.04	0.00	12.77	0.00	***
	2	0.31	0.20	1.55	12.07		0.18	0.01	20.08	0.00	***
	3	-3.76	3.10	-1.21	22.49		0.62	0.03	22.63	0.00	***
	4	-1.31	4.67	-0.28	77.93		0.00	0.01	0.15	87.78	
	5	4.75	5.63	0.84	39.88		-0.83	0.04	-23.38	0.00	***
Education	1	0.00	0.00	0.00	100.00		-0.09	0.01	-11.83	0.00	***
	2	0.00	0.00	0.00	100.00		-0.46	0.03	-16.90	0.00	***
	3	0.00	0.00	0.00	100.00		-1.60	0.09	-18.19	0.00	***
	4	0.00	0.00	0.00	100.00		-0.00	0.02	-0.15	87.78	
	5	0.00	0.00	0.00	100.00		2.16	0.12	18.63	0.00	***
Employed	1	-2.07	0.56	-3.73	0.02	***	-0.35	0.05	-7.22	0.00	***
	2	-1.79	1.39	-1.28	19.89		-1.69	0.21	-8.08	0.00	***
	3	-6.76	11.87	-0.57	56.88		-5.49	0.61	-8.94	0.00	***
	4	4.13	9.18	0.45	65.29		0.57	0.15	3.86	0.01	***
	5	6.49	16.14	0.40	68.74		6.96	0.73	9.52	0.00	***
FamilySize	1	0.08	0.09	0.95	34.06		-0.01	0.01	-0.81	42.00	
	2	-5.07	4.12	-1.23	21.81		-0.04	0.05	-0.81	41.95	
	3	3.81	15.41	0.25	80.45		-0.13	0.16	-0.81	41.94	
	4	2.96	33.98	0.09	93.07		-0.00	0.00	-0.15	87.99	
	5	-1.78	39.79	-0.04	96.43		0.18	0.22	0.81	41.94	
Female	1	-0.01	0.01	-0.58	56.51		0.02	0.03	0.68	49.85	
	2	0.21	0.79	0.27	78.66		0.09	0.13	0.68	49.81	
	3	-2.61	5.01	-0.52	60.28		0.30	0.45	0.68	49.80	
	4	-3.07	10.21	-0.30	76.38		0.00	0.00	0.15	88.10	
	5	5.47	11.73	0.47	64.10		-0.41	0.60	-0.68	49.80	
HealthInsurance	1	-0.00	0.02	-0.01	98.90		-0.09	0.04	-2.17	2.97	**
	2	-1.15	1.26	-0.91	36.12		-0.44	0.20	-2.20	2.77	**
	3	2.96	6.17	0.48	63.14		-1.52	0.68	-2.25	2.43	**
	4	-9.14	16.51	-0.55	57.96		0.05	0.05	0.98	32.83	
	5	7.34	17.48	0.42	67.48		2.01	0.87	2.30	2.16	**
Income	1	0.02	0.01	1.46	14.45		-0.00	0.00	-12.16	0.00	***
	2	-0.36	0.92	-0.39	69.70		-0.00	0.00	-18.01	0.00	***
	3	1.87	10.11	0.18	85.36		-0.00	0.00	-19.87	0.00	***
	4	-4.32	10.98	-0.39	69.37		-0.00	0.00	-0.15	87.78	
	5	2.80	16.54	0.17	86.56		0.00	0.00	20.36	0.00	***
NonWhite	1	0.03	0.03	0.92	35.56		0.30	0.04	6.99	0.00	***
	2	0.96	1.59	0.60	54.76		1.45	0.19	7.81	0.00	***
	3	1.98	6.55	0.30	76.22		4.76	0.56	8.48	0.00	***
	4	0.37	10.52	0.04	97.18		-0.41	0.12	-3.52	0.04	***
	5	-3.34	12.61	-0.27	79.10		-6.09	0.68	-8.94	0.00	***

Significance levels correspond to: $***. < 0.01$, $**. < 0.05$, $*. < 0.1$.

Notes: Table shows the comparison of the marginal effects at mean in % points between the *Ordered Forest* and the ordered logit. The effects are estimated for all classes, together with the corresponding standard errors, t-values and p-values. The standard errors for the *Ordered Forest* are estimated using the weight-based inference and for the ordered logit are obtained via the delta method.

Chapter 3

The Effect of Sport in Online Dating: Evidence from Causal Machine Learning

Co-authors: Daniel Boller, Michael Lechner

Abstract

Online dating emerged as a key platform for human mating. Previous research focused on socio-demographic characteristics to explain human mating in online dating environments, neglecting the commonly recognized relevance of sport. This research investigates the effect of sport activity on human mating by exploiting a unique data set from an online dating platform. Thereby, we leverage recent advances in the causal machine learning literature to estimate the causal effect of sport frequency on the contact chances. We find that for male users, doing sport on a weekly basis increases the probability to receive a first message from a woman by 50%, relatively to not doing sport at all. For female users, we do not find evidence for such an effect. In addition, for male users the effect increases with higher income.

Keywords: Online dating, sports economics, big data, causal machine learning, effect heterogeneity, Modified Causal Forest.

JEL classification: J12, Z29, C21, C45.

3.1 Introduction

Human interactions that have traditionally taken place in physical reality have increasingly shifted to the online world and the Covid-19 pandemic has substantially accelerated this trend. Human mating is also affected by this development, resulting in numerous novel formats of online dating. Indeed, online dating emerged as pivotal instrument for human mating. Rosenfeld, Thomas, and Hausen (2019), for instance, showed, that online dating represents the most common way for heterosexual couples to meet in the US. Cacioppo, Cacioppo, Gonzaga, Ogburn, and Vanderweele (2013) furthermore showed, that more than one-third of marriages in the US (2005-2012) are attributed to an initial contact via online dating.

Understanding the mechanisms that explain human mating in online dating environments is, in turn, decisive to elucidate the structure of societal evolution and to derive algorithms increasing the efficiency of the matching of potential partners. Explaining human mating in online dating environments relies essentially on the information that users share online, including socio-demographic, psychological, and physical traits. Indeed, previous research referred to socio-demographic (e.g., age; Hitsch, Hortaçsu, & Ariely, 2010a) and psychological (e.g., extroversion; Cuperman & Ickes, 2009) traits to explain human mating in online dating environments (for a detailed review, see Eastwick, Luchies, Finkel, & Hunt, 2014). Research considering physical traits, commonly interpreted as sport activity (Schulte-Hostedde, Eys, Emond, & Buzdon, 2012), to explain human mating in online dating environments, however, remains sparse even though few research provides indications that sport activity has substantial effects on human mating (Schulte-Hostedde et al., 2012). However, the effect of sport activity on human mating has not yet been fully understood. This paper attempts to fill this gap. In particular, this paper is, to the best of our knowledge, the first to investigate the causal effect of sport activity on human mating in online dating environments. It is also the first paper to analyze the heterogeneity of this causal effect using the novel causal machine learning methods.

Following this notion, we leverage unique data of more than 16'000 users, forming altogether almost 180'000 interactions. The data allows us not only to map interactions among users on a second-by-second basis, including visiting a user profile and contacting a user via private message, but also to observe more than 600 user characteristics describing the socio-demographic, psychological, and importantly, physical traits, including the frequency of the sport activity. This setting allows us to create a credible research design that eliminates potential sources of endogeneity by focusing on the first, one-way interactions between users, and by observing essentially the very same information, and even beyond, as an actual user. Hence, we can reliably identify the effect of sport activity on contact chances by relying on the conditional independence, i.e. the unconfoundedness research design. Moreover, we exploit recent advances in causal machine learning to estimate the causal effect of sport activity on contact chances in our large-dimensional setting in a very flexible way, while considering potential effect heterogeneities. In particular, we apply the Modified Causal Forest (Lechner, 2018), an estimator that reams the concept of Causal Trees and Forests, by allowing for multiple treatments, as applicable to our measure of sport activity. Furthermore, the Modified Causal Forest improves the splitting rule to account for selection bias and the mean correlated error. Additionally, it allows for estimation and inference on different aggregation levels in one estimation step. All of these aspects are crucial and beneficial for our research. Specifically, we can relax on the functional form assumptions, unlike classical parametric approaches, which is particularly important in large-dimensional settings as ours. Moreover, we can go beyond average effects and can flexibly investigate effect heterogeneities on various aggregation levels.

Leveraging the benefits of the Modified Causal Forest, we find different patterns for males and females. Particularly, for male users, we observe uniformly increasing contact chances by a potential

female partner, for increasing levels of sport activity. Specifically, the contact chances increase by more than 50% if male users practice sport on a weekly basis, relative to no sport at all. However, for female users, we do not find evidence for such an effect. Beyond the average effects, we uncover interesting effect heterogeneities both for males and females. In particular, for male users, we find that the effect of sport frequency on contact chances increases with higher income. This holds true for the income levels of the male users themselves, as well as for the income levels of the potential female partners. This implies that higher income male users enjoy a higher effect of a weekly sport activity, and that higher income female users value the regular sport activity of the potential male partners more. These heterogeneous effects are both statistically precise, as well as substantially relevant. In addition, for female users, we find indications that the effect of sport activity on contact chances increases with a higher sport frequency of the potential male partner. Furthermore, analysing the individualized effects provides additional descriptive evidence for these heterogeneous effects. It reveals further insights for potential heterogeneity mechanisms driven by education level or relationship preferences, among others. Lastly, a placebo test shows the robustness of our results.

This study contributes to research and practice as well as to the society. First, this paper provides new insights for the literature on human mating by demonstrating that sport activity, a key behavioral trait, affects human mating. Second, this paper supports social science research in assessing causal effects in large-dimensional data environments by showcasing an empirical approach, which allows for a very flexible estimation of average effects as well as a systematic assessment of underlying heterogeneities. Third, this paper helps individuals to increase their dating success by exhibiting how sport activities can contribute to the likelihood to be recognized by potential partners, finally highlighting the relevance of sport activity not only from a health but also from a human mating perspective. Finally, this paper serves product developers to improve the architecture of online dating platforms by highlighting the relevance of sport activity, while considering effect heterogeneities (e.g., demographic characteristics) at the same time.

This paper is structured as follows. Section 3.2 provides a short overview on prior work related to our research. Section 3.3 describes the online dating platform and the respective data. Section 3.4 explains the empirical approach, including the identification strategy and the estimation method. Following this, Section 3.5 presents the results, comprising the average and disaggregated effects. Section 3.6 discusses the results and the implications for research, practice, and society.

3.2 Literature

In this section we briefly describe prior work related to our research, comprising literature on sport activity in general as well as literature on sport activity and human mating.

3.2.1 Sport Activity

Sport activity has been ascribed relevant effects on human life, including physical and mental health as well as social outcomes, some of which are summarized next.

First, sport activity was shown to affect health outcomes. For instance Warburton, Nicol, and Bredin (2006) confirmed, based on an extensive review of the literature, that sport activity facilitates the prevention of several chronic diseases (e.g., cardiovascular disease and diabetes). In a similar vein, Humphreys, McLeod, and Ruseski (2014) found, that sport activity reduces self-reported incidences of diabetes, high blood pressure, heart disease, asthma, and arthritis (for a review, see Penedo & Dahn,

2005, and; Eime, Young, Harvey, Charity, & Payne, 2013).

Second, sport activity was demonstrated to enfold effects on mental health. Hillman, Erickson, and Kramer (2008), for instance, showed that sport activity enhances cognition and brain functions (for a review, see Strong et al., 2005, and; Janssen & LeBlanc, 2010). Moreover, sport activity was shown to increase self-reported life satisfaction and happiness (Huang & Humphreys, 2012; Ruseski, Humphreys, Hallman, Wicker, & Breuer, 2014).

Third, sport activity was proven to affect social outcomes. For instance, Caruso (2011) showed that sport activity decreases property and juvenile crime among young adults. Moreover, sport activity was found to enfold positive effects on economic outcomes such as wages and earnings (e.g. Lechner, 2009; Rooth, 2011), human capital (Steckenleiter & Lechner, 2020), and quality of work performance (Pronk et al., 2004). Finally, sport activity has been confirmed to lead to higher academic achievements (Fox, Barr-Anderson, Neumark-Sztainer, & Wall, 2010; Pfeifer & Cornelißen, 2010; Felfe, Lechner, & Steinmayr, 2016; Lechner, 2017; Fricke, Lechner, & Steinmayr, 2018), to positively affect concentration, memory and classroom behavior (Trudeau & Shephard, 2008), and to improve social relations (Stempel, 2005).

The effects of sport activity are, thus, explored in various spheres of human life. The effect of sport activity on human mating, however, is almost unexplored, as discussed next.

3.2.2 Sport Activity and Human Mating

Research on human mating has established in sociology, psychology, economics, and, more recently, computer science, mostly attributable to the range of potential explanatory factors that determine human mating (Eastwick et al., 2014) and novel data opportunities due to computer-mediated approaches for human mating (i.e., online dating). In addition to various studies referring to socio-demographic and psychological characteristics to explain human mating (for a detailed review, see Eastwick et al., 2014), a few studies also consider sport activity as potentially relevant factor in explaining human mating.

Schulte-Hostedde, Eys, and Johnson (2008) studied the effect of males' practiced sport discipline on females' willingness to engage in a relationship, applying an experimental setting. The authors showed that *'[...] team sport athletes were perceived as being more desirable as potential mates than individual sport athletes and non-athletes'* (p. 114). Moreover, the authors argued that *'team sport athletes may have traits associated with good parenting such as cooperation, likeability, and role acceptance'* (p. 114) to explain the positive effect of team sport participation on desirability. However, the authors restrict sport activities to a particular type of sport, namely team vs. individual sport, which, in turn, impedes a valid assessment of the general effect of sport activity on human mating. In a similar vein, Farthing (2005) showed, also applying an experimental setting, that *'[...] females and males preferred heroic sport risk takers as mates, with the preference being stronger for females'* (p. 171) , while interpreting (non-) heroic sport risk as, for example, engaging in (non-) risky sport activities. However, the previously raised concerns apply in the same way to the findings by Farthing (2005).

Further research provides insights on potential indirect effects of sport activity on human mating. In particular, previous research indicated that sport activity improves, inter alia, attractiveness (Park, Buunk, & Wieling, 2007), health (Warburton et al., 2006), and income generation (Lechner, 2009), all of which have been shown to affect human mating (e.g. Hitsch et al., 2010a; Hitsch, Hortaçsu, & Ariely, 2010b; Eastwick et al., 2014). However, these studies remain inconclusive with respect to human mating, given the missing integration of relevant context-factors (i.e., further relevant personal/sport characteristics) affecting human mating.

Taken together, sport activity seems relevant for explaining human mating. However, a conclusive, finally valid, assessment on the effect of sport activity on human mating is missing, given that previous research assessed the effect of sport activity on human mating either in the absence of potentially relevant socio-demographic characteristics or by utilizing a narrowed interpretation, respectively representation, of sport activity. These limitations surprise given that information on sport activity are one of the most articulated and visible features on online dating platforms. Furthermore, as discussed previously, sport activity is ascribed relevant effects on various spheres of human life, including physical and mental health as well as social and economic conditions. Following the above mentioned limitations, we focus on the analysis of the effect of sport activity on human mating.

3.3 Setup and Data

In the course of this research, we collaborated with a German online dating platform operator. The operator provided us both with information on the functionality as well as with data from the online dating platform.

3.3.1 Online Dating

The online dating platform allows a user to virtually meet and communicate with other users. The user has to pay a monthly fixed subscription fee to register and to utilize the online dating platform. The registration at the online dating platform is subdivided into three major sections. First, the user is requested to provide socio-demographic information (e.g., sex, age, education, and income). Second, the user is requested to specify search criteria for potential partners (e.g., sex, age, education, and income). Third, the user is requested to answer a personality test that relates to the users' life style, personality, attitudes and views (79 categories in total). Moreover, the user articulates the language preferences and may include one or more photos on the personal profile page. However, these photos remain fully blurred until the user decides to release the photo for the potential partner.[1] Most importantly, with specific regard to the intended analysis, a user articulates her/his sport preferences and actual sport activities within a total of 27 disciplines, how often she/he actively practices sport, and, finally, which recreational activities dominate in her/his leisure time. A detailed description of the survey questions and the corresponding variables together with descriptive statistics can be found in Appendix 3.D.

Following the registration at the online dating platform, the user can define a query, indicating the preferred sex, age, and geographic location to explore potential partners. The search query returns a shortlist of potential partners, who correspond to the previously defined qualifications. The shortlist includes the potential partners' username, age, a blurred version of the photo, and a matching score, which is computed by the online dating platform operator in order to support users in finding a potentially fitting partner.[2] The user can investigate the potential partner in detail by browsing on the potential partners' profile page, which displays a blurred version of the photo as well as information on the previously described survey. The user can then choose from multiple possible actions. As such, the user can either send a private text message, a 'Smile' icon, or a 'Smile Back' icon (if initially received a 'Smile' icon) to a potential partner. Additionally, a user may leave a 'like' or a text note on a potential partners' profile page. Moreover, the user can initiate a friendship with a potential partner by initiating a profile release

[1] In our analysis we restrict the user interactions by excluding the actions involving the release of the blurred photo. We discuss this point in Detail in Section 3.4.2.

[2] The online dating platform operator does not provide the formula to calculate the matching score. However, it provided us with all data required for its calculation. We elaborate more on this point in Section 3.4.2.

or accepting an initial profile release by a potential partner. Furthermore, a user may request an 'Applet' (game with questions) to a potential partner, which works out similarities/differences between the user and the potential partner. Finally, a user may prevent unwanted users from contacting in any form.

3.3.2 Data

The acquired data consists of two samples. The first, *user sample*, contains personal information about the registered users on the platform. The second, *interaction sample*, contains information about the users' interactions on the platform.

The user sample includes 18'036 newly registered users who joined the platform between January 1st, 2016 and April 30th, 2016.[3] For each registered user, we observe the full information filled upon the registration, which comprises 667 variables in total. For our intended analysis with regard to the sport activity, we exclude the users with daily sport frequency, as these comprise only around 3% of all users, which would prevent a meaningful analysis for this group. Furthermore, we restrict ourselves to the sample of users, whose residency is located in Germany, as only for these users we observe full location information, including the ZIP codes. This restriction affects only about 2% of the observations as the platform provider operates on the German market. Lastly, we exclude users with incomplete information (around 1% of the sample) and those with implausible and inconsistent values (less than 1% of the sample).[4] This leaves us with an available sample consisting of 16'864 users for our analysis. A descriptive summary of selected variables for the user sample is presented in Appendix 3.A.

The interaction sample includes 1'415'645 user actions among the population of newly registered users over the same time period. For each action, we observe the IDs of both users involved in the action, as well as the precise time stamp and the type of action. Each interaction between users must begin with a visit action (invisible to a user), upon which further types of actions are possible, such as a message, like or smile (visible to a user). We refer to the user who initiates an interaction as a *sender* of an action, and the user who gets involved in an interaction as a *recipient* of an action. For the purposes of our analysis, we filter the interactions such that we consider only *one-way* interactions initiated by a visit action, with either no further action at all, or immediately followed by any visible action from the sender, without considering any visible recipient's response to the initial action from the sender. Thus, we select only unique interactions in the sense that the sender was visibly or invisibly active, while the recipient stayed visibly passive. Thereby, we restrict the interactions between the users until the point of a possible reciprocal interaction taking place.[5] This selection of the sample will be later important for the validity of our identification strategy (see Section 3.4.2 for details). We further shape our sample such that each observation represents a valid interaction accompanied by indicators of visible sender actions that have taken place within the particular interaction as well as the sender and recipient user IDs. This leaves us with an available sample consisting of 178'372 valid unique interactions for our analysis.

Lastly, to construct our final estimation sample, we merge the interaction sample with the user sample. As a result, each observation in our estimation sample represents a valid interaction between two users and consists of sender and recipient user IDs together with sender's actions from the interaction sample, and both the sender's as well as recipient's characteristics obtained from the user sample. Furthermore, as the data contains only heterosexual users based on a binary measure for gender, i.e. we never observe a sender and recipient of the same sex in our sample, we split the sample based on gender for a clearer interpretation of the results. Hence, we refer to the sample with only female recipients as

[3] Other empirical studies using online dating data focused on observation periods of similar length (see Hitsch et al., 2010a, and; Hitsch et al., 2010b).
[4] This includes, for example, users with more than one single value for a mutually exclusive answer selection, among others.
[5] For a more detailed definition of valid user interactions with practical examples, see Appendix 3.B.

the *female sample*, as here the females are in the role of an approached user upon receiving a visit action, and possibly further actions, by a male sender of an action. Analogously, we refer to the sample with only male recipients as the *male sample*, as in this case the males are in the role of an approached user upon receiving a visit action, and possibly further actions, by a female sender of an action.

Thus, we are left with 108'456 observations for the female sample and with 69'916 observations for the male sample. The corresponding descriptive statistics of selected variables for the two samples are listed in Appendix 3.A.

3.3.2.1 Sport Activity

In order to investigate the effect of sport in online dating, we leverage the rich information set regarding the sport activities on the user profile. In particular, each profile includes a detailed statement of the user's sport frequency. This information stems from the initial questionnaire filled by the user upon registration. First, the user is asked about the sport types done actively, namely: *'What sports do you do actively?'*, with multiple options (mutually inclusive) such as basketball, fitness, hiking, soccer, tennis, etc., or specifying the option *'none'*. Second, only if the user has not specified the option *'none'*, a further question regarding the particular sport frequency is asked: *'How often do you practice sport?'*. The possible values (mutually exclusive) include the following answers: *'every day', 'several times a week', 'several times a month',* or *'less common'*. Thus, we not only observe the user's binary indication of practicing sport or not, i.e. the extensive margin, but also the particular sport frequency, i.e. the intensive margin. This provides us with a much finer measure of the actual sport activity. Accordingly, we define the sport activity measure to be multi-valued with sport frequencies of *weekly, monthly, rarely* and *never*. We omit the daily frequency for lack of data within this category, as previously mentioned. Furthermore, we leave the sport types out of consideration too, as these include many different and not mutually exclusive values, which prevents a clear separation of the categories.

Table 3.3.1 shows the descriptive statistics for the sport frequency shares in the samples of males and females, respectively as well as the corresponding shares from the innovation sample of the German socioeconomic panel (SOEP-IS; Richter, Schupp, et al., 2015) for a comparison with a representative population sample.[6] First, we see that the sport frequency is unevenly distributed in both samples. Second, we can also observe that the shares are very similar in both samples. Nonetheless, the *never* category is more represented in the female sample, while the *weekly* category is more represented in the male sample. Additionally, we also observe that the subjective sport frequency of the users from the online dating platform is in general much higher than the one of the representative individuals from Germany.[7] Third, with respect to the number of observations in the corresponding samples, we immediately see that even though we have a balanced user sample in terms of gender,[8] females get visited more often than males do.

[6] Similar values for the sport frequency statistics for Germany are documented also in the Eurobarometer Survey (Eurobarometer, 2014), as pointed out by Steckenleiter and Lechner (2020).

[7] Note, that this might be both due to truly higher sport frequency of the registered users as well as due to an overestimation of own actual sport frequency of the users, or the combination of both. Also note, that our sample consists only of singles, which is in contrast to the representative population sample.

[8] The user sample consists of 48% of females and 52% of males. For more descriptive statistics with regard to the user sample, see Appendix 3.A.

