machine learning & knowledge extraction

MDPI

Article

Interaction Difference Hypothesis Test for Prediction Models

Thomas Welchowski [1,*] and Dominic Edelmann [2]

[1] Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Venusberg-Campus 1, 53127 Bonn, North Rhine-Westphalia, Germany
[2] Division of Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Baden-Württemberg, Germany; dominic.edelmann@dkfz-heidelberg.de
* Correspondence: thomas.welchowski@gmx.net

Abstract: Machine learning research focuses on the improvement of prediction performance. Progress was made with black-box models that flexibly adapt to the given data. However, due to their increased complexity, black-box models are more difficult to interpret. To address this issue, techniques for interpretable machine learning have been developed, yet there is still a lack of methods to reliably identify interaction effects between predictors under uncertainty. In this work, we present a model-agnostic hypothesis test for the identification of interaction effects in black-box machine learning models. The test statistic is based on the difference between the variance of the estimated prediction function and a version of the estimated prediction function without interaction effects derived via partial dependence functions. The properties of the proposed hypothesis test were explored in simulations of linear and nonlinear models. The proposed hypothesis test can be applied to any black-box prediction model, and the null hypothesis of the test can be flexibly specified according to the research question of interest. Furthermore, the test is computationally fast to apply, as the null distribution does not require the resampling or refitting of black-box prediction models.

Keywords: prediction models; interpretable machine learning; model-agnostic; hypothesis tests; interaction effects

Citation: Welchowski, T.; Edelmann, D. Interaction Difference Hypothesis Test for Prediction Models. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1298–1322. https://doi.org/10.3390/make6020061

Academic Editor: Ahmad Taher Azar

Received: 11 May 2024
Revised: 26 May 2024
Accepted: 5 June 2024
Published: 14 June 2024

1. Background

In the context of machine learning, one of the main goals is to estimate and tune prediction models in order to optimize predefined performance criteria [1]. In the ongoing academic debate, [2] argued that the attribution of causal factors may require a larger sample size than estimating a prediction model. This is in line with [3], who showed that causality can be linked with prediction robustness. The research area of interpretable machine learning (IML) tries to bridge the gap between prediction and classical statistical inference by making complex black-box predictions more understandable [4]. A black-box model is characterized by an input–output relationship between covariates and a response [5]. In this approach, the internal structure of the box is not explicitly modeled and is regarded as unknown. The Rashomon effect [6] originates from a Japanese movie from 1950. The main plot is about a crime happening in the 12th century, which is shown from the perspectives of multiple people. Differences in those experiences show that it is hard to uncover later what really happened because, for a given set of facts, there are a multitude of compatible stories. Analogously, in machine learning, there are many different models that explain the observed data equally well. The problem of empirical induction has a long history in the philosophy of science. For example, the skepticism of David Hume, dating back to the 18th century [7], or Duhem's theses stated that the falsifiability of a single hypothesis is inconclusive [8]. This work will not address this philosophical problem, but it takes instead a pragmatic approach [9]. It is assumed that the primary goal is to optimize prediction performance [10] in a given context. There is some evidence of a trade-off between prediction performance and interpretability in the

literature [11–14]. However, prediction can benefit from interpretability as well because a deeper qualitative understanding of why a model produces a given output and not another can help generate more robust out-of-sample predictions. Therefore, it is recommended to use interpretability approaches that do not harm prediction performance but help incorporate human considerations into explainable artificial intelligence [15].

IML allows a researcher to benefit from advances in machine learning research and still explore the properties of the model afterwards to increase the interpretability of the model. Example applications include designing regulatorily compliant, fair [16], transparent, and trustworthy prediction models [17]. Another area of IML focuses on the interpretation of the effects of covariates on prediction [18–20]. Here, the focus is on global model interpretability, which means that the prediction function over the whole covariate distribution is the focus of interest instead of explaining single local predictions for specific covariate values [21].

The following sections, Sections 1.1–1.3, introduce the background knowledge required to understand the new proposed interaction difference hypothesis test for prediction models that is defined in Section 2. Firstly, a measure of how a prediction function changes, on average, for different values of a given set of covariates is introduced in Section 1.1. This measure is an essential component used in the definition of interaction effects. Secondly, Section 1.2 describes a general definition of interaction effects for black-box models, which is based on an additive decomposition of the predictions. The decomposition is illustrated using a linear regression example. Thirdly, Section 1.3 provides an overview of existing approaches to quantify interaction effects. Finally, the last introductory section, Section 1.4, describes the null hypothesis of the interaction test and the disadvantages of the previously described, existing approaches that will be addressed in this work.

1.1. Partial Dependence Functions

A global summary of the impact of one covariate on the predictions is the partial dependence (PD) plot [22]. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the observed matrix of p covariates with n independent observations of the multivariate random vector $\boldsymbol{x} \in \mathbb{R}^p$ and $\hat{f}(\boldsymbol{x})$ be the estimated predictions of a statistical model of the prediction function $f(\boldsymbol{x})$ on the population level. $f(\boldsymbol{x})$ does not necessarily equal the covariate–response relationship in the data-generating process. It is assumed that $\hat{f}(\boldsymbol{x})$ was estimated, as well as tuned, prior to IML analysis with respect to prediction performance with the test data. Define $S = \{1, \ldots, p\}$ to be the set of all indices of covariates, and the set $s \subset S$ corresponds to indices of chosen covariates of interest. The term $\mathrm{E}_{\boldsymbol{x}_{S \setminus s}}(\boldsymbol{x}_s)$ is defined as the expectation over the marginal (joint) distribution of all variables not in set s (denoted as $\boldsymbol{x}_{S \setminus s}$) for fixed values $\boldsymbol{x}_s$ of the variables in set s (for a comparison, see [22] Section 8.1). Note that multiple column indices are denoted using set brackets in the subscript; for example, $s = \{1, 2\}$ yields $\boldsymbol{x}_{\{1,2\}}$, and empty subscripts describe all available indices (for example, the second column with all observations, $\boldsymbol{X}_{,2}$). The PD function is given via

$$\mathrm{PD}(\boldsymbol{x}_s) = \mathrm{E}_{\boldsymbol{x}_{S \setminus s}}\Big(f\Big(\boldsymbol{x}_s, \boldsymbol{x}_{S \setminus s}\Big)\Big) \text{and estimated by} \tag{1}$$

$$\widehat{\mathrm{PD}}(\boldsymbol{x}_s) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}\Big(\boldsymbol{x}_s, \boldsymbol{X}_{i,S \setminus s}\Big). \tag{2}$$

For example, in the case of $p = 5, s = \{1, 2, 3\}$, the function $\mathrm{PD}\Big(\boldsymbol{x}_{\{1,2,3\}}\Big)$ is the expected value of the predictions with respect to the covariate distribution $\boldsymbol{x}_{\{4,5\}}$, given the observed covariate values $\boldsymbol{X}_{,\{1,2,3\}}$. If $s = \varnothing$; then, $\mathrm{PD}\big(\boldsymbol{x}_{\varnothing}\big)$ corresponds to the expected marginal prediction over all covariates, $\boldsymbol{x}_{\{1,2,3,4,5\}}$. Note that, in the case of $s = S$, the PD function equals the original model predictions, $\mathrm{PD}(\boldsymbol{x}_s) = f(\boldsymbol{x})$, and the function argument $\boldsymbol{x}_s$ values do not necessarily need to correspond to training data.

1.2. Interactions in Black-Box Models

First, we briefly recap what interaction effects are in a linear model context [23]. Consider the simple case of a linear model prediction function with two independent covariates, $\boldsymbol{x}_1, \boldsymbol{x}_2$, main and interaction effects:

$$f(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{x}_1\boldsymbol{x}_2\beta_{1,2}. \tag{3}$$

The main effects, β_1 and β_2, represent how the prediction function changes linearly, given $\beta_{1,2} = 0$, if the covariate of interest is increased by one unit. In contrast, the interaction effect $\beta_{1,2} \neq 0$ of covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ contributes additional flexibility that goes beyond the main effects and the global intercept β_0. Let the difference term be $d_{\text{linear}} = f(\boldsymbol{x}) - \boldsymbol{x}_1\beta_1 - \boldsymbol{x}_2\beta_2 - \beta_0 = \boldsymbol{x}_1\boldsymbol{x}_2\beta_{1,2}$. If the interaction effect is $\beta_{1,2} \neq 0$, then the variance of the difference term $\text{Var}_{\boldsymbol{x}}(d_{\text{linear}})$ is greater than zero. Similarly, if $\beta_1 \neq 0$, then $\text{Var}_{\boldsymbol{x}}(\boldsymbol{x}_1\beta_1) > 0$ follows.

One advantage of black-box models (for example, neural networks) is their capacity to fit higher-order interaction effects in a data-driven way without the need to explicitly prespecify them. Knowledge of the presence of such interaction effects would increase the scientific understanding of a given phenomenon, and the absence of interaction effects could be used to simplify black-box prediction models with little degradation in performance. In this context, interaction effects can be defined within the functional ANOVA decomposition framework [24]. The prediction function $f(\boldsymbol{x})$ is decomposed into a sum of additive orthogonal terms, $\tilde{f}(\boldsymbol{x}_{\tilde{s}})$, of sets $\tilde{s}$. Each term recursively subtracts all respective previously derived lower-order terms within set $\tilde{s}$. In this work, we use PD functions to define the functional ANOVA terms $\tilde{f}(\boldsymbol{x}_{\tilde{s}})$. In the simple linear regression example in Equation (3), the first ANOVA term would correspond to the expected value of the prediction

$$\begin{aligned} \tilde{f}(\boldsymbol{x}_{\varnothing}) &= \text{PD}(\boldsymbol{x}_{\varnothing}) && (4)\\ &= \beta_0 + \mu_1\beta_1 + \mu_2\beta_2 + \mu_{1,2}\beta_{1,2} && (5) \end{aligned}$$

with μ_j being the expected value of the covariates or, in the case of $\mu_{1,2}$, the expected value of the product of the covariates. By definition, the functional ANOVA main effects, $\tilde{f}(\boldsymbol{x}_1), \tilde{f}(\boldsymbol{x}_2)$, consist of the PD functions of $\boldsymbol{x}_1, \boldsymbol{x}_2$, minus the sum of all possible respective lower-order effects. In the case of one covariate, only the empty set needs to be subtracted:

$$\begin{aligned} \tilde{f}(\boldsymbol{x}_1) &= \text{PD}(\boldsymbol{x}_1) - \tilde{f}(\boldsymbol{x}_{\varnothing}) \text{ and} && (6)\\ \tilde{f}(\boldsymbol{x}_2) &= \text{PD}(\boldsymbol{x}_2) - \tilde{f}(\boldsymbol{x}_{\varnothing}). && (7) \end{aligned}$$

In the concrete scenario, the functional ANOVA main effects are given by

$$\begin{aligned} \text{PD}(\boldsymbol{x}_1) &= \beta_0 + \boldsymbol{x}_1\beta_1 + \mu_2\beta_2 + \boldsymbol{x}_1\mu_2\beta_{1,2} && (8)\\ \Rightarrow \tilde{f}(\boldsymbol{x}_1) &= \boldsymbol{x}_1(\beta_1 + \mu_2\beta_{1,2}) - \mu_1\beta_1 - \mu_{1,2}\beta_{1,2} \text{ and} && (9)\\ \text{PD}(\boldsymbol{x}_2) &= \beta_0 + \mu_1\beta_1 + \boldsymbol{x}_2\beta_2 + \mu_1\boldsymbol{x}_2\beta_{1,2} && (10)\\ \Rightarrow \tilde{f}(\boldsymbol{x}_2) &= \boldsymbol{x}_2(\beta_2 + \mu_1\beta_{1,2}) - \mu_2\beta_2 - \mu_{1,2}\beta_{1,2}. && (11) \end{aligned}$$

If $\beta_{1,2} = 0$, then $\tilde{f}(\boldsymbol{x}_1)$ and $\tilde{f}(\boldsymbol{x}_2)$ correspond analogously to centered main effects in the linear model. In the case of the second-order functional ANOVA term $\tilde{f}\left(\boldsymbol{x}_{\{1,2\}}\right)$, two first-order terms that are contained in set $\{1, 2\}$ need to be substracted, as well as the empty set, to ensure the orthogonality of second- and first-order ANOVA terms. The second-order interaction effect in terms of the functional ANOVA is, then,

$$\begin{aligned} \tilde{f}\left(\boldsymbol{x}_{\{1,2\}}\right) &= \text{PD}\left(\boldsymbol{x}_{\{1,2\}}\right) - \tilde{f}(\boldsymbol{x}_1) - \tilde{f}(\boldsymbol{x}_2) - \tilde{f}(\boldsymbol{x}_{\varnothing}), && (12)\\ \text{PD}\left(\boldsymbol{x}_{\{1,2\}}\right) &= f(\boldsymbol{x}) \text{ and} && (13)\\ \Rightarrow \tilde{f}\left(\boldsymbol{x}_{\{1,2\}}\right) &= \boldsymbol{x}_1\boldsymbol{x}_2\beta_{1,2} - \mu_1\beta_{1,2} - \mu_2\beta_{1,2} + \mu_{1,2}\beta_{1,2}. && (14) \end{aligned}$$

If $\beta_{1,2} \neq 0$, then $\text{Var}_{\boldsymbol{x}}\left[\tilde{f}\left(\boldsymbol{x}_{\{1,2\}}\right)\right] > 0$, similar to the linear model context. If $\beta_{1,2} = 0$, the functional ANOVA main effects have the property $\text{Var}_{\boldsymbol{x}}\left[\tilde{f}\left(\boldsymbol{x}_{\{j\}}\right)\right] > 0$ for $j = \{1,2\}$, which is also shared within the linear model. Note that, if $\beta_{1,2} \neq 0$, then the functional ANOVA main effects include part of the linear model interaction effect in term $\boldsymbol{x}_j \mu_{\{1,2\}\setminus j} \beta_{1,2} : j \in \{1,2\}$. Therefore, we analogously define $\text{Var}_{\boldsymbol{x}}\left[\tilde{f}(\boldsymbol{x}_{s^\star})\right] > 0 : s^\star \subset S \wedge |s^\star| \geqslant 2$ as interaction effects of at least order $|s^\star|$ of the covariates in set $s^\star$ of black-box models.

