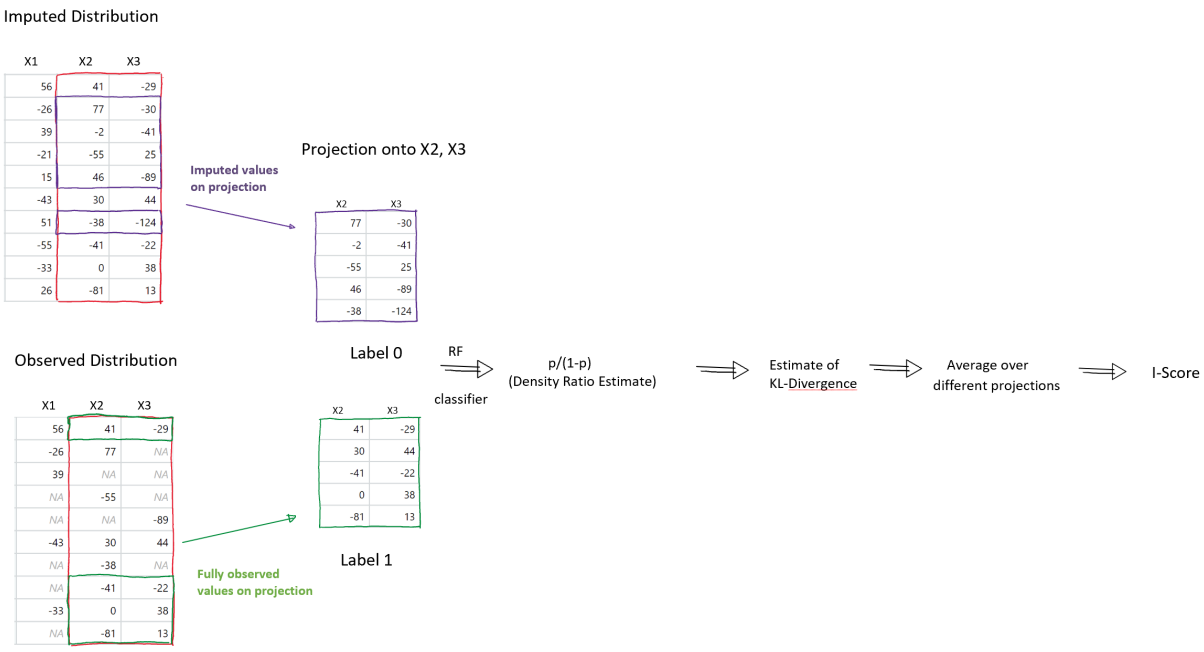
Distributional Prediction, Missing Values, and Tree Ensembles

Jeffrey Näf,

Diss. ETH No. 29019



Diss. ETH No. 29019

Distributional Prediction, Missing Values, and Tree Ensembles

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

Jeffrey Näf

MSc Statistics, ETH Zurich
born June 10, 1989
citizen of Schaffhausen, Switzerland

accepted on the recommendation of
Prof. Dr. Nicolai Meinshausen, examiner
Prof. Dr. Peter Bühlmann, co-examiner

2023

DOI: ???

Abstract

In **Paper A**, we introduce a framework for the construction of high-probability lower bounds on the total variation distance. These bounds are based on a one-dimensional projection, such as a classification or regression method, and can be interpreted as the minimal fraction of samples pointing towards a distributional difference. We further derive asymptotic power and detection rates of two proposed estimators and discuss potential uses through an application to a reanalysis climate dataset.

In **Paper B** we develop a framework called "Imputation Scores" (I-Scores) for assessing missing value imputations. We provide a specific I-Score based on density ratios and projections, that is applicable to discrete and continuous data. It does not require to mask additional observations for evaluations and is also applicable if there are no complete observations. The population version is shown to be proper in the sense that the highest rank is assigned to an imputation method that samples from the correct conditional distribution. The propriety is shown under the *missing completely at random* (MCAR) assumption but is also shown to be valid under *missing at random* (MAR) with slightly more restrictive assumptions. We show empirically on a range of data sets and imputation methods that our score consistently ranks true data high(est) and is able to avoid pitfalls usually associated with performance measures such as RMSE.

In **Paper C**, we develop a fully non-parametric, easy-to-use, and powerful test for the missing completely at random (MCAR) assumption on the missingness mechanism of a data set. The test compares distributions of different missing patterns on random projections in the variable space of the data. The distributional differences are measured with the Kullback-Leibler Divergence, using probability Random Forests [111]. We thus refer to it as "Projected Kullback-Leibler MCAR" (PKLM) test. The use of random projections makes it applicable even if very few or no fully observed observations are available or if the number of dimensions is large. An efficient permutation approach guarantees the level for any finite sample size, resolving a major shortcoming of most other available tests. Moreover, the test can be used on both discrete and continuous data. We show empirically on a range of simulated data distributions and real data sets that our test has consistently high power and is able to avoid inflated type I errors.

In **Paper D** we propose a novel (random) forest construction for multivariate responses based on their joint conditional distribution, independent of the estimation target and the data

model. It uses a new splitting criterion based on the MMD distributional metric, which is suitable for detecting heterogeneity in multivariate distributions. The induced weights define an estimate of the full conditional distribution, which in turn can be used for arbitrary and potentially complicated targets of interest. The method is very versatile and convenient to use, as we illustrate on a wide range of examples.

Zusammenfassung

In **Paper A** entwickeln wir einen Rahmen für die Konstruktion von mit hoher Wahrscheinlichkeit unteren Schranken für die Total Variation (TV) Distance. Diese Schranken basieren auf einer eindimensionalen Projektion, z. B. einer Klassifizierungs- oder Regressionsmethode, und können als der minimale Anteil der Stichproben interpretiert werden, der auf einen Verteilungsunterschied hinweist. Wir leiten ferner die asymptotische Macht und Erkennungsraten von zwei vorgeschlagenen Schätzern her und diskutieren eine Anwendung auf einen Reanalyse-Klimadatensatz.

In **Paper B** entwickeln wir die "Imputation Scores" (I-Scores) zur Bewertung fehlender Werte. Wir stellen einen spezifischen I-Score basierend auf Dichteverhältnissen und Projektionen bereit, der auf diskrete und kontinuierliche Daten anwendbar ist. Es erfordert keine Maskierung zusätzlicher Beobachtungen für Auswertungen und ist auch anwendbar, wenn keine vollständigen Beobachtungen vorliegen. Die Populationsversion erweist sich insofern als korrekt, als der höchste Rang einer Imputationsmethode zugewiesen wird, die Stichproben aus der korrekten bedingten Verteilung erstellt. Die Korrektheit wird unter der Annahme *missing complete at random* (MCAR) gezeigt, ist aber auch unter *missing at random* (MAR) mit etwas restriktiveren Annahmen gültig. Wir zeigen empirisch anhand einer Reihe von Datensätzen und Imputationsmethoden, dass unser Score echte Daten durchweg am höchsten einstuft und in der Lage ist, Fallstricke zu vermeiden, die normalerweise mit Leistungskennzahlen wie RMSE verbunden sind.

In **Paper C** entwickeln wir einen vollständig nicht-parametrischen, benutzerfreundlichen und leistungsstarken Test für die Annahme von missing completely at random (MCAR) für die fehlenden Werte eines Datensatzes. Der Test vergleicht Verteilungen verschiedener fehlender Muster auf zufällige Projektionen im Variablenraum der Daten. Die Verteilungsunterschiede werden mit der Kullback-Leibler-Divergenz unter Verwendung von Random Forests [111] gemessen. Wir bezeichnen ihn daher als "Projected Kullback-Leibler MCAR" (PKLM)-Test. Die Verwendung zufälliger Projektionen macht es anwendbar, selbst wenn sehr wenige oder keine vollständig beobachteten Beobachtungen verfügbar sind oder wenn die Anzahl der Dimensionen gross ist. Ein effizienter Permutationsansatz garantiert das Niveau für jede endliche Stichprobengrösse und löst einen grossen Mangel der meisten anderen verfügbaren Tests. Darüber

hinaus kann der Test sowohl auf diskrete als auch auf kontinuierliche Daten angewendet werden. Wir zeigen empirisch anhand einer Reihe von simulierten Datenverteilungen und realen Datensätzen, dass unser Test eine konstant hohe Aussagekraft hat und in der Lage ist, überhöhte Typ-I-Fehler zu vermeiden.

In **Paper D** entwickeln wir eine neuartige Random Forest Ansatz für multivariate Zielvariablen basierend auf ihrer gemeinsamen bedingten Verteilung, unabhängig vom Schätzziel und dem Datenmodell. Es verwendet ein neues Aufteilungskriterium basierend auf der MMD-Verteilungsmetrik, das geeignet ist, Heterogenität in multivariaten Verteilungen zu erkennen. Die induzierten Gewichte definieren eine Schätzung der vollständigen bedingten Verteilung, die wiederum für potenziell komplizierte Ziele von Interesse verwendet werden kann. Die Methode ist sehr vielseitig und komfortabel in der Anwendung, wie wir an zahlreichen Beispielen verdeutlichen.

Acknowledgments

First and foremost I want to thank Nicolai Meinshausen for being a fantastic supervisor. I always looked forward to our discussions and when I was stuck, I could always count on his deep insights and thoughtful comments to guide the way. I learned many things from him, but most memorably, that a fine-tuned intuition is an invaluable tool in solving problems.

I also want to thank my collaborators, Meta-Lina Spohn, Loris Michel, Domagoj Cévid and Simon Hediger, for an exhilarating time discussing projects, brainstorming and writing papers together. Every aspect of writing papers and doing a PhD was improved through these collaborations.

I also want to thank all of SfS. In particular, I want to thank my co-advisor Peter Bühlmann for always providing helpful advice and a good atmosphere. It has been a pleasure to collaborate with some of the smartest and nicest people I have ever met. I am surely going to miss this time here at SfS.

Neuhausen, 01.12.2022 Jeffrey Näf

Contents

Accompanying papers

A High-Probability Lower Bounds for the Total Variation Distance.
L. Michel, J. Näf.
arXiv.

B Imputation Scores.
J. Näf, M. Spohn, L. Michel, N. Meinshausen.
To appear in the Annals of Applied Statistics.

C PKLM: A flexible MCAR Test Using Classification
J. Näf, M. Spohn, L. Michel, N. Meinshausen.
Major Revision in Psychometrika.

D Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression.
D. Ćevid, L. Michel, J. Näf, P. Bühlmann, N. Meinshausen.
Journal of Machine Learning Research.

1 Introduction

Random Forest (RF) [15] is a versatile method that has been getting a lot of attention over the past two decades. Consider a random variable of interest Y and covariates $\mathbf{X} = (X_1, \ldots, X_p)$. We observe n identically and independently distributed (i.i.d.) draws from their joint distribution and desire to learn the conditional expectation of Y given $\mathbf{X} = \mathbf{x}$. RFs do this by combining N independent trees, whereby each tree splits the data and places them into leaves according to some criterion. Importantly, the splits themselves are a function only of the covariates, for example, an observation goes in the left leaf if $X_1 > 0.5$ and in the right otherwise. However to determine the best splits, a criterion on the dependent variable Y is used. This allows to derive data adaptive splits from a training set and then predict Y from only seeing a new test point $\mathbf{x}$. As a general intuition, the splits are chosen such that the Y_i's in the right and left leaf are as "different" as possible, according to the chosen splitting criterion. Splitting in such a manner several times, one should end up with a leaf in which the data are as "homogeneous" as possible. In more mathematical terms, given some assumptions, a leaf should contain an approximation to an i.i.d. sample from the conditional distribution. An illustration of this approach is given in Figure 1.1. Despite its generality, and apparent ability to model nonlinear relationships, an important feature of the method is that it generally doesn't require tuning. While there are parameters that could be tuned, especially in generalized versions of RF, such as in GRF of [4], RFs usually lead to quite good results. This entails a clear advantage over other nonparametric methods in practice.

This thesis collects several papers, all deeply connected to the RF methodology. Though most of these papers are ultimately of an applied nature, the thesis will focus on their theoretical side and present and further discuss results derived in these papers. Throughout some crucial notation and explanations will be left out to focus on the results and discuss their main message.

A second important idea that permeates the thesis is that two distributions can be compared using a classifier. That is, if we observe an i.i.d. sample from a distribution P and another i.i.d. sample from a distribution Q, we can use the following procedure to test whether $P = Q$: (1) Give the observations from P a label of 1, and those from Q a label of 0, (2) train a classifier to differentiate the two classes and (3) evaluate the classifier using some accuracy measure on a test-set where the labels are known. Step (3) gives a numerical indication of how well the

classifier can differentiate the classes on independent data. Intuitively, the better the classifier can differentiate the two classes on a test set, the more evidence there is against $P = Q$. Various ways can be used to make this idea formally correct. For instance, if the misclassification error on the test set is used as a measure of the accuracy of the classifier, it can be shown that an exact test can be constructed, see e.g. [68]. Otherwise, one can use a permutation approach; randomly permute the labels 0/1 B times, estimate a new classifier on the permuted labels and evaluate on the test set. It can be shown that under weak assumptions, treating the values obtained from the permuted data set as B independent draws from $H_0 : P = Q$ results in a test with valid p-value. We now give an overview of the papers considered in this thesis. For more details, the newest versions of all papers can be found at the end of the thesis.

In Paper A, we take up and extend the idea of using a classifier for two-sample testing. Instead of just providing a decision to reject or not, we use the classifier testing approach to provide a number between $\hat{\lambda} \in [0, 1]$ that indicates how far P and Q are apart. This number represents a high probability lower bound (HPLB) for the total variation distance between P, Q in the sense that it will overshoot the true value with probability less than α. Since this in particular true when the true value is zero, i.e. $P = Q$, $I\{\hat{\lambda} > 0\}$ constitutes a valid test.

Paper B and C discuss new methods in the field of missing values. In paper B we develop a method that scores different imputation methods, while in Paper C introduce a new kind of missing completely at random (MCAR) test. Both again use the idea of classifier-based divergence measures for P and Q. This is used in Paper B to develop a method to check how well a given imputation of the missing values fits the observed data. Though this is not used for formal testing, the resulting classifier-based value obtained on a test set is used as a score indicating how well an imputation method performs. Paper C on the other hand, develops a formal classifier test for multiple classes to differentiate the distribution of different missingness patterns. The idea is that, under the assumption of missing completely at random (MCAR), whereby the probability of a value being missing is unrelated to any observed or unobserved values, the distributions for different patterns of missingness should be the same. This is developed into a classifier-based test of the MCAR assumption. In both papers, random projections also play a crucial role, as detailed below,

Paper D develops a new kind of Random Forest algorithm, the Distributional Random Forest (DRF). It is an outlier in this thesis in two ways; first, it does not use classifier-based divergence measures, but instead the maximum mean discrepancy (MMD) metric. Second, RF is not applied here to solve a given problem, but instead the algorithm itself is generalized. In particular, the usual CART criterion at each split is now performed in the reproducing kernel Hilbert space (RKHS), using a kernel. This allows to greatly generalize the application of Random Forest, as estimating a mean in the RKHS is akin to estimating a representation of the whole conditional distribution. This allows a practitioner to obtain a range of functionals of

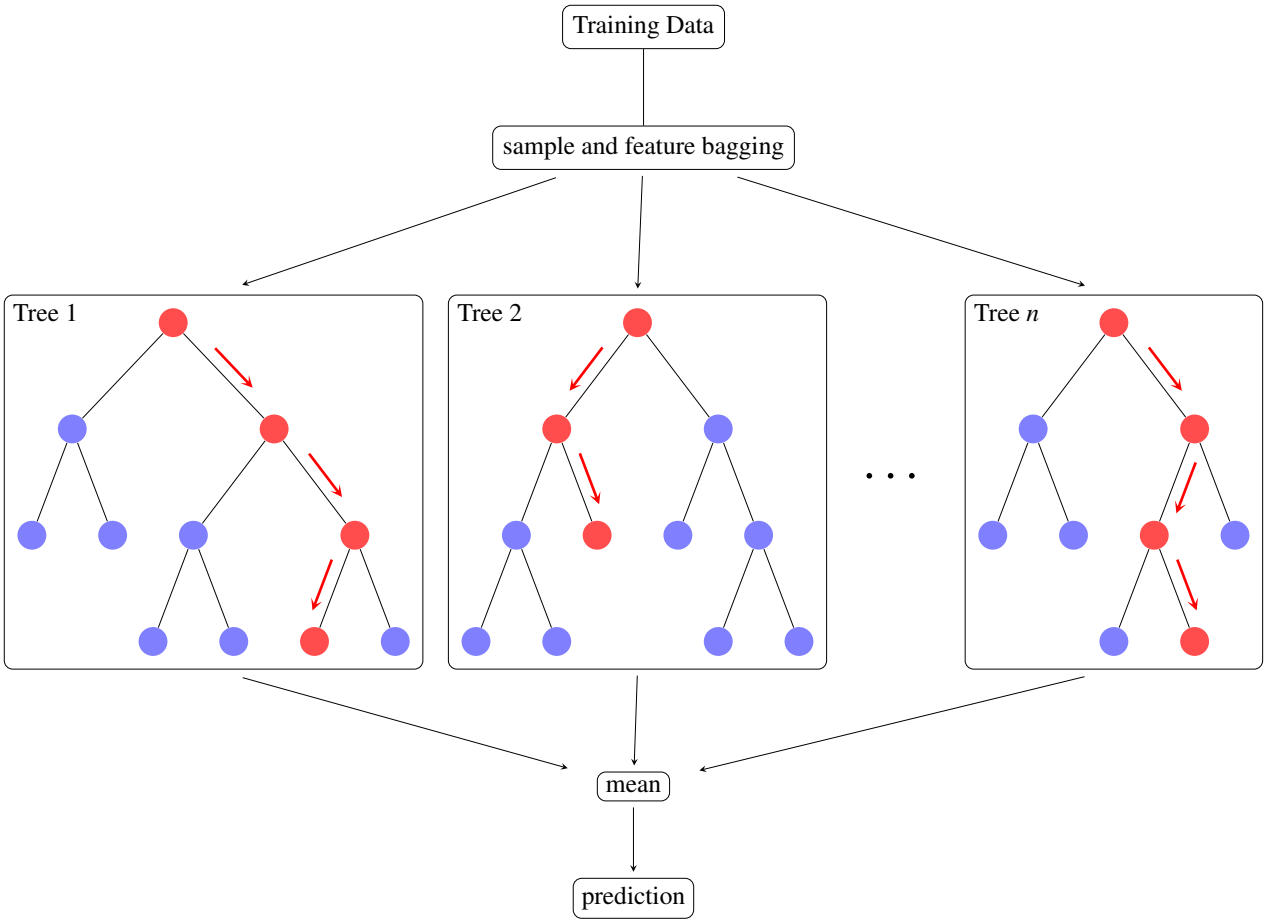


Figure 1.1: Illustration of the workings of a Random Forest, adapted from `https://tikz.net/random-forest/`

interest, even multivariate ones, from one fit of the Random Forest.

2 High Probability Lower Bounds for the Total Variation Distance

2.1 Problem Setup

As mentioned in Chapter 1, one can use a classifier for two-sample testing. That is, for two distributions P, Q, and two i.i.d. samples from each, we can test $H_0 : P = Q$, by using the performance of a classifier as a test statistic. In this paper, we aim to go beyond this and to give an indication of how strong P and Q are different under the alternative.

An interesting measure to quantify the difference between two distributions P and Q is the *total variation (TV) distance*. Assuming P and Q are defined on the measurable space $(\mathcal{X}, \mathcal{A})$ the TV distance is defined as

$$\mathrm{TV}(P, Q) = \sup_{A \in \mathcal{A}} |P(A) - Q(A)| .$$

As $0 \leqslant \mathrm{TV}(P, Q) \leqslant 1$, $\mathrm{TV}(P, Q)$ is a simple to interpret indicator of the difference between P and Q. It also has an interesting intuitive interpretation: $\mathrm{TV}(P, Q)$ is the fraction of mass we need to shift to go from P to Q. Let $\rho : \mathcal{X} \to I$ be any measurable function mapping to $I = [0, 1]$. The following chain of inequalities for $\mathrm{TV}(P, Q)$ will form the starting point of our approach:

$$\mathrm{TV}(P, Q) \geqslant \mathrm{TV}(\rho_{\#}P, \rho_{\#}Q) \geqslant \sup_{t \in I} |\rho_{\#}P((0, t]) - \rho_{\#}Q((0, t])| , \tag{2.1}$$

where $\rho_{\#}P$ is the push-forward measure of P through ρ, that is for all measurable sets A,

$$\rho_{\#}P(A) = P(\rho^{-1}(A)).$$

Connecting back to the concept of testing through classification, any such ρ, one can define a binary classifier $\rho_t : \mathcal{X} \to \{0, 1\}$ based on the cutoff t by $\rho_t(z) := I\{\rho(z) > t\}$. Let $X \sim P$ and $Y \sim Q$ and assume we give a label of 0 to samples from P and a label of 1 to samples from Q.

Then we can rewrite

$$\begin{aligned}\sup_{t\in I}|\rho_{\#}P((0,t]) - \rho_{\#}Q((0,t])| &= \sup_{t\in I}|\mathbb{P}(\rho(X)\leqslant t) + (1-\mathbb{P}(\rho(Y)\leqslant t)) - 1| \\ &= \sup_{t\in I}|\mathbb{P}(\rho_t(X)=0) + \mathbb{P}(\rho_t(Y)=1) - 1| \\ &= \sup_{t\in I}|A_0(t) + A_1(t) - 1|, \end{aligned} \tag{2.2}$$

where $A_j(t)$ is the accuracy of the classifier on class j, $j \in \{0,1\}$. Thus, we can build the best classifier ρ_t from ρ to obtain the lower bound. This is what we describe in the following.

Before diving into more details, let us introduce our setup and some necessary notation. Where not otherwise stated, we assume to observe two independent i.i.d. samples $X_1,\dots,X_m$ from P and $Y_1,\dots,Y_n$ from Q. We define

$$Z_i := \begin{cases} X_i & \text{if } 1\leqslant i\leqslant m, \\ Y_i & \text{if } m+1\leqslant i\leqslant m+n, \end{cases}$$

and attach a label $\ell_i := 0$ for $i = 1,\dots,m$ and $\ell_i := 1$, for $i = m+1,\dots,m+n$. Both m and n are assumed to be non-random with $N = m+n$ such that $m/N \to \pi \in (0,1)$, as $N\to\infty$. For notational convenience, we also assume that $m \leqslant n$. We denote by f and g the densities of P, Q respectively. We define $F := \rho_{\#}P$ and $G := \rho_{\#}Q$ and introduce the empirical measures

$$\hat{F}_m(t) := \frac{1}{m}\sum_{i=1}^{m} I\{\rho(X_i)\leqslant t\}, \quad \hat{G}_n(t) := \frac{1}{n}\sum_{j=1}^{n} I\{\rho(Y_j)\leqslant t\},$$

of all observations $\{\rho_i\}_{i=1,\dots,N} := \{\rho(Z_i)\}_{i=1,\dots,N}$. We can then rewrite (2.2) as

$$\sup_{t\in I}|\rho_{\#}P((0,t]) - \rho_{\#}Q((0,t])| = \sup_{t\in I}|F(t)-G(t)|. \tag{2.3}$$

Denote by $\rho_{(z)}$, $z\in\{1,\dots,N\}$, the z^{th} order statistic of $(\rho_i)_{i=1,\dots,N}$. Throughout, $q_\alpha(p,m)$ is the α-quantile of a binomial distribution with success probability p and number of trials m symbolized by Binomial(p,m). Our analysis centers around the projection $\rho^* : \mathcal{X}\to[0,1]$ given as

$$\rho^*(z) := \frac{g(z)}{f(z)+g(z)}. \tag{2.4}$$

As a remark, if we put a prior probability $\pi = 1/2$ on observing a label ℓ of 1, ρ^* is the posterior probability of observing a draw from Q, referred to as the Bayes probability.

We now formally state the definition of a high-probability lower bound for the total variation distance, using the notation $\lambda := \mathrm{TV}(P,Q)$ from now on:

Definition 2.1. *For a given $\alpha\in(0,1)$, an estimate $\hat{\lambda} = \hat{\lambda}((Z_1,\ell_1),\dots,(Z_N,\ell_N))$ satisfying*

$$\mathbb{P}(\hat{\lambda} > \lambda) \leqslant \alpha \tag{2.5}$$

will be called high-probability lower bound (HPLB) at level α. If instead only the condition

$$\limsup_{N\to\infty} \mathbb{P}(\hat{\lambda} > \lambda) \leqslant \alpha \tag{2.6}$$

holds, we will refer to $\hat{\lambda}$ as asymptotic high-probability lower bound (asymptotic HPLB) at level α.

Thus we require that the true value λ is not overshot with probability $1-\alpha$. Note that an estimator $\hat{\lambda}$ depends on a function ρ. When necessary, this will be emphasized with the notation $\hat{\lambda}^{\rho}$ throughout the text. Whenever ρ is not explicitly mentioned it should be understood that we consider $\rho = \rho^*$.

We now give some more details about the proposed method.

2.2 Proposed Method

In light of relation (2.1), we aim to directly account for the randomness of $\sup_t(\hat{F}_m(t)-\hat{G}_n(t)) = \sup_z(\hat{F}_m(\rho_{(z)})-\hat{G}_n(\rho_{(z)}))$. We define the counting function $V_{m,z} = m\hat{F}_m(\rho_{(z)})$ for each $z \in J_{m,n} := \{1,\dots,m+n-1\}$. Using $m\hat{F}_m(\rho_{(z)}) + n\hat{G}_n(\rho_{(z)}) = z$, it is possible to write:

$$\hat{F}_m(\rho_{(z)}) - \hat{G}_n(\rho_{(z)}) = \frac{m+n}{mn}\left(V_{m,z} - \frac{mz}{m+n}\right). \tag{2.7}$$

A well-know fact (see e.g., [52]) is that under $H_0 : F = G$, $V_{m,z}$ is a hypergeometric random variable, obtained by drawing without replacement z times from an urn that contains m circles and n squares and counting the number of circles drawn. We denote this as $V_{m,z} \sim \text{Hypergeometric}(z, m+n, m)$ and simply refer to the resulting process $z \mapsto V_{m,z}$ as the *hypergeometric process*. As the distribution of $V_{m,z}$ under a general alternative is not known, the discussion in this chapter focuses on the result that one can nonetheless control its behavior, at least asymptotically. We start with the following definition, inspired by [119]:

Definition 2.2 (Bounding function)**.** *A function $J_{m,n} \times [0,1] \ni (z,\tilde{\lambda}) \mapsto Q_{m,n,\alpha}(z,\tilde{\lambda})$ is called a* bounding function *at level α if*

$$\mathbb{P}(\sup_{z\in J_{m,n}} [V_{m,z} - Q_{m,n,\alpha}(z,\lambda)] > 0) \leqslant \alpha. \tag{2.8}$$

It will be called an asymptotic bounding function *at level α if instead*

$$\limsup_{N\to\infty} \mathbb{P}(\sup_{z\in J_{m,n}} [V_{m,z} - Q_{m,n,\alpha}(z,\lambda)] > 0) \leqslant \alpha. \tag{2.9}$$

In other words, for the true value λ, $Q_{m,n,\alpha}(z,\lambda)$ provides an (asymptotic) type 1 error control for the process $z \mapsto V_{m,z}$ (often the dependence on α will be omitted). For $\lambda = 0$ one can obtain a closed-form expression of an asymptotic bounding function from the theory in [52].

Assuming access to an (asymptotic) bounding function $Q_{m,n,\alpha}$, we can define the following estimator,

$$\hat{\lambda}^{\rho} = \inf \left\{ \tilde{\lambda} \in [0,1] : \sup_{z \in J_{m,n}} \left[V_{m,z} - Q_{m,n,\alpha}(z, \tilde{\lambda}) \right] \leqslant 0 \right\}. \tag{2.10}$$

In words, we are looking for the smallest candidate $\tilde{\lambda}$ such that the bounding function $Q_{m,n,\alpha}(z, \tilde{\lambda})$ is larger than $V_{m,z}$ for all z. We will see in the next section why this works.

Since we do not know the true λ, the main challenge in the following is to find bounding functions that would be valid for any potential $\lambda \geqslant 0$. We now introduce a particular type of such a bounding function. With $\tilde{\lambda} \in [0,1]$, $\alpha \in (0,1)$, $m(\tilde{\lambda}) = m - q_{1-\frac{\alpha}{3}}(\tilde{\lambda}, m)$, and $n(\tilde{\lambda}) = n - q_{1-\frac{\alpha}{3}}(\tilde{\lambda}, n)$, we define

$$Q_{m,n,\alpha}(z, \tilde{\lambda}) = \begin{cases} z, \text{ if } 1 \leqslant z \leqslant q_{1-\frac{\alpha}{3}}(\tilde{\lambda}, m) \\ m, \text{ if } m + n(\tilde{\lambda}) \leqslant z \leqslant m + n \\ q_{1-\frac{\alpha}{3}}(\tilde{\lambda}, m) + \mathrm{sq}_{\alpha/3}\left(\tilde{V}_{m(\tilde{\lambda}),z}, z \in \{1, \ldots, m(\tilde{\lambda}) + n(\tilde{\lambda}) - 1\}\right), \text{ otherwise} \end{cases} \tag{2.11}$$

where $\tilde{V}_{m,z}$ denotes the counting function of a hypergeometric process and $\mathrm{sq}_{\alpha}\left(\tilde{V}_{m,z}, z \in J\right)$ is a *simultaneous* confidence band, such that

$$\limsup_{N \to \infty} \mathbb{P}\left(\sup_{z \in J} \left[\tilde{V}_{m,z} - \mathrm{sq}_{\alpha}\left(\tilde{V}_{m,z}, z \in J\right) \right] > 0 \right) \leqslant \alpha. \tag{2.12}$$

We show below that $Q_{m,n,\alpha}$ in (2.11) is really a bounding function. Note that Equation (2.12) includes the case $\mathbb{P}\left(\sup_{z \in J}\left[\tilde{V}_{m,z} - \mathrm{sq}_{\alpha}\left(\tilde{V}_{m,z}, z \in J\right)\right] > 0\right) \leqslant \alpha$ for all N. Depending on which condition is true, we obtain a bounding function or an asymptotic bounding function.

2.3 Results

There are many results in the paper, but we focus on the fact that the level can be guaranteed in this complex setting:

Proposition 2.3. *The estimator $\hat{\lambda}^{\rho}$ is an (asymptotic) HPLB of λ (at level α) for any $\rho : \mathcal{X} \to I$.*

The estimator $\hat{\lambda}^{\rho}$ looks intimidating, but combining the definition of bounding function with the inf properties, Proposition 2.3 follows actually quite easily. It is instructive to see why. Let,

$$G_{m,n} := \left\{ \tilde{\lambda} \in [0,1] : \sup_{z \in J_{m,n}} \left[V_{m,z} - Q_{m,n,\alpha}(z, \tilde{\lambda}) \right] \leqslant 0 \right\}.$$

Then by definition of the infimum, $\hat{\lambda}^\rho > \lambda$ implies $\lambda \in G^c_{m,n}$, or

$$\begin{aligned}\mathbb{P}(\hat{\lambda}^\rho > \lambda) &\leqslant \mathbb{P}(\lambda \in G^c_{m,n}) \\ &= \mathbb{P}(\sup_{z \in J_{m,n}} [V_{m,z} - Q_{m,n,\alpha}(z,\lambda)] > 0).\end{aligned}$$

But by the definition of $Q_{m,n,\alpha}$, this probability is bounded by α. Thus the challenge really lies in finding a valid bounding function. We now establish that:

Proposition 2.4. *$Q_{m,n,\alpha}$ as defined in* (2.11) *is an (asymptotic) bounding function.*

Crucially, the hypergeometric process $\tilde{V}_{m,z}$ has the same distribution, no matter the initial distributions P and Q. This holds more generally for the bounding function $Q_{m,n,\alpha}$ which does not depend on P and Q and is thus *distribution-free*.

We now discuss two interesting concepts that can be used in the proof of Proposition 2.4. The first concept is that of "Distributional Witnesses". We assume to observe two i.i.d. samples of independent random elements X, Y with values in $(\mathcal{X}, \mathcal{A})$ with respective probability measures P and Q. Similar as in [36], let $\mathfrak{C}$ be the set of all random elements $(\tilde{X}, \tilde{Y})$ with values in $(\mathcal{X}^2, \mathcal{A}^2)$, and such that $\tilde{X} \sim P$ and $\tilde{Y} \sim Q$. Following standard convention, we call $(\tilde{X}, \tilde{Y}) \in \mathfrak{C}$ a *coupling* of P and Q. Then $\mathrm{TV}(P, Q)$ may be characterized as

$$\mathrm{TV}(P, Q) = \inf_{\mathfrak{C}} \mathbb{P}(\tilde{X} \neq \tilde{Y}). \tag{2.13}$$

This is in turn equivalent to saying that we minimize $\Pi(x \neq y)$ over all joint distributions Π on $(\mathcal{X}^2, \mathcal{A}^2)$, that have $X_\# \Pi = P$ and $Y_\# \Pi = Q$. Equation (2.13) allows for an interesting interpretation, as detailed (for example) in [36]: The optimal value is attained for a coupling (X^*, Y^*) that minimizes the probability of $X^* \neq Y^*$. The probability that they are different is exactly given by $\mathrm{TV}(P, Q)$. It is furthermore not hard to show that the optimal coupling is given by the following scheme: Let $W \sim \mathrm{Bernoulli}(\mathrm{TV}(P, Q))$ and denote by f the density of P and g the density of Q, both with respect to some measure on $(X, \mathcal{A})$, e.g. $P + Q$. If $W = 0$, draw a random element Z from a distribution with density $\min(f, g)/(1 - \mathrm{TV}(P, Q))$ and set $X^* = Y^* = Z$. If $W = 1$, draw X^* and Y^* independently from $(f - g)_+/\mathrm{TV}(P, Q)$ and $(g - f)_+/\mathrm{TV}(P, Q)$ respectively.

Obviously, X^* and Y^* so constructed are dependent and do not directly relate to the observed X, Y, which are assumed to be independent. However it holds true that marginally, $X \overset{D}{=} X^*$ and $Y \overset{D}{=} Y^*$. In particular, given that $W = 1$, it holds that $X \overset{D}{=} X^* = Y^* \overset{D}{=} Y$, or $X|\{W = 1\} \overset{D}{=} Y|\{W = 1\}$. On the other hand, for $W = 0$, the support of X and Y is disjoint. This suggests that the distribution of X and Y might be split into a part that is common to both and a part that is unique. Indeed, the probability measures P and Q can be decomposed in terms of three probability measures H_P, H_Q, $H_{P,Q}$ such that

$$P = \lambda H_P + (1 - \lambda) H_{P,Q} \quad \text{and} \quad Q = \lambda H_Q + (1 - \lambda) H_{P,Q}, \tag{2.14}$$

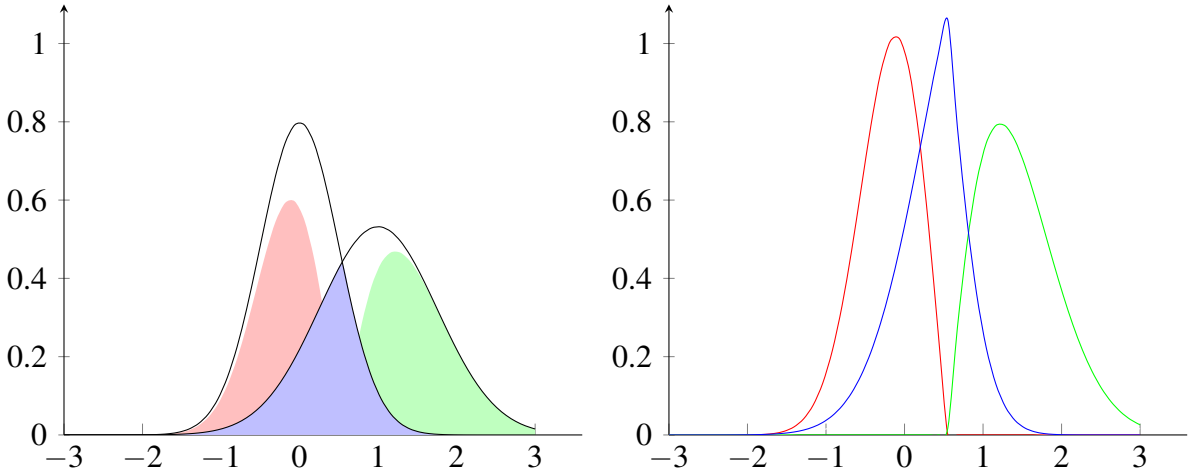


Figure 2.1: Illustration of the distributional witness concept. Left: Densities of a $N(0, 0.5)$ vs $N(1, 0.75)$ with densities of witnesses shaded. Right: Actual densities of witnesses.

where the mixing weight is $\lambda = \mathrm{TV}(P, Q)$. Figure 2.1 illustrates this concept.

Viewed through the lens of random elements, these decompositions allow us to view the generating mechanism of sampling from P and Q respectively as equivalent to sampling from the mixture distributions in (2.14). Indeed we associate to X (equivalently for Y) the latent binary indicator W^P, which takes value 1 if the component specific to P, H_P, is "selected" and zero otherwise. As before, it holds by construction $\mathbb{P}(W^P = 1) = \mathrm{TV}(P, Q)$. Intuitively, an observation X with $W^P = 1$ reveals the distribution difference of P with respect to Q. This fact leads to the following definition:

Definition 2.5 (Distributional Witness)**.** *An observation X from P with latent realization $W^P = 1$ in the representation of P given by* (2.14) *is called a* distributional witness *of the distribution P with respect to Q. We denote by $\mathrm{DW}_m(P; Q)$ the number of witness observations of P with respect to Q out of m independent observations from P.*

From the above sampling scheme, we actually see that

$$\mathrm{DW}_m(P; Q) \sim \mathrm{Binomial}(m, \mathrm{TV}(P, Q)).$$

The second concept is that of a bounding operation: Let $\bar{\Lambda}_P \in \mathbb{N}$, $\bar{\Lambda}_Q \in \mathbb{N}$ be numbers *overestimating* the true number of distributional witnesses from m i.i.d. samples from P and n i.i.d. samples from Q, i.e.

$$\bar{\Lambda}_P \geqslant \Lambda_P := \mathrm{DW}_m(P; Q), \ \bar{\Lambda}_Q \geqslant \Lambda_Q := \mathrm{DW}_n(Q; P). \tag{2.15}$$

Thus, it could be that $\bar{\Lambda}_P, \bar{\Lambda}_Q$ denote the true number of witnesses, but more generally, they need to be larger or equal. If $\bar{\Lambda}_P > \Lambda_P$ or $\bar{\Lambda}_Q > \Lambda_Q$, a *precleaning* is performed: We randomly choose a set of $\bar{\Lambda}_P - \Lambda_P$ non-witnesses from the sample of F and $\bar{\Lambda}_Q - \Lambda_Q$ non-witnesses

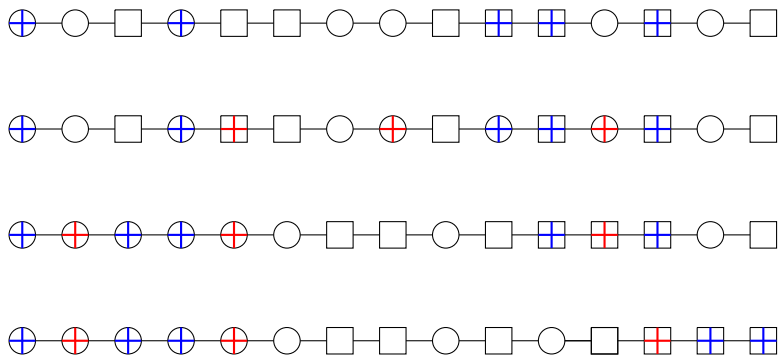


Figure 2.2: Illustration of the bounding operation. The first row from above is the original order statistic shown as circles (coming from F) and squares (coming from G). Witnesses are indicated by blue crosses. In the second, randomly chosen non-witnesses are added to the list of witnesses left and right, indicated by red, until the number of witnesses is $\bar{\Lambda}_P$ and $\bar{\Lambda}_Q$. In the final two rows, the witnesses of F and G are pushed to the left and right respectively, such that the original order of the non-witnesses in the second row is kept intact.

from the sample of G and mark them as witnesses. Thus we artificially increase the number of witnesses left and right to $\bar{\Lambda}_P$, $\bar{\Lambda}_Q$. Given this sample of witnesses and non-witnesses and starting simultaneously from the first and last order statistics $Z_{(1)}$ and $Z_{(N)}$, for $i \in \{1, \dots, N\}$ in the combined sample, we do:

(1) If $i < \bar{\Lambda}_P$ and $Z_{(i)}$ is *not* a witness from F, replace it by a witness from F, randomly chosen out of all the remaining F-witnesses in $\{Z_{(i+1)}, \dots Z_{(N)}\}$. Similarly, if $i < \bar{\Lambda}_Q$ and $Z_{(N-i+1)}$ is *not* a witness from G, replace it by a witness from G, randomly chosen out of all the remaining G-witnesses in $\{Z_{(1)}, \dots Z_{(N-i)}\}$.

(2) Set $i = i + 1$.

We then repeat (1) and (2) until $i = \max\{\bar{\Lambda}_P, \bar{\Lambda}_Q\}$.

This operation is quite intuitive: we move from the left to the right and exchange points that are not witnesses from F (i.e. either non-witnesses or witnesses from G), with witnesses from F that are further to the right. This we do, until all the witnesses from F are aligned in the first $\bar{\Lambda}_P$ positions. We also do the same for the witnesses of G in the other direction of the order statistic. Implementing the same counting process that produced $V_{m,z}$ in the original sample leads to a new counting process $z \mapsto \bar{V}_{m,z}$. This is illustrated in Figure 2.2. Lemma 2.6 collects some properties of this process, which is now much more well-behaved than the original $V_{m,z}$.

Lemma 2.6. *$\bar{V}_{m,z}$ obtained from the bounding operation above has the following properties:*

(i) $\mathbb{P}\left(\forall z \in J_{m,n} : \bar{V}_{m,z} \geqslant V_{m,z}\right) = 1$, *i.e. it stochastically dominates $V_{m,z}$.*

(ii) It increases linearly with slope 1 for the first $\bar{\Lambda}_P$ observations and stays constant for the last $\bar{\Lambda}_Q$ observations.

(iii) If $\bar{\Lambda}_P < m$ and $\bar{\Lambda}_Q < n$ and for $z \in \{\bar{\Lambda}_P + 1, \dots, N - \bar{\Lambda}_Q - 1\}$, it factors into $\bar{\Lambda}_P$ and a process $\tilde{V}_{m-\bar{\Lambda}_P, z-\bar{\Lambda}_P}$, with

$$\tilde{V}_{m-\bar{\Lambda}_P, z-\bar{\Lambda}_P} \sim \textit{Hypergeometric}\left(z - \bar{\Lambda}_P, m + n - \bar{\Lambda}_P - \bar{\Lambda}_Q, m - \bar{\Lambda}_P\right). \tag{2.16}$$

Using these tools, we end this section by sketching the proof of Proposition 2.4.

We aim to prove

$$\limsup_{N\to\infty} \mathbb{P}(\sup_{z\in J_{m,n}} [V_{m,z} - Q_{m,n,\alpha}(z,\lambda)] > 0) \leqslant \alpha. \tag{2.17}$$

Let Λ_P, Λ_Q be the distributional Witnesses of P and Q, as in Definition 2.5. Define the events $A_P := \{\Lambda_P \leqslant q_{1-\frac{\alpha}{3}}(\lambda, m)\}$, $A_Q := \{\Lambda_Q \leqslant q_{1-\frac{\alpha}{3}}(\lambda, n)\}$ and $A = A_P \cap A_Q$, such that $\mathbb{P}(A^c) \leqslant 2\alpha/3$. On A, we overestimate the number of witnesses on each side by construction. In this case we are able to use the bounding operation described above with $\bar{\Lambda}_P = q_{1-\frac{\alpha}{3}}(\lambda, m)$ and $\bar{\Lambda}_Q = q_{1-\frac{\alpha}{3}}(\lambda, n)$ to obtain $\bar{V}_{m,z}$ from Lemma 2.6. The process $\bar{V}_{m,z}$ has

$$\bar{V}_{m,z} = \begin{cases} z, & \text{if } 1 \leqslant z \leqslant q_{1-\frac{\alpha}{3}}(\lambda, m) \\ m, & \text{if } m + n(\lambda) \leqslant z \leqslant m + n \\ \tilde{V}_{m, z-q_{1-\frac{\alpha}{3}}(\lambda,m)} + q_{1-\frac{\alpha}{3}}(\lambda, m), & \text{if } q_{1-\frac{\alpha}{3}}(\lambda, m) < z < m + n(\lambda), \end{cases}$$

where $m(\lambda) = n - q_{1-\frac{\alpha}{3}}(\lambda, m)$, $n(\lambda) = n - q_{1-\frac{\alpha}{3}}(\lambda, n)$, and $\tilde{V}_{m, z-q_{1-\frac{\alpha}{3}}(\lambda,m)} \sim$ Hypergeometric$(z - q_{1-\frac{\alpha}{3}}(\lambda, m), m(\lambda), m(\lambda) + n(\lambda))$. Then:

$$\mathbb{P}(\sup_{z\in J_{m,n}} [V_{m,z} - Q_{m,n,\alpha}(z,\lambda)] > 0) \leqslant \frac{2\alpha}{3} + \mathbb{P}(\sup_{z\in J_{m,n}} [\bar{V}_{m,z} - Q_{m,n,\alpha}(z,\lambda)] > 0 \cap A),$$

Now, $\bar{V}_{m,z} - Q_{m,n,\alpha}(z,\lambda) > 0$ can only happen for $z \in \tilde{J}_{m,n,\lambda} := \{q_{1-\frac{\alpha}{3}}(\lambda, m) + 1, \dots, m + n(\lambda) - 1\}$, as by construction $\bar{V}_{m,z} - Q_{m,n,\alpha}(z,\lambda) = 0$, for $z \notin \tilde{J}_{m,n,\lambda}$. Thus

$$\begin{aligned} &\limsup_{N\to\infty} \mathbb{P}(\sup_{z\in J_{m,n}} [\bar{V}_{m,z} - Q_{m,n,\alpha}(z,\lambda)] > 0 \cap A) \leqslant \frac{2\alpha}{3} + \\ &\limsup_{N\to\infty} \mathbb{P}(\sup_{z\in \tilde{J}_{m,n,\lambda}} \left[\tilde{V}_{m,z-q_{1-\frac{\alpha}{3}}(\lambda,m)} - \mathrm{s}q_{\alpha/3}\left(\tilde{V}_{m(\lambda),z-q_{1-\frac{\alpha}{3}}(\lambda,m)}, z \in \tilde{J}_{m,n,\lambda}\right)\right] > 0) \\ &\leqslant \alpha, \end{aligned}$$

by definition of $\mathrm{s}q_{\alpha/3}\left(\tilde{V}_{m(\lambda),z}, z = 1, \dots, m(\lambda) + n(\lambda) - 1\right)$.

2.4 Discussion

The significance of Proposition 2.4 might not be immediately obvious. The key is that, despite not knowing how $V_{m,z}$ behaves in general, the bounding function in (2.11) allows to bound $V_{m,z}$

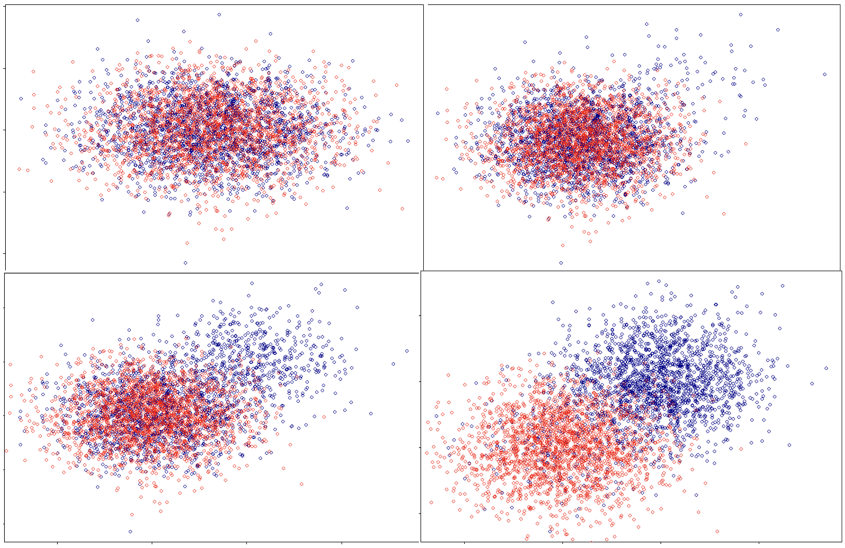


Figure 2.3: Simulate data with $\delta = 0$ (top right), $\delta = 0.05$, (top left), $\delta = 0.3$ (bottom right) and $\delta = 0.8$ (bottom left).

by a quantile that is known (given the candidate $\tilde{\lambda}$) and a process $\tilde{V}_{m,z}$ that is simply the *same hypergeometric process as under the null*. Thus we managed to reduce the problem from the unknown $V_{m,z}$, which in general might depend on P and Q, to a distribution-free hypergeometric process that arises under the null. The way we got there appears moreover quite intuitive with the bounding function described above.

We close this chapter with a small example of an implementation of the estimator in (2.10). We consider P to be a two-dimensional Gaussian distribution with mean zero and identity covariance matrix. On the other hand, Q is a mixture of P and a two-dimensional Gaussian with mean $(2, 2)$. The mixture parameter δ controls the strength of difference between P and Q, as $\delta = 0$ corresponds to $P = Q$ and for $\delta = 1$, Q is the the Gaussian distribution with mean $(2, 2)$. Figure 2.3 illustrates this setting for $\delta \in \{0, 0.05, 0.3, 0.8\}$. We note that a (strong) two-sample test would simply reject the null for any δ large enough. Instead, the estimator in (2.10) based on RF, not only manages to detect the slight difference in P and Q for the case $\delta = 0.05$, but also gets successively larger as δ increases. In particular, it is zero for $\delta = 0$ and thus correctly recognizes that H_0 cannot be rejected, and grows from around 0.0035, for $\delta = 0.05$, up to 0.6435 for $\delta = 0.9$.

3 Scoring Rules for Imputation Methods

3.1 Problem Setup

An important task in actual data applications is dealing with missing values. As such a myriad of ways to impute missing entries with "sensible" values was developed over the last two decades. The 'R-miss-tastic' platform, developed in an effort to collect knowledge and methods to streamline the task of handling missing data, lists over 150 packages [114]. However, while there is no shortage of imputation mehtods, there appears to be no principled way to assess the quality of an imputation for a given dataset. This assessment of quality via "scoring" imputaton methods is the subject of this paper.

The problem is best illustrated with an example: Consider the two-dimensional spiral example, shown in Figure 3.1: We generated 1000 observations of the noisy spiral, each entry having an independent probability of being missing of $p_{miss} = 0.3$. We imputed the missing values using 3 methods: The first "imputation" is simply the true data, followed by "mice-cart", "sample" and "loess". For the sake of this example it is not important what these imputations exactly do, only that (i) by eye, we can identify a seemingly clear ordering of these methods and (ii) "mice-cart", "sample" try to draw from the true conditional distribution of missing given observed, while "loess" simply estimates the mean of X_2 given X_1 and vice-versa. Since this way of assessing an imputation is only possible in such toy examples, we would like to have a general scoring method that reproduces this clear ordering in this example. If the true underlying data are available, say in research settings, then the root mean-squared error or RMSE is often taken, whereby we average the squared difference between any imputed values and its true value. It turns out that this is a dangerously misleading approach. Indeed, this is nicely illustrated by this nonlinear spiral example: RMSE heavily favors loess, which simply imputes conditional means, instead of cart or even true data. This reduces the variance and Figure 3.1 clearly gives an impression that the underlying data distribution is not well-represented with this imputation. Though loess certainly would not look as bad in a multivariate Gaussian distribution, the problem remains: conditional mean imputations artificially reduce the variance and strengthen associations between variables. This can lead to inflated p-values and invalid inference.

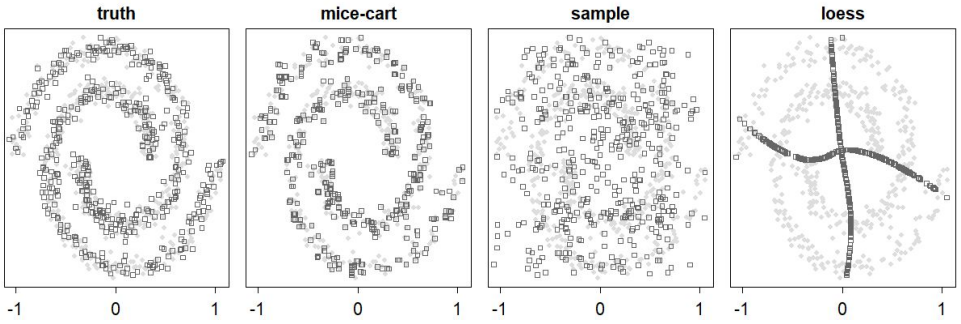


Figure 3.1: Imputations for the spiral example ($n = 1000$, $p = 2$). The complete observations are plotted in gray and the imputed observations in blue.

In contrast, we now present a scoring methodology that avoids these pitfalls, and moreover does not need access to the true data.

Unfortunately there is a lot of notation we need to introduce, before the main results can be discussed. This notation will however also be helpful for the next chapter.

We assume an underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$ on which all random elements are defined. Throughout, we take $\mathcal{P}$ to be a collection of probability measures on $\mathbb{R}^d$, dominated by some σ-finite measure μ. We denote the (unobserved) complete data distribution by $P^* \in \mathcal{P}$ and by P the actually observed distribution with missing values. We assume that P (P^*) has a density p (p^*). We take X (X^*) to be the random vector with distribution P (P^*) and let x_i (x_i^*), $i = 1, \ldots, n$, be realizations of an i.i.d. copy of the random vector X (X^*). Similarly, M is the random vector in $\{0,1\}^d$, encoding the missingness pattern of X, with realization m, whereby for $j = 1, \ldots, d$, $m_j = 0$ means that variable j is observed, while $m_j = 1$ means it is missing. For instance, the observation $(\texttt{NA}, x_2, x_3)$ corresponds to the pattern $(1,0,0)$. We denote the distribution of M as P^M, with support $\mathcal{M}$, so that $\mathbb{P}(M = m) = P^M(m)$.

For a subset $A \subseteq \{1, \ldots, d\}$ and for a random vector X or an observation x in $\mathbb{R}^d$, we denote with X_A (x_A) its projection onto that subset of indices. For instance if $d = 3$ and $A = \{1,2\}$, then $X_A = (X_1, X_2)$ ($x_A = (x_1, x_2)$). The projection onto A of the observation x_i, $(x_i)_A$, is denoted as $x_{i,A}$. Analogously, for a missingness pattern $M \sim P^M$ or an observation m in $\{0,1\}^d$, we denote with M_A (m_A) its projection onto the subset of indices in A. If X has a density p on $\mathbb{R}^d$, we denote by p_A the density of the projection X_A.

To denote assumptions on the missingness mechanism, we use a notation along the lines of [151]. For each realization m of the missingness random vector M we define with $o(X,m) := (X_j)_{j \in \{1,\ldots,d\}: m_j = 0}$ the observed part of X according to m and with $o^c(X,m) := (X_j)_{j \in \{1,\ldots,d\}: m_j = 1}$ the corresponding missing part. Note that this operation only filters the corresponding elements of X according to m, regardless whether or not these elements are actually missing or not. For instance, we might consider the unobserved part $o^c(X,m)$ according to m for the fully observed X, that is $X \sim P | M = \mathbf{0}$, where $\mathbf{0}$ denotes the vector of zeros of length d.

3.2 Proposed Method

The main idea underlying the proposed approach is that "a desirable property of imputation methods is that they should preserve the joint and marginal distributions."[128]. We thus devise a method that measures the distributional difference between an imputed distribution H and the observed data distribution P. An imputation distribution H is one which exactly matches the observed part of P. That is, all $H \in \mathcal{H}_P$, where

$$\mathcal{H}_P := \{H \in \mathcal{P} : h(o(x,m)|M = m) = p(o(x,m)|M = m) \text{ for all } m \in \mathcal{M}\}. \tag{3.1}$$

This just encodes the fact that an imputation is not allowed to change the observed values of a dataset, though here we just ask that the distribution should not be changed. Importantly, this definition includes "imputation" from the true distribution $H = P^*$, which corresponds to sampling from the true conditional distributions of missing given observed.

Since P includes unobserved elements, or `NA`'s, care needs to be taken to use any method in a data-efficient way. We will approach this by using Random Projections onto subsets $A \subset \{1, \ldots, p\}$. The reason to do this is that a projection A always contains at least as many fully observed points, as there are in the overall data set. This is actually true for any pattern m, there are always at least as many observations with pattern m_A on A, then observations with pattern m on the full data.

Let $\mathcal{K}$ be a distribution over random projections, $H_{m_A}(h_{m_A})$ the distribution (density) of $H_A \mid M_A = m_A$ and $p_A(X_A \mid M_A = \mathbf{0})$ the density of the fully observed points on A. We then define the following score:

Definition 3.1. ***Density Ratio I-Score***
We define the DR I-Score of the imputation distribution H by

$$S^*_{NA}(H,P) = \mathbb{E}_{A\sim\mathcal{K}, M_A\sim P^M_A, X_A\sim H_{M_A}}\left[\log\left(\frac{p_A(X_A \mid M_A = \mathbf{0})}{h_{M_A}(X_A)}\right)\right], \tag{3.2}$$

where the integration is over projections $A \in \mathcal{A}$, patterns $M_A \sim P^M_A$ and $X_A \sim H_{M_A}$.

Importantly, this can also be written as

$$S^*_{NA}(H,P) = -\mathbb{E}_{A\sim\mathcal{K}, M_A\sim P^M_A} D_{KL}(h_{M_A} \mid\mid p_A(\cdot \mid M_A = \mathbf{0})),$$

where the Kullback-Leibler divergence (KL divergence) between two distributions $P, Q \in \mathcal{P}$ on $\mathbb{R}^d$ with densities p, q is defined by

$$D_{KL}(p \mid\mid q) := \int p(x) \log\left(\frac{p(x)}{q(x)}\right) d\mu(x).$$

The score thus measures the difference between the fully observed distribution $P_A \mid M_A = \mathbf{0}$ and the imputation distribution given pattern m_A, $H_A \mid M_A = m_A$. Naturally, the closer H is to the true distribution P^* the larger this score should be.

So far the relation to Random Forest might not be clear. The key is that we estimate (3.2) by training a Random Forest classifier, or more accurately, a probability forest [111] to differentiate between fully observed cases on A (p_A) and imputed cases of the pattern m_A. In fact there is a deeper connection: Note that we are taking expectations over random projections A, among other things, and using a probability forest to estimate the ratio in the expectation. This approach can also be seen as an adapted RF, where each tree obtains a potentially different projection of the data. We refer to Paper **B** for more details.

3.3 Results

The main question after defining a score, other than how to estimate it in practice, is whether it is *proper*. In our context propriety of a score means

$$S_{NA}(H, P) \leqslant S_{NA}(P^*, P), \tag{3.3}$$

for all $H \in \mathcal{H}_P$. Thus a proper score should rank an imputation from the true distribution highest. A somewhat surprising result emerged in this context concerning the propriety of the DR I-Score:

Proposition 3.2. *Let $H \in \mathcal{H}_P$, as defined in* (3.1). *If for all $A \in \mathcal{A}$,*

$$\begin{aligned} &p^*(o^c(x_A, m_A) \mid o(x_A, m_A), M = m'_A) = p^*(o^c(x_A, m_A) \mid o(x_A, m_A)), \\ &\textit{for all } m'_A, m_A \in \mathcal{M}_A, \end{aligned} \tag{3.4}$$

then (3.3) *holds for $S^*_{NA}(H, P)$ in* (3.2).

We thus call $S^*_{NA}(H, P)$ a *proper* I-Score, i.e., a score that gives its highest value to the true data imputation $H = P^*$. To illuminate condition (3.4) and the overall approach, consider the following data example with three patterns: $M = (0,0,0)$, $M = (1,0,0)$, $M = (1,1,0)$,

13	10	3
NA	13	7
5	16	9
NA	NA	4
NA	5	8
-1	15	11
NA	16	18
NA	NA	10
9	14	16
NA	14	5

13	10	3
6	13	7
5	16	9
10	8	4
7	5	8
-1	15	11
14	16	18
17	12	10
9	14	16
12	14	5

such that for $A = \{1, 2, 3\}$ and an imputation H, we would have two comparisons:

13	10	3
5	16	9
-1	15	11
9	14	16

to

6	13	7
7	5	8
14	16	18
12	14	5

and

13	10	3
5	16	9
-1	15	11
9	14	16

to

10	8	4
17	12	10

Focussing on the first, we compare,

13	10	3
5	16	9
-1	15	11
9	14	16

to

6	13	7
7	5	8
14	16	18
12	14	5

The left observations have distribution $P \mid M = (0,0,0)$, while the right have $H \mid M = (1,0,0)$. These will be different in general. We now require

$$p^*(o^c(x,m)|o(x,m), M = \mathbf{0}) = p^*(o^c(x,m)|o(x,m), M = m).$$

Thus, while the distribution of $o(x,m)$ and $o(x,m)$ could be different, the conditional distribution of missing (with respect to m) given observed (with respect to m) need to be independent of the actual pattern. If this condition holds more generally over all patterns, as in (3.4), we obtain propriety.

There is a deep connection between (3.4) and traditional missingness analysis: It is simply the missing at random (MAR) condition on the projection A: [151] defines MAR as,

$$\begin{aligned} &\mathbb{P}(M_A = m_A | X_A = x_A) = \mathbb{P}(M_A = m_A | X_A = \tilde{x}_A) \\ &\text{for all } x_A, \tilde{x}_A \text{ s.t. } o(x_A, m_A) = o(\tilde{x}_A, m_A). \end{aligned} \tag{3.5}$$

Lemma 3.3. *Condition* (3.4) *and* (3.5) *are equivalent.*

In fact, as we will argue in the next section, (3.5) is a more natural formulation of the MAR condition, at least in our context.

3.4 Discussion

What makes Proposition 3.2 surprising is that one would expect as a prerequisite for our approach to work that the compared distributions are the same. After all, we are comparing two distributions using the KL divergence, and if we compare different distributions even when $H = P^*$, how can we make sure that P^* is scored highest? The key is that when comparing pattern m to the fully observed pattern, the KL divergence factors as:

$$\begin{aligned} &\text{KL(pattern } m) = \text{KL(observed in pattern } m) + \\ &\mathbb{E}_{\text{observed}}[\text{KL(missing given observed in pattern } m)]. \end{aligned}$$

Since H is not allowed to change the distribution of the observed elements, the first part of this sum will stay the same for all imputation distributions. As in general it will not be zero (i.e. the distribution of observed points may change), it is a kind of irreducible error. The second part compares the conditional distribution of $o^c(x,m)$ given the observed $o(x,m)$. Crucially, we have demanded above that this distribution is not allwed to change, when one goes from $M = m$ to $M = \mathbf{0}$ and thus this term will be minimized at 0, for $H = P^*$. Thus condition (3.4) is a quite natural condition for this comparison. And indeed, if the conditional distributions of missing given observed can vary arbitrarily, one is able to construct counter-examples for which $H = P^*$ is not scored highest. After all, there is no way in general to say something about P^*, if what cannot be observed is able to behave without constraints. This is the same principle underlying the MAR idea. The price we pay for having random projections, is that MAR needs to hold for *all projections A*. But again this is a quite natural condition, and holds in particular if the data is missing completely at random (MCAR). Testing for MCAR is the subject of the next chapter.

4 A useful MCAR Test

4.1 Problem Setup

In the last chapter we introduced a score that allows to evaluate missing data imputations. Taking a step back, before doing any imputation there is some interest in determining whether the missingness mechanism is actually missing completely at random (MCAR). Since the concepts of MCAR and MAR were only discussed briefly in the last chapter, we focus on it more here: Let again the random vector M encode the missingness, i.e. $M_j = 1$ means X_j is missing and $M_j = 0$ means X_j is observed. MCAR means that the missingness mechanism is completely independent from the data, or X^* is independent of M.[1] If $\mathbf{X}$, $\mathbf{M}$ are the matrices collecting n i.i.d. observations from P and P^M respectively in their rows, this also means that these two matrices are independent.

Crucially it turns out that existing MCAR tests either (i) don't conserve the level in all but the most favorable examples and/or (ii) display no power or are computationally not usable on datasets with "reasonable" dimensions.

4.2 Proposed Method

The method shares similarities with the method of Chapter 3, but has its own peculiarities. Recall that a missingness pattern m is defined by a vector of length p, consisting of ones and zeros, indicating which of the p variables are missing in the given pattern. We divide the n observations into $g \in \{1, \dots, G\}$ unique groups, such that the observations of each group share the same missingness pattern. Each group contains n_g observations such that $n_1 + \cdots + n_G = n$. Let P_g denote the joint distribution of the p variables in the missingness pattern group g, such that the n_g observations of group g are i.i.d. draws from P_g. As above, P_g^* is the corresponding (unobserved) underlying distribution, before the missingness mechanism comes into play. Testing MCAR can be reformulated, as testing the equality of the distribution of the

[1] We note that we are not even technically able to test MCAR in general, but instead test for a condition called OAR, but this is not important for this discussion.

different missingness patterns:

$$\begin{aligned} &H_0 : P_1^* = P_2^* = \ldots = P_G^* \\ &\text{v.s.} \\ &H_A : \exists\, i \neq j \in \{1, \ldots, G\} \text{ s.t. } P_i^* \neq P_j^*. \end{aligned} \tag{4.1}$$

In familiar fashion, we test a weaker version of (4.1), by first drawing a set A, as the projection $\mathbf{X}_{\bullet,A}$ contains less patterns than the overall data set. A key difference to Chapter 3 is that we now don't have the luxury of an imputation and thus can only work with the data at hand. As such, we consider the fully observed points on A, which we denote as $\mathbf{X}_{\mathcal{N}_A,A}$. Each observation of $\mathbf{X}_{\mathcal{N}_A,A}$, though belonging to the pattern $\mathbf{0}$ on the projection A, is part of a pattern in the overall data. In the worst case, each observation of $\mathbf{X}_{\mathcal{N}_A,A}$ could belong to a different pattern. We also note that the actual pattern each observation in $\mathbf{X}_{\mathcal{N}_A,A}$ belongs to is determined by the patterns of the variables $\mathbf{X}_{\mathcal{N}_A,A^c}$, with $A^c = \{1, \ldots, p\} \backslash A$. Figure 1 illustrates this: On the projection A, the observations $\mathbf{X}_{\mathcal{N}_A,A}$ are all fully observed. However, looking at the first two columns (A^c), one sees that they actually belong to different patterns, and this is solely determined by the first two columns. Thus to have a sensible labeling, we actually draw a subset $B \subset A^c$, and label the observations in $\mathbf{X}_{\mathcal{N}_A,A}$ according to the patterns of $\mathbf{X}_{\mathcal{N}_A,B}$. This is again illustrated in Figure 1, where the second column is chosen, resulting in two labels, one for the first observation in $\mathbf{X}_{\mathcal{N}_A,A}$ and another one for the second and third.

Thus this is our projection scheme: We first randomly draw $A \subset \{1, \ldots, p\}$ to obtain $\mathbf{X}_{\mathcal{N}_A,A}$ and then $B \subset A^c$ to determine the labels for each observation in $\mathbf{X}_{\mathcal{N}_A,A}$. Given $\mathbf{X}_{\mathcal{N}_A,A}$ with labels, we then train a multiclass classifier that should distinguish the different classes. We thereby again use the KL divergence, as in Chapter 3 and [24], but extend it to multiclass classification. We will leave out the exact details, but the crucial point is that the final test statistic $\hat{U}$ is build by averaging over N different projections:

$$\hat{U} := \frac{1}{N} \sum_{i=1}^{N} \hat{U}^{(A_i,B_i)}, \tag{4.2}$$

where $\hat{U}^{(A_i,B_i)}$ is the classification-based accuracy measure on the ith projection.

To ensure the level of this rather complicated approach is kept for every sample size, we need a permutation approach. That is, we want to randomly shuffle the labels in each projection and calculate the statistic on the shuffled data. Unfortunately, this needs to be done in a way that respects the dependencies inherent in this procedure. For instance, it is not good enough to simply use B permutations independently for each of the N projections. The key is to instead permute the rows of the matrix of missingness patterns $\mathbf{M}$. That is, we redo the analysis using $\mathbf{M}_\sigma$, $\mathbf{M}$ with permuted rows, which results in $\hat{U}_{\sigma_\ell}^{(A_i,B_i)}$, $i = 1, \ldots, N$. Note that the $(A_i, B_i)_{i=1}^N$ remain exactly as before and indeed in practice, the actual statistic and the permuted ones are calculated simultaneously for each (A_i, B_i).

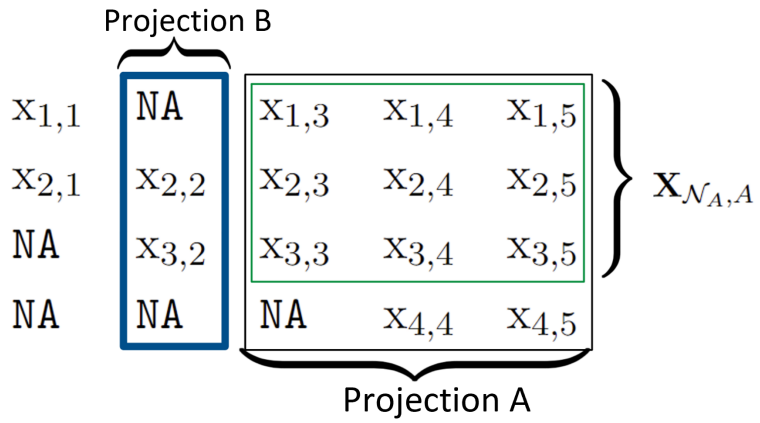


Figure 4.1: Illustration of the projections A, B. In a first step, a projection $A \subset \{1,\dots,p\}$ is drawn, as indicated in black. The fully observed points on A, then form $\mathbf{X}_{\mathcal{N}_A,A}$. In a second step, $B \subset \{1,\dots,p\}\backslash A$ is drawn as indicated in bold blue. The patterns in Projection B then determine which observation in $\mathbf{X}_{\mathcal{N}_A,A}$ have the same label. In this case, the first observation has a different label, than the second and third observations, which share a common label.

Finally, we calculate for $\ell = 1,\dots,L$,

$$\hat{U}_{\sigma_\ell} := \frac{1}{N}\sum_{j=1}^{N} \hat{U}_{\sigma_\ell}^{(A_j,B_j)}. \tag{4.3}$$

The p-value of the test is then obtained as usual by

$$Z := \frac{\sum_{\ell=1}^{L} I\{\hat{U}_{\sigma_\ell} \geqslant \hat{U}\} + 1}{L+1}. \tag{4.4}$$

4.3 Results

The main result here, is that under the aforementioned permutation scheme, the test actually provides a valid p-value. A p-value Z say is valid if, under H_0,

$$\mathbb{P}(Z \leqslant z) \leqslant z \text{ for all } z \in [0,1]. \tag{4.5}$$

That is, a valid p-value is a random variable that, under H_0, stochastically dominates the uniform distribution.

We first discuss a general result for permutation approaches and then use it to show that the p-value in (11) is valid. Let $\mathbf{Y}$ be a $n \times p$ random matrix and let G be a finite set of transformations that forms a group. That is, G contains the identity element, every element of G has an inverse and for $g_1 \in G$ and $g_2 \in G$, the composition $g_2 \circ g_1$ is also in G. If the first column of $\mathbf{Y}$ are $0/1$ labels, G is usually the set of permutations of the labels. That is, each $g \in G$, outputs $g(\mathbf{Y})$, whereby the first column is permuted, while all other columns remain the same.

We also do not consider all elements in G, but instead choose a random subset of size L. Let in the following $g_0 \in G$ be the identity, i.e. $g_0(\mathbf{Y}) = \mathbf{Y}$. Let for randomly selected $g_1, \ldots, g_L \in G$,

$$G' := (g_0, g_1, \ldots, g_L).$$

We also assume to have a statistic T, such that $T(\mathbf{Y})$ takes values in $\mathbb{R}$. Then we obtain the sample $(T(g_0(\mathbf{Y})), T(g_1(\mathbf{Y})), \ldots, T(g_L(\mathbf{Y})))$. Let $T^\ell(\mathbf{Y}, G')$, $\ell \geqslant 1$ be the ℓth order statistics of this sample, such that

$$T^1(\mathbf{Y}, G') \leqslant \ldots \leqslant T^{L+1}(\mathbf{Y}, G').$$

Then

Theorem 4.1. *Assume that under H_0,*

$$\mathbf{Y} \stackrel{D}{=} g(\mathbf{Y}) \text{ for all } g \in G, \tag{4.6}$$

that G is a group and G' defined as above. Then if we define

$$Z := \frac{\sum_{\ell=1}^{L+1} I\{T^\ell(\mathbf{Y}, G') \geqslant T(\mathbf{Y})\}}{B+1}, \tag{4.7}$$

(4.5) *holds for Z under H_0.*

Thus this very general result, a direct consequence of Theorem 2 in [70], guarantees that Z is a valid p-value! Condition (4.6) is actually stronger than what is needed, but it is enough general for our purposes. It is also somewhat intuitive: It implies that under H_0, $T(\mathbf{Y})$ must come from the same distribution as $T^\ell(\mathbf{Y}, G')$, $\ell = 1, \ldots, L+1$. Thus, we should not expect its value to be stochastically higher than the values $T^\ell(\mathbf{Y}, G')$, $\ell = 1, \ldots, L+1$. Care must be taken however with this intuition, because the values in $(T(g_0(\mathbf{Y})), T(g_1(\mathbf{Y})), \ldots, T(g_L(\mathbf{Y})))$ are not (unconditionally) independent.

We now use this to proof that Z as defined in (11) is a valid p-value. In this one case, it might be instructive to include the proof of the paper:

Proposition 4.2. *Under H_0 in* (4.1)*, and Z as defined in* (11)*,* (4.5) *holds.*

Proof

Let $\mathbf{A} = (A_1, \ldots, A_N)$, $\mathbf{B} = (B_1, \ldots, B_N)$. Let G be all possible permutations of the rows of $\mathbf{M}$, that is for all $g \in G$,

$$g(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B}) = (\mathbf{X}^*, \mathbf{M}_{\sigma_\ell}, \mathbf{A}, \mathbf{B}),$$

$\ell = 1, \ldots, L^*$. Note that, since we are only considering fully observed observations for projections $\mathbf{A}$, $\hat{U}$ is actually a function of $\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B}$, while $\hat{U}_{\sigma_\ell}$ is a function of $G_\ell(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B})$. It also holds that:

$$(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B}) \stackrel{D}{=} (\mathbf{X}^*, \mathbf{M}_{\sigma_\ell}, \mathbf{A}, \mathbf{B}) = g(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B}) \ \ \forall g \in G. \tag{4.8}$$

Indeed, under MCAR, $\mathbf{X}^*$ and $\mathbf{M}$ are independent. Since by the i.i.d. assumption also $\mathbf{M}_{\sigma_\ell} \stackrel{D}{=} \mathbf{M}$ for all ℓ and since $\mathbf{A}$, $\mathbf{B}$ are also independent of $\mathbf{M}$, (18) follows. Thus H_0 is such that (4.6) holds and since G is a group, the result follows by Theorem 4.1. ■

4.4 Discussion

It is worth emphasizing how important this result is in this context. Despite the complex non-parametric setting, the p-value obtained through this permutation approach will lead to a valid MCAR test, for any sample size n. As Paper C shows, this is in contrast to other existing MCAR tests, which often grossly inflate the level. Given that a valid p-value, at least for large n, is the basic requirement a statistical test needs to meet, this is a rather significant point. Hence the provocative choice of title in this chapter. Interestingly, our test also displays very high power and is widely applicable, as we detail in the paper.

5 Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression

5.1 Problem Setup

So far we discussed ways of applying (versions) of RF to a variety of problems. However, there has also been substantial work on understand and extending the Random Forest algorithm itself. Among the research that has been done in the RF literature, two stand out: First, the original RF presents a nonparametric method of estimating conditional expectations, and thus researchers looked into whether the same algorithm could be leveraged for different distributional aspects. This presumably started with [117], who used the weight function that RF implicitly produces to construct estimates of conditional quantiles. Second, efforts intensified to understand the behavior of RF on a theoretical level and to use this to provide confidence intervals for its predictions. In particular, [122] first developed an asymptotic normality result using a generalized version of U-statistics, which was refined considerably in [181].

GRF of [4] combined these two directions and proposed a framework within which an adapted Random Forest could be constructed, according to some estimating equations. That is, for a given target, defined through the estimating equations, a splitting criterion has to be developed, which then replaces the usual CART splitting criterion in each tree. Given conditions on the data and on the estimating equations, the (univariate) consistency and asymptotic normality result of [181] then hold for the desired quantity. GRF thus unified various ways of using RF and provided a theoretical framework together with inferential tools.

However the GRF framework, while elegant, has its own shortcomings. First and foremost, while it is never clearly stated in [4], it seems quite clear that the method only works for univariate targets (though they allow for certain "nuisance" parameters). Second, the construction of a new RF method for each new target weakens one of the great strengths of the RF method: that it doesn't need to be adapted or tuned to each new problem or estimator.

Our new method, the Distributional Random Forest (DRF), addresses both shortcomings and delivers a general-purpose method that is able to learn a representation of the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ in a RF-type adaptive fashion. As an additional benefit, the theoretical treatment of our new approach, while more limited in a sense, is more elegant. The complexity of the theory in [4] is greatly increased through the estimating equations. In contrast, our method is simply estimating a mean, albeit in an infinite-dimensional Hilbert space, making the theoretical treatment somewhat straightforward.

5.2 Proposed Method

Consider an i.i.d. sample $(\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_n, \mathbf{X}_n)$, with $\mathbf{Y}$ taking values in $\mathbb{R}^d$ and $\mathbf{X}$ in $\mathbb{R}^p$. We propose the DRF algorithm which estimates the multivariate conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$. More specifically, the data is repeatedly split in different trees, as in RF, but with the splitting criterion adapted to a distributional metric: at each step, we split the data points into two groups based on some feature X_j in such a way that the distribution of $\mathbf{Y}$ for which $X_j \leqslant l$, for some level l, differs the most compared to the distribution of $\mathbf{Y}$ when $X_j > l$, according to the distributional metric. This leads to a partition of $\mathbb{R}^p$, such that a testpoint $\mathbf{x}$ ends up in exactly one of the leaves. Then a weight to each observation $\mathbf{X}_i$ in the sample is assigned as follows: if $\mathbf{X}_i$ and $\mathbf{x}$ are not in the same leaf, the weight is zero. Otherwise, it is $1/N_{\mathbf{x}}$ where $N_{\mathbf{x}}$ is the number of elements in the shared leaf. Repeating this many times with randomization induces a weighting function $w_{\mathbf{x}}(\mathbf{x}_i)$ as in [104, 105], which quantifies the relevance of each training data point $\mathbf{x}_i$ for a given test point $\mathbf{x}$. The conditional distribution is then estimated by an empirical distribution determined by these weights [117].

We propose a splitting criterion based on the Maximal Mean Discrepancy (MMD) statistic [62]. Crucially, this statistic measures the distance between two distributions as the (norm) difference between two expected values in a Hilbert Space. As such, we are able to cast DRF as a (conditional) mean estimation in this Hilbert space, which has several important theoretical implications. First, it allows to extend results from the Random Forest literature to this setting, by generalizing the arguments to Hilbert spaces. While this is not trivial, it is quite natural and immediately implies consistency of a large range of estimators based on the RF. Second, the estimate in the Hilbert space can be easily expressed as a weighted sum of kernels and is of separate interest, as a consistent estimate of the "conditional mean embedding" (CME), see e.g., [132]. Third, while this is less relevant here, the criterion allows for computationally fast approximations, which is crucial for the applicability of the algorithm.

For a positive semi-definite kernel k, let $\mathcal{H}$ be its associated Hilbert space with inner product $\langle , \rangle_{\mathcal{H}}$, see e.g. [75]. This kernel could be tuned in principle, though in typical RF-manner one particular choice, the Gaussian kernel tends to deliver good results for a wide range of

examples. It also has a host of favorable properties, as discussed below. We consider the element $\mu(\mathbf{x}) := \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}] \in \mathcal{H}$. This is called the embedding of the conditional of $\mathbf{Y} \mid \mathbf{X} = \mathbf{x}$, since it can be shown that for all $f \in \mathcal{H}$,

$$\langle f, \mu(\mathbf{x}) \rangle_{\mathcal{H}} = \mathbb{E}[\langle f, k(\mathbf{Y}, \cdot) \rangle_{\mathcal{H}} \mid \mathbf{X} = \mathbf{x}] = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}].$$

Thus $\mu(\mathbf{x})$ defines the conditional distribution in a certain sense. Indeed it can be shown that under certain conditions on the kernel, $\mu(\mathbf{x})$ uniquely defines the distribution $\mathbb{P}_{\mathbf{x}}$. As mentioned above, through the MMD splitting criterion, the DRF estimate $\mu(\mathbf{x})$ at a point $\mathbf{x}$ can be seen as mean estimate in the RKHS $\mathcal{H}$. That is we obtain the estimate

$$\hat{\mu}_n(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) k(\mathbf{y}_i, \cdot). \tag{5.1}$$

As mentioned above, this relates to the literature on the estimation of CMEs, see e.g., [132].

We now study a consistency result in the next section.

5.3 Results

We collect some assumptions on the growing of the forest. For details, the reader is referred to paper D.

(P1) (*Data sampling*) The bootstrap sampling with replacement, usually used in forest-based methods, is replaced by a subsampling step, where for each tree we choose a random subset of size s_n out of n training data points. We consider s_n going to infinity with n, with the rate specified below.

(P2) (*Honesty*) The data used for constructing each tree is split into two partOn the pitfalls of Gaussian scoring for causal discoverys; the first is used for determining the splits and the second for populating the leaves and thus for estimating the response.

(P3) (α-*regularity*) Each split leaves at least a fraction $0 < \alpha \leqslant 0.2$ of the available training sample on each side. Moreover, the trees are grown until every leaf contains between κ and $2\kappa - 1$ observations, for some fixed tuning parameter $\kappa \in \mathbb{N}$.

(P4) (*Symmetry*) The (randomized) output of a tree does not depend on the ordering of the training samples.

(P5) (*Random-split*) At every split point, the probability that the split occurs along the feature X_j is bounded below by π/p, for some $\pi > 0$ and for all $j = 1, \ldots, p$.

This can then be used to show consistency of the method in the Hilbert space norm $\|\cdot\|_{\mathcal{H}}$:

Theorem 5.1. *Suppose that our forest construction satisfies properties **(P1)–(P5)**. Assume additionally that k is a bounded and continuous kernel and that we have a random design with* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *independent and identically distributed on* $[0,1]^p$ *with a density bounded away from* 0 *and infinity. If the subsample size* s_n *is of order* n^β *for some* $0 < \beta < 1$*, the mapping*

$$\mathbf{x} \mapsto \mu(\mathbf{x}) = \mathbb{E}[\mu(\delta_{\mathbf{Y}}) \mid \mathbf{X} = \mathbf{x}] \in \mathcal{H},$$

is Lipschitz and $\sup_{\mathbf{x}\in[0,1]^p} \mathbb{E}[\|\mu(\delta_{\mathbf{Y}})\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}] < \infty$*, we obtain the consistency w.r.t. the RKHS norm:*

$$\|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} = O_p\left(n^{-\gamma}\right), \tag{5.2}$$

for $\gamma = \frac{1}{2}\min\left(1-\beta, \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}\cdot\beta\right)$.

5.4 Discussion

The result shows consistency of the DRF prediction, under quite natural assumptions, with a given rate that depends on the number of dimension of $\mathbf{X}$. The implications of the result strongly depend on the kernel k. Under a Gaussian kernel, and some additional assumptions on $\mathbf{Y}$, it can be shown that (5.2) is equivalent to the weak convergence of the corresponding distributional estimate $\hat{\mathbb{P}}_{\mathbf{x},n}$ to $\mathbb{P}_{\mathbf{x}}$ in probability. This is an interesting concept that we will discuss in more detail: On the set of all probability measures on $\mathbb{R}^d$ weak convergence of P_n to P is defined as

$$\int_{\mathbb{R}^d} f(\mathbf{y}) dP_n(\mathbf{y}) \to \int_{\mathbb{R}^d} f(\mathbf{y}) dP(\mathbf{y}),$$

for all $f : \mathbb{R}^d \to \mathbb{R}$ continuous and bounded. We will denote this as $P_n \xrightarrow{w} P$. A simple transformation of $\hat{\mu}_n$ gives

$$\hat{\mathbb{P}}_{\mathbf{x},n} = \sum_{i=1}^{n} w_i(\mathbf{x})\delta_{\mathbf{y}_i},$$

where for all measurable sets A,

$$\delta_{\mathbf{y}}(A) := \begin{cases} 1, & \text{if } \mathbf{y} \in A \\ 0, & \text{else .} \end{cases}$$

$\hat{\mathbb{P}}_{\mathbf{x},n}$ is a valid probability measure on $\mathbb{R}^d$ and one can define the (semi-metric),

$$d_k(\hat{\mathbb{P}}_{\mathbf{x},n}, \mathbb{P}_{\mathbf{x}}) := \sup_{f\in\mathcal{H}} \|\hat{\mathbb{P}}_{\mathbf{x},n} f - \mathbb{P}_{\mathbf{x}} f\|_{\mathcal{H}} = \|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}},$$

see e.g., [165]. As mentioned under a Gaussian kernel and some mild assumptions,

$$d_k(P_n, P) \to 0 \iff P_n \xrightarrow{w} P.$$

Thus the result above implies that $d_k(\hat{\mathbb{P}}_{\mathbf{x},n}, \mathbb{P}_{\mathbf{x}}) \to 0$ in probability and thus, in this sense $\hat{\mathbb{P}}_{\mathbf{x},n} \xrightarrow{w} \mathbb{P}_{\mathbf{x}}$ in probability. This has in particular the consequence that for all $f : \mathbb{R}^d \to \mathbb{R}$ bounded and continuous,

$$\int_{\mathbb{R}^d} f(\mathbf{y}) d\hat{\mathbb{P}}_{\mathbf{x},n}(\mathbf{y}) \to \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}],$$

in probability, guaranteeing consistency of a range of univariate targets. Other results using "weak convergence in distribution" have been used in the Bootstrap literature, see e.g., [94, Chapter 10].

We note that while the estimating guarantees are relatively weak, so are the assumptions on the data generating process. While this might not seem obvious given the list of assumptions, [117] for example assumes $\mathbf{x} \mapsto F(y \mid \mathbf{X} = \mathbf{x})$ to be Lipschitz with respect to the Kolmogorov metric, i.e. for some $L > 0$,

$$\sup_y |F(y \mid \mathbf{X} = \mathbf{x}) - F(y \mid \mathbf{X} = \mathbf{x}')| \leqslant L\|\mathbf{x} - \mathbf{x}'\|,$$

and achieves convergence in probability in the Kolmogorov metric. In contrast, weak convergence amounts to pointwise convergence of the cdf (in probability). However, we only assume Lipschitz continuity in a much weaker norm (namely in $\|\cdot\|_{\mathcal{H}}$ of the RKHS $\mathcal{H}$) in Theorem 5.1.

As mentioned beforehand, the fact that we still only estimate a mean renders the theory comparatively simple. What is missing from this discussion, compared to [4] are inferential tools. This important discussion is part of current work.

6 Concluding remarks

In this thesis we examined a range of papers based on the Random Forest algorithm. The first used RF to develop a new two-sample test that simultaneously delivers a rejection decision as well as a measure quantifying the strength of difference between P and Q. The second made use of RF in constructing a score indicating how well an imputation of missing values holds up when compared to the observed data. Using similar tools, together with a permutation test, the third project developed an MCAR test with strong level guarantees. Finally, the last project developed a powerful new method to estimate multivariate conditional distributions based on the RF methodology.

We hope that the methods developed in these papers will be of use in the statistical and data science communities, both applied and in research.

Paper

A

High-Probability Lower Bounds for the Total Variation Distance.

L. Michel, J. Näf, N. Meinshausen

arXiv.

High Probability Lower Bounds for the Total Variation Distance

Loris Michel Jeffrey Näf

Nicolai Meinshausen

January 20, 2023

Abstract

The statistics and machine learning communities have recently seen a growing interest in classification-based approaches to two-sample testing. The outcome of a classification-based two-sample test remains a rejection decision, which is not always informative since the null hypothesis is seldom strictly true. Therefore, when a test rejects, it would be beneficial to provide an additional quantity serving as a refined measure of distributional difference. In this work, we introduce a framework for the construction of high-probability lower bounds on the total variation distance. These bounds are based on a one-dimensional projection, such as a classification or regression method, and can be interpreted as the minimal fraction of samples pointing towards a distributional difference. We further derive asymptotic power and detection rates of two proposed estimators and discuss potential uses through an application to a reanalysis climate dataset.