Table 3.3.1: Shares of Sport Frequency for Male and Female Sample

	Never	*Rarely*	*Monthly*	*Weekly*	Observations
Males	0.07	0.08	0.29	0.56	69'916
Females	0.12	0.09	0.29	0.49	108'456
SOEP-IS	0.33	0.23	0.10	0.34	25'544

Note: Color intensity represents the corresponding share sizes for males and females.

Finally, given our definition, the impact of sport activity can be illustrated as follows. The user, here the sender, visits a profile of another user, here the recipient, and gets exposed to an information revealed on the profile. Among other indicators, the sender observes the recipient's indication of the sport frequency, i.e. the variable of interest. Based on the available information, the sender then decides to perform or not to perform a further action.

3.3.2.2 The Interaction between Users

We are interested in the one-way actions of a sender upon visiting a recipient's profile on the website. Even though there are multiple actions a sender can initiate, we focus explicitly on the action of sending a text message for several reasons. First, a text message is the most evident action of showing a serious interest, as in order to compose a text message, the sender has to exhibit a substantial effort, in comparison to other available options, such as simply sending a smile or like. Second, unlike the other generic options, by sending a text message, the sender directly approaches the recipient in an individualized manner. Third, an outcome measure of sending a text message or an email has been previously used in the online dating literature under the assumption that users send a message if and only if the potential utility of the match exceeds some minimum threshold value (compare e.g. Hitsch et al., 2010a; or Bruch, Feinberg, & Lee, 2016). Hence, we define our action of interest as a binary measure of sending (1) or not sending (0) a text message upon a profile visit. Given the binary scale, the natural interpretation as contact chances in terms of message probabilities arises.

Figure 3.3.1: Average Contact Chances according to Sport Activity for Males and Females

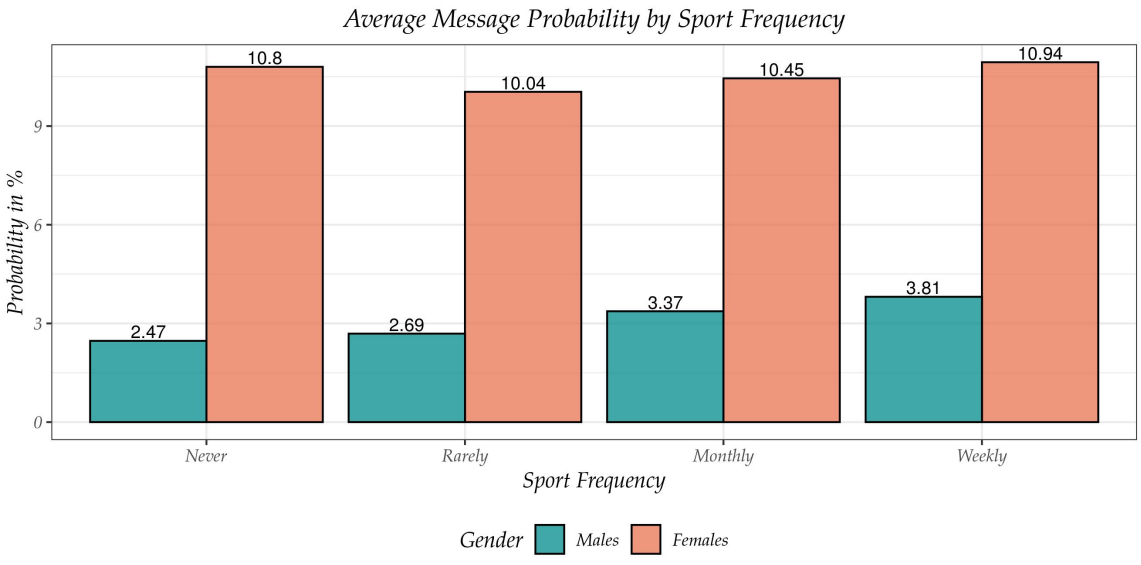


Figure 3.3.1 shows the average message probability in percentages for males and females according to

the sport frequency. First, we see that the levels of females are substantially higher than those of males, i.e. women have unconditionally a higher probability to get messaged than men do. This is in line with previous evidence from studies based on online dating data (Bruch et al., 2016). Second, we observe a slightly increasing message probability with higher sport frequency for males, while for females no clear pattern can be identified.

3.3.2.3 Information about Users

In our sample, we have access to complete information filled by the user upon registration. Hence, we not only observe the condensed information displayed on the main user profile page, but also the expanded information stored in the background of the user profile. Thus, we effectively observe the very same information that a real user observes upon a profile visit of a potential partner, and even beyond. The full information observable to us includes the following components. First, we observe the user's demographic information such as gender, age, height, etc., the socio-demographic information such as education and income level, type of occupation, etc., as well as personal information such as place of residence, smoking habits, or even (self-judged) appearance. Second, in addition to the user specific information, we observe the user's preferences for a potential partner in terms of the search criteria related to the above mentioned socio-demographic information as well. Third, we furthermore observe the user's information stemming from the detailed personality test, which reflects on the user's life style, personality, attitudes and preferences. This includes an extensive information on topics like religion, political views, music and travel preferences, or even partner requirements. The aforementioned user information comprises of an exhaustive list of 663 variables in total. However, given the structure of our data, we include the user information both for the recipient as well as for the sender, resulting in effectively more than thousand variables. Apart from the information coming directly from the platform, we additionally generate a variable measuring the distance between the recipient and the sender, based on the available ZIP codes.[9]

We consider all the aforementioned variables as controls in the sense of potential confounders, i.e. as variables jointly influencing both the recipient's sport activity as well as the recipient's potential outcome of receiving or not receiving a text message, and thus, *de facto* the sender's action to contact or not to contact the recipient. Conditioning on such a large-dimensional covariate space is a challenging estimation task. However, we refrain ourselves from an arbitrary selection of the confounding variables in order to reduce the dimension of the estimation problem. Rather, we apply a novel causal machine learning estimator, which can effectively deal with such large-dimensional setting, performing implicit variable selection in a flexible and data-driven way. The only variable deselection we perform manually is related to endogenous variables.[10] Thus, we remove all variables that could be potentially influenced by the sport frequency. These include mainly variables indicating the specific sport type, but also variables describing sport-related choices such as holiday and leisure time preferences, as well as variables regarding the body type and clothing style. In total, we dismiss 38 endogenous variables. Lastly, we leave out 2 variables without any variation. As a result, we are left with 1247 covariates in total (1229 ordered, including dummies and 18 unordered), reflecting the recipient and sender characteristics.

Apart from the confounding role, the covariates are useful for analysing the effect heterogeneity, too. For this purpose, we pre-specify a small subset of heterogeneity variables, consisting of age, income and education level on both recipient as well as sender side, together with the corresponding distance between the recipient and the sender. We focus on these heterogeneity variables for two main reasons. First, these

[9] The average distance between the recipient and the sender in our sample is 67.32 km. A detailed plot of the distribution of the distance between the users can be found in Appendix 3.A.

[10] We elaborate on this issue more closely when discussing the identifying assumptions in Section 3.4.2.

socio-demographic information are widely recognized in the literature as being the main determinants of the partner choice (for a review of the importance of selected socio-demographic characteristics see Hitsch et al., 2010a; and Eastwick et al., 2014). Second, these are also the main variables that are most visible to the user on the profile summary and thus can potentially impact the shape of the effect. Additionally, we analyze the heterogeneous effects also along the sport frequency for the recipient as well as for the sender. Complementary to the pre-specified subset of heterogeneity variables, the remaining variables might serve for a supplementary descriptive analysis of the effects.

3.4 Empirical Approach

To analyze the effect of sport activity on human mating, we leverage the recent advances in the causal machine learning literature. Below, we outline the parameters of interest together with the identification and estimation thereof.

3.4.1 Parameters of Interest

In order to define the parameters of interest, we rely on the Rubin's (1974) potential outcome framework. We denote the treatment variable of a user i by D_i, which in our case can take on four different integer values, i.e. $D_i \in \{0, 1, 2, 3\}$, corresponding to sport frequencies of *never, rarely, monthly,* and *weekly*, respectively. According to the treatment status, d, we define the potential outcomes for the user i by Y_i^d, which in this case is the action of receiving or not receiving a text message. However, we only observe the potential outcome under the treatment which the user i is associated with (see Holland, 1986, for a discussion of the fundamental problem of causal inference). Thus, the realized outcome can be defined through the observational rule as follows: $Y_i = \sum_{d=0}^{3} \mathbb{I}(D_i = d) \cdot Y_i^d$, which implies that we observe the action of receiving the text message only under a particular sport frequency of the recipient. Further, we denote the observed vector of covariates by X_i, which contains the recipient and sender characteristics, together with a subset of pre-specified heterogeneity variables Z_i, such that $Z_i \subset X_i$.

To analyze the effect of sport frequency on the message probability, we are interested in the following causal parameters. First, the *Average Treatment Effect (ATE)* of treatment $D_i = m$ compared to treatment $D_i = l$ is defined as

$$ATE = \theta = \mathbb{E}[Y_i^m - Y_i^l]$$

and constitutes the classical parameter of interest in microeconometrics, which provides us with an aggregated effect measure (compare e.g. Imbens & Wooldridge, 2009). Second, the *Group Average Treatment Effect (GATE)* is characterized as

$$GATE = \theta(z) = \mathbb{E}[Y_i^m - Y_i^l \mid Z_i = z]$$

and measures the differential effects along the heterogeneity variables Z_i. Thus, it provides us with a disaggregated effect measure according to the specific variables of interest, as in our case is the age, income and education level, distance as well as the sport frequency itself. In the latter case, the GATE corresponds to the *Average Treatment Effect on the Treated (ATET)*. Third, the *Individualized Average Treatment Effect (IATE)* is denoted as

$$IATE = \theta(x) = \mathbb{E}[Y_i^m - Y_i^l \mid X_i = x]$$

and describes the heterogeneous effects based on the full set of observed covariates X_i. As such, the

IATEs present the disaggregated effects on the finest level of granularity and thus provide us with user-type specific effects.

Notice, that both the treatment variable, i.e. the sport frequency, as well as the outcome variable, i.e. receiving a text message, are measured on the recipient side, and hence, also the above defined causal effects refer to the recipient.

3.4.2 Identification Strategy

Given our observational study design, it is not possible to only compare the unconditional message probabilities for different sport frequencies, as displayed in Figure 3.3.1, to infer the causal effects, since the user decision regarding the sport activity is not random. The level of sport frequency might be influenced by other variables representing socio-demographic information, which might also influence the potential outcome of receiving or not receiving a text message. For example, recipients with a higher level of education might have a higher probability of doing sport on a weekly basis, as well as a higher probability of getting messaged. This phenomenon is known as selection bias (Imbens & Wooldridge, 2009). In order to disentangle the causal effect from the selection effect, we need to eliminate such confounding via credible identification strategy.

For the identification of the aforementioned parameters of interest in a multiple treatment case, we rely on the so-called *selection-on-observables* strategy (see Imbens, 2000; or Lechner, 2001). Such identification approach assumes that all confounding variables jointly influencing both the treatment as well as the potential outcomes are observed, and thus, can be conditioned on. Given our rich data on user characteristics and the unique research design, we argue to capture all possible confounding effects for two main reasons. First, for both the recipient and the sender, we observe socio-demographic (e.g., age, education, income) and personal (e.g., family status, smoking habits, place of residence) characteristics, together with the preferences for a potential partner as well as the answers given in a detailed personality test. Thereby we have access to even richer personal information than the actual users when browsing the profiles, and as such, we are able to control for confounding effects stemming from the user's characteristics. Second, given our research design, focusing only on the very first one-way interactions between the recipient and the sender, we effectively eliminate any possible unobserved effects coming from the reciprocal interaction between the users such as sympathy or kindness. By doing so, we explicitly focus only on situations, in which the recipient's profile gets visited by a sender, upon which the recipient does receive or does not receive the very first text message from the sender, without any visible encouragement to do so from the recipient her/him-self. In such a situation, the sender decides solely based on the information visible on the recipient's profile to send or not to send the message. Within our research design, we observe exactly the same information, and even beyond, as the actual sender when facing the decision of sending the first text message. For this reason, we are also able to control for confounding effects stemming from the user's interaction.

Taken together, combining the highly-detailed user information, which exceeds the information directly observable by the actual users, with the unique research design, which eliminates any possible unobservable information, we are confident to capture all confounding effects. In particular, our selection-on-observables strategy relies on the following set of identification assumptions.

First, the so-called *conditional independence* assumption (CIA), states that the potential outcomes and the treatment are independent once conditioned on the covariates. This hinges on the availability of all covariates that jointly influence the potential outcome and the treatment. As we argue, we observe sufficiently rich information on both the recipient as well as the sender side to ensure the plausibility

of the CIA. In addition, our research design eliminates any further influence from a possible reciprocal interaction between the users. Thus, we are confident about the validity of the CIA in this particular case. There are only two potential sources of vulnerability of the CIA in this case. First, it could be caused by the availability of the blurred photo of the user. Even though the photo remains blurred, as we do not allow interactions between the users which would include the action to release the photo, we cannot rule out that information such as the shape of the face or the hair and skin colour could be, nonetheless, inferred. However, despite the information inferred from the blurred photo might possibly affect the outcome, i.e. the message probability of the recipient, we argue that this information should not have an effect on the treatment itself, i.e. the recipient's sport frequency. Thus, it arguably does not qualify as a potential confounder. Nevertheless, limitations in the availability of profile pictures, respectively opportunities to represent the information in profile pictures, are common in the literature on online dating (Fiore, Taylor, Mendelsohn, & Hearst, 2008). Second, it could be caused by the availability of the matching score. However, despite the fact, that we do not observe the score directly, we know that we observe, and indeed condition on, all information which serves for its calculation. Moreover, even though we do not know the exact formula, by using a very flexible estimation approach, we are able to reproduce any arbitrary functional form of the matching score. Nonetheless, if the matching score would consist of the user's sport frequency, the treatment would be indirectly observed as a part of the shortlist of potential partners even before actually visiting the user profile. However, this would not violate the CIA as such, it could rather potentially reduce the size of our effect estimates. For this reason, we conduct a placebo test to provide evidence that this is indeed not the case. We discuss the placebo test in more detail in Section 3.5.3.

Second, the *common support* assumption, ensures that for each value in the support of the covariates, there is a possibility to observe all treatments. This means that we find users with the same age, education, income, etc., for all sport frequency levels. Thus, we are able to check the validity of the common support assumption in the data directly, but do not find any violations thereof (see Lechner & Strittmatter, 2019, for a discussion of common support issues).

Third, the *stable unit treatment value* assumption (see e.g. Rubin, 1991), implies that for each user we observe only one of the potential outcomes based on the treatment status. It further implies that there is no interference among users, hence ruling out any general equilibrium or spillover effects. This means that the sport frequency of one particular user does not affect the message probability of other users. We argue that the SUTVA is plausible in this case, as we analyze only a short time period after the user registration such that general equilibrium or learning effects would not yet emerge.

Fourth, the *exogeneity of confounders* assumptions, indicates that the values of the covariates are not influenced by the treatment. In other words, the user characteristics should not be impacted by the sport frequency. For this reason, we discard all potentially endogenous variables such as indicators of particular sport type, sporty clothing style, preferences for sport holidays or sport club memberships. Therefore, we are confident that the exogeneity assumption holds.

Under the aforementioned assumptions, it can be shown that the above parameters of interest are identified. For technical details, see Lechner (2018).

3.4.3 Estimation Method

In our analysis, we face two major challenges with regard to the estimation of the causal effects of interest. First, we need to deal with a very large conditioning set with an unknown functional form of the covariates. Second, we want to investigate potential effect heterogeneity. In order to overcome these

challenges, we take advantage of the newly developing causal machine learning literature (see Athey, 2018; Athey & Imbens, 2019; or Knaus, Lechner, & Strittmatter, 2021, for overviews). It combines the flexibility and prediction power of machine learning (Hastie, Tibshirani, & Friedman, 2009) with the causal inference from econometrics (Imbens & Wooldridge, 2009). One of the most popular machine learning methods are the so-called regression trees (Breiman, Friedman, Olshen, & Stone, 1984) and random forests (Breiman, 2001). The trees and forests are highly flexible, local nonparametric prediction methods, which can effectively deal with large-dimensional settings (Biau & Scornet, 2016). Adapting these prediction algorithms towards causal inference has lead to developments of Causal Trees (Athey & Imbens, 2016) and Causal Forests (Wager & Athey, 2018), respectively. These methods inherit the advantages of the prediction versions, while flexibly estimating the causal effects with systematically uncovering their heterogeneity. Furthermore, Lechner (2018) extends the Causal Forest for the multiple treatment case, and additionally improves the splitting rule to account for selection bias and for the mean correlated error. The resulting Modified Causal Forest also allows for estimation as well as inference for the parameters of interest at all aggregation levels in one estimation step. Since our application involves multiple treatments with potential confounding, while analyzing various heterogeneity levels of the causal effects, we opt for the latter approach.

In our analysis, we rely on estimating the so-called 'honest' forest, which has been shown to lower the bias of the causal effect estimates and to enable valid statistical inference (Wager & Athey, 2018, and; Lechner, 2018). As such, we randomly split the estimation sample in two equally sized parts and use one sample, i.e. the *training* sample, to build the Modified Causal Forest and the other sample, i.e. the *honest* sample, to estimate the causal effects.[11] Then, the estimation procedure of the Modified Causal Forest can be described as follows. First, the estimator draws a random subsample s of the training sample and subsequently estimates a single causal tree. As such, the subsample gets recursively splitted into smaller subsets, the so-called 'leaves' of the tree $L(x)$. The partitioning follows a splitting rule which removes selection bias and reveals effect heterogeneity. As a result, the observations are homogeneous with regard to the covariate values *within* the leaf, while being heterogeneous *across* the leaves. Then, the treatment effect is estimated within each terminal leaf by simply subtracting the mean outcomes of the respective treatment levels $D_i = m$ and $D_i = l$ from the honest sample as

$$\hat{\theta}_s(x) = \frac{1}{\{i : D_i = m, X_i \in L(x)\}} \sum_{\{i:D_i=m, X_i \in L(x)\}} Y_i \quad - \quad \frac{1}{\{i : D_i = l, X_i \in L(x)\}} \sum_{\{i:D_i=l, X_i \in L(x)\}} Y_i.$$

Second, as a single tree might be quite unstable due to its path-dependent nature, the forest estimates many such trees by drawing S random subsamples in total. The Causal Forest estimate is then given by the ensemble of many causal trees as

$$\widehat{IATE} = \hat{\theta}(x) = \frac{1}{S} \sum_{s=1}^{S} \hat{\theta}_s(x).$$

The additional averaging of the trees helps to reduce the variance and to smooth the edges of the leaves (Bühlmann & Yu, 2002). Conceptionally, the Causal Forest can be thought of as a nearest neighbor matching estimator with an adaptive neighbor choice and can be thus described using a weighted representation, too (Wager & Athey, 2018; Athey, Tibshirani, & Wager, 2019).

Third, the Modified Causal Forest estimates the GATEs by averaging the IATEs in the corresponding

[11] Lechner (2018) shows in a simulation study that the efficiency loss of the 'honest' forest due to sample-splitting is minimal in comparison to the case of 'honest' trees as in Wager and Athey (2018).

subsets defined by the heterogeneity variables Z_i and the ATE by averaging the IATEs in the whole sample as follows

$$\widehat{GATE} = \hat{\theta}(z) = \frac{1}{\{i : Z_i = z\}} \sum_{\{i:Z_i=z\}} \hat{\theta}(X_i)$$

and

$$\widehat{ATE} = \hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}(X_i).$$

Thus, it provides a computationally attractive option to estimate the effects of interest on all desired levels of heterogeneity without the need for re-estimating the whole forest for each single aggregation level.[12]

Fourth, the Modified Causal Forest then explicitly uses the weighted representation of the estimated effects for inference. The weight-based inference can be then conveniently applied to all aggregation levels as well.[13] For an in-depth discussion of the Modified Causal Forest, see Lechner (2018) as well as Cockx, Lechner, and Bollens (2019) and Hodler, Lechner, and Raschky (2020) for empirical applications.

Estimating the effects of sport frequency on the message probability by applying causal machine learning allows us to improve on previous empirical studies in an online dating setting in several dimensions. First of all, we do not have to specify the exact functional relationship between outcome, treatment and covariates, as in the case of using parametric approaches such as the logistic regression (see e.g. Hitsch et al., 2010a; Hitsch et al., 2010b; or Bruch et al., 2016). This is particularly important when dealing with a large-dimensional covariate space, including the characteristics of both the recipient and the sender, as the functional form of the interactions thereof is not *a priori* clear. Furthermore, using causal machine learning also advances the semiparametric approaches used in online dating studies (see e.g. Lee, 2016, for a matching estimation), thanks to more flexible adaptive estimation and its implicit variable selection properties. Lastly, causal machine learning allows us to go beyond the average effects and systematically investigate the effect heterogeneity on various aggregation levels, without the need to specify interactions or to build subsets of data in an *ad-hoc* fashion.

3.5 Results

Below, we present the results for the average and heterogeneous effects of sport activity on contact chances, based on the Modified Causal Forest estimation.

3.5.1 Average Effects

The results for the average effects of the sport activity on the contact chances are summarized in Table 3.5.1. The diagonal presents the potential outcomes, while the corresponding effects are depicted in the lower triangle.

In case of the male sample, for increasing sport frequency, the results show a clear and increasing pattern of the potential outcomes, i.e. of the potential message probability. While the potential message

[12] In our setting, we additionally apply treatment sampling probability weights for the ATE and GATEs aggregation of the IATEs to account for the unbalanced treatment shares.

[13] Athey et al. (2019) further suggest usage of the forest weights for solving many different econometric estimation problems.

probability for users who never practice sport is on average only 2.50%, for users doing sport on a weekly basis, the chances to get messaged increase by more than 50% and amount to 3.82%. Comparing the respective potential outcomes across the sport frequency levels yields the corresponding causal effects measured in percentage points. Accordingly, all effects for all sport frequency comparisons are positive. The most sizeable and the most precise effects are estimated for the most distinct sport frequencies, as one would intuitively expect. Thus, the average effect of a weekly sport activity versus no sport activity at all, is equal to an 1.32 percentage points increase. Similarly, the average effect of a weekly in comparison to only rare sport activity amounts to an 1.20 percentage point increase. Moreover, these effects are both substantively as well as statistically relevant. As such, a male user increasing his sport activity from no sport or only rare sport activity to doing sport on a weekly basis significantly increases the probability of getting messaged by 52.80% and 45.80%, respectively. In practice, this implies receiving 13, respectively, 12 extra messages out of 1000 profile visits. Hence, the contact chances of a male user can be substantially increased solely by becoming more sporty. The remaining effects comparing less distinct sport frequencies lack the statistical relevance, which stems mainly from the substantially lower number of observations for these categories (see Table 3.3.1).

Regarding the female sample, the results do not suggest increasing contact chances with increasing frequency of sport activity, as in the case of the male sample. The potential outcomes thus do not indicate any clear pattern as the message probability firstly drops, when switching from no sport to rare sport activity, and then increases steadily throughout the monthly and weekly sport frequencies, reaching comparable levels with the category of never doing sport. Accordingly, the estimated average effects do not show any explicit structure and lack statistical relevance. The only exception is the precise estimate of the effect of the weekly vs. rare sport activity, with a sizeable increase of an 1.61 percentage points, yet this represents only a minor relative increase of 17.18% in comparison to the effects seen in the male sample. Taken together, based on the overall results, no substantial conclusions can be drawn.