One disadvantage of functional ANOVA is that those derived terms are estimators based on data, and this uncertainty has to be taken into account when conducting inference. A distribution of the functional ANOVA terms under the null hypothesis of no interaction is not available. A second disadvantage is that the complexity to compute the decomposition grows exponentially with the number of covariates to 2^p possible elements. Furthermore, this concept works best with independent covariates, which is unrealistic in practice. A generalized functional ANOVA [25] includes covariate dependencies but requires solving a system of equations that is even more computationally demanding than the functional ANOVA decomposition. This limits the practical application to lower-order interaction terms [18].

1.3. Interaction Measures Based on PD Functions

Based on the concept of PD functions, [22] derived the H^2 statistic to analyze interaction effects. The H^2 statistic measures the variance in the differences between a prediction function and its restricted form under a given null hypothesis normalized by the variance of the prediction function to detect specific interaction effects. Note that the concrete form of H^2 depends on the null hypothesis. For example, to test whether covariates in set s interact with any other covariates of set S, the statistic H_s^2 is defined as

$$\text{H}_s^2 = \frac{\text{Var}_{\boldsymbol{x}_s}\left(f(\boldsymbol{x}) - \text{PD}\left(\boldsymbol{x}_{S\setminus s}\right) - \sum_{j \in s} \text{PD}(\boldsymbol{x}_j)\right)}{\text{Var}_{\boldsymbol{x}}(\text{PD}(\boldsymbol{x}))} \text{ and estimated by} \tag{15}$$

$$\hat{\text{H}}_s^2 = \frac{\sum_{i=1}^n \left[\hat{f}(\boldsymbol{X}_{i,S}) - \widehat{\text{PD}}\left(\boldsymbol{X}_{i,S\setminus s}\right) - \sum_{j \in s} \widehat{\text{PD}}(\boldsymbol{X}_{i,j})\right]^2}{\sum_{i=1}^n \left[\widehat{\text{PD}}(\boldsymbol{X}_{i,S})\right]^2} \tag{16}$$

assuming centered PD functions. Equation (15) is an extension of Equation (45) in [22] to multiple covariates. It was derived by repeatedly applying Equation (42) in [22] for each element of s. Note that the difference of Equation (15) to Equations (43) and (46) by [22] is the hypothesis that is being tested. In the latter case, the hypothesis is to test for the presence of the specific three-way interaction between covariates $\boldsymbol{x}_j, \boldsymbol{x}_k, \boldsymbol{x}_l$ that allows any two-way interaction to be present in the prediction model. This work focuses on testing any interaction effects of covariates in the prediction model specified in set s. In Section 1.4, the hypothesis of this work is described in more detail.

The statistic (15) was developed in the context of rule ensembles, and the flexible specification of interaction effects can be evaluated. The derived hypothesis test is a parametric bootstrap approach that simulates artificial data sets with a prediction model restricted to the null hypothesis of no interaction effects (Section 8.3 in [22]). Rule ensembles can be restricted to not include interaction effects by limiting the tree depth to one, but it does not work for different types of prediction models. Furthermore, the approach is computationally expensive due to the need to refit prediction models to artificial data sets, and the accuracy of the simulated p-value depends on the number of bootstrap replicates. The computational costs rise further due to the tuning process of hyperparameters, which are usually based on resampling methods like k-fold cross-validation. For an overview of recent developments in the field of hyperparameter optimization, we refer to [26].

Another measure to quantify interactions was developed by [27] that quantifies interactions between two covariates, $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$, by estimating the standard deviation of the PD function of the $\boldsymbol{x}_j$ conditional on values of $\boldsymbol{x}_k$. This approach is restricted to two-way inter-

actions. Generalizing this to scale higher-order interaction effects than two would reduce the number of available samples for estimating the standard deviation, and the number of possible combinations of the conditional covariates would increase exponentially. Note that there are also graphical tools to assess interaction effects, for example, [28,29]; however, these can only be meaningfully applied to illustrate lower-dimensional covariate interaction effects than three, and they do not quantify their method uncertainty analytically. Thus, an uncertainty assessment of these methods requires the usage of computer-intensive resampling methods that are not feasible with a large number of covariates.

1.4. Scope of Research

This work explored an interaction hypothesis test in model-agnostic form, meaning that it can be used with any kind of prediction model. It was assumed that the prediction model has enough capacity to potentially estimate interaction effects. In particular, consider the following null hypothesis that there is no interaction effect in the population involving any variable in s:

$$H_0 : f(\boldsymbol{x}) = \text{PD}\left(\boldsymbol{x}_{S\setminus s}\right) + \sum_{j\in s}\text{PD}(\boldsymbol{x}_j) \text{ and respectively} \tag{17}$$

$$H_1 : f(\boldsymbol{x}) \neq \text{PD}\left(\boldsymbol{x}_{S\setminus s}\right) + \sum_{j\in s}\text{PD}(\boldsymbol{x}_j). \tag{18}$$

The set s describes the covariates of interest. For example, if $s = \{1,3\}$ and $S = \{1,2,3\}$, then it tests whether there is any interaction involving the first and third covariates. In this special case, the statistical test includes second- and third-order interaction effects. In general, the number of elements, $|s|$, determines the highest order of interaction effects considered in the hypothesis test.

Generally, one could consider H_s^2; however, using measure H_s^2 as the basis for the interaction test would have some disadvantages in practice:

- Simulations of H_s^2 show increased false positive rates [27,30].
- There is no asymptotic null distribution of the hypothesis test of [22] for the presence of interactions available in model-agnostic form.
- The H_s^2 interaction test is based on Monte Carlo simulations to quantify uncertainty [31], which are computationally runtime-intensive.
- To the best of the authors' knowledge, no systematic power simulation in hypothesis interaction tests based on H_s^2 was conducted.

This work addresses all of these issues. Furthermore, none of the existing IML approaches provide error-rate control [32], and thus, no severe testing is possible. Ref. [33] developed a statistically sophisticated philosophy of science in which the problem of induction is reduced to the practice of severe testing. To believe in a hypothesis is not only a function of the method or data used but also concerns how well the method was critically tested to rule out potential flaws. This work is a first step towards embedding IML methods into this statistical testing framework.

As an alternative to H_s^2, the interaction difference (IAD) and the corresponding hypothesis interaction test are introduced in Section 2. It is shown how the IAD can be transformed into a test statistic that can be embedded into a two-sided, one-sample Z-test. Then, in Section 3, the asymptotic distribution of the test statistic based on test data is derived. Simulations of the proposed method are given in Section 4, which include the distribution of the proposed test statistic (Section 4.1), type 1 error, and power in linear (Section 4.2) models. The advantage of those simulation scenarios is that interaction effects can be more easily incorporated than in more complex black-box models in the design. Section 4.3 covers simulations of $\hat{z}_4$ based on a random forest model. This situation is more realistic than the previous sections because, in linear models, one would not need this interaction test in practice. However, it is harder to control interaction effects in nonlinear simulation

designs. The data analysis example in Section 5 focuses on a variant of the test statistic that includes covariate information.

2. Hypothesis Test of Interactions in Prediction Models

The concept of the proposed statistical test is to compare variances in the estimated prediction model $\hat{f}$ and the estimated prediction model without interactions represented by PD functions. That means both variances are derived from the same data and, hence, dependent. Here, we follow the framework of [34] for robust tests of scale in paired samples. Those tests convert the hypothesis to allow standard asymptotical tests to be used. An advantage of this approach is that these are far more computationally efficient than Monte Carlo permutation tests. This is especially important in high-dimensional prediction tasks to be able to analyze a larger subspace of the exponentially growing number of all possible interaction effects. The key idea is to test whether the interaction difference

$$\mathrm{IAD}_s = \underbrace{\mathrm{Var}_{\boldsymbol{x}}(f(\boldsymbol{x}))}_{\mathrm{IAD}_{f,s}} - \underbrace{\mathrm{Var}_{\boldsymbol{x}}\left(\mathrm{PD}\left(\boldsymbol{x}_{S\setminus s}\right) + \sum_{j\in s}\mathrm{PD}(\boldsymbol{x}_j)\right)}_{\mathrm{IAD}_{\mathrm{PD},s}} \tag{19}$$

equals zero. IAD_s measures the deviation of variability between the original prediction model, $f(\boldsymbol{x})$, and the prediction model under the null hypothesis. Following [22], the prediction model $f(\boldsymbol{x})$ can be decomposed under H_0 into $\mathrm{PD}\left(\boldsymbol{x}_{S\setminus s}\right) + \sum_{j\in s}\mathrm{PD}(\boldsymbol{x}_j)$ if the covariates in set s do not contribute to interaction effects. Proof of this statement based on the functional ANOVA framework is given in Supplementary Materials Section S1. The decomposition of the prediction model for the purpose of testing IAD_s is given via

$$f(\boldsymbol{x}) = \mathrm{PD}\left(\boldsymbol{x}_{S\setminus s}\right) + \sum_{j\in s}\mathrm{PD}(\boldsymbol{x}_j) + \zeta(\boldsymbol{x}). \tag{20}$$

The term $\zeta(\boldsymbol{x})$ includes, for example, additional interaction terms of set s that are not included in $\mathrm{IAD}_{\mathrm{PD},s}$. Under H_0, it holds that $\mathrm{Var}_{\boldsymbol{x}}(\zeta(\boldsymbol{x})) = 0$ and analogous terms in the H_1 scenario, $\mathrm{Var}_{\boldsymbol{x}}(\zeta(\boldsymbol{x})) \neq 0$. For example, in the context of a linear prediction model, $f(\boldsymbol{x}) = \boldsymbol{\beta}_0 + \boldsymbol{x}_1\boldsymbol{\beta}_1 + \boldsymbol{x}_2\boldsymbol{\beta}_2 + \boldsymbol{x}_3\boldsymbol{\beta}_3 + \boldsymbol{x}_2\boldsymbol{x}_3\boldsymbol{\beta}_{2,3}$, under H_0 with no interaction effect of $\boldsymbol{x}_1$ (Supplementary Materials Section S2.1), the error term $\zeta(\boldsymbol{x})$ consists of a linear combination of coefficients and their respective expectations of the covariate terms.

Not all possible specifications of set s are meaningful. For example, using the empty set would give $\mathrm{IAD}_{\mathrm{PD},s} = \mathrm{Var}_{\boldsymbol{x}}(\mathrm{PD}(\boldsymbol{x}_S)) = \mathrm{Var}_{\boldsymbol{x}}(f(\boldsymbol{x}))$, which results in $\mathrm{IAD}_s = 0$. This case is excluded. Furthermore, the cases with a number of elements $|S\setminus s| = 1$ and $|S\setminus s| = 0$ are equivalent. Consider the specific case $S = \{1,2,3\}$. Then, $\mathrm{IAD}_{\mathrm{PD},s=\{1,2\}} = \mathrm{Var}_{\boldsymbol{x}}\left(\mathrm{PD}(\boldsymbol{x}_3) + \sum_{j=1}^{2}\mathrm{PD}(\boldsymbol{x}_j)\right) = \mathrm{Var}_{\boldsymbol{x}}\left(\sum_{j=1}^{3}\mathrm{PD}(\boldsymbol{x}_j)\right)$ that is equal to $\mathrm{IAD}_{\mathrm{PD},s=\{1,2,3\}}$ because $\mathrm{PD}(\boldsymbol{x}_\varnothing)$ does not depend on covariate values and is constant. In this specific case, all combinations of the set s with two covariates are excluded. Instead, the set is described as $s = \{1,2,3\}$.

Consider the following specific example of IAD_s: assuming a linear regression model with three independent, multivariate, standard, normal, distributed covariates and all possible interaction effects under restrictions of H_0, the value of IAD_s is zero, regardless of the set s (see Supplementary Materials Section S2 for details). Deviations from zero in IAD_s are in favor of the alternative hypothesis H_1. In the scenarios under the alternative hypothesis H_1, the test statistic equals the sum of all quadratic interaction coefficients that include the covariates of set s (Supplementary Materials Section S2.5).