Keywords. two-sample testing, distributional difference, classification, higher-criticism

1 Introduction

Two-sample testing is a classical statistical task recurring in various scientific fields. Based on two samples X_i, $i = 1,\dots,m$ and Y_j, $j = 1,\dots,n$ drawn respectively from probability measures P and Q, the goal is to test the hypothesis $H_0 : P = Q$, against potentially any alternative. The trend in the last two decades towards the analysis of more complex and large-scale data has seen the emergence of classification-based approaches to testing. Indeed, the idea of using classification for two-sample testing traces back to the work of [53]. Recently this use of classification has seen a resurgence of interest from the statistics and machine learning

communities with empirical and theoretical work ([88]; [145]; [109]; [68]; [14]; [56]; [87]; [24]) motivated by broader applied scientific work as well-explained in [88].

However, as already pointed out in [53], it is practically very unlikely that two samples come from the exact same distribution. It means that with enough data and using a "universal learning machine" for classification, as Friedman called it, the null will be rejected no matter how small the difference between P and Q is. Therefore, in many situations when a classification-based two-sample test rejects, it would be beneficial to have an additional measure quantifying the actual distributional difference supported by the data.

Practically, one can observe that with two finite samples some fraction of observations will tend to illuminate a distributional difference more than others. At a population level, this translates to the fraction of probability mass one would need to change from P to see no difference with Q. It is well known that this is an equivalent characterization of the total variation distance between P and Q, see e.g. [36]. We recall that for two probability distributions P and Q on measurable space $(\mathcal{X}, \mathcal{A})$ the *total variation (TV) distance* is defined as

$$\mathrm{TV}(P, Q) = \sup_{A \in \mathcal{A}} |P(A) - Q(A)| .$$

Therefore, based on finite samples of P and Q, a finer question than "is P different from Q ?" could be stated as "What is a probabilistic lower bound on the fraction of observations actually supporting a difference in distribution between P and Q ?". This would formally translate into the construction of an estimate $\hat{\lambda}$ satisfying

$$\mathbb{P}(\hat{\lambda} > \mathrm{TV}(P, Q)) \leqslant \alpha,$$

for $\alpha \in (0, 1)$. We call such an estimate $\hat{\lambda}$ a high-probability lower bound (HPLB) for $\mathrm{TV}(P, Q)$. An observation underlying our methodology is that uni-dimensional projections of distributions act monotonically on the total variation distance. Namely for a given (measurable) projection $\rho : \mathcal{X} \to I \subseteq \mathbb{R}$,

$$\mathrm{TV}\left(\rho_{\#}P, \rho_{\#}Q\right) \leqslant \mathrm{TV}\left(P, Q\right), \tag{1}$$

where $\rho_{\#}P$ is the push-forward measure of P, defined as $\rho_{\#}P(A) := P(\rho^{-1}(A))$ for a measurable set A. The construction of an HPLB for a given projection ρ is the focus of our work. They are used as a proxies for $\mathrm{TV}\left(P, Q\right)$ through (1). The gap depends on the informativeness of the selected projection ρ about the distributional difference between P and Q. This naturally established a link with classification and provides insights on how to look for "good projections". Nevertheless, the focus of the present paper is not to derive conditions on how to construct optimal projections ρ, but rather an analysis of the construction and properties of HPLBs for fixed projections. As a by-product, we address an issue that seems to have gone largely unnoticed in

	True Class	
	0	1
0	5121	5150
1	4879	4850

	True Class	
	0	1
0	9987	9925
1	13	75

Table 1: Confusion matrices for 2 different thresholds, $t = 0.5$ (left) and $t = 0.7$ (right). P-values are given as 0.65 and 9×10^{-4} respectively.

the literature on classification and two-sample testing. Namely, given a function $\rho : \mathcal{X} \to [0,1]$ estimating the probability of belonging to the first sample say, what is the "cutoff" t^* allowing for the best possible detection of distributional difference for the binary classifier

$$\rho_t(Z) := I\{\rho(Z) > t\}. \tag{2}$$

In line with the Bayes classifier, $t^* = 1/2$ is often used in classification tasks. However, we show that for the detection of distributional difference, this is not always the best choice. The next section illustrates this issue through a toy example.

1.1 Toy motivating example

As an illustrative example highlighting the importance of the choice of cutoff for a binary classifier, let us consider two probability distributions P and Q on $\mathbb{R}^{12}$ with mutually independent margins defined as follows: $P = \mathcal{N}_{12}(\mu, \Sigma)$ where $\Sigma = I$ is the 12×12 identity matrix and $\mu = (0, \ldots, 0)$ and $Q = (1-\varepsilon)P + \varepsilon C$, where $C = \mathcal{N}_{12}(\mu_C, \Sigma)$, with $\mu_C = (3, 3, 0, \ldots, 0)$ and $\varepsilon = 10^{-2}$. We assume to observe an iid sample $X_1, \ldots, X_n$ from P and $Y_1, \ldots, Y_n$ from Q. Figure 1 shows the projection of samples from P and Q on the first two components.

Consider $\rho : \mathbb{R}^{12} \to [0,1]$ to be a function returning an estimate of the probability that an observation Z belongs to the sample of Q, obtained for instance from a learning algorithm trained on independent data. Assume we would like to test whether there is a significant difference between a sample of P and a sample from Q based on the binary classifier ρ_t as defined in (2).

Table 1 (left) presents the confusion matrix obtained from a Random Forest classifier ρ_t trained on $n = 10'000$ samples of P and Q (as defined above) with the usual cutoff of $t = 0.5$. Based on this matrix, one can use a permutation approach to test $H_0 : P = Q$. The corresponding p-value is 0.65, showing that, despite the high sample size, the classifier is not able to differentiate the two distributions.

However this changes if we instead use a cutoff of $t = 0.7$. Using the same permutation approach, we obtain a p-value of 9×10^{-4}. The corresponding confusion matrix is displayed on Table 1 (right).

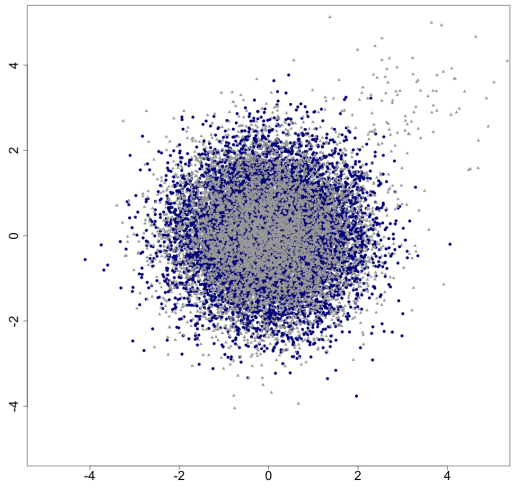


Figure 1: Projections on first two margins of $10'000$ samples from P (blue) and Q (grey).

This observation supports that even though $t = 0.5$ links to the optimal Bayes rate, depending on the choice of alternative (how P and Q differ), different cutoffs induce vastly different detection powers. Put differently, $t = 0.5$ is not always optimal for detecting a change in distribution. In this work, we will explore why this is the case and how this impacts the construction of HPLBs for $\mathrm{TV}(P, Q)$ and their (asymptotic) statistical performances. As an empirical illustration, we show at the end of Section 2.3 that, for the same simulation setting, the HPLB based on a cutoff of 0.5 will be zero, while the one that adaptively chooses the "optimal" cutoff will be positive.

1.2 Contribution and relation to other work

Direct estimators or bounds for the total variation distance have been studied in previous work when the distributions are assumed to be discrete or to belong to a known given class (e.g. [174]; [149]; [80]; [41]; [95]; [130]). Our work aims at constructing lower bounds on the total variation distance based on samples from two unknown distributions. The goal is to provide additional information over the rejection status of classification-based two-sample tests. We summarize our contributions as follows:

Construction of HPLBs: We provide a framework for the construction of high probability lower bounds for the total variation distance based on (potentially unbalanced) samples and propose two estimators derived from binary classification. The first estimator $\hat{\lambda}_{bayes}$, assumes the fixed cutoff $1/2$, whereas the second one $\hat{\lambda}_{adapt}$, is cutoff-agnostic. Despite the somewhat complicated nature of the latter estimator, we show that is a valid HPLB.

Asymptotic detection and power boundaries: We characterize power and detection rates for the proposed estimators for local alternatives with decaying power rate

$$\mathrm{TV}(\rho_{\#}P_N, \rho_{\#}Q_N) \propto N^{-\gamma}$$

, $-1 < \gamma < 0$, for $N = m + n$. We summarize the main result as follows: Consider the minimal rate $-1 < \gamma < 0$ for which a difference in P and Q could still be detected, if the optimal cutoff t^* in (2) for a given P, Q was known – this will be referred as the "oracle rate". The estimator $\hat{\lambda}_{adapt}$ always attains the oracle rate, whereas $\hat{\lambda}_{bayes}$ only attains the oracle rate if the optimal cutoff is actually $t^* = 1/2$. We also obtain the same favorable results for $\hat{\lambda}_{adapt}$ when considering a sequence $\hat{\rho}_N$, estimated on independent data.

Application: We show the potentially use and efficacy of HPLBs on the total variation distance in two different types of applications based on a climate reanalysis dataset.

Software: We provide implementations of the proposed estimators in the R-package `HPLB` available on CRAN.

From a technical point of view, the construction of our lower bound estimators relate to the higher-criticism literature ([43]; [44]) and is inspired by similar methodological constructions of high-probability lower bounds in different setups (e.g. [119]; [116]; [120]). It has also some similarities with the problem of semi-supervised learning in novelty detection ([13]).
The paper is structured as follows. Section 2 introduces the classification framework for constructing lower bounds on total variation distance and describes our proposed estimators. Power and detection rates guarantees of our proposed estimators are presented in Section 3. In Section 4 we generalize our framework beyond binary classification and in Section 5 we present three different applications of our estimators in the context of a climate dataset.

2 Theory and methodology

Let P and Q be two probability measures on $\mathcal{X}$ and $\rho : \mathcal{X} \to I$ be any measurable function mapping to some subset $I \subset \mathbb{R}$. If not otherwise stated, we consider $I = [0,1]$. The following chain of inequalities for $\mathrm{TV}(P,Q)$ will form the starting point of our approach:

$$\mathrm{TV}(P,Q) \geqslant \mathrm{TV}(\rho_{\#}P, \rho_{\#}Q) \geqslant \sup_{t \in I} \left| \rho_{\#}P\left((0,t]\right) - \rho_{\#}Q\left((0,t]\right) \right|. \tag{3}$$

For any such ρ, one can define a binary classifier $\rho_t : \mathcal{X} \to \{0,1\}$ based on the cutoff t by $\rho_t(z) := I\{\rho(z) > t\}$. Before diving into more details, let us introduce our setup and some necessary notation.

Setup and notation: Where not otherwise stated, we assume to observe two independent iid samples $X_1, \dots, X_m$ from P and $Y_1, \dots, Y_n$ from Q. We define

$$Z_i := \begin{cases} X_i & \text{if } 1 \leqslant i \leqslant m, \\ Y_i & \text{if } m+1 \leqslant i \leqslant m+n, \end{cases}$$

and attach a label $\ell_i := 0$ for $i = 1, \dots, m$ and $\ell_i := 1$, for $i = m+1, \dots, m+n$. Both m and n are assumed to be non-random with $N = m+n$ such that $m/N \to \pi \in (0,1)$, as $N \to \infty$. For notational convenience, we also assume that $m \leqslant n$. We denote by f and g the densities of P, Q respectively.[1] We define $F := \rho_{\#}P$ and $G := \rho_{\#}Q$ and introduce the empirical measures

$$\hat{F}_m(t) := \frac{1}{m}\sum_{i=1}^{m} I\{\rho(X_i) \leqslant t\}, \quad \hat{G}_n(t) := \frac{1}{n}\sum_{j=1}^{n} I\{\rho(Y_j) \leqslant t\},$$

of all observations $\{\rho_i\}_{i=1,\dots,N} := \{\rho(Z_i)\}_{i=1,\dots,N}$. Denote by $\rho_{(z)}$, $z \in \{1, \dots, N\}$, the z^{th} order statistic of $(\rho_i)_{i=1,\dots,N}$. Throughout the text, $q_\alpha(p,m)$ is the α-quantile of a binomial distribution with success probability p and number of trials m symbolized by $\text{Binomial}(p,m)$. Similarly, q_α is the α-quantile of a standard normal distribution, denoted $\mathcal{N}(0,1)$. Finally, for two functions $h_1, h_2 : \mathbb{N} \to [0,\infty)$, the notation $h_1(N) \asymp h_2(N)$, as $N \to \infty$ means $\limsup_{N\to\infty} h_1(N)/h_2(N) \leqslant a_1 \in (0,+\infty)$ and $\limsup_{N\to\infty} h_2(N)/h_1(N) \leqslant a_2 \in (0,+\infty)$. The first part of our theoretical analysis centers around the projection $\rho^* : \mathcal{X} \to [0,1]$ given as

$$\rho^*(z) := \frac{g(z)}{f(z)+g(z)}. \tag{4}$$

As a remark, if we put a prior probability $\pi = 1/2$ on observing a label ℓ of 1, ρ^* is the posterior probability of observing a draw from Q, referred to as the Bayes probability.

We now formally state the definition of a high-probability lower bound for the total variation distance, using the notation $\lambda := \text{TV}(P,Q)$ from now on:

Definition 1. *For a given $\alpha \in (0,1)$, an estimate $\hat{\lambda} = \hat{\lambda}\left((Z_1,\ell_1),\dots,(Z_N,\ell_N)\right)$ satisfying*

$$\mathbb{P}(\hat{\lambda} > \lambda) \leqslant \alpha \tag{5}$$

will be called high-probability lower bound (HPLB) at level α. If instead only the condition

$$\limsup_{N\to\infty} \mathbb{P}(\hat{\lambda} > \lambda) \leqslant \alpha \tag{6}$$

holds, we will refer to $\hat{\lambda}$ as asymptotic high-probability lower bound (asymptotic HPLB) at level α.

[1] Wlog, we assume that the densities exist with respect to some common dominating measure.

Note that an estimator $\hat{\lambda}$ depends on a function ρ. When necessary, this will be emphasized with the notation $\hat{\lambda}^\rho$ throughout the text. Whenever ρ is not explicitly mentioned it should be understood that we consider $\rho = \rho^*$.

The above definition is very broad and does not entail any informativeness of the (asymptotic) HPLB. For instance, $\hat{\lambda} = 0$ is a valid HPLB, according to Definition 1. Consequently, for $\varepsilon \in (0,1]$, we study whether for a given (asymptotic) HPLB $\hat{\lambda}$,

$$\mathbb{P}(\hat{\lambda} > (1-\varepsilon)\lambda) \to 1, \tag{C_ε}$$

as $N \to \infty$. This entails several cases: if $\varepsilon = 1$, then (C_ε) means the (detection) power goes to 1. If (C_ε) is true for *all* $\varepsilon \in (0,1]$, it corresponds to consistency of $\hat{\lambda}$. One could also be interested in a non-trivial fixed ε, i.e. in detecting a fixed proportion of λ. In order to quantify the strength of a given (asymptotic) HPLB, we examine how fast λ may decay to zero with N, such that $\hat{\lambda}$ still exceeds a fraction of the the true λ with high probability. More precisely, we assume that the signal vanishes at a rate N^γ, for some $-1 < \gamma < 0$, i.e. $1 > \lambda := \lambda_N \asymp N^\gamma$, as $N \to \infty$. If for a given estimator $\hat{\lambda}$, $\varepsilon \in (0,1]$ and $-1 \leqslant \underline{\gamma}(\varepsilon) < 0$, (C_ε) is true for all $\gamma > \underline{\gamma}(\varepsilon)$, we write $\hat{\lambda}$ attains the rate $\underline{\gamma}(\varepsilon)$. To quantify the strength of an estimator $\hat{\lambda}$, we will study the smallest such rate $\underline{\gamma}(\varepsilon)$ it can attain for a given ε, denoted as $\underline{\gamma}^{\hat{\lambda}}(\varepsilon)$. Formally,

Definition 2. *For a given (asymptotic) HPLB $\hat{\lambda}$ and for $\varepsilon \in (0,1]$, we define* $\underline{\gamma}^{\hat{\lambda}}(\varepsilon) := \inf\{\gamma_0 \in [-1,0) :$ *for all $\gamma > \gamma_0$ and $\lambda \asymp N^\gamma$, $\mathbb{P}(\hat{\lambda} > (1-\varepsilon)\lambda) \to 1$}$\}$.

Of course, attaining the true λ, might be unrealistic in general. In such cases it is also possible to regard λ as the total variation distance between the two distributions after the projection through ρ, $\mathrm{TV}(\rho_\# P, \rho_\# Q)$, as described in more detail in Section 3.

In the following, we aim to construct informative (asymptotic) HPLBs for $\mathrm{TV}(P,Q)$. To put the previously introduced rates into perspective, we first introduce an "optimal" or *oracle rate*. In Section 2.2 we introduce binary classification asymptotic HPLBs focusing on the fixed cutoff $1/2$. Section 2.3 will then introduce a more data adaptive asymptotic HPLB that indeed considers the supremum over all available cutoffs in the sample.

2.1 Oracle rate

In light of (3) and the notation introduced in the last section, for $(t_N)_{N\geqslant 1} \subset I$ a (nonrandom) sequence of cutoffs, we define the estimator

$$\hat{\lambda}^\rho(t_N) = \hat{F}_m(t_N) - \hat{G}_n(t_N) - q_{1-\alpha}\sigma(t_N), \tag{7}$$

where $\sigma(t)$ is the theoretical standard deviation of $\hat{F}_m(t) - \hat{G}_n(t)$,

$$\sigma(t) = \sqrt{\frac{F(t)(1-F(t))}{m} + \frac{G(t)(1-G(t))}{n}}. \tag{8}$$

Using (3), it can be shown that:

Proposition 3. *Let $\lambda > 0$. For any sequence $(t_N)_{N\geqslant 1} \subset I$ of cutoffs, $\hat{\lambda}^{\rho}(t_N)$ defined in* (7) *is an asymptotic HPLB of λ (at level α) for any $\rho : \mathcal{X} \to I$.*

The condition $\lambda > 0$ arises from a technicality – for $\lambda = 0$, one can construct a sequence $(t_N)_{N\geqslant 1}$ such that the level cannot be conserved. Since $\hat{\lambda}^{\rho}(t_N)$ only serves as a theoretical tool, this is not an issue. However the same problem will arise later in Section 3.2.

Naturally, the performance of $\hat{\lambda}^{\rho}(t_N)$ will differ depending on the choice of the sequence $(t_N)_{N\geqslant 1}$ and the choice of ρ. Ideally we would like to choose the "optimal sequence" $(t_N^*)_{N\geqslant 1}$ to reach the lowest rate $\underline{\gamma}$ possible. We might even want to attain the smallest possible rate $\underline{\gamma}$ if we are able to freely choose $(t_N)_N$ for each given $\gamma > \underline{\gamma}$. This rate is technically the rate obtained by a collection of estimators, whereby for each $\gamma > \underline{\gamma}$ a potentially different estimator $\hat{\lambda}(t_N(\gamma))$ may be used. More formally, given $\varepsilon \in (0,1]$, let for the following the *oracle rate* $\underline{\gamma}^{\text{oracle}}(\varepsilon)$ be the smallest rate such that for all $\gamma > \underline{\gamma}^{\text{oracle}}(\varepsilon)$ there exists a sequence $(t_N)_{N\geqslant 1} \subset I$ such that (C_ε) is true for $\hat{\lambda} = \hat{\lambda}^{\rho}(t_N)$. If there exists a sequence $(t_N^*)_{N\geqslant 1} \subset I$ independent of $\gamma > \underline{\gamma}^{oracle}(\varepsilon)$, we may define the *oracle estimator*

$$\hat{\lambda}^{\rho}_{\text{oracle}} = \hat{F}_m(t_N^*) - \hat{G}_n(t_N^*) - q_{1-\alpha}\sigma(t_N^*). \tag{9}$$

In this case $\underline{\gamma}^{oracle}(\varepsilon)$ is the smallest rate attained by $\hat{\lambda}^{\rho^*}_{oracle}$ for a given ε. Clearly, $\underline{\gamma}^{\text{oracle}}(\varepsilon)$ depends on ρ as well, whenever $\rho = \rho^*$ in (4), the dependence on ρ is omitted.

Since $(t_N^*)_{N\geqslant 1}$ corresponds to a specific nonrandom sequence, Proposition 3 ensures that $\hat{\lambda}^{\rho}_{oracle}$ is an asymptotic HPLB. Clearly, even if $\hat{\lambda}^{\rho}_{oracle}$ is defined, it will not be available in practice, as $(t_N^*)_{N\geqslant 1}$ is unknown. However, the oracle rate it attains should serve as a point of comparison for other asymptotic HPLBs. We close this section by considering an example:

Example 1. *Let P, Q be defined by $P = p_N P_0 + (1-p_N)Q_0$ and $Q = (1-p_N)P_0 + p_N Q_0$, where $p_N \in [0,1]$ and P_0, Q_0 have a uniform distribution on $[-1,0]$ and $[0,1]$ respectively. In this example, only p_N is allowed to vary with N, while P_0, Q_0 stay fixed. If we assume $p_N > 0.5$, $\lambda_N = 2p_N - 1$ and $\lambda \asymp N^{\gamma}$ iff $p_N - 1/2 \asymp N^{\gamma}$. Thus*

Proposition 4. *For the setting of Example 1, assume $p_N > 0.5$ for all N, and $p_N - 1/2 \asymp N^{\gamma}$. Then $\underline{\gamma}^{oracle}(\varepsilon) = -1/2$ for all $\varepsilon \in (0,1]$. This rate is attained by the oracle estimator in* (9) *with $t_N^* = 1/2$ for all N.*

2.2 Binary classification bound

Let us fix a cutoff $t \in [0,1]$. From the binary classifier $\rho_t(Z) = I\{\rho(Z) > t\}$ we can define the *in-class accuracies* as $A_0^{\rho}(t) := P(\rho_t(X) = 0)$ and $A_1^{\rho}(t) := Q(\rho_t(Y) = 1)$. From there, relation (3) can be written in a more intuitive form:

$$TV(P,Q) \geqslant \sup_{t\in[0,1]} \left[A_0^{\rho}(t) + A_1^{\rho}(t)\right] - 1. \tag{10}$$

Thus, the (adjusted) maximal sum of in-class accuracies for a given classifier is still a lower bound on $\lambda = TV(P,Q)$. As it can be shown that the inequality in (10) is an equality for $\rho = \rho^*$ and $t = 1/2$, it seems sensible to build an estimator based on $A_0^\rho(1/2) + A_1^\rho(1/2) - 1$. Define the in-class accuracy estimators $\hat{A}_0^\rho(1/2) = \frac{1}{m}\sum_{i=1}^m (1 - \rho_{1/2}(X_i))$ and $\hat{A}_1^\rho(1/2) = \frac{1}{n}\sum_{j=1}^n \rho_{1/2}(Y_j)$. It follows as in Proposition 3, that:

Proposition 5. $\hat{\lambda}^\rho_{bayes} := \hat{A}_0^\rho(1/2) + \hat{A}_1^\rho(1/2) - 1 - q_{1-\alpha}\hat{\sigma}(1/2)$ *with*

$$\hat{\sigma}(1/2) = \sqrt{\frac{\hat{A}_0^\rho(1/2)(1-\hat{A}_0^\rho(1/2))}{m} + \frac{\hat{A}_1^\rho(1/2)(1-\hat{A}_1^\rho(1/2))}{n}} \tag{11}$$

is an asymptotic HPLB of λ (at level α) for any $\rho : \mathcal{X} \to [0,1]$.

It should be noted that if $\hat{\sigma}(1/2)$ in $\hat{\lambda}^\rho_{bayes}$ is replaced by $\sigma(1/2)$ in (8), we obtain $\hat{\lambda}^\rho(1/2)$. Consequently, it should be the case that if $t_N^* = 1/2$, the rate attained by $\hat{\lambda}^\rho_{bayes}$ is the oracle rate. We now demonstrate this in an example:

Example 2. *Compare a given distribution Q with the mixture $P = (1-\delta_N)Q + \delta_N C$, where C serves as a "contamination" distribution and $\delta_N \in (0,1)$. Then, $TV(P,Q) = \delta_N TV(C,Q)$. If we furthermore assume that Q and C are disjoint, then $TV(C,Q) = 1$ and $\lambda_N = \delta_N$. Then the oracle rate $\underline{\gamma}^{oracle}(\varepsilon)$ and $\underline{\gamma}^{\hat{\lambda}_{bayes}}(\varepsilon)$ coincide:*

Proposition 6. *For the setting of Example 2, $\underline{\gamma}^{\hat{\lambda}_{bayes}}(\varepsilon) = \underline{\gamma}^{oracle}(\varepsilon) = -1$, for all $\varepsilon \in (0,1]$.*

Indeed, it can be shown that here $\hat{\lambda}^{\rho^*}(1/2)$ gives rise to the oracle estimator in (9)*. It may therefore not be surprising that $\hat{\lambda}^{\rho^*}_{bayes}$ attains the rate $\underline{\gamma}^{oracle}(\varepsilon)$. A small simulation study illustrating Proposition 6 is given in Figure 6 in 1.1.*

While $\hat{\lambda}^\rho_{bayes}$ is able to achieve the oracle rate in some situations, it may be improved: Taking a cutoff of 1/2, while sensible if no prior knowledge is available, is sometimes suboptimal. This is true, even if ρ^* is used, as we demonstrate with the following example:

Example 3. *Define P and Q by $P = p_1C_1 + (1-p_1)p_2P_0 + (1-p_1)(1-p_2)Q_0$ and $Q = p_1C_2 + (1-p_1)p_2Q_0 + (1-p_1)(1-p_2)P_0$, where C_1, C_2, P_0, Q_0 are probability measures with disjoint support and $p_1, p_2 \in (0,1)$.*

Proposition 7. *For the setting of Example 3, let $p_2 > 0.5$, $p_2 = 0.5 + o(N^{-1})$. Then the oracle rate is $\underline{\gamma}^{oracle}(\varepsilon) = -1$, while $\hat{\lambda}^{\rho^*}_{bayes}$ attains the rate $\underline{\gamma}^{\hat{\lambda}_{bayes}}(\varepsilon) = -1/2$, for all $\varepsilon \in (0,1]$.*

It can be shown that choosing $t_N = 0$ for all N, leads to the oracle rate of -1 in this example. This is entirely missed by $\hat{\lambda}^{\rho^*}_{bayes}$.

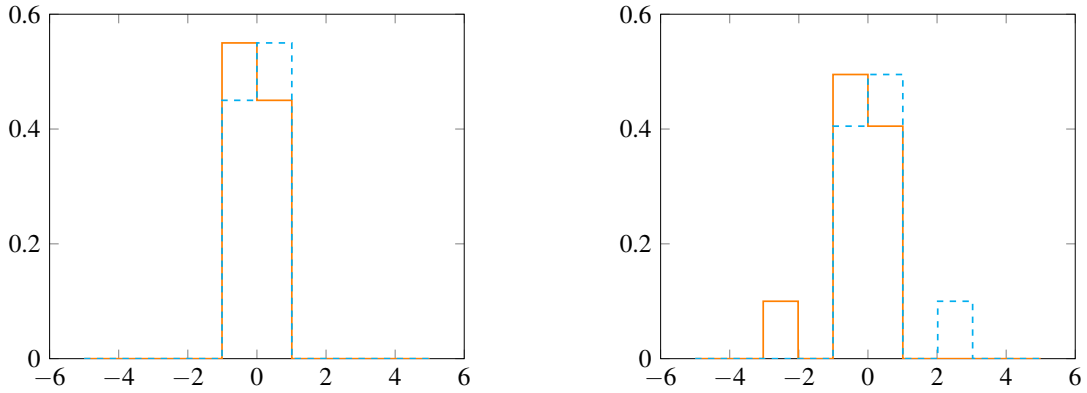


Figure 2: Illustration of Examples 1 and 3. Left: Illustration of Example 1 using $p_N = 0.55$. Right: Illustration of Example 3 using $p_1 = 0.1$, $p_2 = 0.55$ and uniform distributions for C_1, C_2, P_0 and Q_0.

Importantly, $\hat{\lambda}^{\rho^*}_{bayes}$ could still attain the oracle rate in Example 3, if the cutoff of $1/2$ was adapted. In particular, using $\hat{\lambda}^{\rho^*}_{bayes}$ with the decision rule $\rho^*_0(z) = I\{\rho^*(z) > 0\}$ would identify only the examples drawn from C_1 as belonging to class 0. This in turn, would lead to the desired detection rate. Naturally, this cutoff requires prior knowledge about the problem at hand, which is usually not available. In general, if ρ is any measurable function, potentially obtained by training a classifier or regression function on independent data, a cutoff of 1/2 might be strongly suboptimal. We thus turn our attention to an HPLB of the supremum in (3) directly.

2.3 Adaptive binary classification bound

In light of relation (3), we aim to directly account for the randomness of $\sup_t(\hat{F}_m(t) - \hat{G}_n(t)) = \sup_z(\hat{F}_m(\rho_{(z)}) - \hat{G}_n(\rho_{(z)}))$. We follow [52] and define the counting function $V_{m,z} = m\hat{F}_m(\rho_{(z)})$ for each $z \in J_{m,n} := \{1, \ldots, m+n-1\}$. Using $m\hat{F}_m(\rho_{(z)}) + n\hat{G}_n(\rho_{(z)}) = z$, it is possible to write:

$$\hat{F}_m(\rho_{(z)}) - \hat{G}_n(\rho_{(z)}) = \frac{m+n}{mn}\left(V_{m,z} - \frac{mz}{m+n}\right). \tag{12}$$

A well-know fact (see e.g., [52]) is that under $H_0 : F = G$, $V_{m,z}$ is a hypergeometric random variable, obtained by drawing without replacement z times from an urn that contains m circles and n squares and counting the number of circles drawn. We denote this as $V_{m,z} \sim \text{Hypergeometric}(z, m+n, m)$ and simply refer to the resulting process $z \mapsto V_{m,z}$ as the *hypergeometric process*. Though the distribution of $V_{m,z}$ under a general alternative is not known, we will now demonstrate that one can nonetheless control its behavior, at least asymptotically. We start with the following definition, inspired by [119]:

Definition 8 (Bounding function)**.** *A function* $J_{m,n} \times [0,1] \ni (z, \tilde{\lambda}) \mapsto Q_{m,n,\alpha}(z, \tilde{\lambda})$ *is called a* bounding function *at level* α *if*

$$\mathbb{P}(\sup_{z \in J_{m,n}} [V_{m,z} - Q_{m,n,\alpha}(z, \lambda)] > 0) \leqslant \alpha. \tag{13}$$

It will be called an asymptotic bounding function *at level α if instead*

$$\limsup_{N\to\infty} \mathbb{P}(\sup_{z\in J_{m,n}} [V_{m,z} - Q_{m,n,\alpha}(z,\lambda)] > 0) \leqslant \alpha. \tag{14}$$

In other words, for the true value λ, $Q_{m,n,\alpha}(z,\lambda)$ provides an (asymptotic) type 1 error control for the process $z \mapsto V_{m,z}$ (often the dependence on α will be ommited). For $\lambda = 0$ the theory in [52] shows that such an asymptotic bounding function is given by

$$Q_{m,n,\alpha}(z,\tilde{\lambda}) = Q_{m,n,\alpha}(z,0) = \frac{zm}{m+n} + \beta_{\alpha,m} w(z,m,n),$$

with

$$w(z,m,n) = \sqrt{\frac{m}{N}\frac{n}{N}\frac{N-z}{N-1}z}. \tag{15}$$

Assuming access to a bounding function, we can define the estimator presented in Proposition 9.

Proposition 9. *Let $Q_{m,n,\alpha}$ be an (asymptotic) bounding function and define,*

$$\hat{\lambda}^{\rho} = \inf\left\{\tilde{\lambda} \in [0,1] : \sup_{z\in J_{m,n}} \left[V_{m,z} - Q_{m,n,\alpha}(z,\tilde{\lambda})\right] \leqslant 0\right\}. \tag{16}$$

Then $\hat{\lambda}^{\rho}$ is an (asymptotic) HPLB of λ (at level α) for any $\rho : \mathcal{X} \to I$.

The proof of Proposition 9 is given in 3.1.

Since we do not know the true λ, the main challenge in the following is to find bounding functions that would be valid for any potential $\lambda \geqslant 0$. We now introduce a particular type of such a bounding function. With $\tilde{\lambda} \in [0,1]$, $\alpha \in (0,1)$, $m(\tilde{\lambda}) = m - q_{1-\frac{\alpha}{3}}(\tilde{\lambda}, m)$, and $n(\tilde{\lambda}) = n - q_{1-\frac{\alpha}{3}}(\tilde{\lambda}, n)$, we define

$$Q_{m,n,\alpha}(z,\tilde{\lambda}) = \begin{cases} z, \text{ if } 1 \leqslant z \leqslant q_{1-\frac{\alpha}{3}}(\tilde{\lambda}, m) \\ m, \text{ if } m + n(\tilde{\lambda}) \leqslant z \leqslant m+n \\ q_{1-\frac{\alpha}{3}}(\tilde{\lambda}, m) + \mathrm{s}q_{\alpha/3}\left(\tilde{V}_{m(\tilde{\lambda}),z}, z \in \{1,\ldots,m(\tilde{\lambda}) + n(\tilde{\lambda}) - 1\}\right), \text{ otherwise} \end{cases} \tag{17}$$

where $\tilde{V}_{m,z}$ denotes the counting function of a hypergeometric process and $\mathrm{s}q_{\alpha}\left(\tilde{V}_{m,z}, z \in J\right)$ is a *simultaneous* confidence band, such that

$$\limsup_{N\to\infty} \mathbb{P}\left(\sup_{z\in J}\left[\tilde{V}_{m,z} - \mathrm{s}q_{\alpha}\left(\tilde{V}_{m,z}, z \in J\right)\right] > 0\right) \leqslant \alpha. \tag{18}$$

Note that Equation (18) includes the case $\mathbb{P}\left(\sup_{z\in J}\left[\tilde{V}_{m,z} - \mathrm{s}q_{\alpha}\left(\tilde{V}_{m,z}, z \in J\right)\right] > 0\right) \leqslant \alpha$ for all N. Depending on which condition is true, we obtain a bounding function or an asymptotic bounding function:

Proposition 10. *$Q_{m,n,\alpha}$ as defined in* (17) *is an (asymptotic) bounding function.*

The proof of Proposition 10 is given in 3.1.

A valid analytical expression for $\mathrm{sq}_{\alpha/3}\left(\tilde{V}_{m(\tilde{\lambda}),z}, z \in \{1,\ldots,m(\tilde{\lambda})+n(\tilde{\lambda})-1\}\right)$ in (17) based on the theory in [52] is given in Equation (28) of 2. We will denote the asymptotic bounding function when combining (17) with (28) by Q^A. The asymptotic HPLB that arises from (16) with projection ρ and bounding function Q^A will be referred to as $\hat{\lambda}^{\rho}_{adapt}$. Alternatively, we may choose $\mathrm{sq}_{\alpha/3}\left(\tilde{V}_{m(\tilde{\lambda}),z}, z \in \{1,\ldots,m(\tilde{\lambda})+n(\tilde{\lambda})-1\}\right)$ by simply simulating S times from the process $\tilde{V}_{m(\tilde{\lambda}),z}, z = 1,\ldots,m(\tilde{\lambda})+n(\tilde{\lambda})$. For $S \to \infty$, condition (18) then clearly holds true. This is especially important, for smaller sample sizes, where the (asymptotic) Q^A could be a potentially bad approximation.

We close this section by considering once again the introductory example in Section 1.1. Our two proposed estimators applied to this example give $\hat{\lambda}^{\rho}_{bayes} = 0$ and $\hat{\lambda}^{\rho}_{adapt} = 0.0022$. Thus, as one would expect from the permutation test results, $\hat{\lambda}^{\rho}_{adapt}$ is able to detect a difference, whereas $\hat{\lambda}^{\rho}_{bayes}$ is not. While it is difficult in this case to determine the true λ, we can show for another example, that $\hat{\lambda}^{\rho^*}_{adapt}$ attains the rate $\hat{\lambda}^{\rho^*}_{bayes}$ could not:

Proposition 11. *Let P, Q be defined as in Example 3 with $p_2 > 0.5$, $p_2 = 0.5 + o(N^{-1})$. Then $\underline{\gamma}^{\hat{\lambda}_{adapt}}(\varepsilon) = \underline{\gamma}^{oracle}(\varepsilon) = -1$, independent of ε.*

A small simulation study illustrating Propositions 7 and 11 is given in Figure 6 in 1.1. In the next section, we generalize the results in Examples 2 and 3 and show that $\hat{\lambda}^{\rho}_{adapt}$ *always* attains the oracle rate.

3 Theoretical guarantees

This section studies some of the theoretical properties of our proposed lower-bounds. We start in Section 3.1 by assuming access to the "ideal" classifier ρ^* and show that in this case, the $\hat{\lambda}^{\rho^*}_{adapt}$ can asymptotically detect a nonzero TV with a better rate than $\hat{\lambda}^{\rho^*}_{bayes}$. More generally, our main results in Proposition 12 and 14 show that $\hat{\lambda}^{\rho^*}_{adapt}$ achieves the same asymptotic performance as $\hat{\lambda}^{\rho^*}$, which is free to "choose" a sequence of cutoffs $(t_N)_N$. Though we use ρ^* for simplicity, all of the results in this section also hold true for any arbitrary (fixed) $\rho : \mathcal{X} \to [0,1]$, and also if we replace λ by the TV distance on the projection,

$$\lambda(\rho) = \sup_{t\in[0,1]}\left[A^{\rho}_0(t) + A^{\rho}_1(t)\right] - 1 := \sup_{t\in[0,1]}\left[P(\rho(X) \leqslant t) - Q(\rho(Y) \leqslant t)\right] \tag{19}$$

such that $\lambda := \lambda(\rho^*) = TV(P,Q)$.

Section 3.2 then extends the main result of Section 3.1 from ρ^* to a sequence $\hat{\rho} = \hat{\rho}_N$, estimated on independent training data, showing that $\hat{\lambda}^{\hat{\rho}}_{adapt}$ and $\hat{\lambda}^{\hat{\rho}}$ have the same asymptotic

detection power. Finally, we discuss sufficient conditions for the consistency of $\hat{\lambda}^{\hat{\rho}}_{adapt}$. We restrict to $I = [0,1]$ throughout this section.

3.1 Using ρ^*

We start by studying the asymptotic properties of the proposed asymptotic HPLB estimators, assuming access to ρ^* in (4). Recall that for a fixed $\varepsilon \in (0,1]$, $\underline{\gamma}^{oracle}(\varepsilon)$ was defined as the minimal rate such that for all $\gamma > \underline{\gamma}^{oracle}(\varepsilon)$ there exists a sequence $(t_N)_{N\geqslant 1} \subset I$ such that (C_ε), i.e.

$$\mathbb{P}(\hat{\lambda} > (1-\varepsilon)\lambda) \to 1,$$

is true for $\hat{\lambda} = \hat{\lambda}^{\rho^*}(t_N)$. Consider for $\varepsilon \in (0,1]$ the following conditions on $(t_N)_{N\geqslant 1}$:

$$\liminf_{N\to\infty} \frac{\lambda(t_N)}{\lambda_N} \geqslant 1-\varepsilon, \tag{20}$$

and

$$\lim_N \frac{\lambda_N}{\sigma(t_N)} = \infty, \text{ if } \liminf_{N\to\infty} \frac{\lambda(t_N)}{\lambda_N} > 1-\varepsilon, \tag{21a}$$

$$\lim_N \frac{\lambda_N}{\sigma(t_N)} \left(\frac{\lambda(t_N)}{\lambda_N} - (1-\varepsilon) \right) = \infty, \text{ if } \liminf_{N\to\infty} \frac{\lambda(t_N)}{\lambda_N} = 1-\varepsilon, \tag{21b}$$

where $\lambda(t_N) := F(t_N) - G(t_N)$ and $\sigma(t)$ is defined as in (8). We then refer to Condition (21), iff (21a) and (21b) are true:

$$\text{(21a) and (21b) hold.} \tag{21}$$

We now redefine $\underline{\gamma}^{oracle}(\varepsilon)$ as the smallest element of $[0,1]$ with the property that for all $\gamma > \underline{\gamma}^{oracle}(\varepsilon)$ there exists a sequence $(t_N)_{N\geqslant 1} \subset I$ such that (20) and (21) are true. Intuitively, this means that a given rate is achieved for $(t_N)_{N\geqslant 1}$ if either $F(t_N) - G(t_N)$ is strictly larger than $(1-\varepsilon)\lambda_N$ and the variance decreases fast relative to λ_N (Condition (20) and (21a)), or $F(t_N) - G(t_N)$ is exactly equal to $(1-\varepsilon)\lambda_N$ in the limit, which needs to be balanced by an even faster decrease in the variance $\sigma(t_N)$ (Condition (20) and (21b)). As a side remark, (21b) is problematic for $\varepsilon = 1$, if $\sigma(t_N) = 0$ for infinitely many N. In this case, it should be understood that (21b) is taken to be false.

The following proposition confirms that the two definitions of $\underline{\gamma}^{oracle}(\varepsilon)$ coincide:

Proposition 12. *Let $-1 < \gamma < 0$ and $\varepsilon \in (0,1]$ fixed. Then there exists a $(t_N)_{N\geqslant 1}$ such that (C_ε) is true for $\hat{\lambda}^{\rho^*}(t_N)$ iff there exists a $(t_N)_{N\geqslant 1}$ such that* (20) *and* (21) *are true.*

If we consider a classifier with cutoff $t \in I$, $\rho_t(z) = I\{\rho(z) > t\}$ and, as in Section 2.2, define in-class accuracies $A_0(t) := A_0^{\rho^*}(t)$, $A_1(t) := A_1^{\rho^*}(t)$, we may rewrite $\sigma(t)$ in a convenient form

$$\sigma(t) = \sqrt{\frac{A_0(t)(1-A_0(t))}{m} + \frac{A_1(t)(1-A_1(t))}{n}}. \tag{22}$$

Since $\sqrt{N}\lambda_N$ does not go to infinity for $\gamma \leqslant -1/2$, the divergence of the ratio in (21) is only achieved, if both $A_0(t_N)(1-A_0(t_N))$ and $A_1(t_N)(1-A_1(t_N))$ go to zero sufficiently fast. In our context, this is often more convenient to verify directly.

The binary classification estimator $\hat{\lambda}^{\rho^*}_{bayes}$ takes $t_N = 1/2$ and, since $F(1/2) - G_t(1/2) = \lambda_N$, (20) is true for any ε. Thus a given rate $\underline{\gamma}$ is achieved iff (21) is true for $t_N = 1/2$. This is stated formally in the following corollary:

Corollary 13. *$\hat{\lambda}^{\rho^*}_{bayes}$ attains the rate $\underline{\gamma}^{\hat{\lambda}_{bayes}}(\varepsilon) = \underline{\gamma}$ for all $\varepsilon \in (0,1]$, iff* (21) *is true for $t_N = 1/2$ and all $\gamma > \underline{\gamma}$.*

The proof is a direct consequence of Proposition 12 and is given in 3. We thus write $\underline{\gamma}^{\hat{\lambda}_{bayes}}$ instead of $\underline{\gamma}^{\hat{\lambda}_{bayes}}(\varepsilon)$. It should be noted (21) is always true for $\gamma > -1/2$. As such, $\underline{\gamma}^{\hat{\lambda}_{bayes}} \geqslant -1/2$ and only the case of $\underline{\gamma} < -1/2$ is interesting in Corollary 13.

Finally, the adaptive binary classification estimator $\hat{\lambda}^{\rho^*}_{adapt}$ always reaches at least the rate $\underline{\gamma} = -1/2$. In fact, it turns out that it attains the oracle rate:

Proposition 14. *Let $-1 < \gamma < 0$ and $\varepsilon \in (0,1]$ fixed. Then (C_ε) is true for $\hat{\lambda}^{\rho^*}_{adapt}$ iff there exists a $(t_N)_{N\geqslant 1}$ such that* (20) *and* (21) *are true.*

This immediately implies:

Corollary 15. *For all $\varepsilon \in (0,1]$, $\underline{\gamma}^{oracle}(\varepsilon) = \underline{\gamma}^{\hat{\lambda}_{adapt}}(\varepsilon)$.*

The next section shows that this result can be generalized to $\hat{\rho}$ estimated from independent data.

3.2 Estimated ρ

In this section we assume that $\hat{\rho}$ is a "probability function" in $[0,1]$, estimated from data. In that case *sample-splitting* should be used, i.e. the function $\hat{\rho}$ is estimated independently on a *training set* using a learning algorithm which is then used to compute an (asymptotic) HPLB based on an independent *test set*. Sample-splitting is important to avoid spurious correlation between ρ and the (asymptotic) HPLB, not supported by our theory. Formally we assume,

(E1) $\hat{\rho} = \hat{\rho}_{N_{tr}}$ is trained on a sample of size N_{tr}, $(Z_1,\ell_1),\ldots,(Z_{N_{tr}},\ell_{N_{tr}})$, and evaluated on an independent sample $(Z_1,\ell_1),\ldots,(Z_{N_{te}},\ell_{N_{te}})$, with $N_{tr}+N_{te}=N$,

(E2) $N_{te}, N_{tr} \to \infty$, as $N \to \infty$, with $m_{te}/N_{te} \to \pi \in (0,1)$,

where as before m_j denotes the number of draws from P (with label 0) and n_j the number of draws from Q (with label 1), for $j \in \{te, tr\}$.

In practice, most probability estimates try to approximate the Bayes probability (see e.g., [40]):

$$\rho^B(z) = \frac{(1-\pi)g(z)}{\pi f(z) + (1-\pi)g(z)}, \tag{23}$$

with *Bayes classifier* $\rho^B_{1/2}(z) = I\{\rho^B(z) > 1/2\}$. It is the classifier resulting in the maximal overall accuracy, denoted the *Bayes accuracy*: $\pi A_0^{\rho^B}(1/2) + (1-\pi)A_1^{\rho^B}(1/2)$. Clearly, $\rho^B = \rho^*$ for $\pi = 1/2$. More generally, it can be shown that $\rho^B_{1-\pi}(z) = \rho^*_{1/2}(z)$.

Let as before, $\hat{\lambda}^{\hat{\rho}}_{adapt}$ be the estimator obtained when using $\hat{\rho}$ and $\lambda(\hat{\rho})$ be defined as in (19) for $\rho = \hat{\rho}$. Similarly, for a sequence $(t_N)_{N\geqslant 1} \subset I$, we define $\hat{\lambda}^{\hat{\rho}}(t_N)$ to be the theoretical estimator (7) using $\hat{\rho}$. Conditioning on the training data through $\hat{\rho}$, allows for a generalization of the theory in Section 3.1 to estimated ρ.

The first step, is to extend the theory in Section 3.1 to the case of arbitrary (nonrandom) sequences $(\rho_N)_{N\geqslant 1}$. While the proofs of the results in Section 3.1 are applicable almost one-to-one in this case, there is one issue arising from the estimator $\hat{\lambda}^{\rho}_{bayes}$. We exemplify this in the following:

Example 4. *Assume both $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ uniform on $[0,1]$ and ρ_N such that for some $C > 0$,*

$$\rho_{N,1/2}(Z) = \begin{cases} 0, & \text{if } Z \in [0, C/n] \\ 1, & \text{else} \end{cases}$$

Then:

Proposition 16. *For the setting of Example 4, let ξ_1, ξ_2 be independently Poisson distributed, with mean C. Then*

$$\mathbb{P}(\hat{\lambda}^{\rho_N}(1/2) > 0) \to \mathbb{P}(\xi_1 - \xi_2 > q_{1-\alpha}\sqrt{2C}).$$

It can be shown numerically that $\mathbb{P}(\xi_1 - \xi_2 > q_{1-\alpha}\sqrt{2C}) > 1-\alpha$, for some C. Thus, $\hat{\lambda}^{\rho_N}(1/2)$ is not a valid asymptotic HPLB.

The case above appears rather exotic, and might not be realistic. What is more, we used $\hat{\lambda}^{\rho_N}(1/2)$ in the above example, with the true variance included, instead of $\hat{\lambda}^{\rho_N}_{bayes}$. In this case, the accuracies $A_0^{\rho_N}(1/2), A_1^{\rho_N}(1/2)$ cannot even be estimated reliably, so it is not clear what exactly will happen if $\sigma(1/2)$ is estimated. However none of these problems are of concern for $\hat{\lambda}^{\hat{\rho}}_{adapt}$, which conserves the level in any case:

Proposition 17. *Assume (E1) and (E2). Then $\hat{\lambda}^{\hat{\rho}}_{adapt}$ is an (asymptotic) HPLB of λ (at level α).*

Thus, we will focus in this section only on the adaptive estimator. We first generalize Propositions 12 and 14 to this case.

Proposition 18. *Let $-1 < \gamma \leqslant 0$ and $\varepsilon_1 \in (0,1]$ fixed. Assume that $\lambda_N = N_{te}^{\gamma}$ and that (E1) and (E2) hold. Then the following is equivalent*

(i) there exists a sequence $(t_{N_{te}})_{N_{te}\geqslant 1}$ such that (C_ε) is true for $\hat{\lambda}^{\hat{\rho}}(t_{N_{te}})$,

(ii) (C_ε) is true for $\hat{\lambda}^{\hat{\rho}}_{adapt}$.

The main message of Proposition 18 is that for $\hat{\rho}$ estimated on independent training data, $\hat{\lambda}^{\hat{\rho}}_{adapt}$ still has the same asymptotic performance as an estimator that is free to choose its cutoff for any given sample size. And this holds despite the fact that even for $\lambda = 0$, $\hat{\lambda}^{\hat{\rho}}_{adapt}$ is a valid HPLB, which is not clear for $\hat{\lambda}^{\hat{\rho}}(t_N)$, as seen in Example 4.

In practice, one might be more interested under what conditions $\hat{\lambda}^{\hat{\rho}}_{adapt}$ is consistent for a fixed λ. To answer this question, we first restate consistency for a sequence of classifiers, assuming λ does not change:

Definition 19. *A sequence of classifiers $\hat{\rho}_{N,t_N} = \hat{\rho}_{t_N}$, $N \in \mathbb{N}$, is* consistent*, if*

$$A_0^{\hat{\rho}_N}(t_N) + A_1^{\hat{\rho}_N}(t_N) \xrightarrow{p} A_0^{\rho^*}(1/2) + A_1^{\rho^*}(1/2).$$

This is the standard definition of consistency, see e.g., [40, Definition 6.1], with two small modifications: We consider accuracies instead of classification errors and instead of the Bayes accuracy in the limit, we consider the equally weighted accuracy of ρ^*. As such the definition is a special case of the Ψ-consistency of [129].

A simple consequence of Proposition 18 is that a classifier that is consistent for the equally weighted sum of in-class errors, also leads to a consistent estimate of λ.

Corollary 20. *Assume that λ is fixed and that there exists a sequence $(t_{N_{tr}})$, such that the sequence of classifiers $\hat{\rho}_{N_{tr},t_{N_{tr}}}$ is consistent. Then (C_ε) is true for $\hat{\lambda}^{\hat{\rho}}_{adapt}$, for all $\varepsilon > 0$.*

In essence, for a given sequence of estimated $\hat{\rho}$, it is enough that there exists a sequence of cutoffs leading to a consistent classifier, for $\hat{\lambda}_{adapt}$ to be consistent. As is well-known (see e.g., [40] and [129]),

Lemma 7.1. *Assume that $m_{tr}/N_{tr} \to \pi \in (0,1)$ and that*

$$\mathbb{E}[|\hat{\rho}_{N_{tr}} - \rho^B| \mid \hat{\rho}_{N_{tr}}] \xrightarrow{p} 0. \tag{24}$$

Then $\hat{\rho}_{N_{tr},t_{N_{tr}}}$ is consistent for $t_{N_{tr}} = m_{tr}/N_{tr}$.

This is a relatively straightforward sufficient condition for the consistency of $\hat{\lambda}^{\hat{\rho}}_{adapt}$. We would like to note however that, as shown in [40] for the Bayes classifier, (24) usually is much stronger than consistency of the classifier in Definition 19.

We close this section with two examples:

Example 5. *Assume access to the Bayes classifier ρ^B and $\pi \neq 1/2$. In this case, [68] showed that no test based on $\rho^B_{1/2}$ has power higher than its level. In our case, this translate to an inconsistent estimate of $\lambda^{\rho^B}_{bayes}$. On the other hand, it is well-known that*

$$A_0^{\rho^B}(1-\pi) + A_1^{\rho^B}(1-\pi) = A_0^{\rho^*}(1/2) + A_1^{\rho^*}(1/2),$$

so the cutoff of $1-\pi$ fixes the issue and indeed leads to both a consistent estimate and a consistent test.

Example 6. *Combining the arguments in [11, Theorem 3.1] and [40], if X, Y are supported on $[0,1]^d$ and $\hat{\rho}$ is a Random Forest using random splitting, then* (24) *holds. For an adapted version of this Random Forest, the result can also be extended to distributions of X, Y on $\mathbb{R}^d$, see [11, Theorem 3.2].*

Of course, as before, even if $\hat{\rho}$ is not consistent, we might still be able to detect a signal, giving us an indication of the strength of difference between two distributions. That is, the result of Proposition 18 and Corollary 20 hold more generally, as long as $\lambda(\hat{\rho})$ converges in probability to some $\lambda(\rho) \leqslant \lambda$, which may again be seen as the total variation distance on the projected space. In the next section, we move on from the question of consistency and study how one might find a ρ in practice in a more general framework.

4 Practical considerations

In this section we put the methodology introduced in Section 2 in practical perspectives. We first generalize our setting to allow for more flexible projections: Let $\mathcal{P} = \{P_t\,,\ t \in I\}$ be a family of probability measures defined on a measurable space $(\mathcal{X}, \mathcal{A})$ indexed by a totally ordered set $(I, \leq)$. We further assume to have a probability measure μ on I and independent observations $\mathcal{T} = \{(z_i, t_i)\}_{i=1}^N$ such that $z_i \sim P_{t_i}$ conditionally on $t_i \in I$ drawn from μ, for $1 \leqslant i \leqslant N$. Given $s \in I$, and a function $\rho : \mathcal{X} \to I$, we define two empirical distributions denoted $\hat{F}_{m,s}$ and $\hat{G}_{n,s}$ obtained from "cutting" the set of observations $\mathcal{T}$ at s. Namely if we assume that out of the $N = m+n$ observations, m of them have their index t_i smaller or equal to s and n strictly above, we have for $y \in I$

$$\hat{F}_{m,s}(y) = \frac{1}{m}\sum_{i=1}^{N} \mathbb{1}\{\rho(z_i) \leqslant y, t_i \leqslant s\}, \quad \text{and} \quad \hat{G}_{n,s}(y) = \frac{1}{n}\sum_{i=1}^{N} \mathbb{1}\{\rho(z_i) \leqslant y, t_i > s\}.$$

These empirical distributions correspond to the population mixtures $F_s \propto \int_{t \leqslant s} \rho_{\#} P_t \, d\mu(t)$ and $G_s \propto \int_{t > s} \rho_{\#} P_t \, d\mu(t)$. We will similarly denote the measures associated to F_s and G_s as P_s and Q_s respectively and will use the two notations interchangeably. Note that we assumed m, n deterministic so far, which changes in the above framework, where $m \sim \mathrm{Bin}(\pi, N)$, with

$\pi := \mathbb{P}(T \leqslant s)$. Still, with a conditioning argument, one can show that, whenever the level is guaranteed for nonrandom m, n, it will also be once m, n are random.

The question remains how to find a good ρ in practice. As our problems are framed as a split in the ordered elements of I, it always holds that one sample is associated with higher $t \in I$ than the other. Consequently, we have power as soon as we find a $\rho : \mathcal{X} \to I$ that mirrors the relationship between T and Z_T. It therefore makes sense to frame the problem of finding ρ as a loss minimization, where we try to minimize the loss of predicting $T \in I$ from $Z \in \mathcal{X}$: For a given split point s, consider ρ_s that solves

$$\rho_s := \arg\min_{h \in \mathbb{F}} \mathbb{E}[\mathcal{L}_s(h(Z), T)], \tag{25}$$

where $\mathbb{F}$ is a collection of functions $h : \mathcal{X} \to I$ and $\mathcal{L}_s : I \times I \to \mathbb{R}_+$ is some loss function. As before, we assume to have densities f_s, g_s, for P_s, Q_s respectively. For simplicity, we also assume that time is uniform on $I = [0, 1]$. As it is well-known, taking $\mathbb{F}$ to be all measurable functions $h : \mathcal{X} \to I$ and $\mathcal{L}(f(z), t) = \left(f(z) - I_{(s,1]}(t)\right)^2$, we obtain the supremum as

$$\rho_{1,s}(z) := \mathbb{E}[I_{(s,1]}(T)|z] = \frac{(1-s)g_s(z)}{sf_s(z) + (1-s)g_s(z)}, \tag{26}$$

which is simply the Bayes probability in (4). Taking instead $\mathcal{L}(f(z), t) = (f(z) - t)^2$, yields $\mathbb{E}[T|Z]$. Some simple algebra shows that if there is only *one* point of change s^*, i.e. T is independent of Z conditional on the event $T \leqslant s^*$ or $T > s^*$, $\mathbb{E}[T \mid z]$ can be expressed as:

$$\rho_{2,s*}(z) = \frac{1}{2}\left(s^* + \rho_{1,s*}(z)\right), \tag{27}$$

which is a shifted version of $\rho_{1,s*}(z)$. Contrary to $\rho_{1,s}$, the regression version $\rho_{2,s*}$ does not depend on the actual split point s we are considering.

In Section 5 we try to approximate (26) and (27) by using the Random Forest of [15]. That is, the function ρ is fitted on a training set using a learning algorithm which is then used to compute an (asymptotic) HPLB based on an independent test set, as in Section 3.2.

5 Numerical examples

Distributional change detection in climate is a topic of active research (see e.g. [158] and the references therein). We will demonstrate the estimator $\hat{\lambda}_{adapt}$ in three applications using the `NCEP_Reanalysis 2` data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their website at `https://www.esrl.noaa.gov/psd/`. The analyses were run using the R-package `HPLB` (see `https://github.com/lorismichel/HPLB`). We mention that the estimator $\hat{\lambda}_{bayes}$ gives comparable results and is ommited here. This dataset is a worldwide reanalysis containing daily observations of the 4 variables:

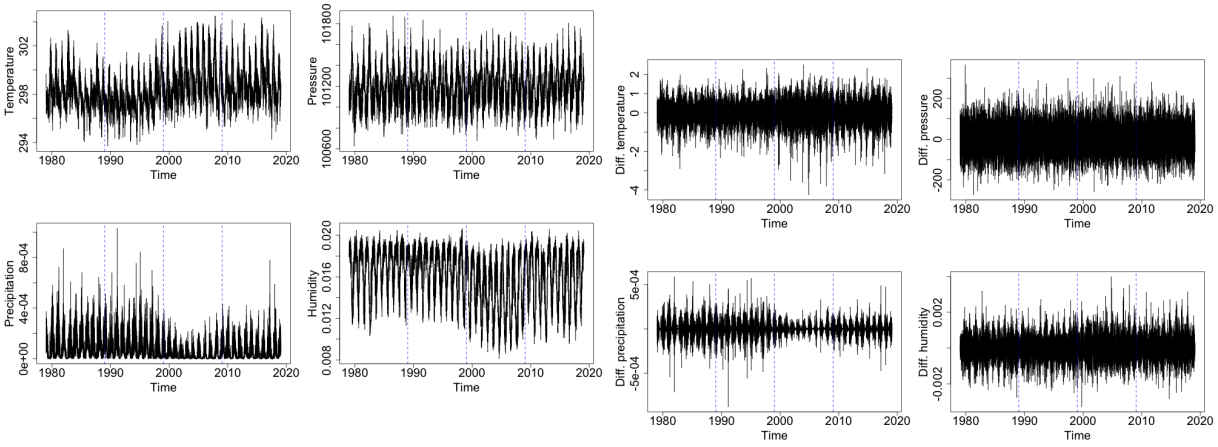


Figure 3: Temperature, pressure, precipitation and humidity at geo-coordinates (-45,-8) (Brazil) over the time period ranging from 1979 to 2019 (on the left). Corresponding differenced series (on the right). The vertical dashed blue lines are the breaks used in analysis (B).

- *air temperature (air)*: daily average of temperature at 2 meters above ground, measured in degree Kelvin;
- *pressure (press)*: daily average of pressure above sea level, measured in Pascal;
- *precipitation (prec)*: daily average of precipitation at surface, measured in kg per m^2 per second;
- *humidity (hum)*: daily average of specific humidity, measured in proportion by kg of air;

over a time span from 1^{st} of January 1979 to 31^{th} January 2019. Each variable is ranging not only over time, but also over $2'592$ locations worldwide, indexed by longitude and latitude coordinates, as (longitude, latitude). All variables are first-differenced to reduce dependency and seasonal effects before running the analyses. Figure 3 displayed the 4 time series corresponding to the geo-coordinates (-45,-8) (Brazil).

The potential changes in distribution present in this dataset could require a refined analysis and simple investigation for mean and/or variance shift might not be enough. Moreover, detecting "small" changes, as $\hat{\lambda}_{adapt}$ is designed to do, could be of interest. In addition, thanks to the equivalent characterization of TV explained in Section 1, $\hat{\lambda}_{adapt}$ represents the minimal percentage of days on which the distribution of the considered variables has changed. We present 2 types of analyses to illustrate the use of the (asymptotic) HPLBs introduced in this paper:

(A) ***temporal climatic change-map***: a study of the change of climatic signals between two periods of time (1^{st} of January 1979 to 15^{th} of January 1999 against 16^{th} of January 1999 to 31^{th} January 2019) across all $2'592$ locations.

(B) ***fixed-location change detection***: a study of the change of climatic signals over several time points for a fixed location.

For analysis (A), compare the first half (years 1979-1999) of the data with the second half (years 1999-2019) over all available locations. That is P_s corresponds to the distribution of the first half of the (differenced) data, while Q_s corresponds to the second. The projection ρ is chosen to be a Random Forest classification. To this end, sample-splitting is applied and the available time-span is equally divided into 4 consecutive time blocks, of which the middle two are used as a training set, while the remaining two are used as a test set. The goal is thereby, as with differencing, to reduce the dependence between observations due to the time-structure of the series. Figure 4 shows the results as a world heatmap. Interestingly, there is an area of very-high estimated TV values in the pacific ocean off the cost of South America. The water temperature in this area is indicative of El Niño.

Analysis (B) illustrates the mixture framework introduced in Section 4 in a time series context where the ordering is given by time. We analyse the change in distribution for the four climatic variables for 3 split points chosen uniformly over the time span. The location is thereby fixed to the coordinates (-45,-8) chosen from the analysis in (A). At each split point s, the distribution of the observations with time points below s is compared to the future observations. In the context of Section 4, a regression model predicting time is an option to quickly evaluate $\hat{\lambda}^{\rho}_{adapt}$ for several different splits. This corresponds to taking the squared error loss in Section 4. Here a Random Forest regression is used to predict time from the four variables. Each data point within a period defined by two splits is allocated into two sets (train and test) as follows: the first and last quartiles of the period are allocated to the training set, the rest (i.e. the middle part) is allocated to the testing set. Single splits through time can be then readily analyzed using $\hat{\lambda}_{adapt}$ on the test data. In addition to the analysis with real data here, 1.2 shows some simulations results.

Figure 5 summarizes the result of the analysis with split points s considered marked in Figure 3 by blue breaks: While Figure 3 indicates that some change might be expected even after differencing, this impression is only confirmed for precipitation in Figure 5. This hints at the fact that a shift is appearing in precipitation while for the other variables no change can marginally be detected. More interesting change is detectable once all 4 variables are considered jointly. This is illustrated on the right in Figure 5, where the estimated TV climbs to a (relatively) high value of around 0.14 between 1995 and and 2000. This corresponds to the high signal observed in Figure 4 for these coordinates, only that here, the regression approach leads to a slightly lower $\hat{\lambda}_{adapt}$.

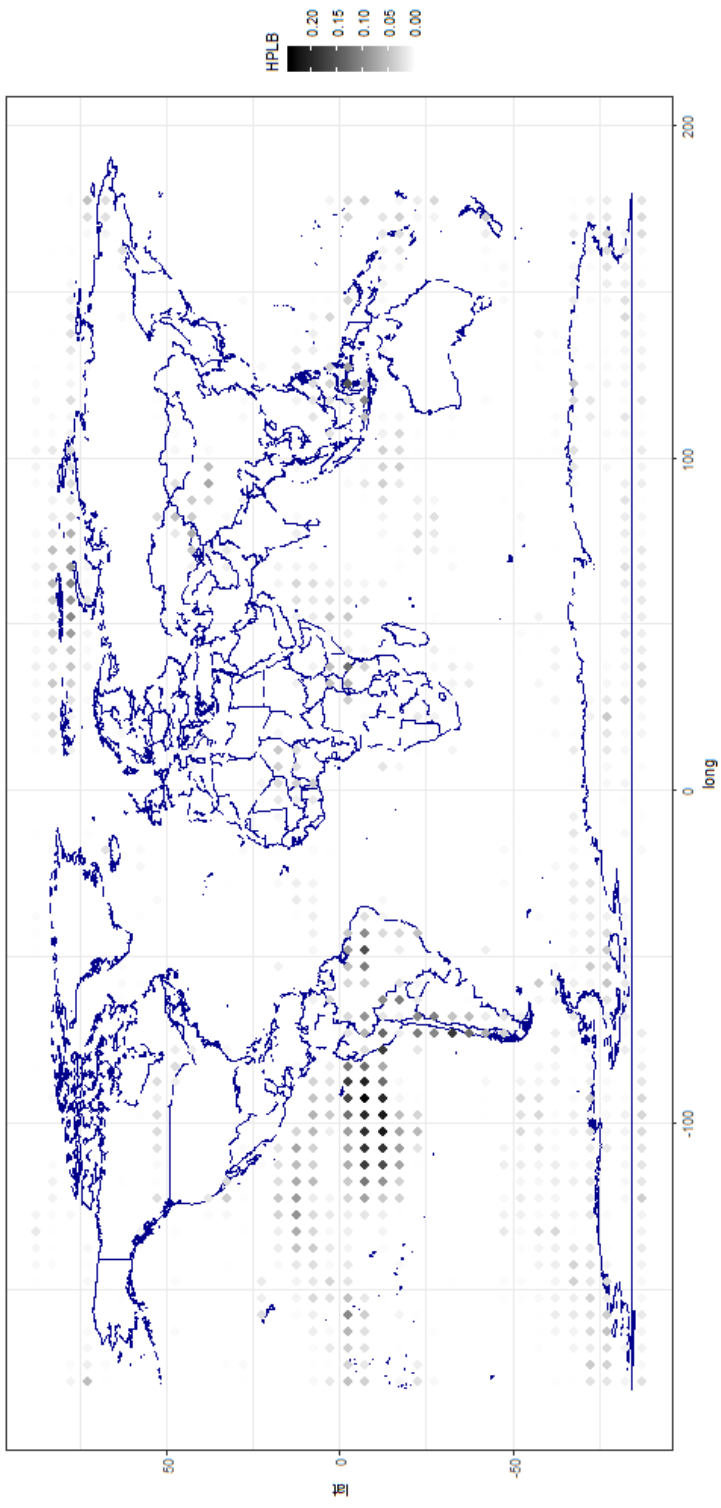


Figure 4: TV estimates by location using $\hat{\lambda}_{adapt}$ when comparing the first half of the available time span with the second half. A Random Forest is fitted using blocks of training and testing on the (differenced) values of the 4 variables. Color changes from white (zero) to black (one).

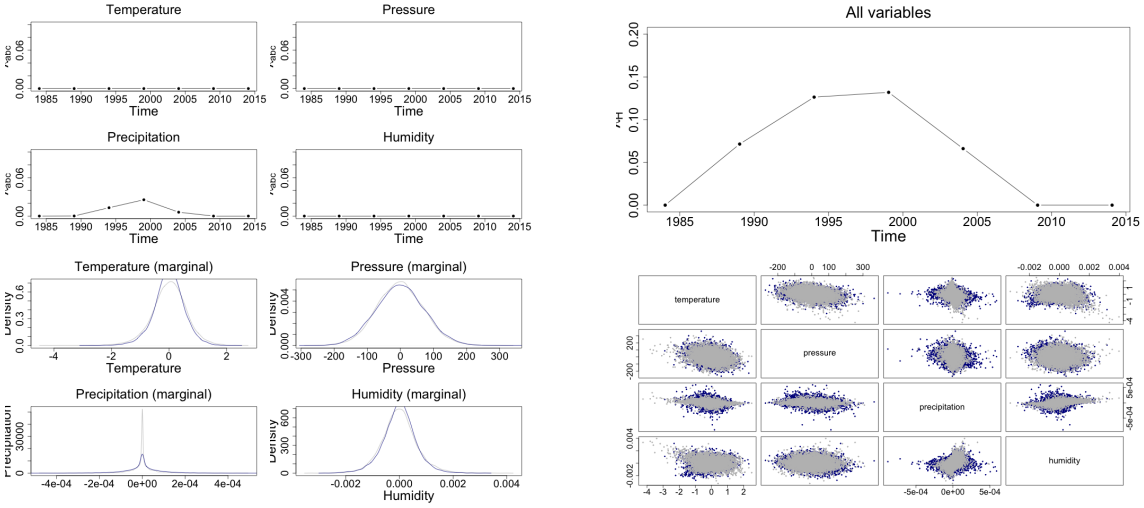


Figure 5: Top rows: High-probability lower bounds on total variation corresponding to 8 breaks for differenced temperature, pressure, precipitation and humidity. On the left the marginal analysis, on the right the joint analysis. Bottom row: Corresponding analysis of marginal density estimates and pair plots.

6 Discussion

We proposed in this paper two probabilistic lower bounds on the total variation distance between two distributions based on a one-dimensional projection. We theoretically characterized power rates given a sequence of (potentially random) projections ρ_N and showed that the adaptive estimator always reaches the best possible rate. Application to a climate reanalysis dataset showcased potential use of these estimators in practice.

1 Simulations

1.1 Illustration of Results in Examples 2 and 3

1.2 Change Detection

We illustrate the change detection described in Section 5 in some simple simulation settings. As in Section 4 we study independent random variables X_t, $t \in I$, with each $X_t \sim P_t$ and μ being the distribution of T on I. In all examples, we take μ to be the uniform distribution on $(0,1)$ and

1) simulate independently first t from μ and then X_t from P_t to obtain a training and test set, each of size $n = 10'000$,

2) train a Random Forest Regression predicting t from X_t on the training data, resulting in the projection ρ,

3) given ρ, evaluate $\hat{\lambda}_{adapt}$ on the test data for 19 s ranging from 0.05 to 0.95 in steps of 0.05.

The first simulation considers 3 settings with univariate random variables X_t:

(a) A mean-shift, with $X_t \sim \mathcal{N}(0,1)$ for $0 \leqslant t \leqslant 1/2$, $X_t \sim \mathcal{N}(1,1)$ for $1/3 < t \leqslant 2/3$ and $X_t \sim \mathcal{N}(2,1)$ for $2/3 < t \leqslant 1$.

(b) A variance shift, with $X_t \sim \mathcal{N}(0,1)$ for $0 \leqslant t \leqslant 1/2$, $X_t \sim \mathcal{N}(0,2)$ for $1/3 < t \leqslant 2/3$ and $X_t \sim \mathcal{N}(0,3)$ for $2/3 < t \leqslant 1$.

(c) A continuous mean-shift, with $X_t \sim \mathcal{N}(2t,1)$.

Results are given in Figure 8.

The second simulation illustrates a covariance change in a bivariate example: For $t \leqslant 0.5$, $X_t = (X_{t,1}, X_{t,2}) \sim \mathcal{N}(0, \Sigma_0)$, while for $t > 0.5$, $X_t = (X_{t,1}, X_{t,2}) \sim \mathcal{N}(0, \Sigma_1)$, with

$$\Sigma_0 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \text{ and } \Sigma_1 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}.$$

The upper and middle part of Figure 9 plots the marginal distributions $X_{t,i}$ against t. In all $T = 1000$ cases, λ^{ρ}_{adapt} (correctly) does not identify any changes in the two marginals. The change is however visible when considering the two variables jointly.

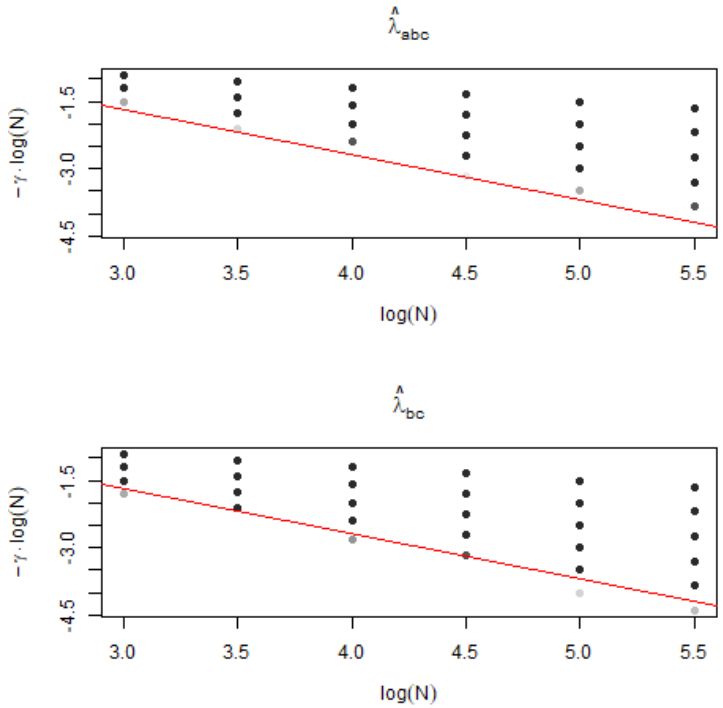


Figure 6: Illustration of Proposition 6 in Example 2. For a range of different γ, $-\gamma \cdot \log(N)$ is plotted against N. For each (γ, N) combination and for 100 repetitions, data was generated from the distribution in Example 2. The dots indicate the number of times the estimator was strictly larger than zero, with points ranging from white (constituting values smaller 0.05) to black (constituting a value of 1). The red line shows a slope of -1 for comparison.

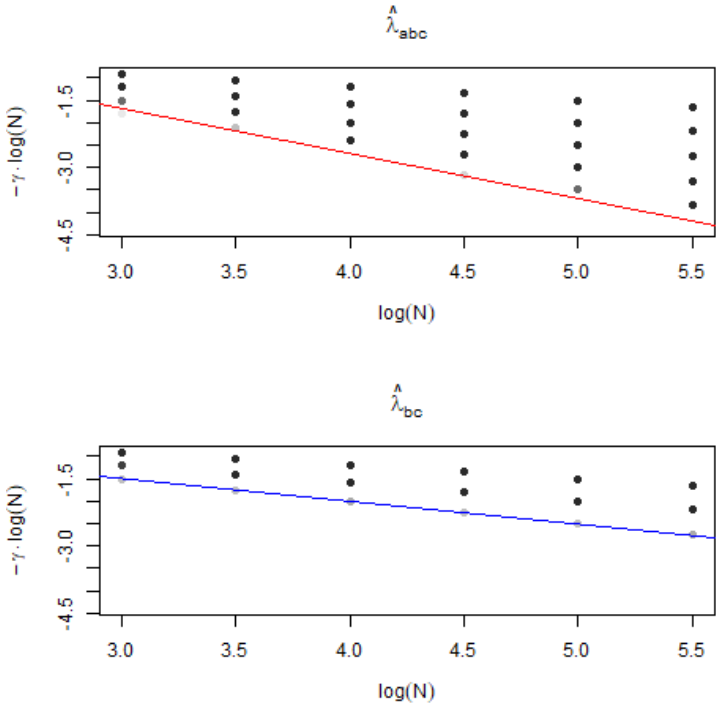


Figure 7: Illustration of Propositions 7 and 11 of Example 3. For a range of different γ, $-\gamma \cdot \log(N)$ is plotted against N. For each (γ, N) combination and for 100 repetitions, data was generated from the distribution in Example 3. The dots indicate the number of times the estimator was strictly larger than zero, with points ranging from white (constituting values smaller 0.05) to black (constituting a value of 1). The red and blue lines show slopes of -1 and -1/2 for comparison.

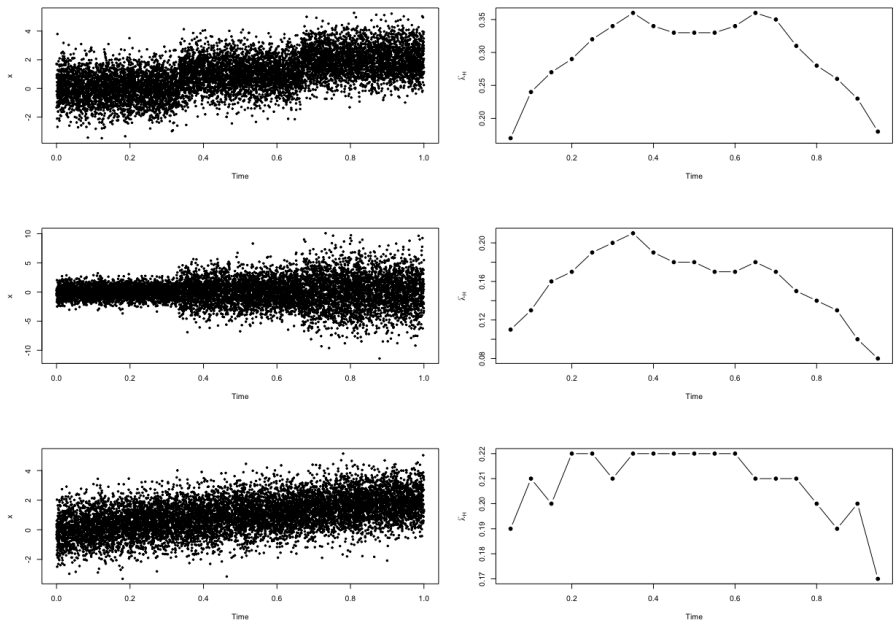


Figure 8: Top row: 3 regimes of mean-shifts. Middle row: 3 regimes of increasing variance, Bottom row: continuous mean-shift

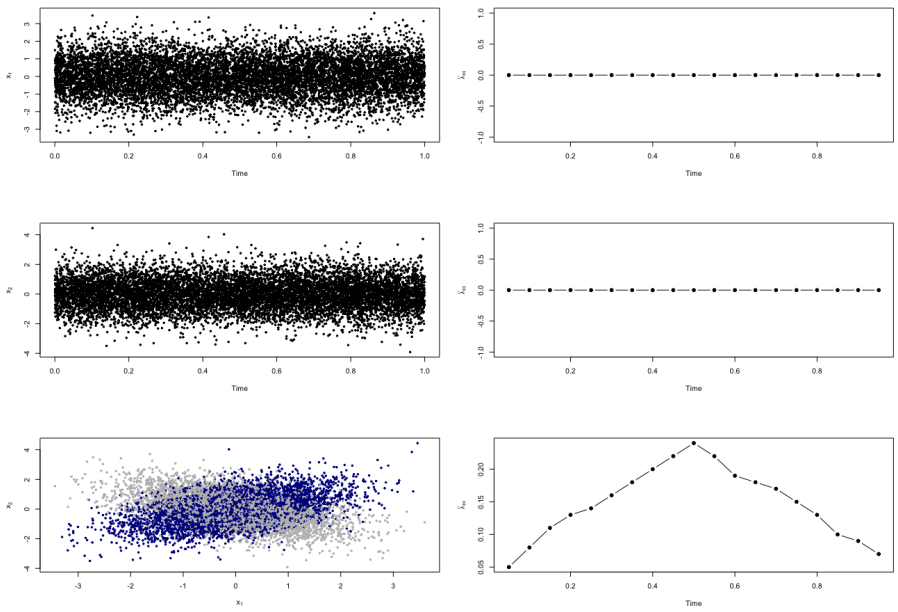


Figure 9: Top and middle row: marginal distributions; Bottom row: joint distribution.

2 Analytical bounding function

Here we give an analytical expression for $\mathrm{sq}_{\alpha/3}\left(\tilde{V}_{m(\tilde{\lambda}),z}, z = 1,\ldots,m(\tilde{\lambda})+n(\tilde{\lambda})-1\right)$ in (17) based on the theory in [52]:

Corollary 21. *The following is a valid simultaneous confidence band in* (17)*:*

$$sq_{\alpha/3}\left(\tilde{V}_{m(\tilde{\lambda}),z}, z = 1,\ldots,m(\tilde{\lambda})+n(\tilde{\lambda})-1\right) = (z - q_{1-\frac{\alpha}{3}}(\tilde{\lambda},m))\frac{m(\tilde{\lambda})}{m(\tilde{\lambda})+n(\tilde{\lambda})} + \beta_{\alpha/3,m(\tilde{\lambda})} w\left(z - q_{1-\frac{\alpha}{3}}(\tilde{\lambda},m), m(\tilde{\lambda}), m(\tilde{\lambda})\right), \tag{28}$$

with

$$\beta_{\alpha,m(\tilde{\lambda})} = \sqrt{2\log(\log(m(\tilde{\lambda})))} + \frac{\log(\log(\log(m(\tilde{\lambda})))) - \log(\pi) + 2x_{\alpha/3}}{2\sqrt{2\log(\log(m(\tilde{\lambda})))}},$$

and $w(z,m,n)$ *defined as in* (15).

Proof Applying Lemma .2 with $p = 1$ and $p_\varepsilon := \lambda$, $m(\lambda)$, $n(\lambda)$ go to infinity as $m,n \to \infty$. Moreover, since we assume $m/N \to \pi \leqslant 1/2$, it holds that

$$\lim_{N\to\infty} \frac{m(\lambda)}{n(\lambda)} = \frac{\pi}{1-\pi} \leqslant 1.$$

Thus for all but finitely many N, it holds that $m(\lambda) \leqslant n(\lambda)$. Combining this together with the fact that $\tilde{V}_{m,z-q_{1-\frac{\alpha}{3}}(\lambda,m)} - Q_{m,n,\alpha}(z,\lambda)$, $z \in \tilde{J}_{m,n,\lambda}$, is just a hypergeometric process adjusted by the

correct mean and variance, it follows from the arguments in [52]:

$$\limsup_{N\to\infty} \mathbb{P}\Big(\sup_{z\in\tilde{J}_{m,n,\lambda}} \Big[\tilde{V}_{m,z-(q_{1-\frac{\alpha}{3}}(\lambda,m))} - (z - q_{1-\frac{\alpha}{3}}(\lambda,m)) \frac{m(\lambda)}{m(\lambda)+n(\lambda)} \\ - \beta_{\alpha/3,m(\lambda)} w\left(z - q_{1-\frac{\alpha}{3}}(\lambda,m), m(\lambda), n(\lambda)\right) \Big] > 0 \Big) \leqslant \frac{\alpha}{3}.$$

Thus (33) indeed holds. ■

3 Proofs

Here we present the proofs of our main results. We start with a few preliminaries: In Section 2, we defined for two functions $h_1, h_2 : \mathbb{N} \to [0, +\infty)$, the notation $h_1(N) \asymp h_2(N)$ to mean that both (1) $\limsup_{N\to\infty} h_1(N)/h_2(N) \leqslant a_1$, for some $a_1 \in \mathbb{R}^+$ and (2) $\limsup_{N\to\infty} h_2(N)/h_1(N) \leqslant a_2$, for some $a_2 \in \mathbb{R}^+$. If instead only (1) is known, we write $h_1(N) = O(h_2(N))$ (translated as "asymptotically larger equal"). If (1) is known to hold for $a_1 = 0$, we write $h_1(N) = o(h_2(N))$ (translated as "asymptotically strictly smaller").

The technical lemmas of Section 3.3 should serve as a basis for the results in Section 3.1 to 3.2. They ensure that we may focus on the most convenient case, when $(t_N)_{N\geqslant 1}$ is such that $N\sigma(t_N) \to \infty$ (Lemma .4) or $m\sigma_F \to \infty$ (Lemma .6) holds. For these sequences, Lemma .3 shows that,

$$\frac{\hat{F}_m(t_N) - \hat{G}_n(t_N) - (F(t_N) - G(t_N))}{\sigma(t_N)} \xrightarrow{D} \mathcal{N}(0,1), \text{ for } N \to \infty. \tag{29}$$

We will now summarize the main proof ideas for the most important results. For Propositions 3 and 12, providing the level and power of $\hat{\lambda}(t_N)$ respectively, we use Lemma .3 and .4 to obtain (29). From this, Proposition 3 directly follows. It moreover implies that Proposition 12 holds iff

$$\frac{\lambda_N[(1-\varepsilon) - \lambda(t_N)/\lambda_N]}{\sigma(t_N)} \to -\infty \iff \text{(20) and (21)}.$$

This is simple, as both (20) and (21) were designed such that this equivalence holds.

We start in a similar manner to obtain the power result for $\hat{\lambda}_{adapt}$ in Proposition 14. We first restate the bounding function Q^A, for $z(t) \in \{q_{1-\alpha}(\tilde{\lambda}, m), \dots, m + n(\tilde{\lambda})\}$,

$$Q_{m,n,\alpha}(z(t),\tilde{\lambda}) = q_{1-\frac{\alpha}{3}}(\tilde{\lambda},m)\frac{n(\tilde{\lambda})}{N(\tilde{\lambda})} + z(t)\frac{m(\tilde{\lambda})}{N(\tilde{\lambda})} + \beta_{\frac{\alpha}{3},m(\tilde{\lambda})}\sqrt{\frac{m(\tilde{\lambda})}{N(\tilde{\lambda})}\frac{n(\tilde{\lambda})}{N(\tilde{\lambda})}\frac{N(\tilde{\lambda})-z(t)}{N(\tilde{\lambda})-1}(z(t) - q_{1-\frac{\alpha}{3}}(\tilde{\lambda},m))}, \tag{30}$$

with $N(\tilde{\lambda}) = m(\tilde{\lambda}) + n(\tilde{\lambda})$. Lemma .6 ensures that we may focus on the case $m\sigma_F \to \infty$. This immediately implies $(V_{m,t_N} - mA_0(t_N))/\sigma_F \xrightarrow{D} \mathcal{N}(0,1)$ due to the Lindeberg-Feller CLT (see

e.g., [177, Chapter 2]). Using Lemma .5 we show that what we would like to prove,

$$\mathbb{P}(V_{m,t_N} > Q_{m,n,\alpha}(z(t_N), \tilde{\lambda}) \ \ \forall \tilde{\lambda} \in [0, \lambda_\varepsilon]) \to 1 \iff \text{(20) and (21)},$$

can be replaced by the much simpler

$$\mathbb{P}(V_{m,t_N} > \tilde{Q}(\varepsilon)) \to 1 \iff \text{(20) and (21)},$$

where

$$\tilde{Q}(\tilde{\lambda}) = m\tilde{\lambda}(1-\pi) + m[\pi A_0(t_N) - (1-\pi)A_1(t_N) + (1-\pi)],$$

can be seen as the "limit" of an appropriately scaled $Q_{m,n,\alpha}(z(t_N), \tilde{\lambda})$. Using the structure of the problem and the asymptotic normality of $(V_{m,t_N} - mA_0(t_N))/\sigma_F$, we show that the result simplifies to showing that

$$\frac{\lambda_N[(1-\varepsilon) - \lambda(t_N)/\lambda_N]}{\sigma(t_N)} \to -\infty \iff \text{(20) and (21)}, \tag{31}$$

which was already done in Proposition 12.

On the other hand, to prove that $\hat{\lambda}_{adapt}$ is an asymptotic HPLB, we need to prove Propositions 9 and 10. The former is immediate with an infimum argument, whereas the latter requires some additional concepts. In particular, we use the bounding operation described in Lemma .7 to bound the original $V_{m,z}$ process pointwise for each z by the well behaved $\bar{V}_{m,z}$. The randomness of this process is essentially the one of the hypergeometric process $\tilde{V}_{m(\tilde{\lambda}),z}, z = 1, \ldots, m(\tilde{\lambda}) + n(\tilde{\lambda}) - 1$, as introduced in Section 2.3. The assumptions put on the bounding function Q, then ensure that we conserve the level.

The next three Section will provide the proofs of the main results, while Section 3.3 collects the aforementioned technical lemmas.

3.1 Proofs for Section 2

In this section, we prove the main results of Section 2, except for Propositions 4, 6, 7 and 11 connected to Examples 2 and 3. Their proofs will be given in Section 3.2.