Table 3.5.1: Average Effects of Sport Activity on the Contact Chances for Males and Females

	Males				Females			
	Never	*Rarely*	*Monthly*	*Weekly*	*Never*	*Rarely*	*Monthly*	*Weekly*
Never	2.50				10.67			
	(0.46)				(0.60)			
Rarely	0.12	2.62			-1.30	9.37		
	(0.63)	(0.43)			(0.88)	(0.63)		
Monthly	0.86	0.74	3.36		-0.29	1.01	10.38	
	(0.52)	(0.50)	(0.26)		(0.71)	(0.72)	(0.40)	
Weekly	**1.32*****	**1.20****	0.46	3.82	0.31	**1.61****	0.60	10.98
	(0.50)	(0.47)	(0.32)	(0.19)	(0.67)	(0.68)	(0.45)	(0.30)

Note: Effects in % points. Potential outcomes on the diagonal. Standard errors in parentheses. Significance levels refer to: $^{***} < 0.01$, $^{**} < 0.05$, $^{*} < 0.1$. Color intensity represents the corresponding level sizes.

In general, based on the results of the average effects, we find sizeable and significant positive effects of a more frequent sport activity, when analyzing the male users, while we find only weak evidence for such effects for the case of female users. It means that for men a higher sport frequency substantially increases the probability of getting messaged by a woman, on average. However, higher sport frequency for women does not seem to consistently lead on average to considerably higher chances of getting messaged by a man.

3.5.2 Heterogeneous Effects

While the average effects provide a general measure for the causal effects of sport activity, a more detailed description of the effect heterogeneity beyond gender, remains unknown. Therefore, we study the heterogeneous effects in respect to the pre-defined set of variables of interest, i.e. the group average treatment effects (GATEs), to uncover possibly differential effects of the sport activity on the contact chances. For the sake of clarity, we focus on the effects comparing the most distinct cases, namely the weekly sport frequency with no sport activity. In this regard, we analyze the effect heterogeneity along age, education and income of the users, on both the recipient as well as the sender side, together with the mutual user distance, based on the following considerations. First, these variables have been previously identified as the main determinants of the partner choice (Hitsch et al., 2010a; Eastwick et al., 2014) and second, these are also the variables which appear on the main profile summary. Thus, we expect these variables to have a higher potential to influence the shape of the effect of the sport activity. Lastly, we investigate the effect heterogeneity based on the particular recipient's as well as sender's sport frequency, which is a natural choice as it corresponds to the effect on the treated, a classical microeconometric parameter of interest (compare e.g. Abadie & Cattaneo, 2018). Essentially, the heterogeneity analysis enables us to investigate if the benefits of the regular sport activity in terms of higher contact chances vary among specific groups of users. Thus, we shed light on the open questions such as if potentially the users with higher age, or with lower education and income level enjoy higher benefits of weekly sport activity than those with lower age, or with higher education and income level, or vice versa.

In order to test for the presence of heterogeneity along the variables of interest, we conduct the Wald test of equality of the estimated GATEs. Additionally, we conduct t-tests for differences of the estimated GATEs from the average effect. Rejection of both tests thus gives support for the existence of heterogeneity with respect to the particular variable.[14]

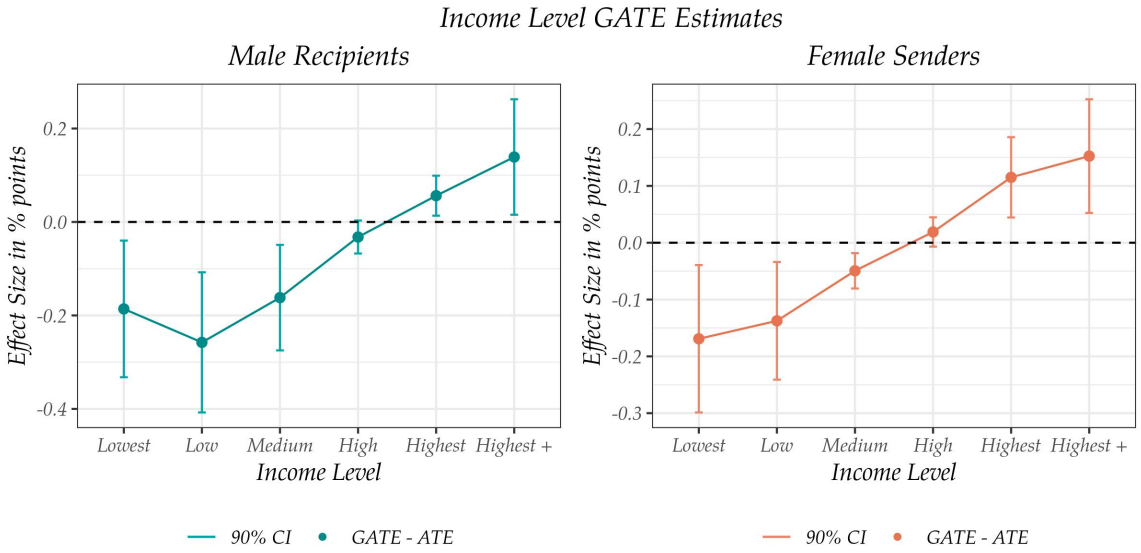
The results of the Wald test suggest heterogeneous effects with regard to the income level for males, both for the recipient as well as the corresponding sender, however, no evidence of heterogeneity in case of females. Furthermore, the heterogeneous effects for males are statistically different from the average effect as well, indicating an explicit pattern, while none of this is the case for females. The respective income level GATE estimates are depicted in Figure 3.5.1 for the recipient and the sender in the male sample. The corresponding results for the female sample are presented in Figure 3.C.1 in Appendix 3.C.

Concerning the male sample, we observe a clear increasing trend of the GATEs for increasing levels of income. As such, for a male recipient, the effect of weekly sport activity in contrast to no sport is greater, the higher the income level of the male recipient himself, and the higher the income level of the female sender, too. As a result, male users with a higher income level, benefit from a regular sport activity on a weekly basis in comparison to no sport, more than male users with a lower income level. This implies that particularly the wealthy males, who earn more than 100'000 EUR in a year, can increase their contact chances the most by practicing sport on a weekly basis. In a similar vein, male users having a potential female partner with high income level, benefit from the higher sport frequency more than the male users, which have a potential female partner with low income level. This pattern suggests also that more wealthy female users value the regular sport activity of a male user more. In addition, not only are these heterogeneous effects statistically relevant, the substantive relevance is documented, too, as the effect sizes are relatively large. As such, the magnitudes of the income level GATEs are ranging from 1.06% points to 1.46% points with respect to the income level of a male recipient, and similarly, from 1.15% points to 1.47% points with respect to the income level of a female sender, in reference to the

[14] Detailed results of the Wald test for equality of the GATEs as well as the tests for differences from the ATE are listed in Appendix 3.C.

average effect of 1.32% points. This implies an increase in the message probability of at least 42.40% for the low income users, up to an increase of 58.80% for the high income users, respectively. This results in a 16.40% difference in message probability solely due to the user's income. A simple back of the envelope calculation reveals this difference in income levels to amount to 4 extra messages out of 1000 profile visits.

Figure 3.5.1: Heterogeneous Effects of Sport Activity based on Income for Males



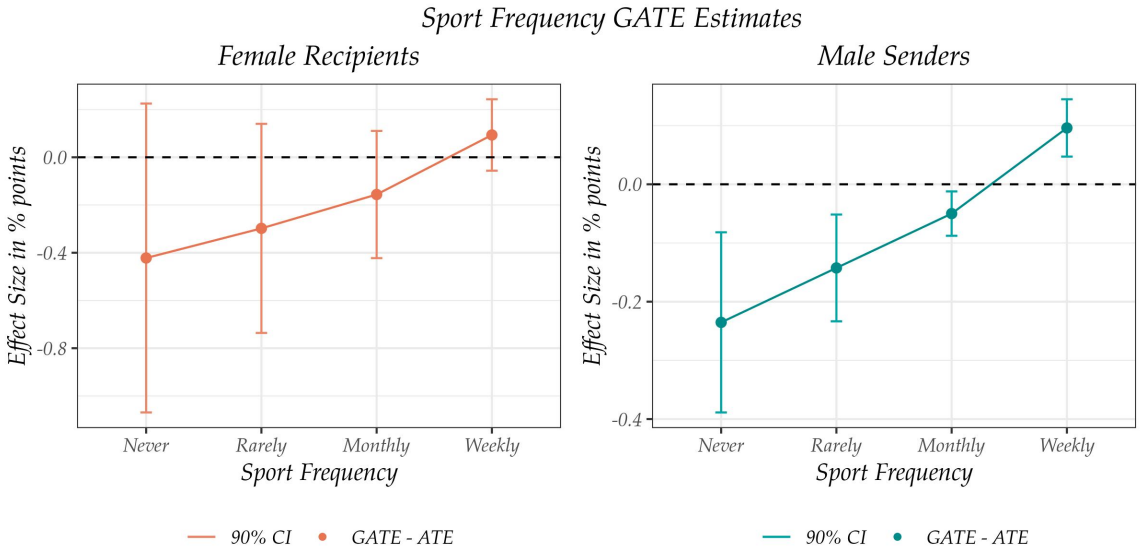
Note: Effects in % points as GATE deviations from the ATE (zero dotted line) with 90% confidence intervals.

As opposed to the male sample, we do not find such evidence of heterogeneity, if we switch the roles of the recipient and the sender (see Figure 3.C.1 in Appendix 3.C). As such, even though we observe a similar increasing pattern for female recipients associated with male senders, the estimated effects lack statistical relevance.

However, in contrast to the results for income heterogeneity, we find supportive evidence for heterogeneity for females in terms of the sport activity, while no such evidence is detected for males. As such, for females, both the Wald test of effect equality as well as the t-tests for differences from the average effect suggest presence of heterogeneity with respect to the level of sport frequency of the male sender with a clear increasing pattern, whilst the heterogeneity with respect to the female recipient lacks the statistical precision. Contrarily, for the male sample, even though we observe a similar increasing pattern as for the female sample, the statistical relevance is, however, absent. The corresponding results for the female sample regarding the sport frequency GATE estimates are presented in Figure 3.5.2 for both the female recipient and the male sender. The respective results for the male sample are depicted in Figure 3.C.2 in Appendix 3.C.

The heterogeneity results with respect to the sport frequency suggest that for a female user, the effect of a weekly sport activity in contrast to no sport is greater, the higher the sport frequency of the potential male partner. Thus, females enjoy a higher effect of their own weekly sport activity, if the sport activity of a potential male partner is on a weekly basis as well. This further suggests that sporty male users appreciate sporty female users more. Nevertheless, despite the clear statistical pattern of the heterogeneity itself, in this case the overall substantive implications remain rather limited as the effect sizes are only moderate, ranging from 0.08% points to 0.41% points, given the average effect of 0.31% points. Additionally, neither for the average effect nor for the respective group effects the presence of an actual null effect can be ruled out.

Figure 3.5.2: Heterogeneous Effects of Sport Activity based on Sport for Females



Note: Effects in % points as GATE deviations from the ATE (zero dotted line) with 90% confidence intervals.

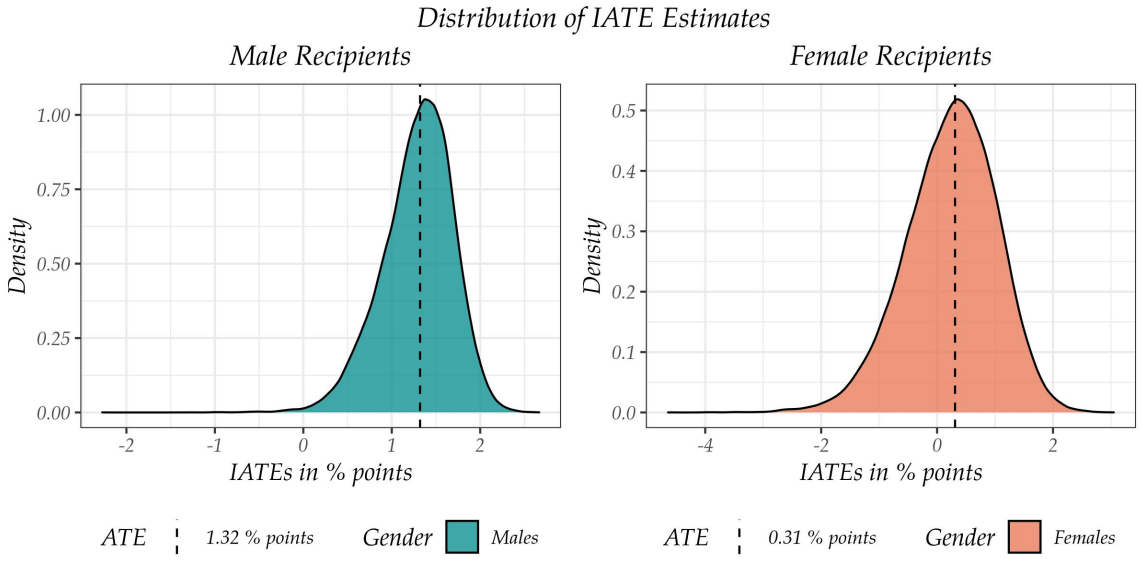
Further results of the Wald tests regarding the remaining heterogeneity variables do not indicate differential effects at conventional significance levels in terms of age or the mutual user distance, concerning both males as well as females. Neither do the differences of the estimated GATEs from the ATE support the evidence for heterogeneous effects. Furthermore, although the Wald test of equality of GATEs based on the education level suggests presence of effect heterogeneity, the differences from the average effect are not statistically relevant and lack an explicit pattern.[15]

Altogether, based on the GATEs analysis, we conclude to find a supporting evidence, both statistical as well as substantive, for heterogeneity in terms of the income level for males and statistical, however, not substantive evidence, in terms of the sport frequency for females, whereas, we find lack of evidence in general, for heterogeneous effects along the age, distance and education level for both males and females.

Additionally, in order to gain more insight for the effect heterogeneity, we analyze the effects on the finest level possible and study the underlying individualized average treatment effects. Figure 3.5.3 provides the distribution of the IATEs for the weekly vs. never comparison, for both the male as well as the female sample, respectively. In both cases, we observe that there is indeed substantial heterogeneity in the considered effects as the effect distributions are noticeably spread out around the mean, i.e. the realized ATE.[16] Additionally, we see that for males, virtually all effects are positive, while for females, about half of the effects are positive and half are negative. This further substantiates the findings on the aggregated levels in terms of the GATEs and the ATE.

[15] The exhaustive results for the effect heterogeneity analysis can be found in Appendix 3.C.
[16] Part of the observed variability is also due to estimation uncertainty: the average standard error for the IATEs is 0.61 for the male and 1.06 for the female sample, respectively.

Figure 3.5.3: Distribution of the Individualized Effects of Sport Activity for Males and Females



Note: Distribution of IATEs smoothed with the Epanechnikov kernel using the Silverman's bandwidth.

To understand these effect distributions more thoroughly, we apply the k-means++ clustering (Arthur & Vassilvitskii, 2007) to provide further descriptive evidence of the dependence of the effects on the heterogeneity variables (see Cockx et al., 2019, for an analogous approach). For this purpose, we perform the clustering by using the IATEs for the weekly vs. never comparison to form distinct clusters, which we sort increasingly according to the mean effect size. We then describe the clusters by the means of the corresponding heterogeneity variables, which however, have not been used to form the clusters. Table 3.5.2 presents the clusters for the IATEs of the male and female sample, respectively.

In general, the clustering reveals consistent patterns with the heterogeneity analysis based on the GATEs. For the male sample, the increasing effects of the sport frequency along the clusters are associated with an increasing level of income on both the recipient as well as the sender side. As such, the lowest effects of sport are clearly for the users with the lowest income level, and vice-versa, the highest effects are evidently for those users with the highest income level. Complementary to the GATEs analysis, the clustering additionally reveals similar increasing patterns in terms of education level and the sport frequency for males. This indicates a further positive relationship, which, however, lacks statistical relevance within the GATEs analysis. Nonetheless, the clustering does not find any particularly clear patterns in terms of age or mutual distance, which is consistent with the GATE estimates.

In case of the female sample, complementary to the GATEs, the clusters suggest an increasing effect for higher sport frequency as documented within the GATEs analysis. However, according to the cluster analysis, this holds true not only for the sender, but also for the recipient side, for which the statistical evidence in terms of the GATEs is missing. In a similar vein, the clusters also suggest a relevant heterogeneity with respect to the income level with an increasing pattern. Furthermore, as for the male clusters, also the female clusters suggest additionally a positive association of the IATEs with the education level, however, no apparent indication of heterogeneity for age or mutual distance.

In addition to the GATE analysis, the clusters further allow for a more detailed description of the IATEs based on the user characteristics, beyond the pre-specified subset of heterogeneity variables. Notably, the cluster analysis reveals a particular relationship between the IATEs and the behavior and preferences of the users, both for males and females. As such, higher IATEs are associated with in-

creasing preference to find the significant other and to have an intimate relationship, as well as with increasing satisfaction of own appearance. In contrast, lower IATEs are associated with increasing smoking frequency, as well as increasing preference for media consumption and comfortable dining. These insights provide not only a better understanding of the specific individualized effects of sport activity on the contact chances, but might serve as a basis and guidance for a selection of relevant heterogeneity variables in future research. An overview of the relevant clusters with variable description is provided in Table 3.C.3 in Appendix 3.C.

Table 3.5.2: Clusters of the Individualized Effects of Sport Activity for Males and Females

	Males					Females				
Clusters	1	2	3	4	5	1	2	3	4	5
IATEs: Weekly vs. Never	0.41	0.88	1.22	1.52	1.85	-1.41	-0.52	0.12	0.71	1.38
Recipient Features										
Age	45.77	45.19	44.57	43.80	43.71	33.53	37.07	38.39	38.15	36.90
Education Level	3.51	3.86	4.24	4.58	4.76	3.65	3.74	3.80	3.93	4.05
Income Level	3.69	3.98	4.26	4.57	4.83	2.98	3.16	3.32	3.46	3.53
Sport Frequency	1.85	2.16	2.37	2.46	2.49	1.55	1.92	2.15	2.34	2.43
Sender Features										
Age	43.94	43.09	42.37	41.53	41.43	36.33	40.15	41.68	41.46	40.23
Education Level	3.46	3.69	3.96	4.16	4.34	3.74	3.82	3.93	4.06	4.25
Income Level	2.94	3.25	3.52	3.79	4.07	3.51	3.82	3.99	4.10	4.14
Sport Frequency	1.86	2.03	2.16	2.24	2.32	1.91	2.06	2.15	2.27	2.44
Shared Features										
Distance	70.39	71.65	70.23	70.16	64.91	64.43	66.22	67.10	66.36	62.97
Observations										
Share	0.07	0.18	0.29	0.31	0.15	0.07	0.20	0.29	0.29	0.15
Total	2288	6439	10153	10826	5252	3753	11048	15896	15484	8047

Note: Means of clustered effects sorted in an increasing order, matched with the heterogeneity variables. Color intensity represents the corresponding effect sizes and highlights the relevant GATEs.

Overall, the cluster analysis of the IATEs emphasizes the results from the GATEs, and as such provides additional evidence for the income heterogeneity for males, as well as the sport heterogeneity for females. Moreover, it reveals further descriptive evidence for increasing effects based on education level, albeit no particular heterogeneity patterns for age or mutual distance. Lastly, it provides valuable insights for additional heterogeneity channels such as relationship preferences.

3.5.3 Placebo Test

In our analysis of the effect of sport activity on contact chances, we assume that the treatment, i.e. the sport frequency is observed once a profile of a recipient has been visited by a sender. However, as discussed in Section 3.4.2, the sport frequency might potentially be entailed in the matching score, which is observable already before the actual profile visit as part of the shortlist of potential partners suggested by the online dating platform. If that would be the case, the sport frequency could potentially indirectly influence already the decision to visit the profile, and not only the decision to send a text message after a profile visit. However, even under such circumstances, this would not violate the CIA *per se*, but rather reduce the size of the estimated effect, which could be then interpreted as a lower bound of the true underlying effect. In order to examine if such mechanism takes place in our setting, we conduct a placebo test inspired by Imbens and Wooldridge (2009) to assess the validity of the CIA by testing for a zero effect

on an outcome variable assumed to be unaffected by the treatment, here the decision to visit the user profile. Accordingly, we redo our main analysis, while swapping the message outcome for a visit outcome. Thus, we estimate the average treatment effects of sport frequency on the visit probability, given the same conditioning set. Therefore, if the sport frequency is, as assumed, not part of the matching score, its effect on the probability to visit a user profile should be equal to zero.

In order to implement such placebo test, we first need to impute the 'potential' visits, as by construction, we only observe the realized visits. For a given user, we consider all registered user profiles with opposite sex and within a specified distance radius as potential visits.[17] We end up with a sample consisting of 38'552'821 observations, out of which 178'372 represent the actual realized and the rest the imputed potential visits. Analogously as in the main analysis, we split the sample into a male and a female sample. Furthermore, due to the computational feasibility and general consistency of the analysis, we randomly draw an identically sized male and female sample as in the main estimation, such that we replicate the corresponding sport frequency shares, too.[18] A similar approach to impute the potential visits has been used also in previous studies focusing on online dating platforms (Bruch et al., 2016).

Table 3.5.3: Average Effects of Sport Activity on the Visit Chances for Males and Females

	Males				Females			
	Never	*Rarely*	*Monthly*	*Weekly*	*Never*	*Rarely*	*Monthly*	*Weekly*
Never	0.23				0.66			
	(0.14)				(0.15)			
Rarely	0.29	0.52			-0.07	0.59		
	(0.24)	(0.20)			(0.22)	(0.16)		
Monthly	0.08	-0.22	0.31		-0.20	-0.13	0.46	
	(0.16)	(0.22)	(0.08)		(0.17)	(0.17)	(0.08)	
Weekly	0.18	-0.12	0.10	0.41	-0.08	-0.01	0.12	0.58
	(0.15)	(0.21)	(0.10)	(0.06)	(0.16)	(0.17)	(0.10)	(0.07)

Note: Effects in % points. Potential outcomes on the diagonal. Standard errors in parentheses. Significance levels refer to: *** < 0.01, ** < 0.05, * < 0.1. Color intensity represents the corresponding level sizes.

Table 3.5.3 summarizes the ATE results of the Modified Causal Forest estimation for the placebo test. First of all, we observe that the potential outcomes for both males and females do not exhibit any particular upward or downward trend as is the case for the main analysis. Furthermore, for neither the male nor the female sample, we find evidence for statistically relevant effects. Moreover, the effect sizes and the levels of potential outcomes are an order of magnitude lower than our main results, being effectively zero in terms of the substantive relevance. Even though the results of such placebo tests do not completely rule out the possibility of a presence of an effect on the visit probability, they provide a supportive evidence that this is, indeed, not the case. Hence, we conclude that our main analysis estimates the full causal effects of sport activity on the contact chances, rather than only lower bounds thereof.

[17] We restrict the potential visits to opposite sex as we observe only heterosexual users in our sample. Furthermore, we restrict the distance of potential users due to dimensionality concerns, as the share of the realized visits would otherwise be almost completely diminished, if unrestricted. Here, we remain rather conservative and set the potential distance to 95% of the maximum observed distance of an actual realized visit.