To test the condition under H_0 that $\mathrm{IAD}_s = 0$, the difference in variances in Equation (19) can be rewritten as covariance using

$$z_1 = f(\boldsymbol{x}) + \mathrm{PD}\left(\boldsymbol{x}_{S\setminus s}\right) + \sum_{j\in s}\mathrm{PD}(\boldsymbol{x}_j) \text{ and} \tag{21}$$

$$z_2 = f(\boldsymbol{x}) - \mathrm{PD}\left(\boldsymbol{x}_{S\setminus s}\right) - \sum_{j\in s}\mathrm{PD}(\boldsymbol{x}_j) \text{ it follows that} \tag{22}$$

$$\mathrm{IAD}_s = \mathrm{Cov}_{\boldsymbol{x}}(z_1, z_2). \tag{23}$$

Proof of this equivalence is given in Supplementary Materials Section S3 that was based on the idea of [35]. The covariance in Equation (23) is the expectation of $z_3 = (z_1 - \mathrm{E}_{\boldsymbol{x}}(z_1))(z_2 - \mathrm{E}_{\boldsymbol{x}}(z_2))$. Let $\hat{z}_{3,i}$ be the estimated value of z_3 evaluated at the i-th observed value in the data set, and $\hat{z}_3 = (\hat{z}_{3,1}, \hat{z}_{3,2}, \ldots, \hat{z}_{3,n})$. The modified Pitman test [34] then evaluates a null hypothesis $\mathrm{E}(z_3) = 0$ in the framework of a one-sample, two-sided Z-test, which is equivalent to testing whether the difference of variances in Equation (19) is zero. In particular, the test statistic is given via

$$\begin{aligned}
\hat{z}_4 &= \frac{\sqrt{n}\bar{z_3}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{z}_{3,i} - \bar{z_3}\right)^2}} \text{ with} \\
\bar{z_3} &= \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{3,i} \text{ that estimate the term} \\
z_4 &= \frac{\mathrm{E}(z_3)}{\sqrt{\mathrm{Var}(z_3)}}.
\end{aligned} \tag{24}$$

Small absolute values around zero indicate H_0, and large absolute values favor H_1. For testing, the value of $\hat{z}_4$ is compared to the respective quantiles of a standard normal distribution.

A related but different question than testing interaction effects is how these influence the prediction performance of the prediction model. Here, we introduce a variant of $\hat{z}_4$ that includes response information. Equation (19) is extended to interaction-difference performance (IADP)

$$\mathrm{IADP}_s = \underbrace{\mathrm{Var}_{\boldsymbol{x}}(y(\boldsymbol{x}) - f(\boldsymbol{x}))}_{\mathrm{IADP}_{f,s}} - \underbrace{\mathrm{Var}_{\boldsymbol{x}}\left(y(\boldsymbol{x}) - \mathrm{PD}\left(\boldsymbol{x}_{S\setminus s}\right) - \sum_{j\in s}\mathrm{PD}(\boldsymbol{x}_j)\right)}_{\mathrm{IADP}_{\mathrm{PD},s}}. \tag{25}$$

The term $\mathrm{IADP}_{f,s}$ is the mean squared error of the prediction model with a quantitative response, $y(\boldsymbol{x})$, or the Brier score in the case of a binary response scale. $\mathrm{IADP}_{\mathrm{PD},s}$ is the mean squared error (MSE) of the restricted prediction model under a null hypothesis of no interaction effects of covariates in set s. A one-sided test is more appropriate here because the interest is whether the interaction effects of covariates s decrease MSE (alternative hypothesis). The terms $z_{P,1}, z_{P,2}, z_{P,3}, z_{P,4}$ for the construction of the interaction test are analogously derived to z_1, z_2, z_3, z_4 via a plugin of Equation (25).

3. Asymptotics of Test Statistics

This section summarizes the asymptotic properties of $\hat{z}_4^{\mathcal{T}}$ evaluated on test data. The PD functions and the prediction model are estimated from the training data. Let f denote the target function and $\widehat{f}$ the corresponding estimate. Moreover, denote

$$g(\boldsymbol{x}) = PD(\boldsymbol{x}_{S\setminus s}) + \sum_{j\in S} PD(\boldsymbol{x}_j)$$

and let $\hat{g}$ denote the corresponding estimate. Then, following (Equation (1)) in Hooker (2004) [24], it holds that $\mathrm{Var}_{\boldsymbol{x}}(f(\boldsymbol{x})) \geqslant \mathrm{Var}_{\boldsymbol{x}}(g(\boldsymbol{x}))$ and $\mathrm{Var}_{\boldsymbol{x}}(f(\boldsymbol{x})) = \mathrm{Var}_{\boldsymbol{x}}(g(\boldsymbol{x}))$ if and only if $g(\boldsymbol{x}) = f(\boldsymbol{x})$ almost everywhere. Hence, testing the equivalence of the variances is, indeed, equivalent to testing $f(\boldsymbol{x}) = g(\boldsymbol{x})$ almost everywhere.

Theorem 1. *Let $n^{\mathcal{T}}$ denote the sample size of the test set, and let n denote the sample size of the training set. Assume that $\sigma_f^2 = Var_{\boldsymbol{x}}(f(\boldsymbol{x}))$ satisfies $0 < \sigma_f^2 < \infty$. Moreover, if $Var_{\boldsymbol{x}}(f(\boldsymbol{x})) = Var_{\boldsymbol{x}}(g(\boldsymbol{x}))$, then assume that, for $n^{\mathcal{T}} \to \infty$ and some $a \in (1,2)$,*

$$\lim_{n^{\mathcal{T}} \to \infty} \left(n^{\mathcal{T}}\right)^a Var_{\boldsymbol{x}}(\widehat{f}(\boldsymbol{x}) - \widehat{g}(\boldsymbol{x})) \xrightarrow{P} c \tag{26}$$

with $0 < c < \infty$. Define

$$\widehat{z_{1,i}} = \widehat{f}(\boldsymbol{X}_{i,}) + \widehat{g}(\boldsymbol{X}_{i,}),$$

and

$$\widehat{z_{2,i}} = \widehat{f}(\boldsymbol{X}_{i,}) - \widehat{g}(\boldsymbol{X}_{i,}).$$

(i) If $Var_{\boldsymbol{x}}(f(\boldsymbol{x})) = Var_{\boldsymbol{x}}(g(\boldsymbol{x}))$, then

$$\left(n^{\mathcal{T}}\right)^{a/2} \sum_{i=1}^{n^{\mathcal{T}}} \left(\left(\widehat{z_{1,i}} - \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \widehat{z_{1,i}}\right) \left(\widehat{z_{2,i}} - \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \widehat{z_{2,i}}\right) \right) \xrightarrow{P} \mathcal{N}(0, c\,\sigma_f^2)$$

for some $0 < \sigma_f^2 < \infty$.

(ii) If $Var_{\boldsymbol{x}}(f(\boldsymbol{x})) \neq Var_{\boldsymbol{x}}(g(\boldsymbol{x}))$, then

$$\left(n^{\mathcal{T}}\right)^{a/2} \sum_{i=1}^{n^{\mathcal{T}}} \left(\left(\widehat{z_{1,i}} - \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \widehat{z_{1,i}}\right) \left(\widehat{z_{2,i}} - \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \widehat{z_{2,i}}\right) \right) \xrightarrow{P} \infty$$

Proof. (ii) is trivial. For (i), note that

$$\begin{aligned} &= \left(\left(\widehat{z_{1,i}} - \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \widehat{z_{1,i}}\right) \left(\widehat{z_{2,i}} - \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \widehat{z_{2,i}}\right) \right) \\ &= \left(\left(\widehat{z_{2,i}} - \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \widehat{z_{2,i}}\right)^2 + 2 \left(\widehat{z_{2,i}} - \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \widehat{z_{2,i}}\right) \widehat{g}(\boldsymbol{X}_{i,}) \right). \end{aligned}$$

It follows from Equation (26) that

$$\left(n^{\mathcal{T}}\right)^{a/2} \sum_{i=1}^{n^{\mathcal{T}}} \left(\widehat{z_{2,i}} - \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \widehat{z_{2,i}}\right)^2 \xrightarrow{P} 0.$$

Moreover, the CLT and Slutzky's lemma yield the result that

$$\left(n^{\mathcal{T}}\right)^{a/2} \left(\widehat{z_{2,i}} - \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \widehat{z_{2,i}}\right) \widehat{g}(\boldsymbol{X}_{i,}) = \left(\left(n^{\mathcal{T}}\right)^{a/2} \left(\widehat{z_{2,i}} - \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \widehat{z_{2,i}}\right) \right) \widehat{g}(\boldsymbol{X}_{i,})$$

converges to a normal distribution with variance $c\,\sigma_f^2$. The crucial assumption of the above theorem is that the convergence rate of the variance of the differences between $\hat{f}$ and $\hat{g}$ is faster than $(n^{\mathcal{T}})^{-1}$, where $n^{\mathcal{T}}$ is the size of the test set. For most models, this will be the case when the size of the training set goes to infinity faster than the size of the test set, i.e., $n^{\mathcal{T}}/n \to 0$ for $n, n^{\mathcal{T}} \to \infty$. A similar result can be derived for the test based on $\hat{z}_{P,4}$, which measures the differences in MSE performance (for a comparison, see Equation (25)). □

Theorem 2. *Let $f : R^p \to R$ and $g : R^p \to R$ be two fixed prediction functions. Moreover, let $(\boldsymbol{X}_{1,}, y(\boldsymbol{X}_{1,})), (\boldsymbol{X}_{2,}, y(\boldsymbol{X}_{2,})), \ldots \left(\boldsymbol{X}_{n^{\mathcal{T}},}, y(\boldsymbol{X}_{n^{\mathcal{T}},})\right)$ denote i.i.d. samples in R^{p+1}. Further, assume*

$$E_{\boldsymbol{x}}[f(\boldsymbol{x})^2] < \infty,$$
$$E_{\boldsymbol{x}}[g(\boldsymbol{x})^2] < \infty,$$
$$E_{\boldsymbol{x}}[y(\boldsymbol{x})^2] < \infty.$$

Then,

$$\frac{1}{\sqrt{n^{\mathcal{T}}}}\left(\sum_{i=1}^{n^{\mathcal{T}}}\left(y(\boldsymbol{X}_{i,}) - f(\boldsymbol{X}_{i,})\right)^2 - \sum_{i=1}^{n^{\mathcal{T}}}\left(y(\boldsymbol{X}_{i,}) - g(\boldsymbol{X}_{i,})\right)^2\right) \to \mathcal{N}(\mu_{diff}, \sigma^2_{diff}),$$

where

$$\mu_{diff} = E_{\boldsymbol{x}}[(y(\boldsymbol{x}) - f(\boldsymbol{x}))^2] - E_{\boldsymbol{x}}[(y(\boldsymbol{x}) - g(\boldsymbol{x}))^2],$$

and σ^2_{diff} can be estimated from the given sample.

If we are interested in showing that f has a smaller expected squared prediction error than g, we can consider the testing problem

$$H_0 : \mu_{diff} \geqslant 0.$$

In particular, in the setting of the paper, we set

$$f = \hat{f}, \qquad g(\boldsymbol{X}_{i,}) = \widehat{\mathrm{PD}}(\boldsymbol{X}_{i,S\setminus s}) + \sum_{j\in s}\widehat{\mathrm{PD}}(\boldsymbol{X}_{i,j})$$

in the above theorem.

Then, the rejection of the null hypothesis provides evidence that the original prediction function, $\hat{f}$, has a smaller prediction error than the "prediction function without" interactions, $\widehat{\mathrm{PD}}(\boldsymbol{X}_{i,S\setminus s}) + \sum_{j\in s}\widehat{\mathrm{PD}}(\boldsymbol{X}_{i,j})$. This, in turn, suggests that there is a meaningful modeling of interaction in $\hat{f}$ and that there are interactions in the target function f. It has to be noted that testing

$$H_0 : \text{Interaction effects of } f \text{ do not improve MSE performance}$$

is not guaranteed to control the nominal level for the two-sample problem. However, simulations indicate that it will typically do so (and even be rather conservative).

4. Simulation

This section summarizes simulation results with the proposed interaction test of Section 2. All simulations use independently generated test data sets to evaluate the interaction test with the same sample size and data-generating process as the respective simulated training data sets. The first simulation analyzes the distribution of $\hat{z}_4$ in the context of linear models while increasing the number of variables (Section 4.1). The second simulation conducts an analysis of type 1 error and power in the context of linear models (Section 4.2). Linear models were used in the first two simulations to demonstrate the empirical behavior in easy-to-understand scenarios where the model allows for the specification of the type of estimated interaction effects. Note that, in practical applications with estimated linear models, there would be no need to conduct the proposed interaction difference test. On the other hand, $\hat{z}_4$ was developed for model-agnostic prediction models, and as such, it is desirable to check whether $\hat{z}_4$ is well behaved in these scenarios, too. Then, in the third simulation, nonlinear

models were explored based on a real data set (Section 4.3). Last but not least, we investigated the proposed modification $\hat{z}_{P,4}$ of the interaction test with responses.

The programming language R for the source code of the complete simulation is available as additional online supplementary material to enhance reproducibility (see the reference after Section 6). The interaction test for prediction models was implemented in the R-package *IADT 1.2.1*, available in the comprehensive R archive network (https://cran.r-project.org (accessed on 26 May 2024)).

4.1. Test Statistic Distribution in Linear Models

To investigate the behavior of the test statistic $\hat{z}_4$ in the context of a linear model, the following data-generating process was specified: The p covariates

$$\begin{aligned}
&\boldsymbol{x} \in \mathbb{R}^p \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \text{follow a multivariate normal distribution with correlations} \\
&\rho_{\text{low},j,k} = 0.25 \text{ over the set } \{j, k \in 1, \ldots, p : j \neq k\}, \\
&\rho_{\text{medium},j,k} = 0.5 \text{ and} \\
&\rho_{\text{high},j,k} = 0.75 \text{ (equi-correlation). The hypothesis is specified with} \\
&S = \{1, \ldots, p\},\ s = \{1\} \text{ and the true linear model with one interaction term is} \\
&f(\boldsymbol{x}) = \boldsymbol{x}_1\boldsymbol{\beta}_1 + \ldots + \boldsymbol{x}_p\boldsymbol{\beta}_p + \boldsymbol{x}_1\boldsymbol{x}_2\boldsymbol{\beta}_{1,2} + \epsilon \text{ with } \epsilon \sim N(0, \sigma^2). \qquad (27)
\end{aligned}$$

This setting was chosen under the alternative hypothesis with a minimal number of interaction terms such that the test statistic was expected to be closer to zero compared to settings with more interaction terms. This simulation was conducted with a different numbers of covariates, $p = \{5, 10, \ldots, 100\}$. The sample size was fixed with 1000 for both simulated training and test data sets. In each scenario, the variance σ^2 of the error term ϵ was set to 0.8 based on prior simulations with $n = 10^6$. The coefficients of the data-generating process were set to $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \boldsymbol{\beta}_{1,2}\right) = (1, \ldots, 1) \in \mathbb{R}^{p+1}$ to study power and $\boldsymbol{\beta} = (1, \ldots, 0)$ to investigate the type I error. The linear model was correctly specified to include all covariates of the data-generating process. Each scenario was independently repeated 100 times. All together, 57,600 test statistics were simulated.