Proposition 3. *Let $\lambda > 0$. For any sequence $(t_N)_{N\geqslant 1} \subset I$ of cutoffs, $\hat{\lambda}^\rho(t_N)$ defined in* (7) *is an asymptotic HPLB of λ (at level α) for any $\rho : \mathcal{X} \to I$.*

Proof From Lemma .4 (III) in Section 3.3, we may assume that for the given $(P, Q, (t_N)_{N\geqslant 1}, \rho)$, $N\sigma(t_N) \to +\infty$, as $N \to \infty$. In this case, we know from Lemma .3 that,

$$\frac{\hat{F}_m(t_N) - \hat{G}_n(t_N) - (F(t_N) - G(t_N))}{\sigma(t_N)} \xrightarrow{D} \mathcal{N}(0,1). \tag{32}$$

Consequently,

$$\limsup_{N\to\infty} \mathbb{P}\left(\hat{\lambda}^{\rho}(t_N) > F(t_N) - G(t_N)\right) = \lim_{N} \mathbb{P}\left(\frac{\hat{F}_m(t_N) - \hat{G}_n(t_N) - (F(t_N) - G(t_N))}{\sigma(t_N)} > q_{1-\alpha}\right)$$
$$= \alpha.$$

Since $\lambda \geqslant F(t_N) - G(t_N)$, the result then follows. ■

The exact same proof can also be used to show that Proposition 3 holds true, for $\lambda = N^{\gamma}$, $-1 < \gamma \leqslant 0$, as long as $\lambda > 0$ for all finite N.

Proposition 5 follows directly from Proposition 3 by exchanging $\sigma(t_N)$ with the consistent estimator used in $\hat{\lambda}^{\rho}_{bayes}$ and after checking that the case (**NC**) cannot happen, for a *fixed* ρ.

Proposition 9. *Let $Q_{m,n,\alpha}$ be an (asymptotic) bounding function and define,*

$$\hat{\lambda}^{\rho} = \inf\left\{\tilde{\lambda} \in [0,1] : \sup_{z\in J_{m,n}} \left[V_{m,z} - Q_{m,n,\alpha}(z,\tilde{\lambda})\right] \leqslant 0\right\}. \tag{16}$$

Then $\hat{\lambda}^{\rho}$ is an (asymptotic) HPLB of λ (at level α) for any $\rho : \mathcal{X} \to I$.

Proof

Let,

$$G_{m,n} := \left\{\tilde{\lambda} \in [0,1] : \sup_{z\in J_{m,n}} \left[V_{m,z} - Q_{m,n,\alpha}(z,\tilde{\lambda})\right] \leqslant 0\right\}.$$

Then by definition of the infimum,

$$\mathbb{P}(\hat{\lambda}^{\rho} > \lambda) \leqslant \mathbb{P}(\lambda \in G^{c}_{m,n})$$
$$= \mathbb{P}(\sup_{z\in J_{m,n}} [V_{m,z} - Q_{m,n,\alpha}(z,\lambda)] > 0).$$

The result follows by definition of $Q_{m,n,\alpha}$. ■

To prove Proposition 10, we need two technical concepts introduced in Section 3.3. In particular we utilize the concept of Distributional Witnesses in Definition 24 and the bounding operation in Lemma .7.

Proposition 10. *$Q_{m,n,\alpha}$ as defined in* (17) *is an (asymptotic) bounding function.*

Proof We aim to prove

$$\limsup_{N\to\infty} \mathbb{P}(\sup_{z\in J_{m,n}} [V_{m,z} - Q_{m,n,\alpha}(z,\lambda)] > 0) \leqslant \alpha. \tag{33}$$

Let Λ_P, Λ_Q be the distributional Witnesses of P and Q, as in Definition 24. Define the events $A_P := \{\Lambda_P \leqslant q_{1-\frac{\alpha}{3}}(\lambda, m)\}$, $A_Q := \{\Lambda_Q \leqslant q_{1-\frac{\alpha}{3}}(\lambda, n)\}$ and $A = A_P \cap A_Q$, such that $\mathbb{P}(A^c) \leqslant 2\alpha/3$. On A, we overestimate the number of witnesses on each side by construction. In this case we are able to use the bounding operation described above with $\bar{\Lambda}_P = q_{1-\frac{\alpha}{3}}(\lambda, m)$ and $\bar{\Lambda}_Q = q_{1-\frac{\alpha}{3}}(\lambda, n)$ to obtain $\bar{V}_{m,z}$ from Lemma .7. The process $\bar{V}_{m,z}$ has

$$\bar{V}_{m,z} = \begin{cases} z, & \text{if } 1 \leqslant z \leqslant q_{1-\frac{\alpha}{3}}(\lambda, m) \\ m, & \text{if } m + n(\lambda) \leqslant z \leqslant m + n \\ \tilde{V}_{m,z-q_{1-\frac{\alpha}{3}}(\lambda,m)} + q_{1-\frac{\alpha}{3}}(\lambda, m), & \text{if } q_{1-\frac{\alpha}{3}}(\lambda, m) < z < m + n(\lambda), \end{cases}$$

where $m(\lambda) = n - q_{1-\frac{\alpha}{3}}(\lambda, m)$, $n(\lambda) = n - q_{1-\frac{\alpha}{3}}(\lambda, n)$, and $\tilde{V}_{m,z-q_{1-\frac{\alpha}{3}}(\lambda,m)} \sim \text{Hypergeometric}(z - q_{1-\frac{\alpha}{3}}(\lambda, m), m(\lambda), m(\lambda) + n(\lambda))$. Then:

$$\mathbb{P}(\sup_{z \in J_{m,n}} [V_{m,z} - Q_{m,n,\alpha}(z, \lambda)] > 0) \leqslant \frac{2\alpha}{3} + \mathbb{P}(\sup_{z \in J_{m,n}} [\bar{V}_{m,z} - Q_{m,n,\alpha}(z, \lambda)] > 0 \cap A),$$

Now, $\bar{V}_{m,z} - Q_{m,n,\alpha}(z, \lambda) > 0$ can only happen for $z \in \tilde{J}_{m,n,\lambda} := \{q_{1-\frac{\alpha}{3}}(\lambda, m) + 1, \ldots, m + n(\lambda) - 1\}$, as by construction $\bar{V}_{m,z} - Q_{m,n,\alpha}(z, \lambda) = 0$, for $z \notin \tilde{J}_{m,n,\lambda}$. Thus

$$\begin{aligned} &\limsup_{N \to \infty} \mathbb{P}(\sup_{z \in J_{m,n}} [\bar{V}_{m,z} - Q_{m,n,\alpha}(z, \lambda)] > 0 \cap A) \leqslant \frac{2\alpha}{3} + \\ &\limsup_{N \to \infty} \mathbb{P}(\sup_{z \in \tilde{J}_{m,n,\lambda}} \left[\tilde{V}_{m,z-q_{1-\frac{\alpha}{3}}(\lambda,m)} - \text{sq}_{\alpha/3}\left(\tilde{V}_{m(\lambda),z-q_{1-\frac{\alpha}{3}}(\lambda,m)}, z \in \tilde{J}_{m,n,\lambda}\right)\right] > 0) \\ &\leqslant \alpha, \end{aligned}$$

by definition of $\text{sq}_{\alpha/3}\left(\tilde{V}_{m(\lambda),z}, z = 1, \ldots, m(\lambda) + n(\lambda) - 1\right)$.

■

3.2 Proofs for Section 3

Proposition 12. *Let $-1 < \gamma < 0$ and $\varepsilon \in (0, 1]$ fixed. Then there exists a $(t_N)_{N \geqslant 1}$ such that (C_ε) is true for $\hat{\lambda}^{\rho^*}(t_N)$ iff there exists a $(t_N)_{N \geqslant 1}$ such that* (20) *and* (21) *are true.*

Proof According to Lemma .4 we are allowed to focus on sequences $(t_N)_{N \geqslant 1}$ such that $N\sigma(t_N) \to \infty$. For $(t_N)_{N \geqslant 1} \subset I$ such a sequence, it holds that

$$\begin{aligned} \mathbb{P}(\hat{\lambda}^{\rho^*}(t_N) > (1 - \varepsilon)\lambda_N) &= \mathbb{P}(\hat{F}_m(t_N) - \hat{G}_n(t_N) - q_{1-\alpha}\sigma(t_N) > (1 - \varepsilon)\lambda_N) \\ &= \mathbb{P}\left(\frac{(\hat{F}_m(t_N) - \hat{G}_n(t_N) - \lambda(t_N))}{\sigma(t_N)} > q_{1-\alpha} - \frac{\lambda_N[(1 - \varepsilon) - \lambda(t_N)/\lambda_N]}{\sigma(t_N)}\right), \end{aligned}$$

where as in Section 3, $\lambda(t_N) = F(t_N) - G(t_N)$. With the same arguments as in Proposition 3, $(\hat{F}_m(t_N) - \hat{G}_n(t_N) - \lambda(t_N))/\sigma(t_N) \xrightarrow{D} \mathcal{N}(0,1)$. Thus, $\mathbb{P}\left(\hat{\lambda}^{\rho^*}(t_N) > (1-\varepsilon)\lambda_N\right) \to 1$, iff

$$\frac{\lambda_N[(1-\varepsilon) - \lambda(t_N)/\lambda_N]}{\sigma(t_N)} \to -\infty. \tag{34}$$

For $\gamma > -1/2$, (21), (20) and (34) are all true for $t_N = 1/2$, so there is nothing to prove in this case.

For $\gamma \leqslant -1/2$, assume (20) and (21) are true for $(t_N)_{N\geqslant 1}$. Then if $\liminf_{N\to\infty} \lambda(t_N)/\lambda_N > 1-\varepsilon$,

$$\frac{\lambda_N[(1-\varepsilon) - \lambda(t_N)/\lambda_N]}{\sigma(t_N)} \leqslant \frac{\lambda_N\left[(1-\varepsilon) - \inf_{M\geqslant N} \lambda(t_M)/\lambda_M\right]}{\sigma(t_N)} \to -\infty,$$

as $\left[(1-\varepsilon) - \inf_{M\geqslant N} \lambda(t_M)/\lambda_M\right] < 0$ for all but finitely many N and $\lambda_N/\sigma(t_N) \to +\infty$, by (21a). If instead $\liminf_{N\to\infty} \lambda(t_N)/\lambda_N = 1-\varepsilon$, the statement follows immediately from (21b). This shows one direction.

On the other hand, assume for all $(t_N)_{N\geqslant 1}$ (20) or (21) is false. We start by assuming the negation of (20), i.e., $\liminf_{N\to\infty} \lambda(t_N)/\lambda_N < 1-\varepsilon$. Then there exists for *all* N an $M \geqslant N$ such that $\lambda(t_M)/\lambda_M \leqslant 1-\varepsilon$, or

$$\frac{\lambda_M[(1-\varepsilon) - \lambda(t_M)/\lambda_M]}{\sigma(t_M)} \geqslant 0.$$

This is a direct contradiction of (34), which by definition means that for large enough N all elements of the above sequence are below zero. Now assume (21a) is wrong, i.e. $\liminf_{N\to\infty} \lambda(t_N)/\lambda_N > 1-\varepsilon$, but $\liminf_{N\to\infty} \lambda_N/\sigma(t_N) < \infty$. Since $\lambda(t_N)/\lambda_N \in [0,1]$ for all N, the lower bound will stay bounded away from $-\infty$ in this case. More specifically,

$$\liminf_{N\to\infty} \frac{\lambda_N}{\sigma(t_N)}[(1-\varepsilon) - \lambda(t_N)/\lambda_N] \geqslant \liminf_{N\to\infty} \frac{\lambda_N}{\sigma(t_N)} \liminf_N [(1-\varepsilon) - \lambda(t_N)/\lambda_N] > -\infty.$$

The negation of (21b) on the other hand, leads directly to a contradiction with (34). Consequently, by contraposition, the existence of a sequence $(t_N)_{N\geqslant 1}$ such that (21) and (20) are true is necessary.

■

Corollary 13. *$\hat{\lambda}^{\rho^*}_{bayes}$ attains the rate $\underline{\gamma}^{\hat{\lambda}_{bayes}}(\varepsilon) = \underline{\gamma}$ for all $\varepsilon \in (0,1]$, iff (21) is true for $t_N = 1/2$ and all $\gamma > \underline{\gamma}$.*

Proof Since $F(t) = A_0(t)$ and $1 - G(t) = A_1(t)$, it holds that $\sigma(1/2)/\hat{\sigma}(1/2) \to 1$ almost surely. Thus, the same arguments as in the proof of Proposition 12 with $t_N = 1/2$ give the result. ■

Proposition 14. *Let $-1 < \gamma < 0$ and $\varepsilon \in (0,1]$ fixed. Then (C_ε) is true for $\hat{\lambda}^{\rho^*}_{adapt}$ iff there exists a $(t_N)_{N\geqslant 1}$ such that* (20) *and* (21) *are true.*

Proof

Let for the following $\varepsilon \in (0,1]$ be arbitrary. The proof will be done by reducing to the case of $\hat{\lambda}(t_N)$. For a sequence $(t_N)_{N\geqslant 1}$ and a given sample of size N we then define the (random) $z(t_N)$, with

$$z(t_N) = \sum_{j=1}^{m} I\{\rho^*(X_j) \leqslant t_N\} + \sum_{i=1}^{n} I\{\rho^*(X_i) \leqslant t_N\} = m\hat{F}(t_N) + n\hat{G}(t_N). \tag{35}$$

Since by definition the observations $\rho^*_{(1)}, \dots \rho^*_{(z(t_N))}$ are smaller t_N, the classifier $\tilde{\rho}_{t_N}(z) := I\{\rho^*(z) > t_N\}$ will label all corresponding observations as zero. As such the number of actual observations coming from P in $\rho^*_{(1)}, \dots \rho^*_{(z(t_N))}$, V_{m,t_N}, will have $V_{m,t_N} \sim \text{Bin}(A_0(t_N), m)$. Recall that

$$A_0(t_N) = A_0^{\rho^*}(t_N) = P(\tilde{\rho}_{t_N}(X) = 0), \;\; A_1(t_N) = A_1^{\rho^*}(t_N) = Q(\tilde{\rho}_{t_N}(Y) = 1),$$

i.e. the true accuracies of the classifier $\tilde{\rho}_{t_N}$.

The goal is to show that we overshoot the quantile $Q_{m,n,\alpha}$:

$$\mathbb{P}(V_{m,t_N} > Q_{m,n,\alpha}(z(t_N), \tilde{\lambda}) \;\; \forall \tilde{\lambda} \in [0, \lambda_\varepsilon]) \to 1, \tag{36}$$

if and only if there exists a (t_N) such that (21) and (20) hold. For this purpose, Lemma .6 emulates Lemma .4 to allow us to focus on $(t_N)_{N\geqslant 1}$ such that $mA_0(t_N)(1 - A_0(t_N)) \to \infty$.

A sufficient condition for (36) is

$$\mathbb{P}\left(\frac{V_{m,t_N} - mA_0(t_N)}{\sqrt{mA_0(t_N)(1 - A_0(t_N))}} > \frac{\sup_{\tilde{\lambda}} Q_{m,n,\alpha}(z(t_N), \tilde{\lambda}) - mA_0(t_N)}{\sqrt{mA_0(t_N)(1 - A_0(t_N))}} \right) \to 1, \tag{37}$$

while a necessary condition is given by

$$\mathbb{P}\left(\frac{V_{m,t_N} - mA_0(t_N)}{\sqrt{mA_0(t_N)(1 - A_0(t_N))}} > \frac{Q_{m,n,\alpha}(z(t_N), \lambda_\varepsilon) - mA_0(t_N)}{\sqrt{mA_0(t_N)(1 - A_0(t_N))}} \right) \to 1. \tag{38}$$

We instead work with a simpler bound:

$$\tilde{Q}(\tilde{\lambda}) = m\tilde{\lambda}(1 - \pi) + m[\pi A_0(t_N) - (1 - \pi)A_1(t_N) + (1 - \pi)]. \tag{39}$$

Note that

$$\sup_{\tilde{\lambda} \in [0, \lambda_\varepsilon]} \tilde{Q}(\tilde{\lambda}) = \tilde{Q}(\lambda_\varepsilon). \tag{40}$$

and

$$\begin{aligned} \tilde{Q}(\lambda_\varepsilon) - mA_0(t_N) &= m(1 - \pi)[\lambda_\varepsilon - (A_0(t_N)(t_N) + A_1(t_N) - 1)] \\ &= m(1 - \pi)[\lambda_\varepsilon - \lambda(t_N)]. \end{aligned}$$

We first show that

$$\mathbb{P}\left(\frac{V_{m,t_N} - mA_0(t_N)}{\sqrt{mA_0(t_N)(1-A_0(t_N))}} > \frac{\tilde{Q}(\lambda_\varepsilon) - mA_0(t_N)}{\sqrt{mA_0(t_N)(1-A_0(t_N))}}\right) \to 1, \tag{41}$$

if and only if there exists a (t_N) such that (21) and (20) hold.

Since again $\frac{V_{m,t_N}-mA_0(t_N)}{\sqrt{mA_0(t_N)(1-A_0(t_N))}} \xrightarrow{D} N(0,1)$, due to the Lindeberg-Feller CLT ([177]), (41) holds iff

$$\frac{\tilde{Q}(\lambda_\varepsilon) - mA_0(t_N)}{\sqrt{mA_0(t_N)(1-A_0(t_N))}} \to -\infty. \tag{42}$$

To prove this claim, we write

$$\frac{\tilde{Q}(\lambda_\varepsilon) - mA_0(t_N)}{\sqrt{mA_0(t_N)(1-A_0(t_N))}} = (1-\pi)\frac{\lambda_N[(1-\varepsilon) - \lambda(t_N)/\lambda_N]}{\sqrt{\frac{A_0(t_N)(1-A_0(t_N))}{m}}} \tag{43}$$

and show that

$$\frac{\lambda_N}{\sqrt{\frac{A_0(t_N)(1-A_0(t_N))}{m}}} \to +\infty \iff \frac{\lambda_N}{\sigma(t_N)} \to +\infty. \tag{44}$$

In this case, (42) is equivalent to (34) and it follows from exactly the same arguments as in the proof of Proposition 12 that (42) is true iff there exists a (t_N) such that (21) and (20) hold.

To prove (44), first assume that

$$\frac{\lambda_N}{\sqrt{\frac{A_0(t_N)(1-A_0(t_N))}{m}}} \to +\infty.$$

This implies that $A_0(t_N)(1-A_0(t_N)) = o(N^{2\gamma+1})$, which means that either $A_0(t_N) = o(N^{2\gamma+1})$ or $(1-A_0(t_N)) = o(N^{2\gamma+1})$. Assume $A_0(t_N) = o(N^{2\gamma+1})$. Since by definition $A_0(t_N)+A_1(t_N)-1 = \lambda(t_N) = O(\lambda_N)$, this means that $1 - A_1(t_N) = O(N^\gamma) + o(N^{2\gamma+1}) = o(N^{2\gamma+1})$ and thus also $A_1(t_N)(1-A_1(t_N)) = o(N^{2\gamma+1})$. The same applies for $1 - A_0(t_N) = o(N^{2\gamma+1})$. Writing $\sigma(t)$ as in (22) this immediately implies $\frac{\lambda_N}{\sigma(t_N)} \to +\infty$. On the other hand, assume $\frac{\lambda_N}{\sigma(t_N)} \to +\infty$. This in turn means

$$A_0(t_N)(1-A_0(t_N)) + A_1(t_N)(1-A_1(t_N)) = o(N^{2\gamma+1}) \tag{45}$$

and thus $A_0(t_N)(1-A_0(t_N)) = o(N^{2\gamma+1})$ and $\lambda_N/\sqrt{\frac{A_0(t_N)(1-A_0(t_N))}{m}} \to +\infty$. This proves (44). Using the arguments of the proof of Proposition 12 this demonstrates that (42) is true iff there exists a (t_N) such that (20) and (21) hold.

It remains to show that (41) implies (37) and is implied by (38). More specifically, as (21) demands that

$$A_0(t_N)(1-A_0(t_N)) = o(N^{2\gamma+1}) \text{ and } A_1(t_N)(1-A_1(t_N)) = o(N^{2\gamma+1}),$$

we may use Lemma .5 below to see that for $c \in (0, +\infty)$,

$$c + o_{\mathbb{P}}(1) \leqslant \frac{Q_{m,n,\alpha}(z(t_N), \lambda_\varepsilon) - mA_0}{\tilde{Q}(\lambda_\varepsilon) - mA_0} \leqslant \frac{\sup_{\tilde{\lambda}} Q_{m,n,\alpha}(z(t_N), \tilde{\lambda}) - mA_0}{\tilde{Q}(\lambda_\varepsilon) - mA_0} \leqslant \frac{1}{c} + o_{\mathbb{P}}(1).$$

For $Z_N \xrightarrow{D} N(0,1)$, $Q_{1,N} \to -\infty$ and $c + O_N \leqslant Q_{2,N}/Q_{1,N}$, with $c > 0$ and $O_N \xrightarrow{p} 0$, it holds that

$$\begin{aligned}
\mathbb{P}(Z_N > Q_{2,N}) &= \mathbb{P}\left(Z_N > \frac{Q_{2,N}}{Q_{1,N}} Q_{1,N}\right) \\
&\geqslant \mathbb{P}\left(Z_N > (c + O_N) Q_{1,N}\right) \\
&= \mathbb{P}\left(Z_N > (c + O_N) Q_{1,N} \cap |O_N| \leqslant \frac{c}{2}\right) + \mathbb{P}\left(Z_N > (c + O_N) Q_{1,N} \cap |O_N| > \frac{c}{2}\right) \\
&\to 1,
\end{aligned}$$

as $Q_{1,N} < 0$ for all but finitely many N and $(c + O_N) > 0$ on the set $|O_N| \leqslant \frac{c}{2}$. Using this argument first with $Q_{1,N} = \tilde{Q}(\lambda_\varepsilon) - mA_0(t_N)$ and $Q_{2,N} = Q_{m,n,\alpha}(z(t_N), \lambda_\varepsilon) - mA_0(t_N)$, and repeating it with $Q_{1,N} = Q_{m,n,\alpha}(z(t_N), \lambda_\varepsilon) - mA_0(t_N)$ and $Q_{2,N} = \tilde{Q}(\lambda_\varepsilon) - mA_0(t_N)$, (58) shows that (41) implies (37) and is implied by (38).

■

We are now able to prove the results in Examples 2 and 3:

Proposition 4. *For the setting of Example 1, assume $p_N > 0.5$ for all N, and $p_N - 1/2 \asymp N^\gamma$. Then $\underline{\gamma}^{oracle}(\varepsilon) = -1/2$ for all $\varepsilon \in (0, 1]$. This rate is attained by the oracle estimator in* (9) *with $t_N^* = 1/2$ for all N.*

Proof

First note that $\rho(z) = (1 - p) I\{-1 \leqslant z \leqslant 0\} + p I\{0 \leqslant z \leqslant 1\}$ and thus

$$A_0(t_N) = pI\{1-p \leqslant t_N \leqslant p\} + I\{t_N > p\}, \;\; A_1(t_N) = pI\{1-p \leqslant t_N \leqslant p\} + I\{t_N < 1-p\} \quad (46)$$

Take any $\gamma \leqslant -1/2$. Then for $\lambda_N/\sigma(t_N) \to \infty$ to be true it is necessary that $A_0(t_N)(1 - A_0(t_N))$ and $A_1(t_N)(1 - A_1(t_N))$ go to zero. But from (46) and the fact that $p \to 0.5$, it is clear that this is only possible for $t_N \in [1 - p, p]^c$ for all but finitely many N. However for such t_N, $\lambda(t_N) = A_0(t_N) + A_1(t_N) - 1 = 0$. Similarly, a sequence $(t_N)_{N \geqslant 1}$ that satisfies (21), cannot satisfy condition (20). Thus for $\gamma \leqslant -1/2$ for any sequence $(t_N)_{N \geqslant 1}$ most one of the two conditions (20) and (21) can be true and thus $-1/2 \leqslant \underline{\gamma}^{oracle}(\varepsilon)$. On the other hand, for $\gamma > -1/2$, taking $t_N = 1/2$ independently of γ, satisfies conditions (20) and (21). ■

Proposition 6. *For the setting of Example 2, $\underline{\gamma}^{\hat{\lambda}_{bayes}}(\varepsilon) = \underline{\gamma}^{oracle}(\varepsilon) = -1$, for all $\varepsilon \in (0, 1]$.*

Proof

We show that $\underline{\gamma}^{\hat{\lambda}_{bayes}} = -1$, from which it immediately follows that $\underline{\gamma}^{oracle} = -1$. Since $A_0^{\rho^*}(1/2) = \lambda_N$ and $A_1^{\rho^*}(1/2) = 1$, it follows for any $\gamma > -1$,

$$\frac{\lambda_N}{\sigma(1/2)} = \frac{\sqrt{m}\lambda_N}{\sqrt{\lambda_N}} \to \infty, \text{ for } N \to \infty.$$

By Proposition (13) this implies $\underline{\gamma}^{\hat{\lambda}_{bayes}} = -1$.

■

Proposition 22. *For the setting of Example 3, let $\varepsilon \in (0,1]$ be arbitrary and $p_2 > 0.5$, $p_2 = 0.5 + o(N^{-1})$. Then $\hat{\lambda}^{\rho^*}_{adapt}$ attains the oracle rate $\underline{\gamma}^{\hat{\lambda}_{adapt}} = \underline{\gamma}^{oracle} = -1$, while $\hat{\lambda}^{\rho^*}_{bayes}$ attains the rate $\underline{\gamma}^{\hat{\lambda}_{bayes}} = -1/2$.*

Proof We first find the expression for λ_N. Since $p_2 > 0.5$

$$\begin{aligned}\lambda_N &= \int (f-g)dx = p_1 + [(1-p_1)p_2 - (1-p_1)(1-p_2)] \int f_0 dx \\ &= p_1 + (1-p_1)[2p_2 - 1]\end{aligned}$$

and, since $p_2 - 1/2 = o(N^{-1})$, it immediately holds that $p_1 \asymp \lambda_N$. Let $\gamma > -1$ be arbitrary and take $t_N = 0$ for all N. Then $\lambda(t_N) = p_1$ and it holds that

$$\frac{\lambda_N}{\lambda(t_N)} = \frac{p_1 + (1-p_1)[2p_2-1]}{p_1} = 1 + \frac{(1-p_1)[2p_2-1]}{p_1} \to 1,$$

as $2p_2 - 1 = o(N^{-1})$ by assumption. Combining this with the fact that $A_0^{\rho^*}(0) = p_1 \asymp \lambda_N$, and $A_1^{\rho^*}(0) = 1$ thus

$$\frac{\lambda_N}{\sigma(t)} = \frac{\sqrt{m}\lambda_N}{\sqrt{A_0^{\rho^*}(0)(1 - A_0^{\rho^*}(0))}} \asymp \sqrt{m\lambda_N} \to \infty,$$

it follows that $\underline{\gamma}^{\text{oracle}} = -1$ and therefore also $\underline{\gamma}^{\hat{\lambda}_{adapt}} = -1$. On the other hand

$$A_0^{\rho^*}(1/2) = A_1^{\rho^*}(1/2) = p_1 + (1-p_1)p_2 \to 0.5,$$

so (21) cannot be true for any $\gamma \leqslant -1/2$. From Corollary 13 it follows that $\hat{\lambda}_{bayes}$ only attains a rate $\underline{\gamma}^{\hat{\lambda}_{bayes}} = -1/2$.

■

Proposition 7 and 11 then immediately follow from Proposition 22.

We continue with the proofs for Section 3.2, by quickly restating assumptions (E1) and (E2):

(E1) $\hat{\rho} = \hat{\rho}_{N_{tr}}$ is trained on a sample of size N_{tr}, $(Z_1, \ell_1), \ldots, (Z_{N_{tr}}, \ell_{N_{tr}})$, and evaluated on an independent sample $(Z_1, \ell_1), \ldots, (Z_{N_{te}}, \ell_{N_{te}})$, with $N_{tr} + N_{te} = N$

(E2) $N_{te}, N_{tr} \to \infty$, as $N \to \infty$, with $m_{te}/N_{te} \to \pi \in (0,1)$.

Let $\lambda(\hat{\rho})$ be defined as in (19):

$$\lambda(\hat{\rho}) = \sup_{t\in[0,1]} \left[A_0^{\hat{\rho}}(t) + A_1^{\hat{\rho}}(t)\right] - 1 := \sup_{t\in[0,1]} \left[P(\hat{\rho}(X) \leqslant t|\hat{\rho}) - Q(\hat{\rho}(Y) \leqslant t|\hat{\rho})\right].$$

We first establish that $\hat{\lambda}^{\hat{\rho}}_{adapt}$ is still an asymptotic HPLB.

Proposition 17. *Assume (E1) and (E2). Then $\hat{\lambda}^{\hat{\rho}}_{adapt}$ is an (asymptotic) HPLB of λ (at level α).*

Proof We first note that Propositions 9 and 10 hold true also for a sequence $(\rho_N)_{N\in\mathbb{N}}$, instead of just a single arbitrary ρ. Thus conditioning on $\hat{\rho}$ trained on independent data, we have in particular that pointwise,

$$\limsup_{N\to\infty} \mathbb{P}(\hat{\lambda}^{\hat{\rho}} > \lambda|\hat{\rho}) \leqslant \alpha.$$

Using Fatou's Lemma [46], it holds that

$$\begin{aligned}\limsup_{N\to\infty} \mathbb{P}(\hat{\lambda}^{\hat{\rho}} > \lambda) &= \limsup_{N\to\infty} \mathbb{E}[\mathbb{P}(\hat{\lambda}^{\hat{\rho}} > \lambda|\hat{\rho})] \\ &\leqslant \mathbb{E}[\limsup_{N\to\infty} \mathbb{P}(\hat{\lambda}^{\hat{\rho}} > \lambda|\hat{\rho})] \\ &\leqslant \alpha,\end{aligned}$$

proving the result. ■

However, for $\hat{\lambda}^{\hat{\rho}}_{bayes}$ (or $\hat{\lambda}^{\rho_N}(1/2)$), we encounter a difficulty when $\lambda = 0$.

Proposition 16. *For the setting of Example 4, let ξ_1, ξ_2 be independently Poisson distributed, with mean C. Then*

$$\mathbb{P}(\hat{\lambda}^{\rho_N}(1/2) > 0) \to \mathbb{P}(\xi_1 - \xi_2 > q_{1-\alpha}\sqrt{2C}).$$

Proof It holds that $A_0^{\rho_N}(1/2) = C/n$ and $1 - A_1^{\rho_N}(1/2) = C/n$ and,

$$\mathbb{P}(\hat{\lambda}^{\rho_N}(1/2) > 0) = \mathbb{P}(n\hat{F}_n^{\rho_N}(1/2) - n\hat{G}_n^{\rho_N}(1/2) > q_{1-\alpha} n\sigma(1/2)).$$

Define $\xi_{01} = n\hat{F}_n^{\rho_N}(1/2) \sim \text{Binomial}(C/n, n)$ and $\xi_{02} = n\hat{G}_n^{\rho_N}(1/2) \sim \text{Binomial}(C/n, n)$. Then by the Poisson convergence theorem and due to independence, $\xi_{01} - \xi_{02}$ converges in distribution to $\xi_1 - \xi_2$. Additionally,

$$n\sigma(1/2) = \sqrt{nA_0^{\rho_N}(1/2)(1 - A_0^{\rho_N}(1/2)) + nA_1^{\rho_N}(1/2)(1 - A_1^{\rho_N}(1/2))} \to \sqrt{2C},$$

proving the result. ■

For $\varepsilon \in (0,1]$ the goal in the following is to establish that for all subsequences, there exists a further subsequence $N(\ell(k)))$ such that

$$\liminf_{k\to\infty} \mathbb{P}(\hat{\lambda} > (1-\varepsilon)\lambda_{N(\ell(k))} | \hat{\rho}_{N_{tr}(\ell(k))}) = 1, \text{ a.s.} \tag{47}$$

This suggests that for a given ε we need to check the following adapted conditions on $(t_{N_{te}})_{N_{te}\geqslant 1}$: For any subsequence $N(\ell)$, we find a further subsequence $N(\ell(k))$, such that

$$\liminf_{k\to\infty} \mathcal{T}_{N(\ell(k))} := \liminf_{k\to\infty} \frac{\lambda^{\hat{\rho}_{N_{tr}(\ell(k))}}\left(t_{N_{te}(\ell(k))}\right)}{\lambda_{N(\ell(k))}} \geqslant 1-\varepsilon \text{ a.s.}, \tag{48}$$

and

$$\lim_{k\to\infty} \frac{\lambda_{N(\ell(k))}}{\sigma^{\hat{\rho}_{N_{tr}(\ell(k))}}\left(t_{N_{te}(\ell(k))}\right)} = \infty \text{ a.s., if } \liminf_{k\to\infty} \mathcal{T}_{N(\ell(k))} > 1-\varepsilon \text{ a.s.}, \tag{49a}$$

$$\lim_{k\to\infty} \frac{\lambda_{N(\ell(k))}}{\sigma^{\hat{\rho}_{N_{tr}(\ell(k))}}\left(t_{N_{te}(\ell(k))}\right)} \left(\mathcal{T}_{N(\ell(k))} - (1-\varepsilon)\right) = \infty \text{ a.s., if } \liminf_{k\to\infty} \mathcal{T}_{N(\ell(k))} = 1-\varepsilon \text{ a.s.}, \tag{49b}$$

where for $F^{\hat{\rho}}(t) := \mathbb{P}(\hat{\rho}(X) \leqslant t|\hat{\rho})$ and $G^{\hat{\rho}}(t) := \mathbb{P}(\hat{\rho}(Y) \leqslant t|\hat{\rho})$

$$\begin{aligned}
\lambda^{\hat{\rho}}(t) &= F^{\hat{\rho}}(t) - G^{\hat{\rho}}(t) \\
\lambda(\hat{\rho}) &= \sup_{t\in[0,1]} \lambda^{\hat{\rho}}(t) \\
\sigma^{\hat{\rho}}(t) &= \left(\frac{F^{\hat{\rho}}(t)(1-F^{\hat{\rho}}(t))}{m_{te}} + \frac{G^{\hat{\rho}}(t)(1-G^{\hat{\rho}}(t))}{n_{te}}\right)^{1/2}.
\end{aligned}$$

We now generalize Propositions 12 and 14 to this case:

Proposition 23. *Let $-1 < \gamma \leqslant 0$ and $\varepsilon_1 \in (0,1]$ fixed. Assume that $\lambda_N = N_{te}^{\gamma}$ and that (E1) and (E2) hold. Then the following is equivalent*

(i) there exists a $(t_{N_{te}})_{N_{te}\geqslant 1}$ such that (48) *and* (49) *are true for ε,*

(ii) (C_ε) is true for $\hat{\lambda}^{\hat{\rho}}(t_N)$,

(iii) (C_ε) is true for $\hat{\lambda}^{\hat{\rho}}_{adapt}$.

Proof The same arguments as in Proposition 12 and 14 show that for a (nonrandom) sequence ρ_N and $(t_N)_{N\geqslant 1}$:

$$\liminf_{N\to\infty} \frac{\lambda^{\rho_N}(t_N)}{\lambda_N} \geqslant 1-\varepsilon, \tag{50}$$

and

$$\lim_N \frac{\lambda_N}{\sigma^{\rho_N}(t_N)} = \infty, \text{ if } \liminf_{N\to\infty} \frac{\lambda^{\rho_N}(t_N)}{\lambda_N} > 1-\varepsilon, \tag{51a}$$

$$\lim_N \frac{\lambda_N}{\sigma^{\rho_N}(t_N)} \left(\frac{\lambda^{\rho_N}(t_N)}{\lambda} - (1-\varepsilon) \right) = \infty, \text{ if } \liminf_{N\to\infty} \frac{\lambda^{\rho_N}(t_N)}{\lambda_N} = 1-\varepsilon, \tag{51b}$$

if and only if

$$\liminf_{k\to\infty} \mathbb{P}(\hat{\lambda}^{\rho_N}(t_N) > (1-\varepsilon)\lambda) = 1,$$

and

$$\liminf_{k\to\infty} \mathbb{P}(\hat{\lambda}^{\rho_N}_{adapt} > (1-\varepsilon)\lambda) = 1.$$

Through conditioning, we now extend this to $\hat{\rho}$. The arguments are the same for $\hat{\lambda}^{\rho_N}_{adapt}$ and $\hat{\lambda}^{\rho_N}(t_N)$ and thus we will write $\hat{\lambda}$ to mean either of them.

First assume (48) and (49) are true for an $\varepsilon \in (0,1]$, $\hat{\rho}$ and sequence $(t_N)_N$. Considering only the chosen subsequence $N(\ell(k))$ and conditioning on $(\hat{\rho}_{N(\ell(k))})_k$, this gives a sequence $\rho_k = \hat{\rho}_{N(\ell(k))}$ such that (50) and (51) are true and by the above this means (47) holds. Since $\mathbb{P}(\hat{\lambda} > (1-\varepsilon)\lambda_N | \hat{\rho}_{N_{tr}})$ is bounded, we can use Fatous lemma to obtain, that every subsequence has a further subsequence with

$$\liminf_{k\to\infty} \mathbb{P}(\hat{\lambda} > (1-\varepsilon)\lambda_{N(\ell(k))}) = 1.$$

An argument by contradiction shows that then the liminf of the overall sequence must be 1 as well. Indeed assume that this is not true. Then we can find a subsequence $N(\ell)$ such that

$$\lim_{\ell\to\infty} \mathbb{P}(\hat{\lambda} > (1-\varepsilon)\lambda_{N(\ell)}) = c < 1.$$

But then any further subsequence will have limsup strictly below 1, contradicting the above.

Now assume (C_ε) is true. Then, by definition, $\mathbb{P}\left(\hat{\lambda} > (1-\varepsilon)\lambda_N\right) \to 1$. But this is also true for any subsequence and thus (47) must also hold. Indeed this simply follows from the fact that

$$\lim_{k\to\infty} \int |f_k| d\mathbb{P} = 0 \implies \lim_{k\to\infty} |f_k| = 0 \text{ a.s.}, \tag{52}$$

applied to $f_k = 1 - \mathbb{P}(\hat{\lambda} > (1-\varepsilon)\lambda_{N(\ell(k))} | \hat{\rho}_{N_{tr}(\ell(k))}) \geqslant 0$. We quickly prove (52) for completeness below. But with that, by the same arguments as above (connecting to a nonrandom sequence ρ_k), (47) implies (48) and (49).

It remains to prove (52). To do so, assume there exists a set B, with $\mathbb{P}(B) > 0$, such that $\liminf_{k\to\infty} |f_k| > 0$ on B. Then again using Fatou's lemma,

$$\begin{aligned}
\liminf_{k\to\infty} \int |f_k| d\mathbb{P} &\geqslant \int \liminf_{k\to\infty} |f_k| d\mathbb{P} \\
&\geqslant \int_B \liminf_{k\to\infty} |f_k| d\mathbb{P} + \int_{B^c} \liminf_{k\to\infty} |f_k| d\mathbb{P} \\
&> 0,
\end{aligned}$$

since $g > 0$ implies that $\int g d\mathbb{P} > 0$. Thus $\liminf_{k\to\infty} \int |f_k| d\mathbb{P} > 0$, proving (52) by contraposition.

■

Again, Proposition 23 would be still valid, if λ was replaced everywhere by $\lambda(\hat{\rho})$, assuming that $\lambda(\hat{\rho})$ converges to a limit $\lambda(\rho) \in [0,1]$ in probability. For instance, if $\lambda(\hat{\rho}) \xrightarrow{p} 0$, at a rate N^γ, $-1 < \gamma < 0$.

Proposition 18 then follows directly from Proposition 23.

Corollary 20. *Assume that λ is fixed and that there exists a sequence $(t_{N_{tr}})$, such that the sequence of classifiers $\hat{\rho}_{N_{tr},t_{N_{tr}}}$ is consistent. Then (C_ε) is true for $\hat{\lambda}^{\hat{\rho}}_{adapt}$, for all $\varepsilon > 0$.*

Proof Due to consistency, it holds for all $\varepsilon \in (0,1]$ that there exists for each subsequence a further subsequence, such that

$$\liminf_{k\to\infty} \frac{\lambda^{\hat{\rho}_{N_{tr}(\ell(k))}}\left(t_{N_{te}(\ell(k))}\right)}{\lambda} \geqslant 1-\varepsilon \text{ a.s.},$$

for $t_{N_{te}(\ell(k))} := t_{N_{tr}(\ell(k))}$. Thus for the sequence $t_{N_{tr}}$ and all $\varepsilon \in (0,1]$, (48) is true. Moreover, since λ is fixed here, (49) is clearly also true, proving the result. ■

3.3 Technical Results

Lemma .2. *Let $p \in [0,1]$, $\alpha \in (0,1)$ with $1-\alpha > 0.5$ and $p_\varepsilon := (1-\varepsilon)p$. Then $mp - q_{1-\alpha}(p_\varepsilon, m) \asymp mp\varepsilon$. More generally, if $p = p_m \asymp m^\gamma$, $-1 < \gamma < 0$, and $p_\varepsilon := (1-\varepsilon)p_m$, then $mp_m - q_{1-\alpha}(p_\varepsilon, m) \asymp mp_m\varepsilon$.*

Proof Let $p = \delta_m \asymp m^\gamma$, for $-1 < \gamma \leqslant 0$, where $\gamma = 0$ indicates the fixed p case. Writing $q_{1-\alpha}(p_\varepsilon, m) = q_{1-\alpha}(\Lambda)$, where $\Lambda \sim \text{Binomial}(p_\varepsilon, m)$, it holds that

$$\frac{q_{1-\alpha}(\Lambda) - mp_\varepsilon}{\sqrt{mp_\varepsilon(1-p_\varepsilon)}} = q_{1-\alpha}(Z_m),$$

where $Z_m := (\Lambda - mp_\varepsilon)/\sqrt{mp_\varepsilon(1-p_\varepsilon)}$ and $q_{1-\alpha}(Z_m)$ is the $1-\alpha$ quantile of the distribution of Z_m. By the Lindenberg-Feller central limit theorem, Z_m converges in distribution to $\mathcal{N}(0,1)$ and is thus uniformly tight, i.e. $Z_m = O_{\mathbb{P}}(1)$. Consequently, it must hold that

$$0 < \frac{q_{1-\alpha}(\Lambda) - mp_\varepsilon}{\sqrt{mp_\varepsilon(1-p_\varepsilon)}} = q_{1-\alpha}(Z_m) \asymp 1,$$

which means $q_{1-\alpha}(\Lambda) - mp_\varepsilon \asymp \sqrt{mp_\varepsilon(1-p_\varepsilon)}$. Writing

$$\Delta_m := mp_m - q_{1-\alpha}(p_\varepsilon, m) = mp_m - mp_\varepsilon - (q_{1-\alpha}(p_\varepsilon, m) - mp_\varepsilon) = mp_m\varepsilon - (q_{1-\alpha}(\Lambda) - mp_\varepsilon),$$

we see that $\Delta_m \asymp m p_m \varepsilon$. ■

As we do not constrain the possible alternatives P, Q and sequences $(t_N)_{N\geqslant 1}$, some proofs have several cases to consider. In an effort to increase readability we will summarize these different cases here for reference: We first introduce a "nuisance condition". This condition arises when $(t_N)_{N\leqslant 1}$ or the sequence of alternatives is such that the variance $\sigma(t_N)$ converges to zero fast, namely if

$$\liminf_{N\to\infty} N\sigma(t_N) < +\infty. \tag{NC}$$

The case in which we are mainly interested is however is the negation of (**NC**),

$$\lim_{N\to\infty} N\sigma(t_N) = +\infty \tag{MC}$$

A special case of that is the following

$$\text{either } F(t_N)(1-F(t_N)) = 0 \text{ or } G(t_N)(1-G(t_N)) = 0 \text{ for infinitely many } N. \tag{MCE}$$

We first show an important limiting result, in the case (**MC**), on which much of our results are based:

Lemma .3. *Let $-1 < \gamma \leqslant 0$, where $\gamma = 0$ constitutes the constant case $\lambda_N = \lambda$. Then for any $\rho : \mathcal{X} \to I$ and any sequence $(t_N)_{N\geqslant 1} \subset I$ such that* (**MC**) *holds,*

$$\mathcal{Z}_N := \frac{\hat{F}_m(t_N) - \hat{G}_n(t_N) - (F(t_N) - G(t_N))}{\sigma(t_N)} \xrightarrow{D} \mathcal{N}(0,1). \tag{53}$$

Proof Let

$$\sigma_F := \sqrt{\frac{F(t_N)(1-F(t_N))}{m}} \quad \text{and} \quad \sigma_G := \sqrt{\frac{G(t_N)(1-G(t_N))}{m}},$$

so that we may write $\sigma(t_N) = \sqrt{\sigma_F^2 + \sigma_G^2}$. From (**MC**) we require $m\sigma_F \to \infty$ or $n\sigma_G \to \infty$.

By the Lindenberg-Feller CLT (see e.g., [177]), it holds for $N \to \infty$ (and thus $m, n \to \infty$),

$$\frac{1}{\sigma_F}(\hat{F}_m(t_N) - F(t_N)) \xrightarrow{D} \mathcal{N}(0,1), \text{ if } m\sigma_F \to \infty$$
$$\frac{1}{\sigma_G}(\hat{G}_m(t_N) - G(t_N)) \xrightarrow{D} \mathcal{N}(0,1), \text{ if } n\sigma_G \to \infty$$

We write

$$\begin{aligned}
\mathcal{Z}_N &= \frac{\hat{F}_m(t_N) - F(t_N) - (\hat{G}_n(t_N) - G(t_N))}{\sigma(t_N)} \\
&= \frac{\hat{F}_m(t_N) - F(t_N)}{\sigma(t_N)} - \frac{(\hat{G}_n(t_N) - G(t_N))}{\sigma(t_N)} \\
&= \frac{\hat{F}_m(t_N) - F(t_N)}{\sigma_F}\frac{\sigma_F}{\sigma(t_N)} - \frac{(\hat{G}_n(t_N) - G(t_N))}{\sigma_G}\frac{\sigma_G}{\sigma(t_N)}.
\end{aligned}$$

Now as,

$$\frac{\sigma_F}{\sigma(t_N)} = \sqrt{\frac{\sigma_F^2}{\sigma_F^2+\sigma_G^2}}, \text{ and } \frac{\sigma_G}{\sigma(t_N)} = \sqrt{\frac{\sigma_G^2}{\sigma_F^2+\sigma_G^2}},$$

we can define $\omega_N := \sigma_F/\sigma(t_N)$, so that

$$\mathcal{Z}_N = \frac{\hat{F}_m(t_N) - F(t_N)}{\sigma_F}\omega_N - \frac{(\hat{G}_n(t_N) - G(t_N))}{\sigma_G}\sqrt{1-\omega_N^2}.$$

Had ω_N a limit, say $\lim_N \omega_N := a \in [0,1]$ and if both $m\sigma_F \to \infty$ and $m\sigma_G \to \infty$ were true, it would immediately follow from classical results (see e.g., [177, Chapter 2]) that $\mathcal{Z}_N \xrightarrow{D} \mathcal{N}(0,1)$. This is not the case as the limit of ω_N might not exist and either $\limsup_{N\to\infty} m\sigma_F < \infty$ or $\limsup_{N\to\infty} n\sigma_G < \infty$. However since $\omega_N \in [0,1]$ for all N, it possesses a subsequence with a limit in $[0,1]$. More generally, every subsequence $(\omega_{N(k)})_k$ possesses a further subsequence $(\omega_{N(k(\ell))})_\ell$ that converges to a limit $a \in [0,1]$. This limit depends on the specific subsequence, but for any such converging subsequence it still holds as above that $\mathcal{Z}_{N(k(\ell))} \xrightarrow{D} \mathcal{N}(0,1)$. Indeed, if both $m\sigma_F \to \infty$ and $n\sigma_G \to \infty$ this is immediate from the above. If, on the other hand, $\liminf_{N\to\infty} m\sigma_F < \infty$, then $[\hat{F}_m(t_{N(k(\ell))}) - F(t_{N(k(\ell))})]/\sigma_F \xrightarrow{D} \mathcal{N}(0,1)$ might not be true. However, if we assume that (**MCE**) does not hold, for the chosen subsequence

$$\frac{\sigma_F^2}{\sigma(t_{N(k(\ell))})^2} = \omega_{N(k(\ell))}^2 \to a^2 \in [0,1],$$

and it either holds that $a = 0$ in which case the first part of $\mathcal{Z}_{N(k(\ell))}$ is negligible or $a > 0$, in which case it must hold that $\sigma_F \asymp \sigma(t_{N(k(\ell))})$ and thus $N(k(\ell))\sigma_F \to \infty$, allowing for $[\hat{F}_m(t_{N(k(\ell))}) - F(t_{N(k(\ell))})]/\sigma_F \xrightarrow{D} \mathcal{N}(0,1)$. The symmetric argument applies if $\liminf_{N\to\infty} m\sigma_G < \infty$. Now assume (**MCE**) holds and for a given subsequence $(t_k)_k$, it is not possible to find a subsequence, such that $F(t_{k(\ell)})(1-F(t_{k(\ell)})) > 0$ for all but finitely many ℓ. Since for all subsequences $k(\ell)\sigma(t_{k(\ell)}) \to \infty$, it must hold that $n(k(\ell))\sigma_G \to \infty$. In particular, we may choose the subsequence such that $F(t_{k(\ell)})(1-F(t_{k(\ell)})) = 0$ for all but finitely many N and in this case:

$$\hat{F}_m(t_{k(\ell)}) - F(t_{k(\ell)}) = 0 \text{ a.s. and } \frac{\hat{G}(t_{k(\ell)}) - G(t_{k(\ell)})}{\sigma_G} \xrightarrow{D} \mathcal{N}(0,1)$$

both are true, implying (53). The symmetric argument holds if instead $G(t_N)(1-G(t_N)) = 0$ for infinitely many N, but $NF(t_N)(1-F(t_N)) \to \infty$.

Thus we have shown that for *any* subsequence of $\mathcal{Z}_N$, there exists a further subsequence converging in distribution to $\mathcal{N}(0,1)$. Assume that despite this, (53) is not true. Then, negating convergence in distribution in this particular instance, means there exists $z \in \mathbb{R}$ such that the cumulative distribution function of $\mathcal{Z}_N$, $F_{\mathcal{Z}_N}$, has $\limsup_{N\to\infty} F_{\mathcal{Z}_N}(z) \neq \Phi(z)$. By the properties of the limsup, there exists a subsequence $\lim_{k\to\infty} F_{\mathcal{Z}_{N(k)}}(z) = \limsup_{N\to\infty} F_{\mathcal{Z}_N}(z) \neq \Phi(z)$. But then no further subsequence of $F_{\mathcal{Z}_{N(k)}}(z)$ converges to $\Phi(z)$, a contradiction. ■

The next lemma ensures that we can for all intents and purposes ignore sequences $(t_N)_{N\geqslant 1}$ for which (**NC**) is true.

Lemma .4. *Let $-1 < \gamma < 0$ and $\varepsilon \in (0,1]$ arbitrary. If for a sequence $(t_N)_{N\geqslant 1}$ and $\rho = \rho^*$,* (**NC**) *holds, then*

(I) (20) *or* (21) *is not true,*

(II) (C_ε) *is not true for* $\lambda = \hat{\lambda}^{\rho^*}(t_N)$.

Furthermore, if for $-1 < \gamma < 0$, a sequence $(t_N)_{N\geqslant 1}$ and $\rho : \mathcal{X} \to I$, (**NC**) *holds then,*

(III) $\limsup_{N\to\infty} \mathbb{P}\left(\hat{\lambda}^{\rho}(t_N) > \lambda\right) = 0$.

(III) is also true for the constant case, $\gamma = 0$, as long as $\lambda > 0$.

Proof (I): If (**NC**) is true, then $N^\beta \sigma(t_N) \to 0$, for any $\beta \in [0,1)$. Indeed, assume there exists $\beta \in [0,1)$ such that $\liminf_{N\to\infty} N^\beta \sigma(t_N) > 0$. Then

$$\liminf_{N\to\infty} N\sigma(t_N) \geqslant +\infty,$$

In particular, it must hold that

$$F(t_N)(1-F(t_N)) = o(N^\zeta) \text{ and } G(t_N)(1-G(t_N)) = o(N^\zeta),$$

for all $\zeta \in [-1,0)$. There are four possibilities for this to be true:

(1) $F(t_N) = o(N^\zeta), G(t_N) = o(N^\zeta)$.

(2) $F(t_N) = o(N^\zeta), 1 - G(t_N) = o(N^\zeta)$.

(3) $(1 - F(t_N)) = o(N^\zeta), (1 - G(t_N)) = o(N^\zeta)$.

(4) $(1 - F(t_N)) = o(N^\zeta), G(t_N) = o(N^\zeta)$.

As (2) and (4) imply that $\lambda(t_N) \to -1$ and $\lambda_N \geqslant \lambda(t_N) \to 1$ respectively, they are not relevant in our framework. Thus (**NC**) directly implies that either (1) or (3) is true and both of them imply $\lambda(t_N) = o(N^\zeta)$ for all $\zeta \in (-1,0)$. Consequently, (20) cannot be true for $\varepsilon < 1$.

For $\varepsilon = 1$ we slightly strengthen the relevant cases (1) and (3):

(1') $\liminf_{N\to\infty} NA_0(t_N) < \infty, \liminf_{N\to\infty} N(1 - A_1(t_N)) < \infty$,

(3') $\liminf_{N\to\infty} N(1 - A_0(t_N)) < \infty, \liminf_{N\to\infty} NA_1(t_N) < \infty$.

It's clear from the above, that if (**NC**) holds, then one of the two has to hold. We will now show that, even though (20) is true in this case, (21b) is not. Indeed, it was mentioned in Section 3 that in case of (**MCE**), (21b) is defined to be false. Thus, we may assume $\sigma(t_N)$ is bounded away from zero for all finite N. Assume that (21b) is true, i.e.

$$\frac{\lambda(t_N)}{\sigma(t_N)} \to \infty.$$

This must then also hold for any subsequence. Now assume (1') is true, and choose a subsequence $(t_k)_{k\in\mathbb{N}} := (t_{N(k)})_{k\in\mathbb{N}}$ with $\lim_{k\to\infty} kA_0(t_k) = \liminf_{N\to\infty} NA_0(t_N) := a \in [0,\infty)$. Then

$$\frac{\lambda(t_k)}{\sigma(t_k)} \leqslant \frac{A_0(t_k)}{\sigma(t_k)} \leqslant \frac{\sqrt{m(k)}A_0(t_k)}{\sqrt{A_0(t_k)(1-A_0(t_k))}} \asymp \sqrt{kA_0(t_k)} \to \sqrt{a} < \infty,$$

a contradiction. Similarly, if (2') is true, we find a subsequence such that $\lim_{k\to\infty} kA_1(t_k) := a \in [0,\infty)$ and bound

$$\frac{\lambda(t_k)}{\sigma(t_k)} \leqslant \frac{A_1(t_k)}{\sigma(t_k)} \leqslant \frac{\sqrt{n(k)}A_1(t_k)}{\sqrt{A_1(t_k)(1-A_1(t_k))}} \asymp \sqrt{kA_1(t_k)} \to \sqrt{a} < \infty.$$

Thus (21b) cannot be true.

(II) and (III): Consider first $\varepsilon \in [0,1)$ and $-1 < \gamma \leqslant 0$. (**NC**) implies for any ρ:

$$\lambda_N^{-1}(\hat{F}(t_N) - \hat{G}(t_N) - \lambda(t_N)) \xrightarrow{p} 0. \tag{54}$$

Indeed by a simple Markov inequality argument for all $\delta > 0$:

$$\mathbb{P}\left(\lambda_N^{-1}(\hat{F}(t_N) - \hat{G}(t_N) - \lambda(t_N)) > \delta\right) \leqslant \frac{\lambda_N^{-2}\sigma(t_N)^2}{\delta} \asymp \frac{(N^{-\gamma}\sigma(t_N))^2}{\delta} \to 0,$$

since $-\gamma \in [0,1)$. Additionally, from the argument in (I), $\lambda_N^{-1}\sigma(t_N) \to 0$ and $\frac{\lambda(t_N)}{\lambda_N} \to 0$. Consequently, for any $\varepsilon \in [0,1)$

$$\begin{aligned}
&\mathbb{P}(\hat{F}(t_N) - \hat{G}(t_N) - q_{1-\alpha}\sigma(t_N) > (1-\varepsilon)\lambda_N) \\
&= \mathbb{P}(\hat{F}(t_N) - \hat{G}(t_N) - \lambda(t_N) - q_{1-\alpha}\sigma(t_N) > (1-\varepsilon)\lambda_N - \lambda(t_N)) \\
&= \mathbb{P}(\lambda_N^{-1}(\hat{F}(t_N) - \hat{G}(t_N) - \lambda(t_N)) - q_{1-\alpha}\lambda_N^{-1}\sigma(t_N) + \frac{\lambda(t_N)}{\lambda_N} > (1-\varepsilon)) \\
&\to 0.
\end{aligned}$$

Consequently, (C_ε) is false for any $\varepsilon \in [0,1)$ and (II) and (III) hold. The case $\varepsilon = 1$ needs special care: Assume that despite (**NC**), $\mathbb{P}(\hat{\lambda}(t_N) > 0) \to 1$ holds true. We consider the two possible cases (1') and (3') in turn: If (1') is true, we write:

$$\mathbb{P}(\hat{\lambda}(t_N) > 0) \leqslant \mathbb{P}(\hat{F}(t_N) > 0) = \mathbb{P}(m\hat{F}(t_N) > 0) = \mathbb{P}(V_{m,t_N} > 0), \tag{55}$$

where $V_{m,t_N} := m\hat{F}(t_N) \sim \text{Binomial}(A_0(t_N), m)$. Since $\mathbb{P}(\hat{\lambda}(t_N) > 0) \to 1$, this is true for any subsequence $\hat{\lambda}(t_{N(k)})$ as well. In particular, we may choose the subsequence $(t_{N(k)})_{k\geqslant 1}$ with $\lim_k N(k)A_0(t_{N(k)}) = \liminf_{N\to\infty} NA_0(t_N) := a \in [0,\infty)$. Renaming the subsequence $(kA_0(t_k))_{k\geqslant 1}$ for simplicity, we find $\limsup_{k\to\infty} kA_0(t_k) \leqslant a$, or $A_0(t_k) = O(k^{-1}) = O(m(k)^{-1})$, since by assumption $m(k)/k \to \pi \in (0,1)$. But then

$$\mathbb{P}(\hat{\lambda}(t_k) > 0) \leqslant \mathbb{P}(V_{m(k),t_k} > 0) = 1 - (1 - A_0(t_k))^{m(k)}$$

and

$$\liminf_{k\to\infty}(1 - A_0(t_k))^{m(k)} \geqslant \liminf_{k\to\infty}(1 - \frac{a}{m(k)})^{m(k)} = \exp(-a) > 0.$$

Thus, $\limsup_{k\to\infty} \mathbb{P}(\hat{\lambda}(t_k) > 0) < 1$, a contradiction.

If (3') is true, then $\liminf_{N\to\infty} NA_1(t_N) < +\infty$ and similar arguments applied to

$$\mathbb{P}(\hat{\lambda}(t_N) > 0) \leqslant \mathbb{P}(1 - \hat{G}(t_N) > 0) = \mathbb{P}(n - n\hat{G}(t_N) > 0) = \mathbb{P}(V_{n,t_N} > 0), \tag{56}$$

where now $V_{n,t_N} := \sum_{i=1}^n I\{\rho(Y_i) > t_N\} \sim \text{Binomial}(A_1(t_N), n)$, give

$$\limsup_{k\to\infty} \mathbb{P}(\hat{\lambda}(t_k) > 0) \leqslant \limsup_{k\to\infty} \mathbb{P}(V_{n(k),t_k} > 0) = 1 - \exp(-a) < 1.$$

This again contradicts $\mathbb{P}(\hat{\lambda}(t_N) > 0) \to 1$.

■

Lemma .4 also immediately implies for $\hat{\lambda}(t_N)$ that if (C_ε) or (20) are true, then $\lim_N N\sigma(t_N) = +\infty$ must hold.

Lemma .5. *Let $-1 < \gamma < 0$ be fixed and as above $\lambda_N \asymp N^\gamma$. Define for $(t_N)_{N\geqslant 1}$, $z(t_N)$ as in* (35) *and $Q_{m,n,\alpha}$, $\tilde{Q}$ as in* (30) *and* (39)*. Assume that for the given γ,*

$$A_0(t_N)(1 - A_0(t_N)) = o(N^{2\gamma+1}) \text{ and } A_1(t_N)(1 - A_1(t_N)) = o(N^{2\gamma+1}), \tag{57}$$

then for $c \in (0, +\infty)$,

$$c + o_{\mathbb{P}}(1) \leqslant \frac{Q_{m,n,\alpha}(z(t_N), \lambda_\varepsilon) - mA_0}{\tilde{Q}(\lambda_\varepsilon) - mA_0} \leqslant \frac{\sup_{\tilde{\lambda}} Q_{m,n,\alpha}(z(t_N), \tilde{\lambda}) - mA_0}{\tilde{Q}(\lambda_\varepsilon) - mA_0} \leqslant \frac{1}{c} + o_{\mathbb{P}}(1). \tag{58}$$

Proof Note that $z(t_N)$ is random, while everything else is deterministic. First,

$$q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon, m)\frac{n(\lambda_\varepsilon)}{N(\lambda_\varepsilon)} + \frac{z(t_N)}{N(\lambda_\varepsilon)}m(\lambda_\varepsilon) + \beta_{\frac{\alpha}{3},m(\lambda_\varepsilon)}\sqrt{\frac{m(\lambda_\varepsilon)}{N(\lambda_\varepsilon)}\frac{n(\lambda_\varepsilon)}{N(\lambda_\varepsilon)}\frac{N(\lambda_\varepsilon) - z(t_N)}{N(\lambda_\varepsilon) - 1}(z(t_N) - q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon, m))}$$

$$\leqslant \sup_{\tilde{\lambda}\in[0,\lambda_\varepsilon]} Q_{m,n,\alpha}(z(t_N), \tilde{\lambda}) \leqslant$$

$$q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon, m)\frac{n}{N(\lambda_\varepsilon)} + \frac{z(t_N)}{N(\lambda_\varepsilon)}m + \beta_{\frac{\alpha}{3},m}\sqrt{\frac{m}{N(\lambda_\varepsilon)}\frac{n}{N(\lambda_\varepsilon)}\frac{N - z(t_N)}{N(\lambda_\varepsilon) - 1}z(t_N)}.$$

Additionally for all $\tilde{\lambda} \in [0, \lambda_\varepsilon]$, with $p_N = [\pi A_0(t_N) - (1-\pi)A_1(t_N) + (1-\pi)]$,

$$\frac{m(\tilde{\lambda})}{N(\tilde{\lambda})} \to \pi, \tag{59}$$

$$\frac{n(\tilde{\lambda})}{N(\tilde{\lambda})} \to 1 - \pi, \tag{60}$$

$$\frac{q_{1-\frac{\alpha}{3}}(\tilde{\lambda}, m)}{m\tilde{\lambda}} \to 1 \tag{61}$$

$$\frac{\frac{z(t_N)}{N(\lambda_\varepsilon)} - A_0(t_N)}{p_N - A_0(t_N)} \xrightarrow{p} 1. \tag{62}$$

The first three assertions follow from Lemma .2 and the assumption that $m/N \to \pi$, as $N \to \infty$. We quickly verify (62). Define

$$S_N = \frac{\frac{z(t_N)}{N(\lambda_\varepsilon)} - A_0(t_N)}{p_N - A_0(t_N)}.$$

By Chebyshev's inequality,

$$\mathbb{P}\left(|S_N - \mathbb{E}[S_N]| > \delta\right) \leqslant \frac{\mathrm{Var}(S_N)}{\delta}, \tag{63}$$

for all $\delta > 0$. Now, $z(t_N)$ may be written as a sum of independent Bernoulli random variables:

$$z(t_N) = \sum_{i=1}^{N} I\{\rho^*(Z_i) \leqslant t_N\} = \sum_{i=1}^{m} I\{\rho^*(X_i) \leqslant t_N\} + \sum_{j=1}^{n} I\{\rho^*(Y_i) \leqslant t_N\},$$

with $I\{\rho^*(X_i) \leqslant t_N\} \sim \text{Bernoulli}(A_0(t_N))$ and $I\{\rho^*(Y_i) \leqslant t_N\} \sim \text{Bernoulli}(1 - A_1(t_N))$. Then

$$\begin{aligned}
\mathrm{Var}(S_N) &= \frac{1}{(p_N - A_0(t_N))^2} \mathrm{Var}\left(\frac{z(t_N)}{N(\lambda_\varepsilon)}\right) \\
&= \frac{1}{(p_N - A_0(t_N))^2 N(\lambda_\varepsilon)^2} \left[mA_0(t_N)(1 - A_0(t_N)) + nA_1(t_N)(1 - A_1(t_N))\right] \\
&\asymp \frac{1}{(p_N - A_0(t_N))^2 N} \left[A_0(t_N)(1 - A_0(t_N)) + A_1(t_N)(1 - A_1(t_N))\right].
\end{aligned}$$

Now, since (i) $p_N - A_0(t_N) = -(1-\pi)[A_0(t_N) + A_1(t_N) - 1] = -(1-\pi)\lambda(t_N)$ and $\lambda(t_N) \leqslant \lambda_N \asymp N^\gamma$ and (ii) (57) holds, it follows that

$$\mathrm{Var}(S_N) \asymp \frac{1}{N^{2\gamma+1}} o(N^{2\gamma+1}) \to 0, \text{ for } N \to \infty.$$

Thus $|S_N - \mathbb{E}[S_N]| \xrightarrow{p} 0$. Moreover, it holds that,

$$\begin{aligned}
\mathbb{E}[S_N] - 1 &= \frac{m/N(\lambda_\varepsilon)A_0(t_N) - n/N(\lambda_\varepsilon)A_1(t_N) + n/N(\lambda_\varepsilon) - A_0(t_N)}{\pi A_0(t_N) - (1-\pi)A_1(t_N) + (1-\pi) - A_0(t_N)} - 1 \\
&= \frac{[(1-\pi) - (1 - m/N(\lambda_\varepsilon))]A_0(t_N) - [(1-\pi) - n/N(\lambda_\varepsilon)](1 - A_1(t_N))}{(\pi-1)[A_0(t_N) - (1 - A_1(t_N))]} \\
&= \frac{o(1)A_0(t_N) - o(1)(1 - A_1(t_N))}{(\pi-1)[A_0(t_N) - (1 - A_1(t_N))]} \\
&= \frac{o(1)[A_0(t_N) - (1 - A_1(t_N))]}{(\pi-1)[A_0(t_N) - (1 - A_1(t_N))]} \to 0, \text{ for } N \to \infty.
\end{aligned}$$

Thus, finally $|S_N - 1| \leqslant |S_N - \mathbb{E}[S_N]| + |\mathbb{E}[S_N] - 1| \xrightarrow{p} 0$.

Continuing, let for the following for two random variables index by N $X_N \preceq Y_N$ mean that $\mathbb{P}(X_N \leqslant Y_N) \to 1$, as $N \to \infty$. Then,

$$\begin{aligned}
&N^{-\gamma} m^{-1} \left(\sup_{\tilde{\lambda}} Q_{m,n,\alpha}(z(t_N), \tilde{\lambda}) - mA_0 \right) \leqslant \\
&N^{-\gamma} q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon, m)/m \frac{n}{N(\lambda_\varepsilon)} + N^{-\gamma} \frac{z(t_N)}{N(\lambda_\varepsilon)} + N^{-\gamma} \frac{\beta_{\frac{\alpha}{3},m}}{m} \sqrt{\frac{m}{N(\lambda_\varepsilon)} \frac{n}{N(\lambda_\varepsilon)} \frac{N - z(t_N)}{N(\lambda_\varepsilon) - 1} \frac{z(t_N)}{m}} - N^{-\gamma} A_0(t_N) \\
&= N^{-\gamma} \left[\frac{q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon, m)/m}{\lambda_\varepsilon} \frac{n}{N(\lambda_\varepsilon)} \lambda_\varepsilon + \frac{z(t_N)}{N(\lambda_\varepsilon)} - A_0(t_N) \right] + N^{-\gamma} \frac{\beta_{\frac{\alpha}{3},m}}{m} \sqrt{\frac{m}{N(\lambda_\varepsilon)} \frac{n}{N(\lambda_\varepsilon)} \frac{N - z(t_N)}{N(\lambda_\varepsilon) - 1} \frac{z(t_N)}{m}} \\
&= N^{-\gamma} \left[\frac{q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon, m)/m}{\lambda_\varepsilon} \frac{n}{N(\lambda_\varepsilon)} \lambda_\varepsilon + \frac{z(t_N)/N(\lambda_\varepsilon) - A_0(t_N)}{p_N - A_0(t_N)} [p_N - A_0(t_N)] \right] + \\
&N^{-\gamma} \frac{\beta_{\frac{\alpha}{3},m}}{m} \sqrt{\frac{m}{N(\lambda_\varepsilon)} \frac{n}{N(\lambda_\varepsilon)} \frac{N - z(t_N)}{N(\lambda_\varepsilon) - 1} \frac{z(t_N)}{m}} \\
&\leq \left[\frac{q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon, m)/m}{\lambda_\varepsilon} \frac{n}{N(\lambda_\varepsilon)} \frac{(1-\varepsilon)\lambda_N}{\lambda_N} - (1-\pi) \frac{z(t_N)/N(\lambda_\varepsilon) - A_0(t_N)}{p_N - A_0(t_N)} \inf_{M \geqslant N} \frac{\lambda(t_M)}{\lambda_M} \right] \sup_{M \geqslant N} (M^{-\gamma} \lambda_M) \\
&+ N^{-\gamma} \frac{\beta_{\frac{\alpha}{3},m}}{m} \sqrt{\frac{m}{N(\lambda_\varepsilon)} \frac{n}{N(\lambda_\varepsilon)} \frac{N - z(t_N)}{N(\lambda_\varepsilon) - 1} \frac{z(t_N)}{m}} \\
&\xrightarrow{p} d_1(1-\pi) \left[(1-\varepsilon) - d_2 \right],
\end{aligned}$$

where $d_1 = \limsup_{N \to \infty} N^{-\gamma} \lambda_N \in (0, \infty)$, $d_2 = \liminf_{N \to \infty} \frac{\lambda(t_N)}{\lambda_N} \in ((1-\varepsilon)\lambda_N, 1]$.

Similarly,

$$
\begin{aligned}
&N^{-\gamma}m^{-1}\left(\sup_{\tilde{\lambda}} Q_{m,n,\alpha}(z(t_N),\tilde{\lambda}) - mA_0\right) \geqslant \\
&N^{-\gamma}\Big(\frac{q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon,m)/m}{\lambda_\varepsilon}\frac{n(\lambda_\varepsilon)}{N(\lambda_\varepsilon)}\lambda_\varepsilon + \frac{z(t_N)}{N(\lambda_\varepsilon)}\frac{m(\lambda_\varepsilon)}{m} + \\
&\frac{\beta_{\frac{\alpha}{3},m(\lambda_\varepsilon)}}{m}\sqrt{\frac{m(\lambda_\varepsilon)}{N(\lambda_\varepsilon)}\frac{n(\lambda_\varepsilon)}{N(\lambda_\varepsilon)}\frac{N(\lambda_\varepsilon)-z(t_N)}{N(\lambda_\varepsilon)-1}\frac{z(t_N)-q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon,m)}{m}} - A_0\Big) \\
&= N^{-\gamma}\lambda_N\Big[\frac{q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon,m)/m}{\lambda_\varepsilon}\frac{n(\lambda_\varepsilon)}{N(\lambda_\varepsilon)}\frac{\lambda_\varepsilon}{\lambda_N} + \frac{z(t_N)/N(\lambda_\varepsilon)(m(\lambda_\varepsilon)/m) - A_0(t_N)}{p_N - A_0(t_N)}\frac{[p_N - A_0]}{\lambda_N}\Big] \\
&+ N^{-\gamma}\frac{\beta_{\frac{\alpha}{3},m(\lambda_\varepsilon)}}{m}\sqrt{\frac{m(\lambda_\varepsilon)}{N(\lambda_\varepsilon)}\frac{n(\lambda_\varepsilon)}{N(\lambda_\varepsilon)}\frac{N(\lambda_\varepsilon)-z(t_N)}{N(\lambda_\varepsilon)-1}\frac{z(t_N)-q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon,m)}{m}} \\
&\geq \Big[\frac{q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon,m)/m}{\lambda_\varepsilon}\frac{n(\lambda_\varepsilon)}{N(\lambda_\varepsilon)}\frac{\lambda_\varepsilon}{\lambda_N} - \frac{z(t_N)/N(\lambda_\varepsilon)(m(\lambda_\varepsilon)/m) - A_0(t_N)}{p_N - A_0(t_N)}(1-\pi)\sup_{M\geqslant N}\frac{\lambda(t_M)}{\lambda_M}\Big]\inf_{M\geqslant N} M^{-\gamma}\lambda_M \\
&+ N^{-\gamma}\frac{\beta_{\frac{\alpha}{3},m(\lambda_\varepsilon)}}{m}\sqrt{\frac{m(\lambda_\varepsilon)}{N(\lambda_\varepsilon)}\frac{n(\lambda_\varepsilon)}{N(\lambda_\varepsilon)}\frac{N(\lambda_\varepsilon)-z(t_N)}{N(\lambda_\varepsilon)-1}\frac{z(t_N)-q_{1-\frac{\alpha}{3}}(\lambda_\varepsilon,m)}{m}} \\
&\xrightarrow{p} d_3(1-\pi)\left[(1-\varepsilon)-d_4\right],
\end{aligned}
$$

where $d_3 = \liminf_{N\to\infty} N^{-\gamma}\lambda_N \in (0,\infty)$, $d_4 = \liminf_{N\to\infty}\frac{\lambda(t_N)}{\lambda_N} \in ((1-\varepsilon)\lambda_N, 1]$. The convergence in probability follows because $m(\lambda_\varepsilon)/m \asymp 1-\lambda_\varepsilon \to 1$ and thus using the same proof as for (62), it holds that

$$
\frac{z(t_N)/N(\lambda_\varepsilon)(m(\lambda_\varepsilon)/m) - A_0(t_N)}{p_N - A_0(t_N)} \xrightarrow{p} 1, \text{ for } N\to\infty.
$$

Additionally,

$$
\begin{aligned}
&N^{-\gamma}m^{-1}(\tilde{Q}(\lambda_\varepsilon) - mA_0) \leqslant (1-\pi)\left[(1-\varepsilon) - \inf_{M\geqslant N}\frac{\lambda(t_M)}{\lambda_M}\right]\sup_{M\geqslant N}(M^{-\gamma}\lambda_M) \\
&\to d_1(1-\pi)\left[(1-\varepsilon)-d_2\right], \text{ for } N\to\infty
\end{aligned}
$$

and

$$
\begin{aligned}
&N^{-\gamma}m^{-1}(\tilde{Q}(\lambda_\varepsilon) - mA_0) \geqslant (1-\pi)\left[(1-\varepsilon) - \sup_{M\geqslant N}\frac{\lambda(t_M)}{\lambda_M}\right]\inf_{M\geqslant N}(M^{-\gamma}\lambda_M) \\
&\to d_3(1-\pi)\left[(1-\varepsilon)-d_4\right], \text{ for } N\to\infty.
\end{aligned}
$$

Thus taking,

$$
c = \frac{d_3(1-\pi)\left[(1-\varepsilon)-d_4\right]}{d_1(1-\pi)\left[(1-\varepsilon)-d_2\right]}
$$

we obtain (58).

■

Lemma .6. *Let* $-1 < \gamma < 0$ *and* $\varepsilon \in (0,1]$ *arbitrary and define* $\sigma_F^2(t_N) = A_0(t_N)(1-A_0(t_N))/m$. *If for a sequence* $(t_N)_{N\geqslant 1}$ *and* $\rho = \rho^*$,

$$\liminf_{N\to\infty} m\sigma_F < +\infty \tag{NC'}$$

then

(I) (20) *or* (21) *is not true,*

(II) (C_ε) *is not true for* $\lambda = \hat{\lambda}^{\rho^*}_{adapt}$.

Proof First note that (**NC'**) implies

$$\sigma_F = o(N^\zeta), \tag{64}$$

for all $\zeta \in (-1,0]$, as in Lemma .4.

(I): With the same arguments as in Lemma .4, (**NC'**) implies two possible cases

(1') $\liminf_{N\to\infty} NA_0(t_N) < \infty$, $\liminf_{N\to\infty} N(1-A_1(t_N)) < \infty$

(2') $\liminf_{N\to\infty} N(1-A_0(t_N)) < \infty$, $\liminf_{N\to\infty} NA_1(t_N) < \infty$

and these in turn imply

(1) $A_0(t_N) = o(N^\zeta)$, $1-A_1(t_N) = o(N^\zeta)$

(2) $1-A_0(t_N) = o(N^\zeta)$, $A_1(t_N) = o(N^\zeta)$,

for all $\zeta \in (-1,0]$. But then (1) and (2) imply, $\lambda(t_N) = o(N^\zeta)$, for all $\zeta \in (-1,0]$, contradicting (20) for $\varepsilon < 1$. For $\varepsilon = 1$ assuming (21b) to be true and following the exact same subsequence argument for (1') and (2') in turn as in Lemma .4 (I) results in a contradiction and thus (21b) cannot be true.

(II): Let V_{m,t_N} be defined as in Proposition 14. (1), (2) make it clear that in our setting, (**NC'**) and (45) are equivalent. Moreover, in the same way as in Lemma .4, for all $\delta > 0$

$$\mathbb{P}(\lambda_N^{-1}(\hat{F}_m(t_N) - A_0(t_N)) > \delta) \leqslant \frac{\lambda_N^{-2}\sigma_F^2}{\delta} \asymp \frac{[N^\gamma \sigma_F]^2}{\delta} \to 0, \text{ for } N \to \infty.$$

Now assume that despite (**NC'**),

$$\mathbb{P}(V_{m,t_N} > Q_{m,n,\alpha}(z(t_N), \tilde{\lambda}) \ \ \forall \tilde{\lambda} \in [0, \lambda_\varepsilon]) \to 1, \text{ for } N \to \infty, \tag{65}$$

holds true. Then, using Lemma .5 and the arguments in Proposition 14, also

$$\mathbb{P}(V_{m,t_N} > \tilde{Q}(\lambda_\varepsilon)) \to 1, \text{ for } N \to \infty, \tag{66}$$

must hold. However, for $\varepsilon < 1$,

$$\begin{aligned}\mathbb{P}(V_{m,t_N} > \tilde{Q}(\lambda_\varepsilon)) &= \mathbb{P}(V_{m,t_N} - mA_0(t_N) > m\lambda_N[(1-\varepsilon)(1-\pi) - \frac{\lambda(t_N)}{\lambda_N}]) \\ &= \mathbb{P}(\lambda_N^{-1}(\hat{F}_m(t_N) - A_0(t_N)) > [(1-\varepsilon)(1-\pi) - \frac{\lambda(t_N)}{\lambda_N}])\end{aligned}$$

and as from (I), $\frac{\lambda(t_N)}{\lambda_N} \to 0$ and $\lambda_N^{-1}(\hat{F}_m(t_N) - A_0(t_N)) \xrightarrow{p} 0$, this probability will converge to zero, contradicting (65) for $\varepsilon \in [0,1)$. For $\varepsilon = 1$, note that (65) also implies that

$$\mathbb{P}(V_{m,t_N} > Q_{m,n,\alpha}(z(t_N),0)) \to 1, \text{ for } N \to \infty. \tag{67}$$

Now by definition of $Q_{m,n,\alpha}(z(t_N),0)$,

$$\mathbb{P}(V_{m,t_N} > Q_{m,n,\alpha}(z(t_N),0)) \leqslant \mathbb{P}(V_{m,t_N} - \frac{m}{N}z(t_N) > 0),$$

and since $V_{m,t_N} = m\hat{F}_m(t_N)$ and $z(t_N) = m\hat{F}_m(t_N) + n\hat{G}_m(t_N)$,

$$\begin{aligned}\mathbb{P}(V_{m,t_N} > Q_{m,n,\alpha}(z(t_N),0)) &\leqslant \mathbb{P}\left(\frac{n}{N}m\hat{F}_m(t_N) - \frac{m}{N}n\hat{G}_m(t_N) > 0\right) \\ &= \mathbb{P}(\hat{F}_m(t_N) - \hat{G}_m(t_N) > 0).\end{aligned}$$

We now can use *exactly* the same argument as in Lemma .4, (II), to obtain that for a correctly chosen subsequences,

$$\limsup_{k\to\infty} \mathbb{P}(\hat{F}_{m(k)}(t_{N(k)}) - \hat{G}_{n(k)}(t_{N(k)}) > 0) < 1,$$

contradicting (65). Thus finally, (C_ε) cannot be true if (**NC'**) is true.

■

Technical tools for Proposition 10: We now introduce two concepts that will help greatly in the proof of Proposition 10. The first concept is that of "Distributional Witnesses". We assume to observe two iid samples of independent random elements X, Y with values in $(\mathcal{X}, \mathcal{A})$ with respective probability measures P and Q. Similar as in [36], let $\mathfrak{C}$ be the set of all random elements $(\tilde{X}, \tilde{Y})$ with values in $(\mathcal{X}^2, \mathcal{A}^2)$, and such that $\tilde{X} \sim P$ and $\tilde{Y} \sim Q$. Following standard convention, we call $(\tilde{X}, \tilde{Y}) \in \mathfrak{C}$ a *coupling* of P and Q. Then $\mathrm{TV}(P,Q)$ may be characterized as

$$\mathrm{TV}(P,Q) = \inf_{\mathfrak{C}} \mathbb{P}(\tilde{X} \neq \tilde{Y}). \tag{68}$$

This is in turn equivalent to saying that we minimize $\Pi(x \neq y)$ over all joint distributions Π on $(\mathcal{X}^2, \mathcal{A}^2)$, that have $X_\#\Pi = P$ and $Y_\#\Pi = Q$. Equation (68) allows for an interesting interpretation, as detailed (for example) in [36]: The optimal value is attained for a coupling

(X^*, Y^*) that minimizes the probability of $X^* \neq Y^*$. The probability that they are different is exactly given by $\mathrm{TV}(P, Q)$. It is furthermore not hard to show that the optimal coupling is given by the following scheme: Let $W \sim \text{Bernoulli}(\mathrm{TV}(P, Q))$ and denote by f the density of P and g the density of Q, both with respect to some measure on $(X, \mathcal{A})$, e.g. $P + Q$. If $W = 0$, draw a random element Z from a distribution with density $\min(f, g)/(1 - \mathrm{TV}(P, Q))$ and set $X^* = Y^* = Z$. If $W = 1$, draw X^* and Y^* independently from $(f - g)_+/\mathrm{TV}(P, Q)$ and $(g - f)_+/\mathrm{TV}(P, Q)$ respectively.

Obviously, X^* and Y^* so constructed are dependent and do not directly relate to the observed X, Y, which are assumed to be independent. However it holds true that marginally, $X \overset{D}{=} X^*$ and $Y \overset{D}{=} Y^*$. In particular, given that $W = 1$, it holds that $X \overset{D}{=} X^* = Y^* \overset{D}{=} Y$, or $X|\{W = 1\} \overset{D}{=} Y|\{W = 1\}$. On the other hand, for $W = 0$, the support of X and Y is disjoint. This suggests that the distribution of X and Y might be split into a part that is common to both and a part that is unique. Indeed, the probability measures P and Q can be decomposed in terms of three probability measures H_P, H_Q, $H_{P,Q}$ such that

$$P = \lambda H_P + (1 - \lambda) H_{P,Q} \quad \text{and} \quad Q = \lambda H_Q + (1 - \lambda) H_{P,Q}, \tag{69}$$

where the mixing weight is $\lambda = \mathrm{TV}(P, Q)$.

Viewed through the lens of random elements, these decompositions allow us to view the generating mechanism of sampling from P and Q respectively as equivalent to sampling from the mixture distributions in (69). Indeed we associate to X (equivalently for Y) the latent binary indicator W^P, which takes value 1 if the component specific to P, H_P, is "selected" and zero otherwise. As before, it holds by construction $\mathbb{P}(W^P = 1) = \mathrm{TV}(P, Q)$. Intuitively an observation X with $W^P = 1$ reveals the distribution difference of P with respect to Q. This fact leads to the following definition:

Definition 24 (Distributional Witness)**.** *An observation X from P with latent realization $W^P = 1$ in the representation of P given by* (69) *is called a* distributional witness *of the distribution P with respect to Q. We denote by $\mathrm{DW}_m(P; Q)$ The number of witness observations of P with respect to Q out of m independent observations from P.*

The second concept is that of a bounding operation: Let $\bar{\Lambda}_P \in \mathbb{N}$, $\bar{\Lambda}_Q \in \mathbb{N}$ be numbers *overestimating* the true number of distributional witnesses from m iid samples from P and n iid samples from Q, i.e.

$$\bar{\Lambda}_P \geqslant \Lambda_P := \mathrm{DW}_m(P; Q), \ \bar{\Lambda}_Q \geqslant \Lambda_Q := \mathrm{DW}_n(Q; P). \tag{70}$$

Thus, it could be that $\bar{\Lambda}_P, \bar{\Lambda}_Q$ denote the true number of witnesses, but more generally, they need to be larger or equal. If $\bar{\Lambda}_P > \Lambda_P$ or $\bar{\Lambda}_Q > \Lambda_Q$, a *precleaning* is performed: We randomly choose a set of $\bar{\Lambda}_P - \Lambda_P$ non-witnesses from the sample of F and $\bar{\Lambda}_Q - \Lambda_Q$ non-witnesses

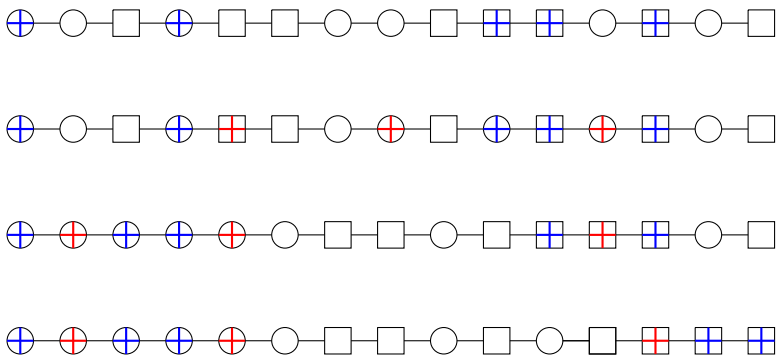


Figure 10: Illustration of the bounding operation. The first row from above is the original order statistics shown as circles (coming from F) and squares (coming from G). Witnesses are indicated by blue crosses. In the second, randomly chosen non-witnesses are added to the list of witnesses left and right, indicated by red, until the number of witnesses is $\bar{\Lambda}_P$ and $\bar{\Lambda}_Q$. In the final two rows, the witnesses of F and G are pushed to the left and right respectively, such that the original order of the non-witnesses in the second row is kept intact.

from the sample of G and mark them as witnesses. Thus we artificially increase the number of witnesses left and right to $\bar{\Lambda}_P$, $\bar{\Lambda}_Q$. Given this sample of witnesses and non-witnesses and starting simultaneously from the first and last order statistics $Z_{(1)}$ and $Z_{(N)}$, for $i \in \{1, \dots, N\}$ in the combined sample, we do:

(1) If $i < \bar{\Lambda}_P$ and $Z_{(i)}$ is *not* a witness from F, replace it by a witness from F, randomly chosen out of all the remaining F-witnesses in $\{Z_{(i+1)}, \dots Z_{(N)}\}$. Similarly, if $i < \bar{\Lambda}_Q$ and $Z_{(N-i+1)}$ is *not* a witness from G, replace it by a witness from G, randomly chosen out of all the remaining G-witnesses in $\{Z_{(1)}, \dots Z_{(N-i)}\}$.

(2) Set $i = i + 1$.

We then repeat (1) and (2) until $i = \max\{\bar{\Lambda}_P, \bar{\Lambda}_Q\}$.

This operation is quite intuitive: we move from the left to the right and exchange points that are not witnesses from F (i.e. either non-witnesses or witnesses from G), with witnesses from F that are further to the right. This we do, until all the witnesses from F are aligned in the first $\bar{\Lambda}_P$ positions. We also do the same for the witnesses of G in the other direction of the order statistics. Figure 10 illustrates this operation. Implementing the same counting process that produced $V_{m,z}$ in the original sample leads to a new counting process $z \mapsto \bar{V}_{m,z}$. Lemma .7 collects some properties of this process, which is now much more well-behaved than the original $V_{m,z}$.

Lemma .7. *$\bar{V}_{m,z}$ obtained from the bounding operation above has the following properties:*

(i) $\mathbb{P}\left(\forall z \in J_{m,n} : \bar{V}_{m,z} \geqslant V_{m,z}\right) = 1$*, i.e. it stochastically dominates $V_{m,z}$.*

(ii) It increases linearly with slope 1 for the first $\bar{\Lambda}_P$ observations and stays constant for the last $\bar{\Lambda}_Q$ observations.

(iii) If $\bar{\Lambda}_P < m$ and $\bar{\Lambda}_Q < n$ and for $z \in \{\bar{\Lambda}_P + 1, \dots, N - \bar{\Lambda}_Q - 1\}$, it factors into $\bar{\Lambda}_P$ and a process $\tilde{V}_{m-\bar{\Lambda}_P, z-\bar{\Lambda}_P}$, with

$$\tilde{V}_{m-\bar{\Lambda}_P, z-\bar{\Lambda}_P} \sim \textit{Hypergeometric}\left(z - \bar{\Lambda}_P, m + n - \bar{\Lambda}_P - \bar{\Lambda}_Q, m - \bar{\Lambda}_P\right). \tag{71}$$

Proof (i) follows, as $\bar{V}_{m,z}$ only counts observations from F and these counts can only increase when moving the witnesses to the left. (ii) follows directly from the bounding operation, through (70).