[18] We repeated the random draw several times, while the results remained qualitatively robust.

3.6 Discussion

The main objective of this paper was to analyze the effect of sport activity on human mating. Following this objective, we examined the effect of sport frequency on contact chances based on a unique dataset from an online dating platform and applying the Modified Causal Forest estimator (Lechner, 2018). We found that for male users, doing sport on a weekly basis increases the probability to receive a first message by more than 50% relatively to not doing sport at all, while for female users, we do not find evidence for such an effect. In addition, we uncover important effect heterogeneities. In particular, the effect of sport frequency on contact chances increases with higher income for male, but not for female, users.

This paper offers notable implications for research and practice. First, this study contributes to the literature on human mating. In particular, we demonstrate that sport activity, as an essential behavioral trait and pivotal information on online dating platforms, enfolds a causal effect on contact chances. In turn, this paper overcomes limitations of previous work that did not consider or comprehensively map the effect of sport activity on human mating. Moreover, this paper expands previous work on the effects of sport activity by demonstrating that sport activity does not only affect physical/mental health and social and economic conditions, as well-documented by prior research (Strong et al., 2005; Lechner, 2009), but also one of the most decisive spheres of human existence, that is human mating.

Second, this paper advances empirical approaches for assessing causal effects in large-dimensional data environments, as applicable, for example, to remote-sensing data. In particular, this research applies a very flexible estimation procedure, which offers not only greater flexibility in considering (interrelated effects of) covariates, but also a systematic analysis of the underlying heterogeneities of the effects on different levels of aggregation (Lechner, 2018). Thus, this paper may support future research in analyzing human behavior in large-dimensional data environments. However, even though the causal machine learning approach is capable of detecting statistically relevant heterogeneities, it is crucial to assess also its substantive relevance. Following this notion, the effect heterogeneities in this research provide different perspectives on practical implications. In particular, the increasing effect of sport activity on contact chances with higher income for male users is both statistically justified as well as substantially relevant, leading to the above mentioned implication. Contrarily, potential implications, resulting from the observation that the effect of sport frequency on contact chances increases with higher sport frequency for females, are limited as the particular evidence in our setting is not substantially relevant, even though it is statistical justified. In addition to the main heterogeneity analysis, the post-estimation descriptive cluster analysis of the most disaggregated effects provides additional insights for possible heterogeneity channels, such as the education level or relationship preferences of the users.

Third, this study may support individuals to increase their chances of finding a mate on online dating platforms by demonstrating if and to what extent sport activity contributes to the likelihood to be recognized. In particular, men may benefit from the insights of this research by being aware that sport on a weekly basis relative to no sport can increase their probability to receive a first message by more than 50%, or even up to 60% in case of higher income individuals, while for women the effect of sport activity on the contact chances is not entirely evident. Thus, this study may incentivise individuals to increase the level of sport activity, not only because of the well-documented effects on, for example, health (Penedo & Dahn, 2005), but also for their chances of finding a mate.

Moreover, from a public health perspective, this paper provides empirical reasoning for justifying and evaluating incentives for public health promotion due to the impact of sport activities for human partnering, family planning, and reproduction.

Finally, this paper may serve practitioners, namely product developers and software engineers, as a foundation to improve the architecture of online dating platforms, including interface designs and matching algorithms. In particular, this study points out the relevance of sport activity for mate evaluation and selection patterns, while considering effect heterogeneities based on established socio-demographic characteristics at the same time. In turn, this research may help practitioners to assess humans' mate evaluation and selection in much more detail and, correspondingly, to evaluate improvements of the architecture of online dating platforms (e.g., customized weighting of sport activity in matching algorithms or specific placement of information on sport activity on individual profile pages). In a similar vein, the insights of this research are applicable to engineer architectures of other platforms with a likewise high degree of interpersonal computer-mediated interaction, for example, social networks.

Acknowledgements

This research project is part of the National Research Programme 'Big Data' (NRP 75) of the Swiss National Science Foundation (SNSF). Further information on the National Research Programme can be found at www.nrp75.ch or https://bigdata-dialog.ch. This work has been co-funded by the GFF-IPF Grant of the Basic Research Fund of the University of St.Gallen. We gratefully acknowledge the data access provided by Parship.de (PE Digital GmbH, Hamburg). A previous version of this paper was presented at research seminars of the University of St.Gallen, the GESIS Spring Seminar in Cologne and the Causal Machine Learning Workshop in St.Gallen. We thank participants, in particular Daniel Goller, Fabian Muny and Anthony Strittmatter for helpful comments. The usual disclaimer applies.

Bibliography

Abadie, A. & Cattaneo, M. D. (2018). Econometric Methods for Program Evaluation.

Arthur, D. & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the annual acm-siam symposium on discrete algorithms* (Vol. 07-09-Janu, pp. 1027–1035).

Athey, S. (2018). The Impact of Machine Learning on Economics. In *The economics of artificial intelligence: An agenda* (pp. 507–547). University of Chicago Press.

Athey, S. & Imbens, G. W. (2016). Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S. & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, *11*(1), 685–725.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148–1178.

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

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. CRC Press.

Bruch, E., Feinberg, F., & Lee, K. Y. (2016). Extracting multistage screening rules from online dating activity data. *Proceedings of the National Academy of Sciences*, *113*(38), 10530–10535.

Bühlmann, P. & Yu, B. (2002). Analyzing bagging.

Cacioppo, J. T., Cacioppo, S., Gonzaga, G. C., Ogburn, E. L., & Vanderweele, T. J. (2013). Marital satisfaction and break-ups differ across on-line and off-line meeting venues. *Proceedings of the National Academy of Sciences*, *110*(25), 10135–10140.

Caruso, R. (2011). Crime and sport participation: Evidence from Italian regions over the period 1997-2003. *Journal of Socio-Economics*, *40*(5), 455–463.

Cockx, B., Lechner, M., & Bollens, J. (2019). Priority to unemployed immigrants?A causal machine learning evaluation of training in Belgium. *arXiv preprint arXiv: 1912.12864*.

Cuperman, R. & Ickes, W. (2009). Big Five Predictors of Behavior and Perceptions in Initial Dyadic Interactions: Personality Similarity Helps Extraverts and Introverts, but Hurts "Disagreeables". *Journal of Personality and Social Psychology*, *97*(4), 667–684.

Eastwick, P. W., Luchies, L. B., Finkel, E. J., & Hunt, L. L. (2014). The predictive validity of ideal partner preferences: A review and meta-analysis. *Psychological Bulletin*, *140*(3), 623–665.

Eime, R. M., Young, J. A., Harvey, J. T., Charity, M. J., & Payne, W. R. (2013). A systematic review of the psychological and social benefits of participation in sport for children and adolescents: Informing development of a conceptual model of health through sport.

Eurobarometer, S. (2014). Sport and Physical Activity. *Brussels: TNS Opinion & Social*.

Farthing, G. W. (2005). Attitudes toward heroic and nonheroic physical risk takers as mates and as friends. *Evolution and Human Behavior*, *26*(2), 171–185.

Felfe, C., Lechner, M., & Steinmayr, A. (2016). Sports and Child Development. *PloS one*, *11*(5), e0151729.

Fiore, A. T., Taylor, L. S., Mendelsohn, G. A., & Hearst, M. (2008). Assessing attractiveness in online dating profiles. In *Conference on human factors in computing systems - proceedings* (pp. 797–806).

Fox, C. K., Barr-Anderson, D., Neumark-Sztainer, D., & Wall, M. (2010). Physical activity and sports team participation: Associations with academic outcomes in middle school and high school students. *Journal of School Health*, *80*(1), 31–37.

Fricke, H., Lechner, M., & Steinmayr, A. (2018). The effects of incentives to exercise on student performance in college. *Economics of Education Review*, *66*, 14–39.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer Science & Business Media.

Hillman, C. H., Erickson, K. I., & Kramer, A. F. (2008). Be smart, exercise your heart: Exercise effects on brain and cognition.

Hitsch, G. J., Hortaçsu, A., & Ariely, D. (2010a). Matching and sorting in online dating. *American Economic Review*, *100*(1), 130–163.

Hitsch, G. J., Hortaçsu, A., & Ariely, D. (2010b). What makes you click?-mate preferences in online dating. *Quantitative Marketing and Economics*, *8*(4), 393–427.

Hodler, R., Lechner, M., & Raschky, P. (2020). Reassessing the Resource Curse using Causal Machine Learning. *CEPR Discussion Paper No. DP15272*.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Huang, H. & Humphreys, B. R. (2012). Sports participation and happiness: Evidence from US microdata. *Journal of Economic Psychology*, *33*(4), 776–793.

Humphreys, B. R., McLeod, L., & Ruseski, J. E. (2014). Physical activity and health outcomes: Evidence from Canada. *Health Economics (United Kingdom)*, *23*(1), 33–54.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710.

Imbens, G. W. & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Janssen, I. & LeBlanc, A. G. (2010). Systematic review of the health benefits of physical activity and fitness in school-aged children and youth.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, *24*(1), 134–161.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of labour market policies* (13, pp. 43–58).

Lechner, M. (2009). Long-run labour market and health effects of individual sports activities. *Journal of Health Economics*, *28*(4), 839–854.

Lechner, M. (2017). Empirical Evidence on Educational Effects of Physical Activity: Four Examples. In T. Pawlowski & M. Fahrner (Eds.), *Arbeitsmarkt und sport – eine ökonomische betrachtung / sport labor economics (sportökonomie 19)* (pp. 12–22). Schorndorf: Hofmann.

Lechner, M. (2018). Modified Causal Forests for Estimating Heterogeneous Causal Effects. *arXiv preprint arXiv: 1812.09487v2*.

Lechner, M. & Strittmatter, A. (2019). Practical procedures to deal with common support problems in matching estimation. *Econometric Reviews*, *38*(2), 193–207.

Lee, S. (2016). Effect of Online Dating on Assortative Mating: Evidence from South Korea. *Journal of Applied Econometrics*, *31*(6), 1120–1139.

Park, J. H., Buunk, A. P., & Wieling, M. B. (2007). Does the face reveal athletic flair? Positions in team sports and facial attractiveness. *Personality and Individual Differences*, *43*(7), 1960–1965.

Penedo, F. J. & Dahn, J. R. (2005). Exercise and well-being: A review of mental and physical health benefits associated with physical activity.

Pfeifer, C. & Cornelißen, T. (2010). The impact of participation in sports on educational attainment-New evidence from Germany. *Economics of Education Review*, *29*(1), 94–103.

Pronk, N. P., Martinson, B., Kessler, R. C., Beck, A. L., Simon, G. E., & Wang, P. (2004). The Association between Work Performance and Physical Activity, Cardiorespiratory Fitness, and Obesity. *Journal of Occupational and Environmental Medicine*, *46*(1), 19–25.

Richter, D., Schupp, J. et al. (2015). The SOEP innovation sample (SOEP IS). *Schmollers Jahrbuch: Journal of Applied Social Science Studies/Zeitschrift für Wirtschafts-und Sozialwissenschaften*, *135*(3), 389–400.

Rooth, D. O. (2011). Work out or out of work - The labor market return to physical fitness and leisure sports activities. *Labour Economics*, *18*(3), 399–409.

Rosenfeld, M. J., Thomas, R. J., & Hausen, S. (2019). Disintermediating your friends: How online dating in the United States displaces other ways of meeting. *Proceedings of the National Academy of Sciences*, *116*(36), 17753–17758.

Rubin, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1991). Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. *Biometrics*, *47*(4), 1213–1234.

Ruseski, J. E., Humphreys, B. R., Hallman, K., Wicker, P., & Breuer, C. (2014). Sport participation and subjective well-being: Instrumental variable results from german survey data. *Journal of Physical Activity and Health*, *11*(2), 396–403.

Schulte-Hostedde, A. I., Eys, M. A., Emond, M., & Buzdon, M. (2012). Sport participation influences perceptions of mate characteristics. *Evolutionary Psychology*, *10*(1), 78–94.

Schulte-Hostedde, A. I., Eys, M. A., & Johnson, K. (2008). Female Mate Choice is Influenced by Male Sport Participation. *Evolutionary Psychology*, *6*(1), 113–124.

Steckenleiter, C. & Lechner, M. (2020). Recent Evidence on the Effects of Physical Activity on Human Capital and Employment. In P. Downward, B. Frick, B. R. Humphreys, T. Pawlowski, J. E. Ruseski, & B. P. Soebbing (Eds.), *The sage handbook of sports economics* (pp. 64–71). SAGE.

Stempel, C. (2005). Adult participation sports as cultural capital: A Test of Bourdieu's Theory of the Field of Sports. *International Review for the Sociology of Sport*, *40*(4), 411–432.

Strong, W. B., Malina, R. M., Blimkie, C. J., Daniels, S. R., Dishman, R. K., Gutin, B., ... Trudeau, F. (2005). Evidence based physical activity for school-age youth. *Journal of Pediatrics*, *146*(6), 732–737.

Trudeau, F. & Shephard, R. J. (2008). Physical education, school physical activity, school sports and academic performance.

Wager, S. & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Warburton, D. E., Nicol, C. W., & Bredin, S. S. (2006). Health benefits of physical activity: The evidence.

Appendix

3.A Descriptive Statistics

Table 3.A.1: Descriptive Statistics for the User Sample

	Mean	SD	Min	Max
Sport Frequency				
Never	0.13	0.34	0.00	1.00
Rarely	0.10	0.31	0.00	1.00
Monthly	0.29	0.45	0.00	1.00
Weekly	0.48	0.50	0.00	1.00
Demographic Features				
Gender (=1 if female)	0.48	0.50	0.00	1.00
Age (in years)	40.05	11.41	18.00	82.00
Income Level				
Lowest	0.12	0.32	0.00	1.00
Low	0.16	0.37	0.00	1.00
Medium	0.20	0.40	0.00	1.00
High	0.24	0.43	0.00	1.00
Highest	0.22	0.42	0.00	1.00
Highest+	0.06	0.24	0.00	1.00
Education Level				
Lowest	0.00	0.06	0.00	1.00
Low	0.08	0.27	0.00	1.00
Medium	0.37	0.48	0.00	1.00
High	0.16	0.37	0.00	1.00
Highest	0.39	0.49	0.00	1.00

Note: Main variables describing the population displayed.

Figure 3.A.1: Distribution of the Distance between Users

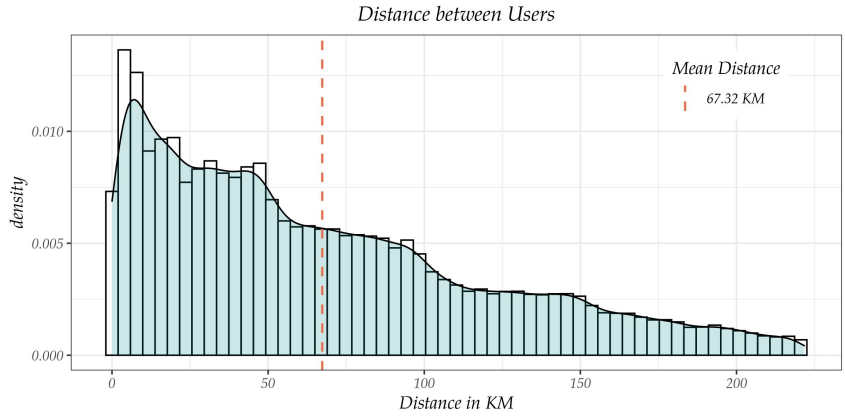


Table 3.A.2: Descriptive Statistics by Sport Frequency for Female Sample

	Never	Rarely	Monthly	Weekly	Total
Outcome					
First Message	0.11	0.10	0.10	0.11	0.11
Recipient Features					
Age	35.49	37.65	37.84	37.83	37.53
Income Level					
Lowest	0.20	0.20	0.10	0.07	0.10
Low	0.28	0.23	0.19	0.14	0.18
Medium	0.28	0.25	0.23	0.22	0.23
High	0.17	0.21	0.27	0.33	0.28
Highest	0.05	0.11	0.18	0.21	0.18
Highest+	0.01	0.01	0.03	0.04	0.03
Education Level					
Lowest	0.00	0.00	0.00	0.00	0.00
Low	0.08	0.06	0.02	0.02	0.03
Medium	0.61	0.52	0.41	0.36	0.42
High	0.20	0.21	0.21	0.22	0.21
Highest	0.11	0.21	0.35	0.41	0.33
Sender Features					
Age	38.57	40.80	41.11	41.07	40.75
Income Level					
Lowest	0.09	0.07	0.05	0.04	0.05
Low	0.16	0.13	0.09	0.07	0.09
Medium	0.23	0.21	0.17	0.16	0.18
High	0.26	0.26	0.27	0.27	0.27
Highest	0.21	0.26	0.32	0.35	0.32
Highest+	0.04	0.07	0.10	0.11	0.09
Education Level					
Lowest	0.00	0.00	0.00	0.00	0.00
Low	0.11	0.10	0.07	0.05	0.06
Medium	0.47	0.41	0.32	0.28	0.33
High	0.14	0.14	0.16	0.17	0.16
Highest	0.28	0.35	0.45	0.50	0.44
Sport Frequency					
Never	0.16	0.14	0.11	0.09	0.11
Rarely	0.13	0.12	0.10	0.09	0.10
Monthly	0.29	0.30	0.31	0.28	0.29
Weekly	0.42	0.44	0.49	0.55	0.50
Observations					
Total Share	0.12	0.09	0.29	0.49	1.00
Total Observations	13'408	98'33	31'801	53'414	108'456

Note: Means of variables displayed in all columns.

Table 3.A.3: Descriptive Statistics by Sport Frequency for Male Sample

	Never	Rarely	Monthly	Weekly	Total
Outcome					
First Message	0.02	0.03	0.03	0.04	0.04
Recipient Features					
Age	44.88	46.21	45.56	43.43	44.37
Income Level					
Lowest	0.06	0.03	0.02	0.02	0.02
Low	0.15	0.08	0.05	0.04	0.05
Medium	0.21	0.19	0.14	0.12	0.14
High	0.22	0.32	0.27	0.25	0.26
Highest	0.30	0.30	0.41	0.42	0.40
Highest+	0.05	0.07	0.11	0.15	0.13
Education Level					
Lowest	0.00	0.01	0.00	0.00	0.00
Low	0.11	0.07	0.03	0.02	0.03
Medium	0.37	0.36	0.24	0.17	0.22
High	0.15	0.20	0.15	0.15	0.16
Highest	0.37	0.37	0.58	0.66	0.59
Sender Features					
Age	43.15	44.20	43.37	41.19	42.19
Income Level					
Lowest	0.13	0.09	0.07	0.06	0.07
Low	0.20	0.18	0.13	0.11	0.13
Medium	0.25	0.25	0.23	0.22	0.23
High	0.26	0.29	0.31	0.33	0.32
Highest	0.14	0.16	0.21	0.23	0.21
Highest+	0.03	0.03	0.04	0.04	0.04
Education Level					
Lowest	0.00	0.00	0.00	0.00	0.00
Low	0.05	0.04	0.03	0.02	0.02
Medium	0.48	0.47	0.39	0.33	0.37
High	0.18	0.20	0.19	0.18	0.19
Highest	0.28	0.29	0.40	0.47	0.42
Sport Frequency					
Never	0.18	0.16	0.12	0.10	0.12
Rarely	0.12	0.11	0.10	0.09	0.09
Monthly	0.30	0.30	0.31	0.30	0.30
Weekly	0.40	0.44	0.47	0.52	0.49
Observations					
Total Share	0.07	0.08	0.29	0.56	1.00
Total Observations	4'690	5'827	19'970	39'429	69'916

Note: Means of variables displayed in all columns.

3.B Online Dating Platform

3.B.1 Valid User Interactions

In our analysis, we restrict ourselves to *one-way* user interactions. These interactions are always initiated by a visit from the sender, which is invisible to the recipient. The visit is then immediately followed by either a visible action from the sender, or possibly no further action at all. However, in both cases a visible reply of the recipient to this initial action by the sender is not permitted. In that sense, we retain only one-way interactions such that the sender was visibly or invisibly active, while the recipient stayed visibly passive. Hence, we do not allow for any visible reciprocal interaction between the sender and the recipient.

For instance, a sender visit followed by a sender message is a valid interaction. Also, two successive sender visits followed by a message is a valid interaction. A single sender visit is valid interaction, too. Further notice, that a sender visit followed by a recipient visit and afterwards a sender message is a valid interaction as well, as the sender has not seen the recipient's visit. However, a sender visit and sender like followed by a recipient visit and like back inducing a sender message is not a valid interaction anymore as the sender message has already been provoked by the recipient. Hence, we always restrict the interactions until the point a possible reciprocal interaction taking place.

3.C Additional Results

3.C.1 Heterogeneous Effects

Table 3.C.1: Wald Tests for Equality of Group Effects for Males and Females

GATEs: Weekly vs. Never	Males		Females	
Wald Test	χ^2	p-Value	χ^2	p-Value
Recipient Features				
Age	23.56	37.08	13.17	96.32
Education Level	14.03	0.72	7.63	10.62
Income Level	22.67	0.04	1.79	87.72
Sport Frequency	9.15	2.74	3.43	32.94
Sender Features				
Age	32.18	7.45	24.18	50.90
Education Level	20.40	0.04	9.75	4.49
Income Level	17.33	0.39	6.49	26.13
Sport Frequency	6.55	8.76	4.41	0.24
Shared Features				
Distance	23.01	40.12	13.59	96.84

Note: Wald tests of Equality of the GATEs. p-Values in %.