Here, the simulation results are shown for the null hypothesis that covariate one does not contribute to interaction effects ($s = \{1\}$). Figure 1 shows the difference $\tilde{d}(\hat{z}_4)$ defined by $\hat{z}_4$, minus the normalized rank quantile of the standard normal distribution on the left side. $\tilde{d}(\hat{z}_4)$ was estimated based on 100 independent replicates of $\hat{z}_4$, given the number of covariates and the correlation of each scenario. All boxplots fluctuate around the value of zero across different number of covariates. Furthermore, the boxplots on the left side, $\tilde{d}(\hat{z}_4)$, of Figure 1 are comparable to those on the right side, $\tilde{d}(\Phi)$, which used a standard, normal, distributed random variable, Φ, instead of $\hat{z}_4$. Note that the volatility in boxplots occurs due to the estimation of ranks, and with increasing sample sizes, the differences in $\tilde{d}(\Phi)$ would converge to zero. The Shapiro–Wilk test [36] is considered the most powerful in detecting non-normality according to [37]. If all 288 scenarios were evaluated with the Shapiro–Wilk test and adjusted for multiple comparisons with a false-discovery rate approach [38] of 0.05, then there would be no case that significantly departed from the normality distribution assumption.

The results for the alternative hypothesis specified in Equation (27) are shown in Figure 2. There is a decreasing trend to shift the distribution of $\hat{z}_4$ more towards zero the higher the number of covariates. With low covariate correlation, the lower quartile of the distribution crosses the zero line with about 30 covariates. When the covariate correlation is higher, this happens with about 20 covariates. In such cases, it is expected that power is reduced because the H_1 distribution becomes more similar to the H_0 distribution. After about 30 covariates, the median of $\hat{z}_4$ does not decrease further. For comparison, the same simulation was conducted using the t-statistic in a linear model of the interaction effect in Figure 3. This figure shows a decreasing trend in the location of the simulated t-value distribution, but the gap of the medians to zero is larger than in Figure 2, and more

covariates are needed so that the lower quartile of the simulated distribution crosses the zero line. The model-specific hypothesis test that was explicitly developed for linear models can be expected to be more efficient in terms of power than a model-agnostic hypothesis test if the assumptions are justified. In conclusion, the proposed test statistic is empirically good when approximated with a normal distribution under H_0, and small effects under H_1 result in similar behavior to *t*-tests with linear models.

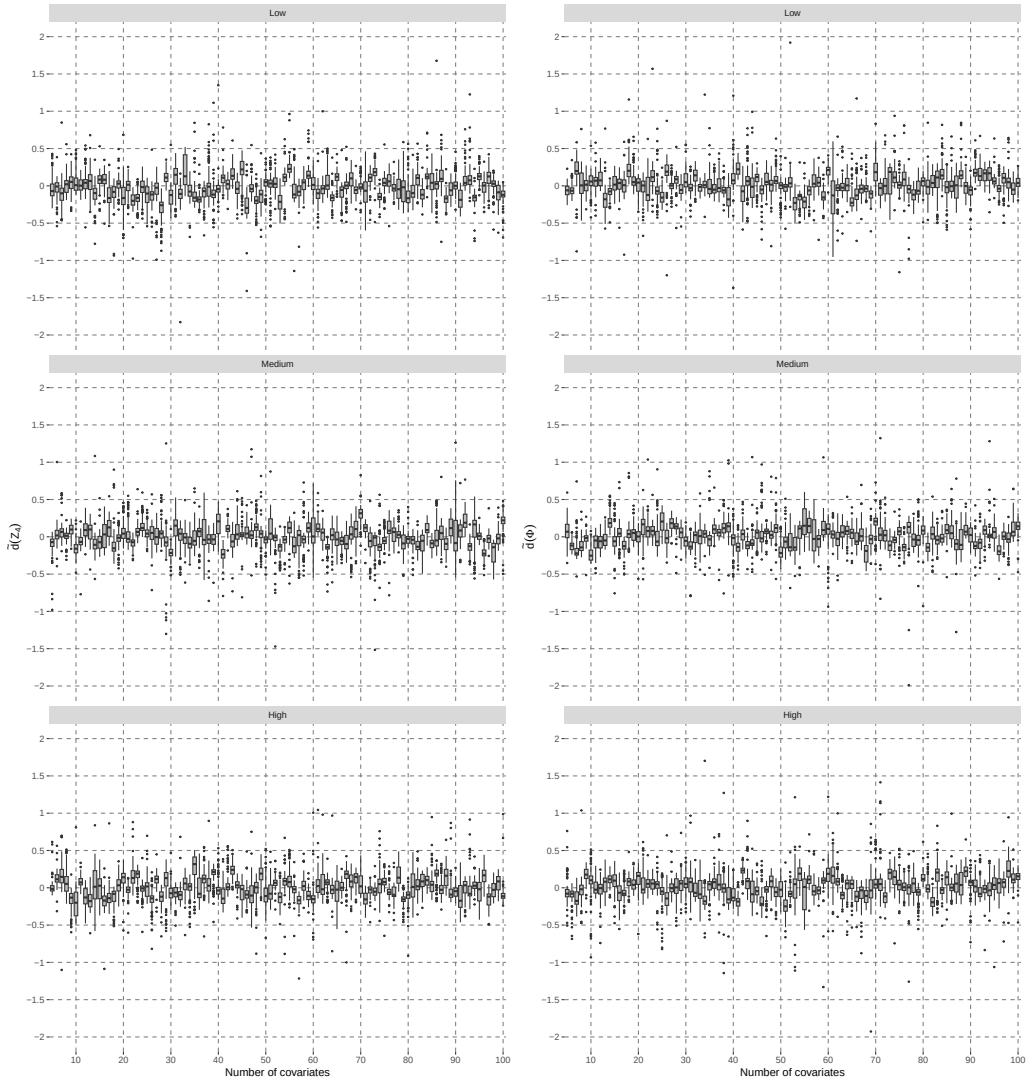

Figure 1. The boxplots on the left side show the distribution of $\tilde{d}$, defined according to $\hat{z}_4$, minus the normalized rank transformation of $\hat{z}_4$ to the standard normal distribution. Instead of $\hat{z}_4$, the standard, normal random variable Φ was used on the right side. The graphs represent different correlation scenarios: low, medium, and high.

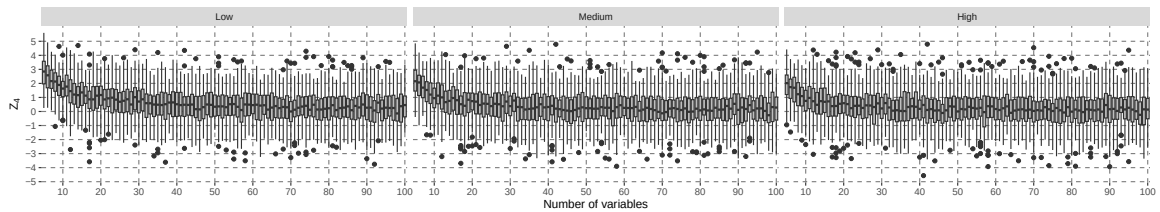

Figure 2. Boxplots of the distribution of $\hat{z}_4$ based on linear models under $H_1, s = \{1\}$ with one interaction term, $\beta_{1,2}$. The graphs represent different correlation scenarios: low, medium, and high.

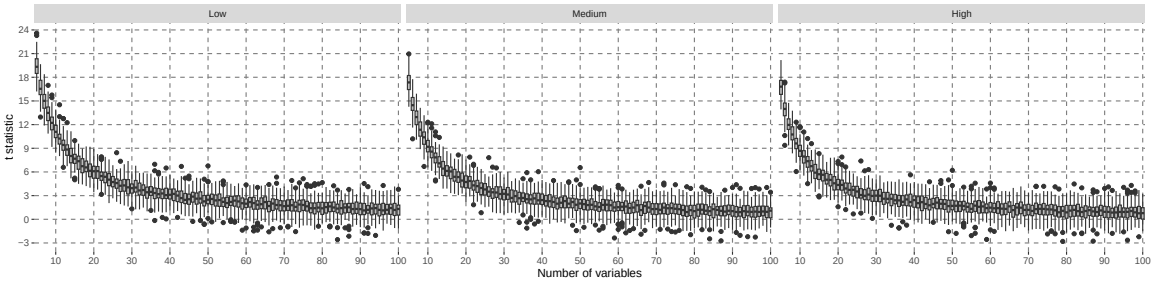

Figure 3. Boxplots of the distribution of the t-statistic of the interaction effect $\beta_{1,2}$ in a linear model under H_1. The graphs represent different correlation scenarios: low, medium, and high.

4.2. Power Simulation in Linear Models

This section focuses on the power and type I error simulation in linear models. Due to the linear structure of the models, interaction effects can be specified separately from main effects, and thus, simulations under both hypotheses H_0 and H_1 can be more easily specified and verified than in more complex prediction models. Therefore, the setting of linear models is a good starting point to explore the properties of the interaction test based on $\hat{z}_4$. Note that, in practice, the proposed interaction test is not needed in linear models because ANOVA methods [23] were developed for the specific case of linear models to test whether the coefficients are zero.

The simulation design of the covariate distribution was the same as in the previous section, Section 4.1, with $p = 5$, except additionally considering the case of no correlation. The data-generating model consisted of three different scenarios with the error term $\epsilon \sim N(0, \sigma^2)$, and it was allowed to differ from the estimated prediction model specification:

$$f(\boldsymbol{x}) = \sum_{j=1}^{p} \boldsymbol{x}_j \boldsymbol{\beta}_j + \epsilon \text{ (main effects),}$$

$$f(\boldsymbol{x}) = \sum_{j=1}^{p} \boldsymbol{x}_j \boldsymbol{\beta}_j + \sum_{j=1}^{p-1} \sum_{k>j} \boldsymbol{x}_j \boldsymbol{x}_k \boldsymbol{\beta}_{j,k} + \epsilon \text{ (main effects, all second order interactions) and}$$

$$f(\boldsymbol{x}) = \sum_{j=1}^{p} \boldsymbol{x}_j \boldsymbol{\beta}_j + \sum_{j=1}^{p-1} \sum_{k>j} \boldsymbol{x}_j \boldsymbol{x}_k \boldsymbol{\beta}_{j,k} + \sum_{j=1}^{p-2} \sum_{k>j} \sum_{l>j,l>k} \boldsymbol{x}_j \boldsymbol{x}_k \boldsymbol{x}_l \boldsymbol{\beta}_{j,k,l} + \epsilon$$

(main effects, all second and third order interactions).

The inference is about the population-model interaction effects (unknown in practice), but in this simulation, the interaction effects are known. The alternative hypothesis is true if the corresponding interaction effects are estimated in the prediction model and simulated in the data-generating process. In the case of the misspecification of the linear predictor, the estimated coefficients converge to the true coefficients of the data-generating process.

The error variance σ^2 was optimized on a data set with $n = 10^6$ prior to the simulation to approximately yield an explained variance of 0.25, 0.5, and 0.75. Sample sizes varied with $n = \{100, 125, \ldots, 300\}$. Lower and upper sample sizes were chosen to avoid instabilities in the estimated coefficients and reach power levels of 1 in at least one scenario. Three different null hypotheses, $s = \{1\}, \{1, 2\}, \{1, 2, 3\}$, were investigated. The linear model was specified under H_1 to estimate all possible main and interaction effects up to the third order. In contrast, under H_0 all interaction effects that included covariates of set s were excluded from the data-generating process. Each combination of the scenarios was repeated independently 1000 times.

The rows of plots in Sections 4.2.1 and 4.2.2 correspond to different covariate correlations, 0, 0.25, 0.5, 0.75, and the columns of plots display varying explained variances, 0.25, 0.5, 0.75. The dotted–dashed lines represent the upper and lower bounds of the exact pointwise 0.95 Clopper–Pearson confidence intervals [39] that were calculated for the type

I error and power proportions. The next two sections, Sections 4.2.1 and 4.2.2, summarize the type I error and power simulation results consisting of 1.08×10^6 hypothesis tests. Additional figures are available in Supplementary Materials Section S4.

4.2.1. Type I Error Results

Figure 4 shows the results for the correctly specified linear model under H_0 with $s = \{1, 2\}$. The estimated linear model includes the main effects of $\boldsymbol{x}_{(1,2)}$ and additional interaction effects of the covariates $\boldsymbol{x}_{(3,4,5)}$ up to the third order. The type I error was controlled with a significance level of $\alpha = 0.05$ in all scenarios, and the hypothesis test is robust to covariate correlations, as well as explained variances.