(iii) According to our assumptions, we deal with the order statistics of two independent iid samples $(X_1, W_1^X), \dots (X_m, W_m^X)$ and $(Y_1, W_1^Y), \dots (Y_n, W_n^Y)$, with $X|W^X = 1$ being equal in distribution to $Y|W^Y = 1$. We consider their order statistics $(Z_{(1)}, W_1^Z), \dots (Z_{(N)}, W_N^Z)$. In the precleaning step, we randomly choose $\bar{\Lambda}_P - \Lambda_P$ i such that $W_i^P = 0$ and $\bar{\Lambda}_Q - \Lambda_Q$ j such that $W_j^P = 0$ and flip their values such that $W_i^P = 1$ and $W_j^Y = 1$. Let $\mathcal{I}(\bar{\Lambda}_P, \bar{\Lambda}_Q)$ denote the index set $\{i : W_i^P = 1 \text{ or } W_i^Y = 1\}$ and let $\mathcal{I}^c := \mathcal{I}(\bar{\Lambda}_P, \bar{\Lambda}_Q)^c = \{1, \dots, N\} \backslash \mathcal{I}(\bar{\Lambda}_P, \bar{\Lambda}_Q)$. "Deleting" all observations, we remain with the order statistics $(Z_{(i)})_{i \in \mathcal{I}^c}$. By construction, up to renaming the indices, we obtain an order statistics $Z_{(1)}, \dots, Z_{(N-\bar{\Lambda}_P-\bar{\Lambda}_Q)}$ drawn from the common distribution $H_{P,Q}$. Therefore the counting process $V_{\mathcal{I},z} = (m - \bar{\Lambda}_P)\hat{F}(Z_{(z)})$ is a hypergeometric process. ■

Paper

B

Imputation Scores.

J. Näf, M. Spohn, L. Michel, N. Meinshausen

To appear in Annals of Applied Statistics.

Imputation Scores

Jeffrey Näf Meta-Lina Spohn
Loris Michel Nicolai Meinshausen

January 20, 2023

Abstract

Given the prevalence of missing data in modern statistical research, a broad range of methods is available for any given imputation task. How does one choose the 'best' imputation method in a given application? The standard approach is to select some observations, set their status to missing, and compare prediction accuracy of the methods under consideration of these observations. Besides having to somewhat artificially mask observations, a shortcoming of this approach is that imputations based on the conditional mean will rank highest if predictive accuracy is measured with quadratic loss. In contrast, we want to rank highest an imputation that can sample from the true conditional distributions. In this paper, we develop a framework called "Imputation Scores" (I-Scores) for assessing missing value imputations. We provide a specific I-Score based on density ratios and projections, that is applicable to discrete and continuous data. It does not require to mask additional observations for evaluations and is also applicable if there are no complete observations. The population version is shown to be proper in the sense that the highest rank is assigned to an imputation method that samples from the correct conditional distribution. The propriety is shown under the *missing completely at random* (MCAR) assumption but is also shown to be valid under *missing at random* (MAR) with slightly more restrictive assumptions. We show empirically on a range of data sets and imputation methods that our score consistently ranks true data high(est) and is able to avoid pitfalls usually associated with performance measures such as RMSE. Finally, we provide the R-package `Iscores` available on CRAN with an implementation of our method.

Keywords. Ranking, Random Projections, Tree Ensembles, Random Forest, KL-Divergence.

1 Introduction

Missing data is a widespread problem in both research and practice. In the words of [121], "imputing incomplete observations is becoming an indispensable intermediate phase between the two traditional phases [...] of collecting data and analyzing data." With the increasing dimensionality and size of modern data sets, the problem of missing data only gets more pronounced. One reason for this is that even for moderate dimensionality and overall fraction of missing entries, the set of complete observations tends to be small if not zero. As such, one would often have to discard a substantial part of the data when keeping only complete observations. In addition, depending on the missingness mechanism, working with complete cases is invalid in many situations.

Consequently, there is a large body of literature on imputation methods, filling in the missing entries in a given data set, see e.g., [108]. Such methods include the very general *Multivariate Imputation by Chained Equations* (mice) approach [176, 38], *missForest* [167], and *multiple imputation in principal component analysis* (mipca) [84]. In more recent work, methods based on non-parametric Bayesian strategies [127], generative adversarial networks [189] and optimal transport [128] were developed. These are just a few examples. The 'R-miss-tastic' platform, developed in an effort to collect knowledge and methods to streamline the task of handling missing data, lists over 150 packages [114].

Despite the broad range of imputation methods, not a lot of attention has been paid in the literature to the question of how to evaluate and choose an imputation method for a given data set. If the true data underlying the missing entries are available (achievable if observations are artificially masked), the imputed values are often simply compared to the true ones through the root mean-squared error (RMSE) or mean absolute error (MAE) (see e.g., [167], [182], [189], [128] and many others). For categorical variables, the percentage of correct predictions (PCP) can be used [103]. To select an imputation method, the one with the lowest overall error value is chosen. This simple method is common, despite having major drawbacks. For example, RMSE (respectively, MAE and PCP) favors methods that impute with the conditional means (respectively, conditional medians and conditional modes), versus samples from the true conditional distributions [57]. As outlined in [175, Chapter 2.5.1], this may lead to a choice of "nonsensical" imputation methods. In particular, they tend to artificially strengthen the association between variables, which can lead to invalid inference. A motivating example of such a case is given in Section 3.

Another approach is to fix a *target quantity*, calculated once on the full data and once on the imputed data and use suitable distances between them for a quality assessment of the imputation. The target quantity can be an expectation, a regression or correlation coefficient, a variance [72, 45, 5, 55, 188] or more involved estimands such as treatment effects using propensity scores

[31]. While this approach is sensible if (one) target quantity can clearly be defined, this may not be ideal in many applications. Often it is not even clear beforehand what the target of interest is, or one deals with several targets of interests, each of which might lead to a different choice of imputation methods [175, Chapter 2.3.4].

As such, there are situations where one might prefer to have a tool that is able to identify a good imputation method for a wide range of targets. A way to achieve this is to simply define the notion of target measure more broadly. As noted in [128]: "A desirable property of imputation methods is that they should preserve the joint and marginal distributions." That is, if x^* is a complete observation and z an observation with imputed values, we want them to be realizations of the same distribution. This can be seen as a special case of the above target quantity approach, with the target being the underlying joint distribution from which the data were generated.

With this goal in mind, more elaborate distributional metrics, such as ϕ-divergences [35] or integral probability metrics (see e.g., [164]) are necessary, even when simply comparing an imputed data set to the true one. This was utilized in [128] with the Wasserstein Distance (WD), but seems otherwise uncommon, despite the drawbacks of measures like RMSE and MAE.

In applications, the true data for the missing values are rarely available. Ad-hoc methods have been proposed to assess the success of the imputation; one such approach is to mask some observed values, impute them, and then compare (via, say, RMSE) the imputations with the observed values. In this paper, we aim to assess more directly the quality of an imputation method. We propose the flexible framework of proper Imputation Scores (I-Scores) to evaluate imputation methods when (i) the target measure is the true data distribution, (ii) the true data underlying the missing values are not available, and (iii) we do not want to artificially mask observations for the evaluation. To overcome the difficulty of observing only incomplete data, we use random projections in the variable space. We propose a specific estimator of an I-Score based on density ratios (DR I-Score). The density ratio is estimated through classification, where the classifier acts as a discriminator between observed and imputed distribution. The result is an easy-to-use imputation scoring method. Under suitable assumptions on the missingness mechanism, we prove propriety of our DR I-Score, meaning that the true underlying data distribution is ranked highest in expectation. The propriety is shown under a *missing completely at random* (MCAR) assumption on the missingness mechanism. It is also shown to be valid under *missing at random* (MAR) if we restrict the random projections in the variable space to always include all variables, which in turn requires access to some complete observations. Empirical results show that indeed true data samples are generally ranked highest by the proposed I-Score if compared with widely used imputation methods. We provide an implementation of the method in the R-package `Iscores`, available on CRAN and GitHub (`https://github.com/missValTeam/Iscores`).

Section 2 introduces the notation and Section 3 showcases a motivating example. Section 3 defines the framework of Imputation Scores (I-Scores) along with a specific I-Score, presenting a way of evaluating imputation methods in the presence of missing values. Section 4 then details the estimation of the I-Score and describes the algorithm. Section 1.2 presents further related work. Section 5 empirically validates the estimated I-Score on a range of data sets and on a real data set with missing values, and Section 8 concludes with a discussion.

2 Notation

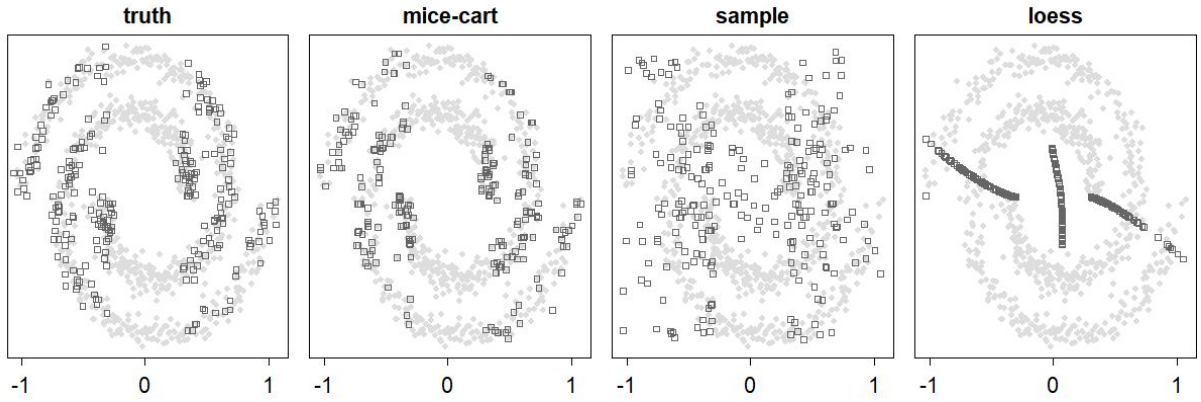
We assume an underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$ on which all random elements are defined. Throughout, we take $\mathcal{P}$ to be a collection of probability measures on $\mathbb{R}^d$, dominated by some σ-finite measure μ. We denote the (unobserved) complete data distribution by $P^* \in \mathcal{P}$ and by P the actually observed distribution with missing values. We assume that P (P^*) has a density p (p^*). We take X (X^*) to be the random vector with distribution P (P^*) and let x_i (x_i^*), $i = 1, \ldots, n$, be realizations of an i.i.d. copy of the random vector X (X^*). Similarly, M is the random vector in $\{0,1\}^d$, encoding the missingness pattern of X, with realization m, whereby for $j = 1, \ldots, d$, $m_j = 0$ means that variable j is observed, while $m_j = 1$ means it is missing. For instance, the observation $(\texttt{NA}, x_2, x_3)$ corresponds to the pattern $(1,0,0)$. We denote the distribution of M as P^M, with support $\mathcal{M}$, so that $\mathbb{P}(M = m) = P^M(m)$.

For a subset $A \subseteq \{1, \ldots, d\}$ and for a random vector X or an observation x in $\mathbb{R}^d$, we denote with X_A (x_A) its projection onto that subset of indices. For instance if $d = 3$ and $A = \{1,2\}$, then $X_A = (X_1, X_2)$ ($x_A = (x_1, x_2)$). The projection onto A of the observation x_i, $(x_i)_A$, is denoted as $x_{i,A}$. Analogously, for a missingness pattern $M \sim P^M$ or an observation m in $\{0,1\}^d$, we denote with M_A (m_A) its projection onto the subset of indices in A. If X has a density p on $\mathbb{R}^d$, we denote by p_A the density of the projection X_A.

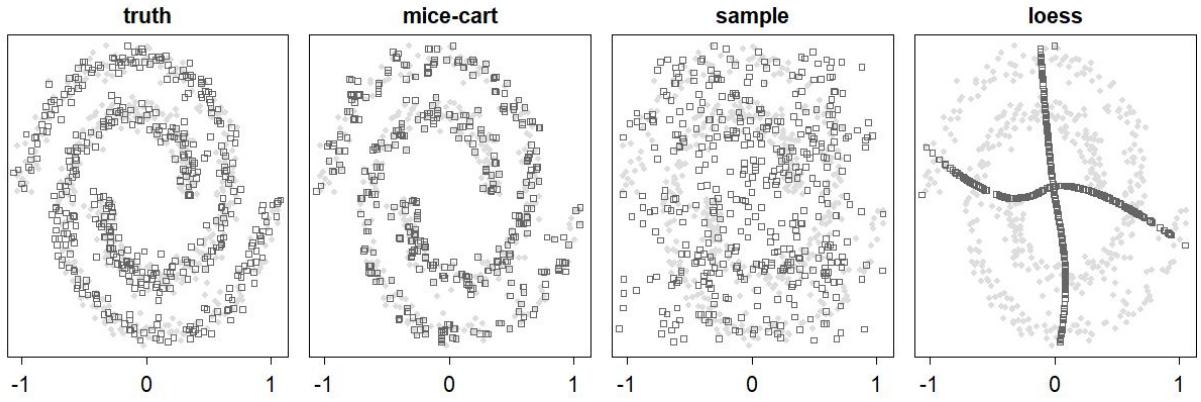
To denote assumptions on the missingness mechanism, we use a notation along the lines of [151]. For each realization m of the missingness random vector M we define with $o(X,m) := (X_j)_{j \in \{1,\ldots,d\}: m_j = 0}$ the observed part of X according to m and with $o^c(X,m) := (X_j)_{j \in \{1,\ldots,d\}: m_j = 1}$ the corresponding missing part. Note that this operation only filters the corresponding elements of X according to m, regardless whether or not these elements are actually missing or not. For instance, we might consider the unobserved part $o^c(X,m)$ according to m for the fully observed X, that is $X \sim P | M = \mathbf{0}$, where $\mathbf{0}$ denotes the vector of zeros of length d.

3 Motivating Example

As a toy motivating example we consider a noisy version of the spiral in two dimensions from the R-package `mlbench` [100]. The concepts and tools introduced here will be detailed and used



(a) MAR



(b) MCAR

Figure 1: Imputations for the spiral example ($n = 1000$, $p = 2$). The complete observations are plotted in light gray with dots and the imputed observations in dark gray with squares. From left to right, true data, mice-cart, sample and loess are shown. The top row shows the MAR case, the bottom row the MCAR case.

in the remainder of the paper. We generated 1000 observations of the noisy spiral, each entry having a probability of being missing of $p_{miss} = 0.3$ with missingness mechanisms MAR and MCAR. In the case of MAR, the variable X_2 is missing with probability p_{miss} if the corresponding X_1 is > 0.3 or < -0.3 and observed otherwise. The variable X_1 is missing with probability p_{miss}, if $X_2 \in [-0.3, 0.3]$. In the case of MCAR, we set every value in the data matrix to `NA` with probability p_{miss}. For MCAR we face already in this low dimensional example an average of around $(1 - (1 - 0.3)^2) * 1000 \approx 500$ observations with at least one missing value, which is half the sample size.

We might decide to impute the missing values and do so with three methods: i) simply estimating the conditional expectation $\mathbb{E}[X_1|X_2]$ and $\mathbb{E}[X_2|X_1]$ on the complete cases using a local regression approach ("loess"), and filling in the missing values by predicting from X_1 (if X_2 is missing) or from X_2 (if X_1 is missing), ii) random sampling an observed value for each missing entry ("sample"), and iii) mice [176] combined with a single tree in each iteration ("mice-cart"). See Appendix 1 for more explanation on sample and mice-cart.

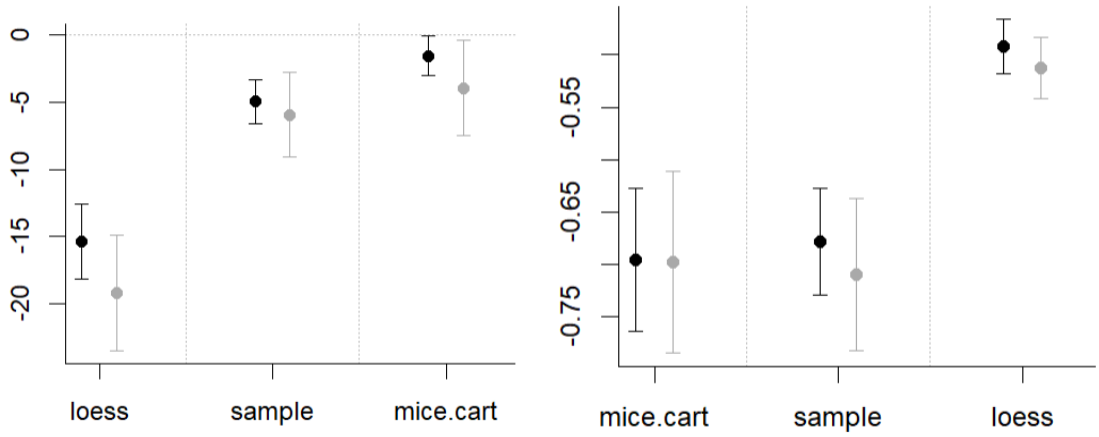
Though a very artificial example, it highlights some interesting features of different evaluation methods of the imputations. As mentioned in the introduction, our target is P^*, the full distribution of the data. In this two dimensional example, a visual evaluation is possible. Figure 1 shows the resulting imputations. While mice-cart and the true underlying data are hard to distinguish, it is apparent that sample is worse than either and loess in turn is much worse than sample.

We may now try to quantify this visually obtained ordering. For the three imputations as well as the true underlying data we compute our DR I-Score (defined later). This score is positively oriented: A higher value indicates a better performance. We additionally compute the negative of RMSE ("negRMSE"), where the negative sign assures the same orientation of higher values indicating better performance. We emphasize that negRMSE is computed using the unobserved full data set, as commonly done in research papers introducing new methods of imputation. We also computed approximated two-sided 95%-confidence intervals (CI) of the DR I-Score and negRMSE, as detailed in Section 4. The results are shown in Figure 3. In the left plot (a), the score of the true data was subtracted from the scores to visualize the comparison to the true data imputation. In the right plot (b), normalizing negRMSE scores, using the true data imputation, is unnecessary, since negRMSE is 0 by definition for the true data imputation.

Maybe unsurprisingly, negRMSE appears to poorly measure the distributional difference between imputed and real data set. In particular, for MCAR and MAR its value is highest for the loess imputation, and significantly so, based on the approximated CI. This is despite the fact that negRMSE is allowed to use the unobserved data. In contrast, our proposed DR I-Score has no access to the unobserved data, and nonetheless ranks true data clearly highest, with mice-cart as a close second, followed by sample and loess. Thus, without using the unobserved samples, our score manages to give a sensible ranking in this example that is in line with the visual impression, for both a MCAR and MAR missingness mechanisms.

4 Scores for Imputations

Our notion of proper Imputation Score (I-Score) is inspired by the classical notion of proper scores (or proper scoring rules) and proper divergence functions (see e.g. [58, 57, 170]). A score assesses whether a probabilistic forecast is close to the true distribution. In a traditional sense, a score S takes a predictive distribution $Q_1 \in \mathcal{P}$ as a first argument, as well as $X \sim Q_2 \in \mathcal{P}$ as a second: $S(Q_1, X)$. The corresponding expectation over the distribution Q_2 is denoted by $S(Q_1, Q_2) := \mathbb{E}_{X\sim Q_2}[S(Q_1, X)]$. Thus, a score value is assigned to Q_1 over the comparison with Q_2. A desirable natural property is that $S(Q_1, Q_2) \leqslant S(Q_2, Q_2)$, for all $Q_1, Q_2 \in \mathcal{P}$. This ensures that the true distribution Q_2 of X is scored highest in expectation. A score that meets this requirement is referred to as proper.



(a) Estimated DR I-Score with CIs.

(b) Negative RMSE with CIs.

Figure 2: Spiral example ($n = 1000$, $p = 2$): Estimation of the proposed DR I-Score (a) and the negRMSE (b) with corresponding CIs for the methods mice-cart, sample and loess under the missingness mechanisms MCAR (black) and MAR (grey). We obtained the CIs by subsampling as described in Section 5.3. In (a) we subtracted the score of the true data from the scores of the methods, thus the line at 0 represents the true data score. We used $p_{miss} = 0.3$ to generate missing values.

4.1 Imputation Score (I-Score)

We now define the notion of a proper I-Score. Despite the analogy to the classical notion of scores, we will need some deviations for the setting of imputation scores as applied to partially observed samples. Recall that P refers to the distribution of X with missing values and correspondingly, $P^* \in \mathcal{P}$ refers to the distribution of X^* without missing values. We denote with P^M the distribution of M with domain $\mathcal{M}$. Naturally, (P^*, P) form a tuple, whereby P is derived from P^* and P^M. We define $\mathcal{H}_P \subset \mathcal{P}$ to be the set of imputation distributions compatible with P, that is

$$\mathcal{H}_P := \{H \in \mathcal{P} : h(o(x,m)|M = m) = p(o(x,m)|M = m) \text{ for all } m \in \mathcal{M}\}, \tag{1}$$

where as above for a pattern m, $o(x,m) = (x_j)_{j\in\{1,\dots,d\}:m_j=0}$ subsets the observed elements of x according to m, while $o^c(x,m) = (x_j)_{j\in\{1,\dots,d\}:m_j=1,}$ subsets the missing elements[1]. Clearly, $P^* \in \mathcal{H}_P$, so that the true distribution P^* can be seen as an imputation.

Definition 25. ***Imputation Score (I-Score)***

Let (P^*, P) be the tuple of distributions as described above and $H \in \mathcal{H}_P$, as given in (1)*. A real-valued function $S_{NA}(H,P)$ is a proper I-Score iff*

$$S_{NA}(H,P) \leqslant S_{NA}(P^*,P),$$

[1]Note that while h and p are densities on $\mathbb{R}^d$, notation is slightly abused by using expressions such as $h(o(x,m)|M = m)$ and $p(o(x,m)|M = m)$, which are densities on $\mathbb{R}^{|\{j:m_j=0\}|}$.

for any imputation distribution $H \in \mathcal{H}_P$. It is strictly proper iff the inequality is strict for $H \neq P^*$.

In practice, we do not have access to H and P but can only draw samples from (H, P) by drawing a sample x from P (with missingness pattern m) and the corresponding coupled imputation z from H that matches x on all observed variables, so that $o(z, m) = o(x, m)$. In the ideal case where (Z, X) are drawn from (P^*, P), the conditional distribution of $o^c(Z, M)$ given $o(Z, M)$ is identical to the conditional distribution of $o^c(X^*, M)$ given $o(X^*, M)$. In principle, we could also score the full conditional distribution of imputation algorithms instead of just supposing that we can draw samples from the conditional distribution. However, most popular imputation algorithms do not specify the conditional distribution explicitly and we thus focus on the latter case. We discuss the sample-based implementation in Section 4.

A key difference to the classical notion of scores [58, 57] is that we do not observe a sample from the 'true' distribution of the missing values of x, given missingness pattern m and observed values $o(x, m)$. We can hence not compare the predictions of the missing values (whether they are explicit distributional predictions or samples from the imputation) to the realization of a sample from the underlying 'true' conditional distribution. It seems perhaps surprising that we can still obtain proper I-Scores in this setting. A central assumption is that the missingness mechanism is such that any subset of variables has a positive probability of being observed (in contrast to a classical regression setting where the target variable is always unobserved before the prediction is made). The variability of the missingness patterns hence allows to construct proper I-Scores. An alternative approach would be to mask observed values as unobserved and score the imputation on these held-out data. This would require to change the distribution of the missingness patterns, however, which would be especially problematic in absence of a MCAR assumption.

We immediately see that negRMSE used in the motivating example of Section 3 does not fit into the framework of proper I-Scores, simply because it requires access to the true underlying values according to P^*.

We now define a specific I-Score that satisfies Definition 25, the density ratio (DR) I-Score.

4.2 Density Ratio I-Score

The goal is to provide a methodological framework to construct an I-Score based on density ratios, the DR I-Score. In addition, we try to circumvent the problem of not observing P^* in a data efficient way, employing *random projections* in the variable space. Considering observations that are projected into a lower-dimensional space allows us to recover more complete cases of the true underlying distribution. As an example, $x = (\texttt{NA}, 1, \texttt{NA}, 2)$ is not complete, however if we project it to the dimensions $A = \{2, 4\}$, $x_A = (1, 2)$ is complete in this lower-dimensional

space. In what follows, we average distributional differences between imputed and complete cases over different projections. In fact, these projections constitute an additional source of randomness over which we integrate to obtain the DR I-score, as we detail now.

Let $\mathcal{A}$ be a subset of the power set $2^{\{1,\dots,d\}}$, that denotes the set of all possible projections, such that each $A \in \mathcal{A}$ describes a set of variables we project onto. Assume the projections are chosen randomly according to a distribution $\mathcal{K}$ with support $\mathcal{A}$ (see Section 5.2 for more details on $\mathcal{K}$). Similarly, we define for each A a distribution P_A^M over the missingness patterns in P_A with support $\mathcal{M}_A$. That is, $m_A \in \mathcal{M}_A$ is a given missingness pattern on the projection with associated probability $\mathbb{P}(M_A = m_A) = P_A^M(m_A)$. For any distribution $H \in \mathcal{H}_P$ we can then consider the conditional distribution $H_A|M_A = m_A$, i.e. the distribution of an imputation H, given the missingness pattern m_A on the projection A. We will abbreviate this distribution with H_{m_A}, so that the density of H_{m_A} is given as $h_{m_A}(x_A) := h_A(x_A|M_A = m_A)$. Denoting with $\mathbf{0}$ the vector of zeros, we similarly write $p_A(x_A|M_A = \mathbf{0})$ for the density of the fully observed points on the projection A.

The following definition specifies the density ratio I-Score as an expected value of the log-ratio of the two densities, where the expectation is also taken over A and M_A.

Definition 26. ***Density Ratio I-Score***
We define the DR I-Score of the imputation distribution H by

$$S^*_{NA}(H,P) = \mathbb{E}_{A\sim\mathcal{K},M_A\sim P_A^M,X_A\sim H_{M_A}}\left[\log\left(\frac{p_A(X_A \mid M_A = \mathbf{0})}{h_{M_A}(X_A)}\right)\right]. \tag{2}$$

If $p_A(X_A \mid M_A = \mathbf{0}) = 0$ for a set of $X_A \sim H_{M_A}$ with nonzero probability, we take $S^*_{NA}(H,P) = -\infty$ as a convention. As an intuition, the DR I-Score given by (2) can be rewritten as

$$S^*_{NA}(H,P) = -\mathbb{E}_{A\sim\mathcal{K},M_A\sim P_A^M} D_{KL}(h_{M_A} \,||\, p_A(\cdot \mid M_A = \mathbf{0})),$$

where the Kullback-Leibler divergence (KL divergence) between two distributions $P, Q \in \mathcal{P}$ on $\mathbb{R}^d$ with densities p, q is defined by

$$D_{KL}(p \,||\, q) := \int p(x)\log\left(\frac{p(x)}{q(x)}\right) d\mu(x).$$

To prove that the DR I-Score is indeed a proper I-Score we need an assumption on the missingness mechanism. This is shown in the following result:

Proposition 27. *Let $H \in \mathcal{H}_P$, as defined in* (1)*. If for all $A \in \mathcal{A}$,*

$$\begin{aligned} &p^*(o^c(x_A,m_A)|o(x_A,m_A), M = m'_A) = p^*(o^c(x_A,m_A)|o(x_A,m_A)),\\ &\text{for all } m'_A, m_A \in \mathcal{M}_A, \end{aligned} \tag{3}$$

then $S^*_{NA}(H,P)$ in (2) *is a proper I-Score.*

The proof is given in in Appendix 4.

Condition (3) is simply the MAR condition on the projection A, see e.g., [151].[2] The key insight is that for any imputation distribution $H \in \mathcal{H}_P$ and $m \in \mathcal{M}$ it holds that

$$h_m(x) = h(o^c(x,m)|o(x,m), M = m)p^*(o(x,m)|M = m), \tag{4}$$

by the definition of $\mathcal{H}_P$ in (1). This can be used to show that the score in (2) factors into (i) an *irreducible part*, stemming from the difference in the observed parts $p^*(o(x,m)|M = m)$ and $p(o(x,m)|M = \mathbf{0})$, and (ii) a score for the distance of the conditional distributions. The latter is minimized for $H = P^*$.

Thus Proposition 27 shows that the proposed I-Score is proper as long as the missingness mechanism is MAR on each projection $A \in \mathcal{A}$. In particular, this holds if

(i) the missingness mechanism is MCAR[3],

(ii) the missingness mechanism is MAR and $\mathcal{A} = \{1, \ldots, p\}$,

(iii) it is known that blocks of data are jointly MAR, and the set of projections $\mathcal{A}$ is chosen such that the blocks are contained as a whole in the projections.

As will be discussed in Section 4, we face a trade-off in practice: The method tends to have more power when allowing for smaller projections while this increases the chance of having a lot of projections violating (3), which may hurt the propriety of the score. Nonetheless, in the empirical validation (Section 7) we do not find evidence that our score violates propriety, even when using random projections without any verification of MAR in the projections.

While our DR I-Score uses a KL-based measurement of the difference between the true distribution and imputation distributions, it would be interesting to look into alternatives such as the multivariate generalization of the integrated quadratic distance of [170], based on the energy score of [59], as suggested by a referee.

4.3 Assessing a Density Ratio through Classification

We assess the density ratio of the proposed DR I-Score (2) by classification, as e.g. in [17] or [24]. Given a projection A and a pattern m_A, we define for x_A in the support of H_{m_A},

$$\pi_{m_A}(x_A) := \frac{p_A(x_A \mid M_A = \mathbf{0})}{p_A(x_A \mid M_A = \mathbf{0}) + h_{m_A}(x_A)}. \tag{5}$$

It then follows that we can rewrite the density ratio in (2) as

$$\frac{p_A(x_A \mid M_A = \mathbf{0})}{h_{m_A}(x_A)} = \frac{\pi_{m_A}(x_A)}{1 - \pi_{m_A}(x_A)}, \tag{6}$$

[2] Lemma 3 in Appendix 4 shows that condition (3) is indeed equivalent to the MAR condition in [151].

[3] Flexible nonparametric methods to test MCAR were developed recently, see e.g., [101, 123].

leaving us with the problem of how to obtain $\pi_{m_A}(x_A)$. Crucially, this term can be assessed through a classifier. To see the link between the definition (5) and classification we introduce further notation. Let $\mathcal{S}^P$ be an i.i.d. sample from P. We denote by $\mathcal{S}^P_A$ the subset of observations in $\mathcal{S}^P$ that is complete on A. Thus, $\mathcal{S}^P_A$ may be seen as an i.i.d. draw from the density $p(\cdot|M_A = \mathbf{0})$. Similarly, we denote by $\mathcal{S}^H_{m_A}$ all observations that originally had missingness pattern m_A, but were imputed with H. We introduce the binary class label Y_{m_A} indicating membership of an observation to the sample $\mathcal{S}^P_A$ by $Y_{m_A} = 1$ and membership to the sample $\mathcal{S}^H_{m_A}$ by $Y_{m_A} = 0$. We then define

$$\begin{aligned} \pi^{Y_{m_A}}(x_A) &:= \mathbb{P}(Y_{m_A} = 1 \mid X_A = x_A), \\ \pi_{m_A} &:= \mathbb{P}(Y_{m_A} = 1), \end{aligned} \tag{7}$$

where π_{m_A} simply denotes the probability of class 1 in a given projection and pattern.

Lemma .8. *Let $\pi_{m_A}(x_A), \pi^{Y_{m_A}}(x_A)$ and π_{m_A} be defined as in* (5) *and* (7) *respectively. If $\pi_{m_A} = 0.5$, then*

$$\pi_{m_A}(x_A) = \pi^{Y_{m_A}}(x_A).$$

The proof is given in Appendix 4 and simply uses Bayes Formula.

Lemma .8 shows that one can access the density ratio (6) through an estimate of the posterior probability $\pi^{Y_{m_A}}(x_A)$. This has a direct connection to classification, as for $\pi_{m_A} = 0.5$, the Bayes classifier with minimal error for this problem is given by $I\{\pi^{Y_{m_A}}(x_A) > 1/2\}$ [40]. In practice, we ensure $\pi_{m_A} = 0.5$ in each projection A and pattern m_A by upsampling the smaller class to the size of the larger class, as detailed in Appendix 2.

5 Practical Aspects

In practice, we take several steps to estimate $S^*_{NA}(H, P)$ based on samples only. For a missingness pattern $m \in \mathcal{M}$, let $\widetilde{\mathcal{A}}_m$ be a set of random projections sampled from $\mathcal{A}$. Note that we sample the set of random projections dependent on the missingness pattern m, as specified in Section 5.2. Let furthermore $\mathcal{N}_{m_A}$ be the set of indices i such that $x_{i,A}$ has missingness pattern m_A whose posterior probability $\pi_{m_A}(x_A)$ in (5) is estimated by $\hat{\pi}_{m_A}(x_A)$. Given an imputation method with $N \geqslant 1$ imputed values $x_i^1, \ldots, x_i^N$ of the incomplete observations, the estimator of $S^*_{NA}(H, P)$ is given by

$$\widehat{S}^*_{NA}(H, P) := \frac{1}{N}\sum_{j=1}^{N} \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{1}{|\widetilde{\mathcal{A}}_m|} \sum_{A \in \widetilde{\mathcal{A}}_m} \frac{1}{|\mathcal{N}_{m_A}|} \sum_{i \in \mathcal{N}_{m_A}} \left[\log\left(\frac{\hat{\pi}_{m_A}(x^j_{i,A})}{1 - \hat{\pi}_{m_A}(x^j_{i,A})} \right) \right], \tag{8}$$

yielding a score of the imputation performance of H, averaged over $N \geqslant 1$ imputations.

For each projection A and pattern m_A, we first split the data into a training and test set. We make sure to have observations with pattern m_A in both the training and test set. We then fit $\hat{\pi}_{m_A}$ on the training set and evaluate it on the test set. We use both halfs of the sample once for training and once for evaluation, to ensure that every observation contributes to the final score (8).

Algorithm 1 in Appendix 2 summarizes the practical estimation of the DR I-Score and gives more details. We now highlight a few more key concepts for the estimator.

5.1 Random Forest Classifier

To estimate $\pi_{m_A}(x^j_{i,A})$ we use a classifier, as detailed in Section 4.3. Our classifier of choice is Random Forest and more specifically, the probability forest of [111]. That is, for each of the `num.proj` projections in $\tilde{\mathcal{A}}_m$, we fit a Random Forest with a small number of trees (say between 5 and 20), a parameter called `num.trees.per.proj`. As such, the overall approach might be seen as one aggregated Random Forest, which restricts the variables in each tree or group of trees to a random subset of variables. This seems natural considering the construction of the RF. In each tree we set `mtry` to the full dimension of the projection to avoid an additional sub-sampling effect. Despite the natural fit of our framework into the Random Forest construction, any other classifier may be chosen to obtain an estimate of π_{m_A}.[4]

5.2 Distribution over Projections and Patterns

The question remains how to choose the set of projections $A \sim \mathcal{K}$ from which we sample the subset $\tilde{\mathcal{A}}$ at random. We group the samples according to their missingness pattern and for each of the groups we sample `num.proj` many projections from $\mathcal{A}$, that we adapt to the given pattern.

Let $O^c_m \subseteq \{1,\dots,d\}$ be the index set of variables with a missing value, such that $o^c(x_i, m) = x_{i,O^c_m}$ and similarly $O_m = \{1,\dots,d\}\backslash O^c_m$ the index set of variables without a missing value. Given a missingness pattern m of a group of samples, we choose $\mathcal{A} = \mathcal{A}_m$ as the set of subsets A that satisfy $A \cap O^c_m \neq \varnothing$ and $A\backslash O^c_m \neq \varnothing$, i.e., a subset of missing indices has to be part of the projection but there must be at least one element in A that is not part of O^c_m. The intuition behind this choice is the following: Since we want to compare a sample of projected imputed observations of a pattern to a sample of projected complete observations, we need to ensure that these samples are not the same. This is the reason for including in each projection at least one index j such that $m_j = 1$, i.e. x_j was missing in the given pattern.

In practice, we select at random a subset $\tilde{\mathcal{A}}_m$ of $\mathcal{A}_m$. To obtain each $A \in \tilde{\mathcal{A}}_m$ we first sample a number r_1 in $\{1,\dots,|O^c_m|\}$ and a number r_2 in $\{1,\dots,d-r_1\}$. Then A is obtained as the union

[4] We also experimented with other classifiers such as generalized linear model (glm), but in practice the resulting score did not have a lot of power discriminating different imputations.

of a random subset of size r_1 from O_m^c and a random subset of size r_2 from O_m.

Having obtained $A \in \tilde{\mathcal{A}}_m$, we choose the support of the projected pattern to be a singleton, $\mathcal{M}_A = \{m_A\}$, with

$$(m_A)_j = \begin{cases} 1, & \text{if } j \in A \cap O_m^c \\ 0, & \text{else.} \end{cases}$$

That is the pattern m_A is simply the pattern m projected to A. We thereby ensure that on the projection A the training set on which the classifier is trained contains the same pattern as the test points.

5.3 Approximate Confidence Intervals

We estimate the variance of our score, if the data would vary, by a jackknife approach as in [153]. We divide $\mathbf{X}$ randomly into two parts and compute the DR I-Score for a given imputation method for each part, obtaining $S^{(1)}$ and $S^{(2)}$. This is repeated B times to obtain scores $S_1^{(1)}, \ldots, S_B^{(1)}$ and $S_1^{(2)}, \ldots, S_B^{(2)}$. Let $\bar{S}_j = 1/2(S_j^{(1)} + S_j^{(2)})$ and let $\widehat{S}$ be the score of the original data set for a given imputation method. We estimate the variance according to the formula of [153], as

$$\widehat{\text{Var}}(\widehat{S}) = \frac{1}{B}\left(\sum_{j=1}^{B}\left(\bar{S}_j - \frac{1}{B}\sum_{j=1}^{B}\bar{S}_j\right)^2\right). \tag{9}$$

The approximate $(1-\alpha)$-Confidence Interval for our score is then given as

$$\hat{S} \pm q_{1-\alpha/2} \cdot \sqrt{\widehat{\text{Var}}(\widehat{S})},$$

where $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal distribution. We choose $\alpha = 0.05$ as default level. As the normality of the score is not guaranteed, a more careful approach would instead try to estimate the quantiles directly in this manner, e.g., using subagging [22]. While this is possible, it is computationally intense as the number of repetitions has to be high to obtain an accurate estimate of the quantiles. In contrast, the variance appears easier to approximate and simulations indicate that the estimate of the variance is reasonable.

6 Further Related Work

There are essentially two strands of literature related to our paper. The first concerns efforts to score imputation methods, as we do in this work. To the best of our knowledge, the research area of scoring and selecting imputation methods has not gained a lot of attention, especially under the more realistic assumption that the true data, or even a simulated ground truth, are not

available for comparison. In [42] the authors state that the performance of an imputation should "preserve the natural relationship between variables in a multivariate data set (...)". The methods they use to assess these properties include comparing densities before and after imputation and in the classification case comparing ROC curves. The authors of [110] consider the scoring of imputation methods with respect to classification tasks. They build a score based on pairwise comparisons of classification performances of two imputation methods using Wilcoxon Tests. Again, this procedure requires the knowledge of the underlying true values.

The second line of literature concerns methods that have very different goals, but are methodologically similar to what we propose: we make use of a classifier able to handle missing values to discriminate between imputed and real data. The key idea is to use projections in the variable space together with a Random Forest (RF) [15] classifier. As such, our method is probably most closely related to the unsupervised Random Forest approach originally proposed in [17] and further developed in [154]. The latter uses an adversarial distribution and a RF classification to achieve a clustering in an unsupervised learning setting. We take the same approach in a different context, whereby our adversarial distribution is the imputation distribution. The classifier approach also has some connections to the famous General Adversarial Network (GAN) [61] and the GAIN imputation method [189] that extends the GAN approach to obtain an imputation method. Aside from the fact that instead of imputation we are concerned about evaluation, there are further differences: First, their discriminator is trained to predict the missingness pattern, while we directly compare imputed and real data. Second, compared to a fully-fledged optimized GAN, our approach based on Random Forest is simple and does not require any backpropagation or tuning. Finally, our approach of obtaining an estimate of the KL divergence as a ratio of estimated class probabilities was introduced in [24], who used it to construct hypothesis tests.

As we use projections in the variable space as a way of adapting a Random Forest to work with missing values, our approach is also connected to the literature of CART and Random Forest algorithms that can handle missing values. We cite and briefly summarize some of the different approaches in the literature. [8] propose an adapted CART algorithm, called Branch Exclusive Splits Trees (BEST), where some predictors are available to split upon only within some regions of the predictor space defined according to other predictors. This structure on the variables needs to exist and be imposed by the researcher. If blind to such a structure, an easy cure is distribution-based imputation (DBI) by [137]. When selecting a predictor and split point, only observations with no missing value in this predictor are considered. An observation with a missing value in the variable of the split is randomized to left and right according to the distribution of the observations that have no missing value in this variable. [18] proposed the approach of so-called surrogate variables. Again, only observations with no missing value are considered when choosing a predictor and split point. After choosing a best (primary) predictor

and split point, a list of surrogate predictors and split points is formed. The first surrogate mimics best the split of the primary split, the second surrogate is second best and so on. This approach makes use of the correlation between the variables. There are also approaches that require fully observed training data, but are able to deal with missing values in prediction, such as [148] or [118], but these are less relevant in our context.

7 Empirical Validation

This section presents an empirical study of the performance of the DR I-Score. We do not aspire to perform an extensive comparison of state-of-the-art imputation methods, but instead employ the different imputations as a tool to validate the DR I-Score. As such, we chose commonly used imputation methods that are easily usable in R. We investigate whether there is evidence against propriety of the DR I-Score and assess the alignment of the ranking induced by the DR I-Score with desired distributional properties. We first list the 9 imputation methods and 15 real world data sets used, covering a range of numbers of observations n as well as numbers of variables d.

7.1 Imputation Methods

We employed the following prevalent single and multiple imputation methods available in R, that can be divided into mice methods ("mice-cart" , "mice-pmm", "mice-midastouch", "mice-rf", "mice-norm.predict") and others ("mipca", "sample", "missForest", "mean").

All methods have in common that they are applicable to the selected data sets without indication of errors or severe warnings. For each method with prefix "mice" we used the R-package `mice` [176]. If a method required specification of parameters, we used the default values. A more detailed description of the methods used can be found in Appendix 1.

7.2 Data Sets

We used the real data sets specified in Table 1 for the assessment of the DR I-Score. They are available in the UCI machine learning repository[5], except for Boston (accessible in R-package `MASS`) and CASchools (accessible in R-package `AER`). We preprocessed the data by cancelling factor variables, in order to be able to run all the assessed imputation methods without errors. However, we kept numerical variables with only few unique values. This preprocessing was done solely for the imputation methods, our proposed score could be used with factor variables as well. Finally, in the data set ecoli we deleted two variables because of multicollinearity issues.

[5] `https://archive.ics.uci.edu/ml/index.php`

data set	n	d
airfoil	1503	6
Boston	506	14
CASchools	420	10
climate.model.crashes	540	19
concrete.compression	1030	9
concrete.slump	103	10
connectionist.bench.vowel	990	10
ecoli	336	5
ionosphere	351	32
iris	150	4
planning.relax	182	12
seeds	210	7
wine	178	13
yacht	308	7
yeast	1484	8

Table 1: Data sets used for performance assessment of the DR I-Score with number of observations n and number of variables d.

7.3 Propriety of DR I-Score

In this section we check empirically whether any imputation is ranked significantly higher than the true underlying data. This is an attempt to empirically assess if the DR I-Score is proper and a lack of significance might also indicate an insufficient sample size to detect a violation. In addition, we can check how well each method performed on each data set with respect to the DR I-Score. To test empirical propriety of the score, that is the non-inferiority of the true data score, we score the fully observed data set by $\widehat{S}^*_{NA}(P^*, P)$ and the imputed data set by $\widehat{S}^*_{NA}(H, P)$ for each imputation distribution H and consider the difference $D_H := \widehat{S}^*_{NA}(H, P) - \widehat{S}^*_{NA}(P^*, P)$. We want to test the following hypotheses for all H:

$$H_0 : D_H = 0 \text{ vs } H_A : D_H > 0. \tag{10}$$

We do this by p-value calculation assuming that approximately

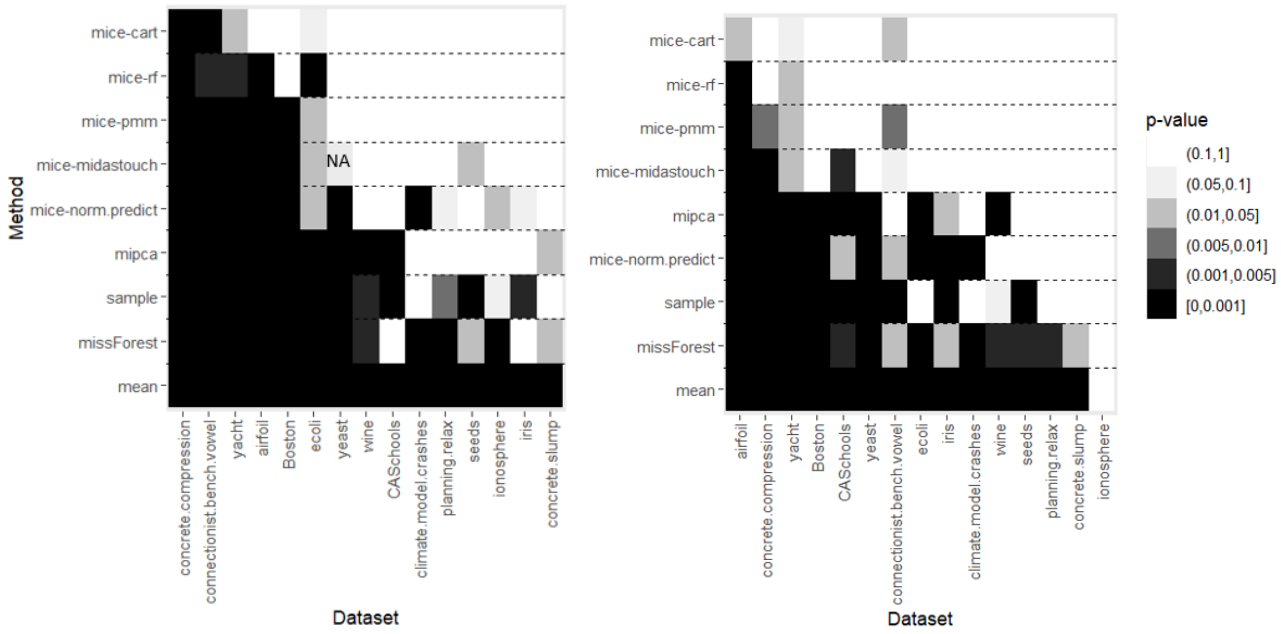
$$D_H \overset{H_0}{\sim} \mathcal{N}(0, \sigma^2(D_H)), \tag{11}$$

where we estimate $\sigma(D_H)$ with the Jackknife variance estimator formula (9) using 30 times $1/2$-subsampling. Details of the whole simulation are given in Appendix 5.

We fixed the overall fraction of missingness to $p_{miss} = 0.2$. When setting data to `NA`, we applied the MCAR and the MAR mechanism: In MCAR, we set each value in the data set to missing with probability p_{miss}. In MAR we created $d/2$ random missingness patterns m for a data set of dimension d. Afterwards, we used the "ampute" function of the package `mice` with these patterns, where all patterns have the same frequency, to create a MAR data set. In the MAR case, p_{miss} might slightly deviate from 0.2.

As parameters for the DR I-Score estimation we chose the number of trees per projection (`num.trees.per.proj`) to be 5 and the minimal node size in each tree to be 10 (the default for a probability RF). We chose the number of projections (`num.proj`) adaptively to the dimension d of the data set: for $d \leqslant 6$ we used 50, for $7 \leqslant d \leqslant 14$ we used 100 and for $d \geqslant 15$ we used 200. We set the number of imputations to $m = 5$ (the default value in the R-package `mice`).

We generated a realization of the `NA`-mask for the MCAR and the MAR case and then computed for each method/data set combination the corresponding p-value of testing (1) under assumption (11). All methods were computable on all data sets, without throwing errors or major warnings, except for mice-midastouch on yeast, indicated with an `NA` in Figure 3. Our findings for testing propriety can be summarized as follows: At level $\alpha = 0.05$ we found no single significant p-value in the MCAR or in the MAR case. At level $\alpha = 0.1$ we found in the MAR case two significant p-values for mice-rf in the data sets yeast and concrete.slump and in the MCAR case one significant p-value for mice-cart in ionosphere. The latter data set has the highest dimension of 32, which for MCAR leads to the extreme case of only observing each



(a) MAR (b) MCAR

Figure 3: Discretized p-values of testing (2) under assumption (11) for the 9 methods applied to the 15 data sets. We used missingness mechanisms MAR (a) and MCAR (b), $p_{miss} = 0.2$ and $m = 5$. The parameter values of the DR I-Score are described in the text. "NA" means that the method was not computable on the respective data set.

missingness pattern once. In summary, these results do not reveal enough evidence against the propriety of our estimated DR I-Score. We point out that, in the MAR case, we did not verify the MAR assumption on the projections.

Reversing the alternative hypothesis we obtain,

$$H_0 : D_H = 0 \text{ vs } H_A : D_H < 0, \tag{12}$$

whose corresponding test may reveal more information about the performance of the methods as well as the difficulty for imputation of each data set. We assume (11) and calculate the corresponding p-values for (2). In Figure 3, we discretized each p-value corresponding to a method applied to a data set into one of 6 batches, which reflect different significance levels. The larger the p-value, the lighter the shade, the better the respective method imputed on the given data set with respect to our score. With a p-value in the batch $(0.1, 1]$ (white) we can not reject the null that $D_H = 0$ even at the level 0.1, i.e., the imputation performed as well as the true data. We sorted the rows and columns to cluster similarly scored methods and data sets together.

Reading the plot row-wise reveals performance information of the methods, the higher up a method appears the better it performed according to the DR I-Score. Interestingly, the MAR and MCAR case reveal almost exactly the same ordering of the methods: Only ranks of mice-norm.predict and mipca are flipped, however these two reveal very similar p-value patterns

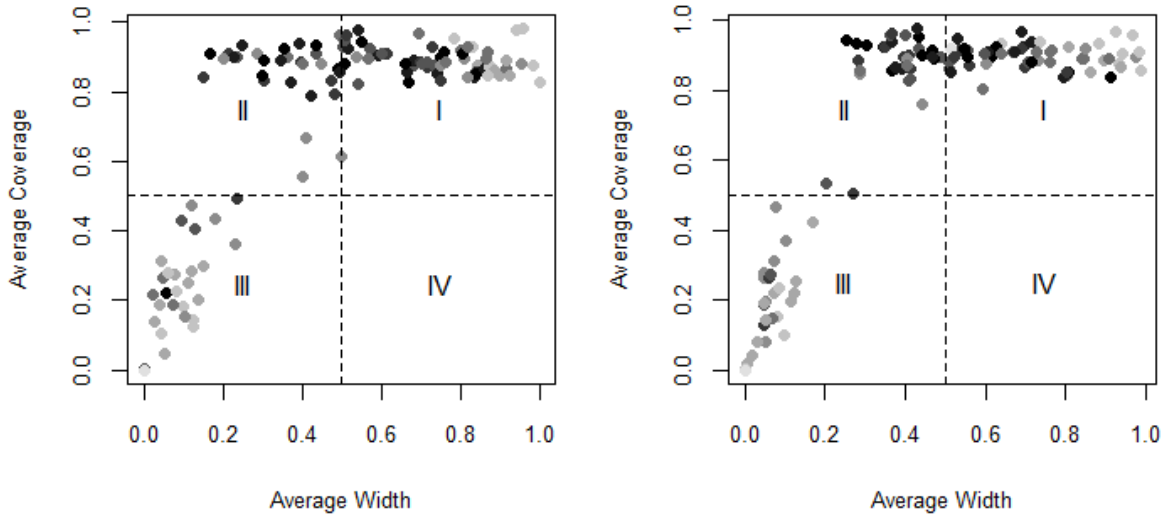
in both plots. If we divide the methods into two groups, we observe in the better group the mice methods (mice-cart, mice-rf, mice-pmm, mice-midastouch). The best method overall in both cases is mice-cart, whose imputations were indistinguishable from the true data in 11 data sets in the MAR case and 12 in the MCAR case, even at level 0.1. In addition, we want to emphasize the suboptimal performance of methods that predict conditional means, without additional randomization to impute values, in particular missForest and mean. This may be surprising as missForest is known to perform very well in the literature, see e.g. [182]. However this impression of good performance is based on measures of accurary, such as RMSE. As laid out in [175], as a prediction method, missForest does not account for the uncertainty caused by the missing data. Contrary to accuracy measures, our score takes the joint distribution into account when assessing performance, hence the comparatively weak performance of missForest over the chosen data sets. The worst method appears to be the mean, a method known to heavily distort the distribution.

Reading the plot column-wise reveals information about how easy or hard the data sets could be imputed by the methods considered. The further to the right a data set appears, the easier it was to impute. For instance, we find that none of the considered methods was able to find an imputation that recovers the joint distribution well enough for the data set concrete.compression (MAR) and airfoil (MCAR).

7.4 Relevancy of DR I-Score

In the last section, we did not discover evidence against the propriety of the estimated DR I-Score. However, a practitioner might want to select the best method with the score, or at least determine the worst performing methods to definitely not employ them. Unfortunately, there is no ground truth to compare to. However, one would hope that the methods chosen by our score perform well on a wide range of targets, even though it was not designed to select for any of these targets specifically. We focus on a target that presumably will be of interest in practice when doing multiple imputation: Average coverage and average width of marginal confidence intervals for each `NA` value, obtained by the m multiple imputations. More specifically, let $\mathcal{N} \subseteq \{1,\dots,n\}$ be the set of i for which $\sum_{j=1}^{p} \mathbf{M}_{ij} > 0$, i.e., those observations that contain at least one missing value. For each observation x_i, let $\mathcal{N}_i \subset \{1,\dots,p\}$ be the missing coordinates of x_i. Given an imputation method H, we obtain for each x_{ij}, $i \in \mathcal{N}$, $j \in \mathcal{N}_i$, m imputed values $x_{ij}^1(H),\dots,x_{ij}^m(H)$. For m large enough, we compute the empirical 0.025- and 0.975-quantiles $x_{ij}^{0.025}(H)$ and $x_{ij}^{0.975}(H)$ and consider the interval spanned in between as the empirical 95%-CI for x_{ij} defined by the method H through $x_{ij}^1(H),\dots,x_{ij}^m(H)$. We denote it by

$$\mathrm{CI}(x_{ij}^1(H),\dots,x_{ij}^m(H)) := [x_{ij}^{0.025}(H), x_{ij}^{0.975}(H)].$$



(a) MAR (b) MCAR

Figure 4: Average coverage plotted against average width for the 9 methods applied to the 15 data sets (total = $9 \times 15 = 135$ points). The darkness indicates the rank induced by the DR I-Score (the darker, the higher the rank). We used the missingness mechanism MAR in (a) and MCAR in (b) with $p_{miss} = 0.2$, $m = 20$ and the in the text described parameter values to compute the DR I-Scores.

We then check whether the true missing data point x_{ij} lies within the CI or not and obtain an average marginal coverage for the method H by averaging over all $i \in \mathcal{N}$, i.e.,

$$\text{Coverage}(H) := \frac{1}{\mathcal{N}} \sum_{i \in \mathcal{N}} \frac{1}{\mathcal{N}_i} \sum_{j \in \mathcal{N}_i} I\{x_{ij} \in \text{CI}(x_{ij}^1(H), \ldots, x_{ij}^m(H))\}.$$

A method that has a large enough variation between the m different imputations will reach a coverage of 1, however the average width of its corresponding CI is large, indicating very little power. We use the average width of the confidence intervals as an indicator of statistical efficiency. A good imputation method will have small average width while still maintaining high average coverage. We define the average marginal width for the method H by

$$\text{Width}(H) := \frac{1}{\mathcal{N}} \sum_{i \in \mathcal{N}} \frac{1}{\mathcal{N}_i} \sum_{j \in \mathcal{N}_i} \left(x_{ij}^{0.975}(H) - x_{ij}^{0.025}(H) \right).$$

For better visualization we constrain the $\text{Width}(H)$ for all methods H to the interval $[0, 1]$: Given a data set, we normalize $\text{Width}(H)$ for all methods H by the maximal width of all methods.

In Figure 4 we plot $\text{Coverage}(H)$ against the normalized $\text{Width}(H)$ for all methods applied to all data sets, leading to totally $15 \times 9 = 135$ points. Not all of them are visible since they can lie on top of each other. For example the method mean always produces the point $(0, 0)$ for all 15 data sets, since there is no variation in the m imputed data sets. The shade of the

method/quadrant	I	II	III
cart	0.47	0.53	0
pmm	0.53	0.47	0
midastouch	0.67	0.33	0
rf	0.73	0.27	0
mipca	0.87	0.13	0
sample	0.87	0.13	0
norm.predict	0	0.13	0.87
mean	0	0	1
missForest	0	0	1

(a) MAR

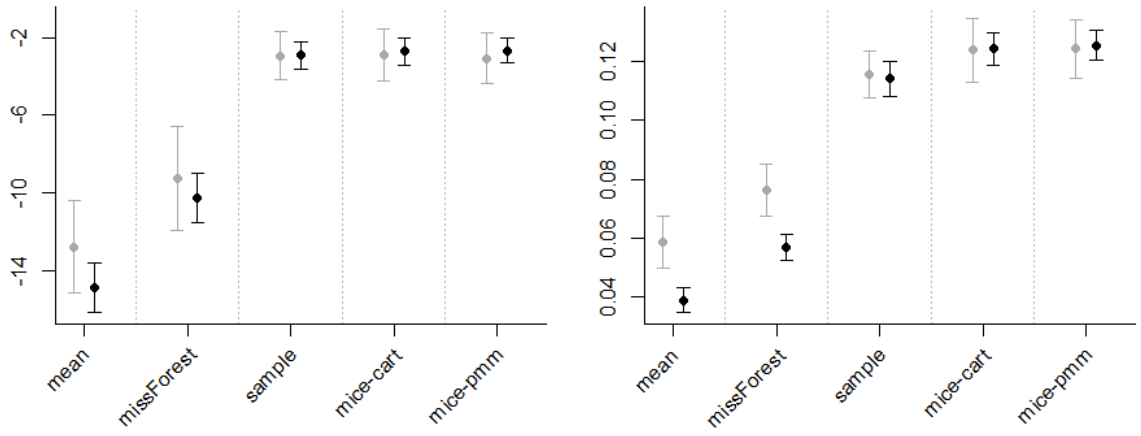
method/quadrant	I	II	III
pmm	0.40	0.60	0
cart	0.47	0.53	0
midastouch	0.53	0.47	0
rf	0.60	0.40	0
mipca	0.87	0.13	0
norm.predict	0	0.13	0.87
sample	0.93	0.07	0
mean	0	0	1
missForest	0	0	1

(b) MCAR

Table 2: The fraction of times each method appeared in the quadrants I, II and III of Figure 4 in the MAR case (a) and the MCAR case (b).

points reflects the induced rank by the DR I-Score, assigning one of the 9 ranks to each method in a given data set: rank 1 to the best scored method (black) up to rank 9 to the worst scored method (light gray). We set the number of imputations to $m = 20$, used missingness mechanism MCAR and MAR and $p_{\text{miss}} = 0.2$ and the same parameter setting as in Section 7.3. For easier interpretation, we name the four quadrants of the square in Figure 4 by the letters I-IV. We observe that quadrant IV, corresponding to high average width and low average coverage, does not contain any points, which makes sense. Considering the shade gradient, the DR I-Score seems to indeed often rank points with high average coverage and low average width the best. Clearly, methods that produce small average width combined with small average coverage as well as high average width combined with high average coverage, are ranked low, which is highly desirable.

Table 2 shows a clearer picture of the roles of the methods in these results. We indicate for each method the fraction of times in the data sets it appeared in the quadrants I, II and III in Figure 4. We observe that all points making up quadrant III are from the methods mice-norm.predict, mean and missForest, i.e., the methods that use a sort of mean imputation. From a distributional point of view, these methods are often not favourable, which is a property nicely captured by the DR I-Score.



(a) Estimated DR I-Score with CIs.

(b) OOB error with CIs.

Figure 6: Births data: Estimation of the proposed DR I-Score (a) and the OOB error of the classification RF (b) with corresponding CIs for the methods mean, missForest, sample, mice-cart and mice-pmm with sample size $n = 500$ (grey) and $n = 2000$ (black). We obtained the CIs by subsampling as described in Section 5.3.

7.5 Real Data Example: Births Data

We further illustrate our method on the natality data of 2020 obtained from the Centers for Disease Control and Prevention (CDC) website.[6] This data set contains information on ≈ 3.5 million births in 2020 in the US. We subsample the data as detailed later and consider 23 variables, listed in Appendix 2. The variables include categorical variables, such as race and education of mother and father and the gender of the newborn, as well as continuous variables, such as the age of the parents and the baby's birth weight. We consider a subset of imputation methods that can deal with these mixed data: mice-pmm, mice-cart, sample and missForest. We also include the mean as a baseline comparison.

We subsample the data to obtain two smaller subsets, one of size 500 and one of size 2000, where we sample at random observations with at least one missing value. As such, we obtain in both samples an overall probability of missingness of $p_{miss} \approx 0.08$, which is lower than in the previous sections. Missing values are encountered predominantly in the variables weight of mother, weight gain of mother and BMI of mother, which are correlated, and variables about the father, such as the age, race or education. We estimate the DR I-Score for the aforementioned imputation methods applied to the two data sets (sizes 500, 2000) with missing values.

Since the births data set also contains a lot of complete cases, it is possible to validate our score in a special way: We compare each of the imputations with a disjoint sample of only complete cases of the same size that we obtained by randomly sampling complete cases from the whole data set. The comparison is assessed via a classification Random Forest, distinguishing

[6] `https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm`

between the sample of complete cases and the imputation, where we report the out-of-bag prediction error (OOB error). We note that in this case, the smaller the OOB error, the easier it was for the classifier to distinguish the imputation from the true data. Hence, the smaller the OOB error, the worse the imputation method. For both the score and the OOB error we also compute 95%-CI with the Jackknife variance estimator formula (9) using 30 times $1/2$-subsampling. We plotted the results in Figure 6 to compare the ranking of methods obtained by our score with the one obtained by the RF. We ordered the methods according to the mean score/OOB error, computed on the sample of size 2000.

First, we observe that the rankings obtained from the DR I-Score and the OOB error of the RF are the same. Second, the ranking is in line with the one we obtained in Section 7.3: mean and missForest appear to be the weakest methods followed by sample, while mice-cart and mice-pmm are ranked best. Third, by increasing the sample size from 500 (gray) to 2000 (black), the CIs get more narrow and the methods can be held apart more significantly for both the score and the RF. We observe that the score cannot distinguish sample from the two best methods as clearly as the RF, which would be desirable. However, the RF is allowed to use an additional (large) complete sample, which is naturally an unfair advantage. Moreover, this method is generally not applicable without this large amount of fully observed points. In contrast, the DR I-Score works with only incomplete observations, and still manages to reproduce the same ordering as the RF.

8 Discussion

In this paper we presented the convenient framework of Imputation Scores to score a given set of imputation methods for an incomplete data set. The widespread assessment of imputation methods via RMSE (of masked observations) favors methods that impute conditional means but do not necessarily reflect the whole conditional distributions. Given the assumption of MAR on each projection, our proposed density ratio I-Score is able to give high scores to methods that replicate the data distribution well.

1 Details of Imputation Methods

1) **missForest** is a multiple imputation method based on iterative use of RF, allowing for continuous and categorical data [167]. After an initial mean-imputation, the variables are sorted according to their amount of missing values, starting with the lowest. For each variable as response, a RF is fitted based on the observed values. The missing values are then predicted with the RF. The imputation procedure is repeated until a stopping criterion is met. We used the R-package `missForest` [166].

2) **mipca** is a multiple imputation method with a PCA model [84]. After an initialization step, an EM algorithm with parametric bootstrap is applied to iteratively update the PCA-parameter estimates and draw imputations from the predictive distribution. The algorithm is implemented in the function `MIPCA` of the R-package `missMDA` [83]. We use the function `estim_ncpPCA` to estimate the number of dimensions for the principal component analysis by cross-validation.

3) **mean** is the simplest single imputation method considered. It imputes with the mean of the observed cases for numerical predictors and the mode of observed cases for categorical predictors. We use the implementation of the R-package `mice`.

4) **sample** is a multiple imputation method sampling at random a value of the observed observations in each variable to impute missing values. We use the implementation of the R-package `mice`.

5) **mice-cart** is a multiple imputation method cycling through the following steps multiple times [45]: After an initial imputation through sampling of the observed values, a classification or regression tree is fitted. For each observation with missing values, the terminal node they end up according to the fitted tree is determined. A random member in this node is selected of which the observed value leads the imputation.

6) **mice-norm.predict** is a multiple imputation method cycling through the same steps as mice-cart with the adaptation that a linear regression is fitted and its predicted value is used as imputation.

7) **mice-pmm** Predictive Mean-Matching is a semi-parametric imputation approach ([106] and [147]). Based on the complete data, a linear regression model is estimated, followed by a parameter update step. Each missing value is filled with the observed value of a donor that is randomly selected among complete observations being close in predicted values to the predicted value of the case containing the missing value.

8) **mice-midastouch** is a multiple imputation method using an adaption of classical predictive mean-matching, where candidate donors have different probabilities to be drawn [55]. The probability depends on the distance between the donor and the incomplete observation. A closeness parameter is specified adaptively to the data.

9) **mice-rf** is a multiple imputation method cycling through the same steps as mice-cart with the adaptations that one tree is fitted for every bootstrap sample. For each observation with missing values, the terminal nodes in each tree are determined. A random member of the union of the terminal nodes is selected of which the observed value leads the imputation.

2 Details on Birth Data

Even though the original data contains a lot of variables, we took only the following variables from the source data:

- mother's age, height, weight before the pregnancy, weight gain during pregnancy and BMI before pregnancy
- mother's race (black, white, asian, NHOPI, AIAN or mixed), marital status (married or unmarried) and the level of education (in total 8 levels)
- father's age, race and level of education
- month of birth
- plurality of the birth (how many babies were born at once)
- whether and when the prenatal care started
- pregnancy duration
- delivery method (vaginal or C-section)
- birth order - the total number of babies born by the same mother (including the current one)
- birth interval - number of months passed since last birth (NA if this is the first child)
- number of cigarettes smoked per day on average before and during the pregnancy
- birth weight (in grams) and gender of the baby
- indicators whether baby had any abnormal condition.

3 Algorithm and Implementation Details of DR I-Score Estimation

Here we present additional details of the implementation of the DR I-Score and the full algorithm in pseudo-code.

Projection Distribution. In Section 5.2 we describe the distribution $\mathcal{K}$ over the projections with restricted support used for the empirical estimation of the DR I-Score. The choice of this distribution is up to the experimenter and can be adapted to the specific patterns of missinginess in a given data set. In particular, the projections can be chosen such that each projection satisfies the MAR assumption, if this can be determined with domain knowledge.

Ensuring Class Balancing. We follow a simple procedure to ensure the same number of observations in the training sets $\mathcal{S}_A^P$ and $\mathcal{S}_{m_A}^H$. First if $\mathcal{S}_{m_A}^H$ has fewer elements than $\mathcal{S}_A^P$, but is "large enough" relative to $\mathcal{S}_A^P$, we simply upsample $\mathcal{S}_{m_A}^H$ with replacement until it contains the same number of elements as $\mathcal{S}_A^P$. The exact same procedure is applied if $\mathcal{S}_A^P$ has fewer elements than $\mathcal{S}_{m_A}^H$. On the other hand, if the set $\mathcal{S}_{m_A}^H$ is smaller than $\mathcal{S}_A^P$, that is if $|\mathcal{S}_{m_A}^H| < \tau|\mathcal{S}_A^P|$ for some $\tau \in (0,1)$, we sample with replacement observations from other patterns and add them to $\mathcal{S}_{m_A}^H$. We found that $\tau = 0.75$ works well empirically. This is done to ensure that we do not upsample one or two observations. In practice it seems adding additional patterns in the training step of the classifier does not hurt propriety.

Numerical Truncation. To avoid numerical issues when calculating the log of the density ratio with Expression (4.7), we apply for each A, m_A and x_A the following truncation function to $\hat{\pi}_{m_A}(x_A)$

$$p(x) = \min(\max(x, 10^{-9}), 1 - 10^{-9}).$$

Thus, we slightly adapt the predicted probabilities $\hat{\pi}_{m_A}(x_A)$ that are close to 0 or 1, such that the log of (4.7) can be computed.

Patterns of Size 1. In our algorithm to estimate the DR I-Score, we split the observations of a given missingness pattern m into a training and a test set. We then use the training set to estimate the RF and use it to predict on the test set. Predominantly in the MCAR case, but also potentially in MAR situations, it often happens that several patterns contain only one observation. In this case, we group these patterns of size 1 together, use all of these samples for training and testing and only fit one RF based on several projections for this new group. This is mainly done due to computational reasons.

Algorithm 1: Algorithm for estimation of DR I-Score

Inputs: data set X containing missing values to impute, a multiple imputation method applied to X yielding N imputed data sets $\hat{X}_i$, i=1,..., N;

Result: DR I-Score s for the imputation method as the average of the N scores $\{s_i\}_{i=1}^N$ for each of the N imputed data sets

Hyper-parameters: number of projections `num.proj`, number of trees per projection `num.trees.per.proj`, standard parameters of the Probability Forests;

- Group observations in X into J different groups according to their unique missing value patterns M_j, $j = 1, \ldots, J$.;

for $i = 1, \ldots, N$ **do**

  for $m = M_1, \ldots, M_J$ **do**

    - Sample a set of `num.proj` projections, A_k, $k = 1, \ldots,$ `num.proj` compatible with the missingness pattern m as described in Section 5.2;

    - Get the projected imputed data $\hat{X}_i$ with pattern m, and split them in two halves $\hat{X}_i^0$ and $\hat{X}_i^1$;

    for $l = 0, 1$ **do**

      for $k = 1, \ldots,$ *num.proj* **do**

        - Get the complete observations $X_{A_k}^{comp}$ from the projected data X_{A_k};

        - Get the projected imputed data $\hat{X}_{i,A_k}^l$;

        - Fit a Probability Forest with `num.trees.per.proj` trees and `mtry` full, discriminating $X_{A_k}^{comp}$ from $\hat{X}_{i,A_k}^l$ (ensuring a balance of classes, see above for details);

      end

      - Form one Probability Random Forest based on the `num.proj` many random forests;

      - Get an estimate of the density ratio, $\frac{\hat{\pi}_{m_A}}{1-\hat{\pi}_{m_A}}$, through this Probability Forest as in Equation (4.7);

      - Compute the individual score contributions, as $\log \frac{\hat{\pi}_{m_A}(x)}{1-\hat{\pi}_{m_A}(x)}$ of the left-out imputations $x \in \hat{X}_{i,A_k}^{1-l}$;

    end

    - Average for $l = 0, 1$ the individual score contributions of all points, leading to score $s_{i,m}$;

  end

  - Average the scores $s_{i,m}$ over all patterns m to get score s_i;

end

- Average the score s_i over all imputations i to get the final score s.

4 Proofs

Proposition 28 (Restatement of Proposition 4.1)**.** *Let $H \in \mathcal{H}_P$, as defined in (4.1). If for all $A \in \mathcal{A}$,*

$$p^*(o^c(x_A, m_A)|o(x_A, m_A), M = m'_A) = p^*(o^c(x_A, m_A)|o(x_A, m_A)),$$
$$\text{for all } m'_A, m_A \in \mathcal{M}_A, \tag{4.4}$$

then $S^*_{NA}(H, P)$ in (4.3) is a proper I-Score.

Proof Since,

$$S^*_{NA}(H, P) = -\mathbb{E}_{A\sim\mathcal{K}, M_A\sim P^M_A} D_{KL}(h_{M_A} \,||\, p_A(\cdot \mid M_A = \mathbf{0})),$$

it is enough to show that for all $A \in \mathcal{A}$, and all $m_A \in \mathcal{M}_A$,

$$D_{KL}(h_{m_A} \,||\, p_A(\cdot|M_A = \mathbf{0})) \geqslant D_{KL}(p^*_{m_A} \,||\, p_A(\cdot|M_A = \mathbf{0})).$$

We thus drop the subscript A for simplicity.

It then holds that for all $m \in \mathcal{M}$,

$$h_m(x) = h(x|M = m) = h(o^c(x, m)|o(x, m), M = m)p^*(o(x, m)|M = m),$$

by the definition of $\mathcal{H}_P$. Similarly,

$$\begin{aligned} p(x|M = \mathbf{0}) &= p^*(o^c(x, m)|o(x, m), M = \mathbf{0})p^*(o(x, m)|M = \mathbf{0}) \\ &= p^*(o^c(x, m)|o(x, m))p^*(o(x, m)|M = \mathbf{0}), \end{aligned}$$

where we used assumption (4.4) in the last step. Crucially, we can then decompose the KL divergence:

$$\begin{aligned} &D_{KL}(h_m \,||\, p(\cdot|M = \mathbf{0})) \\ &= \int \log\left(\frac{h(o^c(x, m)|o(x, m), M = m)p^*(o(x, m)|M = m)}{p^*(o^c(x, m)|o(x, m))p^*(o(x, m)|M = \mathbf{0})}\right) h(x \mid M = m)d\mu(x) \\ &= \int \log\left(\frac{h(o^c(x, m)|o(x, m), M = m)}{p^*(o^c(x, m)|o(x, m))}\right) h(o^c(x, m)|o(x, m), M = m)p^*(o(x, m)|M = m)d\mu(x) \\ &+ \int \log\left(\frac{p^*(o(x, m)|M = m)}{p^*(o(x, m)|M = \mathbf{0})}\right) h(o^c(x, m)|o(x, m), M = m)p^*(o(x, m)|M = m)d\mu(x) \\ &= \mathbb{E}_{o(x,m)\sim p^*_m}\left[\int \log\left(\frac{h(o^c(x, m)|o(x, m), M = m)}{p^*(o^c(x, m)|o(x, m))}\right) h(o^c(x, m)|o(x, m), M = m)d\mu(o^c(x, m))\right] \\ &+ \int \log\left(\frac{p^*(o(x, m)|M = m)}{p^*(o(x, m)|M = \mathbf{0})}\right) p^*(o(x, m)|M = m)d\mu(o(x, m)). \end{aligned}$$

The second summand in the last term is simply the KL divergence between $p^*(o(x,m)|M=\mathbf{0})$ and $p^*(o(x,m)|M=m)$ and represents the irreducible part, as it cannot be changed by any imputation. The first summand is bounded below by zero, and attains zero for $h=p^*$, proving the claim. ■

[151] point out the problematic definition of MAR throughout the literature and define MAR as:

$$\begin{aligned}&\mathbb{P}(M_A=m_A|X_A=x_A)=\mathbb{P}(M_A=m_A|X_A=\tilde{x}_A)\\&\text{for all } x_A,\tilde{x}_A \text{ s.t. } o(x_A,m_A)=o(\tilde{x}_A,m_A).\end{aligned} \tag{S2.1}$$

In the following we ensure that (4.4) and (S2.1) really mean the same:

Lemma .9. *Condition (4.4) and* (S2.1) *are equivalent.*

Proof Throughout we drop the subscript A for simplicity. We start by reformulating (S2.1), for any $x,\tilde{x}$ such that $o(x,m)=o(\tilde{x},m)$,

$$\begin{aligned}&\mathbb{P}(M=m|X=x)=\mathbb{P}(M=m|X=\tilde{x}) \Leftrightarrow\\&\frac{p^*(x|M=m)\mathbb{P}(M=m)}{p^*(x)}=\frac{p^*(\tilde{x}|M=m)\mathbb{P}(M=m)}{p^*(\tilde{x})} \Leftrightarrow\\&\frac{p^*(o(x,m),o^c(x,m)\mid M=m)}{p^*(o(\tilde{x},m),o^c(\tilde{x},m)\mid M=m)}=\frac{p^*(o(x,m),o^c(x,m))}{p^*(o(\tilde{x},m),o^c(\tilde{x},m))} \Leftrightarrow\\&\frac{p^*(o^c(x,m)\mid o(x,m),M=m)}{p^*(o^c(x,m)\mid o(x,m))}=\frac{p^*(o^c(\tilde{x},m)\mid o(x,m),M=m)}{p^*(o^c(\tilde{x},m)\mid o(x,m))} \Leftrightarrow\\&p^*(o^c(x,m)\mid o(x,m),M=m)=\frac{p^*(o^c(\tilde{x},m)\mid o(x,m),M=m)}{p^*(o^c(\tilde{x},m)\mid o(x,m))}p^*(o^c(x,m)\mid o(x,m))\end{aligned} \tag{S2.2}$$

Clearly, (4.4) implies (S2.2). Integrating (S2.2) with respect to the missing part of x, $o^c(x,m)$, only shows that

$$\frac{p^*(o^c(\tilde{x},m)\mid o(x,m),M=m)}{p^*(o^c(\tilde{x},m)\mid o(x,m))}=1,$$

and thus also (4.4). ■

Lemma .10 (Restatement of Lemma 4.1)**.** *Let $\pi_{m_A}(x_A),\pi^{Y_{m_A}}(x_A)$ and π_{m_A} be defined as in (4.6) and (4.8) respectively. If $\pi_{m_A}=0.5$, it holds that*

$$\pi_{m_A}(x_A)=\pi^{Y_{m_A}}(x_A).$$

Proof Given the definition of the class label Y_{m_A}, we can rewrite $p_A(x_A \mid M_A = \mathbf{0})$ and $h_{m_A}(x_A)$ by

$$p_A(x_A \mid M_A = \mathbf{0}) = f(x_A \mid Y_{m_A} = 1),$$
$$h_{m_A}(x_A) = f(x_A \mid Y_{m_A} = 0).$$

By Bayes Formula it now follows

$$\begin{aligned}\pi^{Y_{m_A}}(x_A) &:= \mathbb{P}(Y_{m_A} = 1 \mid X_A = x_A) \\ &= \frac{f(x_A \mid Y_{m_A} = 1)\mathbb{P}(Y_{m_A} = 1)}{f(x_A \mid Y_{m_A} = 1)\mathbb{P}(Y_{m_A} = 1) + f(x_A \mid Y_{m_A} = 0)\mathbb{P}(Y_{m_A} = 0)} \\ &= \frac{p_A(x_A \mid M_A = \mathbf{0})\pi_{m_A}}{p_A(x_A \mid M_A = \mathbf{0})\pi_{m_A} + h_{m_A}(x_A)\pi_{m_A}},\end{aligned}$$

such that if $\pi_{m_A} = 0.5$, we obtain that $\pi^{Y_{m_A}}(x_A) = \pi_{m_A}(x_A)$ for any observation x. ■

5 Details on Propriety Assessment

We used the following pipeline to obtain the results of Section 7.3:

1. For the two missingness mechanisms MAR and MCAR we consider two overall probabilities of missingness ($p_{\text{miss}} = 0.1$ and 0.2). For each of the p_{miss} we do:
2. For each fully observed data set `X` we do:
   (a) We mask `X` respecting the missingness mechanism as well as p_{miss} and obtain a fixed `X.NA`.
   (b) We apply 30 times $1/2$-subsampling on `X` and `X.NA` and obtain 30 subsampled $\texttt{X}^S$ and $\texttt{X.NA}^S$.
   (c) We score `X` with the DR I-Score.
   (d) For each method in methods we do:
      i. We impute `X.NA` $m = 5$ times.
      ii. We score each of the 5 imputed versions of `X.NA` with the DR I-Score and get the final score by averaging.
      iii. We compute the difference D of the DR I-Score of the imputation of `X.NA` and the DR I-Score of `X`.
      iv. For each of the $S = 1, \ldots, 30$ we do:
         A. We apply steps (c) and i. and ii. to $\texttt{X}^S$ and $\texttt{X.NA}^S$.

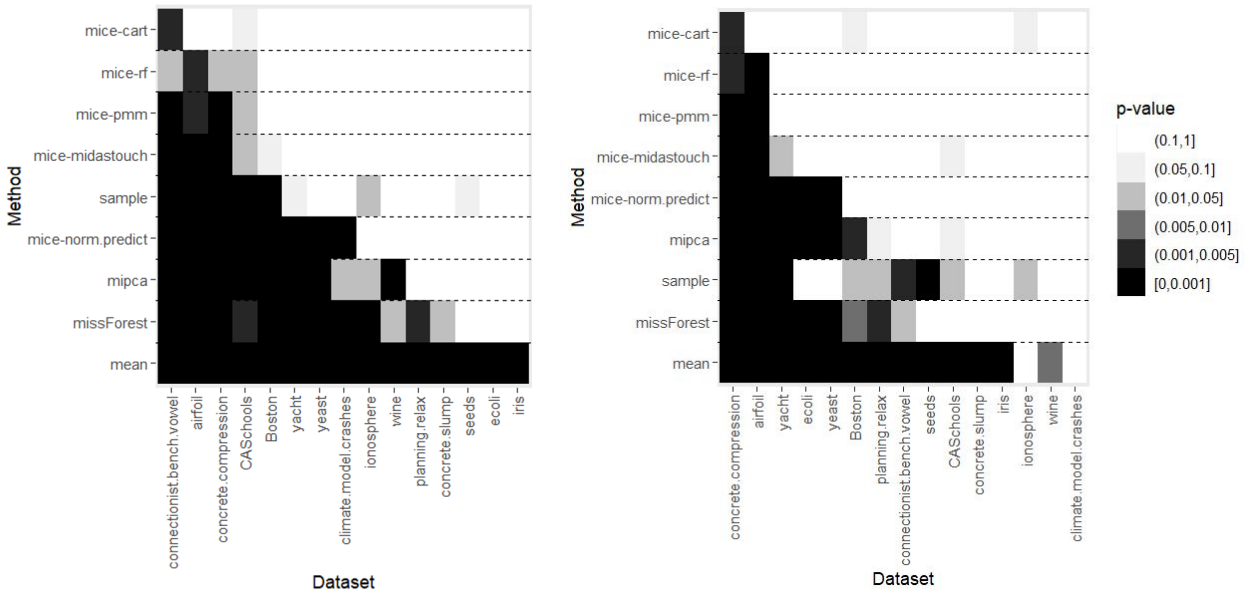
B. We compute the difference D^S of the DR I-Score of the imputation of $\texttt{X.NA}^S$ and the DR I-Score of $\texttt{X}^S$.

v. We estimate the variance of D, $\sigma^2(D)$, with the Jackknife variance estimator formula (5.2) based on D^S.

(e) We compute a p-value by $P_{H_0}(X > D)$, where $X \overset{H_0}{\sim} \mathcal{N}(0, \sigma^2(D))$

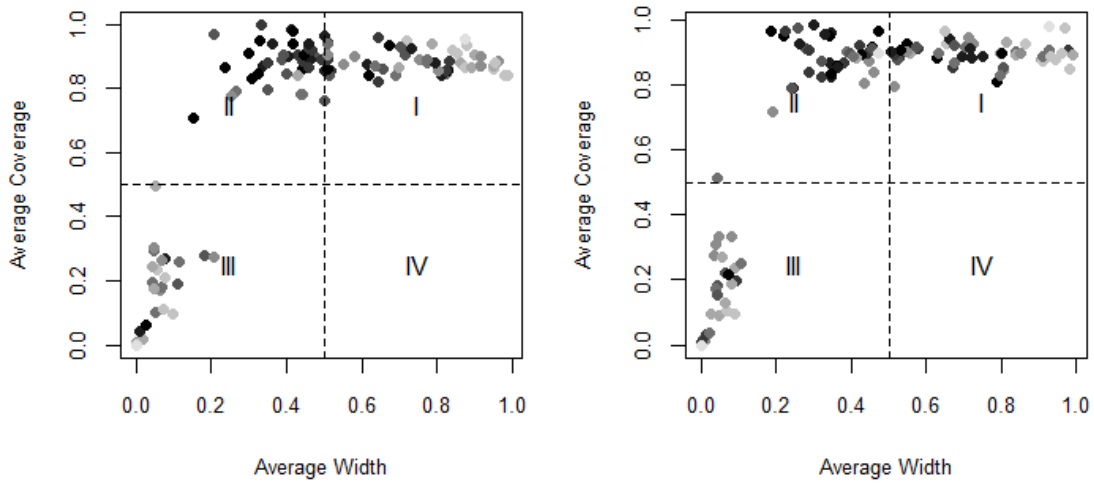
6 Empirical Results for $p_{\text{miss}} = 0.1$

In this section we present additional results for $p_{\text{miss}} = 0.1$.



(a) MAR (b) MCAR

Figure 7: Discretized p-values of testing (7.3) under assumption (7.2) for the 9 methods applied to the 15 data sets. We used missingness mechanisms MAR (a) and MCAR (b), $p_{miss} = 0.1$ and $m = 5$. The parameter values of the DR I-Score are described in the main text.



(a) MAR (b) MCAR

Figure 8: Average coverage plotted against average width for the 9 methods applied to the 15 data sets (total = $9 \times 15 = 135$ points). The darkness indicates the rank induced by the DR I-Score (the darker, the higher the rank). We used the missingness mechanism MAR in (a) and MCAR in (b) with $p_{miss} = 0.1$, $m = 20$ and the in the text described parameter values to compute the DR I-Scores.

method/quadrant	I	II	III
cart	0.20	0.80	0
pmm	0.40	0.60	0
midastouch	0.40	0.60	0
rf	0.67	0.33	0
mipca	0.87	0.13	0
sample	1.00	0	0
norm.predict	0	0	1
mean	0	0	1
missForest	0	0	1

(a) MAR

method/quadrant	I	II	III
pmm	0.33	0.67	0
cart	0.33	0.67	0
midastouch	0.40	0.60	0
rf	0.53	0.47	0
mipca	0.87	0.13	0
sample	0.93	0.07	0
missForest	0	0.07	0.93
norm.predict	0	0	1
mean	0	0	1

(b) MCAR

Table 3: The fraction of times each method appeared in the quadrants I, II and III of Figure 8 in the MAR case (a) and the MCAR case (b).

Paper

C

PKLM: A flexible MCAR Test Using Classification.

J. Näf, M. Spohn, L. Michel, N. Meinshausen

Major Revision in Psychometrika.

PKLM: A flexible MCAR test using Classification

Jeffrey Näf Meta-Lina Spohn
Loris Michel Nicolai Meinshausen

January 20, 2023

Abstract

We develop a fully non-parametric, easy-to-use, and powerful test for the missing completely at random (MCAR) assumption on the missingness mechanism of a dataset. The test compares distributions of different missing patterns on random projections in the variable space of the data. The distributional differences are measured with the Kullback-Leibler Divergence, using probability Random Forests [111]. We thus refer to it as "Projected Kullback-Leibler MCAR" (PKLM) test. The use of random projections makes it applicable even if very few or no fully observed observations are available or if the number of dimensions is large. An efficient permutation approach guarantees the level for any finite sample size, resolving a major shortcoming of most other available tests. Moreover, the test can be used on both discrete and continuous data. We show empirically on a range of simulated data distributions and real datasets that our test has consistently high power and is able to avoid inflated type-I errors. Finally, we provide an R-package `PKLMtest` with an implementation of our test.

Keywords. Ranking, Random Projections, Tree Ensembles, Random Forest, KL-Divergence.

1 Introduction

Dealing with missing values is an integral part of modern statistical analysis. In particular, the assumed mechanism leading to the missing values is of great importance. Based on the work of [146], there are three groups of missingness mechanisms usually considered: The values may be missing completely at random (MCAR), meaning the probability of a value being missing does not depend on the observed or unobserved data. In contrast, the probability of being missing

could depend on observed values (missing at random, MAR) or on unobserved values (missing not at random, MNAR).

As stated in [190], "a formal confirmation of the MCAR missing data mechanism is of great interest, simply because essentially all methods can still yield consistent estimates under MCAR even if the underlying population distribution is unknown". While there is, at least for imputation, a number of approaches that can deal with MAR missing data such as Multivariate Imputation by Chained Equations (mice) [23, 38], many commonly used methods still explicitly rely on the validity of the MCAR assumption. Examples are the easy-to-use listwise-deletion and mean-imputation methods [108]. As such, MCAR testing is still widely employed in the analysis of (psychometric) data; see e.g., [21, 67, 25, 141, 168, 37] for some recent examples.

The testing framework is of an ANOVA-type: when observing a dataset with missing values, there are n observations and G missingness patterns, $g = 1, \ldots, G$. The observations belonging to the missingness pattern g can be seen as a group, such that we observe G groups of observations. The MCAR hypothesis now implies that the distribution of the observed data in all groups is the same, while under the alternative at least two differ. This is technically testing the observed at random (OAR) assumption defined in [144], see also the end of Section 3 for a discussion. This distinction can be avoided by assuming the missingness mechanism is MAR, which is what is usually implicitly done [101].