Table 3.C.2: Tests for Differences of GATEs to ATE for Males and Females

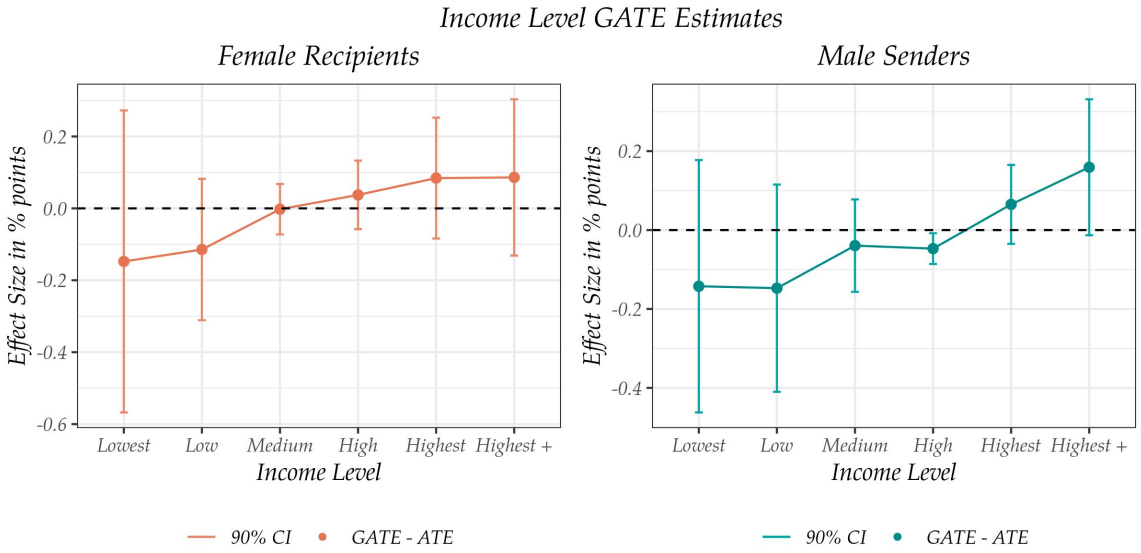
GATEs: Weekly vs. Never		Males				Females		
t-Test	Group	Δ	SE	*p*-Value	Group	Δ	SE	*p*-Value
Recipient Features								
Age	23.50	-0.02	0.08	80.04	21.00	-0.09	0.27	75.02
	30.00	0.02	0.07	83.72	25.00	-0.12	0.23	59.96
	31.50	0.06	0.07	40.97	26.50	0.01	0.21	95.03
	33.00	0.04	0.07	56.74	27.50	-0.10	0.14	48.40
	34.50	0.05	0.06	45.19	28.50	-0.09	0.13	49.29
	35.50	0.03	0.09	76.92	29.50	-0.07	0.09	47.27
	36.50	0.06	0.08	49.74	30.50	-0.04	0.10	66.35
	37.50	-0.01	0.04	83.13	31.50	0.14	0.12	22.45
	39.00	0.04	0.05	40.91	32.50	-0.01	0.09	88.68
	40.50	0.03	0.06	60.45	33.50	0.13	0.10	17.17
	42.50	-0.02	0.03	40.11	34.50	0.06	0.06	36.87
	45.00	-0.05	0.04	14.49	35.50	0.07	0.09	45.64
	46.50	-0.01	0.05	76.75	36.50	0.07	0.06	26.29
	48.00	-0.00	0.05	91.87	37.50	0.12	0.07	9.96
	49.50	0.02	0.05	74.89	38.50	0.03	0.08	69.35
	50.50	0.01	0.06	82.71	40.00	0.01	0.08	93.72
	51.50	0.00	0.06	94.83	42.00	-0.05	0.11	67.89
	52.50	-0.05	0.07	44.77	43.50	-0.09	0.13	48.02
	54.00	0.02	0.07	72.17	45.00	0.07	0.11	51.04
	55.50	-0.06	0.07	41.80	46.50	0.06	0.12	62.45
	57.00	-0.02	0.10	84.60	48.00	0.03	0.12	80.54
	59.50	-0.07	0.12	54.40	49.50	0.02	0.14	89.29
	71.50	-0.09	0.13	52.15	51.00	-0.02	0.14	88.52
					53.50	-0.05	0.15	75.73
					67.00	-0.03	0.16	87.07
Education Level	Lowest	-0.15	0.11	17.30	Lowest	0.11	0.29	70.80
	Low	-0.42	0.19	2.36	Low	-0.22	0.20	27.04
	Medium	-0.27	0.15	7.51	Medium	-0.10	0.07	19.95
	High	0.03	0.02	15.84	High	0.04	0.04	30.37
	Highest	0.09	0.05	9.19	Highest	0.09	0.10	35.32
Income Level	Lowest	-0.19	0.09	3.62	Lowest	-0.15	0.26	56.30
	Low	-0.26	0.09	0.47	Low	-0.11	0.12	33.82
	Medium	-0.16	0.07	1.83	Medium	-0.00	0.04	95.63
	High	-0.03	0.02	13.33	High	0.04	0.06	51.80
	Highest	0.06	0.03	3.07	Highest	0.08	0.10	41.01
	Highest+	0.14	0.08	6.43	Highest+	0.09	0.13	51.46
Sport Frequency	Never	-0.24	0.12	5.62	Never	-0.42	0.39	28.34
	Rarely	-0.16	0.10	14.20	Rarely	-0.30	0.27	26.33
	Monthly	-0.05	0.05	34.31	Monthly	-0.16	0.16	33.56
	Weekly	0.03	0.02	26.99	Weekly	0.09	0.09	30.43
Sender Features								
Age	22.50	-0.03	0.09	76.98	22.00	-0.16	0.29	58.20
	28.00	0.02	0.07	75.51	27.00	-0.08	0.22	71.84
	29.50	0.05	0.08	55.52	28.50	-0.14	0.17	40.21
	30.50	0.06	0.08	43.37	29.50	-0.14	0.15	36.84
	32.00	0.05	0.08	50.57	30.50	-0.01	0.11	90.90
	33.50	0.02	0.08	78.09	31.50	-0.04	0.12	72.55
	34.50	0.06	0.07	43.61	32.50	-0.03	0.09	73.61
	35.50	0.00	0.05	94.99	33.50	0.03	0.08	66.48
	36.50	0.04	0.06	48.48	34.50	0.02	0.07	82.77
	38.00	0.03	0.03	34.29	35.50	0.09	0.07	17.47
	40.00	-0.02	0.03	36.95	36.50	0.16	0.08	4.37
	42.00	-0.00	0.03	96.81	37.50	0.06	0.05	28.04
	44.00	-0.07	0.05	12.42	38.50	0.06	0.06	34.36
	45.50	-0.00	0.04	89.46	40.00	0.10	0.06	8.29
	47.00	-0.01	0.05	77.00	41.50	0.01	0.07	87.26
	48.50	-0.00	0.05	98.54	43.00	0.05	0.07	52.76
	49.50	0.03	0.06	63.18	44.50	0.00	0.08	95.30

continued on next page

t-Test	Group	Δ	SE	*p*-Value	Group	Δ	SE	*p*-Value
	50.50	0.02	0.06	74.85	46.00	0.04	0.09	69.01
	51.50	0.00	0.07	95.22	47.50	-0.01	0.11	95.12
	53.00	-0.01	0.08	93.70	48.50	0.07	0.11	48.48
	55.00	-0.09	0.10	36.01	50.00	0.01	0.12	91.68
	57.50	-0.07	0.11	50.78	51.50	-0.10	0.13	46.68
	68.50	-0.09	0.14	52.80	53.00	0.02	0.12	88.61
					55.00	0.01	0.13	91.37
					57.50	0.02	0.15	87.36
					70.50	-0.05	0.16	73.96
Education Level	Lowest	0.09	0.10	36.18	Lowest	0.04	0.15	75.74
	Low	-0.39	0.16	1.36	Low	-0.22	0.15	13.95
	Medium	-0.10	0.06	8.91	Medium	-0.12	0.08	15.74
	High	0.00	0.02	91.15	High	-0.01	0.03	62.60
	Highest	0.09	0.07	16.89	Highest	0.10	0.06	8.18
Income Level	Lowest	-0.17	0.08	3.20	Lowest	-0.14	0.19	46.40
	Low	-0.14	0.06	2.89	Low	-0.15	0.16	35.62
	Medium	-0.05	0.02	0.91	Medium	-0.04	0.07	58.00
	High	0.02	0.02	22.64	High	-0.05	0.02	4.74
	Highest	0.12	0.04	0.74	Highest	0.07	0.06	28.41
	Highest+	0.15	0.06	1.22	Highest+	0.16	0.11	12.86
Sport Frequency	Never	-0.10	0.06	10.27	Never	-0.24	0.09	1.17
	Rarely	-0.05	0.04	22.08	Rarely	-0.14	0.06	1.00
	Monthly	-0.00	0.01	64.62	Monthly	-0.05	0.02	2.98
	Weekly	0.03	0.02	8.75	Weekly	0.10	0.03	0.12
Shared Features								
Distance	1.57	0.00	0.20	99.86	1.45	0.03	0.11	79.53
	4.60	0.03	0.20	86.16	4.26	0.01	0.11	89.61
	7.79	0.04	0.15	79.16	7.09	-0.02	0.11	88.02
	12.00	0.01	0.08	87.52	10.38	0.05	0.07	48.18
	16.69	0.00	0.04	92.73	14.24	0.03	0.06	58.33
	21.58	0.01	0.03	72.92	18.09	0.06	0.05	19.38
	26.98	0.00	0.02	83.19	22.23	0.01	0.04	85.48
	32.37	-0.01	0.02	62.77	26.85	0.02	0.03	63.64
	37.99	0.00	0.03	92.04	31.26	0.02	0.03	55.16
	43.78	0.01	0.03	72.41	35.62	-0.04	0.03	15.40
	49.55	-0.00	0.04	99.40	40.15	-0.03	0.03	35.87
	56.12	0.00	0.03	87.98	44.60	-0.03	0.04	50.06
	63.60	-0.02	0.04	58.70	49.10	-0.04	0.03	24.65
	71.36	0.01	0.04	72.91	54.62	-0.00	0.03	96.65
	79.30	0.00	0.04	98.34	61.12	-0.03	0.04	38.27
	87.72	-0.02	0.04	70.56	67.78	-0.02	0.03	55.05
	96.77	-0.02	0.04	61.96	74.76	-0.02	0.04	62.73
	108.00	-0.04	0.04	39.33	82.10	0.02	0.04	58.04
	121.75	-0.01	0.04	74.22	89.77	0.01	0.05	84.55
	136.80	-0.00	0.04	90.50	97.78	-0.00	0.04	99.96
	153.36	0.02	0.03	64.38	107.76	-0.01	0.05	86.38
	173.89	-0.02	0.04	67.71	120.60	-0.03	0.05	60.48
	203.71	-0.02	0.04	64.58	134.74	-0.00	0.05	99.07
					149.88	0.02	0.06	75.90
					170.12	-0.01	0.05	92.23
					202.07	-0.02	0.06	78.49

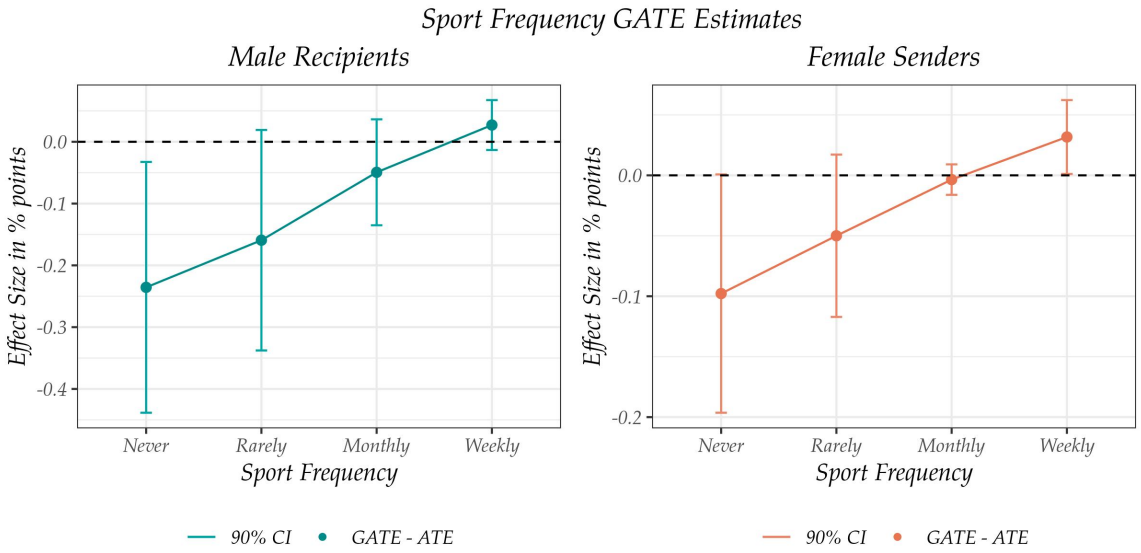
Note: t-tests for Differences of the GATEs from the ATE. Δ in % points. *p*-Values in %.

Figure 3.C.1: Heterogeneous Effects of Sport Activity based on Income for Females



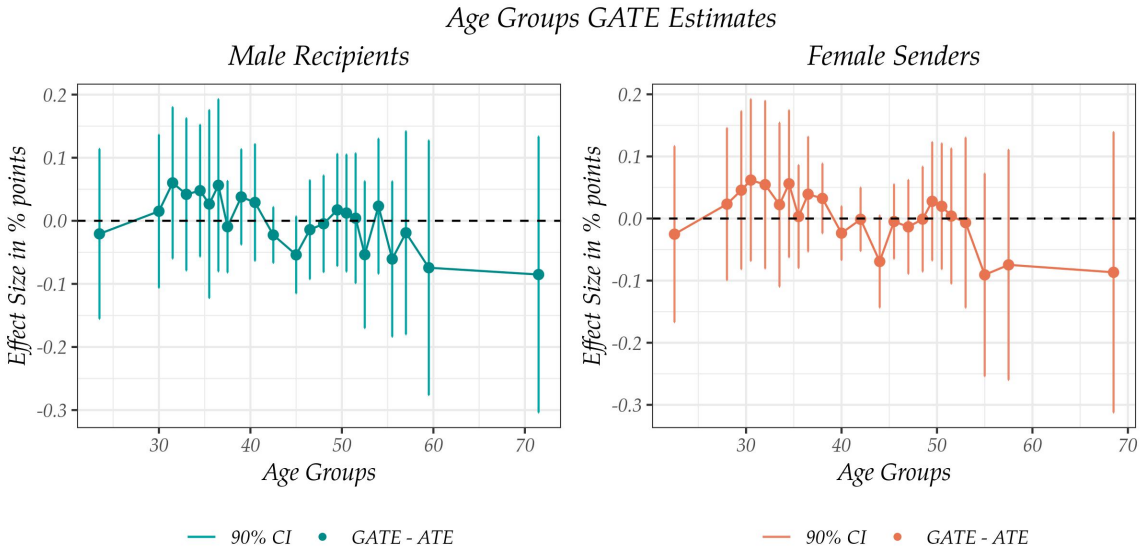
Note: Effects in % points as GATE deviations from the ATE (zero dotted line) with 90% confidence intervals.

Figure 3.C.2: Heterogeneous Effects of Sport Activity based on Sport for Males



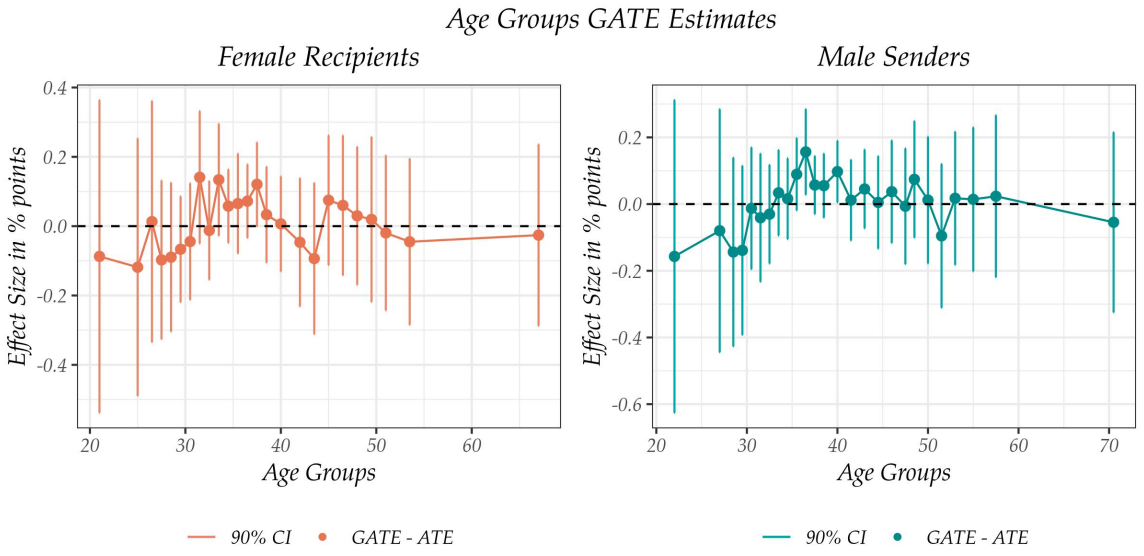
Note: Effects in % points as GATE deviations from the ATE (zero dotted line) with 90% confidence intervals.

Figure 3.C.3: Heterogeneous Effects of Sport Activity based on Age for Males



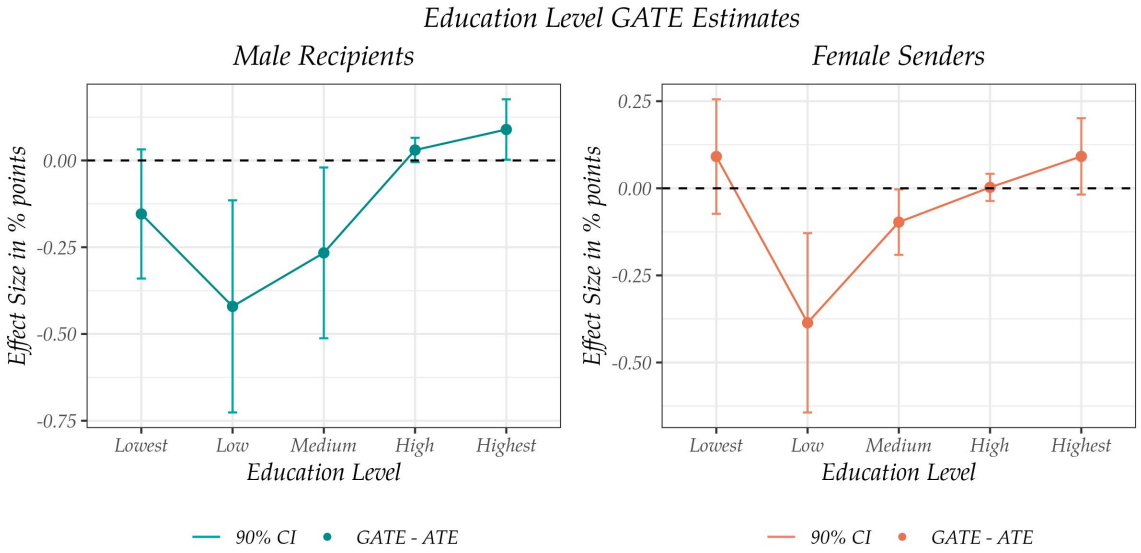
Note: Effects in % points as GATE deviations from the ATE (zero dotted line) with 90% confidence intervals.

Figure 3.C.4: Heterogeneous Effects of Sport Activity based on Age for Females



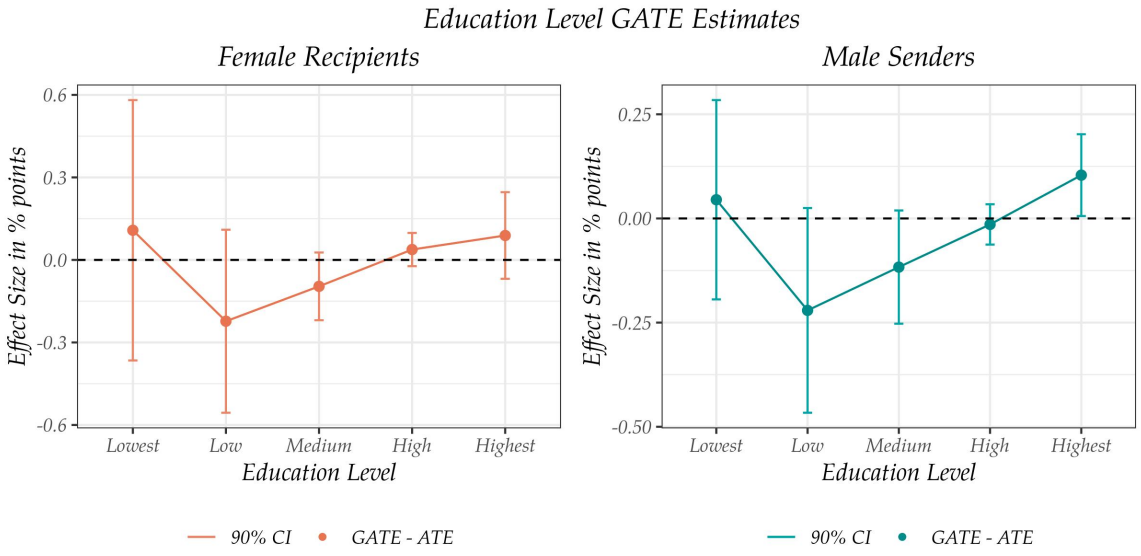
Note: Effects in % points as GATE deviations from the ATE (zero dotted line) with 90% confidence intervals.

Figure 3.C.5: Heterogeneous Effects of Sport Activity based on Education for Males



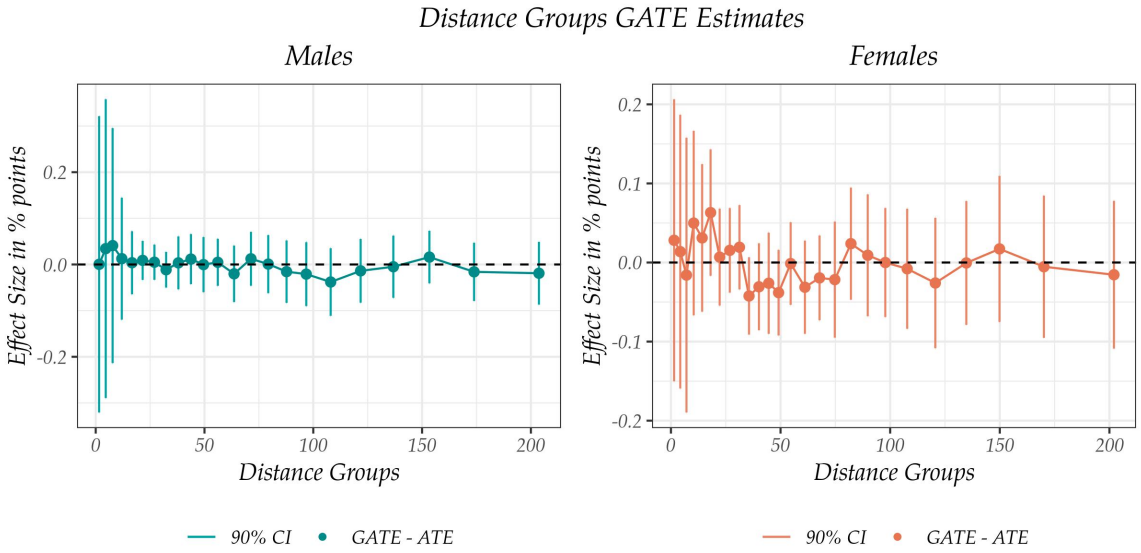
Note: Effects in % points as GATE deviations from the ATE (zero dotted line) with 90% confidence intervals.

Figure 3.C.6: Heterogeneous Effects of Sport Activity based on Education for Females



Note: Effects in % points as GATE deviations from the ATE (zero dotted line) with 90% confidence intervals.

Figure 3.C.7: Heterogeneous Effects of Sport Activity based on Distance



Note: Effects in % points as GATE deviations from the ATE (zero dotted line) with 90% confidence intervals.