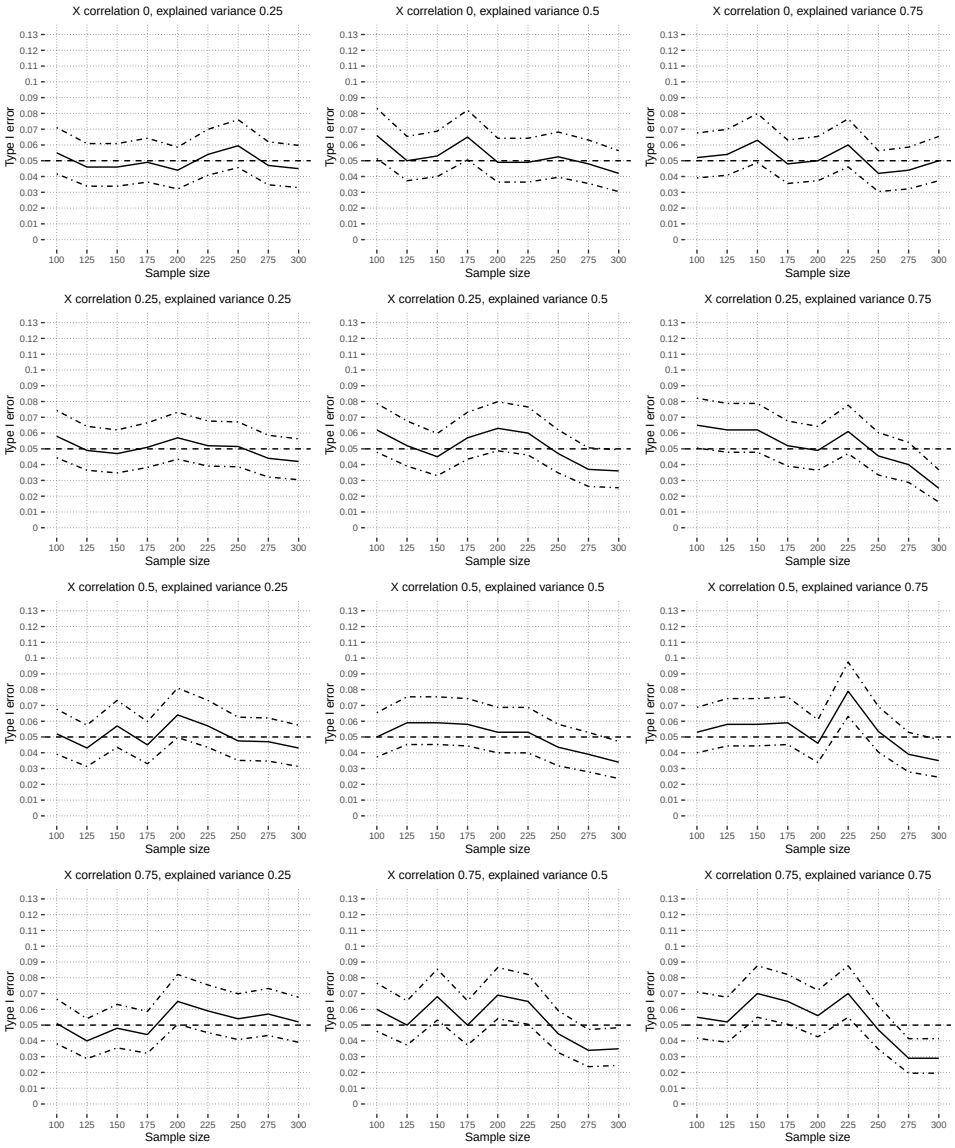

Figure 4. Type I error simulations in scenario of correctly specified estimated linear model with null hypothesis $s = \{1, 2\}$. Dashed lines correspond to the standard alpha 0.05 threshold, dashed-dotted lines represent pointwise 0.95 Clopper-Pearson confidence intervals and full lines show the estimated Type I error.

4.2.2. Power Results

Figure 5 shows the power results under the alternative hypothesis based on $s = \{1, 2\}$ with correctly specified linear models. The hypothesis test reaches power levels around 0.8 in zero- to low-covariate correlation scenarios with at most $n = 200$. The figure shows

that higher covariate correlations reduce the power levels, which are influenced by the instability of the estimated linear models because of multicollinearity in this scenario. Higher explained variances result in slightly higher power. Note that the functional ANOVA decomposition theory [24] does not theoretically work well with strong covariate correlations either because great emphasis is placed on regions with a low probability mass [25].

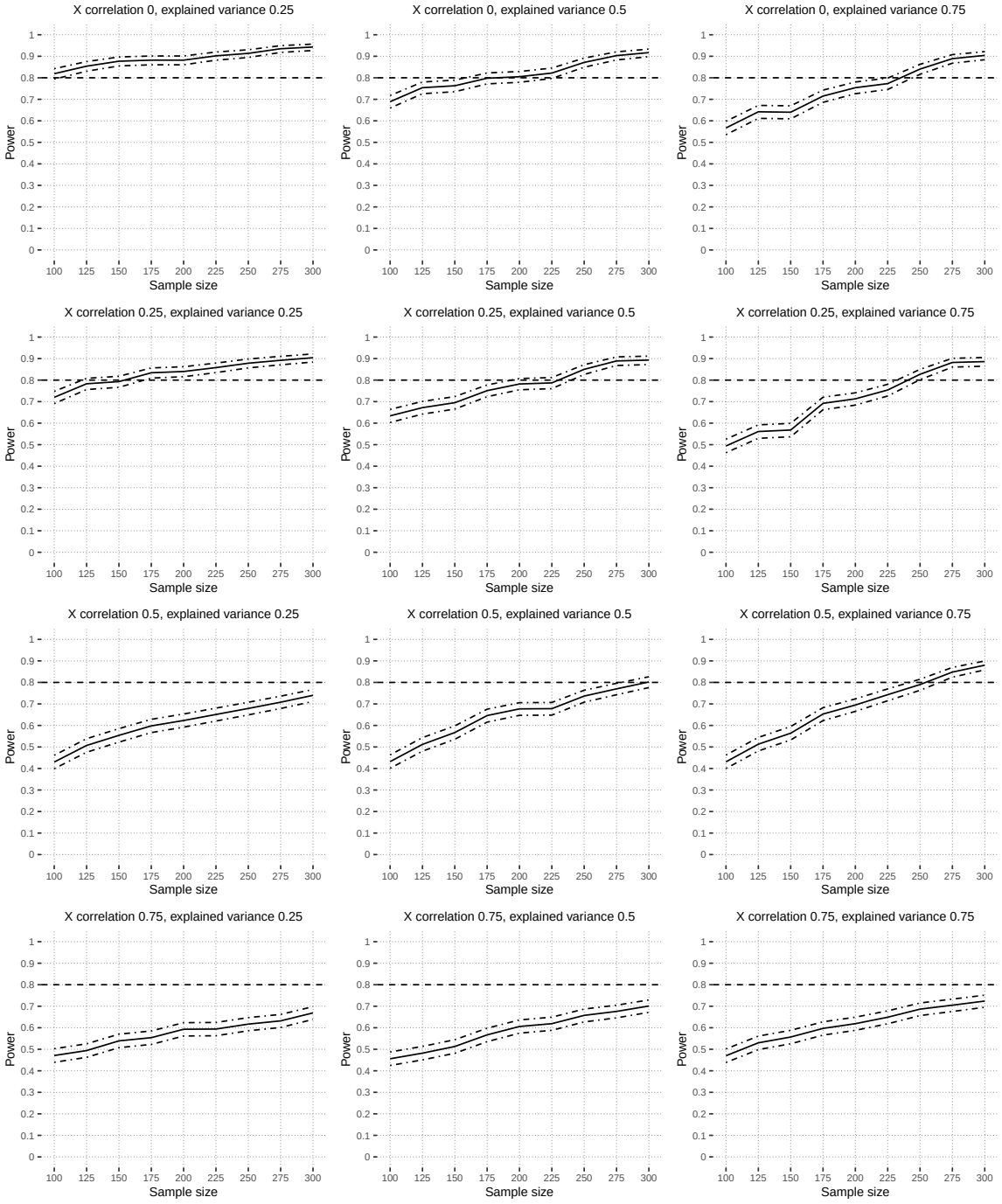

Figure 5. Power simulations with correctly specified estimated linear model, main effects, and all possible interaction effects up to the third order ($s = \{1, 2\}$). Dashed lines correspond to the standard alpha 0.05 threshold, dashed-dotted lines represent pointwise 0.95 Clopper-Pearson confidence intervals and full lines show the estimated Type I error.

Figure 6 shows the power results under H_1 with $s = \{1, 2\}$ in the context of a misspecified linear model. The data-generating model consists of all interaction effects up to order two, except those in $s = \{1, 2\}$, but in the linear model, the main effects and all possible interaction effects up to order three are estimated. Increasing covariate correlations reduces the power, and higher explained variance scenarios yield a higher power. Additional power scenarios are available in Supplementary Materials Section S4.2.

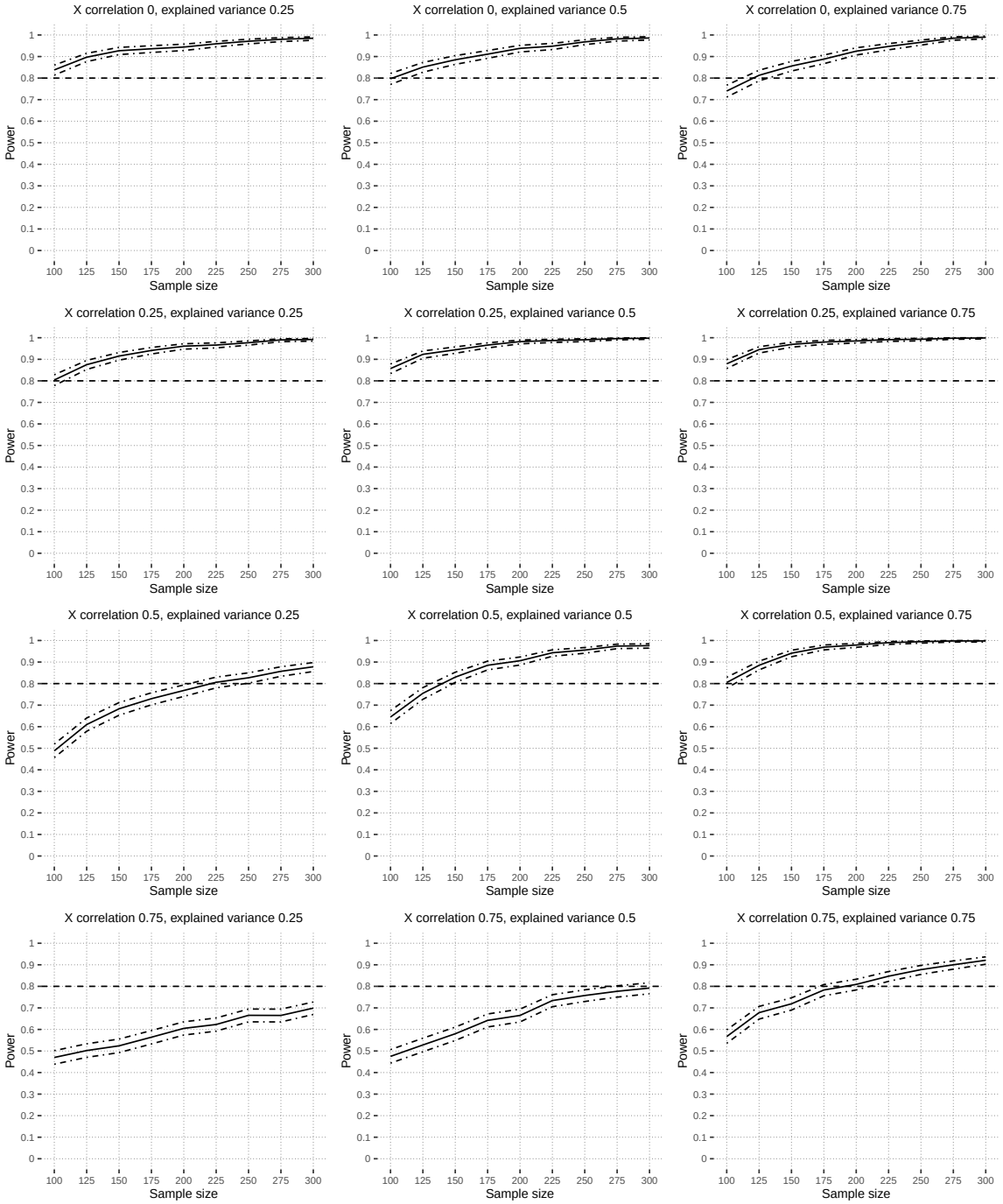

Figure 6. Power simulations with a misspecified estimated linear model with the main effects and all possible interaction effects up to the third order. The data-generating model consists of all interaction effects up to order two except those of $s = \{1, 2\}$. Dashed lines correspond to the standard alpha 0.05 threshold, dashed-dotted lines represent pointwise 0.95 Clopper-Pearson confidence intervals and full lines show the estimated Type I error.

4.3. Power Simulation in Nonlinear Models

In this section, we aim to explore the power of the interaction test in a simulation study based on a data set. As an example data set, the credit approval data from the machine learning repository *OpenML-CC18* [40,41] was used. The response variable was binary with categories for *good* and *bad* credit risks. The data set contains 1000 independent observations, along with 7 numeric and 13 categorical covariates. A descriptive overview of the data is given in Supplementary Materials Section S5.

The data-generating process of the simulation depends on the data set to be more realistic. Covariates were simulated without (*Xind*) and with dependencies (*Xdep*). In the former case, continuous covariates were randomly drawn from the marginal empirical distribution functions of one covariate. Discrete covariates were sampled according to observed relative frequencies. In the design *Xdep*, a Gaussian copula was used to simulate all continuous covariates together, considering their dependencies. The discrete covariate distribution was estimated using relative frequencies of multivariate contingency tables.