The idea of testing the MCAR assumption traces back to [107]. While some more refined versions of this testing idea were developed since then [26, 90, 78], relatively little has been done on distribution-free MCAR tests, able to detect general distributional differences between the missingness patterns. [101] recently made a step in that direction. Their test is completely nonparametric and shown to be consistent. Empirically it is shown to keep the level and to have a high power over a wide range of distributions. An application area where their proposed test struggles is for higher-dimensional data with little or no complete observations. Their testing paradigm is based on "a reasonable amount" of complete cases and all pairwise comparisons between the observed parts of two missingness pattern groups. This is problematic, since, as the dimension p increases, the number of distinct patterns G tends to grow quickly as well. The most extreme case occurs when $G = n$, that is, every observation forms a missingness pattern group on its own. Consequently, their test appears computationally prohibitively expensive for $p > 10$. Additionally, as the dimension increases, both the number of complete cases and the number of observations per pattern tends to decrease, both contributing to a reduction in power for the test in [101].

In this paper, we try to circumvent these problems in a data-efficient way, by employing a one v.s. all-others approach and using *random projections* in the variable space. Considering observations that are projected into a lower-dimensional space allows us to recover more complete cases. As realized by [101], the problem of MCAR testing, as described above, is a

problem of testing whether distributions across missingness patterns are different. The method presented here relies on some of the core ideas of [124] and [24], who do distributional testing using classifiers. We extend the ideas of [24] to be usable for multiclass classification and use the projection idea of [124] to build a test that is usable and powerful even for "high" dimensions. Moreover, using a permutation approach, we are able to provably keep the nominal level α for all n. As outlined later, this is in contrast to other tests, for which the level might be kept only asymptotically, or is even unclear. The approach of random projections together with a permutation test also allows to extract more information than just a global hypothesis test. We make use of this to calculate individual p-values for each variable. Such a partial test for a variable addresses the null hypothesis that, once that variable is removed, the data is MCAR. Together with the test of overall MCAR, this might point towards the potential source of deviation from the null, that is, the variables causing an MCAR violation.

The paper is structured in the following way. Section 2 introduces notation. Section 3 details the testing framework including the null and alternative hypotheses we consider. Section 4 then showcases how to perform this test in practice and details the algorithm. Section 5 shows some numerical comparisons for type-I error control and power. Section 6 explains the extension of partial p-values, while Section 7 concludes. Appendix 1 contains the proofs of all results, while Appendix 2 adds some additional details and shows computation times of the different tests.

1.1 Contributions

Our contributions can be summarized as follows: We develop the PKLM-test, an easy-to-use and powerful non-parametric test for MCAR, that is applicable even in "high" dimensions. We thereby extend the testing approach of [24] to multiclass testing, which in connection with random projections in the variable space and the Random Forest classifier leads to a powerful test for both discrete and continuous types of data. To the best of your knowledge, no other test is as widely applicable and powerful. Moreover, we are able to formally prove the validity of our p-values for any sample size and number of groups G. As we demonstrate in our simulations, this is remarkable for the MCAR testing literature. It appears no other MCAR test has such a guarantee and many have inflated type-I errors, even in realistic cases, see e.g. the discussion in [78].

As an extension, we can compute partial p-values corresponding to each variable, addressing the question of the source of violation of MCAR among the variables. We demonstrate the validity and power of our test on a wide range of simulated and real datasets in conjunction with different MAR mechanisms. Finally, we make our test available through the R-package `PKLMtest`, available on `https://github.com/missValTeam/PKLMtest` and on CRAN.

1.2 Related Work

Previous advances for tests of MCAR were mostly addressed by [107] (referred to as "Little-test") and extensions [26, 90] under the assumption of joint Gaussianity. To the best of our knowledge, the only distribution-free tests are developed in [78], [101] and [194]. The first paper develops a test (referred to as "JJ-test"), which is distribution-free but is only able to spot differences in the covariance matrices between the different patterns. As such, the simulation study in [101] shows that their test (referred to as "Q-test"), which can detect any potential difference, has much more power than the JJ-test. Moreover, the JJ-test requires prior imputation of missing values, which appears undesirable. [194] develop a test that can be used to subsequently also consistently estimate certain estimators under MCAR. Their test requires a set of fully observed "auxiliary" variables that can be used to first test and then estimate properties of some variable of interest. As such their approach and goals are quite different from ours.

Consequently, the test closest to ours is the fully non-parametric method in [101]. However, it is computationally costly or even infeasible to use their test with dimensions typically found in modern datasets ($p \gg 10$), as all pairwise comparisons between missingness patterns are calculated. While this could in principle be avoided by only checking a subset of pairs, we empirically show that, even if all pairwise comparisons are performed, our test has comparable or even higher power than theirs in their own simulation setting. This gap only increases with the number of dimensions or with a decrease in the fraction of fully observed cases.

We also address a major issue in the MCAR testing literature: none of the proposed methods has a finite sample guarantee of producing valid p-values and for some it can even be empirically checked that the produced p-value is not valid in certain settings. If Z is a p-value generated from a statistical test, then it is valid if $\mathbb{P}(Z \leqslant \alpha) \leqslant \alpha$ under H_0 for all $\alpha \in [0, 1]$, see e.g., [99]. Figure 2 in Section 5 shows some example of previous tests violating this validity of p-values. This issue might be surprising since the requirement of a valid p-value might be the most basic demand a statistical test needs to meet. For the Little-test, this is generally true under normality or asymptotically, that is if the number of observations is going to infinity, under some moment conditions and conditions on the group size. Despite this, Section 5 shows that type error rates can strongly exceed the desired level even in samples of 500 observations. The same holds for the JJ-test of [78] for which we sometimes observed a strong inflation of the level. As with the JJ-test, [101] also do not provide a formal guarantee that the level is kept. Though in our own simulation study, which is similar to theirs, we did not find any notable violation of the level for their test.

To conduct our test, we adapt and partially extend the approaches of [24] and [124]. The former develops a two-sample test using classification, an approach that has gained a lot of attention in recent years (see e.g., [89] or [69] for a literature overview). We extend this approach

	PKLM	Q	Little	JJ
Computational Complexity	$O(pn\log(n))$	$O(n^2p)$	$O(np^2)$	$O(n(p^2+\log(n)))$
Can be used without complete observations	Yes	No	No	Yes
Mixed data types possible	Yes	No	No	No
Does not require initial imputation	Yes	Yes	Yes	No
Power beyond differences in first and second moments	Yes	Yes	No	No

Table 1: Illustration of some of the properties of various tests. For details on the calculation of the computational complexities we refer to Appendix 2.

to multiclass testing, to obtain a test statistic akin to [24], but using the out of bag (OOB) probability estimate of the Random Forest (RF) instead of the in-sample probability. This was already hinted in [69] to increase the power of the two-sample testing approach designed by [24]. [124], on the other hand, use random projections to increase the sample efficiency in the presence of missing values. This simple idea makes our test applicable and powerful, even in high dimensions, and even if the number of patterns G is the same as the number of observations. It can also provide additional information together with the rejection decision, as we demonstrate in Section 6. Finally, through an efficient permutation testing approach, we are able to formally guarantee that our test produces valid p-values for *any* n and any number of groups G. It appears that the PKLM-test is the first MCAR test with such a guarantee. Table 1 summarizes some of the properties of different tests. In particular, "mixed data types" refers to a possible combination of continuous data (such as income) and discrete data (such as gender), while "power beyond differences in first and second moments" means the test is able to detect differences between distributions, even if their means or variances are identical. Though this is difficult to show formally, it appears quite clear that the nonparametric nature of our approach allows for the detection of differences in distributions between patterns, even if the missingness groups all share the same mean or covariance matrix. As outlined in [190] this is crucial for the detection of general MCAR deviations and is not the case, for instance, for the widely used Little-test. Appendix 3 studies a simulated MAR example taken from [190], whereby observed means and variances are approximately the same across different groups. Tests such as the Little-test have no power in this example, yet with our approach, we reach a power of 1.

2 Notation

We assume an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random elements are defined. Along the lines of [128] we introduce the following notation: let $\mathbf{X}^* \in \mathbb{R}^{n \times p}$ be a matrix of n complete samples from a distribution P^* on $\mathbb{R}^p$. We denote by $\mathbf{X}$ the corresponding incomplete dataset that is actually observed. Alongside $\mathbf{X}$ we observe the missingness matrix $\mathbf{M} \in \{0,1\}^{n \times p}$, of which an entry $m_{ij} \in \{0,1\}$ is 1, if entry x^*_{ij} is missing, and 0, if it is observed. Each unique combination in $\{0,1\}^p$ in $\mathbf{M}$ is referred to as a missingness pattern and we assume that there are $G \leqslant n$ unique patterns in $\mathbf{M}$. As an example, for $p = 2$, we might have the pattern $(1,0)$ (first value missing, second observed), $(0,1)$ (first value observed, second missing) or $(0,0)$ (both values are observed). We do not consider the completely missing pattern, in this case $(1,1)$.

We assume that each row x_i (x^*_i) of $\mathbf{X}$ ($\mathbf{X}^*$) is a realization of an i.i.d. copy of the random vector X (X^*) with distribution P (P^*). Similarly, M is the random vector in $\{0,1\}^p$ encoding the missingness pattern of X. Furthermore we assume that P (P^*) has a density f (f^*) with respect to some dominating measure. For a random vector X or an observation x in $\mathbb{R}^p$ and subset $A \subseteq \{1,\ldots,p\}$, we denote as X_A (x_A) the projection onto that subset of indices. For instance if $p = 3$ and $A = \{1,2\}$, then $X_A = (X_1, X_2)$ ($x_A = (x_1, x_2)$). For any set $C \subseteq \{1,\ldots,p\}$, we denote by $\mathbf{X}_{\bullet C}$ the matrix of n observations projected onto dimensions in C, so that $\mathbf{X}_{\bullet C}$ is of dimension $n \times |C|$. Similarly, for $R \subseteq \{1,\ldots,n\}$, $\mathbf{X}_{R\bullet}$ denotes the matrix of observations in set R, over all dimensions, so that the dimension of $\mathbf{X}_{R\bullet}$ is given by $|R| \times p$. We denote by F_g (respectively f_g) the complete distribution (density) of the data in the g^{th} missingness pattern group. A quick overview of the notation including the use of indices for the number of missingness patterns, dimensions, observations, projections and permutations is given in Table 2.

3 Testing Framework

In this section, we formulate the specific null and alternative hypotheses for testing MCAR considered by the PKLM-test. Recalling the notation of Section 2, a missingness pattern is defined by a vector of length p, consisting of ones and zeros, indicating which of the p variables are missing in the given pattern. We divide the n observations into $g \in \{1,\ldots,G\}$ unique groups, such that the observations of each group share the same missingness pattern. Each group $g \in \{1,\ldots,G\}$ contains n_g observations such that $n_1 + \ldots + n_G = n$. Let F_g denote the joint distribution of the p variables in the missingness pattern group g, such that the n_g observations of the group g are i.i.d. draws from F_g. As stated in [101], testing MCAR can be

notation	partial	full
distribution	P	P^*
dataset	$\mathbf{X}$	$\mathbf{X}^*$
observation in $\mathbb{R}^p$	x_i	x_i^*
random vector	X	X^*
density	f	f^*
number of missingness patterns	G	
number of dimensions	p	
number of observations	n	
number of projections	N	
number of permutations	L	

Table 2: **Notation**: Summary of the notation used throughout the paper, with ("partial") and without ("full") considering the missing values.

formulated by the hypothesis testing problem

$$\begin{aligned} &H_0 : F_1^* = F_2^* = \ldots = F_G^* \\ &\text{v.s.} \\ &H_A : \exists\, i \neq j \in \{1, \ldots, G\} \text{ s.t. } F_i^* \neq F_j^*. \end{aligned} \tag{1}$$

We want to emphasize the use of F^* in the testing problem (1), indicating that these hypotheses involve distributions we cannot access. Thus, (1) needs to be weakened. Borrowing the notation of [101], for missingness pattern group g we denote with $\boldsymbol{o}_g$ and $\boldsymbol{m}_g$ the subsets of $\{1, \ldots, p\}$ indicating which variables are observed and which are missing, respectively. We denote the induced distributions by $F_{g,\boldsymbol{o}_g}$ and $F_{g,\boldsymbol{m}_g}$. For two groups i and j, we denote by $\boldsymbol{o}_{ij} := \boldsymbol{o}_i \cap \boldsymbol{o}_j$ the shared observed variables of both groups. As mentioned in [101], it is not possible to test (1) reliably, since the distribution $F_{i,\boldsymbol{m}_i}$ of the unobserved variables is inaccessible. Thus, [101] consider the following hypothesis testing problem

$$\begin{aligned} &H_0 : F_{i,\boldsymbol{o}_{ij}} = F_{j,\boldsymbol{o}_{ij}} \; \forall i \neq j \in \{1, \ldots, G\} \\ &\text{v.s.} \\ &H_A : \exists\, i \neq j \in \{1, \ldots, G\} \text{ with } \boldsymbol{o}_{ij} \neq \varnothing \text{ s.t. } F_{i,\boldsymbol{o}_{ij}} \neq F_{j,\boldsymbol{o}_{ij}}. \end{aligned} \tag{2}$$

The null hypothesis H_0 of (2) is implied by H_0 of (1), but not vice-versa. In other words, if we can reject the null hypothesis of (2), we can also reject the null hypothesis of (1). But if the null hypothesis of (2) cannot be rejected, there could still be a distributional change for different groups in the unobserved parts, so that the null hypothesis of (1) is not true. In this

case, the missingness mechanism would be MNAR. Thus, using the terminology of [144], (2) tests the "observed at random" (OAR) hypothesis instead of the MCAR hypothesis. The differentiation can be circumvented by assuming that the missingness mechanism is MAR, which is the approach usually taken, see [101].

The comparison of all pairs of missingness groups in the hypothesis testing problem (2) is problematic however, as laid out in the introduction. In the following, we circumvent this problem in a data-efficient way, considering a one v.s. all-others approach and employing *random projections* in the variable space. Considering observations that are projected into a lower-dimensional space allows us to recover more complete cases. Let $\mathcal{A}$ be the set of all possible subsets of $\{1,\dots,p\}$ with at most $p-1$ elements. For $A \in \mathcal{A}$ we define by $\mathcal{N}_A$ the indices in $1,\dots,n$ of observations that are observed with respect to projection A, i.e., observations of which the projection onto A is fully observed. These observations may belong to different missingness pattern groups $g \in \{1,\dots,G\}$. As an example, $x = (\texttt{NA}, 1, \texttt{NA}, 2, 4)$ and $y = (\texttt{NA}, \texttt{NA}, \texttt{NA}, 1, 3)$ are not complete and not in the same group, however if we project them to the dimensions $A = \{4,5\}$, x_A and y_A are complete in this lower-dimensional space.

Additionally, to circumvent the problem of many groups with only a few members, we assign new *grouping or class labels* to all observations in $\mathcal{N}_A$. To do so, we consider the set of projections $\mathcal{B}(A^c)$, which is defined as the power set of $\{1,\dots,p\}\backslash A$. The set $\mathcal{B}(A^c)$ is never empty since $|A| \leqslant p-1$. For a given projection $B \in \mathcal{B}(A^c)$, we project all observations with indx in $\mathcal{N}_A$ to B and form new collapsed missingness pattern groups $G(A,B)$, where $G(A,B)$ is the set of labels corresponding to distinct missingness patterns among observations with index in $\mathcal{N}_A$ projected to B. This is solely done to determine the grouping or class labels of observations with index in $\mathcal{N}_A$. If two observations with index in $\mathcal{N}_A$ are in the same overall missingness pattern group $g \in \{1,\dots,G\}$, they also end up in the same collapsed group. The other direction is not true, that is the number of collapsed groups $|G(A,B)|$ is at most as large as the initial number of distinct groups G among the observations with index in $\mathcal{N}_A$. Considering again $x = (\texttt{NA}, 1, \texttt{NA}, 2, 4)$ and $y = (\texttt{NA}, \texttt{NA}, \texttt{NA}, 1, 3)$, if $B = \{1,2\}$, then observations x and y are not in the same missingness pattern group. However, if $B = \{1,3\}$, we assign the same class label to x and y. Thus, given the projection A, we obtain a set of fully observed observations $\mathbf{X}_{\mathcal{N}_A,A} = \mathbf{X}^*_{\mathcal{N}_A,A}$, and given the projection B we assign to them the $|G(A,B)|$ different class labels. Figure 1 provides a schematic illustration of projections A and B on a more complicated example with four observations, each corresponding to a different pattern (i.e., $n = G = 4$). According to $B = \{2\}$, the first observation in $\mathbf{X}_{\mathcal{N}_A,A}$ obtains one collapsed class label whereas the second and third observation obtain another, common label, resulting in $|G(A,B)| = 2$.

We are now equipped to formulate our one v.s. all-others approach with the hypothesis

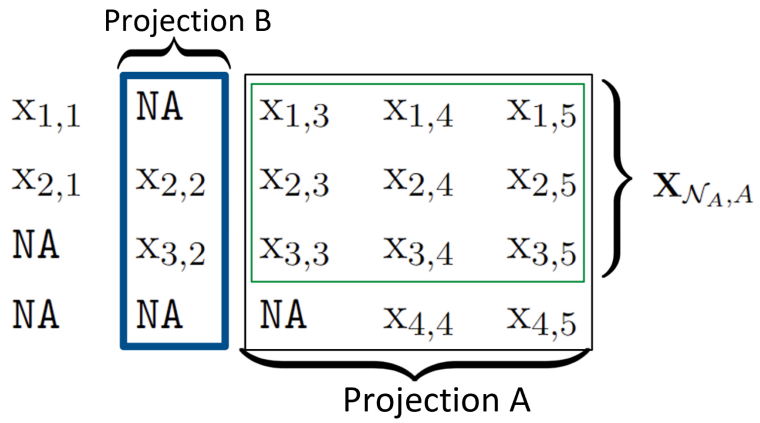


Figure 1: Illustration of the projections A and B in an example with $n = 4$ and $p = 5$. In a first step, a projection $A = \{3,4,5\} \subset \{1,\dots,5\}$ is drawn. The fully observed points on A form $\mathbf{X}_{\mathcal{N}_A,A}$, as indicated in green. In a second step, a projection $B = \{2\} \subset \{1,\dots,5\}\backslash A$ is drawn, as indicated in blue. The patterns in projection B then determine the labels assigned to the observations in $\mathbf{X}_{\mathcal{N}_A,A}$. In this case we obtain two different class labels: the first observation has one label, and the second and third observations share another common label.

testing problem

$$
\begin{aligned}
&H_0 : F_{g,A} = \sum_{j \in G(A,B)\backslash g} \omega_j^g F_{j,A} \\
&\forall g \in G(A,B), \forall B \in \mathcal{B}(A^c), \forall A \in \mathcal{A} \\
&\text{v.s.} \\
&H_A : F_{g,A} \neq \sum_{j \in G(A,B)\backslash g} \omega_j^g F_{j,A}. \\
&\text{for one } g \in G(A,B), B \in \mathcal{B}(A^c), A \in \mathcal{A},
\end{aligned}
\tag{3}
$$

where $F_{g,A}$ is the joint distribution of the observations of class g with index in $\mathcal{N}_A$ and the groups $j \in G(A,B)$ are jointly determined by A and B. Thus, we compare the distribution of the observed part with respect to A of one group g with the mixture of the observed parts of the rest of the groups. The weights ω_j^g are non-negative, sum to 1, and are proportional to the respective fraction of observations in class j.

Example 7. *To give some intuition about the hypothesis testing problem* (3)*, we relate it to the hypothesis testing problem* (2) *with the help of the example of Figure 1. In this example, each observation $i = 1,\dots,4$ has a different pattern and can thus be seen as a draw from a distribution F_i^*. We first assume that the null hypothesis of* (3) *holds and show, as an example, that this implies $F_{1,o_{13}} = F_{3,o_{13}}$. Since the null hypothesis of* (3) *refers to all $A \in \mathcal{A}$, it also includes $A = o_{13} = \{3,4,5\}$, which is what we consider in Figure 1. While we are only interested in $F_{1,A}$ and $F_{3,A}$, taking $B = \{1,2\}$ the observations in $\mathbf{X}_{\mathcal{N}_A,A}$ come from the three*

distributions $F_{1,A}, F_{2,A}, F_{3,A}$. *Due to* (3) *it holds that*

$$\begin{aligned} F_{1,A} &= \omega_2^1 F_{2,A} + \omega_3^1 F_{3,A}, \\ F_{2,A} &= \omega_1^2 F_{1,A} + \omega_3^2 F_{3,A}, \\ F_{3,A} &= \omega_1^3 F_{1,A} + \omega_2^3 F_{2,A}. \end{aligned} \tag{4}$$

Some algebra shows that equation system (4) *is equivalent to* $F_{1,A} = F_{2,A} = F_{3,A}$, *which in particular means* $F_{1,A} = F_{3,A}$, *that we wanted to show. While we took* $i = 1$ *and* $j = 3$ *as an example matching Figure 1, we cycle through all* $A \in \mathcal{A}$ *in* (3) *and thus* $A = \boldsymbol{o}_{ij}$ *for all patterns* i, j *eventually. We now assume that the null hypothesis of* (2) *is true and consider again* $A = \{3, 4, 5\}$ *as an example. Since we only look at the fully observed observations in* $\mathcal{N}_A$ *in* (3), *i.e., leave out the fourth point, we again deal with the three distributions* $F_{1,A}$, $F_{2,A}$, $F_{3,A}$. *Moreover, by construction,* $A \subset \boldsymbol{o}_{12}$ *and* $A \subset \boldsymbol{o}_{13}$ *(even* $A = \boldsymbol{o}_{13}$ *in this case). Thus,* $F_{1,\boldsymbol{o}_{12}} = F_{2,\boldsymbol{o}_{12}}$ *and* $F_{1,\boldsymbol{o}_{13}} = F_{3,\boldsymbol{o}_{13}}$, *implied by the null hypothesis of* (2), *means that* $F_{1,A} = F_{2,A} = F_{3,A}$, *which implies* (4). *Again this might seem constructed, but since by definition,* (3) *only considers the distributions* $F_{i,A}$ *and* $F_{j,A}$ *of fully observed points on A, it will always hold that* $A \subset \boldsymbol{o}_{ij}$.

We make note of an abuse of notation in (3), as the group g in $F_{g,A}$ only corresponds to the same index of F_g in (2), if $B = A^c$, as can be seen in the example of Figure 1: If $B = A^c$, the three observations in $\mathbf{X}_{\mathcal{N}_A,A}$ are drawn from $F_{1,A}, F_{2,A}$ and $F_{3,A}$ respectively. However, if $B = \{2\}$, then observations two and three are now assumed to be drawn from a single distribution, which corresponds to a mixture of $F_{2,A}$ and $F_{3,A}$.

In short, the null hypothesis of (3) implies the null hypothesis of (2) because for $A = \boldsymbol{o}_{ij}$, observations coming from $F_{i,A}$ and $F_{j,A}$ are contained in $\mathbf{X}_{\mathcal{N}_A,A}$. Vice-versa, the null hypothesis of (2) implies the null hypothesis of (3) because A is nested in $\boldsymbol{o}_{ij}$ for all F_i and F_j considered on A. This actually sketches the proof of the following result:

Proposition 29. *Hypothesis testing problem* (3) *is equivalent to* (2).

Tackling hypothesis testing problem (3) would be rather inefficient since we might test many times the same hypothesis when cycling through all $A \in \mathcal{A}$ and $B \in \mathcal{B}(A^c)$. However, the idea is that A and B will only be random draws from $\mathcal{A}$ and $\mathcal{B}(A^c)$. This is discussed in the next section.

4 MCAR test Through Classification

In this section we introduce the classification-based statistic of our test and detail the implementation of our permutation approach, permuting the rows of the missingness matrix $\mathbf{M}$, to obtain a valid test.

4.1 Test Statistic U

Let us fix a projection $A \in \mathcal{A}$ and corresponding projection $B \in \mathcal{B}(A^c)$. We denote the induced collapsed class labels based on projections A and B by $Y^{(A,B)}$, by X_A the projection of the random vector X on A and correspondingly by x_A the projection on A of observation x in $\mathbf{X}_{\mathcal{N}_A,A}$. Furthermore, we define for each $g \in G(A,B)$ and x in $\mathbf{X}_{\mathcal{N}_A,A}$ the following quantities:

$$\begin{aligned} p_g^{(A,B)}(x) &:= P(Y^{(A,B)} = g \mid X_A = x_A), \\ f_g^{(A,B)}(x) &:= P(x_A \mid Y^{(A,B)} = g), \\ \pi_g^{(A,B)} &:= P(Y^{(A,B)} = g). \end{aligned}$$

Let us fix $g \in G(A,B)$ as well. We reformulate the hypothesis testing problem (3):

$$\begin{aligned} H_{0,g}^{(A,B)} &: f_g^{(A,B)} = \frac{1}{1-\pi_g^{(A,B)}} \sum_{j \in \{1,\dots,G(A,B)\}\setminus g} \pi_j^{(A,B)} f_j^{(A,B)} \\ &\text{v.s.} \\ H_{1,g}^{(A,B)} &: f_g^{(A,B)} \neq \frac{1}{1-\pi_g^{(A,B)}} \sum_{j \in \{1,\dots,G(A,B)\}\setminus g} \pi_j^{(A,B)} f_j^{(A,B)}. \end{aligned} \tag{5}$$

Let $S_{f_g^{(A,B)}} \subset \mathcal{N}_A$ denote the indices of observations in $\mathbf{X}_{\mathcal{N}_A,A}$ that belong to class g. For each missingness pattern g, we now define the following statistic in analogy to [24],

$$U_g^{(A,B)} := \frac{1}{|S_{f_g^{(A,B)}}|} \sum_{i \in S_{f_g^{(A,B)}}} \left(\log \frac{p_g^{(A,B)}(x_i)}{1-p_g^{(A,B)}(x_i)} - \log \frac{\pi_g^{(A,B)}}{1-\pi_g^{(A,B)}} \right). \tag{6}$$

This statistic is motivated by the following claim:

Lemma .11. *The logarithm of the density ratio for testing (5) is given by $U_g^{(A,B)}$.*

The main motivation for the form of this test-statistic is that one can use the same arguments as in [24, Proposition 1] to show that a test based on $U_g^{(A,B)}$ will have the highest power among all tests for (5), according to the Neyman-Pearson Lemma. In addition, the test statistic converges to the Kullback-Leibler Divergence between $f_g^{(A,B)}$ and the mixture of the other densities, motivating the name of our MCAR test. A high value of KL-Divergence indicates that the distributions of two samples deviate strongly from each other.

Lemma .12. *$U_g^{(A,B)}$ converges in probability to the Kullback-Leibler Divergence between $f_g^{(A,B)}$ and the mixture of the other densities:*

$$U_g^{(A,B)} \to \mathbb{E}_{f_g} \left[\log \frac{f_g^{(A,B)}(X)(1-\pi_g^{(A,B)})}{\sum_{j \in G(A,B)\setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(X)} \right],$$

as n_g and $\sum_{j \in \{1,\dots,G\}\setminus g} n_j \to \infty$ and $n_g/n \to \pi_g^{(A,B)} \in (0,1)$.

Since the statistic $U_g^{(A,B)}$ is evaluated only on cases $x \in S_{f_g^{(A,B)}}$, it holds that $f_g^{(A,B)}(x) = f_g^{*(A,B)}(x)$ and $p_g^{(A,B)}(x) = p_g^{*(A,B)}(x)$. This means that the projected complete and incomplete distributions coincide on the projected complete samples. Thus we are indeed asymptotically measuring the Kullback-Leibler Divergence between $f_g^{*(A,B)}$ and the mixture of the other densities.

Since there might be only very few observations for a single class g, we symmetrize the KL-Divergence. That is, we use the samples of all classes to evaluate the KL-Divergence and not only the samples of class g. Let $S_{f_{g^c(A,B)}} \subset \mathcal{N}_A$ denote the indices of observations in $\mathbf{X}_{\mathcal{N}_A,A}$ that belong to all other classes $G(A,B)\backslash g$. For each missingness pattern g, we will use, in the following, the difference between two of the above statistics, namely

$$U_g^{(A,B)} - U_{g^c}^{(A,B)} = \frac{1}{|S_{f_g^{(A,B)}}|} \sum_{i \in S_{f_g^{(A,B)}}} \log \frac{p_g^{(A,B)}(x_i)}{1 - p_g^{(A,B)}(x_i)} - \frac{1}{|S_{f_{g^c(A,B)}}|} \sum_{i \in S_{f_{g^c(A,B)}}} \log \frac{p_g^{(A,B)}(x_i)}{1 - p_g^{(A,B)}(x_i)}, \quad (7)$$

where the terms including the class probabilities $\pi_g^{(A,B)}$ cancel out. This difference converges to the symmetrized KL-Divergence between the mixture of $f_g^{(A,B)}$ and the remaining classes and is more sample efficient than only using $U_g^{(A,B)}$. The test statistic for fixed (A,B) is then given by

$$U^{(A,B)} := \sum_{g=1}^{G(A,B)} (U_g^{(A,B)} - U_{g^c}^{(A,B)}),$$

and the final test statistic is defined as

$$U := \mathbb{E}_{A \sim \kappa, B \sim \kappa(A^c)}[U^{(A,B)}]. \quad (8)$$

4.2 Practical Estimation of U

We estimate $p_g^{(A,B)}$ with a multiclass-classifier, yielding $\hat{p}_g^{(A,B)}$. Plugging-in this quantity into (7) yields $\hat{U}_g^{(A,B)} - \hat{U}_{g^c}^{(A,B)}$. We then estimate $U^{(A,B)}$ by

$$\hat{U}^{(A,B)} := \sum_{g=1}^{G(A,B)} (\hat{U}_g^{(A,B)} - \hat{U}_{g^c}^{(A,B)}).$$

Finally, we estimate U by

$$\hat{U} := \frac{1}{N} \sum_{i=1}^{N} \hat{U}^{(A_i,B_i)}, \quad (9)$$

where N is the number of draws of pairs of projections (A_i, B_i), $i = 1, \ldots, N$, with $A \in \mathcal{A}$ according to a distribution κ and $B \in \mathcal{B}(A^c)$ according to a distribution $\kappa(A^c)$.

Our chosen multiclass classifier is Random Forest [19, 16], more specifically, the probability forest of [111]. That is, for each of the N projections, we fit a Random Forest with a specified

number of trees, a parameter called `num.trees.per.proj`. Thus, for each tree (or group of trees) a random subset of variables and labels is chosen based on which the test statistic is computed. In each tree, we set `mtry` to the full dimension of the projection to not have an additional subsampling effect. This approach aligns naturally with the construction of Random Forest, as the overall approach might be seen as one aggregated Random Forest, which restricts the variables in each tree or group of trees to a random subset of variables. We finally use the OOB-samples for predicting $\hat{p}_g^{(A,B)}$.

The question remains how to sample the sets $(A_1, B_1), \dots (A_N, B_N)$ at random. Our chosen approach is quite simple: we first randomly sample a number of dimensions r_1 by drawing uniformly from $\{1, \dots, p-1\}$. We then draw r_1 values without replacement from $\{1, \dots, p\}$ to obtain A. Similarly, we randomly draw a value r_2 from $\{1, \dots, p - r_1\}$ and then draw r_2 values without replacement from $\{1, \dots, p\} \backslash A$ to obtain B. We then consider $\mathbf{M}_{\mathcal{N}_A, B}$, i.e., all patterns for the fully observed observations in A projected to B, and build the labels $Y^{(A,B)}$ based on the patterns in this matrix. This simple approach is used as a default, but one could also employ a more data-adaptive subsampling. In our algorithm, we might restrict the number of collapsed classes by selecting B corresponding to A accordingly. The parameter indicating the maximal number of collapsed classes allowed is given by `size.resp.set`. If set to 2, we reduce the multi-class problem to a two-class problem. In Algorithm 1 we provide the pseudo-code for the estimation of $\hat{U}^{(A,B)}$.

To ensure that the level is kept by a test based on the statistic $\hat{U}$ for any choice of κ and $\kappa(A^c)$, we use a permutation approach, as detailed next.

Algorithm 1: Uhat$(\mathbf{X}, \mathbf{M}, A, B)$

Inputs: incomplete dataset $\mathbf{X}$, missingness indicator $\mathbf{M}$, projections A, B

Result: $\hat{U}^{(A,B)}$

Hyper-parameters: number of trees per projection `num.trees.per.proj`, standard parameters of the Probability Forests `size.resp.set`;

- Recover the complete cases $\mathcal{N}_A$ with respect to A;
- Generate the $G(A, B)$ collapsed class labels $Y^{(A,B)}$ from $\mathbf{M}_{\mathcal{N}_A, B}$;
- Fit a multi-class probability forest with `num.trees.per.proj` trees and `mtry` full;

for $g = 1, \dots, G(A, B)$ **do**

- Estimate $\hat{p}_g^{(A,B)}$ with the fitted forest above using out-of-bag probabilities;
- Return the log-likelihood contribution $\hat{U}_g^{(A,B)} - \hat{U}_{g^c}^{(A,B)}$ for class g;

end

- Average the log-likelihood ratio contributions $\hat{U}_g^{(A,B)} - \hat{U}_{g^c}^{(A,B)}$ from the $G(A, B)$ collapsed classes g to get the statistic $\hat{U}^{(A,B)}$;

4.3 Permutation Test

To ensure the correct level, we follow a permutation approach. Informally speaking, the permutation approach works in this context if the testing procedure can be replicated in exactly the same way on the randomly permuted class labels. This is not completely trivial in this case, as the labels are defined in each projection via the missingness matrix $\mathbf{M}$. It can be shown numerically that permuting the labels at the level of the projection does not conserve the level, as this is blind to the correspondence between the projections across the permutations.

The key to the correct permutation approach is to permute the rows of $\mathbf{M}$. That is, for L permutations σ_ℓ, $\ell = 1, \dots, L$, we obtain L matrices $\mathbf{M}_{\sigma_1}, \dots, \mathbf{M}_{\sigma_L}$ with only the rows permuted. Then we proceed as above: We sample $A \sim \kappa$, $B \sim \kappa(A^c)$ and for each permutation of rows σ_ℓ, $\ell = 1, \dots, L$, we calculate $U^{(A,B)}_{g,\sigma_\ell} - U^{(A,B)}_{g^c,\sigma_\ell}$ as in (7). Using $\hat{p}^{(A,B)}_g$ instead of $p^{(A,B)}_g$ this results in $\hat{U}^{(A,B)}_{g,\sigma_\ell}$ and in the statistic

$$\hat{U}^{(A,B)}_{\sigma_\ell} := \sum_{g=1}^{G(A,B)} \hat{U}^{(A,B)}_{g,\sigma_\ell} - \hat{U}^{(A,B)}_{g^c,\sigma_\ell}.$$

We note that we do not need to refit the forest for this permutation approach to work. Instead, we can directly use $\hat{p}^{(A,B)}_g$ from the original Random Forest that we fitted on the original $\mathbf{M}$.

Finally, we calculate the empirical distribution of the test-statistic under the null, by calculating for $\ell = 1, \dots, L$,

$$\hat{U}_{\sigma_\ell} := \frac{1}{N} \sum_{j=1}^{N} \hat{U}^{(A_j,B_j)}_{\sigma_\ell}. \tag{10}$$

The p-value of the test is then obtained as usual by

$$Z := \frac{\sum_{\ell=1}^{L} I\{\hat{U}_{\sigma_\ell} \geqslant \hat{U}\} + 1}{L + 1}. \tag{11}$$

Then it follows from standard theory on permutation tests that Z is a valid p-value:

Proposition 30. *Under H_0 in* (1)*, and Z as defined in* (11)*, it holds for all $z \in [0, 1]$ that*

$$\mathbb{P}(Z \leqslant z) \leqslant z. \tag{12}$$

Algorithm 2 summarizes the testing procedure.

Algorithm 2: PKLMtest($\mathbf{X}$)

Inputs: incomplete dataset $\mathbf{X}$

Result: p-value

Hyper-parameters: number of pairs of projections N, number of permutations L, number of trees per projection `num.trees.per.proj`, standard parameters of the Probability Forests, maximal number of collapsed classes `size.resp.set`;

- Randomly permute the rows of $\mathbf{M}$ L times to obtain $\mathbf{M}_{\sigma_1}, \dots, \mathbf{M}_{\sigma_L}$;

for $j = 1, \dots, N$ **do**

- Sample a pair of projections (A_j, B_j) hierarchically according to $A_j \sim \kappa$ and $B_j \sim \kappa(A_j)$;
- Calculate $\hat{U}^{(A_j,B_j)} = \text{Uhat}(\mathbf{X}, \mathbf{M}, A_j, B_j)$;
- **for** $\ell = 1, \dots, L$ **do**
  - Calculate $\hat{U}^{(A_j,B_j)}_{\sigma_\ell} = \text{Uhat}(\mathbf{X}, \mathbf{M}_{\sigma_\ell}, A_j, B_j)$;
- **end**

end

- Average the statistics $\hat{U}^{(A_j,B_j)}$, $\hat{U}^{(A_j,B_j)}_{\sigma_\ell}$ over the couples of projections (A_j, B_j) to get the final statistic $\hat{U}$, $\hat{U}_{\sigma_\ell}$, $\ell = 1, \dots, L$;
- Obtain the p-value with (11);

5 Empirical Validation

In this section, we empirically showcase the power of our test in comparison to recent competitors on both simulated and real data. The simulation setting is set up along the lines of [78] and [101] with a common MAR mechanism. For the real datasets we also add a random MAR generation through the function `ampute` of the R-package `mice`, see e.g., [150].

As we did throughout the paper, we refer to our test as "PKLM", the test of [101] as "Q", the test of [107] as "Little" and finally the one of [78] as "JJ". The Little-test is computed with the R-package `naniar` [171], while the JJ-test uses the code of the R-package `MissMech` [79]. Finally, the code for the Q-test was kindly provided to us by the authors.

5.1 Simulated Data

We vary the sample size n, the number of dimensions p, and the number of complete observations, which we denote by r. Cases $1-8$ describe the following different data distributions, similarly as in [101] and in [78]: Throughout, I_p is a covariance matrix with diagonal elements 1 and off-diagonal elements 0 while Σ is a covariance matrix with diagonal elements 1 and off-diagonal elements 0.7:

1. A standard multivariate normal distribution with mean 0 and covariance I_p,
2. a correlated multivariate normal distribution with mean 0 and covariance Σ,
3. a multivariate t-distribution with mean 0, covariance I_p and degree of freedom 4,
4. a correlated multivariate t-distribution with mean 0, covariance Σ and degree of freedom 4,
5. a multivariate uniform distribution which has independent uniform$(0, 1)$ marginal distributions,
6. a correlated multivariate uniform distribution obtained by multiplying $\Sigma^{1/2}$ to the multivariate uniform distribution in 5,
7. a multivariate distribution obtained by generating $W = Z + 0.1Z^3$, where Z is from the standard multivariate normal distribution,
8. a multivariate Weibull distribution which has independent Weibull marginal distribution, and each Weibull marginal distribution has scale parameter 1 and shape parameter 2.

The above implements the fully observed $\mathbf{X}^*$. To compute the type-I error, we then simulate the MCAR mechanism where each value in the p columns of the missingness matrix $\mathbf{M}$ has a probability of $1 - r^{1/p}$ being one and is otherwise zero. To compute the power, we simulate the MAR mechanism following the description in [101]: We generate $\mathbf{M}$ such that the first column consists only of zeros so that the first variable is fully observed. Further, each value in the remaining $p - 1$ columns has a probability of $1 - r^{1/(p-1)}$ being one, while the rest is zero. This results, on average, in r rows in $\mathbf{M}$ with only zeros, and thus in r fully observed rows in $\mathbf{X}$. Next, we sort the rows of $\mathbf{M}$ into two groups, those that will be fully observed (complete group) and those that will have at least one missing value (missing group). So far, the generation is still MCAR. However now, for each row $i = 1, \ldots, n$ we compare $\mathbf{X}^*_{i,1}$ with the mean of $\mathbf{X}^*_{\bullet,1}$, denoted by $\bar{X}_1$. If $\mathbf{X}^*_{i,1} < \bar{X}_1$, the corresponding row i is placed into the complete group with probability 1/6, and with probability 5/6 into the missing group. That is, with probability 1/6, the row i is paired with a row in $\mathbf{M}$ from the complete group, and with probability 5/6, it is paired with a row from the missing group. Thus, in this case it is 5 times more likely that the row is placed in the missing group. On the other hand, if $\mathbf{X}^*_{i,1} \geqslant \bar{X}_1$ the situation reverses, and row i is 5 times more likely to be associated with a row in $\mathbf{M}$ from the complete group. Assigning the rows of $\mathbf{X}^*$ successively to the rows of $\mathbf{M}$ like this results in $\mathbf{X}$ with MAR missingness.

Each experiment was rerun `nsim` $= 300$ times to compute type-I error and power. We used the following default hyperparameter setting for the computation of our PKLM-test: number of permutations `nrep` $= 30$, number of projections `num.proj` $= 100$, minimal node size in a tree

`min.node.size` $= 10$, number of fitted trees per projection `num.trees.per.proj` $= 200$ and maximal number of collapsed classes allowed in a projection `size.resp.set` $= 2$. We note that the choice of these hyperparameters is intriguingly simple: besides `size.resp.set`, it holds that "higher values are better". Thus, as with RF in general, it is mostly a question of computational resources determining how large the values can be chosen. This is especially true for the number of trees for each forest, which should be relatively high in order to minimize additional randomness. We found `num.trees.per.proj` $= 200$ to be a good compromise between speed and accuracy. As the level is guaranteed for any number of permutations, and we desired a choice of hyperparameters that would work for $p = 4$ as well as $p = 40$, we chose the number of permutations low (`nrep` $= 30$), but the number of projections relatively high (`num.proj` $= 100$). The only "difficult" parameter to set is `size.resp.set`, as there appears to be some loss in accuracy when the number of classes is larger than two. We thus found that `size.resp.set` $= 2$, generating two classes, works well in a wide range of examples.

As mentioned throughout the paper, the Q-test could not be calculated for a large range of settings.[1] In particular, computation times were infeasible for the setting $p = 10$ and $r = 0.1$, and for any configuration with $p = 20$ or $p = 40$. For the setting $n = 500$, $p = 10$ and $r = 0.1$ for instance, one test for case 2 took around 20 minutes to finish, implying an approximate overall computation time of $500 \cdot 8 \cdot 2 \cdot 20 = 16000$ minutes or approximately 110 24-hour days. This despite the fact that the R-code of the Q-test we received was well implemented. In the upcoming Tables 3 and 4 of results we always used the nominal level of $\alpha = 0.05$. We boldfaced the results for each row in the tables in the following manner: Whenever the type I error of a test is below or equal to 0.05 and the test has the best power, it will be boldfaced. If this is true for more than one test, they are all boldfaced. Additionally, we boldfaced all the type-I errors that are below or equal to the nominal level $\alpha = 0.05$ to indicate which tests holds the level on average in the given settings.

In the simulation set-up of $n = 200$ and $p = 4$, the Q-test is very powerful, while keeping the nominal level. The PKLM-test is rarely the most powerful here, however the power of the PKLM-test is often relatively close to the best power. As an example, in case 2 for $r = 0.65$, the Q-test has a power of 1 while the PKLM-test has a power 0.93, with both keeping the nominal level $\alpha = 0.05$.

In the set-up of $n = 500$ and $p = 10$, the overall picture changes. The PKLM-test is in all but two of the 24 cases the most powerful test, sometimes leaving the second-best test quite far behind. As an example, in case 3 for $r = 0.65$, the PKLM-test has a power of 0.85 while the Q- and the Little-test exhibit a power of 0.26 and 0.61, respectively. While the Little- and the JJ-test often show inflated levels, this is never a problem for the valid PKLM-test.

In the simulation set-up of $n = 500$ and $p = 20$, it appears as if the Little-test is a strong

[1] The largest number p reported in the paper of [101] is 10, while r is at least 0.35.

competitor. But this is only until one considers its type-I error. Though to a much lesser degree than the JJ-test, the type-I error is often heavily larger than the nominal level. Considering for instance case 4, the power of the Little-test is even slightly less than its actual type-I error for $r = 0.1$. In case 4 with $r = 0.35$, our test displays a power of 0.89 and keeps the level, while the Little-test only has a power of 0.33 despite having a grossly inflated type-I error. All of these problems are worsened for the JJ-test, which often displays an inflated type-I error in almost all cases and simulation set-ups. A similar story plays out in the case $r = 0.65$.

Finally, in the simulation set-up of $n = 1000$ and $p = 40$, the power of our test is again much better than that of all other tests. Interestingly, the PKLM-test tends to have higher power when the components of the distribution are not independent, such as in the cases 2, 4, 6, and 8. For example, in case 1 for $r = 0.65$, PKLM has a power of 0.2, while for case 2 it has a power of 0.95. The main difference between these two cases is the strong positive correlation induced in case 2. This pattern repeats: in all correlated examples and for both $r = 0.65$ and $r = 0.35$, the PKLM has a power nearing 1, whereas in the independent versions, the power is closer to the type-I error. Thus, our test is able to use the dependencies in the data to its advantage, at least for $r = 0.65$ and $r = 0.35$, and can reach a very high power even for comparatively large p.

In summary, our test is very competitive even in small dimensions, where the Q-test is very powerful. It leaves behind all other tests by a wide margin as soon as one increases p. The Q-test remains strong in these situations as well, but becomes quickly infeasible as either p increases or the fraction of complete cases r decreases. Crucially, only the PKLM-test and the Q-test are able to consistently keep the nominal level over all experiments, with the Little- and JJ-test showing blatant inflation of the type-I error in many situations. This is the case despite the fact that simply checking the type-I error for a single level α (0.05 in this case) is far from sufficient to analyse the validity of a p-value.

As an illustration, we randomly chose one of the above experiments in which the Little-test kept the nominal level, e.g., in the simulation set up $n = 500$, $p = 10$, $r = 0.65$ in case 5. In Figure 2 we plot the empirical cumulative distribution functions (ecdf) of 500 p-values under the null (MCAR) of the four different tests. The red line is the $x = y$ line. In blue we plotted 100 ecdfs of a uniform$(0, 1)$-distribution. As described in Equation (12), a valid p-value has the property that the corresponding black ecdf values do not lie above the region defined by the blue lines. As Proposition 30 predicts, this is clearly the case for the PKLM-test. That the p-values appear rather discrete stems from the fact that we chose a low number of permutations (`nrep` $= 30$). The Q-test is sometimes overshooting the red line, though this appears to mostly stem from estimation error. In general, it is remarkable how closely the ecdfs of p-values from both the Q- and PKLM-test resemble the ecdf of a uniform sample. The JJ-test appears to consistently have $P(Z \leqslant z) \geqslant z$. The Little-test finally appears to produce a valid p-value

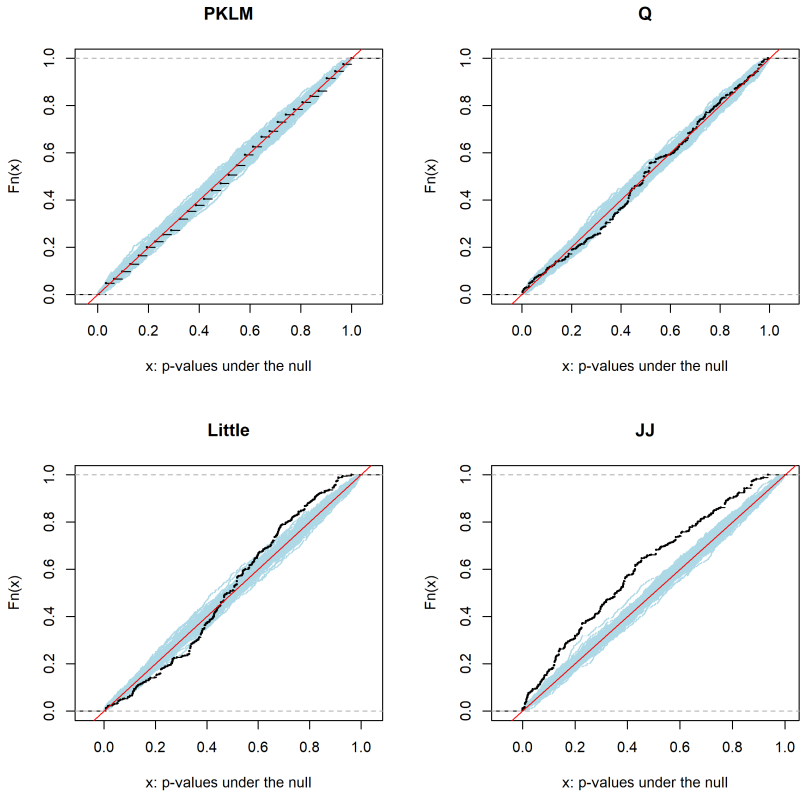


Figure 2: Example plot of cumulative distribution function values of the p-values under the null (MCAR) of the four different tests. The simulation set up is $n = 500$, $p = 10$, $r = 0.65$ in case 5, with 500 repetitions. The red line is the $x = y$ line, while the blue lines show 100 ecdfs of 500 simulated uniform random variables.

as long as only values $z < 0.5$ are considered. For $z \geqslant 0.5$, the the ecdf clearly violates the requirement of a valid p-value. If there is no theoretical guarantee, it is thus important to not just check the type-I error at $\alpha = 0.05$, but to instead consider other levels, e.g., $\alpha = 0.1$.

5.2 Real Data

We used 13 real datasets with varying number of observations n and dimensions p for further empirical assessment of the PKLM-test and comparison to the other three tests. The datasets are available in the UCI machine learning repository[2]. We preprocessed the data by cancelling factor variables, in order to be able to run all other three tests. However, we kept numerical variables with only few unique values.

For the generation of the `NA`s, we use an overall probability of missingness of $p_{miss} = 0.3$ (not to be confused with r from the last subsection, denoting the number of complete cases). We used a random MAR generation through the function `ampute` of the R-package `mice`. This function can randomly generate realistic MAR mechanisms, see e.g., [150]. Each experiment was run `nsim` $= 300$ times to compute the type-I error and power. We used the following hyperparameter setting for the computation of our PKLM-test: number of permutations `nrep` $=$ 30, number of projections `num.proj` $= 300$, minimal node size in a tree `min.node.size` $= 10$, number of fitted trees per projection `num.trees.per.proj` $= 200$ and maximal number of collapsed classes allowed in a projection `size.resp.set` $= 2$. The results are shown in Table 5. Our test is again very competitive with the best power in 7 out of 13 datasets, conditional on valid type-I errors. The Little-test shows also often good performance, though given the problematic level displayed in the previous section, this has to be considered with some care. The Q-test also has relatively high power in the situations where it can be calculated. However, due to computational time we only run the Q-test for $p \leqslant 10$. All in all, we see that the Q-test quickly gets infeasible for large p and n and the advantage of the PKLM-test strengthens with increasing p.

6 Extension

In addition to the "global test" of MCAR, we can study the effect of single variables: For any given variable $k = 1, \ldots, p$, we can calculate

$$\hat{U}^{-k} = \frac{1}{|\mathcal{P}_{-k}|} \sum_{i \in \mathcal{P}_{-k}} \hat{U}^{(A_i, B_i)},$$

where $\mathcal{P}_{-k}$ are all pairs of projections (A_i, B_i) from the N randomly chosen ones, with B_i not containing variable k. We can use the analogous calculation based on the permuted missingness

[2] `https://archive.ics.uci.edu/ml/index.php`

				Power				Type-I Error			
n	p	r	case	PKLM	Q	Little	JJ	PKLM	Q	Little	JJ
200	4	0.65	1	0.73	**0.98**	0.98	0.12	**0.03**	**0.03**	0.06	**0.04**
			2	**0.93**	1.00	0.96	0.04	**0.03**	0.06	0.06	**0.05**
			3	0.81	**0.94**	0.92	0.05	**0.03**	**0.02**	**0.04**	0.08
			4	0.89	**0.97**	0.91	0.05	**0.01**	**0.03**	**0.05**	**0.05**
			5	0.79	**1.00**	**1.00**	0.19	**0.03**	**0.04**	**0.04**	0.06
			6	0.90	**1.00**	0.99	0.20	**0.03**	**0.04**	**0.03**	0.13
			7	**0.80**	0.93	0.95	0.04	**0.04**	0.06	0.09	0.08
			8	0.72	**0.92**	0.90	0.26	**0.03**	**0.05**	**0.04**	0.08
200	4	0.35	1	0.79	**0.98**	0.97	0.04	**0.03**	**0.04**	**0.04**	0.13
			2	0.87	**0.98**	0.97	0.08	**0.03**	**0.03**	**0.03**	0.08
			3	0.82	**0.97**	0.90	0.16	**0.03**	**0.03**	0.06	0.12
			4	0.87	**0.99**	0.92	0.10	**0.03**	**0.02**	0.08	0.11
			5	0.79	**0.99**	**0.99**	0.10	**0.04**	**0.05**	**0.05**	0.08
			6	0.80	**1.00**	0.97	0.12	**0.03**	**0.04**	0.06	0.11
			7	0.79	**0.98**	0.92	0.09	**0.03**	**0.05**	0.07	0.06
			8	0.83	**0.99**	0.99	0.10	**0.05**	**0.05**	0.06	**0.05**
200	4	0.10	1	0.30	**0.40**	0.26	0.20	0.06	**0.03**	**0.05**	0.22
			2	**0.35**	0.50	0.27	0.12	**0.03**	0.10	**0.05**	0.18
			3	0.25	**0.29**	0.18	0.21	**0.04**	**0.01**	**0.04**	0.24
			4	0.37	**0.42**	0.17	0.19	**0.03**	**0.03**	**0.03**	0.17
			5	0.27	**0.51**	0.33	0.26	**0.05**	**0.02**	**0.05**	0.20
			6	0.31	**0.40**	0.27	0.24	**0.03**	**0.03**	**0.04**	0.17
			7	0.26	**0.42**	0.22	0.20	**0.04**	**0.04**	0.09	0.31
			8	0.31	**0.39**	0.32	0.23	**0.03**	**0.03**	**0.04**	0.18
500	10	0.65	1	**0.93**	0.89	0.88	0.09	**0.05**	0.06	0.06	**0.05**
			2	**0.99**	1.00	0.84	0.08	**0.02**	0.06	**0.05**	**0.05**
			3	**0.85**	0.26	0.61	0.12	**0.02**	**0.05**	0.18	0.10
			4	**0.99**	0.96	0.60	0.10	**0.04**	0.06	0.19	0.12
			5	0.89	**0.98**	0.96	0.16	**0.04**	**0.05**	**0.03**	0.10
			6	**0.99**	1.00	0.91	0.15	**0.04**	0.07	**0.02**	0.13
			7	**0.90**	0.61	0.68	0.09	**0.02**	0.07	0.12	0.07
			8	**0.79**	0.76	0.76	0.18	**0.03**	**0.04**	**0.05**	0.09
500	10	0.35	1	**0.89**	0.74	0.66	0.07	**0.02**	**0.02**	**0.02**	0.08
			2	**0.99**	0.99	0.69	0.09	**0.03**	0.06	**0.03**	0.11
			3	**0.88**	0.33	0.51	0.14	**0.04**	**0.05**	0.18	0.11
			4	**0.98**	0.91	0.48	0.12	**0.04**	0.08	0.20	0.10
			5	**0.91**	0.92	0.83	0.12	**0.04**	0.06	**0.04**	0.12
			6	**0.98**	1.00	0.75	0.09	**0.03**	0.08	**0.04**	0.11
			7	**0.89**	0.46	0.52	0.05	**0.03**	**0.03**	0.08	0.11
			8	**0.92**	0.78	0.74	0.10	**0.05**	0.06	0.06	0.07
500	10	0.10	1	**0.31**	–	0.06	0.12	**0.02**	–	**0.03**	0.10
			2	**0.45**	–	0.07	0.12	**0.03**	–	**0.03**	0.07
			3	**0.34**	–	0.18	0.16	**0.03**	–	0.19	0.14
			4	**0.45**	–	0.20	0.16	**0.02**	–	0.22	0.11
			5	0.33	–	**0.04**	0.12	0.06	–	**0.02**	0.14
			6	**0.45**	–	0.03	0.08	**0.05**	–	**0.01**	0.12
			7	**0.34**	–	0.12	0.09	**0.05**	–	0.09	0.15
			8	**0.34**	–	0.04	0.16	**0.03**	–	**0.05**	0.13

Table 3: Simulated power and type-I error of PKLM, Q, Little and JJ for $n = 200$, $p = 4$ and $n = 500$, $p = 10$. We use $r = 0.65$, 0.35 and 0.1. Cases $1 - 8$ describe different data distributions. The experiments were repeated 300 times and the parameter setting for PKLM described above was used.

				Power				Type-I Error			
n	p	r	case	PKLM	Q	Little	JJ	PKLM	Q	Little	JJ
500	20	0.65	1	**0.39**	–	0.36	0.06	**0.02**	–	**0.05**	0.09
			2	**0.91**	–	0.48	0.08	**0.03**	–	**0.05**	0.10
			3	**0.33**	–	0.49	0.20	**0.03**	–	0.24	0.11
			4	**0.90**	–	0.40	0.14	**0.04**	–	0.22	0.11
			5	0.32	–	**0.64**	0.14	**0.04**	–	**0.04**	0.08
			6	**0.93**	–	0.39	0.25	**0.04**	–	**0.01**	0.09
			7	**0.33**	–	0.37	0.07	**0.03**	–	0.09	0.10
			8	**0.23**	–	0.25	0.14	**0.04**	–	0.06	**0.03**
500	20	0.35	1	**0.45**	–	0.22	0.08	**0.03**	–	**0.04**	0.09
			2	**0.90**	–	0.22	0.09	**0.03**	–	**0.04**	0.08
			3	**0.43**	–	0.35	0.18	**0.02**	–	0.34	0.12
			4	**0.89**	–	0.33	0.20	**0.03**	–	0.31	0.15
			5	**0.46**	–	0.24	0.09	**0.02**	–	**0.02**	0.12
			6	**0.91**	–	0.14	0.14	**0.02**	–	**0.03**	0.10
			7	**0.41**	–	0.22	0.11	**0.02**	–	0.11	0.10
			8	**0.52**	–	0.18	0.08	**0.03**	–	**0.04**	0.07
500	20	0.10	1	**0.13**	–	0.00	0.14	**0.03**	–	**0.00**	0.10
			2	**0.24**	–	0.01	0.14	**0.04**	–	**0.01**	0.12
			3	0.08	–	0.21	0.16	0.06	–	0.22	0.10
			4	**0.26**	–	0.27	0.08	**0.04**	–	0.31	0.13
			5	**0.12**	–	0.00	0.10	**0.03**	–	**0.00**	0.19
			6	**0.19**	–	0.00	0.11	**0.05**	–	**0.00**	0.18
			7	**0.07**	–	0.08	0.12	**0.04**	–	0.07	0.12
			8	**0.07**	–	0.02	0.11	**0.04**	–	**0.00**	0.16
1000	40	0.65	1	**0.20**	–	0.00	0.09	**0.05**	–	**0.00**	0.15
			2	**0.95**	–	0.00	0.12	**0.03**	–	**0.00**	0.14
			3	**0.23**	–	0.00	0.29	**0.02**	–	**0.00**	0.17
			4	**0.94**	–	0.00	0.26	**0.05**	–	**0.00**	0.17
			5	**0.16**	–	0.00	0.30	**0.02**	–	**0.00**	0.19
			6	**0.97**	–	0.00	0.26	**0.02**	–	**0.00**	0.19
			7	**0.23**	–	0.00	0.11	**0.02**	–	**0.00**	0.10
			8	**0.13**	–	0.00	0.17	**0.03**	–	**0.00**	0.12
1000	40	0.35	1	**0.35**	–	0.00	0.12	**0.02**	–	**0.00**	0.11
			2	**0.97**	–	0.00	0.13	**0.05**	–	**0.00**	0.10
			3	**0.37**	–	0.00	0.30	**0.03**	–	**0.00**	0.30
			4	**0.96**	–	0.00	0.33	**0.04**	–	**0.00**	0.27
			5	**0.32**	–	0.00	0.14	**0.04**	–	**0.00**	0.11
			6	**0.98**	–	0.00	0.16	**0.03**	–	**0.00**	0.10
			7	**0.36**	–	0.00	0.11	**0.02**	–	**0.00**	0.08
			8	**0.30**	–	0.00	0.16	**0.02**	–	**0.00**	0.10
1000	40	0.10	1	**0.08**	–	0.00	0.15	**0.02**	–	**0.00**	0.12
			2	**0.32**	–	0.00	0.12	**0.02**	–	**0.00**	0.10
			3	**0.06**	–	0.00	0.13	**0.05**	–	**0.00**	0.20
			4	**0.25**	–	0.00	0.25	**0.03**	–	**0.00**	0.28
			5	**0.08**	–	0.00	0.11	**0.03**	–	**0.00**	0.09
			6	**0.27**	–	0.00	0.09	**0.04**	–	**0.00**	0.11
			7	**0.07**	–	0.00	0.16	**0.03**	–	**0.00**	0.13
			8	**0.07**	–	0.00	0.15	**0.04**	–	**0.00**	0.08

Table 4: Simulated power and type-I error of PKLM, Q, Little and JJ for $n = 500$, $p = 20$ and $n = 1000$, $p = 40$. We use $r = 0.65$, 0.35 and 0.1. Cases $1 - 8$ describe different data distributions. The experiments were repeated 300 times and the parameter setting for PKLM described above was used.

			Power				Type-I Error			
dataset	n	p	PKLM	Q	Little	JJ	PKLM	Q	Little	JJ
iris	150	4	0.41	**0.91**	0.84	0.27	**0.03**	**0.04**	**0.03**	0.16
blood.transfusion	748	4	0.48	0.97	**1.00**	NA	**0.01**	0.06	**0.04**	NA
airfoil	1503	6	**0.92**	0.13	0.17	0.09	**0.02**	**0.03**	0.06	0.42
seeds	210	7	0.64	**0.74**	0.57	0.24	**0.05**	**0.02**	**0.02**	0.10
yacht	308	7	0.60	0.56	**0.76**	0.24	**0.03**	0.07	**0.05**	0.24
yeast	1484	8	**0.82**	0.52	0.15	0.14	**0.05**	0.06	0.23	0.85
glass	214	9	0.10	0.02	**0.20**	0.20	**0.01**	**0.00**	**0.03**	0.33
concrete.compression	1030	9	0.64	0.48	**0.81**	0.47	**0.04**	**0.04**	**0.05**	0.41
wine.quality.red	1599	11	**0.81**	–	0.72	0.80	**0.04**	–	0.15	0.52
wine.quality.white	4898	11	**0.98**	–	0.96	0.87	**0.04**	–	0.10	0.79
planning.relax	182	12	**0.29**	–	0.20	0.14	**0.00**	–	**0.00**	NA
climate.model.crashes	540	19	**0.18**	–	0.22	0.47	**0.00**	–	**0.00**	NA
ionosphere	351	32	**0.45**	–	0.97	0.18	**0.00**	–	0.06	NA

Table 5: Simulated power and level of PKLM, Q, Little and JJ for 13 real datasets. We use $p_{miss} = 0.3$. The experiments were repeated 300 times and the parameter setting for PKLM described above was used. The NAs for some values of the JJ-test indicate that the test was not computable in any of the 300 repetitions due to not enough observations in enough usable missingness groups.

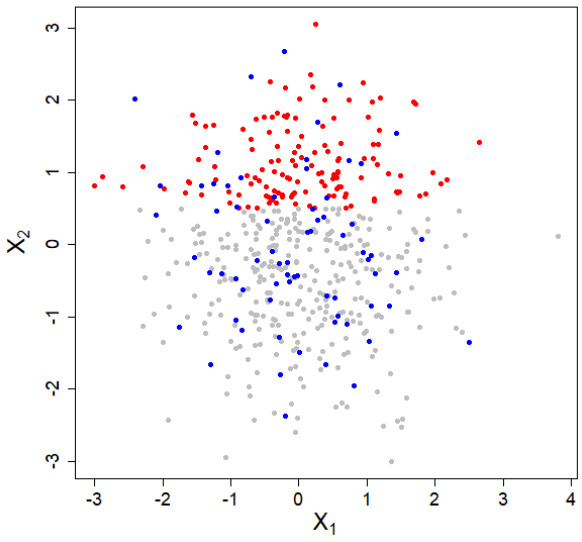


Figure 3: X_1 and X_2 of the fully observed data in the simulated example of Section 6. In red: Points with missing values in X_1, in blue: points with missing values in X_2. The blue points are randomly scattered, independently of the value of X_1, while in the red points, there is a visible trend towards having more missing values in X_1 for higher values of variable X_2.

matrix $\mathbf{M}$

$$\hat{U}^{-k}_{\sigma_\ell} = \frac{1}{|\mathcal{P}_{-k}|} \sum_{j \in \mathcal{P}_{-k}} \hat{U}^{(A_j,B_j)}_{\sigma_\ell},$$

to obtain the p-value as in (11). This "partial" p-value is valid and corresponds to the effect of removing the patterns induced by variable k. Indeed, assume the difference in the distribution of two patterns stems from a variable j alone. If $j \in B$, a perfect classifier will be able to reliably differentiate the two, leading to a high value for $\hat{U}^{-k}$ relative to the permutation values. If j is not forming the labels, we will not test these two classes against each other and thus not be able spot this difference. As such, we might expect to see a high p-value for $\hat{U}^{-j}$, when variable j is removed, but a tendency to low p-values for $\hat{U}^{-k}$, $k \neq j$.

We illustrate the usefulness of partial p-values with an example. Let $C_{-k} = \{1, \ldots, p\} \backslash \{k\}$. We assume $\mathbf{X}_{\bullet, C_{-k}}$ has a MCAR missingness structure, in particular, we simulate below the MCAR mechanism described in Section 5.1 with $r = 0.65$. Let $k = 1$ and assume that this first column of observations $\mathbf{X}_{\bullet,1}$ has missingness depending on the observed values of $\mathbf{X}_{\bullet,2}$. For instance, each value is missing if the mean of the corresponding row $\mathbf{X}_{j,2}$ is larger than 0.5. In this simple example $\mathbf{X}$ is MAR, but $\mathbf{X}_{\bullet, C_{-1}}$ is MCAR. We simulate this example, with $p = 4$ and $n = 500$, $\mathbf{X}_{i,\bullet}$ being independent standard Gaussian and the MAR/MCAR mechanism as described above. The first two fully observed components, X_1 and X_2, are shown in Figure 3. As before, we set `num.trees.per.proj`=200 and use 100 projections. In this example, we are only able to spot any difference when $j = 1$ is used to build the labels.

Our test reliably delivers small p-values ($\leqslant 0.05$) for the three partial tests based on projections potentially including variable 1, i.e., sets of projections $\mathcal{P}_{-2}$, $\mathcal{P}_{-3}$, and $\mathcal{P}_{-4}$ and a high p-value for the partial test based on $\mathcal{P}_{-1}$. Thus in this sense, the test detects that the main culprit of the MAR mechanism lies in the first variable.

7 Concluding Remarks

In this paper we presented the powerful, flexible and easy-to-use PKLM-test for the MCAR assumption on the missingness mechanism of a dataset. We proved the validity of the p-value of the test and showed its power over a wide range of distributions. We also provided an extension allowing to do partial tests, that may shed light on the source of the violation of the MCAR assumption. Naturally, with some slight adaptations the test can be used as a general test of homogeneity of G different groups in the sense that it tests whether G different groups have the same distribution.

1 Proofs

Proposition 29. *Hypothesis testing problem* (3) *is equivalent to* (2).

Proof We first show H_0 of (2) implies H_0 of (3). Let A, B be arbitrary. If they are such that there is only one label, there is nothing to test, so we may assume to have $|G(A,B)| \geqslant 2$ patterns in $\mathbf{X}_{\mathcal{N}_A,A}$. This means that $A \subset \boldsymbol{o}_{ij}$ for *all* patterns $i, j \in G(A,B)$. This simply follows because, by construction, each of the $|G(A,B)|$ patterns in $\mathbf{X}_{\mathcal{N}_A,A}$ has the elements in A fully observed. But since by assumption for all $i, j \in \{1,\dots,G\}$, $F_{i,\boldsymbol{o}_{ij}} = F_{j,\boldsymbol{o}_{ij}}$ and $A \subset \boldsymbol{o}_{ij}$, this immediately implies that $F_{i,A} = F_{j,A}$ for all $i, j \in \{1,\dots,G\}$ and thus $F_{g,A} = \sum_{j \in G(A,B)\setminus g} \omega_j^g F_{j,A}$. Since A, B were arbitrary, one direction follows.

We now show that H_0 of (3) implies H_0 of (2). The proof is based on the following claim: Consider G arbitrary distribution functions $F_1,\dots,F_G$ and weights $(\omega_j^g)_{j=1}^{G-1}$, $j = 1,\dots,G$ such that $\sum_{j=1}^{G-1} \omega_j^g = 1$ for all j. Then

$$F_g = \sum_{j \in \{1,\dots,G\}\setminus g} \omega_j^g F_j,\ \forall g \in \{1,\dots,G\} \implies F_i = F_j,\ \forall i \neq j \in \{1,\dots,G\}. \tag{13}$$

We prove the implication by induction: Consider first $G = 3$. Assuming the LHS of (13) and plugging the equation for F_2 into the equation for F_1, we obtain:

$$\begin{aligned} F_1 &= w_2^1 w_1^2 F_1 + w_2^1 w_3^2 F_3 + w_3^1 F_3 \\ &= w_2^1 w_1^2 F_1 + (w_2^1 w_3^2 + w_3^1) F_3, \end{aligned}$$

which implies $(1 - w_1^2 w_2^1) F_1 = (w_2^1 w_3^2 + w_3^1) F_3$. Since

$$1 = w_2^1 + w_3^1 = w_2^1 (w_3^2 + w_1^2) + w_3^1 = w_2^1 w_3^2 + w_2^1 w_1^2 + w_3^1,$$

we have the equality $(1 - w_1^2 w_2^1) = (w_2^1 w_3^2 + w_3^1)$ and thus $F_1 = F_3$. Plugging this back into the equivalent equation for F_2, we obtain $F_1 = F_2 = F_3$. Now assume (13) is true for G distributions $F_1,\dots,F_G$ and we now would like to prove it for $G+1$. Assume wlog that the weight of F_2 in the equation of F_1 is nonzero (there will always be at least one such distribution $F_2,\dots,F_G$). Using the same trick as above, we may plug say the equation for F_2 into F_1, thereby reducing the number of equations/distributions to G. By the induction assumption this implies that $F_1 = F_3 = \ldots = F_G$. But immediately this also implies that $F_2 = F_1$ and implies (13). With this result we can now proof the that H_0 of (3) implies H_0 of (2).

Take two arbitrary groups i, j and $A = \boldsymbol{o}_{ij}$ and take $B = A^c$. To ease notation we just wlog take $i = 1$ and $j = 2$. Then $A = \boldsymbol{o}_{12}$ contains the dimensions for which patterns 1 and 2 have fully observed values. Thus, observations in $\mathbf{X}_{\mathcal{N}_A,A}$ contain draws from $F_{1,\boldsymbol{o}_{12}}$ and $F_{2,\boldsymbol{o}_{12}}$. Since by assumption

$$H_0 : F_{g,A} = \sum_{j \in G(A,B)\setminus g} \omega_j^g F_{j,A}\ , \forall g \in G(A,B), \tag{14}$$

it follows by (13), that $F_{i,A} = F_{j,A}$ for all $i, j \in G(A,B)$ and thus in particular, $F_{1,A} = F_{2,A}$. Since we will have $A = \boldsymbol{o}_{ij}$ for all groups $i \neq j$, H_0 of (2) holds. ■

Lemma .11. *The logarithm of the density ratio for testing (5) is given by* $U_g^{(A,B)}$.

Proof Based on the definitions of $p_g^{(A,B)}(x)$, $f_g^{(A,B)}(x)$ and $\pi_g^{(A,B)}$ we obtain by Bayes Rule,

$$p_g^{(A,B)}(x) = \frac{f_g^{(A,B)}(x)\pi_g^{(A,B)}}{\sum_{j\in G(A,B)} \pi_j^{(A,B)} f_j^{(A,B)}(x)}, \tag{15}$$

assuming the existence of densities f_g of distributions F_g for each $g \in G(A,B)$. Following the same steps as in [24], we get that the logarithm of the (joint) density ratio for testing H_0 vs H_1 of (5), given by

$$\log \frac{f_g^{(A,B)}(x)(1-\pi_g^{(A,B)})}{\sum_{j\in G(A,B)\setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(x)}. \tag{16}$$

We reformulate the fraction in (16) in terms of $p_g^{(A,B)}$, starting from (15):

$$\begin{aligned} p_g^{(A,B)}(x) \sum_{j\in G(A,B)\setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(x) &= (\pi_g^{(A,B)} - p_g^{(A,B)}(x)\pi_g^{(A,B)}) f_g^{(A,B)}(x) \\ &= \pi_g^{(A,B)}(1 - p_g^{(A,B)}(x)) f_g^{(A,B)}(x). \end{aligned}$$

Thus, the inside of the logarithm of (16) is given by the following function of $p_g^{(A,B)}$:

$$\frac{f_g^{(A,B)}(x)(1-\pi_g^{(A,B)})}{\sum_{j\in G(A,B)\setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(x)} = \frac{1-\pi_g^{(A,B)}}{\pi_g^{(A,B)}} \frac{p_g^{(A,B)}(x)}{1-p_g^{(A,B)}(x)}.$$

■

Lemma .12. $U_g^{(A,B)}$ *converges in probability to the Kullback-Leibler Divergence between* $f_g^{(A,B)}$ *and the mixture of the other densities:*

$$U_g^{(A,B)} \to \mathbb{E}_{f_g}\left[\log \frac{f_g^{(A,B)}(X)(1-\pi_g^{(A,B)})}{\sum_{j\in G(A,B)\setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(X)}\right],$$

as n_g *and* $\sum_{j\in\{1,\dots,G\}\setminus g} n_j \to \infty$ *and* $n_g/n \to \pi_g^{(A,B)} \in (0,1)$.

Proof From the proof of Lemma .11, we know that $U_g^{(A,B)}$ can be rewritten as

$$\begin{aligned} U_g^{(A,B)} &:= \frac{1}{|S_{f_g^{(A,B)}}|} \sum_{i\in S_{f_g^{(A,B)}}} \left(\log \frac{p_g^{(A,B)}(x_i)}{1-p_g^{(A,B)}(x_i)} - \log \frac{\pi_g^{(A,B)}}{1-\pi_g^{(A,B)}}\right) \\ &= \frac{1}{n_g} \sum_{i\in S_{f_g^{(A,B)}}} \log \frac{f_g^{(A,B)}(x_i)(1-\pi_g^{(A,B)})}{\sum_{j\in G(A,B)\setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(x_i)}. \end{aligned} \tag{17}$$

Since $n_g/n \to \pi_g^{(A,B)} \in (0,1)$ and the x_i are i.i.d., the result follows from the law of large numbers. ■

Proposition 30. *Under H_0 in* (1)*, and Z as defined in* (11)*, it holds for all $z \in [0,1]$ that*

$$\mathbb{P}(Z \leqslant z) \leqslant z. \tag{12}$$

Proof

Let $\mathbf{A} = (A_1, \ldots, A_N)$ and $\mathbf{B} = (B_1, \ldots, B_N)$ be two sets of N projections. Let $G_1, \ldots G_{L^*}$ be all possible permutations of the rows of the missingness matrix $\mathbf{M}$, such that

$$G_\ell(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B}) = (\mathbf{X}^*, \mathbf{M}_{\sigma_\ell}, \mathbf{A}, \mathbf{B}),$$

for $\ell = 1, \ldots, L^*$. Note that, since we are only considering fully observed observations for all projections in $\mathbf{A}$, $\hat{U}$, a function of $(\mathbf{X}, \mathbf{M}, \mathbf{A}, \mathbf{B})$, is indeed a function of $(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B})$, while $\hat{U}_{\sigma_\ell}$ is a function of $G_\ell(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B})$. It also holds that under the null, that is under MCAR, that

$$(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B}) \overset{D}{=} (\mathbf{X}^*, \mathbf{M}_{\sigma_\ell}, \mathbf{A}, \mathbf{B}) = G_\ell(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B}) \;\; \forall \ell = 1, \ldots L^*. \tag{18}$$

This is true because, under MCAR, $\mathbf{M}$ and $\mathbf{X}^*$ are independent. Since by the i.i.d. assumption also $\mathbf{M}_{\sigma_\ell} \overset{D}{=} \mathbf{M}$ for all $\ell = 1, \ldots, L^*$ and since $\mathbf{A}$, $\mathbf{B}$ are also independent of $\mathbf{M}$, (18) follows. As outlined for example in [70], this implies that under H_0,

$$\mathbb{P}(Z \leqslant z \mid \mathbf{A}, \mathbf{B}) \leqslant z.$$

Integrating over $(\mathbf{A}, \mathbf{B})$, results in (12). ■

2 Additional Details and Computation Times

Here we provide more implementation details, discuss the complexity calculations in Table 1 and show computation times of the different tests in the experiments.

Numerical truncation. In order to avoid numerical issues when calculating the density ratio with Expression (6) or the log thereof, if we get predicted probabilities $\hat{p}_A$ close to 0 or 1, we apply the following truncation function to $\hat{p}_A$:

$$p(x) = \min(\max(x, 10^{-9}), 1 - 10^{-9}).$$

Hyperparameter Selection. Generally speaking, it holds that "the more the better", certainly for the parameters N, L and `num.trees.per.proj`. As such, the choice of those three parameters depends mostly on the computational power available to the user. For `size.resp.set`, this is not quite as clear, though we found a value of two to work well in most situations.

PLKM Test. We first consider the complexity of one Random Forest, which is in this case

$$\texttt{num.tree} \cdot pn\log(n).$$

Note that this includes the calculation of $\hat{p}$ on the test sample through the OOB-error. In total we do this `num.proj` times. However, we consider `num.tree` and `num.proj` independent of n and p and thus treating it as constant. In this case we end up with $pn\log(n)$. Finally we need to calculate the statistics U and repeat this number of calculations a fixed number of times. This would add a factor Bn, where again we assume that B does not grow with n and p. As this is neglible compared to $pn\log(n)$, the complexity is given as $O(pn\log(n))$

Q-test. The Q-test compares all groups leading to a complexity of G^2 to compare each group with any other. Additionally, the statistic used is an MMD type, so the complexity is $(n_1+n_2)^2$, where n_1, n_2 are the respective group sizes. The group size can be at worst n/G, which together results in $O(n^2)$. The bootstrap on the other hand can also be ignored, as it simply results in a constant factor multiplied to n^2.

JJ and Little-test. Both JJ- and Little-test rely on covariance estimation which scales as np^2. This gives the $O(np^2)$ complexity for the Little-test. For the JJ-test one also needs an ordering operation to obtain the test statistics, with complexity $n\log(n)$, which results in overall complexity $O(n(p^2+\log(n)))$.

As mentioned above, Table 1 just shows how the complexity scales in n and p and, in case of our test, treats the number of projections as constants. One might argue that the number of projections should be a function of p as well. Similarly, for "small" p and small number of groups G, the Q-test can be faster than ours. Still the complexities provide a good illustration of how quickly the Q-test can become infeasible, when the number of groups (often a function of p) and/or the number of observation increases.

3 Example of [190]

[190] study settings where group means and variances are approximately equal across missingness patterns, such that MCAR tests based on differences in means and variances, such as the Little-test, have no power. We study one such example here: Let $p = 2$ and (Z_1, Z_2) be jointly multivariate normal with correlation zero and let $X_1 = Z_1$ and

$$X_2 = 0.5Z_1 + (1-0.25)^{1/2}Z_2.$$

We set X_2 to `NA` if

$$X_1 \in (-\infty, -1.932] \cup (-0.314, 0.314] \cup (1.932, \infty).$$

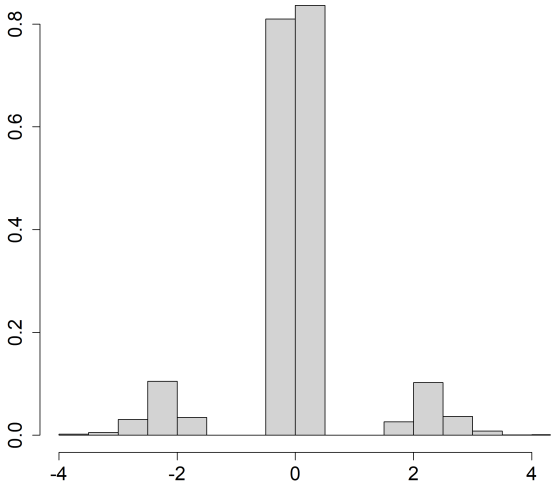


Figure 4: Histogram with relative frequencies of X_1 if the corresponding X_2 is `NA`.

This corresponds to around 30% missing values. Figure 4 displays a histogram, plotting all observations of X_1 with X_2 missing for a simulation of $n = 10'000$. This corresponds to the MAR example used in [190, Section 3] and we refer to their paper for more details.

We simulate the above distribution for $n = 1000$ and run our PKLM-test with the same parameters as described in Section 5.1. Though the deviation from MCAR cannot be detected through the first two moments in this example, our test reliably reaches a power of 1.

Paper

D

Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression.

D. Ćevid, L. Michel, J. Näf, P. Bühlmann, N. Meinshausen

Journal of Machine Learning Research.

Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression

Domagoj Ćevid Loris Michel Jeffrey Näf
Peter Bühlmann Nicolai Meinshausen

January 20, 2023

Abstract

Random Forest [15] is a successful and widely used regression and classification algorithm. Part of its appeal and reason for its versatility is its (implicit) construction of a kernel-type weighting function on training data, which can also be used for targets other than the original mean estimation. We propose a novel forest construction for multivariate responses based on their joint conditional distribution, independent of the estimation target and the data model. It uses a new splitting criterion based on the MMD distributional metric, which is suitable for detecting heterogeneity in multivariate distributions. The induced weights define an estimate of the full conditional distribution, which in turn can be used for arbitrary and potentially complicated targets of interest. The method is very versatile and convenient to use, as we illustrate on a wide range of examples. The code is available as `Python` and `R` packages **`drf`**.

Keywords. causality, distributional regression, fairness, Maximal Mean Discrepancy, Random Forests, two-sample testing

1 Introduction

In practice, one often encounters heterogeneous data, whose distribution is not constant, but depends on certain covariates. For example, data can be collected from several different sources, its distribution might differ across certain subpopulations or it could even change with time, etc. Inferring valid conclusions about a certain target of interest from such data can be very challenging as many different aspects of the distribution could potentially change. As an example, in medical studies, the effectiveness of a certain treatment might not be constant throughout the

population but depend on certain patient characteristics such as age, race, gender, or medical history. Another issue could be that different patient groups were not equally likely to receive the same treatment in the observed data.

Obviously, pooling all available data together can result in invalid conclusions. On the other hand, if for a given test point of interest one only considers similar training data points, i.e. a small homogeneous subpopulation, one may end up with too few samples for accurate statistical estimation. In this paper, we propose a method based on the Random Forest algorithm [15] which in a data-adaptive way determines for any given test point which training data points are relevant for it. This in turn can be used for drawing valid conclusions or for accurately estimating any quantity of interest.

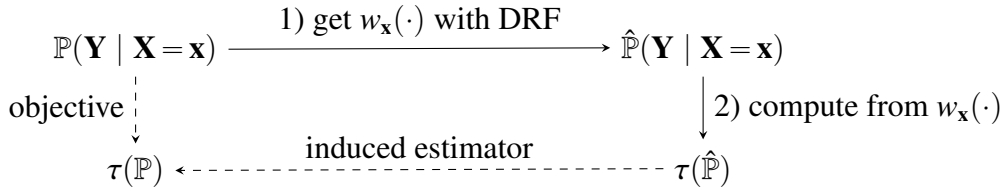
Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d) \in \mathbb{R}^d$ be a multivariate random variable representing the data of interest, but whose joint distribution is heterogeneous and depends on some subset of a potentially large number of covariates $\mathbf{X} = (X_1, X_2, \ldots, X_p) \in \mathbb{R}^p$. Throughout the paper, vector quantities are denoted in bold. We aim to estimate a certain target object $\tau(\mathbf{x})$ that depends on the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) = \mathbb{P}(\mathbf{Y} \mid X_1 = x_1, \ldots, X_p = x_p)$, where $\mathbf{x} = (x_1, \ldots, x_p)$ is an arbitrary point in $\mathbb{R}^p$. The estimation target $\tau(\mathbf{x})$ can range from simple quantities, such as the conditional expectations $\mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}]$ [15] or quantiles $Q_\alpha[f(\mathbf{Y}) \mid \mathbf{X}]$ [117] for some function $f : \mathbb{R}^d \to \mathbb{R}$, to some more complicated aspects of the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, such as conditional copulas or conditional independence measures. Given the observed data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, the most straightforward way of estimating $\tau(\mathbf{x})$ nonparametrically would be to consider only the data points in some neighborhood $\mathcal{N}_\mathbf{x}$ around $\mathbf{x}$, e.g. by considering the k nearest neighbors according to some metric. However, such methods typically suffer from the curse of dimensionality even when p is only moderately large: for a reasonably small neighborhood, such that the distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} \in \mathcal{N}_\mathbf{x})$ is close to the distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, the number of training data points contained in $\mathcal{N}_\mathbf{x}$ will be very small, thus making the accurate estimation of the target $\tau(\mathbf{x})$ difficult. The same phenomenon occurs with other methods which locally weight the training observations such as kernel methods [155], local MLE [51] or weighted regression [33] even for the relatively simple problem of estimating the conditional mean $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ for fairly small p. For that reason, more importance should be given to the training data points $(\mathbf{x}_i, \mathbf{y}_i)$ for which the response distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)$ at point $\mathbf{x}_i$ is similar to the target distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, even if $\mathbf{x}_i$ is not necessarily close to $\mathbf{x}$ in every component.

In this paper, we propose the Distributional Random Forest (DRF) algorithm which estimates the multivariate conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ in a locally adaptive fashion. This is done by repeatedly dividing the data points in the spirit of the Random Forest algorithm [15]: at each step, we split the data points into two groups based on some feature X_j in such a way that the distribution of $\mathbf{Y}$ for which $X_j \leqslant l$, for some level l, differs the most compared to the

distribution of $\mathbf{Y}$ when $X_j > l$, according to some distributional metric. One can use any multivariate two-sample test statistic, provided it can detect a wide variety of distributional changes. As the default choice, we propose a criterion based on the Maximal Mean Discrepancy (MMD) statistic [62] with many interesting properties. This splitting procedure partitions the data such that the distribution of the multivariate response $\mathbf{Y}$ in the resulting leaf nodes is as homogeneous as possible, thus defining neighborhoods of relevant training data points for every $\mathbf{x}$. Repeating this many times with randomization induces a weighting function $w_{\mathbf{x}}(\mathbf{x}_i)$ as in [104, 105], described in detail in Section 2, which quantifies the relevance of each training data point $\mathbf{x}_i$ for a given test point $\mathbf{x}$. The conditional distribution is then estimated by an empirical distribution determined by these weights [117]. This construction is data-adaptive as it assigns more weight to the training points $\mathbf{x}_i$ that are closer to the test point $\mathbf{x}$ in the components which are more relevant for the distribution of $\mathbf{Y}$.

Our forest construction does not depend on the estimation target $\tau(\mathbf{x})$, but it rather estimates the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ directly and the induced forest weights can be used to estimate $\tau(\mathbf{x})$ in a second step. This approach has several advantages. First, only one DRF fit is required to obtain estimates of many different targets, which has a big computational advantage. Furthermore, since those estimates are obtained from the same forest fit, they are mutually compatible. For example, if the conditional correlation matrix $\{\mathrm{Cor}(Y_i,\ Y_j \mid \mathbf{X}=\mathbf{x})\}_{i,j=1}^{d}$ is estimated componentwise using some other method, the resulting matrix might not be positive semidefinite, and as another example, the CDF estimates $\hat{\mathbb{P}}(\mathbf{Y} \leqslant \mathbf{y} \mid \mathbf{X}=\mathbf{x})$ might not be monotone in $\mathbf{y}$, see Figure 6. Finally, it could be extremely difficult to tailor forest construction to some complex targets $\tau(\mathbf{x})$. The induced weighting function can thus be used not only for obtaining simple distributional aspects such as, for example, the conditional quantiles, conditional correlations, or joint conditional probability statements, but also to obtain more complex objectives, such as conditional independence tests [193], heterogeneous regression (see also Section 4.4 for more details) [96, 181] or semiparametric estimation by fitting a parametric model for $\mathbf{Y}$, having nonparametrically adjusted for $\mathbf{X}$ [12]. Representation of the conditional distribution via the weighting function has a great potential for applications in causality such as causal effect estimation or as a way of implementing do-calculus [133] for finite samples, as we discuss in Section 4.4.