3.C.2 Clustering Analysis

Table 3.C.3: Descriptive Clusters of IATEs based on the k-means++ Clustering

Code		Males					Females				
	Clusters	1	2	3	4	5	1	2	3	4	5
	IATEs: Weekly vs. Never	0.41	0.88	1.22	1.52	1.85	-1.41	-0.52	0.12	0.71	1.38
	Recipient Features										
5	Smoking Frequency	1.30	0.85	0.32	0.10	0.03	0.50	0.47	0.45	0.37	0.29
14	Relevance of Sexuality	0.38	0.44	0.47	0.49	0.52	0.21	0.27	0.31	0.32	0.31
108	TV in Leisure Time	0.27	0.26	0.25	0.25	0.28	0.32	0.27	0.21	0.15	0.11
223	Radio/TV at Home	0.62	0.62	0.60	0.59	0.60	0.73	0.68	0.65	0.64	0.67
284	Appearance Satisfaction	0.18	0.19	0.20	0.21	0.22	0.11	0.16	0.18	0.20	0.21
292	Importance of Sexuality	0.29	0.33	0.36	0.38	0.40	0.21	0.24	0.28	0.29	0.30
303	Comfortable Dining	0.67	0.63	0.57	0.55	0.55	0.73	0.66	0.60	0.57	0.58
324	Wish Significant Other	0.27	0.28	0.29	0.32	0.33	0.26	0.26	0.26	0.27	0.30
	Sender Features										
5	Smoking Frequency	0.69	0.52	0.36	0.31	0.27	0.48	0.50	0.49	0.42	0.32
14	Relevance of Sexuality	0.24	0.28	0.29	0.33	0.40	0.35	0.40	0.44	0.45	0.46
108	TV in Leisure Time	0.23	0.21	0.20	0.19	0.17	0.29	0.30	0.29	0.27	0.24
223	Radio/TV at Home	0.65	0.61	0.60	0.58	0.55	0.67	0.66	0.64	0.63	0.61
284	Appearance Satisfaction	0.18	0.17	0.17	0.19	0.22	0.12	0.16	0.17	0.19	0.23
292	Importance of Sexuality	0.21	0.24	0.23	0.25	0.28	0.29	0.32	0.33	0.35	0.38
303	Comfortable Dining	0.68	0.65	0.60	0.56	0.54	0.63	0.62	0.60	0.58	0.54
324	Wish Significant Other	0.23	0.23	0.25	0.25	0.27	0.28	0.29	0.31	0.32	0.35
	Observations										
	Share	0.07	0.18	0.29	0.31	0.15	0.07	0.20	0.29	0.29	0.15
	Total	2288	6439	10153	10826	5252	3753	11048	15896	15484	8047

Note: Means of clustered effects sorted in an increasing order, matched with selected user characteristics. Variable codes refer to the exact questions from the registration questionnaire presented in Table 3.D.1.

3.D Supplementary Material

3.D.1 Registration Questionnaire and Descriptive Statistics

Table 3.D.1: Summary of the Registration Questionnaire and the Descriptive Statistics

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 1	0.48	0.50	0.00	1.00	Gender.	Dummy	Female
No. 2	40.05	11.41	18.00	82.00	Age (in years).	Ordered	
No. 3	3.47	1.45	1.00	6.00	Gross annual income (EUR).	Ordered	Lowest income level (15,000 Euro)
						Ordered	Low income level (15,000 – 25,000 Euro)
						Ordered	Medium income level (25,000 – 35,000 Euro)
						Ordered	High income level (35,000 – 50,000 Euro)
						Ordered	Highest income level (50,000 – 100,000 Euro)
						Ordered	Highest income level (plus) (above 100,000 Euro)
No. 4	3.86	1.04	1.00	5.00	Education.	Ordered	Lowest education level (Lower Completion)
						Ordered	Low education level (Graduation)
						Ordered	Medium education level (Commercial- / Technical School Diploma)
						Ordered	High education level (High School)
						Ordered	Highest education level (Completed Studies)

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 5	0.47	0.75	0.00	2.00	Smoking.	Ordered	No
						Ordered	Sometimes
						Ordered	Yes
No. 6	174.95	9.23	133.00	208.00	Body height (in cm).	Ordered	
No. 7	0.61	0.49	0.00	1.00	Family status.	Unordered	Single.
	0.13	0.34	0.00	1.00		Unordered	Separated.
	0.21	0.40	0.00	1.00		Unordered	Divorced.
	0.05	0.21	0.00	1.00		Unordered	Widowed.
No. 8	0.70	1.04	0.00	30.00	Number of children.	Ordered	
No. 9	0.27	0.68	0.00	30.00	Number of children in own household.	Ordered	
No. 10	0.46	0.50	0.00	1.00	Desire to have children.	Unordered	No information.
	0.05	0.23	0.00	1.00		Unordered	No.
	0.45	0.50	0.00	1.00		Unordered	Irrelevant.
	0.03	0.17	0.00	1.00		Unordered	Yes.
No. 11	0.38	0.49	0.00	1.00	Apart from love and affection, what are the main reasons for your desire for partnership? A maximum of 3 answers is possible.	Dummy	Life is easier to master when there are two of you.
No. 12	0.48	0.50	0.00	1.00		Dummy	A partner would give me emotional comfort.
No. 13	0.41	0.49	0.00	1.00		Dummy	I need someone I trust completely.
No. 14	0.34	0.47	0.00	1.00		Dummy	I want to live regular sexuality.
No. 15	0.45	0.50	0.00	1.00		Dummy	I would like to spend much of my free time together with a partner.
No. 16	0.30	0.46	0.00	1.00		Dummy	I do not want to grow old alone.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 17	0.11	0.31	0.00	1.00		Dummy	A partnership offers more security in every respect.
No. 18	0.03	0.18	0.00	1.00	Which statement should apply most to your preferred partner?	Dummy	We fit together based on our external appearance.
No. 19	0.42	0.49	0.00	1.00		Dummy	We both have the same interests.
No. 20	0.55	0.50	0.00	1.00		Dummy	He / she has a strong appeal to me.
No. 21	0.19	0.40	0.00	1.00	What would you be most interested in if you found someone attractive? Exact two answers required.	Dummy	What he / she does professionally.
No. 22	0.16	0.37	0.00	1.00		Dummy	Whether he / she lives in secure financial circumstances.
No. 23	0.30	0.46	0.00	1.00		Dummy	Health and vitality.
No. 24	0.79	0.41	0.00	1.00		Dummy	Warm-heartedness.
No. 25	0.55	0.50	0.00	1.00		Dummy	The external appearance.
No. 26	0.32	0.47	0.00	1.00	Suppose you and your partner are invited to a wedding party of friends. You are just getting ready. Just as you know yourself: Which thoughts are going through your mind most likely?	Dummy	Whether we make a good impression.
No. 27	0.27	0.44	0.00	1.00		Dummy	Whether what we bring along is appropriate.
No. 28	0.19	0.39	0.00	1.00		Dummy	Whether there are not too many people, I do not know.
No. 29	0.23	0.42	0.00	1.00		Dummy	I am starting to realize once again that dress codes are not for me.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 30	0.30	0.46	0.00	1.00	Why do you think you have not yet found the right partner?	Dummy	I am very demanding with respect to the future partner.
No. 31	0.17	0.37	0.00	1.00		Dummy	I simply was not ready yet.
No. 32	0.18	0.38	0.00	1.00		Dummy	I am too shy or too inhibited.
No. 33	0.10	0.29	0.00	1.00		Dummy	I probably had too little time or opportunities to make more deep contacts.
No. 34	0.26	0.44	0.00	1.00		Dummy	I have closed myself too much in the past for some reason.
No. 35	0.47	0.50	0.00	1.00	If you really liked a book or magazine article, would you like your partner to read it too?	Dummy	Yes, I would have more pleasure with it.
No. 36	0.53	0.50	0.00	1.00		Dummy	I do not care.
No. 37	0.84	0.37	0.00	1.00	Suppose you live together with your partner in a two-room apartment. How would you furnish the apartment?	Dummy	In any case a shared bedroom.
No. 38	0.16	0.37	0.00	1.00		Dummy	Everyone should have an own room, but at least one of them should have space for shared nights.
No. 39	0.47	0.50	0.00	1.00	How do you react to lovesickness?	Dummy	I lose the joy of eating.
No. 40	0.12	0.32	0.00	1.00		Dummy	I eat more.
No. 41	0.41	0.49	0.00	1.00		Dummy	Neither nor.
No. 42	0.68	0.46	0.00	1.00	Which statement about sexual loyalty in the partnership comes closest to your attitude?	Dummy	Absolute loyalty without exception!

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 43	0.18	0.39	0.00	1.00		Dummy	It is important to always strive to be loyal.
No. 44	0.04	0.20	0.00	1.00		Dummy	To be loyal in the heart is much more important than physical loyalty.
No. 45	0.05	0.22	0.00	1.00		Dummy	Especially in a long partnership a gaffe can happen.
No. 46	0.04	0.20	0.00	1.00		Dummy	To demand absolute loyalty is possessive thinking.
No. 47	0.34	0.47	0.00	1.00	What is your idea of the external form of a marriage?	Dummy	In any case some kind of ritual, like a church wedding can be.
No. 48	0.09	0.29	0.00	1.00		Dummy	A legally binding contract is sufficient for me.
No. 49	0.29	0.45	0.00	1.00		Dummy	Nothing special, I would fully agree with the wishes of my partner in that case.
No. 50	0.28	0.45	0.00	1.00		Dummy	I have no concept of it.
No. 51	0.03	0.18	0.00	1.00	Which terms describe characteristics that you would like the other person to appreciate in you? A maximum of 5 answers is possible.	Dummy	Seriously.
No. 52	0.06	0.23	0.00	1.00		Dummy	Cheerful.
No. 53	0.53	0.50	0.00	1.00		Dummy	Humorous.
No. 54	0.24	0.42	0.00	1.00		Dummy	Uncomplicated.
No. 55	0.26	0.44	0.00	1.00		Dummy	Naturally.
No. 56	0.11	0.31	0.00	1.00		Dummy	Justice-loving.
No. 57	0.07	0.25	0.00	1.00		Dummy	Adaptable.
No. 58	0.30	0.46	0.00	1.00		Dummy	Sensitive.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 59	0.28	0.45	0.00	1.00		Dummy	Tender.
No. 60	0.08	0.27	0.00	1.00		Dummy	Spirited.
No. 61	0.04	0.19	0.00	1.00		Dummy	Restrained.
No. 62	0.01	0.11	0.00	1.00		Dummy	Frugal.
No. 63	0.06	0.23	0.00	1.00		Dummy	Domesticated.
No. 64	0.13	0.34	0.00	1.00		Dummy	Close to nature.
No. 65	0.17	0.37	0.00	1.00		Dummy	Optimistic.
No. 66	0.05	0.22	0.00	1.00		Dummy	Capable.
No. 67	0.17	0.37	0.00	1.00		Dummy	Fond of children.
No. 68	0.13	0.33	0.00	1.00		Dummy	Strong of character.
No. 69	0.11	0.32	0.00	1.00		Dummy	Handsome.
No. 70	0.30	0.46	0.00	1.00		Dummy	Warm-hearted.
No. 71	0.15	0.36	0.00	1.00		Dummy	Educated.
No. 72	0.07	0.26	0.00	1.00		Dummy	Value-conscious.
No. 73	0.08	0.28	0.00	1.00		Dummy	Good manners.
No. 74	0.05	0.23	0.00	1.00		Dummy	Thoughtful.
No. 75	0.09	0.28	0.00	1.00		Dummy	Independent.
No. 76	0.11	0.32	0.00	1.00		Dummy	Tolerant.
No. 77	0.14	0.34	0.00	1.00		Dummy	Spontaneous.
No. 78	0.15	0.36	0.00	1.00		Dummy	Self-confident.
No. 79	0.07	0.26	0.00	1.00		Dummy	Imaginative.
No. 80	0.03	0.16	0.00	1.00		Dummy	Career conscious.
No. 81	0.41	0.49	0.00	1.00		Dummy	Reliable.
No. 82	0.09	0.29	0.00	1.00		Dummy	Calm.
No. 83	0.18	0.38	0.00	1.00		Dummy	Sympathetic.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 84	0.25	0.43	0.00	1.00	What do you think people who know you well are most likely to think about you? Exact two answers required.	Dummy	Is ready for any fun.
No. 85	0.23	0.42	0.00	1.00		Dummy	Gets the bright side out of life.
No. 86	0.26	0.44	0.00	1.00		Dummy	Thinks a lot and seriously about life.
No. 87	0.21	0.41	0.00	1.00		Dummy	Is always in a good mood and happy.
No. 88	0.11	0.31	0.00	1.00		Dummy	A little dreamy.
No. 89	0.33	0.47	0.00	1.00		Dummy	Approaches the problem objectively and deliberately.
No. 90	0.29	0.46	0.00	1.00		Dummy	Finds a good solution even in unpleasant situations for herself / himself.
No. 91	0.19	0.40	0.00	1.00		Dummy	Nothing can upset her / him.
No. 92	0.13	0.34	0.00	1.00		Dummy	Takes lively part in everything.
No. 93	0.29	0.45	0.00	1.00	What seems most important to you in a partnership? Exact two answers required.	Dummy	Give each other plenty of space.
No. 94	0.65	0.48	0.00	1.00		Dummy	To coordinate the wishes of the individual with each other.
No. 95	0.38	0.49	0.00	1.00		Dummy	Do not always weigh everything on the gold scale.
No. 96	0.16	0.37	0.00	1.00		Dummy	Steer life in a calmer direction.
No. 97	0.30	0.46	0.00	1.00		Dummy	Also let five be straight sometimes.
No. 98	0.15	0.36	0.00	1.00		Dummy	Taking completely new paths.
No. 99	0.08	0.27	0.00	1.00		Dummy	Preserving the tried and tested.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 100	0.23	0.42	0.00	1.00	Imagine family and friends: What reaction would you attach particular importance to when it comes to your choice of your partner?	Dummy	I would value that my family agrees with my choice of partner.
No. 101	0.31	0.46	0.00	1.00		Dummy	That my friends are happy about my choice of partner.
No. 102	0.24	0.43	0.00	1.00		Dummy	I do not care what families or friends think of my choice of partner.
No. 103	0.22	0.41	0.00	1.00		Dummy	I would value that my partner's family likes me.
No. 104	0.22	0.41	0.00	1.00	Do you drink alcohol?	Dummy	Yes, for example at meals, in society, for relaxation.
No. 105	0.68	0.47	0.00	1.00		Dummy	Occasionally.
No. 106	0.10	0.30	0.00	1.00		Dummy	No.
No. 107	0.32	0.47	0.00	1.00	What do you like to do in your leisure time? A maximum of 3 answers is possible.	Dummy	Reading.
No. 108	0.25	0.43	0.00	1.00		Dummy	Watching TV.
No. 109	0.43	0.50	0.00	1.00		Dummy	Relaxing.
No. 110	0.66	0.47	0.00	1.00		Dummy	Going out.
No. 111	0.30	0.46	0.00	1.00		Dummy	Cinema.
No. 112	0.58	0.49	0.00	1.00		Dummy	Pursuing my hobbies.
No. 113	0.12	0.33	0.00	1.00		Dummy	Playing in convivial gatherings.
No. 114	0.42	0.49	0.00	1.00	What is your favorite way to spend your leisure time?	Dummy	At home—or I visit a friend.
No. 115	0.32	0.47	0.00	1.00		Dummy	In the free nature.
No. 116	0.26	0.44	0.00	1.00		Dummy	In convivial gatherings.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 117	0.26	0.44	0.00	1.00	Do you like cooking?	Dummy	Yes, I really enjoy cooking.
No. 118	0.39	0.49	0.00	1.00		Dummy	Yes, very much.
No. 119	0.13	0.33	0.00	1.00		Dummy	I only cook when I have to.
No. 120	0.10	0.30	0.00	1.00		Dummy	I only like to cook when I want to host visitors.
No. 121	0.12	0.33	0.00	1.00		Dummy	I cannot cook well.
No. 122	0.02	0.15	0.00	1.00	What special interests / hobbies do you have? A maximum of 6 answers is possible.	Dummy	Theater.
No. 123	0.55	0.50	0.00	1.00		Dummy	Photography.
No. 124	0.10	0.30	0.00	1.00		Dummy	Film / Video.
No. 125	0.04	0.20	0.00	1.00		Dummy	Literature.
No. 126	0.06	0.23	0.00	1.00		Dummy	Art.
No. 127	0.04	0.20	0.00	1.00		Dummy	Music.
No. 128	0.07	0.26	0.00	1.00		Dummy	Cooking.
No. 129	0.16	0.37	0.00	1.00		Dummy	Cinema.
No. 130	0.13	0.34	0.00	1.00		Dummy	Architecture.
No. 131	0.08	0.27	0.00	1.00		Dummy	History.
No. 132	0.00	0.06	0.00	1.00		Dummy	Carpentry / crafts.
No. 133	0.13	0.34	0.00	1.00		Dummy	Pottery.
No. 134	0.25	0.43	0.00	1.00		Dummy	Handworks.
No. 135	0.02	0.14	0.00	1.00		Dummy	Collecting.
No. 136	0.17	0.37	0.00	1.00	What kind of music do you like to listen to? Multiple answers are possible.	Dummy	Musicals / Operettas.
No. 137	0.09	0.29	0.00	1.00		Dummy	Operas.
No. 138	0.17	0.37	0.00	1.00		Dummy	Symphony concerts.
No. 139	0.05	0.21	0.00	1.00		Dummy	Chamber music.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 140	0.04	0.18	0.00	1.00		Dummy	Folk Music.
No. 141	0.22	0.41	0.00	1.00		Dummy	Schlager.
No. 142	0.11	0.31	0.00	1.00		Dummy	Chansons / Songs.
No. 143	0.03	0.17	0.00	1.00		Dummy	Ethno.
No. 144	0.25	0.43	0.00	1.00		Dummy	Jazz.
No. 145	0.82	0.38	0.00	1.00		Dummy	Rock.
No. 146	0.24	0.43	0.00	1.00		Dummy	Metal / Hard Rock.
No. 147	0.17	0.38	0.00	1.00		Dummy	Reggae.
No. 148	0.26	0.44	0.00	1.00		Dummy	Rap.
No. 149	0.35	0.48	0.00	1.00		Dummy	Dance.
No. 150	0.30	0.46	0.00	1.00		Dummy	House.
No. 151	0.00	0.01	0.00	1.00		Dummy	Other
No. 152	0.78	0.41	0.00	1.00	Do you play an instrument?	Dummy	No.
No. 153	0.22	0.41	0.00	1.00		Dummy	Yes.
No. 154	0.51	0.50	0.00	1.00	What is your favorite form of holiday? Multiple answers are possible.	Dummy	Sun and beach.
No. 155	0.10	0.30	0.00	1.00		Dummy	Study trips.
No. 156	0.04	0.19	0.00	1.00		Dummy	Meditation.
No. 157	0.13	0.34	0.00	1.00		Dummy	Boat trips.
No. 158	0.17	0.38	0.00	1.00		Dummy	At home.
No. 159	0.54	0.50	0.00	1.00		Dummy	Cities, culture and art.
No. 160	0.49	0.50	0.00	1.00		Dummy	Relaxation holidays.
No. 161	0.70	0.46	0.00	1.00		Dummy	At the sea.
No. 162	0.38	0.48	0.00	1.00		Dummy	In the mountains.
No. 163	0.17	0.38	0.00	1.00		Dummy	Camping.
No. 164	0.25	0.43	0.00	1.00		Dummy	Adventure holidays.
No. 165	0.27	0.45	0.00	1.00		Dummy	Beauty / wellness holidays.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 166	0.06	0.24	0.00	1.00		Dummy	Group tours.
No. 167	0.10	0.29	0.00	1.00	How do you plan your holiday?	Dummy	As little as possible: I prefer to drive into the blue.
No. 168	0.41	0.49	0.00	1.00		Dummy	I plan and organize my holiday carefully and early.
No. 169	0.49	0.50	0.00	1.00		Dummy	Once the date and destination are fixed, I like to leave everything else to the moment.
No. 170	0.83	0.37	0.00	1.00	Do you like to take longer walks?	Dummy	Yes.
No. 171	0.17	0.37	0.00	1.00		Dummy	No.
No. 172	0.00	0.02	0.00	1.00	How do you proceed when you have private plans?	Dummy	I am proceeding fairly systematically.
No. 173	0.00	0.01	0.00	1.00		Dummy	I think it will work out somehow.
No. 174	0.47	0.50	0.00	1.00	Do you usually feel more comfortable at home than in society?	Dummy	Yes.
No. 175	0.53	0.50	0.00	1.00		Dummy	No.
No. 176	0.81	0.39	0.00	1.00	How must a living room be tempered so that you feel really comfortable?	Dummy	Well warm (21C [69.8F] or slightly more).
No. 177	0.19	0.39	0.00	1.00		Dummy	Rather cool (19C [66.2F] or a little less).
No. 178	0.11	0.31	0.00	1.00	Do you sleep with the window open?	Dummy	Yes, absolutely.
No. 179	0.49	0.50	0.00	1.00		Dummy	Yes, if it is possible.
No. 180	0.17	0.38	0.00	1.00		Dummy	No, I find that uncomfortable.
No. 181	0.23	0.42	0.00	1.00		Dummy	I do not really care.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 182	0.24	0.43	0.00	1.00	There are people who are very lively in the morning; others only become really active in the evening. How is it with you?	Dummy	Alive in the morning.
No. 183	0.34	0.47	0.00	1.00		Dummy	Rather lively in the evening.
No. 184	0.41	0.49	0.00	1.00		Dummy	It makes no difference to me.
No. 185	0.73	0.44	0.00	1.00	(In the next section you will see pairs of images. Please choose spontaneously the image you like most. Our tip: Just follow your gut feeling.) Which one do you like more?	Dummy	(Image 1)
No. 186	0.27	0.44	0.00	1.00		Dummy	(Imag 2)
No. 187	0.65	0.48	0.00	1.00	(In the next section you will see pairs of images. Please choose spontaneously the image you like most. Our tip: Just follow your gut feeling.) Which arrangement appeals to you more emotionally?	Dummy	(Image 3)
No. 188	0.35	0.48	0.00	1.00		Dummy	(Image 4)
No. 189	0.71	0.45	0.00	1.00	(In the next section you will see pairs of images. Please choose spontaneously the image you like most. Our tip: Just follow your gut feeling.) Which image appeals to you more?	Dummy	(Image 5)
No. 190	0.29	0.45	0.00	1.00		Dummy	(Image 6)

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 191	0.55	0.50	0.00	1.00	(In the next section you will see pairs of images. Please choose spontaneously the image you like most. Our tip: Just follow your gut feeling.) Do not think about it for long, decide for a shape!	Dummy	(Image 7)
No. 192	0.45	0.50	0.00	1.00		Dummy	(Image 8)
No. 193	0.69	0.46	0.00	1.00	(In the next section you will see pairs of images. Please choose spontaneously the image you like most. Our tip: Just follow your gut feeling.) Which of these two arrangements do you prefer more?	Dummy	(Image 9)
No. 194	0.31	0.46	0.00	1.00		Dummy	(Image 10)
No. 195	0.48	0.50	0.00	1.00	(In the next section you will see pairs of images. Please choose spontaneously the image you like most. Our tip: Just follow your gut feeling.) Which movement can you empathize with better?	Dummy	(Image 11)
No. 196	0.52	0.50	0.00	1.00		Dummy	(Image 12)
No. 197	0.51	0.50	0.00	1.00	(In the next section you will see pairs of images. Please choose spontaneously the image you like most. Our tip: Just follow your gut feeling.) Which image do you prefer more?	Dummy	(Image 13)
No. 198	0.49	0.50	0.00	1.00		Dummy	(Image 14)