Ensemble methods like random forest were among the top-performing prediction methods with tabular data in a recent comparison to deep learning [42], and results from Kaggle competition challenges show similar trends (for example, [43]). Additionally, random forests are easy to tune, and usually, tuning the number of randomly available covariates at each split (*mtry*) suffices [44]. First, a random forest model was tuned via 10-fold cross-validation of the original data regarding out-of-sample, binomial log-likelihood function with the tuning parameter *mtry* (model $\text{RF}_{\text{interact}}$). Then, the absolute values of the interaction test statistic were evaluated for this model separately with each covariate. The three covariates with the highest values were chosen (age, employment, and existing credits). Among these sets, all possible pairwise sets with other covariates (excluding age, employment, and existing credits) were analyzed to determine the strongest two-way interaction effects in the data. These were "age of person interacts with housing finance", "employment status interacts with housing finance", and "number of existing credits interacts with job qualification". The sets correspond to the covariates

$s = \{1\} \leftrightarrow$ "age of person"

$s = \{1, 2\} \leftrightarrow$ "age of person", "employment status"

$s = \{1, 2, 3\} \leftrightarrow$ "age of person", "employment status", "number of existing credits"

To evaluate the power and type I error rates, it is necessary to be able to specify the data-generating process under both the H_0 and H_1 hypotheses. It is known that, if the random forests are restricted to only include tree stumps (only one covariate split), then there are no interaction effects. In this simulation, all data-generation processes were identical to the specification of the estimated random forest models. Under H_0, all sets, s, were restricted to tree stumps depending on all covariates with the tuned parameter *mtry* (RF_0). For each strong interaction effect, separate random forests ($\text{RF}_{\text{age, housing}}$, $\text{RF}_{\text{employment, housing}}$, $\text{RF}_{\text{credits, job}}$) were estimated with an unrestricted tree depth but only including the two variables of the previously determined interaction effect with $mtry = 2$. If there was a strong signal of two interaction covariates in the data and the random forest model had only the option to estimate the response with those covariates, then it was quite likely that the interaction effect would be estimated in the model. Under H_1 with set $s = \{1\}$, the predictions of RF_0 and $\text{RF}_{\text{age, housing}}$ were averaged with the mean. Analogously, in the case of $s = \{1, 2\}$, the random forest models RF_0, $\text{RF}_{\text{age, housing}}$ and $\text{RF}_{\text{employment, housing}}$ were averaged, and if $s = \{1, 2, 3\}$, then the average predictions of RF_0, $\text{RF}_{\text{age, housing}}$, $\text{RF}_{\text{employment, housing}}$ and $\text{RF}_{\text{credits, job}}$ were calculated. After data generation, the estimated random forest models were tuned using simulated test data analogously as model $\text{RF}_{\text{interact}}$. All together, there were 120 scenarios (10 sample sizes, two covariate designs, three sets s, and two different hypotheses) that were independently repeated 1000 times.

4.3.1. Type I Error Results

In Figure 7, the estimated type I errors, based on random forests, are shown for independent covariate simulation. The curves fluctuate around the prespecified alpha level of 0.05. In the case of dependent covariates, Figure 8 shows that the type I error is controlled for $s = \{1\}$. Larger sets indicate a small positive trend for increasing sample sizes. This could indicate that covariate dependencies have a small influence on the type I error in nonlinear models. This is in contrast to the observed results of Section 4.2.1, where even strong covariate correlations overall did not have much of an effect on the estimated type I errors. In the design *Xdep*, the strongest correlation in the Gaussian copula between "credit amount" and "credit duration" was 0.6174 in the original data set. All other numeric covariates had less absolute correlation than 0.3. The simulated interaction effect between "employment" and "housing finance", measured using the corrected contingency coefficient [45], was 0.2909. The previous value is above the 0.95 empirical simulated quantile 0.1527 under independence, and thus, this case can be interpreted as low-dependency. Another difference compared to linear models is that random forests do not have continuous predictions, which means that, for certain ranges of the covariates, the prediction function stays constant.

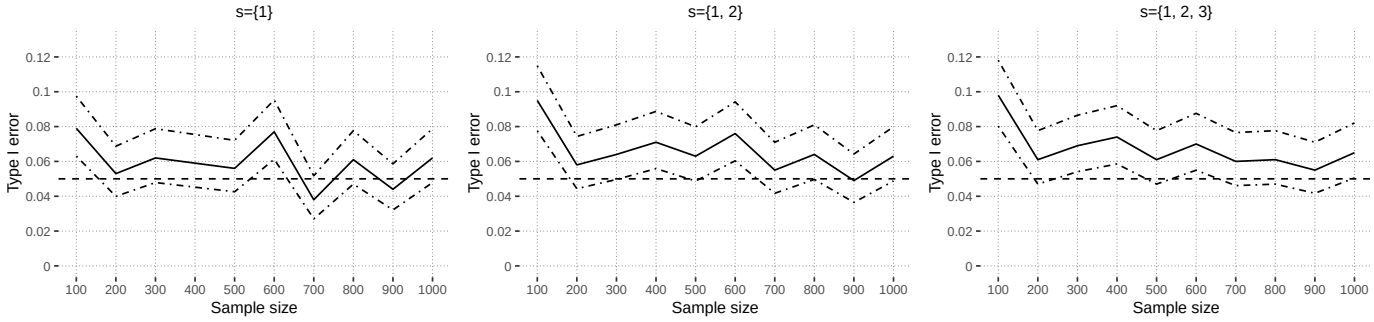

Figure 7. Estimated alpha error of the interaction test based on random forests with independent simulated covariates under different H_0 hypotheses. The dashed line represents the standard 0.05 significance threshold. Overall, the interaction test controls the prespecified alpha error. The dotted–dashed lines represent the upper and lower bounds of the exact pointwise 0.95 Clopper–Pearson confidence intervals. Full lines represent estimated type I error.

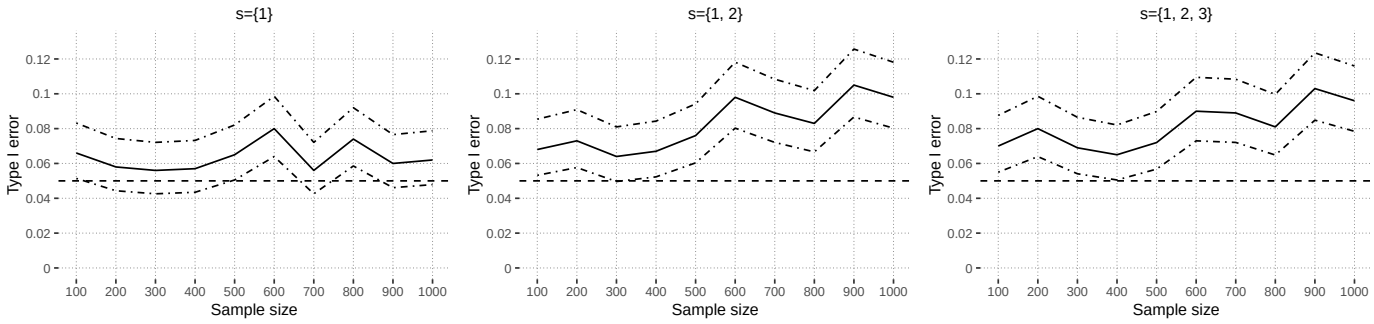

Figure 8. Estimated alpha error of the interaction test based on random forests with dependent simulated covariates under different H_0 hypotheses. The dashed line represents the standard 0.05 significance threshold. The interaction test controls the prespecified alpha error in scenario $s = \{1\}$. In the other two graphs, there is a slightly anti-conservative trend for higher sample sizes. The dotted–dashed lines represent the upper and lower bounds of the exact pointwise 0.95 Clopper–Pearson confidence intervals. Full lines represent estimated type I error.

4.3.2. Power Results

Figure 9 shows the estimated power based on random forest models. Power increases with the sample size, and the curve gradients decline. Several hundred observations

are sufficient to ensure commonly used power levels of 0.8 [46]. In contrast to Figure 9, the scenarios of $\|s\| > 1$ in Figure 10 show somewhat lower power levels at sample size $n = 1000$. It is analogous to the previous section, Section 4.3.1, that the performance using the *Xdep* design is a little bit worse than that using the *Xind* design.

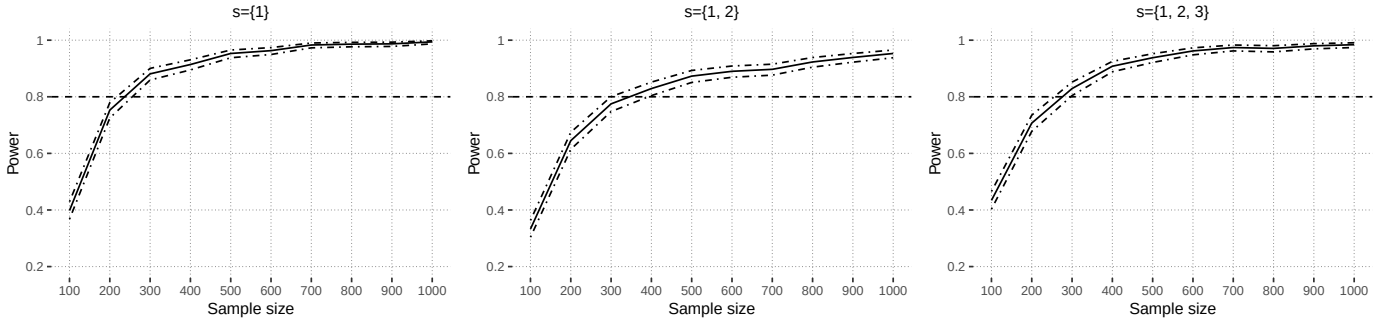

Figure 9. Estimated power of the interaction test based on random forests with independent simulated covariates under the H_1 hypothesis s. The dashed line represents a standard power level, 0.8, assumed in sample-size planning. Two hundred to three hundred observations suffice for acceptable power levels. The dotted–dashed lines represent the upper and lower bounds of the exact pointwise 0.95 Clopper–Pearson confidence intervals. Full lines represent estimated power.

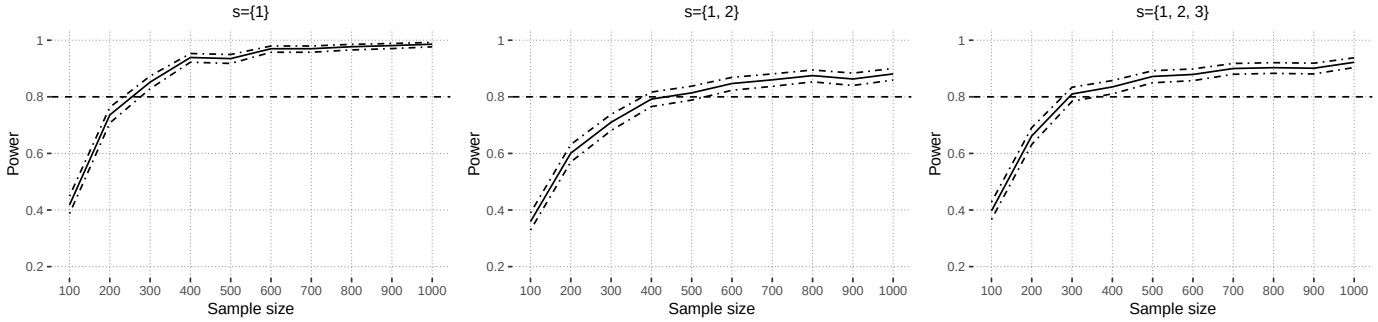

Figure 10. Estimated power of the interaction test based on random forests with dependent simulated covariates under different null hypotheses, s. The dashed line represents a standard power level of 0.8 assumed in sample size planning. Two hundred and fifty to four hundred observations suffice for acceptable power levels. The dotted–dashed lines represent the upper and lower bounds of the exact pointwise 0.95 Clopper–Pearson confidence intervals. Full lines represent estimated power.

4.4. Interaction Test Statistic with Response

In this section, we explore the proposed extension in Equation (25) to include response information in $\hat{z}_{P,4}$ as a sensitivity analysis. The simulation design was based on the example given in [27,47]. The response function takes the form of H_0

$$g(\boldsymbol{x}) = 5\sin(\pi \boldsymbol{x}_1) + 5\sin(\pi \boldsymbol{x}_2) + 20(\boldsymbol{x}_3 - 0.5)^2 + 10\boldsymbol{x}_4 + 5\boldsymbol{x}_5 + \epsilon \text{ and under } H_1 \quad (28)$$

$$g(\boldsymbol{x}) = 10\sin(\pi \boldsymbol{x}_1 \boldsymbol{x}_2) + 20(\boldsymbol{x}_3 - 0.5)^2 + 10\boldsymbol{x}_4 + 5\boldsymbol{x}_5 + \epsilon \quad (29)$$

with $\boldsymbol{x} \in \mathbb{R}^{10}$ and $\epsilon \sim N(0, \sigma^2)$. Both under H_0 and H_1, the error variance was set to achieve an explained variance of 95% based on the average of 25 independent simulated data sets of size 10^6. The sample sizes varied from $n = 100, 200, \ldots, 1000$. For each simulated training data set, a multivariate adaptive regression spline (MARS) was fitted [47] with a maximal degree of two. Type I error results are shown in Figure 11. Overall, the estimated type I error held the specified alpha level 0.05, but it was slightly conservative. In this example, at least 100 observations were sufficient to achieve power levels above 80% (Figure 12). The results demonstrate that the modified test statistic with the response information $\hat{z}_{P,4}$ is also able to control the type I error, and it achieves reasonable power levels similar to $\hat{z}_4$.