Therefore, DRF is used in two steps: in the first step, we obtain the weighting function $w_{\mathbf{x}}(\cdot)$ describing the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ in a target- and model-free way, which is then used as an input for the second step. Even if the method used in the second step does not directly support weighting of the training data points, one can easily resample the data set with the sampling probabilities equal to $\{w_{\mathbf{x}}(\mathbf{x}_i)\}_{i=1}^{n}$. This two-step approach is visualized in the following diagram:



1.1 Related work and our contribution

Several adaptations of the Random Forest algorithm have been proposed for targets beyond the original one of the univariate conditional mean $\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$: for survival analysis [73], conditional quantiles [117], density estimation [136], CDF estimation [74] or heterogeneous treatment effects [181]. Almost all such methods use the weights induced by the forest, as described in Section 2, rather than averaging the estimates obtained per tree. This view of Random Forests as a powerful adaptive nearest neighbor method is well known and dates back to [104, 105]. It was first used for targets beyond the conditional mean in [117], where the original forest construction with univariate Y was used [15]. However, the univariate response setting considered there severely restricts the number of interesting targets $\tau(\mathbf{x})$ and DRF can thus be viewed as an important generalization of this approach to the multivariate setting.

In order to be able to perform certain tasks or to achieve a better accuracy, many forest-based methods adapt the forest construction by using a custom splitting criterion tailored to their specific target, instead of relying on the standard CART criterion. In [192] and [74], a parametric model for the response $\mathbf{Y} \mid \mathbf{X} = \mathbf{x} \sim f(\theta(\mathbf{x}), \cdot)$ is assumed and recursive splitting is performed based on a permutation test which uses the user-provided score functions. Similarly, [4] estimate certain univariate targets for which there exist corresponding score functions defining the local estimating equations. The data is split so that the estimates of the target in resulting child nodes differ the most. This is different, though, to the target-free splitting criterion of DRF, which splits so that the distribution of $\mathbf{Y}$ in child nodes is as different as possible.

Since the splitting step is extensively used in the algorithm, its complexity is crucial for the overall computational efficiency of the method, and one often needs to resort to approximating the splitting criterion [136, 4] to obtain good computational run time. We propose a splitting criterion based on a fast random approximation of the MMD statistic [63, 195], which is commonly used in practice for two-sample testing as it is able to detect any change in the multivariate distribution of $\mathbf{Y}$ with good power [62]. DRF with the MMD splitting criterion also has interesting theoretical properties as shown in Section 3 below.

The multivariate response case has not received much attention in the Random Forest literature. Most of the existing forest-based methods focus on either a univariate response Y or on a certain univariate target $\tau(\mathbf{x})$. One interesting line of work considers density estimation [136] and uses aggregation of the CART criteria for different response transformations. Another ap-

proach [92, 152, 77] is based on aggregating standard univariate CART splitting criteria for $Y_1, \ldots, Y_d$ and targets only the conditional mean of the responses, a task which could also be solved by separate regression fits for each Y_i. In order to capture any change in the distribution of the multivariate response $\mathbf{Y}$, one needs to not only consider the marginal distributions for each component Y_i, but also to determine whether their dependence structure changes, see e.g. Figure 8.

There is an increasing number of methods that nonparametrically estimate the joint multivariate conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ in the statistics and machine learning literature. In addition to a few simple classical methods such as k-nearest neighbors and kernel regression, there exist methods based on normalizing flows such as Inverse Autoregressive Flow [91] or Masked Autoregressive Flow [131] and also conditional variants of several popular generative models such as Conditional Generative Adversarial Networks [125] or Conditional Variational Autoencoder [160]. The focus of these methods is more on the settings with large response dimension d and small covariate dimension p, such as image or text generation. Another interesting and related line of research focuses on estimating the conditional mean embedding (CME), as described e.g., in [162, 126, 161, 132], rather than estimating the conditional distribution directly. CMEs generalize the concept of embedding (marginal) probability distributions into a Reproducing Kernel Hilbert Space (RKHS) to the conditional case. Interestingly, DRF with the MMD-based splitting criterion can also be viewed as a method for estimating the CME, as discussed in Section 3 below. This viewpoint provides a natural connection between the Random Forest and kernel embedding literature. A comparison of DRF with the methods for distributional estimation listed above can be found in Section 4.1.

Our contribution, resulting in the proposal of the Distributional Random Forest (DRF), can be summarized as follows: First, we introduce the idea of forest construction based on sequential multivariate two-sample test statistics. It does not depend on a particular estimation target and is completely nonparametric, which makes its implementation and usage very simple and universal. Not only does it not require additional user input such as the log-likelihoods or score functions, but it can be used even for complicated targets for which there is no obvious forest construction. Furthermore, it has a computational advantage as only a single forest fit is needed for producing estimates of many different targets that are additionally compatible with each other. Second, we propose an MMD-based splitting criterion with good statistical and computational properties, for which we also derive interesting theoretical results in Section 3. It underpins our implementation, which we provide as `R` and `Python` packages `drf`. Finally, we show on a broad range of examples in Section 4 how many different statistical estimation problems, some of which not being easily tractable by existing forest-based methods, can be cast to our framework, thus illustrating the usefulness and versatility of DRF.

2 Method

In this section we describe the details of the Distributional Random Forest (DRF) algorithm. We closely follow the implementations of the `grf` [4] and `ranger` [187] `R`-packages. A detailed description of the method and its implementation and the corresponding pseudocode can be found in the Appendix 1.

2.1 Forest Building

The trees are grown recursively in a model-free and target-free way as follows: For every parent node P, we determine how to best split it into two child nodes of the form $C_L = \{X_j \leqslant l\}$ and $C_R = \{X_j > l\}$, where the variable X_j is one of the randomly chosen splitting candidates and l denotes its level based on which we perform the splitting. The split is chosen such that we maximize a certain (multivariate) two-sample test statistic

$$\mathcal{D}\left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\}\right), \tag{1}$$

which measures the difference of the empirical distributions of the data $\mathbf{Y}$ in the two resulting child nodes C_L and C_R. Therefore, in each step we select the candidate predictor X_j which seems to affect the distribution of $\mathbf{Y}$ the most, as measured by the metric $\mathcal{D}(\cdot,\cdot)$. Intuitively, in this way we ensure that the distribution of the data points in every leaf of the resulting tree is as homogeneous as possible, which helps mitigate the bias caused by pooling the heterogeneous data together. A related idea can be found in GRF [4], where one attempts to split the data so that the resulting estimates $\hat{\tau}_L$ and $\hat{\tau}_R$, obtained respectively from data points in C_L and C_R, differ the most:

$$\frac{n_L n_R}{n_P^2}\left(\hat{\tau}_L - \hat{\tau}_R\right)^2, \tag{2}$$

where we write $n_P = |\{i \mid \mathbf{x}_i \in P\}|$ and n_L, n_R are defined analogously.

One could construct the forest using any metric $\mathcal{D}(\cdot,\cdot)$ for empirical distributions. However, in order to have a good accuracy of the overall method, the corresponding two-sample test using $\mathcal{D}(\cdot,\cdot)$ needs to have a good power for detecting any kind of change in distribution, which is a difficult task in general, especially for multivariate data [6, 169]. Another very important aspect of the choice of distributional metric $\mathcal{D}(\cdot,\cdot)$ is the computational efficiency; one needs to be able to sequentially compute the values of $\mathcal{D}\left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\}\right)$ for every possible split very fast for the overall algorithm to be computationally feasible, even for moderately large data sets. Below, we propose a splitting criterion based on the MMD two-sample test statistic [62] which has both good statistical and computational properties.

In contrast to other forest-based methods, we do not use any information about our estimation target τ in order to find the best split of the data, which comes with a certain trade-off.

On one hand, it is sensible that tailoring the splitting criterion to the target should improve the estimation accuracy; for example, some predictors might affect the conditional distribution of $\mathbf{Y}$, but not necessarily the estimation target τ and splitting on such predictors unnecessarily reduces the number of training points used for estimating τ. On the other hand, our approach has multiple benefits: it is easier to use as it does not require any user input such as the likelihood or score functions and it can also be used for very complicated targets for which one could not easily adapt the splitting criterion. Furthermore, only one DRF fit is necessary for producing estimates of many different targets, which has both computational advantage and the practical advantage that the resulting estimates are mutually compatible (see e.g. Figure 5).

Interestingly, sometimes it could even be beneficial to split based on a predictor which does not affect the target of estimation, but which affects the conditional distribution. This is illustrated by the following toy example. Suppose that for a bivariate response (Y_1, Y_2) we are interested in estimating the slope of the linear regression of Y_2 on Y_1 conditionally on $p = 30$ predictors $\mathbf{X}$, i.e. our target is $\tau(\mathbf{x}) = \mathrm{Cov}(Y_1, Y_2 \mid \mathbf{X} = \mathbf{x})/\mathrm{Var}(Y_1 \mid \mathbf{X} = \mathbf{x})$. This is one of the main use cases for GRF and its variant which estimates this target is called Causal Forest [181, 4]. Let us assume that the data has the following distribution:

$$\mathbb{P}\left(\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \,\middle|\, \mathbf{X} = \mathbf{x}\right) \sim N\left(\begin{bmatrix} x_1 \\ x_1 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}\right) \qquad \mathbf{X} \sim N(\mathbf{0}, I_p), \tag{3}$$

i.e. X_1 affects only the mean of the responses, while the other $p-1$ predictors have no effect. In Figure 1 we illustrate the distribution of the data when $n = 300, p = 30, \sigma = 0.2$, together with the DRF and GRF splitting criteria. The true value of the target is $\tau(\mathbf{x}) = 0$, but when σ is not too big, the slope estimates $\hat{\tau}$ on pooled data will be closer to 1. Therefore, the difference of $\hat{\tau}_L$ and $\hat{\tau}_R$ between the induced slope estimates for a candidate split, which is used for splitting criterion (2) of GRF, might not be large enough for us to decide to split on X_1, or the resulting split might be too unbalanced. This results in worse forest estimates for this toy example, see Figure 1.

2.2 Weighting Function

Having constructed our forest, just as the standard Random Forest [15] can be viewed as the weighted nearest neighbor method [104], we can use the induced weighting function to estimate the conditional distribution at any given test point $\mathbf{x}$ and thus any other quantity of interest $\tau(\mathbf{x})$. This approach is commonly used in various forest-based methods for obtaining predictions, see e.g., [74, 136, 4].

Suppose that we have built N trees $\mathcal{T}_1, \ldots, \mathcal{T}_N$. Let $\mathcal{L}_k(\mathbf{x})$ be the set of the training data points which end up in the same leaf as $\mathbf{x}$ in the tree $\mathcal{T}_k$. The weighting function $w_{\mathbf{x}}(\mathbf{x}_i)$ is

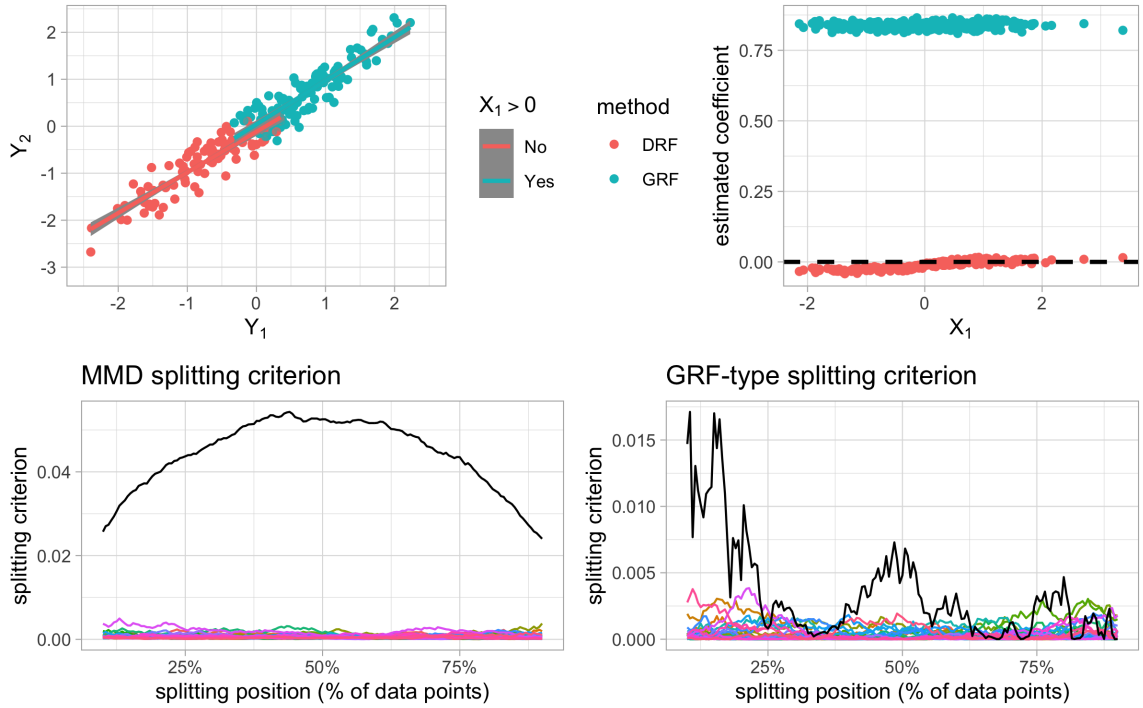


Figure 1: Top left: Illustration of data distribution for the toy example (3) when $n = 300$, $p = 30$. Bottom: The corresponding MMD (12) (left) and GRF (2) splitting criteria (right) at the root node. The curves of different colors correspond to different predictors, with X_1 denoted in black. Top right: Comparison of the estimates of DRF and Causal Forest [4] which respectively use those splitting criteria. Test points were randomly generated from the same distribution as the training data. Black dashed line indicates the correct value of the target quantity.

defined as the average of the corresponding weighting functions per tree [105]:

$$w_{\mathbf{x}}(\mathbf{x}_i) = \frac{1}{N} \sum_{k=1}^{N} \frac{\mathbb{1}\left(\mathbf{x}_i \in \mathcal{L}_k(\mathbf{x})\right)}{|\mathcal{L}_k(\mathbf{x})|}. \tag{4}$$

The weights are positive and add up to 1: $\sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) = 1$. In the case of equally sized leaf nodes, the assigned weight to a training point $\mathbf{x}_i$ is proportional to the number of trees where the test point $\mathbf{x}$ and $\mathbf{x}_i$ end up in the same leaf node. This shows that forest-based methods can in general be viewed as adaptive nearest neighbor methods. The sets $\mathcal{L}_k(\mathbf{x})$ of DRF will contain data points $(\mathbf{x}_i, \mathbf{y}_i)$ such that $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)$ is close to $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, thus removing bias due to heterogeneity of $\mathbf{Y}$ caused by $\mathbf{X}$. On the other hand, since the trees are constructed randomly and are thus fairly independent [15], the leaf sets $\mathcal{L}_k(\mathbf{x})$ will be different enough so that the induced weights $w_{\mathbf{x}}(\mathbf{x}_i)$ are not concentrated on a small set of data points, which would lead to high estimation variance. Such good bias-variance tradeoff properties of forest-based methods are also implied by their asymptotic properties [10, 179], even though this is a still active area of research and not much can be shown rigorously.

One can estimate the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ from the weighting function by

using the corresponding empirical distribution:

$$\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \cdot \delta_{\mathbf{y}_i}, \tag{5}$$

where $\delta_{\mathbf{y}_i}$ is the point mass at $\mathbf{y}_i$.

2.2.0.1 Two-step approach using weights. The weighting function $w_{\mathbf{x}}(\mathbf{x}_i)$ can directly be used for any target $\tau(\mathbf{x})$ in a second step and not just for estimating the conditional distribution. For example, the estimated conditional joint CDF is given by

$$\hat{F}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) = \hat{\mathbb{P}}(Y_1 \leqslant t_1, \ldots, Y_d \leqslant t_d \mid \mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \mathbb{1}(\cap_{j=1}^{d} \{(\mathbf{y}_i)_j \leqslant t_j\}). \tag{6}$$

It is important to point out that using the induced weighting function for locally weighted estimation is different than the approach of averaging the noisy estimates obtained per tree [181], originally used in standard Random Forests [15]. Even though the two approaches are equivalent for conditional mean estimation, the former approach is often much more efficient for more complex targets [4], since the number of data points in a single leaf is very small, leading to large variance of the estimates.

For the univariate response, the idea of using the induced weights for estimating targets different than the original target of conditional mean considered in [15] dates back to Quantile Regression Forests (QRF) [117], where a lot of emphasis is put on the quantile estimation, as the number of interesting targets is quite limited in the univariate setting. In the multivariate case, on the other hand, many interesting quantities such as, for example, conditional quantiles, conditional correlations or various conditional probability statements can easily be directly estimated from the weights.

By using the weights as an input for some other method, we can accomplish some more complicated objectives, such as conditional independence testing, causal effect estimation, semi-parametric learning, time series prediction or tail-index estimation in extreme value analysis. As an example, suppose that our data $\mathbf{Y}$ come from a certain parametric model, where the parameter θ is not constant, but depends on $\mathbf{X}$ instead, i.e. $\mathbf{Y} \mid \mathbf{X} = \mathbf{x} \sim f(\theta(\mathbf{x}), \cdot)$, see also [192]. One can then estimate the parameter $\theta(\mathbf{x})$ by using weighted maximum likelihood estimation:

$$\hat{\theta}(\mathbf{x}) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \log f(\theta, \mathbf{y}_i).$$

Another example is heterogeneous regression, where we are interested in the regression fit of an outcome $Y \in \mathbb{R}$ on certain predicting variables $\mathbf{W} \in \mathbb{R}^s$ conditionally on some event $\{\mathbf{X} = \mathbf{x}\}$. This can be achieved by weighted regression of Y on $\mathbf{W}$, where the weights $w_{\mathbf{x}}(\mathbf{x}_i)$ assigned to each data point $(\mathbf{w}_i, y_i)$ are obtained from DRF with the multivariate response $(Y, \mathbf{W}) \in \mathbb{R}^{s+1}$ and predictors $\mathbf{X} \in \mathbb{R}^p$, for an illustration see Section 4.4.

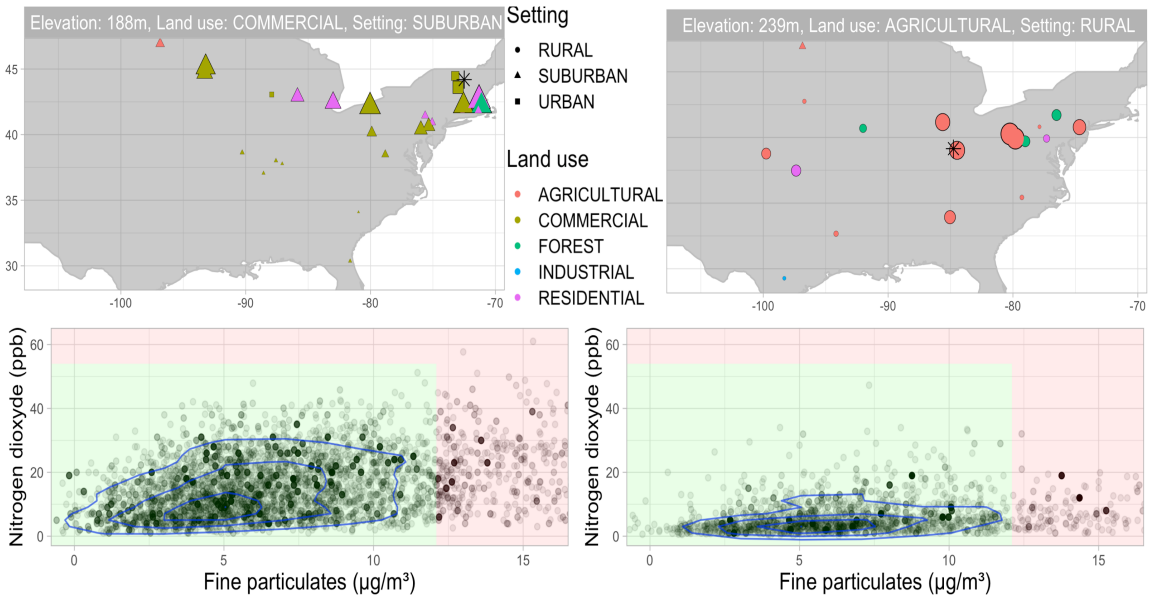


Figure 2: Top: the characteristics of the important training sites, for a fixed test site whose position is indicated by a black star and whose characteristics are indicated in the title. The total weight assigned corresponds to the symbol size. Bottom: estimated joint conditional distribution of two pollutants NO_2 and PM2.5, where the weights correspond to the transparency of the data points. Green area corresponds to 'Good' air quality category (AQI $\leqslant$ 50).

The weighting function of DRF is illustrated on the air quality data in Figure 2. Five years ($2015-2019$) of air pollution measurements were obtained from the US Environmental Protection Agency (EPA) website. Six main air pollutants (nitrogen dioxide (NO_2), carbon monoxide (CO), sulphur dioxide (SO_2), ozone (O_3) and coarse and fine particulate matter (PM10 and PM2.5)) that form the air quality index (AQI) were measured at many different measuring sites in the US for which we know the longitude, latitude, elevation, location setting (rural, urban, suburban) and how the land is used within a $1/4$ mile radius. Suppose we would want to know the distribution of the pollutant measurements at some new, unobserved, measurement site. We train DRF with the measurements (intraday maximum) of the two pollutants PM2.5 and NO_2 as the responses, and the site longitude, latitude, elevation, land use and location settings as the predictors and choose two decommissioned measurement sites as test points. For each test point we obtain the weights to all training measurements. We further combine the weights for all measurements corresponding to the same site. The top row illustrates for a given test site, whose characteristics are indicated in the plot title, how much weight in total is assigned to the measurements from a specific training site. We see that the important sites share many characteristics with the test site and that DRF determines the relevance of each characteristic in a data-adaptive way. The bottom row shows the corresponding estimates of the joint conditional distribution of the pollutants (we choose 2 of them for visualization purposes), where the transparency of each training point reflects the assigned weight. One can clearly see how

the estimated pollution levels are larger for the suburban site than for the rural site. The forest weights can be used, for example, for estimating the joint density (whose contours can be seen in the plot) or for estimating the probability that the AQI is below a certain value by summing the weights in the corresponding region of space.

2.3 Distributional Metric

In order to determine the best split of a parent node P, i.e. such that the distributions of the responses $\mathbf{Y}$ in the resulting child nodes C_L and C_R differ the most, one needs a good distributional metric $\mathcal{D}(\cdot,\cdot)$ (see Equation (1)) which can detect change in distribution of the response $\mathbf{Y}$ when additionally conditioning on an event $\{X_j > l\}$. Testing equality of distributions from the corresponding samples is an old problem in statistics, called two-sample testing problem. For univariate data, many good tests exist such as Wilcoxon rank test [185], Welch's t-test [183], Wasserstein two-sample testing [143], Kolmogorov-Smirnov test [112] and many others, but obtaining an efficient test for multivariate distributions has proven to be quite challenging due to the curse of dimensionality [54, 7].

Additional requirement for the choice of distributional metric $\mathcal{D}(\cdot,\cdot)$ used for data splitting is that it needs to be computationally very efficient as splitting is used extensively in the algorithm. If we construct N trees from n data points and in each node we consider mtry candidate variables for splitting, the complexity of the standard Random Forest algorithm [15] in the univariate case is $O(N \times \text{mtry} \times n \log n)$ provided our splits are balanced. It uses the CART splitting criterion, given by:

$$\frac{1}{n_P}\left(\sum_{\mathbf{x}_i \in C_L}(y_i - \overline{y}_L)^2 + \sum_{\mathbf{x}_i \in C_R}(y_i - \overline{y}_R)^2\right), \tag{7}$$

where $\overline{y}_L = \frac{1}{n_L}\sum_{\mathbf{x}_i \in C_L} y_i$ and $\overline{y}_R$ is defined analogously. This criterion has an advantage that not only it can be computed in $O(n_P)$ complexity, but this can be done for all possible splits $\{X_j \leqslant l\}$ as cutoff level l varies, since updating the splitting criterion when moving a single training data point from one child node to the other requires only $O(1)$ computational steps (most easily seen by rewriting the CART criterion as in (13)).

If the time complexity of evaluating the DRF splitting criterion (1) for a single splitting candidate X_j and all cutoffs l of interest (usually taken to range over all possible values) is at least n^c for some $c > 1$, say $O(f(n))$ for some function $f : \mathbb{R} \to \mathbb{R}$, then by solving the recursive relation we obtain that the overall complexity of the method is given by $O(N \times \text{mtry} \times f(n))$ [3], which can be unfeasible even for moderately large n if f grows too fast.

The problem of sequential two-sample testing is also central to the field of change-point detection [186, 20], with the slight difference that in the change-point problems the distribution is assumed to change abruptly at certain points in time, whereas for our forest construction we

only are interested in finding the best split of the form $\{X_j \leqslant l\}$ and the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \{\mathbf{X} \in P\} \cap \{X_j \leqslant l\})$ usually changes gradually with l. The testing power and the computational feasibility of the method play a big role in change-point detection as well. However, the state-of-the-art change-point detection algorithms [102, 113] are often too slow for our purpose as sequential testing is done $O(N \times \text{mtry} \times n)$ times for forest construction, much more frequently than in change-point problems.

2.3.1 MMD splitting criterion

Even though DRF could in theory be constructed with any distributional metric $\mathcal{D}(\cdot, \cdot)$, as a default choice we propose splitting criterion based on the Maximum Mean Discrepancy (MMD) statistic [62]. Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be the RKHS of real-valued functions on $\mathbb{R}^d$ induced by some positive-definite kernel k, and let $\varphi : \mathbb{R}^d \to \mathcal{H}$ be the corresponding feature map satisfying that $k(\mathbf{u}, \mathbf{v}) = \langle \varphi(\mathbf{u}), \varphi(\mathbf{v}) \rangle_{\mathcal{H}}$.

The MMD statistic $\mathcal{D}_{\text{MMD}(k)}(U, V)$ for kernel k and two samples $U = \{\mathbf{u}_1, \ldots, \mathbf{u}_{|U|}\}$ and $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_{|V|}\}$ is given by:

$$\mathcal{D}_{\text{MMD}(k)}(U, V) = \frac{1}{|U|^2} \sum_{i,j=1}^{|U|} k(\mathbf{u}_i, \mathbf{u}_j) + \frac{1}{|V|^2} \sum_{i,j=1}^{|V|} k(\mathbf{v}_i, \mathbf{v}_j) - \frac{2}{|U||V|} \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} k(\mathbf{u}_i, \mathbf{v}_j). \quad (8)$$

MMD compares the similarities, described by the kernel k, within each sample with the similarities across samples and is commonly used in practice for two-sample testing. It is based on the idea that one can assign to each distribution $\mathcal{P}$ its embedding $\mu(\mathcal{P})$ into the RKHS $\mathcal{H}$, which is the unique element of $\mathcal{H}$ given by

$$\mu(\mathcal{P}) = \mathbb{E}_{\mathbf{Y} \sim \mathcal{P}}[\varphi(\mathbf{Y})]. \quad (9)$$

The MMD two-sample statistic (8) can then equivalently be written as the squared distance between the embeddings of the empirical distributions with respect to the RKHS norm $\|\cdot\|_{\mathcal{H}}$:

$$\mathcal{D}_{\text{MMD}(k)}(U, V) = \left\| \mu\left(\frac{1}{|U|} \sum_{i=1}^{|U|} \delta_{\mathbf{u}_i} \right) - \mu\left(\frac{1}{|V|} \sum_{i=1}^{|V|} \delta_{\mathbf{v}_i} \right) \right\|_{\mathcal{H}}^2, \quad (10)$$

recalling that $\delta_{\mathbf{y}}$ is the point mass at $\mathbf{y}$.

As the sample sizes $|U|$ and $|V|$ grow, the MMD statistic (10) converges to its population version, which is the squared RKHS distance between the corresponding embeddings of the data-generating distributions of U and V. Since the embedding map μ is injective for a characteristic kernel k, we see that MMD is able to detect any difference in the distribution. Even though the power of the MMD two sample test also deteriorates as the data dimensionality grows, since the testing problem becomes intrinsically harder [142], it still has good empirical power compared to other multivariate two-sample tests for a wide range of k [63].

2.3.1.1 Fast random splitting criterion approximation. The $O((|U| + |V|)^2)$ complexity for computing $\mathcal{D}_{\text{MMD}(k)}(U, V)$ from (8) is nevertheless too large for many applications. For that reason, several fast approximations of MMD have been suggested in the literature [63, 65, 191, 32, 81]. As already mentioned, the complexity of the distributional metric $\mathcal{D}(\cdot, \cdot)$ used for DRF is crucial for the overall method to be computationally efficient, since the splitting step is used extensively in the forest construction. We therefore propose splitting based on an MMD statistic computed with an approximate kernel $\tilde{k}$, which is also a fast random approximation of the original MMD statistic [195].

Bochner's theorem (see e.g. [184, Theorem 6.6]) gives us that any bounded shift-invariant kernel can be written as

$$k(\mathbf{u}, \mathbf{v}) = \int_{\mathbb{R}^d} e^{i\omega^T(\mathbf{u}-\mathbf{v})} d\nu(\boldsymbol{\omega}), \tag{11}$$

i.e. as a Fourier transform of some measure ν. Therefore, by randomly sampling the frequency vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_B$ from normalized ν, we can approximate our kernel k by another kernel $\tilde{k}$ (up to a scaling factor) as follows:

$$k(\mathbf{u}, \mathbf{v}) = \int_{\mathbb{R}^d} e^{i\omega^T(\mathbf{u}-\mathbf{v})} d\nu(\boldsymbol{\omega}) \approx \frac{1}{B} \sum_{b=1}^{B} e^{i\omega_b^T(\mathbf{u}-\mathbf{v})} = \tilde{k}(\mathbf{u}, \mathbf{v}),$$

where we define $\tilde{k}(\mathbf{u}, \mathbf{v}) = \langle \boldsymbol{\varphi}(\mathbf{u}), \boldsymbol{\varphi}(\mathbf{v}) \rangle_{\mathbb{C}^B}$ as the kernel function with the feature map given by

$$\boldsymbol{\varphi}(\mathbf{u}) = \frac{1}{\sqrt{B}} \left(\tilde{\varphi}_{\omega_1}(\mathbf{u}), \ldots, \tilde{\varphi}_{\omega_B}(\mathbf{u})\right)^T = \frac{1}{\sqrt{B}} \left(e^{i\omega_1^T \mathbf{u}}, \ldots, e^{i\omega_B^T \mathbf{u}}\right)^T,$$

which is a random vector consisting of the Fourier features $\tilde{\varphi}_\omega(\mathbf{u}) = e^{i\omega^T \mathbf{u}} \in \mathbb{C}$ [139]. Such kernel approximations are frequently used in practice for computational efficiency [140, 98]. As a default choice of k we take the Gaussian kernel with bandwidth σ, since in this case we have a convenient expression for the measure ν and we sample $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_B \sim N_d(\mathbf{0}, \sigma^{-2} I_d)$. The bandwidth σ is chosen as the median pairwise distance between all training responses $\{\mathbf{y}_i\}_{i=1}^n$, commonly referred to as the 'median heuristic' [66].

From the representation of MMD via the distribution embeddings (10), we can obtain that MMD two-sample test statistic $\mathcal{D}_{\text{MMD}(\tilde{k})}$ using the approximate kernel $\tilde{k}$ is given by

$$\mathcal{D}_{\text{MMD}(\tilde{k})}\left(\{\mathbf{u}_i\}_{i=1}^{|U|}, \{\mathbf{v}_i\}_{i=1}^{|V|}\right) = \frac{1}{B} \sum_{b=1}^{B} \left| \frac{1}{|U|} \sum_{i=1}^{|U|} \tilde{\varphi}_{\omega_b}(\mathbf{u}_i) - \frac{1}{|V|} \sum_{i=1}^{|V|} \tilde{\varphi}_{\omega_b}(\mathbf{v}_i) \right|^2.$$

Interestingly, $\mathcal{D}_{\text{MMD}(\tilde{k})}$ is not only an MMD statistic on its own, but can also be viewed as a random approximation of the original MMD statistic $\mathcal{D}_{\text{MMD}(k)}$ (8) using kernel k; by using the kernel representation (11), it can be written as

$$\mathcal{D}_{\text{MMD}(k)}\left(\{\mathbf{u}_i\}_{i=1}^{|U|}, \{\mathbf{v}_i\}_{i=1}^{|V|}\right) = \int_{\mathbb{R}^d} \left| \frac{1}{|U|} \sum_{i=1}^{|U|} \tilde{\varphi}_{\omega}(\mathbf{u}_i) - \frac{1}{|V|} \sum_{i=1}^{|V|} \tilde{\varphi}_{\omega}(\mathbf{v}_i) \right|^2 d\nu(\boldsymbol{\omega}).$$

Finally, our DRF splitting criterion $\mathcal{D}(\cdot,\cdot)$ (1) is then taken to be the (scaled) MMD statistic $\frac{n_L n_R}{n_P^2}\mathcal{D}_{\mathrm{MMD}(\tilde{k})}\left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\}\right)$ with the approximate random kernel $\tilde{k}$ used instead of k, which can thus be conveniently written as:

$$\frac{1}{B}\sum_{b=1}^{B}\frac{n_L n_R}{n_P^2}\left|\frac{1}{n_L}\sum_{\mathbf{x}_i\in C_L}\tilde{\varphi}_{\omega_b}(\mathbf{y}_i)-\frac{1}{n_R}\sum_{\mathbf{x}_i\in C_R}\tilde{\varphi}_{\omega_b}(\mathbf{y}_i)\right|^2, \tag{12}$$

where we recall that $n_P = |\{i \mid \mathbf{x}_i \in P\}|$ and n_L, n_R are defined analogously. The additional scaling factor $\frac{n_L n_R}{n_P^2}$ in (12) occurs naturally and compensates the increased variance of the test statistic for unbalanced splits; it also appears in the GRF (2) and CART (see representation (13)) splitting criteria.

The main advantage of the splitting criterion based on $\mathcal{D}_{\mathrm{MMD}(\tilde{k})}$ is that by using the representation (1) it can be easily computed for every possible splitting level l in $O(Bn_P)$ complexity, whereas the MMD statistic $\mathcal{D}_{\mathrm{MMD}(k)}$ using kernel k would require $O(n_P^2)$ computational steps, which makes the overall complexity of the algorithm $O\left(B \times N \times \mathrm{mtry} \times n\log n\right)$ instead of much slower $O\left(N \times \mathrm{mtry} \times n^2\right)$.

We do not use the same approximate random kernel $\tilde{k}$ for different splits; for every parent node P we resample the frequency vectors $\{\omega_b\}_{b=1}^{B}$ defining the corresponding feature map $\tilde{\varphi}$. Using different $\tilde{k}$ at each node might help to better detect different distributional changes. Furthermore, having different random kernels for each node agrees well with the randomness of the Random Forests and helps making the trees more independent. Since the MMD statistic $\mathcal{D}_{\mathrm{MMD}(\tilde{k})}$ used for our splitting criterion is not only an approximation of $\mathcal{D}_{\mathrm{MMD}(k)}$, but is itself an MMD statistic, it inherits good power for detecting any difference in distribution of $\mathbf{Y}$ in the child nodes for moderately large data dimensionality d, even when B is reasonably small. One could even consider changing the number of random Fourier features B at different levels of the tree, as n_P varies, but for simplicity we take it to be fixed.

2.3.1.2 Relationship to CART. There is some similarity of our MMD-based splitting criterion (12) with the standard variance reduction CART splitting criterion (7) when $d = 1$, which can be rewritten as:

$$\frac{n_L n_R}{n_P^2}\left(\frac{1}{n_L}\sum_{\mathbf{x}_i\in C_L} y_i-\frac{1}{n_R}\sum_{\mathbf{x}_i\in C_R} y_i\right)^2. \tag{13}$$

The derivation can be found in Appendix 2. From this representation, we see that the CART splitting criterion (7) is also equivalent to the GRF splitting criterion (2) when our target is the univariate conditional mean $\tau(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ which is estimated for C_L and C_R by the sample means $\hat{\tau}_L = \overline{y}_L$ and $\hat{\tau}_R = \overline{y}_R$. Therefore, as it compares the means of the univariate response Y in the child nodes, the CART criterion can only detect changes in the response mean well, which is sufficient for prediction of Y from $\mathbf{X}$, but might not be suitable for more complex targets. Similarly, for multivariate applications, aggregating the marginal CART criteria [92, 152]

across different components Y_i of the response can only detect changes in the means of their marginal distributions. However, it is possible in the multivariate case that the pairwise correlations or the variances of the responses change, while the marginal means stay (almost) constant. For an illustration on simulated data, see Figure 7. Additionally, aggregating the splitting criteria over d components of the response $\mathbf{Y}$ can reduce the signal size if only the distribution of a few components change. Our MMD-based splitting criterion (12) is able to avoid such difficulties as it implicitly inspects all aspects of the multivariate response distribution.

If one takes a trivial kernel $k_{\text{id}}(y_i, y_j) = y_i y_j$ with the identity feature map $\varphi_{\text{id}}(y) = y$, the distributional embedding (9) is given by $\mu(\mathcal{P}) = \mathbb{E}_{Y\sim\mathcal{P}}[Y]$ and thus the corresponding splitting criterion based on $\mathcal{D}_{\text{MMD}(k_{\text{id}})}$ (10) is exactly equal to the CART splitting criterion (7), which can be seen from its equivalent representation (13). Interestingly, Theorem 31 in Section 3 shows that the MMD splitting criterion with general kernel k can also be viewed as the abstract version of the CART criterion in the RKHS $\mathcal{H}$ corresponding to k [50], with the response variable being the feature map $\varphi(\mathbf{Y}) \in \mathcal{H}$. Therefore, DRF with the MMD splitting criterion can also be viewed as a forest-based method for estimation of the conditional embedding, which further justifies the proposed method. In Section 3 below, we use this relationship to derive interesting theoretical properties of DRF with the MMD splitting criterion.

3 Theoretical Results

In this section we first use the properties of the kernel mean embedding in order to relate DRF with the MMD splitting criterion to an abstract version of the standard Random Forest with the CART splitting criterion [15], where the response is taking values in the corresponding RKHS. This representation reveals that DRF with the MMD splitting criterion can be viewed as a Random Forest estimator of the conditional mean embedding (CME) [132], similarly as the standard Random Forest estimates the conditional mean. This relationship is further exploited to adapt the existing theoretical results from the Random Forest literature to show that our estimate (5) of the conditional distribution of the response is consistent with respect to the MMD metric for probability measures and with a good rate. Finally, we show that this implies consistency of the induced DRF estimates for a range interesting targets $\tau(\mathbf{x})$, such as conditional CDFs or quantiles. The proofs of all results can be found in the Appendix 2.

3.1 Casting DRF as a Random Forest in an RKHS

Recalling the notation from above, let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be the Reproducing kernel Hilbert space induced by the positive definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and let $\varphi : \mathbb{R}^d \to \mathcal{H}$ be its corresponding feature map. The kernel embedding function $\mu : \mathcal{M}_b(\mathbb{R}^d) \to \mathcal{H}$ maps any bounded signed Borel

measure $\mathcal{P}$ on $\mathbb{R}^d$ to an element $\mu(\mathcal{P}) \in \mathcal{H}$ defined by

$$\mu(\mathcal{P}) = \int_{\mathbb{R}^d} \varphi(\mathbf{y})\, d\mathcal{P}(\mathbf{y}), \tag{14}$$

see also (9). Boundedness of k ensures that μ is indeed defined on all of $\mathcal{M}_b(\mathbb{R}^d)$, while continuity of k ensures that $\mathcal{H}$ is separable [75].

By considering the kernel embedding $\mu(\cdot)$ and using its linearity, the embedding of the distributional estimate $\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$ of DRF (5) can be written as the average of the embeddings of the empirical distributions of $\mathbf{Y}$ in the leaves containing $\mathbf{x}$ over all trees:

$$\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) = \frac{1}{N}\sum_{k=1}^{N} \mu\left(\frac{1}{|\mathcal{L}_k(\mathbf{x})|}\sum_{\mathbf{x}_i \in \mathcal{L}_k(\mathbf{x})} \delta_{\mathbf{y}_i}\right) = \frac{1}{N}\sum_{k=1}^{N} \frac{1}{|\mathcal{L}_k(\mathbf{x})|}\sum_{\mathbf{x}_i \in \mathcal{L}_k(\mathbf{x})} \mu(\delta_{\mathbf{y}_i}). \tag{15}$$

This is analogous to the prediction of the response for the standard univariate Random Forest, but where we average the embeddings $\mu(\delta_{\mathbf{y}_i}) = \varphi(\mathbf{y}_i) \in \mathcal{H}$ instead of the response values y_i themselves.

Furthermore, one can relate the MMD splitting criterion to the original CART criterion (7), which measures the mean squared prediction error for splitting a certain parent node P into children C_L and C_R. On one hand, from Equation (13) we see that the CART criterion also measures the squared distance between the response averages $\frac{1}{n_L}\sum_{\mathbf{x}_i \in C_L} y_i$ and $\frac{1}{n_R}\sum_{\mathbf{x}_i \in C_R} y_i$ in the child nodes, but on the other hand, Equation (10) shows that the MMD splitting criterion measures the RKHS distance between the embeddings of the empirical response distributions in C_L and C_R. This is summarized in the following theorem, which not only shows that the MMD splitting criterion can be viewed as the abstract CART criterion in the RKHS $\mathcal{H}$ [50], but also that DRF with the MMD splitting criterion can be viewed asymptotically as a greedy minimization of the average squared MMD distance between our estimate $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ and the truth $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$:

Theorem 31. *For any split of a parent node P into child nodes C_L and C_R, let $\hat{\mathbb{P}}_{split}(\mathbf{x}) = \sum_{j\in\{L,R\}} \mathbb{1}(\mathbf{x} \in C_j)\frac{1}{n_j}\sum_{\mathbf{x}_i \in C_j} \delta_{\mathbf{y}_i}$ denote the resulting estimate of the distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ when $\mathbf{x} \in P$. Then the MMD splitting criterion can be viewed as the version of the CART criterion* (7) *on $\mathcal{H}$:*

$$\begin{aligned}
&\arg\max_{split} \frac{n_L n_R}{n_P^2} \mathcal{D}_{MMD(k)}\left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\}\right) \\
&= \arg\min_{split} \frac{1}{n_P}\sum_{\mathbf{x}_i \in P} \left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{\mathbb{P}}_{split}(\mathbf{x}_i))\right\|_{\mathcal{H}}^2.
\end{aligned}$$

Moreover, for any node P and any fixed distributional estimator $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, we have:

$$\begin{aligned}
&\frac{1}{n_P}\sum_{\mathbf{x}_i \in P} \left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i))\right\|_{\mathcal{H}}^2 \\
&= V_P + \mathbb{E}\left[\|\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X})) - \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}))\|_{\mathcal{H}}^2 \mid \mathbf{X} \in P\right] + O_p(n^{-1/2}),
\end{aligned}$$

where $V_P = \mathbb{E}\left[\|\mu(\delta_{\mathbf{Y}}) - \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}))\|_{\mathcal{H}}^2 \mid \mathbf{X} \in P\right]$ *is a deterministic term not depending on the estimates* $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$.

In conclusion, from the above results we see that by applying the kernel embedding (14), we can shift the perspective to the RKHS $\mathcal{H}$ and view DRF as the analogue of the original Random Forest for estimation of the CME $\mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) = \mathbb{E}[\varphi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]$ in an abstract Hilbert space $\mathcal{H}$. Like some traditional CME estimators [162, 161, 126, 132] it is also of the form

$$\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) = \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \cdot k(\mathbf{y}_i, \cdot). \tag{16}$$

It can be shown that, since DRF produces nonnegative weights that sum to one in (16), there is one-to-one correspondence between the resulting estimate in $\mathcal{H}$ and the empirical probability distribution in $\mathcal{M}_b(\mathbb{R}^d)$. Thus DRF can be seen as a CME estimator through (16), or directly as an estimator for the conditional distribution through (5). By contrast, other CME estimators of the form (16) have weights that are unconstrained and can be negative. Finding an appropriate distribution on $\mathbb{R}^d$ for a given mean-embedding (sometimes referred to as "distributional inverse image problem", see e.g. [115, 126]) is not straightforward in general. For certain tasks, such as sampling from the estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ or obtaining the plug-in estimates of some target functionals $\tau(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$, this is crucial.

3.2 Convergence of Conditional Distribution Estimates

As we have seen, DRF can be viewed as the abstract version of the standard Random Forest when the response takes value in an RKHS. In principle, one could thus derive properties of DRF by adapting any existing theoretical result from the literature to the RKHS case. However, a lot of care is needed for making the results rigorous in this abstract setup, as many useful properties of $\mathbb{R}$ need not hold for infinite-dimensional $\mathcal{H}$. This section is inspired by the results from [181].

We suppose that the forest construction satisfies the following properties, which significantly facilitate the theoretical considerations of the method and ensure that our forest estimator is well behaved, as stated in [181]:

(P1) (*Data sampling*) The bootstrap sampling with replacement, usually used in forest-based methods, is replaced by a subsampling step, where for each tree we choose a random subset of size s_n out of n training data points. We consider s_n going to infinity with n, with the rate specified below.

(P2) (*Honesty*) The data used for constructing each tree is split into two partOn the pitfalls of Gaussian scoring for causal discoverys; the first is used for determining the splits and the second for populating the leaves and thus for estimating the response.

(P3) (*α-regularity*) Each split leaves at least a fraction $0 < \alpha \leqslant 0.2$ of the available training sample on each side. Moreover, the trees are grown until every leaf contains between κ and $2\kappa - 1$ observations, for some fixed tuning parameter $\kappa \in \mathbb{N}$.

(P4) (*Symmetry*) The (randomized) output of a tree does not depend on the ordering of the training samples.

(P5) (*Random-split*) At every split point, the probability that the split occurs along the feature X_j is bounded below by π/p, for some $\pi > 0$ and for all $j = 1, \ldots, p$.

The validity of the above properties are easily ensured by the forest construction used.

From Equation (15), the prediction of DRF for a given test point $\mathbf{x}$ can be viewed as an element of $\mathcal{H}$. If we denote the i-th training observation by $\mathbf{Z}_i = (\mathbf{x}_i, \mu(\delta_{\mathbf{y}_i})) \in \mathbb{R}^p \times \mathcal{H}$, then by (15) we estimate the embedding of the true conditional distribution $\mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$ by the average of the corresponding estimates per tree:

$$\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) = \frac{1}{N} \sum_{j=1}^{N} T(\mathbf{x}; \varepsilon_j, \mathcal{Z}_j),$$

where $\mathcal{Z}_k$ is a random subset of $\{\mathbf{Z}_i\}_{i=1}^n$ of size s_n chosen for constructing the j-th tree $\mathcal{T}_j$ and ε_j is a random variable capturing all randomness in growing $\mathcal{T}_j$, such as the choice of the splitting candidates. $T(\mathbf{x}; \varepsilon, \mathcal{Z})$ denotes the output of a single tree: i.e. the average of the terms $\mu(\delta_{\mathbf{Y}_i})$ over all data points $\mathbf{Z}_i$ contained in the leaf $\mathcal{L}(\mathbf{x})$ of the tree constructed from ε and $\mathcal{Z}$.

Since one can take the number of trees N to be arbitrarily large, we consider an "idealized" version of our estimator, as done in [180], which we denote as $\hat{\mu}_n(\mathbf{x})$:

$$\hat{\mu}_n(\mathbf{x}) = \binom{n}{s_n}^{-1} \sum_{i_1 < i_2 < \ldots < i_{s_n}} \mathbb{E}_\varepsilon \, T(\mathbf{x}; \varepsilon; \{\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}\}), \tag{17}$$

where the sum is taken over all $\binom{n}{s_n}$ possible subsets of $\{\mathbf{Z}_i\}_{i=1}^n$.We have that $\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) \to \hat{\mu}_n(\mathbf{x})$ as $N \to \infty$, while keeping the other variables constant, and thus we assume for simplicity that those two quantities are the same.

Our main result shows that, under similar assumptions as in [180], the embedding of our conditional distribution estimator $\hat{\mu}_n(\mathbf{x}) = \mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$ consistently estimates $\mu(\mathbf{x}) := \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$ with respect to the RKHS norm with a certain rate:

Theorem 32. *Suppose that our forest construction satisfies properties **(P1)–(P5)**. Assume additionally that k is a bounded and continuous kernel and that we have a random design with $\mathbf{X}_1, \ldots, \mathbf{X}_n$ independent and identically distributed on $[0,1]^p$ with a density bounded away from 0 and infinity. If the subsample size s_n is of order n^β for some $0 < \beta < 1$, the mapping*

$$\mathbf{x} \mapsto \mu(\mathbf{x}) = \mathbb{E}[\mu(\delta_{\mathbf{Y}}) \mid \mathbf{X} = \mathbf{x}] \in \mathcal{H},$$

is Lipschitz and $\sup_{\mathbf{x}\in[0,1]^p} \mathbb{E}[\|\mu(\delta_{\mathbf{Y}})\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}] < \infty$*, we obtain the consistency w.r.t. the RKHS norm:*

$$\|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} = O_p\left(n^{-\gamma}\right), \tag{18}$$

for $\gamma = \frac{1}{2}\min\left(1-\beta, \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}\cdot\beta\right)$.

Remark 33. *The rate in* (18) *is analogous to the one from [181], who used it further to derive the asymptotic normality of the Random Forest estimator in* $\mathbb{R}$*. Unfortunately, this alone is not enough to establish asymptotic normality of* $(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}))/\sigma_n$ *as an element of* $\mathcal{H}$*. To do so, one needs to prove a functional central limit theorem with a Gaussian limiting process in the Hilbert space* $\mathcal{H}$*. This then allows to deduce asymptotic normality of smooth real-valued functionals. We will provide the detailed derivations in future work.*

3.3 Convergence of the Induced Estimates

The above result shows that DRF estimate $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ converges fast to the truth $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ in the MMD distance, i.e. the RKHS distance between the corresponding embeddings. Even though this is interesting on its own, ultimately we want to relate this result to estimation of certain distributional targets $\tau(\mathbf{x}) = \tau(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x}))$.

For any $f \in \mathcal{H}$, we have that the DRF estimate of the target $\tau(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}]$ equals the dot product $\langle f, \hat{\mu}_n(\mathbf{x})\rangle_{\mathcal{H}}$ in the RKHS:

$$\langle f, \hat{\mu}_n(\mathbf{x})\rangle_{\mathcal{H}} = \left\langle f, \int_{\mathbb{R}^d} \varphi(\mathbf{y}) d\hat{\mathbb{P}}(\mathbf{y} \mid \mathbf{X}=\mathbf{x}) \right\rangle_{\mathcal{H}} = \int_{\mathbb{R}^d} f(\mathbf{y})\, d\hat{\mathbb{P}}(\mathbf{y} \mid \mathbf{X}=\mathbf{x}) = \sum_{i=1}^n w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i),$$

where we recall the weighting function $w_{\mathbf{x}}(\cdot)$ induced by the forest (4). Therefore, the consistency result (18) in Theorem 32 directly implies that

$$\sum_{i=1}^n w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i) = \langle f, \hat{\mu}_n(\mathbf{x})\rangle_{\mathcal{H}} \xrightarrow{p} \langle f, \mu(\mathbf{x})\rangle_{\mathcal{H}} = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}] \quad \text{for any } f \in \mathcal{H}, \tag{19}$$

i.e. that the DRF consistently estimates the targets of the form $\tau(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}]$, for $f \in \mathcal{H}$. From (18) we also obtain the rate of convergence when $s_n \asymp n^{\beta}$:

$$\left|\sum_{i=1}^n w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i) - \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}]\right| = O_p\left(n^{-\gamma}\|f\|_{\mathcal{H}}\right),$$

for γ as in Theorem 32. When k is continuous, it is well known that all elements of $\mathcal{H}$ are continuous, see e.g. [75]. Under certain assumptions on the kernel and its input space, holding for several popular kernels, (e.g. the Gaussian kernel) [163], we can generalize the convergence result (19) to any bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$, as the convergence of measures $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x}) \to \mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ in the MMD metric will also imply their weak convergence, i.e. k metrizes weak convergence [163, 157, 156]:

Corollary 34. *Assume that one of the following two sets of conditions holds:*

(a) The kernel k is bounded, (jointly) continuous and has

$$\int\int k(\mathbf{x},\mathbf{y})d\mathcal{P}(\mathbf{x})d\mathcal{P}(\mathbf{y}) > 0 \;\; \forall \mathcal{P} \in \mathcal{M}_b(\mathbb{R}^d)\backslash\{0\}. \tag{20}$$

Moreover, $\mathbf{y} \mapsto k(\mathbf{y}_0,\mathbf{y})$ is vanishing at infinity, for all $\mathbf{y}_0 \in \mathbb{R}^d$.

(b) The kernel k is bounded, shift-invariant, (jointly) continuous and ν in the Bochner representation in (11) *is supported on all of $\mathbb{R}^d$. Moreover, $\mathbf{Y}$ takes its values almost surely in a closed and bounded subset of $\mathbb{R}^d$.*

Then, under the conditions of Theorem 32, we have for any bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$ that DRF consistently estimates the target $\tau(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}]$ for any $\mathbf{x} \in [0,1]^p$:

$$\sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i) \xrightarrow{p} \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}].$$

Recalling the Portmanteau Lemma on separable metric spaces, see e.g. [46, Chapter 11], this has several other interesting consequences, such as the consistency of CDF and quantile estimates; Let $F_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\cdot)$ be the conditional CDF of $\mathbf{Y}$ and for any index $1 \leqslant i \leqslant d$, let $F_{Y_i|\mathbf{X}=\mathbf{x}}(\cdot)$ be the conditional CDF of Y_i and $F^{-1}_{Y_i|\mathbf{X}=\mathbf{x}}(\cdot)$ its generalized inverse, i.e. the quantile function. Let $\hat{F}_{Y_i|\mathbf{X}=\mathbf{x}}(\cdot)$ and $\hat{F}^{-1}_{Y_i|\mathbf{X}=\mathbf{x}}(\cdot)$ be the corresponding DRF estimates via weighting function (6). Then we have the following result:

Corollary 35. *Under the conditions of Corollary 34, for any $1 \leqslant i \leqslant d$, we have*

$$\hat{F}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) \xrightarrow{p} F_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{t})$$
$$\hat{F}^{-1}_{Y_i|\mathbf{X}=\mathbf{x}}(t) \xrightarrow{p} F^{-1}_{Y_i|\mathbf{X}=\mathbf{x}}(t),$$

for all points of continuity $\mathbf{t} \in \mathbb{R}^d$ and $t \in \mathbb{R}$ of $F_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\cdot)$ and $F^{-1}_{Y_i|\mathbf{X}=\mathbf{x}}(\cdot)$ respectively.

4 Applications and Numerical Experiments

The goal of this section is to demonstrate the versatility and applicability of DRF for many practical problems. We show that DRF can be used not only as an estimator of the multivariate conditional distribution, but also as a two-step method to easily obtain out-of-the box estimators for various, and potentially complex, targets $\tau(\mathbf{x})$.

Our main focus lies on the more complicated targets which cannot be that straightforwardly approached by conventional methods. However, we also illustrate the usage of DRF for certain applications for which there already exist several well-established methods. Whenever possible

in such cases, we compare the performance of DRF with the specialized, task-specific methods to show that, despite its generality, there is at most a very small loss of precision. However, we should point out that for many targets such as, that can not be written in a form of a conditional mean or a conditional quantile, for example, conditional correlation, direct comparison of the accuracy is not possible for real data, since no suitable loss function exists and the ground truth is unknown. Finally, we show that, in addition to directly estimating certain targets, DRF can also be a very useful tool for many different applications, such as causality and fairness.

Detailed descriptions of all competing methods, data sets and the corresponding analyses can be found in Appendix 3, and some additional simulations can be found in the Appendix 4.

4.1 Estimation of Conditional Multivariate Distributions

In order to provide good estimates for any target $\tau(\mathbf{x}) = \tau(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x}))$, our method needs to estimate the conditional multivariate distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ well. Therefore, we first investigate here the accuracy of the DRF estimate (5) of the full conditional distribution and compare its performance with the performance of several existing methods.

In addition to a few simple methods such as the k-nearest neighbors or the kernel regression, which locally weight the training points, we also consider the CME estimator of [132] and several advanced machine learning methods such as the Conditional Generative Adversarial Network (CGAN) [125, 2], Conditional Variational Autoencoder (CVAE) [160] and Masked Autoregressive Flow [131]. It is worth mentioning that the focus in the machine learning literature has been more on applications where d is very large (e.g. pixels of an image) and p is very small (such as image labels). Even though some methods do not provide the estimated conditional distribution in a form as simple as DRF, one is still able to sample from the estimated distribution and thus perform any subsequent analysis and make fair comparisons between the methods. For the CME estimator we simply set the negative weights to zero and renormalize, such that the weights are nonnegative and sum to one.

We first illustrate the estimated distributions by the above methods on a toy example where $n = 1000, p = 10, d = 2$ and

$$Y_1 \perp\!\!\!\perp Y_2 \mid \mathbf{X}=\mathbf{x}, \quad Y_1 \mid \mathbf{X}=\mathbf{x} \sim U(x_1, x_1+1), \quad Y_2 \mid \mathbf{X}=\mathbf{x} \sim U(0, x_2), \quad \mathbf{X} \sim U(0,1)^p. \tag{21}$$

In the above example X_1 affects the mean of Y_1, whereas X_2 affects the both mean and variance of Y_2, and $X_3, \ldots, X_p$ have no impact. The results can be seen in Figure 3. We see that, unlike some other methods, DRF is able to balance the importance of the predictors X_1 and X_2 and thus to estimate the distributions of Y_1 and Y_2 well.

One can do a more extensive comparison on a collection of real data sets. We use the benchmark data sets from the multi-target regression literature [172] together with some additional ones created from the data sets described throughout this paper. The performance of

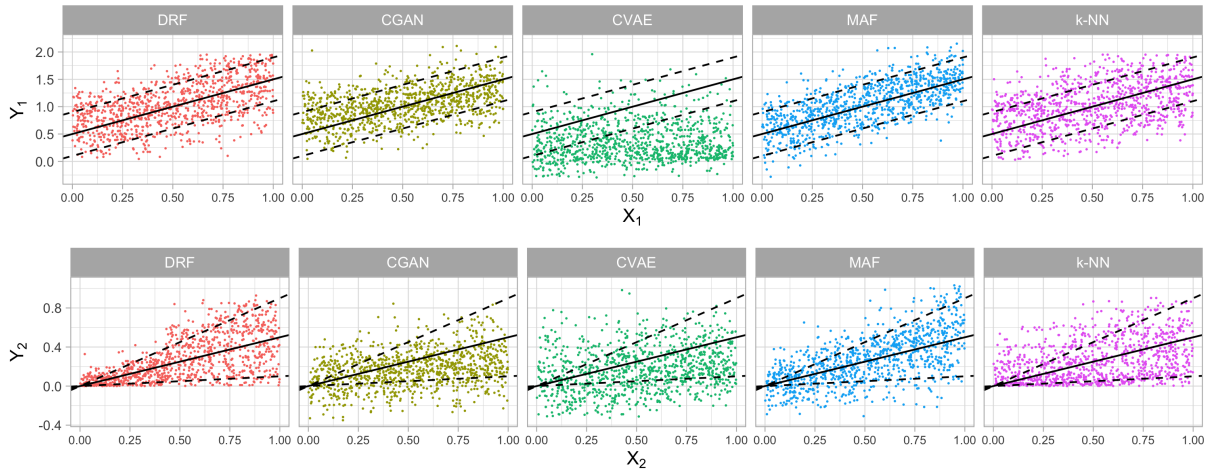


Figure 3: The illustration of the estimated joint conditional distribution obtained by different methods for the toy example (21). For 1000 randomly generated test points $\mathbf{X}_{\text{test}} \sim U(0,1)^p$ the top row shows the estimated distribution of the response component Y_1, whereas the bottom row shows the estimated distribution of Y_2. The 0.1 and 0.9 quantiles of the true conditional distribution are indicated by a dashed black line, whereas the conditional mean is shown as a black solid line.

DRF is compared with the performance of other existing methods for nonparametric estimation of multivariate distributions by using the Negative Log Predictive Density (NLPD) loss, which evaluates the logarithm of the induced multivariate density estimate [138]. As the number of test points grows to infinity, NLPD loss becomes equivalent to the average KL divergence between the estimated and the true conditional distribution and is thus able to capture how well one estimates the whole distribution, instead of only its mean.

In addition to the methods mentioned above, we also include some methods that are intended only for mean prediction, by assuming that the distribution of the response around its mean is homogeneous, i.e. that the conditional distribution $\mathbb{P}(\mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{X}] \mid \mathbf{X} = \mathbf{x})$ does not depend on $\mathbf{x}$. This is fitted by regressing each component of $\mathbf{Y}$ separately on $\mathbf{X}$ and using the pooled residuals. We consider the standard nonparametric regression methods such as Random Forest [15], XGBoost [27], and Deep Neural Networks [60].

The results are shown in Table 1. We see that DRF performs well for a wide range of sample size and problem dimensionality, especially in problems where p is large and d is moderately big. It does so without the need for any tuning or involved numerical optimization.

4.2 Estimation of Statistical Functionals

Because DRF represents the estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) = \sum_i w_{\mathbf{x}}(\mathbf{x}_i) \cdot \delta_{\mathbf{y}_i}$ in a convenient form by using weights $w_{\mathbf{x}}(\mathbf{x}_i)$, a plug-in estimator $\tau(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$ of many common real-valued statistical functionals $\tau(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) \in \mathbb{R}$ can be easily constructed from $w_{\mathbf{x}}(\cdot)$.

	jura	slump	wq	enb	atp1d	atp7d	scpf	sf1	sf2	copula	wage	births1	births2	air
n	359	103	1K	768	337	296	143	323	1K	5K	10K	10K	10K	10K
p	15	7	16	8	370	370	8	21	22	10	73	23	24	15
d	3	3	14	2	6	6	3	3	6	2	2	2	4	6
DRF	3.9	**4.0**	**22.5**	2.1	7.3	**7.0**	**2.0**	-24.2	**-24.3**	**2.8**	**2.8**	2.5	**4.2**	**8.5**
CGAN	10.8	5.3	27.3	3.5	10.4	363	4.8	9.8	21.1	5.8	360	**2.4**	>1K	11.8
CVAE	4.8	37.8	36.8	2.6	>1K	>1K	108.8	8.6	>1K	2.9	>1K	>1K	49.7	9.6
MAF	4.6	4.5	23.9	3.0	8.0	8.1	2.6	4.7	3.8	2.9	3.0	2.5	>1K	8.5
k-NN	4.5	5.0	23.4	2.4	8.8	8.6	4.1	-22.4	-19.7	2.9	**2.8**	2.7	4.4	8.8
kernel	4.1	4.2	23.0	**2.0**	**6.6**	7.1	2.9	-23.0	-20.6	2.8	2.9	2.6	4.3	8.4
RF	7.1	12.1	35.2	5.7	12.7	13.3	16.7	3.9	2.2	5.8	6.1	5.0	8.3	13.9
XGBoost	11.4	38.3	25.9	3.0	>1K	>1K	>1K	0.3	1.6	3.5	2.9	>1K	>1K	12.8
DNN	4.0	4.2	23.3	2.6	8.6	8.7	2.6	2.3	2.2	2.9	3.0	2.6	5.4	8.6
CME	**3.2**	4.9	23.2	2.9	8.5	8.4	2.5	**-24.4**	**-24.3**	2.8	3.5	3.8	15.2	8.8

Table 1: NLPD loss computed on out-of-sample observations for the estimated conditional distributions obtained by several different methods (corresponding to rows) for many real data sets (corresponding to columns). The best method is indicated in bold.

We first investigate the performance for the classical problem of univariate quantile estimation on simulated data. We consider the following three data generating mechanisms with $p = 40, n = 2000$ and $\mathbf{X}_i \overset{i.i.d.}{\sim} U(-1,1)^p$:

- Scenario 1: $Y \sim N(0.8 \cdot \mathbb{1}(X_1 > 0), 1)$ (mean shift based on X_1)
- Scenario 2: $Y \sim N(0, (1 + \mathbb{1}(X_1 > 0))^2)$ (variance shift based on X_1)
- Scenario 3: $Y \sim \mathbb{1}(X_1 \leqslant 0) \cdot N(1,1) + \mathbb{1}(X_1 > 0) \cdot \text{Exp}(1)$ (distribution shift based on X_1, constant mean and variance)

The first two scenarios correspond exactly to the examples given in [4].

In Figure 4 we can see the corresponding estimates of the conditional quantiles for DRF, Quantile Regression Forest (QRF) [117], which uses the same forest construction with CART splitting criterion as the original Random Forest [15] but estimates the quantiles from the induced weighting function, Generalized Random Forests (GRF) [4] with a splitting criterion specifically designed for quantile estimation and Transformation Forests (TRF) [74]. We see that DRF is performing very well even compared to methods that are specifically tailored to quantile estimation.

The multivariate setting is however more interesting, as one can use DRF to compute much more interesting statistical functionals $\tau(\mathbf{x})$. We illustrate this in Figure 5 for the air quality data set, described in Section 2.2. The left plot shows one value of the estimated multivariate CDF,

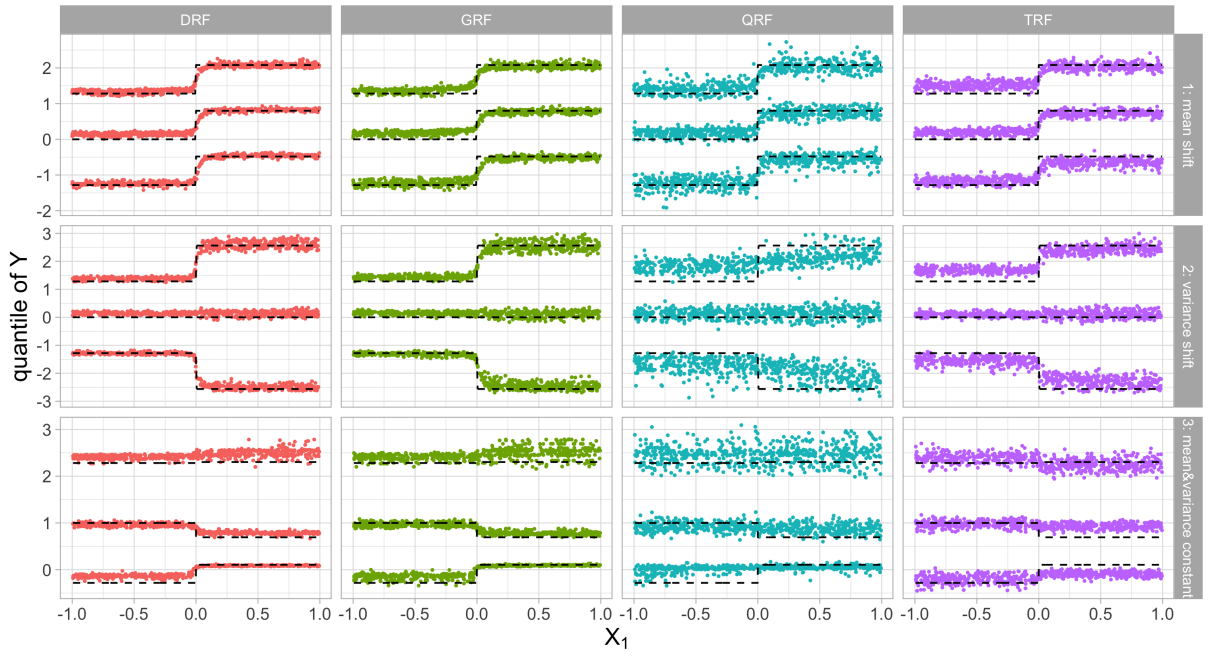


Figure 4: Scatter plot of predictions of the $0.1, 0.5$ and 0.9 quantiles against X_1 for randomly generated 500 test data points $\mathbf{X}_{\text{test}} \sim U(-1,1)^p$. The true values of the quantiles are displayed by black dashed lines. The columns corresponds to different methods DRF (red), GRF (green), QRF (blue), TRF (purple). The rows correspond to different simulation scenarios. The first two are taken from [4].

specifically the estimated probability of the event that the air quality index (AQI) is at most 50 at a given test site. This corresponds to the "Good" category and means that the amount of every air pollutant is below a certain threshold determined by the EPA. Such probability estimates can be easily obtained by summing the weights of the training points belonging to the event of interest. For both plots in Figures 2 and 5, we train the single DRF with the same set of predictor variables and take the three pollutants O_3, SO_2 and PM2.5 as the responses. In this way we still have training data from many different sites.

In order to investigate the accuracy of the conditional CDF obtained by DRF, we compare the estimated probabilities with estimates of the standard univariate classification forest [15]

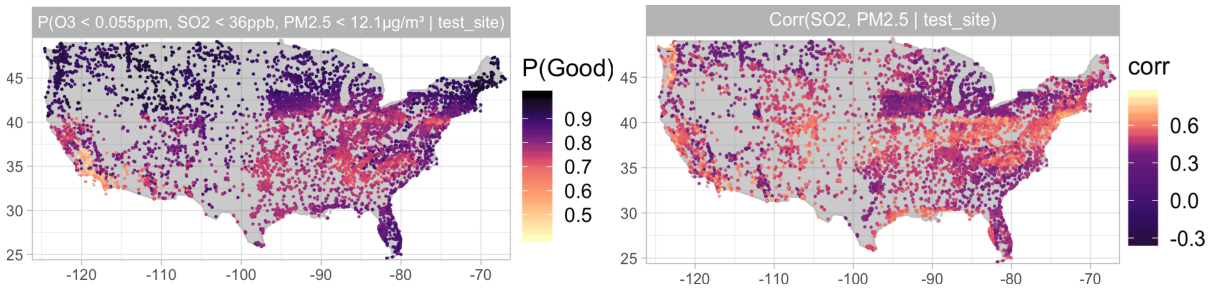


Figure 5: Estimates of the probability $\mathbb{P}(\text{AQI} \leqslant 50 \mid \text{test site})$ (left) and the conditional correlation (right) derived from the DRF estimate of the multivariate conditional distribution.

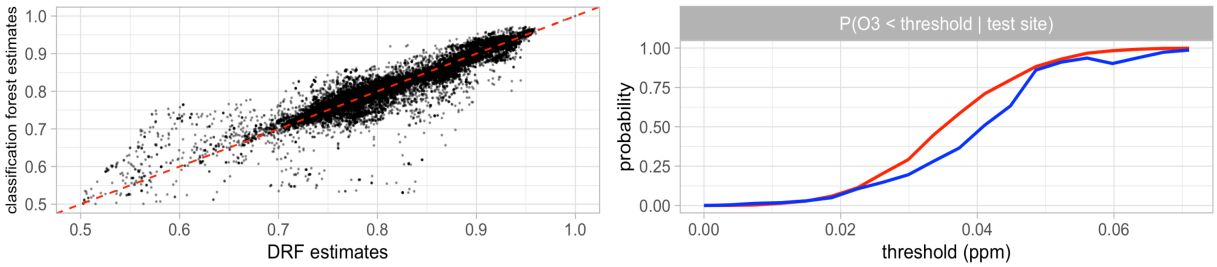


Figure 6: Left: Comparison of the CDF estimates obtained by DRF (displayed also in the left plot of Figure 5) and by the classification forest. Right: Example how the CDF estimated by using the classification forest (blue) need not be monotone, whereas the DRF estimates (red) are well-behaved.

with the response $\mathbb{1}(\text{AQI} \leqslant 50)$. In the left plot of Figure 6, we can see that the DRF estimates of the $\mathbb{P}(\text{AQI} \leqslant 50 \mid \mathbf{X} = \mathbf{x})$ (also visualized in Figure 5) are quite similar to the estimates of the classification forest predicting the outcome $\mathbb{1}(\text{AQI} \leqslant 50)$. Furthermore, the cross-entropy loss evaluated on the held-out measurements equals 0.4671 and 0.4663 respectively, showing almost no loss of precision. In general, estimating the simple functionals from the weights provided by DRF comes usually at a small to no loss compared to the classical methods specifically designed for this task.

In addition to the classical functionals $\tau(\mathbf{x})$ in the form of an expectation $\mathbb{E}(f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x})$ or a quantile $Q_\alpha(f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x})$ for some function $f : \mathbb{R}^d \to \mathbb{R}$, which can also be computed by solving the corresponding one-dimensional problems, additional interesting statistical functionals with intrinsically multivariate nature that are not that simple to estimate directly are accessible by DRF, such as, for example, the conditional correlations $\text{Cor}(Y_i, Y_j \mid \mathbf{X} = \mathbf{x})$. As an illustration, the estimated correlation of the sulfur dioxide (SO_2) and fine particulate matter (PM2.5) is shown in the right plot of Figure 5. The plot reveals also that the local correlation in many big cities is slightly larger than in its surroundings, which can be explained by the fact that the industrial production directly affects the levels of both pollutants.

A big advantage of the target-free forest construction of DRF is that all subsequent targets are computed from same the weighting function $w_{\mathbf{x}}$ obtained from a single forest fit. First, this is computationally more efficient, since we do not need for every target of interest to fit the method specifically tailored to it. For example, estimating the CDF with classification forests requires fitting one forest for each function value. Secondly and even more importantly, since all statistical functionals are plug-in estimates computed from the same weighting function, the obtained estimates are mathematically well-behaved and mutually compatible. For example, if we estimate $\text{Cor}(Y_i, Y_j \mid \mathbf{X} = \mathbf{x})$ by separately estimating the terms $\text{Cov}(Y_i, Y_j \mid \mathbf{X} = \mathbf{x})$, $\text{Var}(Y_i \mid \mathbf{X} = \mathbf{x})$, and $\text{Var}(Y_j \mid \mathbf{X} = \mathbf{x})$, one can not in general guarantee the estimate to be in the range $[-1, 1]$, but this is possible with DRF. Alternatively, the correlation or covariance matrices that are estimated entrywise are guaranteed to be positive semi-definite if one uses DRF. As an additional

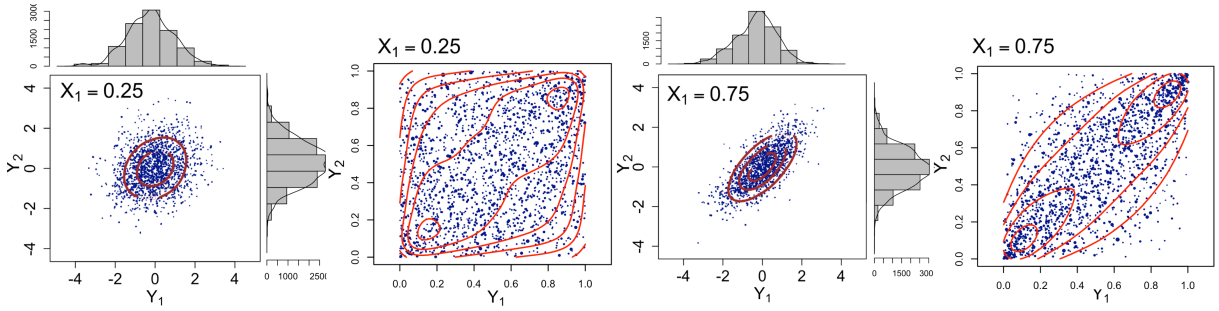


Figure 7: Estimated conditional joint distribution of (Y_1, Y_2) and conditional copulas obtained by DRF at different test points $\mathbf{x}$, where x_1 equals 0.25 and 0.75 respectively. The red lines are the contours of the true multivariate density function.

illustration, Figure 6 shows that the estimated (univariate) CDF using the classification forest need not be monotone due to random errors in each predicted value, which can not happen with the DRF estimates.

4.3 Conditional Copulas and Conditional Independence Testing

One can use the weighting function not only to estimate certain functionals, but also to obtain more complex objects, such as, for example, the conditional copulas. The well-known Sklar's theorem [159] implies that at a point $\mathbf{x} \in \mathbb{R}^p$, the conditional CDF $\mathbb{P}(\mathbf{Y} \leqslant \mathbf{y} \mid \mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_1 \leqslant y_1, \ldots, Y_d \leqslant y_d \mid \mathbf{X} = \mathbf{x})$ can be represented by a CDF $C_{\mathbf{x}}$ on $[0,1]^d$, the conditional copula at $\mathbf{x}$, and d conditional marginal CDFs $F_{Y_i|\mathbf{X}=\mathbf{x}}(y) = \mathbb{P}(Y_i \leqslant y \mid \mathbf{X} = \mathbf{x})$ for $1 \leqslant i \leqslant d$, as follows:

$$\mathbb{P}(\mathbf{Y} \leqslant \mathbf{y} \mid \mathbf{X} = \mathbf{x}) = C_{\mathbf{x}}\left(F_{Y_1|\mathbf{X}=\mathbf{x}}(y_1), \ldots, F_{Y_d|\mathbf{X}=\mathbf{x}}(y_d)\right). \tag{22}$$

Copulas capture the dependence of the components Y_i by the joint distribution of the corresponding quantile levels of the marginal distributions: $F_{Y_i|\mathbf{X}=\mathbf{x}}(Y_i) \in [0,1]$. Decomposing the full multivariate distribution to marginal distributions and the copula is a very useful technique used in many fields such as risk analysis or finance [30]. Using DRF enables us to estimate copulas conditionally, either by fitting certain parametric model or nonparametrically, directly from the weights.

To illustrate this, consider an example where the 5-dimensional $\mathbf{Y}$ is generated from the equicorrelated Gaussian copula $\mathbf{Y} = (Y_1, \ldots, Y_5) \mid \mathbf{X} = \mathbf{x} \sim C^{\text{Gauss}}_{\rho(\mathbf{x})}$ conditionally on the covariates $\mathbf{X}$ with distribution $\mathbf{X}_i \overset{i.i.d.}{\sim} U(0,1)^p$, where $p = 30$ and $n = 5000$. All Y_i have a $N(0,1)$ distribution marginally, but their conditional correlation for $i \neq j$ is given by $\text{Cor}(Y_i, Y_j) = \rho(\mathbf{x}) = x_1$. Figure 7 shows that DRF estimates the full conditional distribution at different test points $\mathbf{x}$ quite accurately and thus we can obtain a good nonparametric estimate of the conditional copula as follows. First, for each component Y_i, we compute the corresponding marginal CDF estimate $\hat{F}_{Y_i|\mathbf{X}=\mathbf{x}}(\cdot)$ from the weights. Second, we map each response $\mathbf{y}_i \rightarrow$

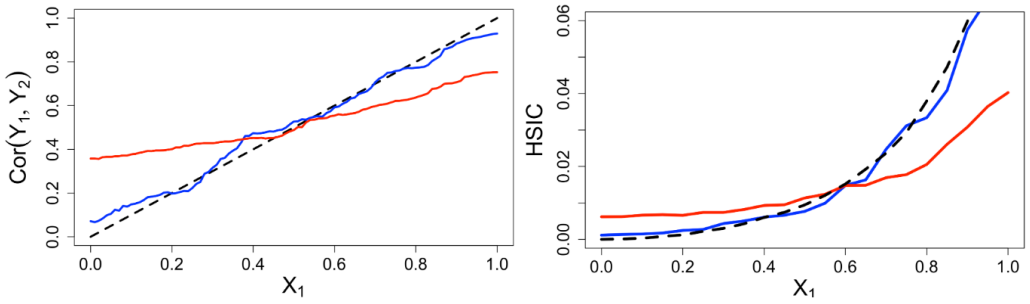


Figure 8: Estimated conditional correlation of Y_1 and Y_2 (left) and estimated conditional dependence quantified by HSIC statistic (right), obtained by DRF_{MMD} (blue) and DRF_{CART} (red) respectively. For every test point, we set $X_j = 0.5$, $j \neq 1$. Black dashed curve indicates the population values.

$\mathbf{u}_i := \left(\hat{F}_{Y_1|\mathbf{X}=\mathbf{x}}\left((\mathbf{y}_i)_1\right), \ldots, \hat{F}_{Y_d|\mathbf{X}=\mathbf{x}}\left((\mathbf{y}_i)_d\right)\right)$. The copula estimate is finally obtained from the weighted distribution $\sum_{i=1}^n w_{\mathbf{x}}(\mathbf{x}_i)\delta_{\mathbf{u}_i}$, from which we sample the points in Figure 7 in order to visualize the copula.

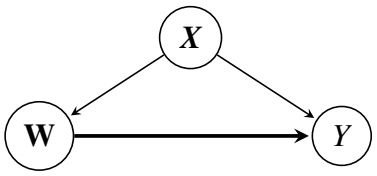
If we want to instead estimate the copula parametrically, we need to find the choice of parameters for a given model family which best matches the estimated conditional distribution, e.g. by weighted maximum likelihood estimation (MLE). For the above example, the correlation parameter of the Gaussian copula can be estimated by computing the weighted correlation with weights $\{w_{\mathbf{x}}(\mathbf{x}_i)\}_{i=1}^n$. The left plot in Figure 8 shows the resulting estimates of the conditional correlation $\text{Cor}\left(Y_1, Y_2 \mid \mathbf{X} = \mathbf{x}\right)$ obtained from DRF_{MMD}, which uses the MMD splitting criterion (12) described in Section 2.3.1, and DRF_{CART}, which aggregates the marginal CART criteria [92, 152]. We see that DRF_{MMD} is able to detect the distributional heterogeneity and provide good estimates of the conditional correlation. On the other hand, DRF_{CART} cannot detect the change in distribution of $\mathbf{Y}$ caused by X_1 that well. The distributional heterogeneity can not only occur in marginal distribution of the responses (a case extensively studied in the literature), but also in their interdependence structure described by the conditional copula $C_{\mathbf{x}}$, as one can see from decomposition (22). Since DRF_{MMD} relies on a distributional metric for its splitting criterion, it is capable of detecting any change in distribution [62], whereas aggregating marginal CART criteria for $Y_1, \ldots, Y_d$ in DRF_{CART} only captures the changes in the marginal means.

This is further illustrated for a related application of conditional independence testing, where we compute some dependence measure from the obtained weights. For example, we can test the independence $Y_1 \perp\!\!\!\perp Y_2$ conditionally on the event $\mathbf{X} = \mathbf{x}$ by using the Hilbert Schmidt Independence Criterion (HSIC) [64], which measures the difference between the joint distribution and the product of the marginal distributions. The right plot of Figure 8 shows that the DRF_{MMD} estimates are quite close to the population value of the HSIC, unlike the ones obtained by DRF_{CART}.

4.4 Heterogeneous Regression and Causal Effect Estimation

In this and the following section, we illustrate that, in addition to direct estimation of certain targets, DRF can also be a useful tool for complex statistical problems and applications, such as causality.

Suppose we would like to investigate the relationship between some (univariate) quantity of interest Y and certain predictors $\mathbf{W}$ from heterogeneous data, where the change in distribution of $(\mathbf{W}, Y)$ can be explained by some other covariates $\mathbf{X}$. Very often in causality applications, $\mathbf{W}$ is a (multivariate) treatment variable, Y is the outcome, which is commonly, but not necessarily, binary, and $\mathbf{X}$ is a set of observed confounding variables for which we need to adjust if we are interested in the causal effect of $\mathbf{W}$ on Y. This is illustrated by the following causal graph:



The problem of nonparametric confounding adjustment is hard; not only can the marginal distributions of Y and $\mathbf{W}$ be affected by $\mathbf{X}$, thus inducing spurious associations due to confounding, but the way how $\mathbf{W}$ affects Y can itself depend on $\mathbf{X}$, i.e. the treatment effect might be heterogeneous. The total causal effect can be computed by using the adjustment formula [133]:

$$\begin{aligned}\mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w})] &= \int \mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w}), \mathbf{X}=\mathbf{x}]\, \mathbb{P}(\mathbf{X}=\mathbf{x} \mid do(\mathbf{W}=\mathbf{w}))d\mathbf{x} \\ &= \int \mathbb{E}[Y \mid \mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x}]\, \mathbb{P}(\mathbf{X}=\mathbf{x})d\mathbf{x}. \end{aligned} \tag{23}$$

In general, implementing do-calculus for finite samples and potentially non-discrete data might not be straightforward and comes with certain difficulties. In this case, the standard approach would be to estimate the conditional mean $\mathbb{E}[Y \mid \mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x}]$ nonparametrically by regressing Y on $(\mathbf{X}, \mathbf{W})$ with some method of choice and to average out the estimates over different $\mathbf{x}$ sampled from the observed distribution of $\mathbf{X}$. Using DRF for this approach is not necessary, but has an advantage that one can easily estimate the full interventional distribution $\mathbb{P}(Y \mid do(\mathbf{W}=\mathbf{w}))$ and not only the interventional mean $\mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w})]$.

Another way of computing the causal effect, which allows to add more structure to the problem, is explained in the following: We use DRF to first fit the forest with the multivariate response $(\mathbf{W}, Y)$ and the predictors $\mathbf{X}$. In this way, one can for any point of interest $\mathbf{x}$ obtain the joint distribution of $(\mathbf{W}, Y)$ conditionally on the event $\mathbf{X}=\mathbf{x}$ and then the weights $\{w_{\mathbf{x}}(\mathbf{x}_i)\}_{i=1}^n$ can be used as an input for some regression method for regressing Y on $\mathbf{W}$ in the second step. This conditional regression fit might be of an independent interest, but it can also be used for estimating the causal effect $\mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w})]$ from (23), by averaging the estimates

$\mathbb{E}[Y \mid \mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x}]$ over $\mathbf{x}$, where $\mathbf{x}$ is sampled from the empirical observation of $\mathbf{X}$. In this way one can efficiently exploit and incorporate any prior knowledge of the relationship between $\mathbf{W}$ and Y, such as, for example, monotonicity, smoothness or that it satisfies a certain parametric regression model, without imposing any assumptions on the effect of $\mathbf{X}$ on $(\mathbf{W}, Y)$. Furthermore, one might be able to better extrapolate to the regions of space where $\mathbb{P}(\mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x})$ is small, compared to the standard approach which computes $\mathbb{E}[Y \mid \mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x}]$ directly, by regressing Y on $(\mathbf{W}, \mathbf{X})$. Extrapolation is crucial for causal applications, since for computing $\mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w})]$ we are interested in what would happen with Y when our treatment variable $\mathbf{W}$ is set to be $\mathbf{w}$, regardless of the value achieved by $\mathbf{X}$. However, it can easily happen that for this specific combination of $\mathbf{X}$ and $\mathbf{W}$ there are very few observed data points, thus making the estimation of the causal effect hard [133].

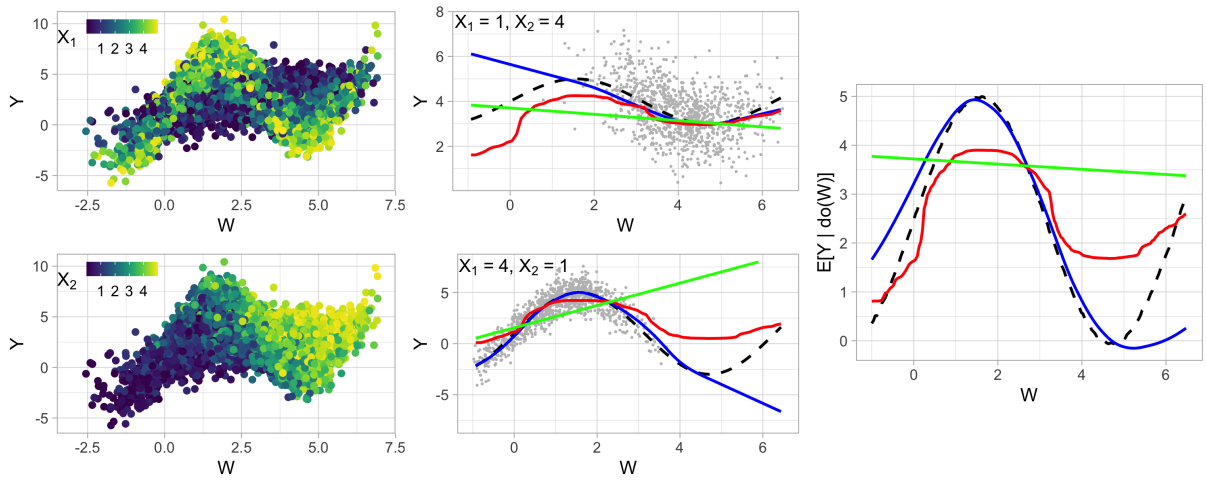


Figure 9: Left: Visualization of heterogeneous synthetic example (24). Middle: Gray points depict joint distribution of (W, Y) conditionally on $\mathbf{X}=\mathbf{x}$, for some choices of $\mathbf{x}$ indicated in the top left corner. Black curve indicates the true conditional mean $\mathbb{E}[Y \mid W=w, \mathbf{X}=\mathbf{x}]$, the blue curve represents the estimate obtained by DRF with response (W, Y) and predictors $\mathbf{X}$ in combination with smoothing splines regression, the red curve represents the estimate obtained by standard Random Forest, whereas the green line shows the estimate of the Causal Forest [4] which makes the linearity assumption and is thus misspecified. Right: The corresponding estimates for all the methods of the causal effect $\mathbb{E}[Y \mid do(W=w)]$ computed from (23). The true causal effect is denoted by a black dashed curve.