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 199	0.11	0.31	0.00	1.00	(In the next section you will see pairs of images. Please choose spontaneously the image you like most. Our tip: Just follow your gut feeling.) Get a feel for the different directions of movement. Decide on one.	Dummy	(Image 15)
No. 200	0.89	0.31	0.00	1.00		Dummy	(Image 16)
No. 201	0.41	0.49	0.00	1.00	(In the next section you will see pairs of images. Please choose spontaneously the image you like most. Our tip: Just follow your gut feeling.) Which one do you like more?	Dummy	(Image 17)
No. 202	0.59	0.49	0.00	1.00		Dummy	(Image 18)
No. 203	0.73	0.44	0.00	1.00	(In the next section you will see pairs of images. Please choose spontaneously the image you like most. Our tip: Just follow your gut feeling.) Which representation do you prefer more?	Dummy	(Image 19)
No. 204	0.27	0.44	0.00	1.00		Dummy	(Image 20)
No. 205	0.54	0.50	0.00	1.00	(In the next section you will see pairs of images. Please choose spontaneously the image you like most. Our tip: Just follow your gut feeling.) Which image appeals to you more emotionally?	Dummy	(Image 21)

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 206	0.46	0.50	0.00	1.00		Dummy	(Image 22)
No. 207	0.77	0.42	0.00	1.00	Do you get excited about something easily?	Dummy	No, not necessarily.
No. 208	0.23	0.42	0.00	1.00		Dummy	Yes, very much.
No. 209	0.26	0.44	0.00	1.00	If you like a track or song well: Why is that mostly?	Dummy	I like the text.
No. 210	0.60	0.49	0.00	1.00		Dummy	I like the rhythm.
No. 211	0.14	0.34	0.00	1.00		Dummy	I like the melody.
No. 212	0.26	0.44	0.00	1.00	Which tones appeal to you most?	Dummy	Saxophone tones.
No. 213	0.16	0.37	0.00	1.00		Dummy	Violin sounds.
No. 214	0.58	0.49	0.00	1.00		Dummy	Piano playing.
No. 215	0.75	0.43	0.00	1.00	Regardless of what is fashionable at the moment: You choose your clothes in terms of style and color tone...	Dummy	Covered and discreet.
No. 216	0.25	0.43	0.00	1.00		Dummy	Bold and expressive.
No. 217	0.27	0.44	0.00	1.00	What type of house appeals to you most?	Dummy	Image 23 (Country house)
No. 218	0.47	0.50	0.00	1.00		Dummy	Image 24 (City villa)
No. 219	0.26	0.44	0.00	1.00		Dummy	Image 25 (Architect house)
No. 220	0.25	0.43	0.00	1.00	Which of these three plants do you prefer to look at most?	Dummy	Image 26 (Orchid)
No. 221	0.60	0.49	0.00	1.00		Dummy	Image 27 (Strelitzie)
No. 222	0.15	0.36	0.00	1.00		Dummy	Image 28 (Rose)
No. 223	0.64	0.48	0.00	1.00	When you get home and are alone, do you habitually turn on the radio / TV / music?	Dummy	Yes.
No. 224	0.36	0.48	0.00	1.00		Dummy	No.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 225	0.36	0.48	0.00	1.00	How do you prefer to dress most? A maximum of 2 answers is possible.	Dummy	Casual.
No. 226	0.21	0.41	0.00	1.00		Dummy	Practical.
No. 227	0.19	0.39	0.00	1.00		Dummy	Elegant.
No. 228	0.29	0.45	0.00	1.00		Dummy	Fashionable.
No. 229	0.32	0.47	0.00	1.00		Dummy	Correct and adapted to the situation.
No. 230	0.11	0.31	0.00	1.00		Dummy	Very personal and unconventional.
No. 231	0.05	0.22	0.00	1.00	Imagine: You slip on a banana peel on the sidewalk. You have not hurt yourself, but people turn and stop. One want to help you. What could be your first reaction?	Dummy	I am annoyed that there are people who throw a banana peel on the sidewalk without thinking about it.
No. 232	0.21	0.41	0.00	1.00		Dummy	I get up and carry the banana peel to the nearest waste bin so that the same does not happen to others.
No. 233	0.14	0.34	0.00	1.00		Dummy	≪Ouch! Been lucky again!≫
No. 234	0.19	0.39	0.00	1.00		Dummy	I downplay the incident because I find it unpleasant to draw so much attention to myself.
No. 235	0.11	0.31	0.00	1.00		Dummy	While I am still sitting, I look at the people from below and say ≪What a show, I could play with it, right?≫
No. 236	0.30	0.46	0.00	1.00		Dummy	I get up and say ≪Nothing happens!≫ and keep going.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 237	0.15	0.36	0.00	1.00	Imagine yourself: You live in a larger apartment building. At two thirty in the morning, your doorbell rings. Someone answers the intercom and asks if a Mr. Müller lives in the house. That is indeed the case—and of course your neighbor also has his own bell. What could you say?	Dummy	≪Try again with another bell.≫
No. 238	0.02	0.15	0.00	1.00		Dummy	≪For that you woke me from the deepest sleep!≫ And I end the conversation.
No. 239	0.04	0.19	0.00	1.00		Dummy	I do not get up at all, but pull the blanket over my head and try to continue sleeping.
No. 240	0.05	0.23	0.00	1.00		Dummy	I think something could have happened. When then asked for Müller, I swear into the apparatus: ≪Are you crazy! What do you think of disturbing strangers in the middle of the night?≫
No. 241	0.34	0.47	0.00	1.00		Dummy	≪If you like the Müller, let him sleep.≫
No. 242	0.18	0.39	0.00	1.00		Dummy	I think something could have happened. When then asked for Müller, I let myself explain what he / she wants from Mr. Müller at night.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 243	0.22	0.41	0.00	1.00		Dummy	≪Here is not Müller, but I know that the name badges are actually not easily recognizable.≫
No. 244	0.16	0.37	0.00	1.00	Imagine yourself: A friend buys a new car that is far too expensive—far beyond his / her circumstances. It is the car you have always dreamed of. What could you say to your friend?	Dummy	≪One should not live above one's means. When will you finally grow up?≫
No. 245	0.19	0.39	0.00	1.00		Dummy	≪Wonder of technology! I am glad you allowed yourself this. One could get jealous.≫
No. 246	0.05	0.23	0.00	1.00		Dummy	≪It is so beautiful, I would be afraid to park it in public and get a bump or a scratch directly.≫
No. 247	0.06	0.24	0.00	1.00		Dummy	≪Think about it, once you have driven a meter with it, the car is only worth half . . . ≫
No. 248	0.46	0.50	0.00	1.00		Dummy	≪Oh cool! Let us go for a spin!≫
No. 249	0.08	0.26	0.00	1.00		Dummy	≪I think you need a chauffeur for that—I will break the car in for you!≫
No. 250	0.12	0.32	0.00	1.00	Imagine yourself: You and a friend were very upset about another person. Your friend makes the suggestion to pay it back to the other person. What could be your first reaction?	Dummy	≪I would think twice about it. Anyone digging a pit for others falls into it herself / himself.≫

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 251	0.04	0.20	0.00	1.00		Dummy	≪I do not know. When that comes out. . . I want to be left alone.≫
No. 252	0.37	0.48	0.00	1.00		Dummy	≪Forget it, we will laugh about it in a year.≫
No. 253	0.11	0.32	0.00	1.00		Dummy	≪That is nasty, I do not take part in it.≫
No. 254	0.06	0.24	0.00	1.00		Dummy	≪I think it could be quite funny.≫
No. 255	0.29	0.46	0.00	1.00		Dummy	≪Let it be, otherwise he / she is quite alright.≫
No. 256	0.40	0.49	0.00	1.00	What is your first impulse when you get very angry about the behavior of a person close to you?	Dummy	I clearly say that I am angry.
No. 257	0.27	0.45	0.00	1.00		Dummy	I stay calm and try to clarify the situation.
No. 258	0.11	0.31	0.00	1.00		Dummy	I think to myself: It does not happen that often.
No. 259	0.22	0.41	0.00	1.00		Dummy	I swallow the anger and grit my teeth.
No. 260	0.24	0.42	0.00	1.00	Sometimes it happens that one is offended by a person. How do you react to that?	Dummy	I think maybe it was not meant that way.
No. 261	0.19	0.39	0.00	1.00		Dummy	I am sure that I will find a way to deal with it.
No. 262	0.53	0.50	0.00	1.00		Dummy	I have been gnawing on it for a while.
No. 263	0.05	0.21	0.00	1.00		Dummy	I would like to pay back something like this immediately.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 264	0.23	0.42	0.00	1.00	If someone contradicts you even though you know you are right, how do you usually react?	Dummy	I am annoyed by the bossiness of the other, but I leave it at that.
No. 265	0.14	0.35	0.00	1.00		Dummy	It is not so important to me to be right.
No. 266	0.50	0.50	0.00	1.00		Dummy	I try to convince the other.
No. 267	0.13	0.34	0.00	1.00		Dummy	I will clarify who is right.
No. 268	0.36	0.48	0.00	1.00	Imagine you are at a party with a man / a woman you love. Suddenly you see that he / she is flirting with another person. How do you react to that?	Dummy	I try to disturb the flirt.
No. 269	0.29	0.45	0.00	1.00		Dummy	I suffer silently and say nothing.
No. 270	0.18	0.39	0.00	1.00		Dummy	I also flirt.
No. 271	0.16	0.37	0.00	1.00		Dummy	I do not mind. I allow him / her the fun.
No. 272	0.16	0.37	0.00	1.00	A scene from a dream. Select the title of the image that you think best expresses the content. Image 29.	Dummy	Adventure in Scotland.
No. 273	0.59	0.49	0.00	1.00		Dummy	Lost passion.
No. 274	0.25	0.43	0.00	1.00		Dummy	Power of conscience.
No. 275	0.33	0.47	0.00	1.00	Take a close look at this dream scene: Which title best reflects the image content for you? Image 30.	Dummy	Dance of the Vampires.
No. 276	0.59	0.49	0.00	1.00		Dummy	Voices from the afterlife.
No. 277	0.08	0.28	0.00	1.00		Dummy	Fun society.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 278	0.19	0.39	0.00	1.00	What title would you give this dream scene? Image 31.	Dummy	Cool sensuality.
No. 279	0.15	0.36	0.00	1.00		Dummy	Goddess of lust.
No. 280	0.66	0.48	0.00	1.00		Dummy	End of freedom.
No. 281	0.57	0.50	0.00	1.00	One last dream scene. Which image title best expresses the content for you? Image 32.	Dummy	Opera ball.
No. 282	0.30	0.46	0.00	1.00		Dummy	The joy of the game.
No. 283	0.14	0.34	0.00	1.00		Dummy	The winner.
No. 284	0.18	0.38	0.00	1.00	Are you satisfied with your external appearance?	Dummy	Yes.
No. 285	0.51	0.50	0.00	1.00		Dummy	Yes, on the whole.
No. 286	0.18	0.39	0.00	1.00		Dummy	That fluctuates.
No. 287	0.13	0.34	0.00	1.00		Dummy	I am also sometimes dissatisfied.
No. 288	0.27	0.45	0.00	1.00	Do you believe in the good in humans?	Dummy	Yes, always.
No. 289	0.43	0.49	0.00	1.00		Dummy	I try it.
No. 290	0.11	0.31	0.00	1.00		Dummy	Sometimes I find it hard to believe.
No. 291	0.19	0.39	0.00	1.00		Dummy	It depends on the context in which.
No. 292	0.28	0.45	0.00	1.00	How important is sexuality to you?	Dummy	Very important.
No. 293	0.61	0.49	0.00	1.00		Dummy	Important.
No. 294	0.11	0.31	0.00	1.00		Dummy	Less important.
No. 295	0.01	0.11	0.00	1.00		Dummy	Not so important.
No. 296	0.45	0.50	0.00	1.00	What is your basic view of the institution of marriage?	Dummy	If two really love each other, then they should marry.
No. 297	0.29	0.46	0.00	1.00		Dummy	Whoever wants to found a family should also marry.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 298	0.26	0.44	0.00	1.00		Dummy	Marriage as an institution is completely unnecessary.
No. 299	0.32	0.47	0.00	1.00	Is it important for you that everything is always in the place where it actually belongs?	Dummy	Yes, absolutely.
No. 300	0.43	0.50	0.00	1.00		Dummy	Not necessarily.
No. 301	0.25	0.43	0.00	1.00		Dummy	What does right place mean, it can mean something completely different for everyone.
No. 302	0.23	0.42	0.00	1.00	What significance does food have for you?	Dummy	My main priority is to eat well.
No. 303	0.61	0.49	0.00	1.00		Dummy	I love to eat comfortably.
No. 304	0.16	0.37	0.00	1.00		Dummy	The most important thing for me is a healthy diet.
No. 305	0.54	0.50	0.00	1.00	Do you attach importance to regular meals?	Dummy	As far as possible for me, I eat regularly and at fixed times.
No. 306	0.46	0.50	0.00	1.00		Dummy	No, not at all. I eat when I am hungry.
No. 307	0.19	0.40	0.00	1.00	How do you find it when advertisements on television or in newspapers are sexually emphasized?	Dummy	Disturbing, sometimes tasteless.
No. 308	0.41	0.49	0.00	1.00		Dummy	Quite pleasant indeed.
No. 309	0.40	0.49	0.00	1.00		Dummy	Uninteresting.
No. 310	0.43	0.50	0.00	1.00	Which of the following values are the most important for you in life? A maximum of 2 answers is possible.	Dummy	True friendship.
No. 311	0.40	0.49	0.00	1.00		Dummy	Happiness in love.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 312	0.24	0.43	0.00	1.00		Dummy	Letting myself get involved in something I like.
No. 313	0.08	0.28	0.00	1.00		Dummy	Professional success.
No. 314	0.18	0.38	0.00	1.00		Dummy	To be valued and respected by the people in my environment.
No. 315	0.12	0.32	0.00	1.00		Dummy	Social security.
No. 316	0.09	0.29	0.00	1.00		Dummy	Self-realization.
No. 317	0.44	0.50	0.00	1.00		Dummy	A familiar home with a partner.
No. 318	0.17	0.37	0.00	1.00	What maxim do you think is best to live by?	Dummy	Work before pleasure!
No. 319	0.21	0.41	0.00	1.00		Dummy	Love thy neighbor as thyself!
No. 320	0.62	0.49	0.00	1.00		Dummy	Live and let live!
No. 321	0.15	0.36	0.00	1.00	What is currently your greatest wish? A maximum of 2 answers is possible.	Dummy	Expand professional opportunities.
No. 322	0.15	0.36	0.00	1.00		Dummy	Get to know nice and interesting people.
No. 323	0.07	0.26	0.00	1.00		Dummy	Make good friends.
No. 324	0.27	0.44	0.00	1.00		Dummy	Discover the great love.
No. 325	0.53	0.50	0.00	1.00		Dummy	Finding a partner for life.
No. 326	0.14	0.35	0.00	1.00		Dummy	Start over again.
No. 327	0.53	0.50	0.00	1.00		Dummy	Build a stable relationship that makes me feel safe.
No. 328	0.33	0.47	0.00	1.00	Does it bother you if people in your area use their mobile phones without hesitation?	Dummy	No, I have got used to it by now.
No. 329	0.48	0.50	0.00	1.00		Dummy	Actually yes, but one has to live with that today.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 330	0.16	0.36	0.00	1.00		Dummy	Yes, I think that is terrible.
No. 331	0.03	0.18	0.00	1.00		Dummy	I then take the opportunity to have conversations myself.
No. 332	0.29	0.45	0.00	1.00	There is a lot of discussion about climate change, environmental protection, energy sources, etc. Which statement is closest to your opinion?	Dummy	Considering our advanced technology, we should be able to come up with something reasonable.
No. 333	0.61	0.49	0.00	1.00		Dummy	We should simply accept that we have to be more careful with nature.
No. 334	0.11	0.31	0.00	1.00		Dummy	I would like not to think at all about what is still to come.
No. 335	0.04	0.19	0.00	1.00	What would most disturb you in your partner's environment during the introduction phase?	Dummy	A dominant father-in-law (in the making).
No. 336	0.07	0.25	0.00	1.00		Dummy	An overprotective mother-in-law (in the making).
No. 337	0.21	0.41	0.00	1.00		Dummy	Too much influence of old friends.
No. 338	0.60	0.49	0.00	1.00		Dummy	Bad moods of the partner that pull me down.
No. 339	0.09	0.28	0.00	1.00		Dummy	Too freaky types in the circle of acquaintances.
No. 340	0.10	0.30	0.00	1.00	Regardless of your current place of residence, where would you most like to live?	Dummy	In a large city with a metropolitan feeling.
No. 341	0.30	0.46	0.00	1.00		Dummy	In the environment of a larger city.
No. 342	0.16	0.37	0.00	1.00		Dummy	In a more tranquil small city.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 343	0.18	0.38	0.00	1.00		Dummy	A little quieter or completely in the rural area.
No. 344	0.27	0.44	0.00	1.00		Dummy	It does not matter—I can feel comfortable at many places...
No. 345	3.24	2.27	0.00	12.00	Profile images (count).	Ordered	
No. 346	0.00	0.06	0.00	1.00	Language(s).	Dummy	Danish.
No. 347	0.58	0.49	0.00	1.00		Dummy	German.
No. 348	0.57	0.50	0.00	1.00		Dummy	English.
No. 349	0.06	0.24	0.00	1.00		Dummy	Spanish.
No. 350	0.00	0.03	0.00	1.00		Dummy	Finnish.
No. 351	0.14	0.34	0.00	1.00		Dummy	French.
No. 352	0.03	0.16	0.00	1.00		Dummy	Italian.
No. 353	0.01	0.10	0.00	1.00		Dummy	Dutch.
No. 354	0.00	0.04	0.00	1.00		Dummy	Norwegian.
No. 355	0.01	0.10	0.00	1.00		Dummy	Polish.
No. 356	0.01	0.07	0.00	1.00		Dummy	Portuguese.
No. 357	0.02	0.15	0.00	1.00		Dummy	Russian.
No. 358	0.00	0.07	0.00	1.00		Dummy	Swedish.
No. 359	0.00	0.07	0.00	1.00		Dummy	Turkish.
No. 360	0.09	0.29	0.00	1.00	Occupation.	Unordered	Office and Administrative Support Occupations.
	0.06	0.24	0.00	1.00		Unordered	Business and Financial Operations Occupations.
	0.04	0.19	0.00	1.00		Unordered	Community and Social Service Occupations.
	0.03	0.18	0.00	1.00		Unordered	Sales and Related Occupations.
	0.03	0.17	0.00	1.00		Unordered	Healthcare Support Occupations.
	0.05	0.22	0.00	1.00		Unordered	Production Occupations.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
	0.07	0.26	0.00	1.00		Unordered	Healthcare Practitioners and Technical Occupations.
	0.03	0.17	0.00	1.00		Unordered	Life, Physical, and Social Science Occupations.
	0.05	0.21	0.00	1.00		Unordered	Computer and Mathematical Occupations.
	0.02	0.15	0.00	1.00		Unordered	Installation, Maintenance, and Repair Occupations.
	0.07	0.25	0.00	1.00		Unordered	No Information.
	0.01	0.12	0.00	1.00		Unordered	Legal Occupations.
	0.01	0.10	0.00	1.00		Unordered	Protective Service Occupations.
	0.09	0.29	0.00	1.00		Unordered	Management Occupations.
	0.12	0.32	0.00	1.00		Unordered	Education, Training, and Library Occupations.
	0.02	0.15	0.00	1.00		Unordered	Construction and Extraction Occupations.
	0.05	0.22	0.00	1.00		Unordered	Arts, Design, Entertainment, and Media Occupations.
	0.02	0.13	0.00	1.00		Unordered	Personal Care and Service Occupations.
	0.01	0.08	0.00	1.00		Unordered	Building and Grounds Cleaning and Maintenance Occupations.
	0.03	0.17	0.00	1.00		Unordered	Transportation and Material Moving Occupations.
	0.07	0.26	0.00	1.00		Unordered	Architecture and Engineering Occupations.
	0.01	0.08	0.00	1.00		Unordered	Military Specific Occupations.

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
	0.01	0.11	0.00	1.00		Unordered	Food Preparation and Serving Related Occupations.
	0.00	0.05	0.00	1.00		Unordered	Farming, Fishing, and Forestry Occupations.
No. 361	0.19	0.39	0.00	1.00	Search criteria (partner): Children.	Unordered	No, please not.
	0.06	0.23	0.00	1.00		Unordered	No matter if the children do not live in the household.
	0.02	0.12	0.00	1.00		Unordered	Yes, in any case.
	0.74	0.44	0.00	1.00		Unordered	No matter.
No. 362	0.79	0.41	0.00	1.00	Search criteria (partner): Income.	Unordered	Same.
	0.21	0.41	0.00	1.00		Unordered	No matter.
	0.00	0.04	0.00	1.00		Unordered	No information.
No. 363	0.78	0.41	0.00	1.00	Search criteria (partner): Education.	Unordered	Same.
	0.19	0.39	0.00	1.00		Unordered	No matter.
	0.03	0.17	0.00	1.00		Unordered	Just my education level.
	0.00	0.04	0.00	1.00		Unordered	No information.
No. 364	0.22	0.41	0.00	1.00	Search criteria (partner): Smoking.	Unordered	No.
	0.22	0.41	0.00	1.00		Unordered	Occasionally.
	0.01	0.10	0.00	1.00		Unordered	Yes.
	0.55	0.50	0.00	1.00		Unordered	No information.
No. 365	0.98	0.12	0.00	1.00	Search criteria (partner): Minimum Age (in years).	Ordered	
No. 366	0.99	0.11	0.00	1.00	Search criteria (partner): Maximum Age (in years).	Ordered	

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 367	0.52	0.50	0.00	1.00	Search criteria (partner): Minimum Height (in cm).	Ordered	
No. 368	0.39	0.49	0.00	1.00	Search criteria (partner): Maximum Height (in cm).	Ordered	
No. 369	692.37	332.32	20.00	873.00	Search criteria (partner): Maximum distance (in km).	Ordered	
No. 370	0.46	0.50	0.00	1.00	Search criteria (partner): Distance search.	Unordered	No information.
	0.35	0.48	0.00	1.00		Unordered	Yes.
	0.18	0.39	0.00	1.00		Unordered	No.
No. 371	0.00	0.02	0.00	1.00	Search criteria (partner): Location.	Dummy	Country (code): Not defined
No. 372	0.01	0.12	0.00	1.00		Dummy	Country (code): AT; Region: Wien
No. 373	0.01	0.11	0.00	1.00		Dummy	Country (code): AT; Region: RDW
No. 374	0.01	0.11	0.00	1.00		Dummy	Country (code): AT; Region: Niederösterreich
No. 375	0.02	0.13	0.00	1.00		Dummy	Country (code): AT; Region: Vorarlberg
No. 376	0.02	0.13	0.00	1.00		Dummy	Country (code): AT; Region: Oberösterreich
No. 377	0.02	0.14	0.00	1.00		Dummy	Country (code): AT; Region: Salzburg
No. 378	0.02	0.14	0.00	1.00		Dummy	Country (code): AT; Region: Tirol
No. 379	0.01	0.11	0.00	1.00		Dummy	Country (code): AT; Region: Burgenland
No. 380	0.01	0.11	0.00	1.00		Dummy	Country (code): AT; Region: Steiermark