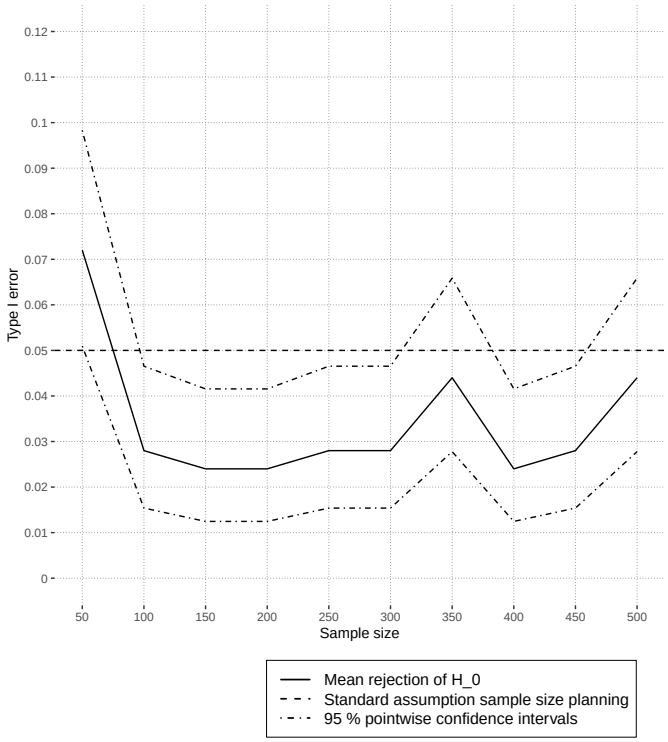

Figure 11. Estimated alpha error of the one-sided interaction test $\hat{z}_{P,4}$ based on MARS with $s = 1$. The dashed line represents a standard alpha level of 0.05 assumed in sample size planning. The dotted–dashed lines represent the upper and lower bounds of the exact pointwise 0.95 Clopper–Pearson confidence intervals.

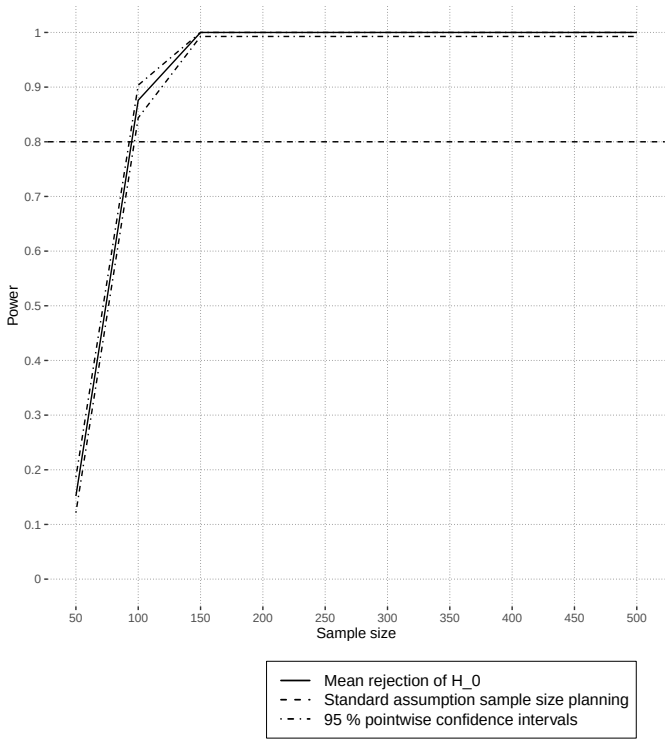

Figure 12. Estimated power of the one-sided interaction test $\hat{z}_{P,4}$ based on MARS with $s = 1$. The dashed line represents a standard power level of 0.8 assumed in sample size planning. One hundred observations suffice for acceptable power levels. The dotted–dashed lines represent the upper and lower bounds of the exact pointwise 0.95 Clopper–Pearson confidence intervals.

5. Data Analysis

This section summarizes the results of the data analysis example. The Boston Housing prices data set from the US census in 1970 [48] was explored for comparison to the data set investigated by [22]. The median value of owner-occupied homes in 1000s of USD was the quantitative response. All available other variables were used as covariates in an extreme gradient-boosting model [49]. The data set was split randomly into tuning data (50%) and a test data set (50%). The tuning data were split again with five times repeated 25-fold cross-validation to tune all possible pairs of the number of boosting iterations $1000, 1100, \ldots, 2000$ and the maximal tree depth $1, 2, \ldots, 14$. The learning rate was set constant to 0.01, and subsampling of the rows and columns was done with a probability of 0.5. The tuning parameters with the lowest MSE were 2000 boosting iterations and a tree depth of 4. Let the performance measure $\xi(M)$ be the average absolute prediction error of the model M divided by the average absolute prediction error of the median response. Evaluating $\xi(M)$ on the test set with the model results in 0.4323. Note that the mean of $\xi(M)$ over all tuning grid values, 0.3407, was comparable to the results of [22]. Testing the null hypothesis of no interaction between all covariates gave a p-value of 0.0107. Thus, interaction effects have an impact. To assess which covariates contribute to interaction effects, all sets $[s = \{1\}, s = \{2\}, \ldots, s = \{14\}]$ were investigated in Figure 13. All covariates above or below the dashed line per capita crime rate by town (CRM), nitric oxides concentration with parts per 10 million (NOX), average number of rooms per dwelling (RM), index of accessibility to radial highways (RAD), full-value property tax rate per 10,000 USD (TAX), and the pupil–teacher ratio by town (PTRATIO) contribute to interaction effects for Boston housing prices. All of those covariates have positive values for the test statistic, which means that those interaction effects overall increase the variability of the prediction model.

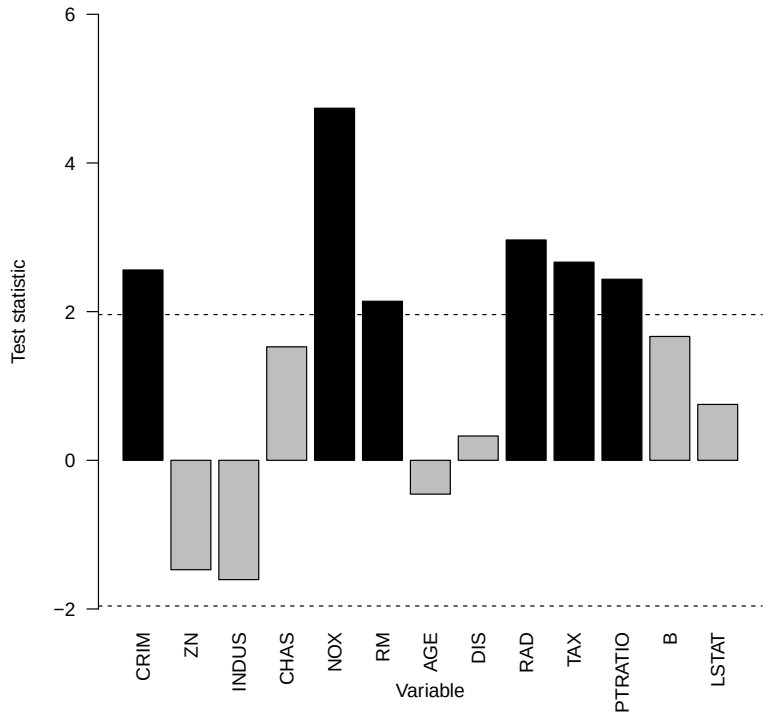

Figure 13. Test statistic z_4 values of the gradient-boosting model for each covariate separately. The black bars highlight the passing significance threshold $\alpha \leqslant 0.05$ of the two-sided test with the null hypothesis that each covariate does not contribute to interaction effects. The dotted lines indicate positive and negative H_0 rejection thresholds.

In the next step, the impact of the previously identified covariate interaction effects can be evaluated. First, covariates with interaction effects $s = \{1, 5, 6, 9, 10, 11\}$ were tested one-sided with the null hypothesis that the prediction model with possible interaction

effects has an equal or higher MSE. Overall, the p-value was 0.0155, and we concluded that the interaction effects of those covariates reduce the MSE. The MSE was reduced by 5.46% relative to the prediction model without interaction effects. The next question is: Which interaction effects associated covariates are responsible for this reduction? It is answered in Figure 14. In this particular case of Boston housing prices, interaction effects with covariates NOX, RAD, TAX, and PTRATIO led to statistically significant MSE improvements in the prediction model. This means that the covariates influence the Boston Housing prices with two-way or higher-order interaction effects, and those identified interaction effects improve the prediction performance.

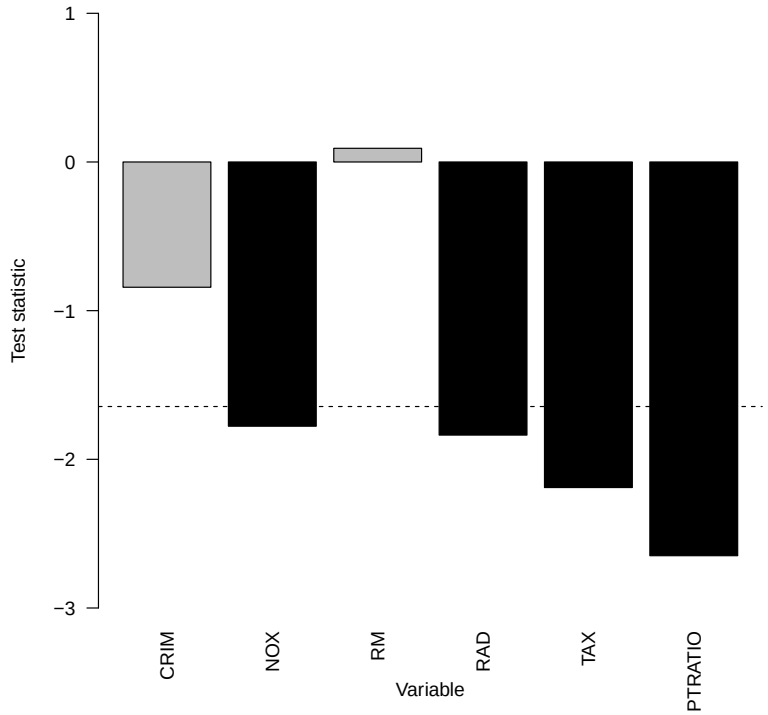

Figure 14. Test statistic z_4 values of the gradient-boosting model for each covariate separately. The black bars highlight the passing significance threshold $\alpha \leqslant 0.05$ of the one-sided test with the null hypothesis that interaction effects associated with a specific covariate do not contribute to MSE reduction. Dotted lines indicate H_0 rejection thresholds.

6. Discussion

This work introduced a model-agnostic statistical interaction test that a hypothesis set can be flexibly specified. An asymptotic distribution of the test statistic was derived (Section 3). The interaction test neither required the refitting of the prediction model nor the resampling of the original data. The low computational runtime cost of the interaction test allows for the exploration of multiple sets of covariates. Our recommendation is to evaluate the test statistic with test data. The distribution of the test statistic behaved well in linear models even in the case of strong covariate correlations (Section 4.1). Simulations with linear (Section 4.2) and nonlinear models (Section 4.3) show that, overall, the type I error is bounded by the prespecified alpha level in most cases and that the test achieves reasonable power levels for several hundred observations in the simulations. The interaction test can be used for black-box models along with other measures of interpretability to better understand interaction effects. Low deviations of the test statistic from zero may indicate that the prediction model could be approximated well using a simpler model without covariate interaction effects in set s.

In addition to Section 3, the evaluation of $\hat{z}_4$ under the training data $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ was discussed. In this case, the observations $\hat{z}_{3,1}, \hat{z}_{3,2}, \ldots, \hat{z}_{3,n}$ are dependent because each observed value of $\hat{z}_{3,i}$ includes all training data in the estimation of the PD function in

$\hat{z}_{1,i}, \hat{z}_{2,i}$. The prediction model $\hat{f}(\boldsymbol{x})$ is not constant and changes if the training sample size increases because it is estimated from the same data. As such, the uniform convergence speed of $\hat{f}(\boldsymbol{x})$ and the PD functions $\widehat{\text{PD}}(\boldsymbol{x}_s)$ would need to be faster than $n^{-1/2}$, which corresponds to the convergence speed of the mean according to the Berry–Essens theorem (see, for example, [50]). However, especially nonparametric machine learning models usually have a lower convergence speed than $n^{-1/2}$ [51], and there is no guarantee that multiplications of $\hat{z}_1, \hat{z}_2$ in $\hat{z}_3$ yield faster convergence rates. Additionally, the CLT would require extensions to work under dependence between observations such as those presented in [52,53]. That specific theory would require the supremum of the maximal correlation coefficient (SMCC) [54] for all possible sets of observations $\hat{z}_{3,i_1}, \hat{z}_{3,i_2}$ with lag $\mathcal{L} = |i_1 - i_2|$ to converge at least linearly to zero as $\mathcal{L} \to \infty$. This assumption is difficult to investigate with simulations and, to the best of the authors' knowledge, impossible to prove because the number of available observations with a specific lag depends on the sample size, while the supremum of the maximal correlation depends on the number of comparisons. Note that, in the case of iid random variables, higher dimensions of the covariate matrix (more comparisons) affect the distribution of the maximal estimated Pearson correlation (see [55] for asymptotic results).