As an illustration, we consider the following synthetic data example, with continuous outcome Y, continuous univariate treatment W, $n = 5000$ and $p = 20$:

$$\mathbf{X} \sim U(0,5)^p, \quad W \mid \mathbf{X} \sim N(X_2, 1), \quad Y \mid \mathbf{X}, W \sim N(X_2 + X_1 \sin(W), 1). \tag{24}$$

A visualization of the data can be seen on the left side of Figure 9; treatment W affects Y nonlinearly, X_2 is a confounding variable that affects the marginal distributions of Y and W and X_1 makes the treatment effect heterogeneous. The middle part of Figure 9 shows the conditional regression fits, i.e. the estimates of $\mathbb{E}[Y \mid W=w, \mathbf{X}=\mathbf{x}]$ as w varies and $\mathbf{x}$ is fixed. In general,

the conditional regression fit is related to the concept of the conditional average treatment effect (CATE) as it quantifies the effect of $\mathbf{W}$ on Y for the subpopulation for which $\mathbf{X} = \mathbf{x}$. We see that combination of DRF with response (Y, W) and predictors $\mathbf{X}$ with the smoothing splines regression of Y on W (blue curve) is more accurate than the estimates obtained by standard Random Forest [15] with response Y and predictors $(W, \mathbf{X})$ (red curve). Furthermore, we see that the former approach can extrapolate better to regions with small number of data points, which enables us to better estimate the causal effect $\mathbb{E}[Y \mid do(W = w)]$ from (23), by averaging the corresponding estimates of $\mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]$ over observed $\mathbf{x}$, as shown in the right plot of Figure 9.

There exist many successful methods in the literature for estimating the causal effects and the (conditional) average treatment effects for a wide range of settings [1, 28, 181, 96]. However, some methods are not designed for the most general case and make certain modeling assumptions or are designed specifically for the (very common) case where the treatment variable is univariate or even binary. Due to its versatility, DRF can easily be used when the underlying assumptions of conventional methods are violated, when some additional structure is given in the problem or for the general, nonparametric, settings [76, 49, 85]. Appendix 4 contains additional comparisons with some existing methods for causal effect estimation.

4.4.1 Births data

We further illustrate the applicability of DRF for causality-related problems on the natality data obtained from the Centers for Disease Control and Prevention (CDC) website, where we have information about all recorded births in the USA in 2018. We investigate the relationship between the pregnancy length and the birthweight, an important indicator of baby's health. Not only is this relationship complex, but it also depends on many different factors, such as parents' race, baby's gender, birth multiplicity (single, twins, triplets...) etc. In the left two plots of Figure 10 one can see the estimated joint distribution of birthweight and pregnancy length conditionally on many different covariates, as indicated in the plot. The black curves denote the subsequent regression fit, based on smoothing splines. In addition to the estimate of the mean, indicated by the solid curve, we also include the estimates of the conditional 0.1- and 0.9-quantiles, indicated by dashed curves, which is very useful in practice for determining whether a baby is large or small for its gestational age. Notice how DRF assigns less importance to the mother's race when the point of interest is a twin (middle plot), as in this case more weight is given to twin births, regardless of the race of the parents.

Suppose now we would like to understand how a twin birth T causally affects the birthweight B, but ignoring the obvious indirect effect due to shorter pregnancy length L. For example, sharing of resources between the babies might have some effect on their birthweight. We additionally need to be careful to adjust for other confounding variables $\mathbf{X}$, such as, for exam-

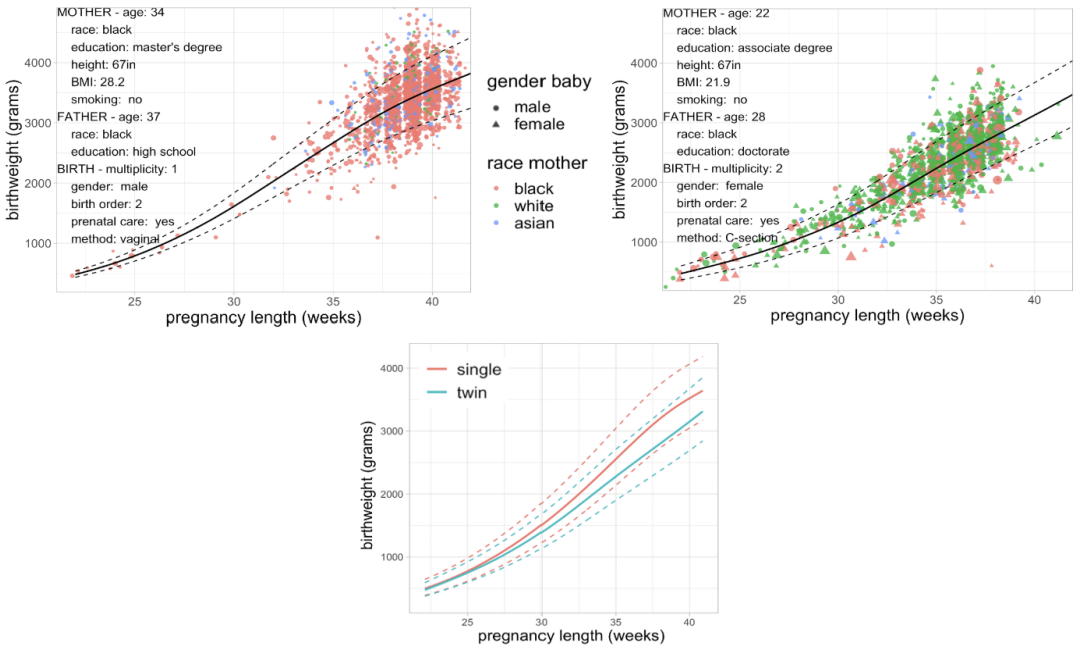
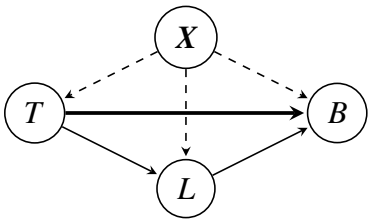


Figure 10: Above: estimated relationship of pregnancy length and birthweight, conditionally on the criteria indicated in the upper left corner. Below: estimated interventional effect of twin birth on the birthweight for a fixed pregnancy length. In all plots the solid curves denote the estimated conditional mean and the dashed denote the estimated 0.1 and 0.9 quantiles.

ple, the parents' race, which can affect B, T and L. We assume that this is represented by the following causal graph:



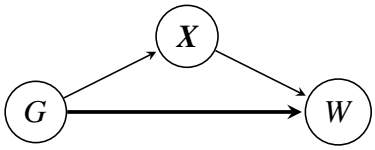
In order to answer the above question, we investigate the causal quantity $\mathbb{P}(B \mid do(T = t, L = l))$. Even though one cannot make such do-intervention in practice, this quantity describes the total causal effect if the birth multiplicity and the length of the pregnancy could be manipulated and thus for a fixed pregnancy length l, we can see the difference in birthweight due to T. We compute this quantity as above, by using DRF with subsequent regression fits, which has the advantage of better extrapolating to regions with small probability, such as long twin pregnancies (see the middle plot of Figure 10). In the right plot of Figure 10 we show the mean and quantiles of the estimated interventional distribution and we see that, as one might expect, a twin birth causes smaller birthweight on average, with the difference increasing with the length of the pregnancy.

4.5 Fairness

Being able to compute different causal quantities with DRF could prove useful in a range of applications, including fairness [97]. We investigate the data on approximately 1 million full-time employees from the 2018 American Community Survey by the US Census Bureau from which we have extracted the salary information and all covariates that might be relevant for salaries. In the bottom left plot of Figure 11 one can see the distribution of hourly salary of men and women (on the logarithmic scale). The overall salary was scaled with working hours to account for working part-time and for the fact that certain jobs have different working hours. We can see that men are paid more in general, especially for the very high salaries. The difference between the median hourly salaries, a commonly used statistic in practice, amounts 17% for this data set.

We would like to answer whether the observed gender pay gap in the data is indeed unfair, i.e. only due to the gender, or whether it can at least in part be explained by some other factors, such as age, job type, number of children, geography, race, attained education level and many others. Hypothetically, it could be, for example, that women have a preference for jobs that are paid less, thus causing the gender pay gap.

In order to answer this question, we assume that the data is obtained from the following causal graph, where G denotes the gender, W the hourly wage and all other factors are denoted by $\mathbf{X}$:



i.e. G is a source node and W is a sink node in the graph. In order to determine the direct effect of the gender on wage that is not mediated by other factors, we would like to compute the distribution of the nested counterfactual $W(\text{male}, \mathbf{X}(\text{female}))$, which is interpreted as the women's wage had they been treated in same way as men by their employers for determining the salary, but without changing their propensities for other characteristics, such as the choice of occupation [29]. Therefore, it can be obtained from the observed distribution as follows:

$$\begin{aligned}\mathbb{P}\left(W(\text{male}, \mathbf{X}(\text{female}))\right) &= \int \mathbb{P}\left(W(G=\text{male}, \mathbf{X}=\mathbf{x})\right) \mathbb{P}(\mathbf{X}=\mathbf{x} \mid G=\text{female}) d\mathbf{x} \\ &= \int \mathbb{P}\left(W \mid G=\text{male}, \mathbf{X}=\mathbf{x}\right) \mathbb{P}(\mathbf{X}=\mathbf{x} \mid G=\text{female}) d\mathbf{x}, \end{aligned} \tag{25}$$

Put in the language of the fairness literature, it quantifies the unfairness when all variables $\mathbf{X}$ are assumed to be resolving [86], meaning that any difference in salaries directly due to factors $\mathbf{X}$ is not viewed as gender discrimination. For example, one does not consider unfair if people

with low education level get lower salaries, even if the gender distribution in this group is not balanced.

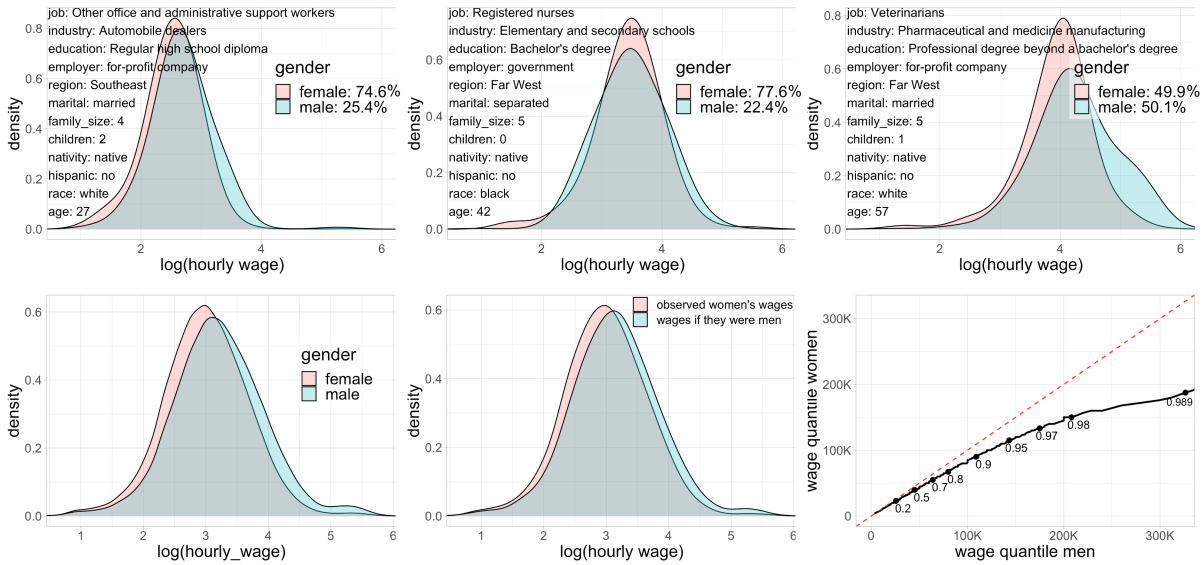


Figure 11: Top row: Estimated joint distribution of wage and gender for some fixed values of other covariates $\mathbf{X}$ indicated in the top left part of each plot. Bottom row: observed overall distribution of salaries (left), estimated counterfactual distribution $\mathbb{P}\left(W(\text{male}, \mathbf{X}(\text{female}))\right)$ of women's salaries (middle) and the quantile comparison of the counterfactual distribution of women's salaries and the observed distribution of men's salaries (right).

There are several ways how one can compute the distribution of $W(\text{male}, \mathbf{X}(\text{female}))$ from (25) with DRF. The most straightforward option is to take W as the response and $(G, \mathbf{X})$ as predictors in order to compute the conditional distribution $\mathbb{P}\left(W \mid G = \text{male}, \mathbf{X} = \mathbf{x}\right)$. However, with this approach it could happen that for predicting $\mathbb{P}\left(W \mid G = \text{male}, \mathbf{X} = \mathbf{x}\right)$ we also assign weight to training data points for which $G = \text{female}$. This happens if in some trees we did not split on variable G, which is likely, for example, if $\mathbb{P}(G = \text{male} \mid \mathbf{X} = \mathbf{x})$ is low. Using salaries of both genders to estimate the distribution of men's salaries might be an issue if our goal is to objectively compare how women and men are paid.

Another approach is to take (W, G) as a multivariate response and $\mathbf{X}$ as the predictors for DRF and thus obtain joint distribution of (W, G) conditionally on the event $\mathbf{X} = \mathbf{x}$. In this way we can also quantify the gender discrimination of a single individual with characteristics $\mathbf{x}$ by comparing his/her salary to the corresponding quantile of the salary distribution of people of the opposite gender with the same characteristics $\mathbf{x}$ [135]. This is interesting because the distribution of salaries, and thus also the gender discrimination, can be quite different depending on other factors such as the industry sector or job type, as illustrated for a few choices of $\mathbf{x}$ in the top row of Figure 11.

Finally, by averaging the DRF estimates of $\mathbb{P}\left(W \mid \mathbf{X} = \mathbf{x}, G = \text{male}\right)$, conveniently represented via the weights, over different $\mathbf{x}$ sampled from the distribution $\mathbb{P}(\mathbf{X} \mid G = \text{female})$, we

can compute the distribution of the nested counterfactual $W(\text{male}, \mathbf{X}(\text{female}))$ [29]. In the middle panel in the bottom row of Figure 11 a noticeable difference in the means, also called natural direct effect in the causality literature [133], is still visible between the observed distribution of women's salaries and the hypothetical distribution of their salaries had they been treated as men, despite adjusting for indirect effects of the gender via covariates $\mathbf{X}$. By further matching the quantiles of the counterfactual distribution $\mathbb{P}\left(W(\text{male}, \mathbf{X}(\text{female}))\right)$ with the corresponding quantiles of the observed distribution of men's salaries in the bottom right panel of Figure 11, we can also see that the adjusted gender pay gap even increases for larger salaries. Median hourly wage for women is still 11% lower than the median wage for the hypothetical population of men with exactly the same characteristics $\mathbf{X}$ as women, indicating that only a minor proportion of the actually observed hourly wage difference of 17% can be explained by other demographic factors.

5 Conclusion

We have shown that DRF is a flexible, general and powerful tool, which exploits the well-known properties of the Random Forest as an adaptive nearest neighbor method via the induced weighting function. Not only does it estimate multivariate conditional distributions well, but it constructs the forest in a model- and target-free way and is thus an easy to use out-of-the-box algorithm for many, potentially complex, learning problems in a wide range of applications, including also causality and fairness, with competitive performance even for problems with existing tailored methods.

1 Implementation Details

Here we present the implementation of the Distributional Random Forests (DRF) in detail. The code is available in the R-package `drf` and the Python package `drf`. The implementation is based on the implementations of the R-packages `grf` [4] and `ranger` [187]. The largest difference is in the splitting criterion itself and the provided user interface. Algorithm 1 gives the pseudocode for the forest construction and computation of the weighting function $w_{\mathbf{x}}(\cdot)$.

- Every tree is constructed based on a random subset of size s (taken to be 50% of the size of the training set by default) of the training data set, similar to [181]. This differs from the original Random Forest algorithm [15], where the bootstrap subsampling is done by drawing from the original sample with replacement.

- The principle of honesty [10, 39, 181] is used for building the trees (line 4), whereby for each tree one first performs the splitting based on one random set of data points $\mathcal{S}_{\text{build}}$, and then populates the leaves with a disjoint random set $\mathcal{S}_{\text{populate}}$ of data points for determining the weighting function $w_{\mathbf{x}}(\cdot)$. This prevents overfitting, since we do not assign weight to the data points which we used to built the tree.

- We borrow the method for selecting the number of candidate splitting variables from the `grf` package [4]. This number is randomly generated as $\min(\max(\text{Poisson}(\text{mtry}), 1), p)$, where mtry is a tuning parameter. This differs from the original Random Forests algorithm, where the number of splitting candidates is fixed to be mtry.

- The number of trees built is $N = 2000$ by default.

- The factor variables in both the responses and the predictors are encoded by using the one-hot encoding, where we add an additional indicator variable for each level l of some factor variable X_k. This implies that in the building step, if we split on this indicator variable, we divide the current set of data points in the sets where $X_k = l$ and $X_k \neq l$. This works well if the number of levels is not too big, since otherwise one makes very uneven splits and the dimensionality of the problem increases significantly. Handling of categorical problems is a general challenge for the forest based methods and is an area of active research [82]. We will leave improving on this approach for the future development.

- We try to enforce splits where each child has at least a fixed percentage (chosen to be 10% as the default value) of the current number of data points. In this way we achieve balanced splits and reduce the computational time. However, we cannot enforce this if we are trying to split on the variable X_i with only a few unique values, e.g. indicator variable for a level of some factor variable.

Algorithm 1 Pseudocode for Distributional Random Forest

procedure BUILDFOREST(set of samples $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, number of trees N)
  for $i = 1, \ldots, N$ **do**
    $\mathcal{S}_{\text{subsample}} = \text{SUBSAMPLE}(\mathcal{S})$
    $\mathcal{S}_{\text{build}}, \mathcal{S}_{\text{populate}} \leftarrow \text{SPLITSAMPLES}(\mathcal{S}_{\text{subsample}})$ ▷ Honesty principle, see above
    $\mathcal{T}_i \leftarrow \text{CREATENEWTREE}(\mathcal{S}_{\text{build}})$ ▷ Samples $\mathcal{S}_{\text{build}}$ used for building the tree
    $\text{BUILDTREE}(\text{ROOTNODE}(\mathcal{T}_i))$ ▷ Start recursion from the root node
    $\text{POPULATELEAVES}(\mathcal{T}_i, \mathcal{S}_{\text{populate}})$ ▷ Samples $\mathcal{S}_{\text{populate}}$ used for computing $w_{\mathbf{x}}(\cdot)$
  end for
  return $\mathcal{F} = \{\mathcal{T}_1, \ldots, \mathcal{T}_N\}$
end procedure

procedure BUILDTREE(current node $\mathcal{N}$) ▷ Recursively constructs the trees
  if $\text{STOPPINGCRITERION}(\mathcal{N})$ **then** ▷ E.g. if only a few samples left
    return
  end if
  $\mathcal{S} \leftarrow \text{GETSAMPLES}(\mathcal{N})$
  $\mathcal{I} \leftarrow \text{GETSPLITVARIABLES}()$ ▷ Random set of candidate variables
  $\mathcal{C} \leftarrow \text{INITIALIZESPLITS}()$ ▷ Here we store info about candidate splits
  for idx $\in \mathcal{I}$, level l **do** ▷ l iterates over all values of variable X_{idx}
    $\mathcal{S}_L, \mathcal{S}_R \leftarrow \text{CHILDSAMPLES}(\mathcal{S}, \text{idx}, l)$ ▷ Split samples based on $(\mathbf{x}_i)_{\text{idx}} \leq l$
    test statistic $v = \text{SPLITTINGCRITERION}(\mathcal{S}_L, \mathcal{S}_R)$ ▷ Two-sample test of choice
    $\text{ADDNEWSPLITCANDIDATE}(\mathcal{C}, v, \mathcal{S}_L, \mathcal{S}_R, \text{idx}, l)$
  end for
  $\mathcal{S}_L, \mathcal{S}_R, \text{idx}, l \leftarrow \text{FINDBESTSPLIT}(\mathcal{C})$
  $\mathcal{N}_L \leftarrow \text{CREATENODE}(\mathcal{S}_L)$ ▷ Create new node with set of samples $\mathcal{S}_L$
  $\mathcal{N}_R \leftarrow \text{CREATENODE}(\mathcal{S}_R)$ ▷ Create new node with set of samples $\mathcal{S}_R$
  $\text{BUILDTREE}(\mathcal{N}_L), \quad \text{BUILDTREE}(\mathcal{N}_R)$ ▷ Proceed building recursively
  $\text{CHILDREN}(\mathcal{N}) \leftarrow \mathcal{N}_L, \mathcal{N}_R$
  $\text{SPLIT}(\mathcal{N}) \leftarrow \text{idx}, l$ ▷ Store the split
  return
end procedure

procedure GETWEIGHTS(forest $\mathcal{F}$, test point $\mathbf{x}$) ▷ Computes the weighting function
  vector of weights $w = \text{ZEROS}(n)$ ▷ n is the training set size
  for $i = 1, \ldots, |\mathcal{F}|$ **do**
    $\mathcal{L} = \text{GETLEAFSAMPLES}(\mathcal{T}_i, \mathbf{x})$ ▷ indices of training samples in same leaf as $\mathbf{x}$
    for idx $\in \mathcal{L}$ **do**
      $w[\text{idx}] = w[\text{idx}] + 1/(|\mathcal{L}| \cdot |\mathcal{F}|)$
    end for
  end for
  return w
end procedure

- All components of the response Y are scaled for the building step (but not when we populate the leaves). This ensures that each component of the response contributes equally to the kernel values, and consequently to the MMD two-sample test statistic. Plain usage of the MMD two-sample test would scale the components of Y at each node. However, this approach favors always splitting on the same variables, even though their effect will diminish significantly after having split several times.

- By default, in step 20 of the Algorithm 1, we use the MMD-based splitting criterion given by

$$\frac{1}{B}\sum_{k=1}^{B}\frac{|\mathcal{S}_L||\mathcal{S}_L|}{(|\mathcal{S}_L|+|\mathcal{S}_R|)^2}\left|\frac{1}{|\mathcal{S}_L|}\sum_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{S}_L}\varphi_{\omega_k}(\mathbf{y}_i)-\frac{1}{|\mathcal{S}_R|}\sum_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{S}_R}\varphi_{\omega_k}(\mathbf{y}_i)\right|^2.$$

  The Gaussian kernel $k(\mathbf{x},\mathbf{y}) = \frac{1}{(\sqrt{2\pi}\sigma)^d}e^{\frac{-\|x-y\|_2^2}{2\sigma^2}}$ is used as the default choice, with the bandwidth σ chosen as the median pairwise distance between all training responses $\{\mathbf{y}_i\}_{i=1}^n$, commonly referred to as the 'median heuristic' [66]. However the algorithm can be used with any choice of kernel, or in fact with any two-sample test.

- The number B of random Fourier features is fixed and taken to be 20 by default. The performance of the trees empirically shows stability for large range of B. Smaller values of B help making the trees more independent which could improve the performance. One could even use an adaptive strategy of choosing B, possibly increasing B as the depth of the tree increases, but we decided to keep B fixed for simplicity.

- We compute variable importance similarly as for the original Random Forest algorithm [15, 187], by sequentially permuting each variable and investigating the decrease in the performance. However, since we target the full conditional distribution of the multivariate response, as our performance measure we use for every test point $(\mathbf{x},\mathbf{y})$ the MMD distance between the estimated joint distribution $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$, described by the DRF weights, and the point mass $\delta_{\mathbf{y}}$.

2 Derivations and Proofs

In this section we present proofs and further details for the results in Sections 2.3.1 and 3, in the order in which they appear. Sections 2.1 and 2.3 derive the splitting criterion presented in Equation (12) in the main text, with Section 2.3 showing that the CART criterion can be written in an analogous way. Section 2.4 provides background and proves to the statements in Section 3.

2.1 Expressing MMD test statistic as an integral in the feature space

The biased MMD two-sample statistic is given as

$$\begin{aligned}
&\mathcal{D}_{\text{MMD}}\left(\{\mathbf{u}_i\}_{i=1}^m, \{\mathbf{v}_i\}_{i=1}^n\right) \\
&= \frac{1}{m^2}\sum_{i,j} k(\mathbf{u}_i, \mathbf{u}_j) + \frac{1}{n^2}\sum_{i,j} k(\mathbf{v}_i, \mathbf{v}_j) - \frac{2}{mn}\sum_i\sum_j k(\mathbf{u}_i, \mathbf{v}_j) \\
&= \frac{1}{m^2}\sum_{i,j} k(\mathbf{u}_i, \mathbf{u}_j) + \frac{1}{n^2}\sum_{i,j} k(\mathbf{v}_i, \mathbf{v}_j) - \frac{1}{mn}\sum_i\sum_j k(\mathbf{u}_i, \mathbf{v}_j) - \frac{1}{mn}\sum_i\sum_j k(\mathbf{v}_j, \mathbf{u}_i).
\end{aligned}$$

Assume that the kernel k is bounded and shift-invariant, then by Bochner's theorem there exist a measure ν such that k can be written as $k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{i\omega^T(\mathbf{x}-\mathbf{y})} d\nu(\boldsymbol{\omega})$.

Let us write $\varphi_\omega^U = \frac{1}{m}\sum_i e^{i\omega^T \mathbf{u}_i}$ and $\varphi_\omega^V = \frac{1}{n}\sum_i e^{i\omega^T \mathbf{v}_i}$. We can now write $\mathcal{D}_{\text{MMD}}$ as

$$\begin{aligned}
\mathcal{D}_{\text{MMD}}\left(\{\mathbf{u}_i\}_{i=1}^m, \{\mathbf{v}_i\}_{i=1}^n\right) &= \int_{\mathbb{R}^d} \left(\varphi_\omega^U \overline{\varphi_\omega^U} + \varphi_\omega^V \overline{\varphi_\omega^V} - \varphi_\omega^U \overline{\varphi_\omega^V} - \varphi_\omega^V \overline{\varphi_\omega^U}\right) d\nu(\boldsymbol{\omega}) \\
&= \int_{\mathbb{R}^d} \left|\varphi_\omega^U - \varphi_\omega^V\right|^2 d\nu(\boldsymbol{\omega}) \\
&= \int_{\mathbb{R}^d} \left|\frac{1}{m}\sum_{i=1}^m \varphi_\omega(\mathbf{u}_i) - \frac{1}{n}\sum_{i=1}^n \varphi_\omega(\mathbf{v}_i)\right|^2 d\nu(\boldsymbol{\omega}),
\end{aligned}$$

where $\varphi_\omega(\mathbf{y}) = e^{i\omega^T \mathbf{y}} \in \mathbb{C}$ are the corresponding Fourier features, which is what we wanted to show.

2.2 Approximate kernel and its MMD

When the kernel k is bounded and shift invariant, we have seen that it can be written as $k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{i\omega^T(\mathbf{x}-\mathbf{y})} d\nu(\boldsymbol{\omega})$. This integral can be approximated by sampling from ν: Let $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_B \sim \nu$ be a random sample from the measure ν. Then we can write

$$k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{i\omega^T(\mathbf{x}-\mathbf{y})} d\nu(\boldsymbol{\omega}) \approx \frac{1}{B}\sum_{b=1}^B e^{i\omega_b{}^T(\mathbf{x}-\mathbf{y})} = \frac{1}{B}\langle \boldsymbol{\varphi}(\mathbf{u}), \boldsymbol{\varphi}(\mathbf{v})\rangle_{\mathbb{C}^B} := \tilde{k}(\mathbf{u}, \mathbf{v}),$$

where $\boldsymbol{\varphi}(\mathbf{u}) = (\varphi_{\omega_1}(\mathbf{u}), \ldots, \varphi_{\omega_B}(\mathbf{u}))^T$ is a random complex vector consisting of the Fourier features $\varphi_\omega(\mathbf{u}) = e^{i\omega^T \mathbf{u}} \in \mathbb{C}$. The kernel $\tilde{k}$ is analogous to the kernel k, but where the measure ν is replaced by the empirical measure $\tilde{\nu} = \frac{1}{B}\sum_{b=1}^B \delta_{\omega_b}$:

$$\tilde{k}(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{i\omega^T(\mathbf{x}-\mathbf{y})} d\tilde{\nu}(\boldsymbol{\omega}).$$

Analogously as in the section 2.1, we can now write the MMD for the kernel $\tilde{k}$ as:

$$\begin{aligned}\mathcal{D}_{\mathrm{MMD}(\tilde{k})} &= \int_{\mathbb{R}^d} \left| \frac{1}{m}\sum_{i=1}^{m} \varphi_{\omega}(\mathbf{u}_i) - \frac{1}{n}\sum_{i=1}^{n} \varphi_{\omega}(\mathbf{v}_i) \right|^2 d\tilde{\nu}(\boldsymbol{\omega}) \\ &= \frac{1}{B}\sum_{b=1}^{B} \left| \frac{1}{m}\sum_{i=1}^{m} \varphi_{\omega_b}(\mathbf{u}_i) - \frac{1}{n}\sum_{i=1}^{n} \varphi_{\omega_b}(\mathbf{v}_i) \right|^2,\end{aligned}$$

which can also additionally be interpreted as the approximation of $\mathcal{D}_{\mathrm{MMD}}$. Therefore, our splitting criterion is obtained as the MMD of the random approximate kernel $\tilde{k}$:

$$\frac{1}{B}\sum_{b=1}^{B} \frac{n_L n_R}{n_P^2} \left| \frac{1}{n_L}\sum_{\mathbf{x}_i \in C_L} \varphi_{\omega_b}(\mathbf{y}_i) - \frac{1}{n_R}\sum_{\mathbf{x}_i \in C_R} \varphi_{\omega_b}(\mathbf{y}_i) \right|^2.$$

The scaling factor $\frac{n_L n_R}{n_P^2}$ occurs naturally and penalizes the increased variance of the sample MMD statistic when n_L or n_R are small: it appears when we rewrite the CART criterion in the related form, see section 2.3.

This representation of the MMD is the key why we use the approximate kernel $\tilde{k}$ instead of k. This splitting criterion can be computed in $O(Bn_P)$ complexity, by updating the sums $\sum_{\mathbf{x}_i \in C_L} \varphi_{\omega_k}(\mathbf{y}_i)$ and $\sum_{\mathbf{x}_i \in C_R} \varphi_{\omega_k}(\mathbf{y}_i)$ in $O(1)$ computations, whereas this is not possible for $\mathcal{D}_{\mathrm{MMD}}$.

2.3 CART criterion rewritten

Standard CART criterion used in Random Forests [15] is the following: we repeatedly choose to split the parent node P of size n_P in two children C_L and C_R, of sizes n_L and n_R respectively, such that the expression

$$\frac{1}{n_P}\left(\sum_{i \in C_L} (Y_i - \overline{Y}_L)^2 + \sum_{i \in C_R} (Y_i - \overline{Y}_R)^2 \right) \tag{26}$$

is minimized, where $\overline{Y}_L = \frac{1}{n_L}\sum_{i \in C_L} Y_i$ and Y_R is defined similarly.

We now have $\overline{Y} = \frac{1}{n_P}\sum_{i \in P} Y_i = \frac{n_L}{n_P}\overline{Y}_L + \frac{n_R}{n_P}\overline{Y}_R$, which gives $\overline{Y} - \overline{Y}_L = \frac{n_R}{n_P}(\overline{Y}_R - \overline{Y}_L)$, so we

can write

$$\begin{aligned}\sum_{i\in C_L}(Y_i-\overline{Y}_L)^2 &= \sum_{i\in C_L}(Y_i-\overline{Y}+\overline{Y}-\overline{Y}_L)^2 = \sum_{i\in C_L}(Y_i-\overline{Y}+\frac{n_R}{n_P}(\overline{Y}_R-\overline{Y}_L))^2\\ &= \sum_{i\in C_L}(Y_i-\overline{Y})^2+2\frac{n_R}{n_P}(\overline{Y}_R-\overline{Y}_L)\sum_{i\in C_L}(Y_i-\overline{Y})+\frac{n_Ln_R^2}{n_P^2}(\overline{Y}_R-\overline{Y}_L)^2\\ &= \sum_{i\in C_L}(Y_i-\overline{Y})^2+2\frac{n_R}{n_P}(\overline{Y}_R-\overline{Y}_L)\cdot n_L(\overline{Y}_L-\overline{Y})+\frac{n_Ln_R^2}{n_P^2}(\overline{Y}_R-\overline{Y}_L)^2\\ &= \sum_{i\in C_L}(Y_i-\overline{Y})^2+2\frac{n_Rn_L}{n_P}(\overline{Y}_R-\overline{Y}_L)\cdot\frac{n_R}{n_P}(\overline{Y}_L-\overline{Y}_R)+\frac{n_Ln_R^2}{n_P^2}(\overline{Y}_R-\overline{Y}_L)^2\\ &= \sum_{i\in C_L}(Y_i-\overline{Y})^2-\frac{n_Ln_R^2}{n_P^2}(\overline{Y}_R-\overline{Y}_L)^2.\end{aligned}$$

Similarly we obtain

$$\sum_{i\in C_R}(Y_i-\overline{Y}_R)^2 = \sum_{i\in C_R}(Y_i-\overline{Y}+\overline{Y}-\overline{Y}_R)^2 = \sum_{i\in C_R}(Y_i-\overline{Y})^2-\frac{n_L^2n_R}{n_P^2}(\overline{Y}_R-\overline{Y}_L)^2,$$

which gives us that the CART criterion (26) can be written as

$$\begin{aligned}&\frac{1}{n_P}\left(\sum_{i\in C_L}(Y_i-\overline{Y})^2-\frac{n_Ln_R^2}{n_P^2}(\overline{Y}_R-\overline{Y}_L)^2+\sum_{i\in C_R}(Y_i-\overline{Y})^2-\frac{n_L^2n_R}{n_P^2}(\overline{Y}_R-\overline{Y}_L)^2\right)\\ &\qquad=\frac{1}{n_P}\sum_{i\in P}(Y_i-\overline{Y})^2-\frac{n_Ln_R}{n_P^2}(\overline{Y}_R-\overline{Y}_L)^2,\end{aligned}$$

since $n_L+n_R=n_P$. Since the first term depends only on the parent node and not on the chosen split, we conclude that minimizing the CART criterion (26) is equivalent to maximizing the following expression

$$\frac{n_Ln_R}{n_P^2}(\overline{Y}_L-\overline{Y}_R)^2. \tag{27}$$

This equivalent criterion can be interpreted as comparing the difference in the means of the resulting child nodes, i.e. we will choose the split such that the means in the child nodes are as heterogeneous as possible. The scaling factor $\frac{n_Ln_R}{n_P^2}$ appears naturally, penalizing uneven splits due to the increased variance of $\overline{Y}_L$ or $\overline{Y}_R$.

2.4 Proofs for Section 3

2.4.0.1 Preliminaries. We first set notation and define basic probabilistic concepts on the separable Hilbert space $(\mathcal{H},\langle\cdot,\cdot\rangle_{\mathcal{H}})$. We thereby mostly refer to [75] and [134]. The initial results derived here parallel some of the results derived in [132], but where derived independently. Let $(\Omega,\mathcal{A},\mathbb{P})$ be the underlying probability space. Let $(\mathcal{H},\langle\cdot,\cdot\rangle_{\mathcal{H}})$ be the Hilbert space induced

by the kernel k and $\mu : \mathcal{M}_b(\mathbb{R}^d) \to \mathcal{H}$ be the embedding function of, $\mu(P) \in \mathcal{H}$ for all bounded signed Borel measures P on $\mathbb{R}^d$. Throughout we assume that k is *bounded* and *continuous* in its two arguments. Boundedness of k ensures that μ is indeed defined on all of $\mathcal{M}_b(\mathbb{R}^d)$, while continuity of $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ ensures $\mathcal{H}$ is *separable*. Thus measurability issues can be avoided, in particular, a map $\xi : (\Omega, \mathcal{A}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is measurable iff $\langle \xi, f \rangle_{\mathcal{H}}$ is measurable for all $f \in \mathcal{H}$. Moreover, a quick check reveals that $\mu(P)$ is linear on $\mathcal{M}_b(\mathbb{R}^d)$. If $\mathbb{E}[\|\xi\|_{\mathcal{H}}] < \infty$, we define

$$\mathbb{E}[\xi] := \int_\Omega \xi d\mathbb{P} \in \mathcal{H},$$

where the integral is meant in a Bochner sense. Separability and $\mathbb{E}[\|\xi\|_{\mathcal{H}}] < \infty$ mean this integral is well-defined and moreover

$$F(\mathbb{E}[\xi]) = \mathbb{E}[F(\xi)],$$

for any continuous linear function $F : \mathcal{H} \to \mathbb{R}$.[1] In particular, $\mathbb{E}[\langle \xi, f \rangle_{\mathcal{H}}] = \langle \mathbb{E}[\xi], f \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Define moreover for $q \geqslant 1$, and $\xi, \xi_1, \xi_2 \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H})$,

$$\begin{aligned}
\mathcal{L}^q(\Omega, \mathcal{A}, \mathcal{H}) &= \{\xi : (\Omega, \mathcal{F}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H})) \text{ measurable, with } \mathbb{E}[\|\xi\|^q] < \infty]\} \\
\mathbb{L}^q(\Omega, \mathcal{A}, \mathcal{H}) &= \text{Set of equivalence classes in } \mathcal{L}^q(\Omega, \mathcal{A}, \mathcal{H}) \\
\mathrm{Var}(\xi) &:= \mathbb{E}[\|\xi - \mathbb{E}[\xi]\|^2] = \mathbb{E}[\|\xi\|^2] - \|\mathbb{E}[\xi]\|^2, \ \ \xi \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}) \\
\mathrm{Cov}(\xi_1, \xi_2) &= \mathbb{E}[\langle \xi_1 - \mathbb{E}[\xi_1], \xi_2 - \mathbb{E}[\xi_2] \rangle_{\mathcal{H}}] = \mathbb{E}[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}}] - \langle \mathbb{E}[\xi_1], \mathbb{E}[\xi_2] \rangle_{\mathcal{H}}.
\end{aligned}$$

It is well-known, that $(\mathbb{L}^q, \|\cdot\|_{\mathbb{L}^q(\mathcal{H})})$ is a Banach space, with

$$\|\xi\|_{\mathbb{L}^q(\mathcal{H})} = \mathbb{E}[\|\xi\|_{\mathcal{H}}^q]^{1/q}.$$

We can then also define *conditional* expectation. For a sub $\sigma-$ algebra $\mathcal{F} \subset \mathcal{A}$, $\xi \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H})$, $\mathbb{E}[\xi \mid \mathcal{F}]$ is the (a.s.) unique element such that

(C1) $\mathbb{E}[\xi \mid \mathcal{F}] : (\Omega, \mathcal{F}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is measurable and $\mathbb{E}[\xi \mid \mathcal{F}] \in \mathbb{L}^1(\Omega, \mathcal{F}, \mathcal{H})$,

(C2) $\mathbb{E}[\xi \mathbb{1}_F] = \mathbb{E}[\mathbb{E}[\xi \mid \mathcal{F}] \mathbb{1}_F]$ for all $F \in \mathcal{F}$.

See e.g. [173] or [134, Chapter 1]. (C2) in particular means that $\mathbb{E}[\mathbb{E}[\xi \mid \mathcal{F}]] = \mathbb{E}[\mathbb{E}[\xi \mid \mathcal{F}] \mathbb{1}_\Omega] = \mathbb{E}[\xi]$, since $\Omega \in \mathbb{F}$ for any σ-algebra. It can also be shown that $F(\mathbb{E}[\xi \mid \mathcal{F}]) = \mathbb{E}[F(\xi) \mid \mathcal{F}]$ for all linear and continuous $F : \mathcal{H} \to \mathbb{R}$ and that $\|\mathbb{E}[\xi \mid \mathcal{F}]\|_{\mathcal{H}} \leqslant \mathbb{E}[\|\xi\|_{\mathcal{H}} \mid \mathcal{F}]$ [134, Chapter 1]. Moreover,

(C3) For $\xi \in \mathbb{L}^2(\Omega, \mathcal{A}, \mathcal{H})$, $\mathbb{E}[\xi \mid \mathcal{F}]$ is the orthogonal projection into $\mathbb{L}^2(\Omega, \mathcal{F}, \mathcal{H})$,

[1] Here and later $F(\xi)$ is meant to mean $F(\xi(\omega))$ for all $\omega \in \Omega$.

again we refer to [173]. We note that, as with conditional expectation on $\mathbb{R}$, $\mathbb{E}[\xi \mid \mathcal{F}]$ is only defined uniquely a.s. As such all (in)equalitie statements hold only a.s. However, we will often not explicitly write this going forward.

We then define $\mathbb{E}[\xi \mid \mathbf{X}] = \mathbb{E}[\xi \mid \sigma(\mathbf{X})]$. The following Proposition shows that this notion is well-defined and some further properties of Hilbert space-valued conditional expectation, in addition to (C1) – (C3):

Proposition 36. *Let $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_1)$, $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_2)$ be two separable Hilbert spaces, $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2 \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_1)$ and $\xi_1, \xi_2, \xi \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_2)$.*[2]

(C4) There exists a measurable function $h : (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1)) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$, such that $\mathbb{E}[\xi \mid \sigma(\mathbf{X})] = h(\mathbf{X}) = \mathbb{E}[\xi \mid \mathbf{X}]$,

(C5) If $\xi_1 \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_1)$, $\xi_2 \in \mathcal{L}^2(\Omega, \sigma(\mathbf{X}), \mathcal{H}_1)$, then $\mathbb{E}[\langle \xi_1, \xi_2\rangle_{\mathcal{H}_1} \mid \mathbf{X}] = \langle \mathbb{E}[\xi_1 \mid \mathbf{X}], \xi_2\rangle_{\mathcal{H}_1}$,

(C6) If $\mathbf{X}_2$ and $(\xi, \mathbf{X}_1)$ are independent, then $\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[\xi \mid \mathbf{X}_1]$,

(C7) $\mathbb{E}[\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2] \mid \mathbf{X}_1] = \mathbb{E}[\mathbb{E}[\xi \mid \mathbf{X}_1] \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[\xi \mid \mathbf{X}_1]$.

Proof We will use the following fact in the proof: Under the assumption of separability, all relevant notions of measurability are the same, see e.g. [75, Chapter 2, 7]. In particular, $\xi \in \mathbb{L}^q(\Omega, \mathcal{F}, \mathcal{H}_2)$, for any $q \geqslant 1$, means that $\xi : (\Omega, \mathcal{F}) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ is measurable, which in turns means there exists a sequence of simple functions

$$f_n = \sum_{k=1}^{m_n} g_k \mathbb{1}_{A_k}, \tag{28}$$

with $g_k \in \mathcal{H}_2$ and $A_k \in \mathcal{F}$ for all k and such that $f_n \to \xi$ a.s. on $\mathcal{H}_2$ and even $\|f_n - \xi\|_{\mathbb{L}^q(\mathcal{H})} \to 0$, see e.g. [134, Proposition 1.2].

For (C4), we note that by (C1), $\mathbb{E}[\xi \mid \sigma(\mathbf{X})] \in \mathbb{L}^1(\Omega, \sigma(\mathbf{X}), \mathcal{H}_2)$ and thus there exists a sequence of functions $f_n : (\Omega, \sigma(\mathbf{X})) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ of the form (28), such that $f_n \to \mathbb{E}[\xi \mid \sigma(\mathbf{X})]$ a.s. on $\mathcal{H}_2$. Since $A_k \in \sigma(\mathbf{X})$, $A_k = \{\omega : \mathbf{X}(\omega) \in B_k\}$ for some $B_k \in \mathcal{B}(\mathcal{H}_1)$, we may transform f_n from a function on Ω to a function on $\mathcal{H}_1$ into $\mathcal{H}_2$:

$$f_n(\omega) = \sum_{k=1}^{m_n} g_k \mathbb{1}_{A_k}(\omega) = \sum_{k=1}^{m_n} g_k \mathbb{1}_{B_k}(\mathbf{X}(\omega)) := h_n(\mathbf{X}(\omega)).$$

This defines a sequence of measurable functions $h_n : (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1)) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ with $h(\mathbf{X}) = \lim_n h_n(\mathbf{X}) = \mathbb{E}[\xi \mid \sigma(\mathbf{X})]$ a.s., proving the result.

[2] We again note that all equalities technically only hold a.s.

We first show (C5) for simple functions and then extend this to $\mathcal{L}^2(\Omega, \sigma(\mathbf{X}), \mathcal{H}_1)$, using the fact at the beginning of the proof. Let thus f_n be of the form (28). Then for all $F \in \sigma(\mathbf{X})$,

$$\begin{aligned}
\mathbb{E}[\langle \mathbb{E}[f_n \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] &= \sum_{k=1}^{m_n} \mathbb{E}[\langle \mathbb{E}[\mathbb{1}_{A_k} \mid \mathbf{X}] g_k, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] \\
&= \sum_{k=1}^{m_n} \mathbb{E}[\mathbb{E}[\mathbb{1}_{A_k} \mid \mathbf{X}] \langle g_k, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] \\
&= \sum_{k=1}^{m_n} \mathbb{E}[\mathbb{E}[\mathbb{1}_{A_k} \langle g_k, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F \mid \mathbf{X}]] \\
&= \mathbb{E}[\langle f_n, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F],
\end{aligned}$$

from the properties of real-valued conditional expectation. As additionally $\langle \mathbb{E}[f_n \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}}$ is clearly $\sigma(\mathbf{X})$ measurable, (C1) and (C2) are met for this candidate. Since conditional expectation is (a.s.) uniquely defined by (C1) and (C2), (C5) holds true for the special case of simple functions. For general $\xi_1 \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_1)$, let f_n have $\|f_n - \xi_1\|_{\mathbb{L}^2(\mathcal{H})} \to 0$. The goal is to show that

$$|\mathbb{E}[\langle \mathbb{E}[f_n \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] - \mathbb{E}[\langle \mathbb{E}[\xi_1 \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| \to 0, \tag{29}$$

$$|\mathbb{E}[\langle f_n, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] - \mathbb{E}[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| \to 0. \tag{30}$$

We can bound both terms by the same quantity, using CauchyâSchwarz:

$$\begin{aligned}
|\mathbb{E}[\langle \mathbb{E}[f_n \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] - \mathbb{E}[\langle \mathbb{E}[\xi_1 \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| &= |\mathbb{E}[\langle \mathbb{E}[f_n - \xi_1 \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| \\
&\leqslant \mathbb{E}[\mathbb{E}[\|f_n - \xi_1\|_{\mathcal{H}} \mathbb{1}_F \|\xi_2\|_{\mathcal{H}} \mid \mathbf{X}]] \\
&= \mathbb{E}[\|f_n - \xi_1\|_{\mathcal{H}} \mathbb{1}_F \|\xi_2\|_{\mathcal{H}}],
\end{aligned}$$

as the random variable $\mathbb{1}_F \|\xi_2\|_{\mathcal{H}}$ is $\sigma(\mathbf{X})$ measurable by assumption and

$$\begin{aligned}
|\mathbb{E}[\langle f_n, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] - \mathbb{E}[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| &= |\mathbb{E}[\langle f_n - \xi_1, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| \\
&\leqslant \mathbb{E}[\|f_n - \xi_1\|_{\mathcal{H}} \|\xi_2\|_{\mathcal{H}} \mathbb{1}_F].
\end{aligned}$$

The result thus follows from the Hölder inequality,

$$\mathbb{E}[\|f_n - \xi_1\|_{\mathcal{H}} \|\xi_2\|_{\mathcal{H}} \mathbb{1}_F] \leqslant \|f_n - \xi_1\|_{\mathbb{L}^2(\mathcal{H})} \cdot \|\xi_2\|_{\mathbb{L}^2(\mathcal{H})} \to 0.$$

Finally (C6) and (C7) is easily proven with the technique of "scalarization" [134, Chapter 1]. We do the full argument for (C6), (C7) can be shown analogously. That is, we show that for any $F : \mathcal{H}_2 \to \mathbb{R}$ linear and continuous,

$$F(\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2](\omega)) = F(\mathbb{E}[\xi \mid \mathbf{X}_1](\omega)). \tag{31}$$

for almost all ω and some representation of $\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2]$. This then immediately implies the result. Now, using the property of real-valued conditional expectations

$$F(\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2]) = \mathbb{E}[F(\xi) \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[F(\xi) \mid \mathbf{X}_1],$$

since $F(\xi)$ is real-valued and independent of $\mathbf{X}_2$. Since $\mathbb{E}[F(\xi) \mid \mathbf{X}_1] = F(\mathbb{E}[\xi \mid \mathbf{X}_1])$, we obtain (31). ■

(C4) in particular allows to see $\mathbb{E}[\xi \mid \sigma(\mathbf{X})]$ as a function in $\mathbf{X}$ and thus justifies the notation $\mathbb{E}[\xi \mid \mathbf{X}]$ and all the subsequent derivations. We may also define *conditional independence* through conditional expectation: With the notation of Proposition 36, ξ and $\mathbf{X}_1$ are conditionally independent given $\mathbf{X}_2$, if $\mathbb{E}[f(\xi) \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[f(\xi) \mid \mathbf{X}_1]$ for all $f : (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ bounded and measurable, see e.g., [34, Proposition 2.3]. This leads to two further important properties:

Proposition 37. *Let $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_1)$, $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_2)$ be two separable Hilbert spaces, $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2 \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_1)$ and $\xi_1, \xi_2, \xi \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_2)$.*

(C8) If ξ and $\mathbf{X}_2$ are conditionally independent given $\mathbf{X}_1$, then $\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[\xi \mid \mathbf{X}_1]$,

(C9) If ξ_1, ξ_2 are conditionally independent given $\mathbf{X}$, $\mathbb{E}[\langle \xi_1, \xi_2 \rangle \mid \mathbf{X}] = \langle \mathbb{E}[\xi_1 \mid \mathbf{X}], \mathbb{E}[\xi_2 \mid \mathbf{X}] \rangle$.

Proof We prove (C8) again using the scalarization trick: For any $F : \mathcal{H}_2 \to \mathbb{R}$ continuous and linear, it holds that

$$F_n(f) := F(f)\mathbb{1}\{|F(f)| \leqslant n\} \quad \forall f \in \mathcal{H}_2,$$

is a bounded and measurable function. Thus by assumption,

$$\mathbb{E}[F_n(\xi) \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[F_n(\xi) \mid \mathbf{X}_1].$$

We now show that $\mathbb{E}[F_n(\xi) \mid \mathbf{X}_1, \mathbf{X}_2] \to \mathbb{E}[F(\xi) \mid \mathbf{X}_1, \mathbf{X}_2]$ and $\mathbb{E}[F_n(\xi) \mid \mathbf{X}_1] \to \mathbb{E}[F(\xi) \mid \mathbf{X}_1]$ a.s. Let $\mathcal{F}$ stand for either $\sigma(\mathbf{X}_1, \mathbf{X}_2)$ or $\sigma(\mathbf{X}_1)$. Then, as $F(f) = F_n(f) + F(f)\mathbb{1}\{|F(f)| > n\}$,

$$\begin{aligned} |\mathbb{E}[F_n(\xi) \mid \mathcal{F}] - \mathbb{E}[F(\xi) \mid \mathcal{F}]| &\leqslant \mathbb{E}[|F_n(\xi) - F(\xi)| \mid \mathcal{F}] \\ &= \mathbb{E}[|F(f)|\mathbb{1}\{|F(f)| > n\} \mid \mathcal{F}]. \end{aligned}$$

Now, since for all n, $|F(f)|\mathbb{1}\{|F(f)| > n\} \geqslant 0$ and $\mathbb{E}[|F(f)|\mathbb{1}\{|F(f)| > n\} \mid \mathcal{F}] \leqslant \mathbb{E}[|F(f)| \mid \mathcal{F}] < \infty$ a.s. an application of Fatou's Lemma for (real-valued) conditional expectation (see e.g., [46, Problem 10.7]) to $|F(f)| - |F(f)|\mathbb{1}\{|F(f)| > n\}$ implies

$$\limsup_n |\mathbb{E}[F_n(\xi) \mid \mathcal{F}] - \mathbb{E}[F(\xi) \mid \mathcal{F}]| \leqslant \mathbb{E}[\limsup_n |F(f)|\mathbb{1}\{|F(f)| > n\} \mid \mathcal{F}] = 0.$$

Thus, we have shown that for all $F : \mathcal{H}_2 \to \mathbb{R}$ continuous and linear, $F(\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2]) = F(\mathbb{E}[\xi \mid \mathbf{X}_1])$, proving the claim.

Finally, combining (C5), (C7) and (C8), we obtain for ξ_1, ξ_2 conditionally independent given $\mathbf{X}$

$$\begin{aligned}\mathbb{E}[\langle \xi_1, \xi_2 \rangle \mid \mathbf{X}] &= \mathbb{E}[\mathbb{E}[\langle \xi_1, \xi_2 \rangle \mid \mathbf{X}, \xi_2] \mid \mathbf{X}] \\ &= \mathbb{E}[\langle \mathbb{E}[\xi_1 \mid \mathbf{X}, \xi_2], \xi_2 \rangle \mid \mathbf{X}] \\ &= \mathbb{E}[\langle \mathbb{E}[\xi_1 \mid \mathbf{X}], \xi_2 \rangle \mid \mathbf{X}] \\ &= \langle \mathbb{E}[\xi_1 \mid \mathbf{X}], \mathbb{E}[\xi_2 \mid \mathbf{X}] \rangle.\end{aligned}$$

■

Let for $\mathbf{x} \in \mathcal{X}$, $P_{\mathbf{x}}$ be the conditional distribution of Y given $\mathbf{x}$ on $\mathbb{R}^d$ and similarly with $P_{\mathbf{X}}$ (i.e. the regular conditional probability measure). Note that $P_{\mathbf{x}} \in \mathcal{H}$, while $P_{\mathbf{X}}$ is a random element mapping into $\mathcal{H}$.

As in [181], we define for f, g two functions, with $\liminf_{s\to\infty} g(s) > 0$, $f(s) = O(g(s))$ if

$$\limsup_{s\to\infty} \frac{|f(s)|}{g(s)} \leqslant C,$$

for some $C > 0$. If $C = 1$, then we write $f(s) \precsim g(s)$. For a sequence of random variables $X_n : \Omega \to \mathbb{R}$, and $a_n \in (0, +\infty)$, $n \in \mathbb{N}$, we write as usual $X_n = O_p(a_n)$, if

$$\lim_{M\to\infty} \sup_n \mathbb{P}(a_n^{-1}|X_n| > M) = 0,$$

i.e. if X_n is bounded in probability. We write $X_n = o_p(a_n)$, if $a_n^{-1}X_n$ converges in probability to zero. Similarly, for (S, d) a separable metric space, $\mathbf{X}_n : (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$, $n \in \mathbb{N}$ and $\mathbf{X} : (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$ measurable, we write $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$, if $d(\mathbf{X}_n, \mathbf{X}) = o_p(1)$.

Finally let $\mathbf{X} \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_1)$, $\xi \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_2)$ and assume that $A \subset \Omega$ depends on $\mathbf{X}$, $A = A(\mathbf{X})$. Thus for $\mathbf{X}$ fixed to a certain value, A is a fixed set. If $\mathbb{P}(A \mid \mathbf{X}) > 0$ almost everywhere, we define

$$\mathbb{E}[\xi \mid A] = \mathbb{E}[\xi \mid \mathbf{X}, A] := \frac{\mathbb{E}[\xi \mathbb{1}_A \mid \mathbf{X}]}{\mathbb{P}(A \mid \mathbf{X})} \in \mathcal{L}^2(\Omega, \sigma(\mathbf{X}), \mathcal{H}_2).$$

It then holds by construction that

$$\mathbb{E}[\xi \mathbb{1}_A \mid \mathbf{X}] = \mathbb{E}[\xi \mid \mathbf{X}, A] \cdot \mathbb{P}(A \mid \mathbf{X}). \tag{32}$$

Let again $\mu(\mathbf{x}) := \mu(P_{\mathbf{x}})$ be the embedding of the true conditional distribution into $\mathcal{H}$. We first state 3 preliminary results:

Lemma .13. *It holds that* $\mathbb{E}[\mu(\delta_{\mathbf{Y}}) \mid \mathbf{X} = \mathbf{x}] = \mu(P_{\mathbf{x}})$.

Proof We first note that $P_{\mathbf{x}}$ exists and is a probability measure on $\mathbb{R}^d$. Since k is bounded $\mu(P_{\mathbf{x}}) \in \mathcal{H}$ exists and is uniquely defined by the relation

$$\langle f, \mu(P_{\mathbf{x}})\rangle_{\mathcal{H}} = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}] \quad \forall f \in \mathcal{H}.$$

On the other hand, by the above $\mathbb{E}[\mu(\delta_{\mathbf{Y}}) \mid \mathbf{X}] \in \mathbb{L}^2(\Omega, \mathcal{A}, \mathcal{H})$ exists and for all $f \in \mathcal{H}$,

$$\langle \mathbb{E}[\mu(\delta_{\mathbf{Y}}) \mid \mathbf{X}], f\rangle_{\mathcal{H}} = \mathbb{E}[\langle \mu(\delta_{\mathbf{Y}}), f\rangle_{\mathcal{H}} \mid \mathbf{X}] = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}],$$

since $F : H \to \mathbb{R}$, $F(g) = \langle g, f\rangle_{\mathcal{H}}$ defines a continuous linear function. In particular, for all $f \in \mathcal{H}$,

$$\langle \mathbb{E}[\mu(\delta_{\mathbf{Y}}) \mid \mathbf{X} = \mathbf{x}], f\rangle_{\mathcal{H}} = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}],$$

or $\mathbb{E}[\mu(\delta_{\mathbf{Y}}) \mid \mathbf{X} = \mathbf{x}] = \mu(P_{\mathbf{x}})$. ■

Since $\mu(\delta_{\mathbf{Y}}) = k(\mathbf{Y}, \cdot)$, Lemma .13 in fact corresponds to Lemma 3.2 in [132].

For a more compact notation in the following Lemma, let $N = \{1, \ldots, n\}$ and let for $A \subset N$ and $k \leqslant |A|$, $C_k(A)$ be the set of all subsets of size k drawn from A without replacement, with $C_0 := \varnothing$.

Lemma .14 (H-Decomposition of a Hilbert-space valued Kernel)**.** *Let* $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_1)$, $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_2)$ *be two separable Hilbert spaces,* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be i.i.d. copies of a random element* $\mathbf{X} : (\Omega, \mathcal{A}) \to (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1))$. *Write* $\mathcal{X}_n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ *and let* $T : (\mathcal{H}_1^n, \mathcal{B}(\mathcal{H}_1^n)) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ *measurable with* $\mathbb{E}[\|T(\mathcal{X}_n)\|_{\mathcal{H}_2}^2] < \infty$. *If* T *is symmetric, there exists functions* T_j, $j = 1, \ldots, n$, *such that*

$$T(\mathcal{X}_n) = \mathbb{E}[T(\mathbf{X})] + \sum_{i=1}^{n} T_1(\mathbf{X}_i) + \sum_{i_1 < i_2} T_2(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) + \ldots T_n(\mathcal{X}_n), \tag{33}$$

and it holds that

$$Var(T(\mathcal{X}_n)) = \sum_{i=1}^{n} \binom{n}{i} Var(T_i(\mathbf{X}_1, \ldots, \mathbf{X}_i)), \tag{34}$$

and

$$T_1(\mathbf{X}_i) = \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_i] - \mathbb{E}[T(\mathcal{X}_n)].$$

Proof Composition (33) was proven in a range of different ways for real-valued T, see e.g. [71], [47], [48], or [178]. We consider and slightly extend the elegant proof of [47] to also prove (34).

See also [178]. Let

$$\begin{aligned} T_1(\mathbf{X}_i) &= \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_i] - \mathbb{E}[T(\mathcal{X}_n)] \\ T_2(\mathbf{X}_i, \mathbf{X}_j) &= \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_i, \mathbf{X}_j] - \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_i] - \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_j] + \mathbb{E}[T(\mathcal{X}_n)] \\ &\vdots \\ T_\ell(\pi_{NA_\ell}(\mathcal{X}_n)) &= \sum_{k=0}^{\ell-1} (-1)^{\ell-k} \sum_{B \in C_k(A_\ell)} \mathbb{E}[T(\mathcal{X}_n) \mid \pi_{NB}(\mathcal{X}_n)]. \end{aligned}$$

We note that T_ℓ does not depend on the exact indices in A_ℓ thanks to the assumed symmetry of T. Since $\mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_1, \ldots, \mathbf{X}_n] = T(\mathcal{X}_n)$ is part of $T_n(\mathcal{X}_n)$, this leads to a telescoping sum, already proving (33).

Adapting the approach of [47], consider now $\mathcal{L}^2(\Omega, \sigma(\mathcal{X}_n), \mathcal{H})$ and let Q_i be the projection operator into $\mathcal{L}^2(\Omega, \sigma(\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n), \mathcal{H})$. That is

$$(Q_i T)(\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n) = \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n],$$

by (C3). Now as in [178],

(I) the Q_i commute,

(II) For $A_\ell = \{i_1, \ldots, i_\ell\} \subset N$,

$$(Q_{i_1} \cdots Q_{i_\ell} T)(\pi_{NA_\ell^c}(\mathcal{X}_n)) = \mathbb{E}[T(\mathcal{X}_n) \mid \pi_{NA_\ell^c}(\mathcal{X}_n)].$$

In particular,

$$(Q_1 \cdots Q_\ell T)(\mathbf{X}_{\ell+1}, \ldots, \mathbf{X}_n) = \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_{\ell+1}, \ldots, \mathbf{X}_n].$$

Moreover, it holds that

$$T_\ell(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_\ell}) = ([I - Q_{i_1}][I - Q_{i_2}] \ldots [I - Q_{i_\ell}] Q_{i_\ell+1} \cdots Q_{i_n} T)(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_\ell}).$$

Expanding the identity

$$T(\mathcal{X}_n) = (I^n T)(\mathcal{X}_n) = ([(I - Q_1) + Q_1][(I - Q_2) + Q_2] \cdots [(I - Q_n) + Q_n] T)(\mathcal{X}_n),$$

as in [178] and using $Q_i(1 - Q_i) = 0$, proves (33). Furthermore the above implies that for any subset $A_l \subset N$ that intersects with $A_\ell = \{i_1, \ldots, i_\ell\}$, we must have $\mathbb{E}[T_\ell(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_\ell}) \mid \pi_{NA_l}(\mathcal{X}_n)] = 0$. Indeed assume $A_l \cap A_\ell = \{i_2, \ldots, i_\ell\}$, then since the elements of $\mathcal{X}_n$ are independent, by (C6),

$$\mathbb{E}[T_\ell(\pi_{NA_\ell}(\mathcal{X}_n)) | \pi_{NA_l}(\mathcal{X}_n)] = \mathbb{E}[T_\ell(\pi_{NA_\ell}(\mathcal{X}_n)) \mid \pi_{N(A_l \cap A_\ell)}(\mathcal{X}_n)],$$

i.e. all elements outside the intersection are irrelevant. Moreover,

$$\begin{aligned}&\mathbb{E}[T_\ell(\pi_{NA_\ell}(\mathcal{X}_n)) \mid \pi_{N(A_l\cap A_\ell)}(\mathcal{X}_n)]\\&= (Q_{i_1}[I-Q_{i_1}][I-Q_{i_2}]\dots[I-Q_{i_\ell}]Q_{i_\ell+1}\cdots Q_{i_n}T)(\pi_{N(A_l\cap A_\ell)}(\mathcal{X}_n))\end{aligned}$$

and clearly this projection can only be 0. The same argument can be made for any other intersection set $A_l \cap A_\ell$, even if it is the empty set.

Thus combining the above with (C5), for $A_\ell \neq A_l$, it holds that

$$\begin{aligned}\mathbb{E}[\langle T_\ell(\pi_{NA_\ell}(\mathcal{X}_n)), T_l(\pi_{NA_l}(\mathcal{X}_n))\rangle_{\mathcal{H}}] &= \mathbb{E}[\mathbb{E}[\langle T_\ell(\pi_{NA_\ell}(\mathcal{X}_n)), T_l(\pi_{NA_l}(\mathcal{X}_n))\rangle_{\mathcal{H}} \mid \pi_{NA_l}(\mathcal{X}_n)]]\\&= \mathbb{E}[\langle \mathbb{E}[T_\ell(\pi_{NA_\ell}(\mathcal{X}_n)) \mid \pi_{NA_l}(\mathcal{X}_n)], T_l(\pi_{NA_l}(\mathcal{X}_n))\rangle_{\mathcal{H}}]\\&= 0.\end{aligned}$$

In other words, the covariance between any two elements in the decomposition of $T(\mathcal{X}_n) - \mathbb{E}[T(\mathbf{X})]$ in (33) is uncorrelated. Finally then

$$\begin{aligned}\mathrm{Var}(T(\mathbf{X}_1,\dots,\mathbf{X}_n)) &= \mathbb{E}[\langle T(\mathcal{X}_n) - \mathbb{E}[T(\mathcal{X}_n)], T(\mathcal{X}_n) - \mathbb{E}[T(\mathcal{X}_n)]\rangle]\\&= \sum_{i=1}^{n}\binom{n}{i}\mathrm{Var}(T_i(\mathbf{X}_1,\dots,\mathbf{X}_i)).\end{aligned}$$

■

In order to proceed, we first prove Theorem 31, which is somewhat separate from the remainder of this section:

Theorem 31. *For any split of a parent node P into child nodes C_L and C_R, let $\hat{\mathbb{P}}_{split}(\mathbf{x}) = \sum_{j\in\{L,R\}} \mathbb{1}(\mathbf{x} \in C_j)\frac{1}{n_j}\sum_{\mathbf{x}_i\in C_j}\delta_{\mathbf{y}_i}$ denote the resulting estimate of the distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ when $\mathbf{x} \in P$. Then the MMD splitting criterion can be viewed as the version of the CART criterion* (7) *on $\mathcal{H}$:*

$$\begin{aligned}&\arg\max_{split} \frac{n_L n_R}{n_P^2}\mathcal{D}_{MMD(k)}\left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\}\right)\\&= \arg\min_{split} \frac{1}{n_P}\sum_{\mathbf{x}_i\in P}\left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{\mathbb{P}}_{split}(\mathbf{x}_i))\right\|_{\mathcal{H}}^2.\end{aligned}$$

Moreover, for any node P and any fixed distributional estimator $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$, we have:

$$\begin{aligned}&\frac{1}{n_P}\sum_{\mathbf{x}_i\in P}\left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x}_i))\right\|_{\mathcal{H}}^2\\&= V_P + \mathbb{E}\left[\|\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X})) - \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}))\|_{\mathcal{H}}^2 \mid \mathbf{X} \in P\right] + O_p(n^{-1/2}),\end{aligned}$$

where $V_P = \mathbb{E}\left[\|\mu(\delta_{\mathbf{Y}}) - \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}))\|_{\mathcal{H}}^2 \mid \mathbf{X} \in P\right]$ is a deterministic term not depending on the estimates $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$.

Proof The first part of the Theorem is shown analogously to the proof in Section 2.3 of the appendix, but where we replace the standard dot product in $\mathbb{R}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ associated with $(\mathcal{H}, k)$ and use the induced RKHS norm $\|\cdot\|_{\mathcal{H}}$. Also, since k is bounded, the embedding $\mu(\mathcal{D})$ into RKHS $\mathcal{H}$ exists for any distribution $\mathcal{D}$, so everything is well-defined.

For the second statement of the Theorem 1, note that $n_P \sim \text{Binomial}(\pi, n)$, where $\pi := \mathbb{P}(\mathbf{X} \in P) > 0$. Let $\hat{P}_{\mathbf{x}} = \hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ be a fixed conditional distribution estimator and recall that $P_{\mathbf{x}} = \mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$. We now write

$$\sum_{\mathbf{x}_i \in P} \left\| \mu(\delta_{\mathbf{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i}) \right\|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \left\| \mu(\delta_{\mathbf{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i}) \right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{x}_i \in P\}.$$

Then it holds that

$$\mathbb{E}\left[\left\| \mu(\delta_{\mathbf{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i}) \right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X}_i \in P\} \right] = \mathbb{E}\left[\mathbb{E}\left[\left\| \mu(\delta_{\mathbf{Y}}) - \mu(\hat{P}_{\mathbf{X}}) \right\|_{\mathcal{H}}^2 \mid \mathbf{X} \right] \mathbb{1}\{\mathbf{X} \in P\} \right],$$

and

$$\begin{aligned} \mathbb{E}\left[\left\| \mu(\delta_{\mathbf{Y}}) - \mu(\hat{P}_{\mathbf{X}}) \right\|_{\mathcal{H}}^2 \mid \mathbf{X} \right] = \mathbb{E}\Big[& \mathbb{E}\left\| \mu(\delta_{\mathbf{Y}}) - \mu(P_{\mathbf{X}}) \right\|_{\mathcal{H}}^2 + \mathbb{E}\left\| \mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}}) \right\|_{\mathcal{H}}^2 \mid \mathbf{X} \Big] \\ & + 2\mathbb{E}\left[\langle \mu(\delta_{\mathbf{Y}}) - \mu(P_{\mathbf{X}}), \mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}}) \rangle_{\mathcal{H}} \mid \mathbf{X} \right]. \end{aligned}$$

It follows with Lemma .13 and (C5) that,

$$\begin{aligned} & \mathbb{E}\left[\langle \mu(\delta_{\mathbf{Y}}) - \mu(P_{\mathbf{X}}), \mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}}) \rangle_{\mathcal{H}} \mid \mathbf{X} \right] = \\ & \langle \mathbb{E}\left[\mu(\delta_{\mathbf{Y}}) - \mu(P_{\mathbf{X}}) \mid \mathbf{X} \right], \mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}}) \rangle_{\mathcal{H}} = 0. \end{aligned}$$

Combining the three equations, this means

$$\begin{aligned} & \mathbb{E}\left[\left\| \mu(\delta_{\mathbf{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i}) \right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X}_i \in P\} \right] \\ & = \mathbb{E}\left[\left\| \mu(\delta_{\mathbf{Y}}) - \mu(P_{\mathbf{X}}) \right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X} \in P\} + \left\| \mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}}) \right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X} \in P\} \right] \\ & = \pi \left(V_P + \mathbb{E}\left[\left\| \mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}}) \right\|_{\mathcal{H}}^2 \mid \mathbf{X} \in P \right] \right), \end{aligned} \tag{35}$$

using that $\mathbb{E}[g(\mathbf{X}) \mid \mathbf{X} \in P] = \mathbb{E}[g(\mathbf{X})\mathbb{1}\{\mathbf{X} \in P\}]/\mathbb{P}(\mathbf{X} \in P)$. We now show that the difference between $\frac{1}{n_P} \sum_{\mathbf{x}_i \in P} \left\| \mu(\delta_{\mathbf{y}_i}) - \mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)) \right\|_{\mathcal{H}}^2$ and the expectation on the left of Equation (35) is $O_p(n^{-1/2})$, using standard CLT arguments. Define $K = \sup_{z,z'} \mid k(z,z') \mid < \infty$, as we have assumed that k is bounded. For any two distributions $\mathcal{D}_1, \mathcal{D}_2$ we now obtain

$$\left\| \mu(\mathcal{D}_1) - \mu(\mathcal{D}_2) \right\|_{\mathcal{H}}^2 = \mathbb{E}[k(\mathbf{Z}_1, \mathbf{Z}_1')] - 2\mathbb{E}[k(\mathbf{Z}_1, \mathbf{Z}_2)] + \mathbb{E}[k(\mathbf{Z}_2, \mathbf{Z}_2')] \leqslant 4K,$$

where $\mathbf{Z}_1, \mathbf{Z}_1' \sim \mathcal{D}_1$ and $\mathbf{Z}_2, \mathbf{Z}_2' \sim \mathcal{D}_2$ are independent random variables. Thus,

$$\mathbb{E}\left[\left\| \mu(\delta_{\mathbf{Y}}) - \mu(\hat{P}_{\mathbf{X}}) \right\|_{\mathcal{H}}^2 \right] \leqslant 4K, \quad \mathbb{E}\left[\left\| \mu(\delta_{\mathbf{Y}}) - \mu(\hat{P}_{\mathbf{X}}) \right\|_{\mathcal{H}}^4 \right] \leqslant 16K^2,$$

implying that both first and second moments of the random variable $\left\|\mu(\delta_{\mathbf{Y}}) - \mu(\hat{P}_{\mathbf{X}})\right\|^2_{\mathcal{H}} \mathbb{1}\{\mathbf{X} \in P\}$ are finite. Moreover, since $\left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i})\right\|^2_{\mathcal{H}} \mathbb{1}\{\mathbf{x}_i \in P\}$ for $i = 1, \ldots, n$ are its i.i.d. realizations, it follows directly from the CLT that:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i})\right\|^2_{\mathcal{H}} \mathbb{1}\{\mathbf{x}_i \in P\} - \mathbb{E}\left[\left\|\mu(\delta_{\mathbf{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i})\right\|^2_{\mathcal{H}} \mathbb{1}\{\mathbf{X}_i \in P\}\right]\right) = O_p(1).$$

By multiplying the above equation with $n/n_P = (1/\pi + o_p(1)) = O_p(1)$, it also holds that

$$\sqrt{n}\left(\frac{1}{n_P}\sum_{\mathbf{x}_i \in P}\left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i})\right\|^2_{\mathcal{H}} - \frac{n}{n_P}\mathbb{E}\left[\left\|\mu(\delta_{\mathbf{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i})\right\|^2_{\mathcal{H}} \mathbb{1}\{\mathbf{X}_i \in P\}\right]\right) = O_p(1). \quad (36)$$

Thus,

$$\begin{aligned}
&\sqrt{n}\left(\frac{1}{n_P}\sum_{\mathbf{x}_i \in P}\left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i})\right\|^2_{\mathcal{H}} - \frac{1}{\pi}\mathbb{E}\left[\left\|\mu(\delta_{\mathbf{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i})\right\|^2_{\mathcal{H}} \mathbb{1}\{\mathbf{X}_i \in P\}\right]\right) \\
&= \sqrt{n}\left(\frac{1}{n_P}\sum_{\mathbf{x}_i \in P}\left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i})\right\|^2_{\mathcal{H}} - \frac{n}{n_P}\mathbb{E}\left[\left\|\mu(\delta_{\mathbf{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i})\right\|^2_{\mathcal{H}} \mathbb{1}\{\mathbf{X}_i \in P\}\right]\right) \\
&\quad - \sqrt{n}\left(1 - \frac{n\pi}{n_P}\right)\frac{1}{\pi}\mathbb{E}\left[\left\|\mu(\delta_{\mathbf{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i})\right\|^2_{\mathcal{H}} \mathbb{1}\{\mathbf{X}_i \in P\}\right]. \quad (37)
\end{aligned}$$

Now both terms in (37) are $O_p(1)$: For the first term this follows from (36). For the second term it holds, since $\mathbb{E}\left[\left\|\mu(\delta_{\mathbf{Y}}) - \mu(\hat{P}_{\mathbf{X}})\right\|^2_{\mathcal{H}}\right] \leqslant 4K$ and

$$\sqrt{n}\left(1 - \frac{n\pi}{n_P}\right) = O_p(1),$$

which in turn is true by another application of the CLT on random variables $\mathbb{1}\{\mathbf{X}_i \in P\}$ and the fact that $n/n_P = O_p(1)$:

$$\sqrt{n}\left(1 - \frac{n\pi}{n_P}\right) = \sqrt{n}\frac{n_P - n\pi}{n_P} = \frac{n}{n_P}\frac{n_P - n\pi}{\sqrt{n}} = O_p(1).$$

Combining the fact that the expression in (37) is $O_p(1)$ with (35) gives the result. ■

Let $\hat{\mu}_n(\mathbf{x})$ be defined as in (17):

$$\hat{\mu}_n(\mathbf{x}) = \binom{n}{s_n}^{-1} \sum_{i_1 < i_2 < \ldots < i_{s_n}} \mathbb{E}_\varepsilon\left[T(\mathbf{x}, \varepsilon; \mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}})\right], \quad (38)$$

where the sum is taken over all $\binom{n}{s_n}$ possible subsamples $\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}$ of $\mathbf{Z}_1, \ldots \mathbf{Z}_n$ and $s_n \to \infty$ with n. Moreover, a single tree is given as

$$T(\mathbf{x}; \epsilon_k, \mathbf{Z}_{k_1}, \ldots \mathbf{Z}_{k_{s_n}}) = \sum_{j=1}^{s_n} \frac{\mathbb{1}(\mathbf{X}_{k_j} \in \mathcal{L}_k(\mathbf{x}))}{|\mathcal{L}_k(\mathbf{x})|}\mu(\delta_{\mathbf{Y}_{k_j}}). \quad (39)$$

Given these preliminary results, the proofs we use are for the most part analogous to the ones in [181]. Thus the proofs are given mostly for completeness and sometimes omitted altogether. We introduce the following additional notation, similar to Section 3: Let $\mathcal{Z}_s = (\mathbf{Z}_1, \ldots, \mathbf{Z}_s)$ collect s i.i.d. copies of $\mathbf{Z}$ and define for $j = 1, \ldots, s_n$,

$$\begin{aligned} \mathrm{Var}(T) &= \mathrm{Var}(T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})) \\ \mathrm{Var}(T_j) &= \mathrm{Var}(\mathbb{E}[T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n}) | \mathbf{Z}_1, \ldots, \mathbf{Z}_j]). \end{aligned}$$

The index s will take the role of s_n or n, depending on the situation. We note that, due to i.i.d. sampling, it doesn't matter for variance or expectation what kind of subset $\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_s}$ we are considering. In particular, we might just take $\mathcal{Z}_s$ each time.

Before going on to the main proofs, we repeat here the assumed properties of the trees for completeness:

(P1) (*Data sampling*) The bootstrap sampling with replacement, usually used in forest-based methods, is replaced by a subsampling step, where for each tree we choose a random subset of size s_n out of n training data points. We consider s_n going to infinity with n, with the rate specified below.

(P2) (*Honesty*) An observation $\mathbf{Z} = (\mathbf{X}, \mu(\delta_{\mathbf{Y}}))$ is either used to place the splits in a tree or to estimate the response, but never both.

(P3) (α*-regularity*) Each split leaves at least a fraction $\alpha \leqslant 0.2$ of the available training sample on each side. Moreover, the trees are grown until every leaf contains between κ and $2\kappa - 1$ observations, for some fixed tuning parameter $\kappa \in \mathbb{N}$.

(P4) (*Symmetry*) The (randomized) output of a tree does not depend on the ordering of the training samples.

(P5) (*Random-split*) At every split point, the probability that the split occurs along the feature X_j is bounded below by π/p, for some $\pi > 0$ and for all $j = 1, \ldots, p$.

Note that assumption **(P2)** differentiates from assumption **(P2)** in the main text. We refer to it as

(P2') (*Double Sampling*) The data used for constructing each tree is split into two parts; the first is used for determining the splits and the second for populating the leaves and thus for estimating the response.

(P2) will allow us to assume that all s_n observations of the tree are used to estimate the response, as in (39) and the trees are built with some auxiliary data. This is done for simplicity

of exposition, the results can be extended to hold in case of **(P2')** as well. In fact, since the (random) division into the two data sets can be seen as part of ε_k, the adaptation simply involves changing s_n to $s_n/2$.

Now, we may directly apply Lemma (.14) to the U-statistics $\hat{\mu}_n(\mathbf{x})$:

Lemma .15. *Let $\hat{\mu}_n(\mathbf{x})$ be as in* (38) *and assume T satisfies **(P4)** and*

$$Var(T) < \infty.$$

Then

$$\begin{aligned} Var(\hat{\mu}_n(\mathbf{x})) &\leqslant \frac{s_n^2}{n} Var(T_1) + \frac{s_n^2}{n^2} Var(T) \\ &\leqslant \left(\frac{s_n}{n} + \frac{s_n^2}{n^2} \right) Var(T). \end{aligned}$$

Proof Using the composition in (33) on $T(\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_s}) := \mathbb{E}_\varepsilon \left[T(\mathbf{x}, \varepsilon; \mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_s}) \right]$, we have for $A \subset N = \{1, \ldots, n\}$, $|A| = s_n$,

$$T(\pi_{NA}(\mathcal{Z}_n)) = \mathbb{E}[T(\pi_{NA}(\mathcal{Z}_n))] + \sum_{\ell=1}^{s} \sum_{B \in C_\ell(A)} T_\ell(\pi_{NB}(\mathcal{Z}_n)), \tag{40}$$

Moreover, it holds by symmetry and i.i.d. sampling, that for $A_1 = A_2 \subset N$, $|A_1| = |A_2| = \ell$, $T_\ell(\pi_{A_1}(\mathcal{Z}_n)) = T_\ell(\pi_{A_2}(\mathcal{Z}_n))$. Thus we obtain,

$$\begin{aligned} \hat{\mu}_n(\mathbf{x}) = \mathbb{E}[T(\mathcal{Z}_{s_n})] + \binom{n}{s_n}^{-1} \Bigg(&\binom{n-1}{s_n-1} \sum_{i=1}^{n} T_1(\mathbf{Z}_i) + \binom{n-2}{s_n-2} \sum_{i_1<i_2} T_2(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}) \\ &+ \ldots + \sum_{i_1<i_2<\ldots<i_{s_n}} T_s(\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}) \Bigg). \end{aligned}$$

Now

$$\begin{aligned} \binom{n}{s_n}^{-1} \binom{n-j}{s_n-j} &= \frac{s_n!}{n!} \frac{(n-j)!}{(s_n-j)!} \\ &= \frac{s_n \cdot (s_n-1) \cdots (s_n-j+1)}{n \cdot (n-1) \cdots (n-j+1)} \\ &= \frac{(s_n)_j}{(n)_j}, \end{aligned}$$

where $(s_n)_j = s_n(s_n-1) \cdot (s_n-(j-1)) = s_n!/(s_n-j)!$. In particular

$$\binom{n}{s_n}^{-1} \binom{n-1}{s_n-1} = \frac{s_n}{n}.$$

Consequently,

$$\hat{\mu}_n(\mathbf{x}) = \frac{s_n}{n}\sum_{i=1}^{n} T_1(\mathbf{Z}_i) + \frac{(s_n)_2}{(n)_2}\sum_{i_1<i_2} T_2(\mathbf{Z}_{i_1},\mathbf{Z}_{i_2}) + \frac{(s_n)_3}{(n)_3}\sum_{i_1<i_2<i_3} T_3(\mathbf{Z}_{i_1},\mathbf{Z}_{i_2},\mathbf{Z}_{i_3}) + \\ \ldots + \frac{(s_n)_{s_n}}{(n)_{s_n}}\sum_{i_1<i_2<\ldots<i_{s_n}} T_{s_n}(\mathbf{Z}_{i_1},\ldots,\mathbf{Z}_{i_{s_n}}),$$

with covariances between terms equal to 0, as in Lemma .14. Thus

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_n(\mathbf{x})) &= \frac{s_n^2}{n^2} n\mathrm{Var}(T_1) + \sum_{i=2}^{s_n}\left(\frac{(s_n)_i}{(n)_i}\right)^2 \binom{n}{i}\mathrm{Var}(T_i) \\
&= \frac{s_n^2}{n}\mathrm{Var}(T_1) + \sum_{i=2}^{s_n}\left(\frac{(s_n)_i}{(n)_i}\right)\binom{s_n}{i}\mathrm{Var}(T_i) \\
&\leqslant \frac{s_n^2}{n}\mathrm{Var}(T_1) + \frac{(s_n)_2}{(n)_2}\sum_{i=2}^{s_n}\binom{s_n}{i}\mathrm{Var}(T_i) \\
&\leqslant \frac{s_n^2}{n}\mathrm{Var}(T_1) + \frac{s_n^2}{n^2}\mathrm{Var}(T),
\end{aligned}$$

where the last step followed from (34). This proves the first inequality. On the other hand, we have from Lemma .14 that

$$\mathrm{Var}(T) = \sum_{i=1}^{s_n}\binom{s_n}{i}\mathrm{Var}(T_i) \geqslant s_n\mathrm{Var}(T_1),$$

leading to the second inequality. ■

Lemma .16. *[Lemma 2 from [181]] Let T is a tree satisfying **(P3)**, **(P5)** trained on $\mathcal{Z}_s = (\xi_1,\mathbf{X}_1),\ldots(\xi_s,\mathbf{X}_s)$ and let $L(x,\mathcal{Z}_s)$ be the leaf containing $\mathbf{x}$. Suppose that $\mathbf{X}_1,\ldots,\mathbf{X}_s$ are i.i.d. on $[0,1]^p$ independently with a density f bounded away from 0 and infinity. Then,*

$$\mathbb{P}\left(diam(L(\mathbf{x},\mathcal{Z}_s)) \geqslant \sqrt{d}\left(\frac{s}{2k-1}\right)^{-0.51\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}\right) \leqslant d\left(\frac{s}{2k-1}\right)^{-1/2\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}. \quad (41)$$

Lemma .17. *Let T be a tree satisfying **(P1)**, **(P2)** and $L(\mathbf{x},\mathcal{Z}_s)$ be the leaf containing $\mathbf{x}$. Then*

$$\mathbb{E}[T(\mathcal{Z}_s)] = \mathbb{E}[\mathbb{E}[\xi_1 \mid \mathbf{X}_1 \in L(\mathbf{x},\mathcal{Z}_s)]], \quad (42)$$

and

$$Var(T(\mathcal{Z}_s)) \leqslant \sup_{\mathbf{x}\in[0,1]^p}\mathbb{E}[\|\xi_1\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}]. \quad (43)$$

Proof We want to prove

$$\mathbb{E}[T(\mathcal{Z}_s)] = \mathbb{E}[\mathbb{E}[T(\mathcal{Z}_s) \mid L(\mathbf{x}, \mathcal{Z}_s)]] = \mathbb{E}[\mathbb{E}[\xi_1 \mid \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), L(\mathbf{x}, \mathcal{Z}_s)]]. \tag{44}$$

Let for the following $N_{\mathbf{x}} = \sum_{i=1}^{s} \mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x}, \mathcal{Z}_s)\}$. Then due to i.i.d. sampling:

$$\mathbb{E}[\mathbb{E}[T(\mathcal{Z}_s) \mid L(\mathbf{x}, \mathcal{Z}_s)]] = \mathbb{E}[\mathbb{E}[\sum_{i=1}^{s} S_i \xi_i \mid L(\mathbf{x}, \mathcal{Z}_s)]] = s\mathbb{E}[S_1 \xi_1 \mid L(\mathbf{x}, \mathcal{Z}_s)].$$

The last expression can be broken into

$$\begin{aligned}
s\mathbb{E}[\mathbb{E}[S_1 \xi_1 \mid L(\mathbf{x}, \mathcal{Z}_s)]] &= s\mathbb{E}\left[\mathbb{E}[\mathbb{E}[S_1 \xi_1 \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)] \mid L(\mathbf{x}, \mathcal{Z}_s)]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{s}{N_{\mathbf{x}}}\mathbb{E}[\mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)\}\xi_1 \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)] \mid L(\mathbf{x}, \mathcal{Z}_s)\right]\right] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\frac{s}{N_{\mathbf{x}}}\mathbb{E}[\xi_1 \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)] \\
&\quad \mathbb{P}\left(\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s) \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)\right) \mid L(\mathbf{x}, \mathcal{Z}_s)\Big]\Big].
\end{aligned} \tag{45}$$

Now, by honesty, given the knowledge that $\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)$, ξ_1 is independent of $N_{\mathbf{x}}$, thus:

$$\begin{aligned}
&s\mathbb{E}[S_1 \xi_1 \mid L(\mathbf{x}, \mathcal{Z}_s)] \\
&= \mathbb{E}\left[\mathbb{E}[\xi_1 \mid L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)]\mathbb{E}\left[\frac{s}{N_{\mathbf{x}}}\mathbb{P}\left(\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s) \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)\right) \mid L(\mathbf{x}, \mathcal{Z}_s)\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}[\xi_1 \mid L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)]s\mathbb{E}\left[S_1 \mid L(\mathbf{x}, \mathcal{Z}_s)\right]\right].
\end{aligned}$$

Now it holds by i.i.d. sampling that,

$$s\mathbb{E}\left[S_1 \mid L(\mathbf{x}, \mathcal{Z}_s)\right] = \sum_{i=1}^{s} \mathbb{E}\left[S_i \mid L(\mathbf{x}, \mathcal{Z}_s)\right] = \mathbb{E}\left[\sum_{i=1}^{s} S_i \mid L(\mathbf{x}, \mathcal{Z}_s)\right] = 1,$$

as $\sum_{i=1}^{s} S_i = 1$ by definition.