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 381	0.01	0.11	0.00	1.00		Dummy	Country (code): AT; Region: Kärnten
No. 382	0.00	0.06	0.00	1.00		Dummy	Country (code): BE; Region: Brussels Hoofdstedelijk Gewest
No. 383	0.00	0.05	0.00	1.00		Dummy	Country (code): BE; Region: Flandre occidentale
No. 384	0.00	0.05	0.00	1.00		Dummy	Country (code): BE; Region: Oost-Vlaanderen
No. 385	0.00	0.05	0.00	1.00		Dummy	Country (code): BE; Region: RDW
No. 386	0.00	0.05	0.00	1.00		Dummy	Country (code): BE; Region: Brabant wallon
No. 387	0.00	0.05	0.00	1.00		Dummy	Country (code): BE; Region: Brabant flamand
No. 388	0.00	0.06	0.00	1.00		Dummy	Country (code): BE; Region: Anvers
No. 389	0.00	0.06	0.00	1.00		Dummy	Country (code): BE; Region: Limburg
No. 390	0.00	0.06	0.00	1.00		Dummy	Country (code): BE; Region: Liège
No. 391	0.00	0.05	0.00	1.00		Dummy	Country (code): BE; Region: Namen
No. 392	0.00	0.05	0.00	1.00		Dummy	Country (code): BE; Region: Hainaut
No. 393	0.00	0.06	0.00	1.00		Dummy	Country (code): BE; Region: Luxemburg
No. 394	0.02	0.12	0.00	1.00		Dummy	Country (code): CH; Region: Aargau

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 395	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Graubünden
No. 396	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Jura
No. 397	0.01	0.12	0.00	1.00		Dummy	Country (code): CH; Region: Luzern
No. 398	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Neuchâtel
No. 399	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Nidwalden
No. 400	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Obwalden
No. 401	0.02	0.13	0.00	1.00		Dummy	Country (code): CH; Region: St.Gallen
No. 402	0.02	0.13	0.00	1.00		Dummy	Country (code): CH; Region: Schaffhausen
No. 403	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Solothurn
No. 404	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Schwyz
No. 405	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Appenzell Innerrhoden
No. 406	0.01	0.12	0.00	1.00		Dummy	Country (code): CH; Region: Thurgau
No. 407	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Ticino
No. 408	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Uri
No. 409	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Vaud

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 410	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Valais
No. 411	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Zug
No. 412	0.02	0.14	0.00	1.00		Dummy	Country (code): CH; Region: Zürich
No. 413	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: RDW
No. 414	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Appenzell Ausserrhoden
No. 415	0.02	0.13	0.00	1.00		Dummy	Country (code): CH; Region: Basel-Land
No. 416	0.02	0.13	0.00	1.00		Dummy	Country (code): CH; Region: Basel-Stadt
No. 417	0.01	0.12	0.00	1.00		Dummy	Country (code): CH; Region: Bern
No. 418	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Fribourg
No. 419	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Genève
No. 420	0.01	0.11	0.00	1.00		Dummy	Country (code): CH; Region: Glarus
No. 421	0.25	0.43	0.00	1.00		Dummy	Country (code): DE; Region: Baden-Württemberg
No. 422	0.28	0.45	0.00	1.00		Dummy	Country (code): DE; Region: Nordrhein-Westfalen
No. 423	0.18	0.38	0.00	1.00		Dummy	Country (code): DE; Region: Rheinland-Pfalz

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 424	0.09	0.29	0.00	1.00		Dummy	Country (code): DE; Region: Saarland
No. 425	0.12	0.32	0.00	1.00		Dummy	Country (code): DE; Region: Sachsen
No. 426	0.10	0.30	0.00	1.00		Dummy	Country (code): DE; Region: Sachsen-Anhalt
No. 427	0.12	0.33	0.00	1.00		Dummy	Country (code): DE; Region: Schleswig-Holstein
No. 428	0.11	0.31	0.00	1.00		Dummy	Country (code): DE; Region: Thüringen
No. 429	0.07	0.25	0.00	1.00		Dummy	Country (code): DE; Region: RDW
No. 430	0.26	0.44	0.00	1.00		Dummy	Country (code): DE; Region: Bayern
No. 431	0.16	0.37	0.00	1.00		Dummy	Country (code): DE; Region: Berlin
No. 432	0.12	0.32	0.00	1.00		Dummy	Country (code): DE; Region: Brandenburg
No. 433	0.11	0.31	0.00	1.00		Dummy	Country (code): DE; Region: Bremen
No. 434	0.16	0.37	0.00	1.00		Dummy	Country (code): DE; Region: Hamburg
No. 435	0.21	0.41	0.00	1.00		Dummy	Country (code): DE; Region: Hessen
No. 436	0.10	0.30	0.00	1.00		Dummy	Country (code): DE; Region: Mecklenburg-Vorpommern
No. 437	0.19	0.39	0.00	1.00		Dummy	Country (code): DE; Region: Niedersachsen

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 438	0.00	0.07	0.00	1.00		Dummy	Country (code): DK; Region: Bornholm
No. 439	0.01	0.07	0.00	1.00		Dummy	Country (code): DK; Region: Vestjylland og det sydlige Østjylland
No. 440	0.00	0.07	0.00	1.00		Dummy	Country (code): DK; Region: Vestsjælland, Lolland-Falster og Møn
No. 441	0.00	0.07	0.00	1.00		Dummy	Country (code): DK; Region: RDW
No. 442	0.00	0.07	0.00	1.00		Dummy	Country (code): DK; Region: Færøerne
No. 443	0.00	0.07	0.00	1.00		Dummy	Country (code): DK; Region: Fyn og øerne
No. 444	0.00	0.07	0.00	1.00		Dummy	Country (code): DK; Region: Grønland
No. 445	0.00	0.07	0.00	1.00		Dummy	Country (code): DK; Region: København, Frederiksberg og omegn
No. 446	0.00	0.07	0.00	1.00		Dummy	Country (code): DK; Region: Nordjylland
No. 447	0.00	0.07	0.00	1.00		Dummy	Country (code): DK; Region: Nordsjælland
No. 448	0.00	0.07	0.00	1.00		Dummy	Country (code): DK; Region: Østjylland
No. 449	0.01	0.07	0.00	1.00		Dummy	Country (code): DK; Region: Sønderjylland samt dele af Sydjylland og dele af Vestjylland

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 450	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Andalucía
No. 451	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Navarra
No. 452	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Castilla-La Mancha
No. 453	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: País Vasco
No. 454	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Cataluña
No. 455	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Valencia
No. 456	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Extremadura
No. 457	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Ciudad Autónoma de Melilla
No. 458	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Galicia
No. 459	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Ciudad Autónoma de Ceuta
No. 460	0.01	0.08	0.00	1.00		Dummy	Country (code): ES; Region: Islas Baleares
No. 461	0.01	0.08	0.00	1.00		Dummy	Country (code): ES; Region: Islas Canarias
No. 462	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: RDW
No. 463	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Aragón

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 464	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: La Rioja
No. 465	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Asturias
No. 466	0.01	0.08	0.00	1.00		Dummy	Country (code): ES; Region: Madrid
No. 467	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Cantabria
No. 468	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Murcia
No. 469	0.01	0.07	0.00	1.00		Dummy	Country (code): ES; Region: Castilla y León
No. 470	0.01	0.09	0.00	1.00		Dummy	Country (code): FR; Region: Alsace
No. 471	0.00	0.06	0.00	1.00		Dummy	Country (code): FR; Region: DOM-TOM
No. 472	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Franche-Comté
No. 473	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Haute-Normandie
No. 474	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Ile-de-France
No. 475	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Languedoc-Roussillon
No. 476	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Limousin
No. 477	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Lorraine

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 478	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Midi-Pyrénées
No. 479	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Monaco
No. 480	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Nord-Pas-de-Calais
No. 481	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Aquitaine
No. 482	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Pays-de-la-Loire
No. 483	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Picardie
No. 484	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Poitou-Charentes
No. 485	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Provence-Alpes-Côte-d'Azur
No. 486	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Rhône-Alpes
No. 487	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: RDW
No. 488	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Auvergne
No. 489	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Basse-Normandie
No. 490	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Bourgogne
No. 491	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Bretagne

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 492	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Centre
No. 493	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Champagne-Ardenne
No. 494	0.00	0.07	0.00	1.00		Dummy	Country (code): FR; Region: Corse
No. 495	0.01	0.08	0.00	1.00		Dummy	Country (code): GB; Region: East Midlands
No. 496	0.01	0.08	0.00	1.00		Dummy	Country (code): GB; Region: Wales
No. 497	0.01	0.08	0.00	1.00		Dummy	Country (code): GB; Region: North West
No. 498	0.01	0.08	0.00	1.00		Dummy	Country (code): GB; Region: West Midlands
No. 499	0.01	0.08	0.00	1.00		Dummy	Country (code): GB; Region: RDW
No. 500	0.01	0.08	0.00	1.00		Dummy	Country (code): GB; Region: Northern Ireland
No. 501	0.01	0.08	0.00	1.00		Dummy	Country (code): GB; Region: East Anglia
No. 502	0.01	0.09	0.00	1.00		Dummy	Country (code): GB; Region: Scotland
No. 503	0.01	0.09	0.00	1.00		Dummy	Country (code): GB; Region: Greater London
No. 504	0.01	0.09	0.00	1.00		Dummy	Country (code): GB; Region: South East
No. 505	0.01	0.08	0.00	1.00		Dummy	Country (code): GB; Region: Yorkshire The Humber

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 506	0.01	0.09	0.00	1.00		Dummy	Country (code): GB; Region: South West
No. 507	0.01	0.08	0.00	1.00		Dummy	Country (code): GB; Region: North East
No. 508	0.00	0.06	0.00	1.00		Dummy	Country (code): IE; Region: Leinster
No. 509	0.00	0.06	0.00	1.00		Dummy	Country (code): IE; Region: Munster
No. 510	0.00	0.06	0.00	1.00		Dummy	Country (code): IE; Region: Connacht
No. 511	0.00	0.06	0.00	1.00		Dummy	Country (code): IE; Region: Ulster
No. 512	0.00	0.06	0.00	1.00		Dummy	Country (code): IE; Region: RDW
No. 513	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Abruzzo
No. 514	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Marche
No. 515	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Molise
No. 516	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Piemonte
No. 517	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Puglia
No. 518	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Sardegna
No. 519	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Sicilia

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 520	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Toscana
No. 521	0.01	0.08	0.00	1.00		Dummy	Country (code): IT; Region: Trentino-Alto Adige
No. 522	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Umbria
No. 523	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Valle d'Aosta
No. 524	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Basilicata
No. 525	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Veneto
No. 526	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: RDW
No. 527	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Calabria
No. 528	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Campania
No. 529	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Emilia-Romagna
No. 530	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Friuli-Venezia Giulia
No. 531	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Lazio
No. 532	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Liguria
No. 533	0.01	0.07	0.00	1.00		Dummy	Country (code): IT; Region: Lombardia

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 534	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Aguascalientes
No. 535	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Durango
No. 536	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Guanajuato
No. 537	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Guerrero
No. 538	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Hidalgo
No. 539	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Jalisco
No. 540	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: México
No. 541	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Michoacán de Ocampo
No. 542	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Morelos
No. 543	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Nayarit
No. 544	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Nuevo León
No. 545	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Baja California
No. 546	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Oaxaca
No. 547	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Puebla

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 548	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Qerétaro de Arteaga
No. 549	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Quintana Roo
No. 550	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: San Luís Potosí
No. 551	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Sinaloa
No. 552	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Sonora
No. 553	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Tabasco
No. 554	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Tamaulipas
No. 555	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Tlaxcala
No. 556	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Baja California Sur
No. 557	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Veracruz
No. 558	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Yucatán
No. 559	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Zacatecas
No. 560	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: RDW
No. 561	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Campeche

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 562	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Coahuila de Zaragoza
No. 563	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Colima
No. 564	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Chiapas
No. 565	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Chihuahua
No. 566	0.00	0.03	0.00	1.00		Dummy	Country (code): MX; Region: Distrito Federal
No. 567	0.01	0.08	0.00	1.00		Dummy	Country (code): NL; Region: Drenthe
No. 568	0.01	0.09	0.00	1.00		Dummy	Country (code): NL; Region: Utrecht
No. 569	0.01	0.08	0.00	1.00		Dummy	Country (code): NL; Region: Zeeland
No. 570	0.01	0.09	0.00	1.00		Dummy	Country (code): NL; Region: Zuid-Holland
No. 571	0.01	0.08	0.00	1.00		Dummy	Country (code): NL; Region: RDW
No. 572	0.01	0.08	0.00	1.00		Dummy	Country (code): NL; Region: Flevoland
No. 573	0.01	0.08	0.00	1.00		Dummy	Country (code): NL; Region: Friesland
No. 574	0.01	0.09	0.00	1.00		Dummy	Country (code): NL; Region: Gelderland
No. 575	0.01	0.08	0.00	1.00		Dummy	Country (code): NL; Region: Groningen

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 576	0.01	0.09	0.00	1.00		Dummy	Country (code): NL; Region: Limburg
No. 577	0.01	0.08	0.00	1.00		Dummy	Country (code): NL; Region: Overijssel
No. 578	0.01	0.08	0.00	1.00		Dummy	Country (code): NL; Region: Noord-Brabant
No. 579	0.01	0.08	0.00	1.00		Dummy	Country (code): NL; Region: Noord-Holland
No. 580	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Akershus
No. 581	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Oppland
No. 582	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Oslo
No. 583	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Østfold
No. 584	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Rogaland
No. 585	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Sogn og fjordane
No. 586	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Sør-Trøndelag
No. 587	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Svalbard
No. 588	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Telemark
No. 589	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Troms

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 590	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Vest-Agder
No. 591	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Aust-Agder
No. 592	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Vestfold
No. 593	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: RDW
No. 594	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Buskerud
No. 595	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Finnmark
No. 596	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Hedmark
No. 597	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Hordaland
No. 598	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Møre og Romsdal
No. 599	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Nordland
No. 600	0.00	0.06	0.00	1.00		Dummy	Country (code): NO; Region: Nord-Trøndelag
No. 601	0.01	0.08	0.00	1.00		Dummy	Country (code): SE; Region: Stockholm
No. 602	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Blekinge
No. 603	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Skåne

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 604	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Halland
No. 605	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Västra Götaland
No. 606	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Värmland
No. 607	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Örebro
No. 608	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Västmanland
No. 609	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Dalarna
No. 610	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Gävleborg
No. 611	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Västernorrland
No. 612	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Jämtland
No. 613	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Västerbotten
No. 614	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Norrbotten
No. 615	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: RDW
No. 616	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Uppsala
No. 617	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Södermanland

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
No. 618	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Östergötland
No. 619	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Jönköping
No. 620	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Kronoberg
No. 621	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Kalmar
No. 622	0.01	0.07	0.00	1.00		Dummy	Country (code): SE; Region: Gotland
No. 623	2.11	1.05	0.00	3.00	How often do you practice sport?	Ordered	Never.
						Ordered	Rarely.
						Ordered	Monthly.
						Ordered	Weekly.
No. 624	0.01	0.11	0.00	1.00	ZIP code area (first two digits of five-digit German ZIP code; note that German ZIP code areas do not necessarily correspond to administrative units).	Unordered	ZIP: 01
	0.00	0.05	0.00	1.00		Unordered	ZIP: 02
	0.00	0.05	0.00	1.00		Unordered	ZIP: 03
	0.02	0.12	0.00	1.00		Unordered	ZIP: 04
	0.01	0.11	0.00	1.00		Unordered	ZIP: 06
	0.01	0.07	0.00	1.00		Unordered	ZIP: 07
	0.00	0.07	0.00	1.00		Unordered	ZIP: 08
	0.01	0.09	0.00	1.00		Unordered	ZIP: 09
	0.02	0.15	0.00	1.00		Unordered	ZIP: 10
	0.02	0.14	0.00	1.00		Unordered	ZIP: 12

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
	0.01	0.12	0.00	1.00		Unordered	ZIP: 13
	0.02	0.12	0.00	1.00		Unordered	ZIP: 14
	0.01	0.08	0.00	1.00		Unordered	ZIP: 15
	0.01	0.08	0.00	1.00		Unordered	ZIP: 16
	0.00	0.07	0.00	1.00		Unordered	ZIP: 17
	0.01	0.09	0.00	1.00		Unordered	ZIP: 18
	0.00	0.07	0.00	1.00		Unordered	ZIP: 19
	0.01	0.08	0.00	1.00		Unordered	ZIP: 20
	0.01	0.12	0.00	1.00		Unordered	ZIP: 21
	0.03	0.17	0.00	1.00		Unordered	ZIP: 22
	0.01	0.10	0.00	1.00		Unordered	ZIP: 23
	0.01	0.12	0.00	1.00		Unordered	ZIP: 24
	0.01	0.09	0.00	1.00		Unordered	ZIP: 25
	0.01	0.10	0.00	1.00		Unordered	ZIP: 26
	0.01	0.08	0.00	1.00		Unordered	ZIP: 27
	0.01	0.09	0.00	1.00		Unordered	ZIP: 28
	0.01	0.07	0.00	1.00		Unordered	ZIP: 29
	0.01	0.12	0.00	1.00		Unordered	ZIP: 30
	0.01	0.10	0.00	1.00		Unordered	ZIP: 31
	0.01	0.09	0.00	1.00		Unordered	ZIP: 32
	0.01	0.09	0.00	1.00		Unordered	ZIP: 33
	0.01	0.09	0.00	1.00		Unordered	ZIP: 34
	0.01	0.10	0.00	1.00		Unordered	ZIP: 35
	0.01	0.07	0.00	1.00		Unordered	ZIP: 36
	0.01	0.08	0.00	1.00		Unordered	ZIP: 37
	0.01	0.12	0.00	1.00		Unordered	ZIP: 38
	0.01	0.09	0.00	1.00		Unordered	ZIP: 39
	0.02	0.13	0.00	1.00		Unordered	ZIP: 40

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
	0.01	0.10	0.00	1.00		Unordered	ZIP: 41
	0.01	0.09	0.00	1.00		Unordered	ZIP: 42
	0.01	0.11	0.00	1.00		Unordered	ZIP: 44
	0.01	0.12	0.00	1.00		Unordered	ZIP: 45
	0.01	0.10	0.00	1.00		Unordered	ZIP: 46
	0.01	0.12	0.00	1.00		Unordered	ZIP: 47
	0.01	0.11	0.00	1.00		Unordered	ZIP: 48
	0.01	0.11	0.00	1.00		Unordered	ZIP: 49
	0.02	0.14	0.00	1.00		Unordered	ZIP: 50
	0.01	0.10	0.00	1.00		Unordered	ZIP: 51
	0.01	0.10	0.00	1.00		Unordered	ZIP: 52
	0.02	0.13	0.00	1.00		Unordered	ZIP: 53
	0.01	0.08	0.00	1.00		Unordered	ZIP: 54
	0.01	0.11	0.00	1.00		Unordered	ZIP: 55
	0.01	0.09	0.00	1.00		Unordered	ZIP: 56
	0.01	0.08	0.00	1.00		Unordered	ZIP: 57
	0.01	0.09	0.00	1.00		Unordered	ZIP: 58
	0.01	0.10	0.00	1.00		Unordered	ZIP: 59
	0.01	0.12	0.00	1.00		Unordered	ZIP: 60
	0.01	0.09	0.00	1.00		Unordered	ZIP: 61
	0.02	0.12	0.00	1.00		Unordered	ZIP: 63
	0.01	0.10	0.00	1.00		Unordered	ZIP: 64
	0.02	0.13	0.00	1.00		Unordered	ZIP: 65
	0.01	0.12	0.00	1.00		Unordered	ZIP: 66
	0.01	0.10	0.00	1.00		Unordered	ZIP: 67
	0.01	0.09	0.00	1.00		Unordered	ZIP: 68
	0.01	0.09	0.00	1.00		Unordered	ZIP: 69
	0.01	0.12	0.00	1.00		Unordered	ZIP: 70

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
	0.02	0.12	0.00	1.00		Unordered	ZIP: 71
	0.01	0.12	0.00	1.00		Unordered	ZIP: 72
	0.01	0.11	0.00	1.00		Unordered	ZIP: 73
	0.01	0.11	0.00	1.00		Unordered	ZIP: 74
	0.01	0.08	0.00	1.00		Unordered	ZIP: 75
	0.02	0.13	0.00	1.00		Unordered	ZIP: 76
	0.00	0.07	0.00	1.00		Unordered	ZIP: 77
	0.01	0.09	0.00	1.00		Unordered	ZIP: 78
	0.01	0.11	0.00	1.00		Unordered	ZIP: 79
	0.02	0.13	0.00	1.00		Unordered	ZIP: 80
	0.01	0.12	0.00	1.00		Unordered	ZIP: 81
	0.01	0.11	0.00	1.00		Unordered	ZIP: 82
	0.01	0.10	0.00	1.00		Unordered	ZIP: 83
	0.01	0.09	0.00	1.00		Unordered	ZIP: 84
	0.02	0.13	0.00	1.00		Unordered	ZIP: 85
	0.01	0.12	0.00	1.00		Unordered	ZIP: 86
	0.01	0.08	0.00	1.00		Unordered	ZIP: 87
	0.01	0.09	0.00	1.00		Unordered	ZIP: 88
	0.01	0.09	0.00	1.00		Unordered	ZIP: 89
	0.01	0.11	0.00	1.00		Unordered	ZIP: 90
	0.01	0.11	0.00	1.00		Unordered	ZIP: 91
	0.00	0.06	0.00	1.00		Unordered	ZIP: 92
	0.01	0.08	0.00	1.00		Unordered	ZIP: 93
	0.01	0.07	0.00	1.00		Unordered	ZIP: 94
	0.00	0.06	0.00	1.00		Unordered	ZIP: 95
	0.00	0.07	0.00	1.00		Unordered	ZIP: 96
	0.01	0.10	0.00	1.00		Unordered	ZIP: 97
	0.00	0.06	0.00	1.00		Unordered	ZIP: 98

continued on next page

Code	Mean	SD	Min	Max	Question	Coding	Answer
	0.01	0.10	0.00	1.00		Unordered	ZIP: 99

Note: First column lists a unique identifier for each variable. Second, third, fourth and fifth column report the corresponding mean, standard deviation, minimum and maximum values for the variable, respectively. Sixth column contains the specific questions from the registration questionnaire. Seventh column indicates the variable encoding: *dummy* stands for a binary variable equal to 1 if the respective answer has been chosen (mutually inclusive); *ordered* stands for a numeric value with a clear inherent ordering (both continuous and categorical), directly filled by the user (mutually exclusive); *unordered* stands for a text value without an ordered structure (categorical), directly filled by the user (mutually exclusive). Last column lists the corresponding answers available in the registration questionnaire.

Curriculum Vitae

Education

2017 - 2021	Doctoral Degree in Economics and Finance (PhD), University of St.Gallen Dissertation: Essays in Predictive and Causal Machine Learning Supervisor: Prof. Dr. Michael Lechner
2014 - 2016	Master Degree in Economics (MSc), Vienna University of Economics and Business
2016	Exchange Semester, Free University of Berlin
2011 - 2014	Bachelor Degree in Management (Bc), Comenius University in Bratislava
2013	Exchange Semester, Upper Austria University of Applied Sciences

Experience

2021 - present	Postdoctoral Researcher, Swiss Federal Institute of Technology in Lausanne
2017 - 2021	Research Assistant, Swiss Institute for Empirical Economic Research
2016 - 2017	Research Intern, ifo Institute for Economic Research
2016	Summer Intern, Erste Asset Management
2015 - 2016	Teaching Assistant, Vienna University of Economics and Business

St.Gallen, November 9, 2021

Gabriel Okasa