Whether to use $\hat{z}_4$ or $\hat{z}_{P,4}$ with a response should be decided according to the goals of data analysis. The choice may also consider the characteristics of the data-generating process of the application. For example, if the signal-to-noise ratio is low, then $\hat{z}_4$ would be preferable to $\hat{z}_{P,4}$ regarding statistical power because, in this case, the usage of the response information would add more noise that would make it harder to differentiate between H_0 and H_1. In the reverse situation with a high signal-to-noise ratio, the additional information of the response in $\hat{z}_{P,4}$ could reduce the variability of the terms $\text{IADP}_{f,s}$ and $\text{IADP}_{PD,s}$, and thus, hypotheses H_0 and H_1 could be more easily distinguished compared to the test statistic $\hat{z}_4$. Future research may investigate the behavior of both statistics, $\hat{z}_4, \hat{z}_4 P, 4$, in other settings that were not considered in this work (for example, other data sets and different black-box prediction models).

From a general perspective, the choice of whether to apply IML to training or test data depends on the goals of statistical analysis [56]. If the influence of covariates on the prediction model at the population level is the focus of interest, it does not matter whether training or test data are used, as long as data sets originate from the same data-generating process. The more data are available, the more powerful the proposed interaction test is, provided that all other conditions stay constant. In contrast, if the goal is to analyze the impact of covariates on prediction performance, then it is reasonable to apply IML methods to test data sets. This is in line with [18], who recommends the usage of test data in the case of permutation variable importance. Test data usage in the interaction difference test has better theoretical properties and, thus, is recommended for applications.

An alternative to H_s^2 was proposed by [57] that uses accumulated local effect functions instead of PD functions. ALE curves are more computationally efficient and avoid the extrapolation problem to non-observed covariate combinations. However, ALE curves attribute part of the interaction effect to the main effect if there are interactions between correlated features [58]. Extrapolations can be investigated graphically via the stratification of PD plots regarding other covariates. Furthermore, PD plots can be enhanced using individual conditional expectation curves [28], which plot each observed predicted value to investigate variability and possible interaction effects. This graphical representation is not available for ALE. Therefore, this paper focused on the analysis of PD functions.

7. Conclusions

This work has proposed a new model-agnostic hypothesis test to detect interaction effects in prediction models. The null hypothesis states that a given set of covariates does not contribute to any interaction effects. The concept is based on the interaction difference between the variances of the original model predictions and predictions under restricted interaction effects with the null hypothesis. The restricted form of the prediction model is

given via functional ANOVA decomposition, combined with partial dependence functions. The interaction difference was then embedded into the framework of a two-sided, one-sample Z-test. The resulting test statistic is asymptotically normally distributed if it is evaluated using test data. Various simulations showed that, in most cases, the type I error was controlled, and several hundred observations yielded reasonable power levels.

The extended test statistic $\hat{z}_{P,A}$ was explored to incorporate response information into $\hat{z}_4$. If interaction effects were detected with $\hat{z}_4$, the modification $\hat{z}_{P,A}$ could be used to assess whether these interaction effects contributed to MSE prediction performance. In this case, the null hypothesis is that the MSE of the original model with interaction effects is equal to or worse than the prediction model without those interaction effects.

Overall, this work has extended the existing IML methodology to better explain black-box prediction models' interaction effects. It is computationally run time-efficient due to the derived asymptotic distribution and available on CRAN as the R-package `IADT`.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/make6020061/s1: Supplementary materials with the R source code to this article are available online https://www.imbie.uni-bonn.de/cloud/index.php/s/DACosJQ2N8Df9pD (accessed on 26 May 2024).

Author Contributions: T.W.: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, visualization, and writing—original draft. D.E.: methodology, supervision, validation, and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in the study are openly available in the R-package *mlbench* 2.1-3.1. The help file is available with the command *?BostonHousing*. It was originally published by [48].

Acknowledgments: Special thanks are extended to Matthias Schmid for fruitful discussions about asymptotic statistics, IML, and machine learning, as well as proofreading earlier versions of the manuscript.

Conflicts of Interest: There are no conflicts of interests/competing interests to declare.

References

1. Clarke, B.S.; Clarke, J.L. *Predictive Statistics*; Cambridge University Press: Cambridge, UK, 2018. [CrossRef]
2. Efron, B. Prediction, Estimation, and Attribution. *J. Am. Stat. Assoc.* **2020**, *115*, 636–655. [CrossRef]
3. Buehlmann, P. Invariance, Causality and Robustness. *Stat. Sci.* **2020**, *35*, 404–426. [CrossRef]
4. Murdoch, W.J.; Singh, C.; Kumbier, K. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef] [PubMed]
5. Bunge, M. A general black box theory. *Philos. Sci.* **1963**, *30*, 346–358. [CrossRef]
6. Anderson, R. The Rashomon Effect and Communication. *Can. J. Commun.* **2016**, *41*, 249–270. [CrossRef]
7. Wright, J.P. *Hume's 'A Treatise of Human Nature': An Introduction*; Cambridge University Press: Cambridge, UK, 2009.
8. Grünbaum, A. *Can Theories be Refuted? Essays on the Duhem-Quine Thesis*; Chapter The Duhemian Argument; Springer: Dordrecht, The Netherlands, 1976; pp. 116–131. [CrossRef]
9. James, W. *Pragmatism: A New Name for Some Old Ways of Thinking*; Project Gutenberg: Salt Lake City, UT, USA, 1922.
10. Breiman, L. Statistical Modelling: The Two Cultures. *Stat. Sci.* **2001**, *16*, 199–231. [CrossRef]
11. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the KDD '15: 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 1721–1730.
12. Choi, E.; Bahadori, M.T.; Kulas, J.A.; Schuetz, A.; Stewart, W.F.; Sun, J. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3512–3520.
13. Lakkaraju, H.; Bach, S.H.; Leskovec, J. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1675–1684. [CrossRef]
14. Dziugaite, G.K.; Ben-David, S.; Roy, D.M. Enforcing Interpretability and its Statistical Impacts: Trade-offs between Accuracy and Interpretability. *arXiv* **2020**, arXiv:cs.LG/2010.13764.

15. Retzlaff, C.O.; Angerschmid, A.; Saranti, A. Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cogn. Syst. Res.* **2024**, *86*, 101243. [CrossRef]
16. Ferry, J.; Aivodji, U.; Gambs, S.; Huguet, M.J.; Siala, M. Improving fairness generalization through a sample-robust optimization method. *Mach. Learn.* **2023**, *112*, 2131–2192. [CrossRef]
17. Mukhopadhyay, S. InfoGram and admissible machine learning. *Mach. Learn.* **2022**, *111*, 205–242. [CrossRef]
18. Molnar, C. *Interpretable Machine Learning*; Leanpub: Victoria, BC, Canada, 2023. Available online: https://christophm.github.io/interpretable-ml-book/ (accessed on 26 May 2024).
19. Burns, C.; Thomason, J.; Tansey, W. Interpreting Black Box Models via Hypothesis Testing. In Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference FODS'20, Seattle, WA, USA, 18–20 October 2020; pp. 47–57. [CrossRef]
20. Messner, W. From black box to clear box: A hypothesis testing framework for scalar regression problems using deep artificial neural networks. *Appl. Soft Comput.* **2023**, *146*, 110729. [CrossRef]
21. Carvalho, D.; Pereira, E.; Cardoso, J. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* **2019**, *8*, 832. [CrossRef]
22. Friedman, J.H.; Popescu, B.E. Predictive learning via rule ensembles. *Ann. Appl. Stat.* **2008**, *2*, 916–954. [CrossRef]
23. Rao, C.R.; Toutenburg, H. *Linear Models: Least Squares and Alternatives*, 2nd ed.; Springer: New York, NY, USA, 1999. [CrossRef]
24. Hooker, G. Discovering Additive Structure in Black Box Functions. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '04, Seattle, WA, USA, 22–25 August 2004; pp. 575–580. [CrossRef]
25. Hooker, G. Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *J. Comput. Graph. Stat.* **2007**, *16*, 709–732. [CrossRef]
26. Bischl, B.; Binder, M.; Lang, M.; Pielok, T.; Richter, J.; Coors, S.; Thomas, J.; Ullmann, T.; Becker, M.; Boulesteix, A.L.; et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Min. Knowl. Discov.* **2023**, *13*, e1484. [CrossRef]
27. Greenwell, B.M.; Boehmke, B.C.; McCarthy, A.J. A Simple and Effective Model-Based Variable Importance Measure. *arXiv* **2018**, arXiv:stat.ML/1805.04755.
28. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *J. Comput. Graph. Stat.* **2015**, *24*, 44–65. [CrossRef]
29. Herbinger, J.; Bischl, B.; Casalicchio, G. REPID: Regional Effect Plots with implicit Interaction Detection. In *Proceedings of Machine Learning Research, Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, Virtual, 28–30 March 2022*; Camps-Valls, G., Ruiz, F.J.R., Valera, I., Eds.; PMLR: 2022; Volume 151, pp. 10209–10233.
30. Henninger, M.; Debelak, R.; Rothacher, Y.; Strobl, C. Interpretable machine learning for psychological research: Opportunities and pitfalls. *Psychol. Methods* **2023**. [CrossRef] [PubMed]
31. Molnar, C.; König, G.; Herbinger, J.; Freiesleben, T.; Dandl, S.; Scholbeck, C.A.; Casalicchio, G.; Grosse-Wentrup, M.; Bischl, B. *xxAI—Beyond Explainable AI*; Chapter General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models; Springer International Publishing: Cham, Switzerland, 2022; pp. 39–68. [CrossRef]
32. Watson, D.S. Conceptual challenges for interpretable machine learning. *Synthese* **2022**, *200*, 65. [CrossRef]
33. Mayo, D.G. *Statistical Inference as Severe Testing*; Cambridge University Press: Cambridge, UK, 2018. [CrossRef]
34. Grambsch, P.M. Simple robust tests for scale differences in paired data. *Biometrika* **1994**, *81*, 359–372. [CrossRef]
35. Pitman, E.J.G. A Note on Normal Correlation. *Biometrika* **1939**, *31*, 9–12. [CrossRef]
36. Royston, P. Algorithm AS 181: The W Test for Normality. *J. R. Stat. Soc. Ser. C* **1982**, *31*, 176–180. [CrossRef]
37. Razali, N.; Wah, Y. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model. Anal.* **2011**, *2*, 21–33.
38. Benjamini, Y.; Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **2001**, *29*, 1165–1188. [CrossRef]
39. Clopper, C.J.; Pearson, E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **1934**, *26*, 404–413. [CrossRef]
40. Vanschoren, J.; van Rijn, J.N.; Bischl, B.; Torgo, L. OpenML: Networked Science in Machine Learning. *ACM SIGKDD Explor. Newsl.* **2014**, *15*, 49–60. [CrossRef]
41. Dua, D.; Graff, C. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2017. Available online: http://archive.ics.uci.edu/ml (accessed on 26 May 2024).
42. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 1–48. Available online: https://openreview.net/ (accessed on 26 May 2024).
43. Bojer, C.S.; Meldgaard, J.P. Kaggle forecasting competitions. *Int. J. Forecast.* **2021**, *37*, 587–603. [CrossRef]
44. Probst, P.; Wright, M.N.; Boulesteix, A.L. Hyperparameters and tuning strategies for random forest. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1301. [CrossRef]
45. Fahrmeir, L.; Kuenstler, R.; Pigeot, I.; Tutz, G. *Statistik—Der Weg zur Datenanalyse*, 8th ed.; Springer: Berlin/Heidelberg, Germany, 2016. [CrossRef]
46. Lenth, R.V. Some Practical Guidelines for Effective Sample Size Determination. *Am. Stat.* **2001**, *55*, 187–193. [CrossRef]
47. Friedman, J.H. Multivariate Adaptive Regression Splines. *Ann. Stat.* **1991**, *19*, 1–67. [CrossRef]

48. Harrison, D.; Rubinfeld, D.L. Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manag.* **1978**, *5*, 81–102. [CrossRef]
49. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
50. Shorack, G.A. *Probability for Statisticians*; Springer: New York, NY, USA, 2000. [CrossRef]
51. Hall, P. On Convergence Rates in Nonparametric Problems. *Int. Stat. Rev.* **1989**, *57*, 45–58. [CrossRef]
52. Doukhan, P.; Winterberger, O. An invariance principle for weakly dependent stationary general models. *Probab. Math. Stat.* **2007**, *27*, 45–73. [CrossRef]
53. Dedecker, J.; Doukhan, P.; Lang, G.; León, R.J.R.; Louhichi, S.; Prieur, C. *Weak Dependence*; Springer: New York, NY, USA, 2007. [CrossRef]
54. Renyi, A. On measures of dependence. *Acta Math. Acad. Sci. Hung.* **1959**, *10*, 441–451. [CrossRef]
55. Ding, X. Limit Properties of the Largest Entries of High-Dimensional Sample Covariance and Correlation Matrices. *Hindawi Math. Probl. Eng.* **2021**, *2021*, 8. [CrossRef]
56. Altmann, T.; Bodensteiner, J.; Dankers, C.; Dassen, T.; Fritz, N.; Gruber, S.; Kopper, F.; Kronseder, V.; Wagner, M.; Renkl, E. *Limitations of Interpretable Machine Learning Methods*; Leanpub: Victoria, BC, Canada, 2020. Available online: https://slds-lmu.github.io/iml_methods_limitations/ (accessed on 26 May 2024).
57. Molnar, C.; Casalicchio, G.; Bischl, B. Quantifying Interpretability of Arbitrary Machine Learning Models Through Functional Decomposition. Technical Report. *arXiv* **2019**, arXiv:1904.03867.
58. Groemping, U. Model-Agnostic Effects Plots for Interpreting Machine Learning Models. Technical Report 1, Beuth Hochschule füer Technik Berlin, Reports in Mathematics, Physics and Chemistry. 2020. Available online: http://www.data2intelligence.de/BHT_FBII_reports/Report-2020-001.pdf (accessed on 26 May 2024).