For (43), we write

$$\begin{aligned}
&\mathrm{Var}(T(\mathcal{Z}_s)) \leqslant \mathbb{E}\left[\left\|\sum_{i=1}^{s} S_i \xi_i\right\|_{\mathcal{H}}^2\right] = \mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2}\sum_{i=1}^{s} \mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x}, \mathcal{Z}_s)\}\|\xi_i\|_{\mathcal{H}}^2\right] + \\
&\mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2}\sum_{i=1}^{s}\sum_{j\neq i} \mathbb{1}\{\mathbf{X}_j \in L(\mathbf{x}, \mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x}, \mathcal{Z}_s)\}\langle \xi_i, \xi_j\rangle_{\mathcal{H}}\right].
\end{aligned}$$

We focus on the second term. For the first, the bound follows by analogous arguments. Similar as before,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2}\sum_{i=1}^{s}\sum_{j\neq i} \mathbb{1}\{\mathbf{X}_j \in L(\mathbf{x}, \mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x}, \mathcal{Z}_s)\}\langle \xi_i, \xi_j\rangle_{\mathcal{H}}\right] \\
&= s(s-1)\mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2}\mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)\}\langle \xi_1, \xi_2\rangle_{\mathcal{H}}\right] \\
&= s(s-1)\mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2}\mathbb{E}\left[\mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)\}\langle \xi_1, \xi_2\rangle_{\mathcal{H}} \mid N_{\mathbf{x}}\right]\right].
\end{aligned} \tag{46}$$

By the same argument as above

$$\begin{aligned}
&\mathbb{P}(\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s) \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)) \\
&= \frac{1}{s(s-1)} \mathbb{E}\left[\sum_{i=1}^{s} \sum_{j \neq i} \mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x}, \mathcal{Z}_s)\} \mathbb{1}\{\mathbf{X}_j \in L(\mathbf{x}, \mathcal{Z}_s)\} \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s) \right] \\
&= \frac{1}{s(s-1)} \mathbb{E}\left[N_{\mathbf{x}}(N_{\mathbf{x}} - 1) \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s) \right] \\
&= \frac{N_{\mathbf{x}}(N_{\mathbf{x}} - 1)}{s(s-1)},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)\} \mathbb{1}\{\mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)\} \langle \xi_1, \xi_2 \rangle_{\mathcal{H}} \mid N_{\mathbf{x}} \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)\} \mathbb{1}\{\mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)\} \langle \xi_1, \xi_2 \rangle_{\mathcal{H}} \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s) \right] \mid N_{\mathbf{x}} \right] \\
&= \mathbb{E}\Big[\mathbb{E}\left[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}} \mid N_{\mathbf{x}}, A, \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s) \right] \\
&\mathbb{P}(\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s) \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)) \mid N_{\mathbf{x}} \Big] \\
&= \mathbb{E}\left[\mathbb{E}\left[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}} \mid \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s), L(\mathbf{x}, \mathcal{Z}_s) \right] \mid N_{\mathbf{x}} \right] \frac{N_{\mathbf{x}}(N_{\mathbf{x}} - 1)}{s(s-1)}. \qquad (47)
\end{aligned}$$

Thus combining (46) and (47),

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2} \sum_{i=1}^{s} \sum_{j \neq i} \mathbb{1}\{\mathbf{X}_j \in L(\mathbf{x}, \mathcal{Z}_s)\} \mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x}, \mathcal{Z}_s)\} \langle \xi_i, \xi_j \rangle_{\mathcal{H}} \right] \\
&= \mathbb{E}\Big[\frac{N_{\mathbf{x}}(N_{\mathbf{x}} - 1)}{N_{\mathbf{x}}^2} \mathbb{E}\left[\mathbb{E}\left[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}} \mid \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s), L(\mathbf{x}, \mathcal{Z}_s) \right] \mid N_{\mathbf{x}} \right] \Big] \\
&\leqslant \mathbb{E}\left[\mathbb{E}\left[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}} \mid \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s), L(\mathbf{x}, \mathcal{Z}_s) \right] \right] \\
&= \mathbb{E}\left[\langle \mathbb{E}[\xi_1 \mid \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)], \mathbb{E}[\xi_2 \mid \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)] \rangle_{\mathcal{H}} \right],
\end{aligned}$$

where in the last step we used independence of $(\xi_1, \mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)\})$, $(\xi_2, \mathbb{1}\{\mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)\})$ conditionally on $L(\mathbf{x}, \mathcal{Z}_s)$ and (C9). Finally,

$$\begin{aligned}
\mathbb{E}\left[\langle \mathbb{E}[\xi_1 \mid \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)], \mathbb{E}[\xi_2 \mid \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)] \rangle_{\mathcal{H}} \right] &\leqslant \sup_{\mathbf{x} \in [0,1]^p} \|\mathbb{E}[\xi_1 \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}^2 \\
&\leqslant \sup_{\mathbf{x} \in [0,1]^p} \mathbb{E}[\|\xi_1\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}],
\end{aligned}$$

proving the claim.

■

Corollary 38. *Under the conditions of Lemma .16, assume*

$$\mathbf{x} \mapsto \mu(\mathbf{x}) = \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}] \in \mathcal{H},$$

*is Lipschitz and that the trees T in the forest satisfy **(P2)** and **(P3)**. Then*

$$\|\mathbb{E}[\hat{\mu}_n(\mathbf{x})] - \mu(\mathbf{x})\|_{\mathcal{H}} = \mathcal{O}\left(s^{-1/2\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}\right), \tag{48}$$

and

$$\|\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)]\|_{\mathcal{H}} \xrightarrow{p} \|\mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}. \tag{49}$$

If moreover,

$$\mathbf{x} \mapsto \mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}] \in \mathbb{R},$$

is Lipschitz, then:

$$\mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] \xrightarrow{p} \mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}]. \tag{50}$$

Proof By (42), it holds as in [181]

$$\|\mathbb{E}[T(x, \mathbf{Z})] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}} = \|\mathbb{E}[\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]]\|_{\mathcal{H}}.$$

Let

$$s_1^* = \sqrt{d}\left(\frac{s}{2k-1}\right)^{-0.51\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}, s_2^* = d\left(\frac{s}{2k-1}\right)^{-0.5\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}$$

Then it follows from Lemma .16 that

$$\mathbb{P}(L(\mathbf{x}, \mathcal{Z}_s) \geqslant s_1^*) \leqslant s_2^*,$$

while by Lipschitz continuity: $\mathbb{E}[\|\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}] \leqslant L\mathbb{E}[\mathrm{diam}(L(\mathbf{x}, \mathcal{Z}_s))]$, where L is the Lipschitz constant. Thus,

$$\begin{aligned}
&\|\mathbb{E}[\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]]\|_{\mathcal{H}} \\
&\leqslant \|\mathbb{E}[(\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}])\, \mathbb{1}\{\mathrm{diam}(L(\mathbf{x}, \mathcal{Z}_s)) \geqslant s_1^*\}]\|_{\mathcal{H}} \\
&+ \|\mathbb{E}[(\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}])\, \mathbb{1}\{\mathrm{diam}(L(\mathbf{x}, \mathcal{Z}_s)) < s_1^*\}]\|_{\mathcal{H}} \\
&\leqslant \mathbb{E}[\|\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}} \mathbb{1}\{\mathrm{diam}(L(\mathbf{x}, \mathcal{Z}_s)) \geqslant s_1^*\}] \\
&+ \mathbb{E}[\|\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}} \mathbb{1}\{\mathrm{diam}(L(\mathbf{x}, \mathcal{Z}_s)) < s_1^*\}] \\
&\leqslant \left(\sup_{\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^p} \|\mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_1] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_2]\|_{\mathcal{H}}\right) \mathbb{P}(\mathrm{diam}(L(\mathbf{x}, \mathcal{Z}_s)) \geqslant s_1^*) + L s_1^* \\
&\leqslant \left(\sup_{\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^p} \|\mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_1] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_2]\|_{\mathcal{H}}\right) s_2^* + L s_1^* \\
&\lesssim \left(\sup_{\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^p} \|\mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_1] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_2]\|_{\mathcal{H}}\right) s_2^*,
\end{aligned}$$

since $s_1^*/s_2^* \to 0$. Due to the Lipschitz condition

$$\sup_{\mathbf{x}_1,\mathbf{x}_2\in[0,1]^p} \|\mathbb{E}[\xi \mid \mathbf{X}=\mathbf{x}_1] - \mathbb{E}[\xi \mid \mathbf{X}=\mathbf{x}_2]\|_{\mathcal{H}} \leqslant L \sup_{\mathbf{x}_1,\mathbf{x}_2\in[0,1]^p} \|\mathbf{x}_1-\mathbf{x}_2\|_{\mathbb{R}^d} = O(1).$$

Finally by the reverse triangle inequality

$$\begin{aligned}\big|\|\mathbb{E}[\xi \mid \mathbf{X}\in L(\mathbf{x},\mathcal{Z}_s)]\|_{\mathcal{H}} - \|\mathbb{E}[\xi \mid \mathbf{X}=\mathbf{x}]\|_{\mathcal{H}}\big| &\leqslant \|\mathbb{E}[\xi \mid \mathbf{X}\in L(\mathbf{x},\mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X}=\mathbf{x}]\|_{\mathcal{H}} \\ &\leqslant L\mathrm{diam}(L(\mathbf{x},\mathcal{Z}_s)),\end{aligned}$$

and if $\mathbf{x}\mapsto \mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}]$ is Lipschitz as well, also

$$|\mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X}\in L(\mathbf{x},\mathcal{Z}_s)] - \mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}]| \leqslant C\mathrm{diam}(L(\mathbf{x},\mathcal{Z}_s)).$$

Since $\mathrm{diam}(L(\mathbf{x},\mathcal{Z}_s)) \xrightarrow{p} 0$, as $s\to\infty$, (49), respectively (50) hold true.

As the expectation of the forest is the same as that of one tree:

$$\mathbb{E}[\hat{\mu}_n(\mathbf{x})] = \mathbb{E}\left[T(\mathbf{x},\varepsilon;\mathbf{Z}_{i_1},\ldots,\mathbf{Z}_{i_{s_n}})\right],$$

the result follows. ■

This leads us to the proof of of Theorem 32 in the main text.

Theorem 32. *Suppose that our forest construction satisfies properties **(P1)–(P5)**. Assume additionally that k is a bounded and continuous kernel and that we have a random design with $\mathbf{X}_1,\ldots,\mathbf{X}_n$ independent and identically distributed on $[0,1]^p$ with a density bounded away from 0 and infinity. If the subsample size s_n is of order n^β for some $0<\beta<1$, the mapping*

$$\mathbf{x}\mapsto \mu(\mathbf{x}) = \mathbb{E}[\mu(\delta_{\mathbf{Y}}) \mid \mathbf{X}=\mathbf{x}] \in \mathcal{H},$$

is Lipschitz and $\sup_{\mathbf{x}\in[0,1]^p}\mathbb{E}[\|\mu(\delta_{\mathbf{Y}})\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}] < \infty$, we obtain the consistency w.r.t. the RKHS norm:

$$\|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} = O_p\left(n^{-\gamma}\right), \tag{18}$$

for $\gamma = \frac{1}{2}\min\left(1-\beta, \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}\cdot\beta\right)$.

Proof We first note that $\sup_{\mathbf{x}\in[0,1]^p}\mathbb{E}[\|\mu(\delta_{\mathbf{Y}})\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}] < \infty$ together with (43) implies $\mathrm{Var}(T) < \infty$. Thus, from Markov's inequality and Lemma .15,

$$\mathbb{P}\left(n^\gamma||\hat{\mu}_n(\mathbf{x}) - \mathbb{E}[\hat{\mu}_n(\mathbf{x})]||_{\mathcal{H}} > \varepsilon\right) \leqslant \frac{n^{2\gamma}}{\varepsilon^2}(s/n + s^2/n^2)\mathrm{Var}(T) = \frac{1}{\varepsilon^2}O(n^{2\gamma+\beta-1}).$$

Thus

$$n^\gamma||\hat{\mu}_n(\mathbf{x}) - \mathbb{E}[\hat{\mu}_n(\mathbf{x})]||_{\mathcal{H}} = O_p(1),$$

for $\gamma \leqslant (1-\beta)/2$. In particular, it goes to zero for any $\varepsilon > 0$, if $\gamma < (1-\beta)/2$. Since,

$$n^{\gamma} \left\| \hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}) \right\|_{\mathcal{H}} \leqslant n^{\gamma} \left\| \hat{\mu}_n(\mathbf{x}) - \mathbb{E}[\hat{\mu}(\mathbf{x})] \right\| + n^{\gamma} \left\| \mathbb{E}[\hat{\mu}(\mathbf{x})] - \mu(\mathbf{x}) \right\|_{\mathcal{H}},$$

the result follows as soon as the second expression goes to zero. Now from Theorem 38, with $C_\alpha = \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}$,

$$\left\| \mathbb{E}[\hat{\mu}_n(\mathbf{x})] - \mu(\mathbf{x}) \right\|_{\mathcal{H}} = O\left(s_n^{-1/2C_\alpha \frac{\pi}{p}} \right) = O\left(n^{-1/2\beta C_\alpha \frac{\pi}{p}} \right).$$

This goes to zero provided that,

$$1/2\beta C_\alpha \frac{\pi}{p} > \gamma.$$

■

To prove Corollaries 34 and 35, we first need another auxiliary result:

Lemma .18. *Let (S,d) be a separable metric space and $\mathbf{X}_n : (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$, $n \in \mathbb{N}$ and $\mathbf{X} : (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$ be measurable. Then $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ if and only if for every subsequence $n(k)$ there exists a further subsequence $n(k(l))$ such that*

$$\mathbf{X}_{n(k(l))} \to \mathbf{X} \text{ a.s.} \tag{51}$$

Proof If $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$, then so does any subsequence $\mathbf{X}_{n(k)}$. By well-known results, see e.g. [177, Chapter 2], this implies that there exists a further subsequence $\mathbf{X}_{n(k(l))}$ such that a.s. convergence holds.

As is well known, there exists a metric ρ on $\mathcal{P}(S)$ such that $\rho(\mathbf{X}_n, \mathbf{X}) \to 0$ iff $\mathbf{X}_n \xrightarrow{p} X$, see e.g., [46, Chapter 11]. Now assume that for any subsequence we can find a further subsequence such that (51) holds, but the overall sequence does not converge in probability. Then we can build a subsequence such that for some $\varepsilon > 0$,

$$\rho(\mathbf{X}_{n(k)}, \mathbf{X}) \geqslant \varepsilon$$

for all elements of that subsequence. Thus any further subsequence will also not convergence in probability and consequently cannot converge a.s. This proves the claim. ■

We note that the set A with $P(A) = 1$ on which (51) holds is allowed to depend on the subsequence. Corollary 34 and 35 are finally proven jointly in the following Corollary. The proof is motivated by the tools used in [9].

Corollary 39. *Assume that one of the following two sets of conditions holds:*

(a) The kernel k is bounded, (jointly) continuous and has

$$\int\int k(\mathbf{x},\mathbf{y})d\mathcal{P}(\mathbf{x})d\mathcal{P}(\mathbf{y}) > 0 \;\; \forall \mathcal{P} \in \mathcal{M}_b(\mathbb{R}^d)\backslash\{0\}. \tag{52}$$

Moreover, $\mathbf{y} \mapsto k(\mathbf{y}_0,\mathbf{y})$ is vanishing at infinity, for all $\mathbf{y}_0 \in \mathbb{R}^d$.

(b) The kernel k is bounded, shift-invariant, (jointly) continuous and ν in the Bochner representation in (11) *is supported on all of $\mathbb{R}^d$. Moreover, $\mathbf{Y}$ takes its values almost surely in a closed and bounded subset of $\mathbb{R}^d$.*

Then, under the conditions of Theorem 32, we have for any bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$ that DRF consistently estimates the target $\tau(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}]$ for any $\mathbf{x} \in [0,1]^p$:

$$\sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i) \xrightarrow{p} \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}].$$

Moreover,

$$\hat{F}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) \xrightarrow{p} F_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{t})$$
$$\hat{F}^{-1}_{Y_i|\mathbf{X}=\mathbf{x}}(t) \xrightarrow{p} F^{-1}_{Y_i|\mathbf{X}=\mathbf{x}}(t),$$

for all points of continuity $\mathbf{t} \in \mathbb{R}^d$ and $t \in \mathbb{R}$ of $F_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\cdot)$ and $F^{-1}_{Y_i|\mathbf{X}=\mathbf{x}}(\cdot)$ respectively.

Proof As shown in [163, Theorem 3.2], (a) implies that k metrizes weak convergence. Similarly, from Theorem 9 in [165], it follows that k is characteristic on the compact subspace of $\mathbb{R}^d$ in which Y takes its value almost surely. Thus, ignoring the Null set, Theorem 23 in [165] implies that k metrizes the weak convergence in this case as well. Thus in both cases $\|\hat{\mu}_{n(k(l))}(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} \to 0$ implies weak convergence of $\hat{\mu}_{n(k(l))}(\mathbf{x})$ to $\mu(\mathbf{x})$.

From Theorem .18, for any subsequence, we can choose a further subsequence, such that

$$\|\hat{\mu}_{n(k(l))}(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} \to 0 \text{ , a.s.}$$

and since it is assumed that $\|\cdot\|_{\mathcal{H}}$ metrizes weak convergence, $\hat{\mu}_{n(k(l))}(\mathbf{x})$ converges weakly to $\mu(\mathbf{x})$ on a set A, depending on the subsequence, with $\mathbb{P}(A) = 1$. Let $C_b(\mathbb{R}^d)$ denote the space of all bounded continuous functions on $\mathbb{R}^d$. By the Portmanteau theorem (see e.g. [46]), this implies that on A

(I) $\int f d\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x}) \to \int f d\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ for all $f \in C_b(\mathbb{R}^d)$

(II) $\hat{F}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{t}) \to F_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{t})$ for all continuity points $\mathbf{t} \in \mathbb{R}^d$ of $F_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\cdot)$,

where we omitted the dependence on the subsequence. But, since the subsequence $n(k)$ was arbitrary, this immediately implies

(I') $\int f d\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) \xrightarrow{p} \int f d\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ for all $f \in C_b(\mathbb{R}^d)$

(II') $\hat{F}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(t) \xrightarrow{p} F_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(t)$ for all continuity points t of $F_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\cdot)$,

for the overall sequence.

On the other hand, (II) implies that on A, for the given subsequence $n(k(l))$, $\hat{F}_{Y_i|\mathbf{X}=\mathbf{x}}(t) \to F_{Y_i|\mathbf{X}=\mathbf{x}}(t)$ for all $t \in \mathbb{R}$ at which $F_{Y_i|\mathbf{X}=\mathbf{x}}(\cdot)$ is continuous. Using for each $\omega \in A$ the arguments in [177, Chapter 21], this implies that

$$\hat{F}^{-1}_{Y_i|\mathbf{X}=\mathbf{x}}(t) \to F^{-1}_{Y_i|\mathbf{X}=\mathbf{x}}(t), \text{ for all continuity points } t \text{ of } F^{-1}_{Y_i|\mathbf{X}=\mathbf{x}}(t)$$

on A for the given subsequence. Again, as the subsequence $n(k)$ was arbitrary, this implies the result. ■

3 Simulation Details

In this section we describe in detail all our simulations shown in the main paper, together with the data used in the analysis. The data sets are available in the R-package `drf` as well.

3.1 Air quality data

3.1.0.1 Data. This data is obtained from the website of the Environmental Protection Agency website (`https://aqs.epa.gov/aqsweb/airdata/download_files.html`). We have daily measurements for 5 years of data (2015-2019) for 6 'criteria' pollutants that form the Air Quality Index (AQI):

- O_3 - ground ozone (8 hours' average, expressed in pieces per million (ppm))
- SO_2 - sulfur dioxide (1 hour average, expressed in pieces per billion (ppb))
- CO - carbon monoxide (8 hours' average, expressed in pieces per million (ppm))
- NO_2 - nitrogen dioxide (1 hour average, expressed in pieces per billion (ppb))
- PM2.5 - fine particulate matter smaller than 2.5 micrometers (24 hours' average, expressed in $\nu g/m^3$)
- PM10 - large particulate matter, smaller than 10 micrometers (24 hours' average, expressed in $\mu g/m^3$)

For the above quantities, we have the maximal and mean value within the same day. In our analysis we have used only the maximal intraday values.

The pollutants are measured at different measurement sites. For each site we have information about

- site address (street, city, county, state, zip code)
- site coordinates (longitude and latitude)
- site elevation
- location setting (rural, urban, suburban)
- how the land is used within a $1/4$ mile radius (agricultural, forest, desert, industrial, commercial, residential, blighted area, military reservation, mobile)
- date when the measurement site was put in operation
- date when the measurement site was decommissioned (NA if the site is still operational)

We have information about $19'739$ sites, much more than the number of $2'419$ sites from which we have measurements in years 2015-2019, since many sites were only operating in the past and are decommissioned.

In total there is $5'305'859$ pollutant measurements. Many pollutants are measured at the same site, but it is important to note that not every site measures every pollutant, so there is a lot of 'missing' measurements. It can also occur that there are several measuring devices for the same pollutant at the same site, in which case we just average the measurements across the devices and do not report those measurements separately.

3.1.0.2 Analysis. Since we have a lot of missing data, we use only the data points (identified by the measurement date and the measurement site) for which we have measurements of all the pollutants chosen as the responses. For that reason we also do not train DRF with all 6 pollutants as the responses, but only those that we are interested in, since only 64 sites measure all pollutants. For computational feasibility, we only use $50'000$ of the available measurements for the training step. We also omit the states Alaska and Hawaii and the US territories for plotting purposes.

To obtain the results displayed in Figure 2, we train the DRF with the measurements (intraday maximum) of the two pollutants PM2.5 and NO_2 as the responses, and the site longitude, latitude, elevation, land use and location settings as the predictors. We manually choose two decommissioned measurement sites (for which we have no measurements in years 2015-2019) as the test points. For each test point we obtain the weights to all training measurements. We

further combine the weights for all measurements corresponding to the same site, which is represented by the symbol size in the top row. The bottom row shows the estimated distribution of the response, where the transparency (alpha) each training point corresponds to the assigned weight. We also add some estimated contours.

For all plots in Figures 2 and 5, we train the single DRF with the same set of predictor variables and take the three pollutants O_3, SO_2 and PM2.5 as the responses. In this way we still have training data from many different sites (see the above discussion on missing data) and moreover, those are the 3 pollutants that most likely cross the threshold for the "Good" AQI category set by the EPA. Carbon monoxide (CO), for example, almost never crosses this threshold.

In left plot of Figure 6, we compare the estimated CDF value with the standard classification forest which has the indicator $\mathbb{1}(O_3 < 0.055\text{ppm}, SO_2 < 36\text{ppb}, \text{PM2.5} < 12.1\mu g/m^3)$ as the univariate response. In the right plot, we obtain the estimated CDF by fitting for each threshold a separate classification forest with an indicator $\mathbb{1}(O_3 \leqslant \text{threshold})$. We pick a test point such that the classification performs bad, just to illustrate that its estimated CDF need not be monotone, which cannot happen with DRF. In most of the cases, the estimated CDFs are very similar, as can also be seen from the left plot in Figure 6.

3.2 Benchmark Analysis

In this part we compare the resulting distributional estimates of DRF with several benchmark methods on a number of data sets. Because our target of estimation is now the whole conditional distribution one needs to use a distributional loss, and there appears to be no well-established choice in the literature. Furthermore, for any test point $\mathbf{x}_i$ we only have one observation $\mathbf{y}_i$ from $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, which makes performance evaluation of our estimator $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ hard. We thus use the following performance measure:

- **(NLPD loss)** For a fixed conditional distribution estimator $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)$, we sample a set of $m = 500$ observations from which we estimate the conditional density via a Gaussian kernel estimator, using the L_2 loss with scale components and the median heuristic for the choice of the bandwidth parameter. We then evaluate the negative log-likelihood of the test observation $\mathbf{y}_i$ implied by the kernel estimate of the distribution and average over the test set (consisting of multiple pairs $(\mathbf{x}_i, \mathbf{y}_i)$). To reduce the dependence of these results on single large values of the log-likelihood, we use an $0.05-$trimmed mean to average the losses over the training set.

This loss definition provides a fair way to compare the ability to estimate the conditional distribution since most of the candidate methods only allow for sampling from the estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)$.

3.2.1 Competing methods

We compare DRF that uses the MMD splitting criterion with many existing methods that can be used for estimation of the conditional distribution.

- **Nearest Neighbor (k-NN)**: The standard k-nearest neighbors algorithm with the Euclidean metric. An estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ at a test point $\mathbf{x}$ is defined by a uniform distribution over the k nearest observations in the training set. k is chosen to be the square root of the training set size.

- **Gaussian kernel (kernel)**: The estimate of the conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ at a test point $\mathbf{x}$ is obtained by assigning to each training observation $(\mathbf{x}_i, \mathbf{y}_i)$ the weight proportional to the Gaussian kernel $k(\mathbf{x}, \mathbf{x}_i)$, analogously to usual kernel estimation methods. Median heuristic is used for bandwidth selection.

- **Homogeneous distribution models**: This method makes the homogeneity assumption that the residuals have constant distribution and only the conditional mean changes. The estimate of the conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is obtained by first fitting a regression method of choice, computing the residuals, assigning the same weight to every residual and then adding those residuals to the predicted mean. We chose three different methods for the mean estimation:

  1. **Random Forests (RF)**, a classical univariate regression forest is fitted independently for each response component;
  2. **Extreme Gradient Boosting (XGBoost)**, a tree gradient boosting model (as described in [27]) is fitted independently for each response;
  3. **Deep Neural Network (DNN)**, a single deep neural network is fitted to predict the conditional mean of each response.

- **Conditional Generative Adversarial Neural Network (CGAN)**: The estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is obtained through sampling from the discriminator with conditional feature $\mathbf{x}$. The implementation of the CGAN is taken from [2]. The architecture of the neural networks was taken to be the best one for the considered data sets among a set of candidates.

- **Conditional Variational Auto-Encoder (CVAE)**: The estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is obtained through sampling from the decoder of the CVAE with conditional feature $\mathbf{x}$. The implementation of the CVAE follows the one in [160]. The architecture of the neural networks was taken to be the best one for the considered data sets among a set of candidates.

- **Masked Autoregressive Flow (MAF)**: The estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is obtained through sampling from the normalizing flow model with conditional feature $\mathbf{X}$. The implementation of the model follows the one presented in [131]. The number of layers is chosen to be the best value from a set $\{5, 10\}$ for the considered data set.

- **Conditional Mean Embedding (CME)**: The CME is calculated as in (16) with the weights given as in e.g., [162, 126, 132]. We choose both kernels to be Gaussian kernels with $\sigma = 0.01$, as in [132] and set $\lambda = 0.01$. The estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is obtained through sampling from the obtained weights, renormalized such that they lie in $[0, 1]$ and sum up to one.

3.2.2 Benchmark data sets

Many benchmark data sets used come from the multiple target regression literature, where only the conditional means of the multivariate response is considered. We have used the data sets: **jura**, **slump**, **wq**, **enb**, **atp1d**, **atp7d**, **scpf**, **sf1** and **sf2** collected in the `Mulan` [172] library. Description about the dimensionality of the data sets, together with the descriptions of the outcomes and the regressors can be found in [172] with links to the relevant papers introducing these data sets. In each data set categorical variables have been represented by the one-hot dummy encoding, the observations with missing data were removed together with constant regressors.

We additionally added 5 data sets obtained from the data sets used in the main paper:

- **copula**: Simulated Gaussian copula example where the response Y is bivariate and whose marginal distribution is $N(0, 1)$, but the correlation between Y_1 and Y_2 depends on X_1.

- **birth1**: This data set is created from the CDC natality data and contains many covariates as predictors and the pregnancy length and birthweights as the responses.

- **birth2**: This data set is similar as the above one, but we take pregnancy length as the predictor and add 3 more measures of baby's health as the response: APGAR score measured 5 minutes after birth and indicators whether there were any abnormal conditions and congenital anomalies.

- **wage**: This data set is created from the 2018 American Community Survey. We take the logarithmic hourly wage and gender as the response, as it was done in the fairness example in the main paper.

- **air**: This data set is obtained from the EPA air quality data. All six pollutants were taken as the response and we add both the information about the measuring site (location, which

setting it is in, etc.), as well as the temporal information when the measurement has taken place (month, day of the week).

3.3 Births data

3.3.0.1 Data. This data set is obtained from the CDC Vital Statistics Data Online Portal (`https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm`) and contains the information about the $\approx$ 3.8 million births in 2018. However, as we do not need this many data points, we subsample $300'000$ of them. Even though the original data contains a lot of variables, we have taken only the following variables from the source data:

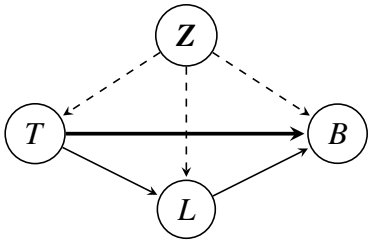
- mother's age, height, weight before the pregnancy and BMI before pregnancy
- mother's race (black, white, asian, NHOPI, AIAN or mixed), marital status (married or unmarried) and the level of education (in total 8 levels)
- father's age, race and education level
- month and year of birth
- plurality of the birth (how many babies were born at once)
- whether and when the prenatal care started
- length of the pregnancy
- delivery method (vaginal or C-section)
- birth order - the total number of babies born by the same mother (including the current one)
- birth interval - number of months passed since last birth (NA if this is the first child)
- number of cigarettes smoked per day on average during the pregnancy
- birthweight (in grams) and gender of the baby
- APGAR score (taken after 5min and 10min)
- indicators whether baby had any abnormal condition or some congenital anomalies

3.3.0.2 Analysis. After removing the data points with any missing entries and taking only the data points where the race of both parents is either black, white or Asian (for nicer plotting), we are left with $183'881$ data points. We use randomly chosen $100'000$ data points for training the DRF. We take the birthweight and the pregnancy length as the bivariate response and for the predictors we take: mother's age, race, education, marital status, height, BMI; father's age, race and education level; birth plurality, birth order, delivery method, baby's gender, number of cigarettes and indicator whether prenatal care took place.

For arbitrary test points from the data we can get the estimated weights by the fitted DRF, thus estimating the joint distribution of birthweight and pregnancy length conditional on all other variables mentioned above. Two such distributions are shown in Figure 10. In addition we use the weights to fit a parametric model for the mean and 0.1 and 0.9 quantiles. This is done as follows:

- We slightly upweight the data points where the pregnancy length is significantly above or below the usual range. This is to avoid the bulk of the data points to dominate the fit obtained for very long or short pregnancies.
- We apply the transformation $f(\cdot) = \log(\log(\cdot))$ on both the pregnancy length and the birthweight since then the scatterplots look much nicer.
- We estimate the mean with smoothing splines with a small manually chosen number of degrees of freedom.
- The fitted mean is subtracted from the response (birthweight). The residuals seem well behaved with maybe slight, seemingly linear trend in standard deviation.
- We fit the 0.1 and 0.9 quantiles as the best linear functions that minimize the sum of quantile losses, by using the `quantreg` package [93].
- The data is transformed back on the original scale by using the function $f^{-1}(\cdot) = \exp(\exp(\cdot))$.

For the right plot in Figure 10, we have the following causal graph, as mentioned in the main paper:



We want to determine the direct effect (indicated in bold) of the twin pregnancy T on the birthweight B that is due to sharing of resources by the babies (space, food etc.) and is not

due to the fact that twin pregnancy causes shorter pregnancy length L, which in turn causes the smaller birthweight. Another big issue is that we have confounding factors $\mathbf{Z}$ which can directly affect B, L and T. For example, the number of twin pregnancies significantly depends on the parents' race, but so do the pregnancy length and the birthweight, e.g. black people have more twins, shorter pregnancies and smaller babies. We take all other variables as the potential confounders $\mathbf{Z}$ and adjust for all of them (mother's age, race, education, marital status, height, BMI; father's age, race and education level; birth plurality, birth order, baby's gender, number of cigarettes and indicator whether prenatal care took place). In order to do it, we fit the same DRF as before, where $\mathbf{Z}$ and T are the predictors and (B, L) is the bivariate response for which we can fit the parametric model described above. We compute then the interventional distribution $\mathbb{P}(B \mid do(T = t, L = l))$ for all values of t and l, by using the do-calculus to adjust the confounding $\mathbf{Z}$ via the backdoor criterion [133], where we also use the obtained parametric regression fit. In this way we can generalize the fit well, which is important when doing the do-calculus, since we are interested in some hypothetical combinations of covariates which might not occur frequently in the observed data, such as very long twin pregnancies.

3.4 Wage data

3.4.0.1 Data. The PUMS (Public Use Microdata Area) data from the 2018 1-Year American Community Survey is obtained from the US Census Bureau API (`https://www.census.gov/content/dam/Census/data/developers/api-user-guide/api-guide.pdf`). The survey is sent to $\approx$ 3.5 million people annually and aims to give more up to date data than the official census that is carried out every decade. The 2018 data set has $3'214'539$ anonymized data points for the 51 states and District of Columbia. Even though the original survey contains many questions, we have retrieved only the subset of variables that might be relevant for the salaries:

- person's gender, age, race (AIAN, black, white, asian, mix, NHOPI, other), indicator of hispanic origin, state of residence, US citizenship indicator (5 ordered levels), indicator whether the person is foreign-born
- person's marital status, number of own children in the same household and the number of family members in the same household
- person's education level (24 ordered levels) and level of English knowledge (5 ordered levels)
- person's employment status (employed, not at work, not in workforce, unemployed)

- for employed people we have annual salary earnings, number of weeks worked in a year and average number of hours worked per week
- for employed people we have employer type (government, non-profit company, for-profit company, self-employed), occupation (530 levels), industry where the person works (271 levels) and the geographical unit where the person works (59 levels)
- statistical weight determined by the US Census Bureau which aims to correct sampling bias

For our purposes, since we want to analyze the unfairness of the gender pay gap, we consider only employed people that are at least 17 years of age, have worked full-time (at least 48 weeks in a year) and have worked at least 16 hours a week on average. We also omit the self-employed persons, since they often report zero annual salary and the pay gap there, if exists, cannot be called unfair as the salary is not determined by any employer. Since there are no missing data which would need to be omitted, we finally end up with $1'071'866$ data points.

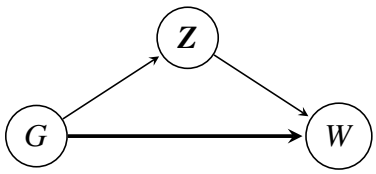
3.4.0.2 Analysis. We scale the salary with the amount of time spent working (determined from the number of weeks worked and average hours worked per week) to compute the logarithm of the hourly wages. The scaling with the time spent working is necessary, since full-time employed men spend on average 11% more time working than women. The logarithmic transformation is used since the salaries are very skewed (positively) and logarithmic wages show nice behavior.

We also reduce the large number of levels of some of the categorical variables: for the occupation we use the group of 530 jobs into 20 categories provided in the SOC system (`https://www.bls.gov/soc/`); for the industry information we group the 271 possibilities in 23 categories as is done in the NIACS classification (`https://www.bls.gov/bls/naics.htm`); for the work place we group the 59 US states and foreign territories into 9 economic regions (including the "abroad" category), as determined by the Bureau of the Economic Analysis (`https://apps.bea.gov/regional/docs/regions.cfm`).

We want to investigate how the logarithmic hourly wage W is affected by the gender G, depending on the other factors $\mathbf{Z}$: age, race, hispanic origin, citizenship, being foreign-born, marital status, family size, number of children, education level, knowledge of English, occupation, industry type and place of work. To do this, we train DRF with bivariate response (W, G) and predictors $\mathbf{Z}$ on a subsample of $300'000$ data points. With it we can answer the following: For fixed values of the covariates $\mathbf{Z} = \mathbf{z}$, what are the distributions of salaries of men and women. In addition, we can determine the "propensities", i.e. the proportion of men and women corresponding to $\mathbf{Z} = \mathbf{z}$. This information is displayed in the top row of Figure 11 for a

combination of covariates corresponding to some person in the left-out data. It illustrates how the distribution of salaries and their relationship can vary with different covariates $\mathbf{Z}$.

We do not only want to determine how different covariates $\mathbf{Z}$ affect the salary distribution, but we want to quantify the overall fairness of the pay, after appropriate adjustments. In Figure 11, we can see that the observed salaries of men and women differ noticeably, and this difference in the logarithmic wages means that an average woman has 17% smaller salary than an average men. However, the question is how much of this difference is "fair". For example, the effect of the gender on the salary can be mediated through some variables such as, for example, the occupation, workplace or the level of education and we are only interested in the direct effect. This is illustrated in the following causal graph:



If we assume that people have the freedom to choose such variables themselves, the pay gap which arises from such different choices for men and women is fair and those variables are resolving variables [86]. Another way that the pay gap can be explained is that some of the variables are not statistically independent of the gender in the population of full-time employed people (e.g. the race or the age), but they themselves have an effect on the salary.

In order to address those issues, we compute the distribution of the nested counterfactual $W(\text{male}, \mathbf{Z}(\text{female}))$, corresponding to the wages of a person that has characteristics $\mathbf{Z}$ as a woman, but which was treated as a man for obtaining the salary. Such distribution can be computed from the DRF, as described in the main paper: we randomly draw a female person and for its characteristics $\mathbf{z}$ we obtain the conditional distribution of wages of men with those characteristics $\mathbb{P}(W \mid G = \text{male}, \mathbf{Z} = \mathbf{z})$ via the weights. Those distributions are averaged over random draw of $1'000$ women (that were not used in the training step of the DRF). In case that the difference in salary is fair, the distribution of the counterfactual salary $W(\text{male}, \mathbf{Z}(\text{female}))$ should be exactly the same as the observed distribution of women's wages. However, we can see that this is not the case and that the median salaries of the two distributions differ by 11%. Even though this is smaller than the 17% we obtain by comparing only the observational distributions, it still shows that women are paid less compared to men.

4 Additional Synthetic Examples

4.1 Univariate distributional regression.

The univariate response (case $d = 1$) is by far the most studied case in the regression literature. However, at the level of the whole conditional distribution and compared to the multivariate case, the range of practically interesting targets $\tau(\mathbf{x})$ is quite reduced, e.g. conditional mean of some functional $\mathbb{E}(f(Y) \mid \mathbf{X})$ or conditional quantiles $Q_\alpha(Y \mid \mathbf{X})$. In Figure 4, we have compared the performance of DRF (which uses the MMD splitting criterion) with 3 different tree-based univariate methods that can estimate the conditional quantiles in the univariate case:

- **QRF**: the quantile regression forest introduced in [117], which is equivalent to DRF_{CART} in the univariate case and uses the standard forest construction [15] to get the weights.
- **GRF**: the quantile forest proposed in [4] based on the generalized random forest algorithm.
- **TRF**: the transformation forest, a model-based recursive partitioning approach, introduced in [74].

Additionally to the visual inspection of the performance given in Figure 4, we present here a formal performance comparison for the three simulation scenarios also described in the main paper. The first two scenarios correspond exactly to the examples given in [4] for the quantile version of the GRF, which serve to illustrate its advantage compared to the conventional quantile regression forest (QRF) [117]. Scenario 3, in addition, aims at assessing the ability to detect a change of distribution that does not relate to a change in the first two moments.

method	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
DRF	**0.180**	**0.353**	**0.402**	**0.349**	**0.177**	**0.267**	**0.518**	**0.589**	**0.514**	**0.264**	0.140	**0.298**	0.371	**0.351**	**0.198**
QRF	0.182	0.357	0.482	0.351	0.179	0.285	0.526	0.592	0.521	0.281	0.144	0.299	0.376	0.357	0.204
GRF	0.183	0.359	0.409	0.354	0.180	0.278	0.522	0.590	0.517	0.274	**0.139**	0.299	**0.371**	**0.351**	0.200
TRF	0.183	0.358	0.408	0.353	0.180	0.272	0.519	0.590	0.516	0.268	0.145	0.300	0.373	**0.351**	0.200
5-NN	0.232	0.402	0.452	0.404	0.239	0.354	0.587	0.657	0.584	0.340	0.187	0.348	.424	0.406	0.260
20-NN	0.192	0.368	0.418	0.366	0.192	0.290	0.535	0.606	0.533	0.283	0.146	0.310	0.382	0.365	0.211
40-NN	0.187	0.364	0.413	0.360	0.185	0.283	0.528	0.596	0.522	0.273	0.141	0.303	0.376	0.357	0.204

Table 2: Average quantile losses for scenarios 1 (left), 2 (middle), 3 (right) over the repeated out-of-sample validations.

The performance of each method is evaluated as follows: We consider the quantile (pinball) loss for the resulting quantile estimates provided by each candidate method for the different per-

centiles $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The losses are presented and computed based on repeated (10 times) out-of-sample validation (with a $70 - 30\%$ ratio between the training and testing sets sizes). The results are presented respectively for each scenario in Table 2. We additionally include the estimates obtained by k-nearest neighbor algorithm for several different values of k.

Table 3: Average mean squared errors (MSE) for the three scenarios described above over 10 repeated out-of-sample validations for estimating the conditional mean.

method	**SC1**	**SC2**	**SC3**
RF	1.0545	2.4940	0.9624
DRF	**1.0412**	**2.4561**	**0.9340**

Furthermore, Table 3 shows non-inferiority of DRF compared to the standard Random Forest for the classical task of estimating the conditional mean. We observe that DRF has a good relative performance that makes it on par with existing algorithms, some of which specially designed for the problem of estimating conditional quantiles. Furthermore, it seems that the MMD splitting criterion improves the CART criterion for distributional regression in a general heterogeneous case (see e.g. scenarios 2 and 3), since the CART criterion is suitable only for detecting the change in the conditional mean, unlike MMD.

Dependence of the estimated quantiles on X_1 for each method (except the k-nearest neighbors) is displayed in the main paper in Figure 4. In addition, the estimates of 2-Wasserstein distance to the true conditional distribution, quantifying the difference in the estimated CDFs, are shown in Figure 12.

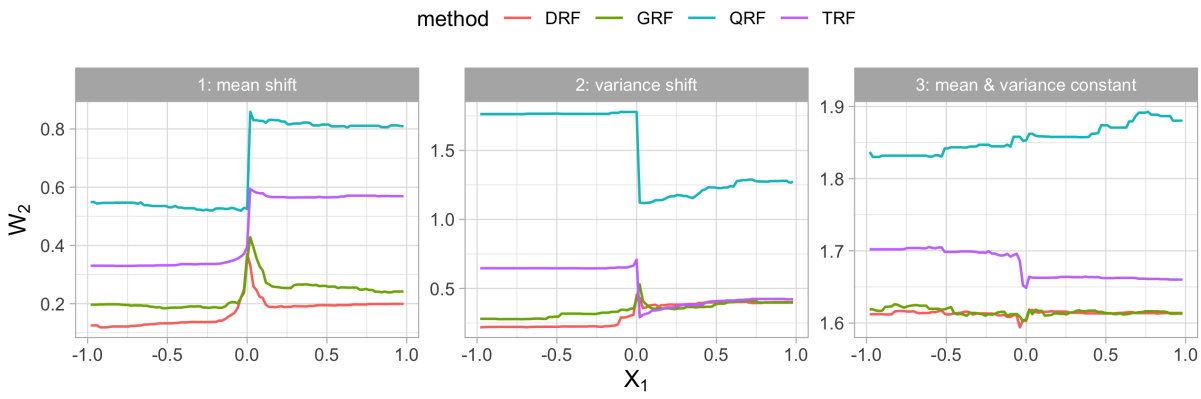
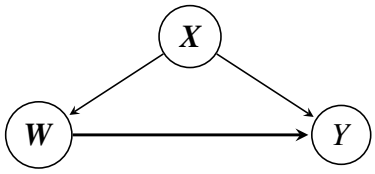


Figure 12: Scatter plot of discrete estimates of the 2-Wasserstein distance between the estimated and true conditional distribution against X_1 for a grid of test points of the form $(x_1, 0, \ldots, 0)$. The 2-Wasserstein distance is estimated over a grid of 100 quantiles with levels equally spaced on $[0, 1]$. Different colors corresponds to different methods: DRF (red), GRF (green), QRF (blue), TRF (purple).

4.2 Heterogeneous regression and causal effects

We explore here the performance of DRF on the synthetic data for the setup of heterogeneous regression, where we want to obtain the regression fit of Y on the explanatory (or treatment) variables $\mathbf{W}$, but where this fit might change depending on values of $\mathbf{X}$. This can be done by DRF by using $\mathbf{X}$ as predictors and $(\mathbf{W}, Y)$ as the response and then using some standard regression method for regressing Y on $\mathbf{W}$ in the second step, having already obtained the weights that describe the conditional distribution $\mathbb{P}((\mathbf{W}, Y) \mid \mathbf{X} = \mathbf{x})$.

The most important such setup is when the data come from the following causal graph:



In the case of such a causal graph, $\mathbf{X}$ are confounding variables, which we need to adjust for to understand the causal effect of $\mathbf{W}$ on Y. Not only can the marginal distributions of Y and $\mathbf{W}$ be affected by $\mathbf{X}$, but also the regression fit (e.g. the regression coefficients).

4.2.1 CATE and ATE

One special case of this setup that is intensively studied in the causal literature is when W is a (univariate) binary treatment variable. In this case we are interested in the distribution of the potential outcomes $Y(W = 0)$ and $Y(W = 1)$ and especially in their difference. It is commonly measured by using the Conditional Average Treatment Effect

$$\text{CATE}(\mathbf{x}) = \mathbb{E}[Y(W = 1) - Y(W = 0) \mid \mathbf{X} = \mathbf{x}]$$

and the Average Treatment Effect

$$\text{ATE} = \mathbb{E}[Y(W = 1) - Y(W = 0)] = \mathbb{E}[\text{CATE}(\mathbf{X})].$$

4.2.1.1 Competing methods. We will compare the performance of the DRF with the following methods, specially designed for estimation of the CATE (or ATE)

- Double Machine Learning (DML) of [28], which assumes the model $Y = m(\mathbf{X}) + W\theta + \epsilon$ with constant treatment effect and can thus only be used for estimating ATE and not CATE.
- X-learner (XL) introduced in [96] (the version with RF learners)
- Causal Forest (CF) introduced in [181, 4] (we use the GRF version [4] with local centering that substantially improves on the version in [181])

In order to make the comparison fair, we use the local centering approach for DRF as well.

4.2.1.2 Data. We will use the following data models for our simulations, where the first three are taken directly from [4]:

1. In this model X_3 is a confounder affecting both W and Y:

$$\mathbf{X} \sim U(0,1)^p, \quad W \mid \mathbf{X} \sim \text{Bernoulli}\left(\frac{1}{4}(1+\beta_{2,4}(X_3))\right),$$
$$Y \mid \mathbf{X}, W \sim 2\left(X_3 - \frac{1}{2}\right) + N(0,1),$$

   where $\beta_a, b(x)$ is the density of the beta random variable with parameters a and b.

2. In this model the treatment effect is heterogeneous, i.e. how W affects Y changes with X_1 and X_2:

$$\mathbf{X} \sim U(0,1)^p, \quad W \mid \mathbf{X} \sim \text{Bernoulli}(0.5),$$
$$Y \mid \mathbf{X}, W \sim \left(W - \frac{1}{2}\right)\eta(X_1)\eta(X_2) + N(0,1),$$

   where $\eta(x) = 1 + \left(1 + e^{-20(x-\frac{1}{3})}\right)^{-1}$.

3. This model is a combination of the previous two, so the treatment effect is heterogeneous and we have confounding:

$$\mathbf{X} \sim U(0,1)^p, \quad W \mid \mathbf{X} \sim \text{Bernoulli}\left(\frac{1}{4}(1+\beta_{2,4}(X_3))\right),$$
$$Y \mid \mathbf{X}, W \sim 2\left(X_3 - \frac{1}{2}\right) + \left(W - \frac{1}{2}\right)\eta(X_1)\eta(X_2) + N(0,1).$$

4. The following model is similar to above, with slightly different structure, where X_2 induces the confounding effects and X_1 makes the treatment heterogeneous:

$$\mathbf{X} \sim U(0,1)^p, \quad W \mid \mathbf{X} \sim \text{Bernoulli}\left(\text{expit}(4X_2 - 2)\right),$$
$$Y \mid \mathbf{X}, W \sim 100X_2^2 + \left(W - \frac{1}{2}\right)\sin(3X_1) + N(0,1).$$

4.2.1.3 Results. For every model we generate n data points $(X_1, \ldots, X_p, W, Y)_{i=1,\ldots,n}$. We run all methods and compute the root mean squared error of the obtained CATE estimate on a randomly generated test set $\mathbf{X}_{\text{test}}$ containing 1000 data points. CATE corresponds to the coefficient of W in the data generating mechanism of Y. We repeat the same procedure 100 times and report the average result. For methods other than the DML, we estimate ATE by averaging the CATE estimates over the randomly generated test set. The results can be seen in Table 4 and Figure 13. Even though DRF is performing less well in general compared to the methods that are specially designed for the task of estimating CATE, we can still see that its estimates are fairly good.

Table 4: RMSE for the CATE, averaged over 1000 test points and 100 overall repetitions.

model	n	p	**DRF**	**CF**	**XL**
1	800	10	0.140	**0.109**	0.149
1	1600	10	0.119	**0.085**	0.122
1	800	20	0.125	**0.094**	0.128
1	1600	20	0.105	**0.076**	0.107
2	800	10	0.452	0.319	**0.288**
2	1600	10	0.285	0.234	**0.228**
2	800	20	0.568	0.336	**0.306**
2	1600	20	0.341	0.254	**0.241**
3	800	10	0.621	0.328	**0.319**
3	1600	10	0.453	0.243	**0.237**
3	800	20	0.708	**0.343**	0.346
3	1600	20	0.533	0.257	**0.256**
4	800	10	0.320	**0.273**	0.682
4	1600	10	0.285	**0.228**	0.389
4	800	20	0.316	**0.289**	0.722
4	1600	20	0.291	**0.248**	0.412

Table 5: RMSE for the ATE, averaged over 100 repetitions.

model	n	p	**DRF**	**CF**	**XL**	**DML**
1	800	10	0.0841	**0.0806**	0.0858	0.0843
1	1600	10	0.0526	**0.0517**	0.0523	0.0538
1	800	20	0.0786	**0.0762**	0.0802	0.0785
1	1600	20	**0.0585**	0.0588	0.0609	0.0625
2	800	10	**0.0844**	0.0880	0.0877	0.0891
2	1600	10	**0.0567**	0.0587	0.0593	0.0584
2	800	20	0.0783	**0.0767**	0.0788	0.0872
2	1600	20	**0.0645**	0.0657	0.0665	0.0656
3	800	10	**0.0914**	0.0916	0.0932	0.1116
3	1600	10	**0.0573**	0.0581	0.0599	0.0778
3	800	20	**0.0858**	0.0917	0.0922	0.1114
3	1600	20	**0.0623**	0.0673	0.0599	0.0925
4	800	10	**0.1061**	0.1075	0.2554	0.9734
4	1600	10	**0.0677**	0.0665	0.1028	0.5542
4	800	20	**0.1008**	0.1046	0.2308	2.7805
4	1600	20	**0.0655**	0.0660	0.0947	1.834

4.2.2 Continuous W, linear treatment effect

When the treatment variable W is continuous, many methods designed for binary treatments, such as the X-learner [96] cannot be used. However, many important real-word examples fall within this framework. As an example, we might be interested in how the amount of medicine W affects some biological parameter of interest Y (conditionally on $\mathbf{X}$). When W affects Y linearly conditionally on $\mathbf{X}$, one can still use the Causal Forest (CF) [4, 181] method, which makes the splits based on the slope of the conditional linear fit $Y \sim W$. Due to its generality and versatility, DRF can trivially be used in such setting as well.

To illustrate this, we consider the Model 3, as described in the previous section, which is also taken from [4], but where we change the distribution of the binary treatment variable so that it is continuous and it has a normal distribution with the same mean and variance, which

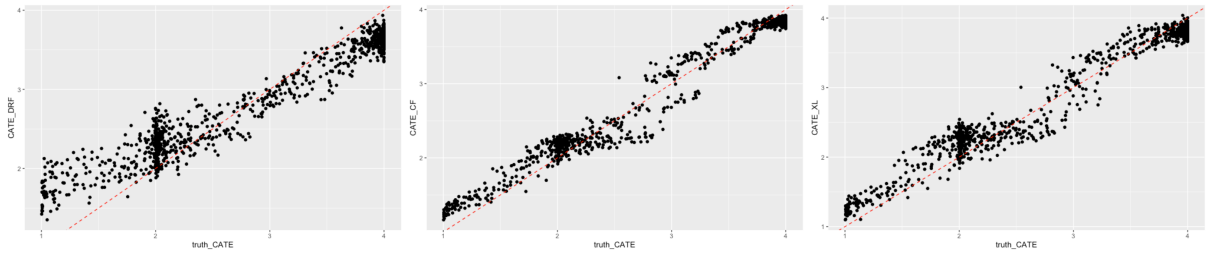


Figure 13: Estimates of the CATE for DRF (left), Causal Forest (middle), and X-learner (right) plotted against the true cate on the x-axis for Model 3 with $n = 1600$, $p = 20$.

depend on $\mathbf{X}$. In this model W affects Y linearly, which is a crucial assumption for the CF approach to work. We take $n = 10000$ and $p = 10$. The concept of CATE does not exist in this form in such setup and therefore we consider how the forest obtained by each method estimates both the intercept and the slope of the fit $Y \sim W$, conditionally on $\mathbf{X}$. The results can be seen in Figure 14. We see that the estimate of the slope for DRF is slightly worse than for the CF, whose forest construction is specially designed for estimating the conditional slope. However, DRF estimates the intercept significantly better than the CF, especially in combination the local centering approach, which uses the centered data $Y - \widehat{Y}(\mathbf{X})$ and $W - \widehat{W}(\mathbf{X})$ instead. In this example the slope depends only on X_2, whereas the intercept depends on both X_1 and X_2. Since CF targets only the slope for forest construction, it will split mostly on X_2 and not on X_1, which leads to poor estimate of the intercept term. On the other hand, DRF splits both on X_1 and X_2, depending on the size of their effect on the joint distribution of (W, Y). For many applications, especially in causality (see also the example in next section), it is essential to know the whole conditional distribution $\mathbb{P}(Y \mid W, \mathbf{X})$ so the DRF approach might be more beneficial than the CF.

4.2.3 Nonlinear treatment effect

There are very few methods that can estimate the treatment effect when the treatment variable W is continuous and affects Y nonlinearly, as it is commonly the case in real world settings. Here we demonstrate that DRF can be easily used in such setup as well, as opposed to the CF, which assumes a linear, though heterogeneous, treatment effect of W on Y. We further show how the obtained regression fits $Y \sim W$ conditional on $\mathbf{X}$ can be used to estimate the causal effect $\mathbb{E}[Y \mid do(W = w)]$, as it is done in the main paper in the birth data example:

$$\mathbb{E}[Y \mid do(W = w)] = \int \mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]\mathbb{P}(\mathbf{X} = \mathbf{x})d\mathbf{x}.$$

We compare the performance of DRF with the straightforward and commonly used approach, where we first regress Y on $(W, \mathbf{X})$ and use this regression fit which estimates $\mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]$ together with the above formula to estimate the causal effect.

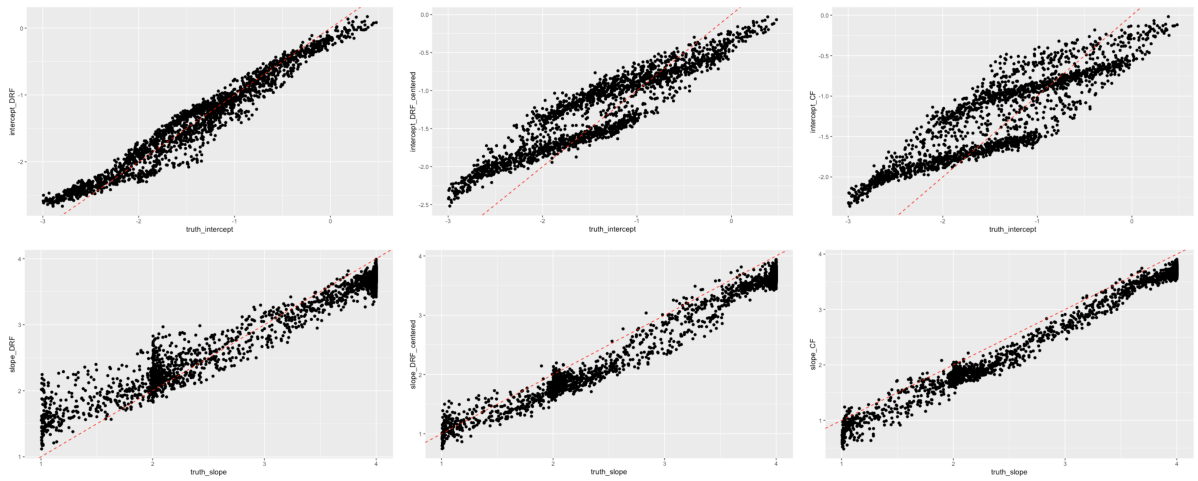


Figure 14: The estimates of the intercept (top row) and the slope (bottom row) for the linear fit conditional on $\mathbf{X}$, against the true values on the x-axis, obtained for the plain DRF (left), DRF with local centering as in [4] (middle) and CF (right), which uses the local centering approach. The data is generated from Model 3 described in the previous section, with continuous treatment variable W with the same mean and variance conditionally on $\mathbf{X}$.

4.2.3.1 Data We consider the following example, similar to the previous examples:

$$\mathbf{X} \sim U(0,1)^p, \quad W \mid \mathbf{X} \sim \frac{1}{2}\left|1 + 4X_3 + N(0,1)\right|,$$
$$Y \mid \mathbf{X}, W \sim 3\left(X_3 - \frac{1}{2}\right) + 3X_1 \sin(3W) + X_2 N(0,1).$$

Therefore, W affects Y highly nonlinearly through a sine function. X_3 is a confounding variable that affects the marginal distributions of Y and W. X_2 regulates the error level for Y, whereas X_1 makes the treatment effect heterogeneous. This is illustrated in the following plot:

4.2.3.2 Results In Figure 16 we can see the estimated joint distribution of (Y, W) conditionally on $\mathbf{X} = \mathbf{x}$, where the values of X_1 and X_3 vary, while the rest are fixed (even though X_2 also affects the conditional distribution, the effect is much weaker than for X_1 of X_3, see Figure 16). We see that the estimated distribution matches the true regression line, denoted in red, very well. The estimated distribution induced by the DRF weights enables us to fit some specialised regression method for regressing Y on W for every fixed value of $\mathbf{X}$. The blue line indicates the fit obtained by using smoothing splines. Compared to the green line, which shows the predicted values for regression $Y \sim (W, \mathbf{X})$, it is nicer looking and also is able to extrapolate much better to values of W which have low probability conditionally on X.

This extrapolation is crucial for causal applications, since for computing $\mathbb{E}[Y \mid do(W = w)]$ we are interested in what would happen with Y when our treatment variable W is fixed to be w, regardless of which values are achieved by $\mathbf{X}$. However, it can easily happen that for this specific combination of $\mathbf{X}$ and W there are very few observed data points, which makes the

estimation hard [133]. In this example, W tends to be small for small values of X_3 and vice-versa and thus is hard to say what would happen with Y when X_3 is large and W is set to a small value by an outside intervention.

In Figure 17, we indeed see that the estimates of the causal effect $\mathbb{E}[Y \mid do(W = w)]$ by DRF are much better. One can still see that the error increases for the border values of W, which have small probability for some values of X, since the estimation there is much harder, but this error is much less pronounced for DRF than for the standard regression approach.

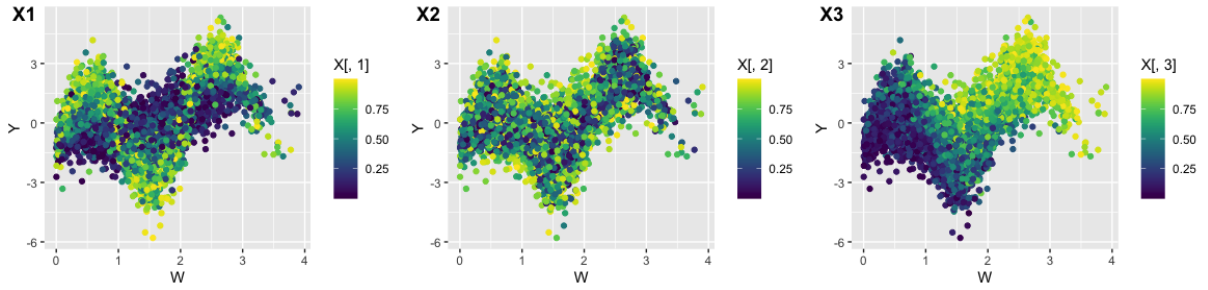


Figure 15: Visualization how X_1, X_2, X_3 affect the conditional nonlinear regression fit $Y \sim W$. X_1 changes the effect size, X_2 changes the noise level, whereas X_3 is a confounding variable which affects the means of W and Y.

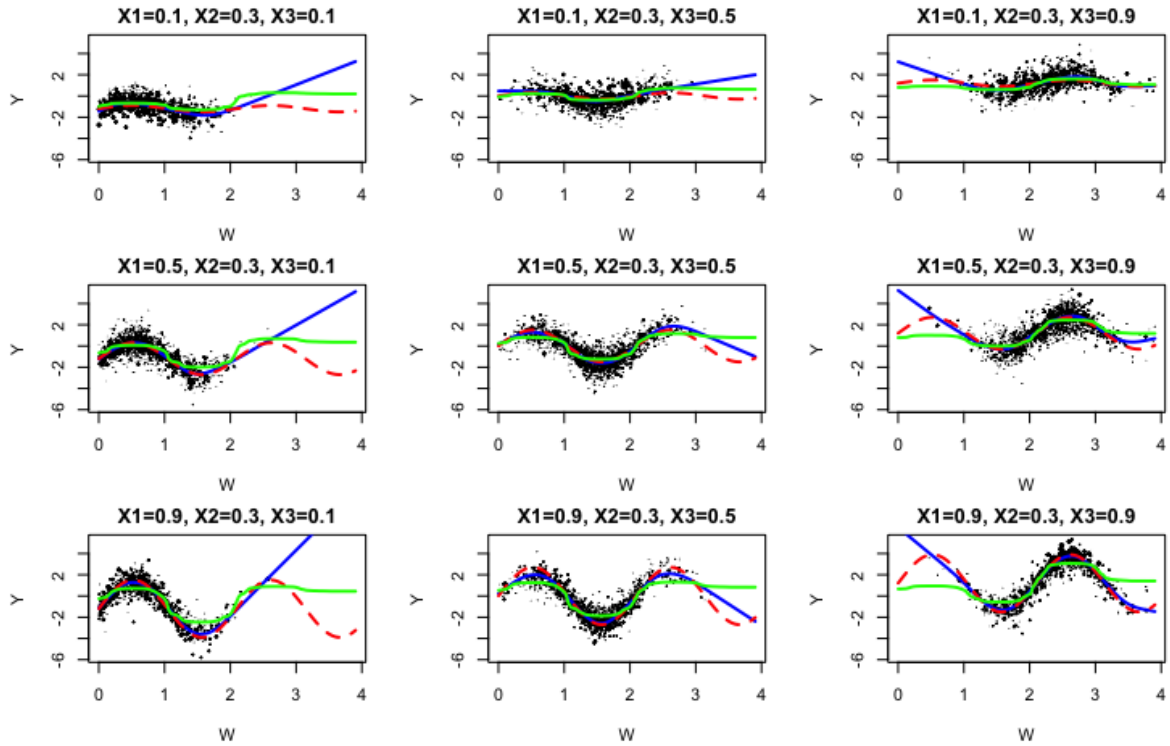


Figure 16: For a grid of values for the test point $\mathbf{x}$, the scatterplot illustrates the estimated joint distribution (Y, W) by DRF. The subsequent regression fit using smoothing splines is denoted in blue, whereas the true conditional mean $\mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]$ is denoted with red dashed line. Green line shows the estimate of the conditional mean $\mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]$ with plain random forest.

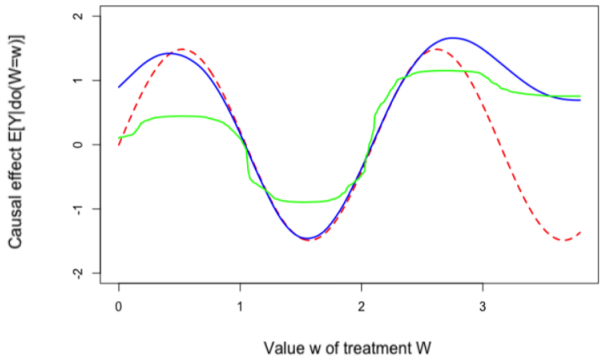


Figure 17: The estimated causal effect $\mathbb{E}[Y \mid do(W = w)]$ with DRF (blue) and with conventional method which regresses Y on $(W, \mathbf{X})$ using plain Random Forest (green). The true value is denoted by a red dashed line.

Bibliography

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

[2] Karan Aggarwal, Matthieu Kirchmeyer, Pranjul Yadav, Sathiya S. Keerthi, and Patrick Gallinari. Benchmarking regression methods: A comparison with CGAN. *arXiv preprint arXiv:1905.12868*, 2019.

[3] Mohamad Akra and Louay Bazzi. On the solution of linear recurrence equations. *Computational Optimization and Applications*, 10(2):195–210, 1998.

[4] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized Random Forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019.

[5] Vincent Audigier, François Husson, and Julie Josse. Multiple imputation for continuous variables using a bayesian principal component analysis. *Journal of Statistical Computation and Simulation*, 86(11):2140–2156, 2016.

[6] Zhidong Bai and Hewa Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, pages 311–329, 1996.

[7] Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

[8] Cédric Beaulac and Jeffrey S. Rosenthal. BEST: a decision tree algorithm that handles missing values. *Computational Statistics*, 35(3):1001–1026, September 2020.

[9] Patrizia Berti, Luca Pratelli, and Pietro Rigo. Almost sure weak convergence of random probability measures. *Stochastics*, 78(2):91–97, 2006.

[10] Gérard Biau. Analysis of a Random Forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

[11] Gerard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 09 2008.

[12] Peter J. Bickel, Chris AJ Klaassen, Yaâacov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

[13] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973â3009, December 2010.

[14] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41 – 65, 2019.

[15] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[16] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[17] Leo Breiman. Manual on setting up, using, and understanding Random Forests. Technical report, Berkeley CA, 2003.

[18] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Wadsworth and Brooks, 1984.

[19] Leo Breiman, Joseph H Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. 1984.

[20] Boris E. Brodsky and Boris S. Darkhovsky. *Nonparametric Methods in Change Point Problems*, volume 243. Springer Science & Business Media, 2013.

[21] Samantha M. Brown, Jenalee R. Doom, Stephanie Lechuga-PeÃ±a, *Sarah Enos Watamura, and Tiffany Koppels. Stress and parenting during the global covid-19 pandemic. Child Abuse & Neglect*, 110 : 104699, 2020. *Protecting children from maltreatment during COVID-19 : First volume*.

[22] Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

[23] Stef Buuren and Catharina Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45, 12 2011.

[24] Haiyan Cai, Bryan Goggin, and Qingtang Jiang. Two-sample test based on classification probability. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(1):5–13, 2020.

[25] Nora E. Charles, Stephanie J. Strong, Lauren C. Burns, Margaret R. Bullerjahn, and Katherine M. Serafine. Increased mood disorder symptoms, perceived stress, and alcohol use among college students during the covid-19 pandemic. *Psychiatry Research*, 296:113706, 2021.

[26] HY Chen and R Little. A test of missing completely at random for generalised estimating equations with missing data. *Biometrika*, 86(1):1–13, 03 1999.

[27] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 16, pages 785–794, New York, NY, USA, 2016.

[28] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[29] Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

[30] Umberto Cherubini, Elisa Luciano, and Walter Vecchiato. *Copula Methods in Finance*. John Wiley & Sons, 2004.

[31] Jungyeon Choi, O. Dekkers, and S. le Cessie. A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology*, 34:23–36, 2018.

[32] Kacper P. Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, volume 28, pages 1981–1989. 2015.

[33] William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[34] Panayiota Constantinou and Alexander P. Dawid. Extended conditional independence and applications in causal inference. *The Annals of Statistics*, 45(6):2618 – 2653, 2017.

[35] I. Csiszar. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

[36] Yuval Peres David A. Levin and Elizabeth L. Wilmer. Lecture notes on markov chains and mixing times, March 2019.

[37] Jan Alexander de Vos, Mirjam Radstaak, Ernst T. Bohlmeijer, and Gerben J. Westerhof. Modelling trajectories of change in psychopathology and well-being during eating disorder outpatient treatment. *Psychotherapy Research*, 0(0):1–13, 2022.

[38] Yi Deng, Changgee Chang, Moges Seyoum Ido, and Qi Long. Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific reports*, 6:21689–21689, Feb 2016.

[39] Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In *31st International Conference on Machine Learning (ICML)*, volume 32 of *ICML'14*, pages 665–673, 2014.

[40] Luc Devroye, Laszlo Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[41] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians, 2018.

[42] Yaohui Ding and Arun Ross. A comparison of imputation methods for handling missing scores in biometric fusion. *Pattern Recognition*, 45(3):919–933, 2012.

[43] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 06 2004.

[44] David Donoho and Jiashun Jin. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, 30(1):1–25, 02 2015.

[45] L.L. Doove, S. Van Buuren, and E. Dusseldorp. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72(C):92–104, 2014.

[46] Richard M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002.

[47] Eugene B. Dynkin and Avishai Mandelbaum. Symmetric statistics, Poisson point processes, and multiple Wiener integrals. *The Annals of Statistics*, 11(3):739 – 745, 1983.

[48] Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 05 1981.

[49] Jan Ernest and Peter Bühlmann. Marginal integration for nonparametric causal inference. *Electronic Journal of Statistics*, 9(2):3155 – 3194, 2015.

[50] Guangzhe Fan, Zhou Wang, and Jiheng Wang. Cw-ssim kernel based random forest for image classification. In *Visual Communications and Image Processing 2010*, volume 7744, page 774425. International Society for Optics and Photonics, 2010.

[51] Jianqing Fan, Mark Farmen, and Irene Gijbels. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):591–608, 1998.

[52] Helmut Finner and Veronika Gontscharuk. Two-sample Kolmogorov-Smirnov-type tests revisited: Old and new tests in terms of local levels. *The Annals of Statistics*, 46(6A):3014–3037, 12 2018.

[53] Jerome H. Friedman. On multivariate goodness-of-fit and two-sample testing. 2004.

[54] Jerome H. Friedman and Lawrence C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

[55] Philipp Gaffert, Florian Meinfelder, and V. Bosch. Towards an MI-proper predictive mean matching. 2016.

[56] Johann Gagnon-Bartsch and Yotam Shem-Tov. The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *The Annals of Applied Statistics*, 13(3):1464–1483, 09 2019.

[57] Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[58] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[59] Tilmann Gneiting, Larissa I. Stanberry, Eric P. Grimit, Leonhard Held, and Nicholas A. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):211–235, Aug 2008.

[60] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep Learning*, volume 1. MIT press Cambridge, 2016.

[61] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672â2680, Cambridge, MA, USA, 2014. MIT Press.

[62] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19, pages 513–520, 2007.

[63] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[64] Arthur Gretton, Kenji Fukumizu, Choon H. Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, page 585â592, 2007.

[65] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, volume 25, pages 1205–1213. 2012.

[66] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, volume 25, pages 1205–1213, 2012.

[67] Mariah T. Hawes, Aline K. Szenczy, Daniel N. Klein, Greg Hajcak, and Brady D. Nelson. Increases in depression and anxiety symptoms in adolescents and young adults during the COVID-19 pandemic. *Psychological Medicine*, page 1â9, 2021.

[68] Simon Hediger, Loris Michel, and Jeffrey Näf. On the use of random forest for two-sample testing. *Computational Statistics & Data Analysis*, 170:107435, 2022.

[69] Simon Hediger, Loris Michel, and Jeffrey Näf. On the use of random forest for two-sample testing. *Computational Statistics & Data Analysis*, 170:107435, 2022.

[70] Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *TEST*, 27(4):811–825, Dec 2018.

[71] Wassily Hoeffding. The strong law of large numbers for U-statistics. Mimeograph Series 302, Institute of Statistics, University of Noth Carolina, 1961.

[72] Nicholas Horton and Stuart Lipsitz. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, 55:244–254, 09 2001.

[73] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

[74] Torsten Hothorn and Achim Zeileis. Predictive distribution modeling using transformation forests. *Journal of Computational and Graphical Statistics*, 0(0):1–16, 2021.

[75] Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 2015.

[76] Guido W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

[77] Hemant Ishwaran and Udaya B. Kogalur. Package 'randomForestSRC', 2022.

[78] Mortaza Jamshidian and Siavash Jalal. Tests of homoscedasticity, normality, and missing completely atÂ random for incomplete multivariate data. *Psychometrika*, 75(4):649–674, Dec 2010.

[79] Mortaza Jamshidian, Siavash Jalal, and Camden Jansen. MissMech: An R package for testing homoscedasticity, multivariate normality, and missing completely at random (mcar). *Journal of Statistical Software*, 56(6):1–31, 2014.

[80] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the L_1 distance. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 750–754, July 2016.

[81] Wittawat Jitkrittum, Zoltán Szabó, Kacper P. Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, volume 29, pages 181–189. 2016.

[82] Jonathan Johannemann, Vitor Hadad, Susan Athey, and Stefan Wager. Sufficient representations for categorical variables. *arXiv preprint arXiv:1908.09874*, 2019.

[83] Julie Josse and François Husson. missMDA: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

[84] Julie Josse, Jérôme Pagès, and Francois Husson. Multiple imputation in principal component analysis. *Adv. Data Analysis and Classification*, 5:231–246, 10 2011.

[85] Edward H. Kennedy, Zongming Ma, Matthew D. McHugh, and Dylan S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(4):1229, 2017.

[86] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, volume 31, pages 656–666, 2017.

[87] Ilmun Kim, Ann B. Lee, and Jing Lei. Global and local two-sample tests via regression. *Electronic Journal of Statistics*, 13(2):5253–5305, 2019.

[88] Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411 – 434, 2021.

[89] Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411 – 434, 2021.

[90] Kevin H. Kim and Peter M. Bentler. Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, 67(4):609–623, Dec 2002.

[91] Durk P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, volume 29, page 4743â4751, 2016.

[92] Dragi Kocev, Celine Vens, Jan Struyf, and Sašo Džeroski. Ensembles of multi-objective decision trees. In *European Conference on Machine Learning*, pages 624–631. Springer, 2007.

[93] Roger Koenker, Stephen Portnoy, Pin T. Ng, Achim Zeileis, Philip Grosjean, and Brian D. Ripley. Package quantreg, 2012.

[94] M.R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer New York, 2008.

[95] Egor Kosov. Total variation distance estimates via L^2-norm for polynomials in log-concave random vectors. 2018.

[96] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[97] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076, 2017.

[98] Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood: Approximating kernel expansions in log-linear time. In *30th International Conference on Machine Learning (ICML)*, volume 28 of *ICML'13*, page 244â252, 2013.

[99] E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[100] Friedrich Leisch and Evgenia Dimitriadou. *mlbench: Machine learning benchmark problems*, 2021. R package version 2.1-3.

[101] Jun Li and Yao Yu. A nonparametric test of missing completely at random for incomplete multivariate data. *Psychometrika*, 80(3):707–726, Sep 2015.

[102] Shuang Li, Yao Xie, Hanjun Dai, and Le Song. Scan *B*-statistic for kernel change-point detection. *Sequential Analysis*, 38(4):503–544, 2019.

[103] W. Lin and Chih-Fong Tsai. Missing value imputation: a review and analysis of the literature (2006â2017). *Artificial Intelligence Review*, 53:1487–1509, 2019.

[104] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors (technical report no. 1055). *University of Wisconsin*, 2002.

[105] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

[106] Roderick J. A. Little. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, (6):287–296, 1988.

[107] Roderick J. A. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.

[108] Roderick J A Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., USA, 1986.

[109] David Lopez-Paz and Maxime Oquab. Revisiting Classifier Two-Sample Tests. 2017.

[110] Julián Luengo, Salvador GarcÃa, and Francisco Herrera. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems - KAIS*, 32:1–32, 07 2011.

[111] James Malley, J Kruppa, Abhijit Dasgupta, Karen Malley, and Andreas Ziegler. Probability machines consistent probability estimation using nonparametric learning machines. *Methods of information in medicine*, 51:74–81, 09 2011.

[112] Frank J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[113] David S. Matteson and Nicholas A. James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

[114] Imke Mayer, Aude Sportisse, Julie Josse, Nicholas Tierney, and Nathalie Vialaneix. R-miss-tastic: a unified platform for missing values methods and workflows, 2021.

[115] Lachlan McCalman, Simon T. O'Callaghan, and Fabio Ramos. Multi-modal estimation with kernel embeddings for learning motion models. In *IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, May 6-10, 2013*, pages 2845–2852, 2013.

[116] Nicolai Meinshausen. False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics*, 33(2):227–237, 2006.

[117] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

[118] Nicolai Meinshausen. Node harvest. *The Annals of Applied Statistics*, 4(4):2049â2072, 2010.

[119] Nicolai Meinshausen and Peter Bühlmann. Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika*, 92(4):893–907, 12 2005.

[120] Nicolai Meinshausen and John Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics*, 34(1):373–393, 02 2006.

[121] Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558, 1994.

[122] Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(1):841–881, January 2016.

[123] Loris Michel, Jeffrey Näf, Meta-Lina Spohn, and Nicolai Meinshausen. PKLM: A flexible MCAR test using classification. *arXiv preprint arXiv:2109.10150*, 2021.

[124] Loris Michel, Jeffrey Näf, Meta-Lina Spohn, and Nicolai Meinshausen. Proper scoring rules for missing value imputation, 2021.

[125] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[126] Krikamol Muandet, Kenji Fukumizu, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

[127] Jared S. Murray and Jerome P. Reiter. Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479, 2016.

[128] Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 7130–7140. PMLR, 2020.

[129] Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[130] Frank Nielsen and Ke Sun. Guaranteed deterministic bounds on the total variation distance between univariate mixtures. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Sep. 2018.

[131] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30, page 2335â2344, 2017.

[132] Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[133] Judea Pearl. *Causality*. Cambridge university press, 2009.

[134] Gilles Pisier. *Martingales in Banach Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016.

[135] Drago Plečko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44, 2020.

[136] Taylor Pospisil and Ann B. Lee. Rfcde: Random forests for conditional density estimation. *arXiv preprint arXiv:1804.05753*, 2018.

[137] J. Ross Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., 1993.

[138] Joaquin Quinonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. Evaluating predictive uncertainty challenge. In *Machine Learning Challenges Workshop*, pages 1–27. Springer, 2005.

[139] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pages 1177–1184, 2008.

[140] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*, volume 21, pages 1313–1320, 2009.

[141] Mehdi Rajeb, Yurou Wang, Kaiwen Man, and Laura M. Morett. Students' acceptance of online learning in developing nations: scale development and validation. *Educational technology research and development*, Nov 2022.

[142] Aaditya Ramdas, Sashank J. Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. AAAI'15, page 3571â3577, 2015.

[143] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2), 2017.

[144] Christopher H. Rhoads. Problems with tests of the missingness mechanism in quantitative policy studies. *Statistics, Politics, and Policy*, 3(1), 2012.

[145] Jonathan Rosenblatt, Roee Gilron, and Roy Mukamel. Better-than-chance classification for signal detection. *Biostatistics (Oxford, England)*, 08 2016.

[146] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[147] Donald B. Rubin. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, (4):87–94, 1986.

[148] Maytal Saar-Tsechansky and Foster Provost. Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8:1623â1657, 2007.

[149] Igal Sason and Sergio Verdu. Upper bounds on the relative entropy and rényi divergence as a function of total variation distance for finite alphabets, 2015.

[150] Rianne Margaretha Schouten, Peter Lugtig, and Gerko Vink. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930, 2018.

[151] Shaun Seaman, John Galati, Dan Jackson, and John Carlin. What is meant by "missing at random" ? *Statistical Science*, 28(2):257–268, 2013.

[152] Mark Segal and Yuanyuan Xiao. Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):80–87, 2011.

[153] Jun Shao and C. F. J. Wu. A general theory for jackknife variance estimation. *The Annals of Statistics*, 17(3):1176–1197, 1989.

[154] Tao Shi and Steve Horvath. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1):118–138, 2006.

[155] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC press, 1986.

[156] Carl-Johann Simon-Gabriel, Alessandro Barp, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *arXiv preprint arXiv:2006.09268*, 2020.

[157] Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *The Journal of Machine Learning Research*, 19(1):1708–1736, 2018.

[158] Sebastian Sippel, Nicolai Meinshausen, Erich Fischer, Eniko Székely, and Reto Knutti. Climate change now detectable from any single day of global weather. *Nature Climate Change*, 10:35–41, 01 2020.

[159] Abe Sklar. *Fonctions de Répartition À N Dimensions Et Leurs Marges*. Université Paris 8, 1959.

[160] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 28, pages 3483–3491, 2015.

[161] Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

[162] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *26th Annual International Conference on Machine Learning (ICML)*, ICML '09, page 961â968, 2009.

[163] Bharath K. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839 – 1893, 2016.

[164] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[165] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, April 2010.

[166] Daniel J. Stekhoven. *missForest: Nonparametric Missing Value Imputation using Random Forest*, 2013. R package version 1.4.

[167] Daniel J. Stekhoven and Peter Bühlmann. MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 10 2011.

[168] Xiaoyue Sun and Mengtong Chen. Associations between perceived material deprivation, social support and violent victimization among chinese children. *Child Abuse & Neglect*, 127:105583, 2022.

[169] Gábor J. Székely and Maria L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

[170] Thordis L. Thorarinsdottir, Tilmann Gneiting, and Nadine Gissibl. Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):522–534, 2013.

[171] Nicholas J Tierney and Dianne H Cook. Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations, 2020.

[172] Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. MULAN: A Java Library for Multi-Label Learning. *Journal of Machine Learning Research*, 12(71):2411–2414, 2011.

[173] Hisaharu Umegaki and Albert T. Bharucha-Reid. Banach space-valued random variables and tensor products of Banach spaces. *Journal of Mathematical Analysis and Applications*, 31(1):49–67, 1970.

[174] Paul Valiant and Gregory Valiant. Estimating the unseen: Improved estimators for entropy and other properties. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2157–2165. Curran Associates, Inc., 2013.

[175] S. van Buuren. *Flexible Imputation of Missing Data. Second Edition*. Chapman & Hall/CRC Press, Boca Raton, FL, 2018.

[176] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

[177] Aad van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[178] Richard A. Vitale. Covariances of symmetric statistics. *Journal of Multivariate Analysis*, 41(1):14–26, 1992.

[179] Stefan Wager. Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*, 2014.

[180] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*, 2017.

[181] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[182] Akbar K. Waljee, Ashin Mukherjee, Amit G. Singal, Yiwei Zhang, Jeffrey Warren, Ulysses Balis, Jorge Marrero, Ji Zhu, and Peter Dr Higgins. Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, 3(8):e002847, Aug 2013.

[183] Bernard L. Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[184] Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.

[185] Frank Wilcoxon. Individual comparisons of grouped data by ranking methods. *Journal of Economic Entomology*, 39(2):269–270, 1946.

[186] Douglas A. Wolfe and Edna Schechtman. Nonparametric statistical procedures for the change-point problem. *Journal of Statistical Planning and Inference*, 9(3):389–396, 1984.

[187] Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1â17, 2017.

[188] Xueying Xu, Leizhen Xia, Qimeng Zhang, Shaoning Wu, Mingcheng Wu, and Hongbo Liu. The ability of different imputation methods for missing values in mental measurement questionnaires. *BMC Medical Research Methodology*, 20(42):1–16, 02 2020.

[189] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5689–5698. PMLR, 10–15 Jul 2018.

[190] Ke-Hai Yuan, Mortaza Jamshidian, and Yutaka Kano. Missing data mechanisms and homogeneity of means and variances–covariances. *Psychometrika*, 83(2):425–442, Jun 2018.

[191] Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 26, pages 755–763, 2013.

[192] Achim Zeileis, Torsten Hothorn, and Kurt Hornik. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514, 2008.

[193] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence*, page 804â813, 2011.

[194] Shixiao Zhang, Peisong Han, and Changbao Wu. A unified empirical likelihood approach for testing mcar and subsequent estimation. *Scandinavian Journal of Statistics*, 46(1):272–288, 2019.

[195] Ji Zhao and Deyu Meng. Fastmmd: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6):1345–1372, 2015.

JEFFREY NÄF

Researcher - Statistician

CONTACT

jeffrey@my-sh.ch

Linkedin

Medium

Orcid

Google Scholar

SKILLS

Python 1+ yrs

R 6+ yrs

Matlab 6+ yrs

TEACHING

Teaching Assistant 2018-21
Fundamentals of Mathematical Statistics

ETH Zürich

Lecturer 2017/18
Fundamental Probability for Finance

University of Zurich

Teaching Assistant 2012-14
Introductory Econometrics

University of Zurich

AWARDS

Rigour and Relevance Research Award 2022
Swiss Academy of Marketing Science

Prize for outstanding Master Thesis 2015
University of Zurich

RESEARCH INTERESTS

High-dimensional statistics, multivariate modeling, mixed models, kernel methods, nonparametric statistics, CLV modelling, testing

EDUCATION

PhD in Statistics, ETH Zurich 2018 - 2023
Supervisor: Nicolai Meinshausen

My thesis is centered around Random Forest (RF) and various applications. In particular, we developed non-parametric two-sample tests and a measure of quality of imputation methods for missing data. We finally developed a natural generalization of the RF that allows for an efficient non-parametric estimation of the whole conditional distribution.

Master of Science in Statistics ETH Zurich 2015 - 2017
Final GPA – 5.85 out of 6

Master Thesis: *Review of Asymptotic Results in Empirical Process Theory*, with Prof. Dr. Sara van de Geer.

Master of Arts in Business Administration University of Zurich 2013 - 2018
Final GPA – 5.80 out of 6

Master Thesis: *Getting out of the COMFORT Zone: The MEXI Distribution for Asset Returns*, with Prof. Dr. Marc Paolella.

Bachelor of Arts in Economics, University of Zurich 2010 - 2014
Final GPA – 5.35 out of 6

Bachelor Thesis: *Re-evaluating Takahashi Korekiyo's Role in Japan's Recovery from the Great Depression*, with Prof. Dr. Mathias Hoffmann, Dr Alexander Rathke.

WORK EXPERIENCE

Doctoral Student and Group Coordinator of the Seminar for Statistics May 18 - Feb 23
ETH Zurich, Switzerland

Research combined with teaching and organizing the courses and exams at the Seminar for Statistics at ETHZ, as the Group Coordinator.

Research Assistant at the Chair of Empirical Finance Jun 15 - May 18
University of Zurich, Switzerland

Working on research projects, such as the development of new multivariate distributions for portfolio applications.

Research Assistant at the Chair of Marketing and Market Research Jun 14 - May 18
University of Zurich, Switzerland

Working on research projects such as the extension of the Pareto/NBD model to allow for time-varying covariates and implementation of the derivations in R.

OUTREACH

Author and Contributor to various Medium articles in Towards Data Science, such as ongoing

- DRF: A Random Forest for (almost) everything
- CLVTools: A powerful R package to evaluate your customers
- A/B testing with Random Forest
- I-Scores: How to choose the best method to fill in NAs in your data set
- Nonlinear Shrinkage: An Introduction

SOFTWARE

CLVTools R package ongoing

Part of the developer community of the CLVTools R Package for customer evaluation. The package contains efficiently implemented versions of some of the most important models for customer lifetime evaluation.

MISCELLANEOUS

First Dan (Black Belt) Shotokan Karate

Coffee lover with advanced Barista skills

Assistant at the Chair of Microeconomics/Industrial Organization Jun 14 - May 18
University of Zurich, Switzerland

Administrative tasks and programming in Mathematica and LaTeX.

PUBLICATIONS

Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression JMLR
Journal of Machine Learning Research, 2022

Authors: D. Cévid, L. Michel, J. Näf, N. Meinshausen, P. Bühlmann

R-NL: Fast and Robust Covariance Estimation for Elliptical Distributions in High Dimensions arXiv
arXiv preprint, 2022

Authors: S. Hediger, J. Näf, M. Wolf

Combining the MGHyp Distribution with Nonlinear Shrinkage in Modeling Financial Asset Returns SSRN
SSRN preprint, 2022

Authors: S. Hediger, J. Näf

Heterogeneous Tail Generalized Common Factor Modeling SSRN
SSRN preprint, 2021

Authors: S. Hediger, J. Näf, M. S. Paolella, P. Polak

On the Use of Random Forest for Two-Sample Testing CSDA
Computational Statistics and Data Analysis, 2021

Authors: S. Hediger, L. Michel, J. Näf

The Role of Time-Varying Contextual Factors in Latent Attrition Models for Customer Base Analysis MS
Marketing Science, 2021

Authors: P. Bachmann, M. Meierer, J. Näf

PKLM: A flexible MCAR test using Classification arXiv
Major Revision in Psychometrika

Authors: L. Michel, J. Näf, M. Spohn, N. Meinshausen

Imputation Scores AOAS
To appear in the Annals of Applied Statistics

Authors: J. Näf, M. Spohn,L. Michel, N. Meinshausen

High Probability Lower Bounds for the Total Variation Distance arXiv
arXiv preprint, 2020

Authors: L. Michel, J. Näf, N. Meinshausen

Heterogeneous tail generalized COMFORT modeling via Cholesky decomposition JMVA
Journal of Multivariate Analysis, 2019

Authors: J. Näf, M. S. Paolella, P. Polak

SERVICE

Reviewer
Annals of Statistics

TALKS & CONFERENCES

EcoSta — **June 2022**
5th International Conference on Econometrics and Statistics, Japan

Presenting the working paper "Shrinking in COMFORT"

Invited Seminar Talk — **Mai 2022**
University of Montpellier

Presenting the paper "Imputation Scores"

EcoSta — **June 2019**
3rd International Conference on Econometrics and Statistics, Taiwan

Presenting the paper "Heterogeneous tail generalized COMFORT modeling via Cholesky decomposition"

REFERENCES

Nicolai Meinshausen — ETH
Seminar for Statistics, ETH Zurich

meinshausen@stat.math.ethz.ch

Peter Bühlmann — ETH
Seminar for Statistics, ETH Zurich

buhlmann@stat.math.ethz.ch

Markus Meierer — UNIGE
Institute of Management

markus.meierer@unige.ch