


On the use of random forest for two-sample testing

Journal Article**Author(s):**

Hediger, Simon; Michel, Loris; [Näf, Jeffrey](#) 

Publication date:

2022-06

Permanent link:

<https://doi.org/10.3929/ethz-b-000530959>

Rights / license:

[Creative Commons Attribution 4.0 International](#)

Originally published in:

Computational Statistics & Data Analysis 170, <https://doi.org/10.1016/j.csda.2022.107435>



On the use of random forest for two-sample testing

Simon Hediger^b, Loris Michel^a, Jeffrey Näf^{a,*}

^a Seminar for Statistics, ETH Zürich, Switzerland

^b Department of Banking and Finance, University of Zurich, Switzerland



ARTICLE INFO

Article history:

Received 4 August 2020

Received in revised form 14 October 2021

Accepted 15 January 2022

Available online 24 January 2022

Keywords:

Random forest

Distribution testing

Classification

Kernel two-sample test

MMD

Total variation distance

U-statistics

ABSTRACT

Following the line of classification-based two-sample testing, tests based on the Random Forest classifier are proposed. The developed tests are easy to use, require almost no tuning, and are applicable for *any* distribution on \mathbb{R}^d . Furthermore, the built-in variable importance measure of the Random Forest gives potential insights into which variables make out the difference in distribution. An asymptotic power analysis for the proposed tests is conducted. Finally, two real-world applications illustrate the usefulness of the introduced methodology. To simplify the use of the method, the R-package “hypoRF” is provided.

© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Two-sample testing via classification methods is an old idea tracing back to the work of Friedman (2004). Generally speaking, one adapts the output of a classifier to construct a two-sample test. Let $\mathbf{X}_1, \dots, \mathbf{X}_{n_0}$ and $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_1}$ be a collection of \mathbb{R}^d -valued random vectors, such that $\mathbf{X}_i \stackrel{iid}{\sim} P$ and $\mathbf{Y}_i \stackrel{iid}{\sim} Q$, where P and Q are some Borel probability measure on \mathbb{R}^d . The goal is to test

$$H_0 : P = Q, \quad H_A : P \neq Q. \quad (1)$$

Given these iid samples of vectors, we define labels $\ell_i = 1$ for each \mathbf{X}_i and $\ell_i = 0$ for each \mathbf{Y}_i to obtain the data (\mathbf{Z}_j, ℓ_j) , $j = 1, \dots, N$, for $N = n_0 + n_1$, and $\mathbf{Z}_j = \mathbf{X}_i$ or $\mathbf{Z}_j = \mathbf{Y}_i$. On this data, we train a classifier $\hat{g} : \mathbb{R}^d \rightarrow \{0, 1\}$. If \hat{g} is able to “accurately” predict ℓ on some test samples, it is taken as evidence against H_0 . In this work, we assume the data is generated from a mixture distribution

$$\mathbf{Z}_j \stackrel{iid}{\sim} (1 - \pi)P + \pi Q,$$

such that $n_1 \sim \text{Bin}(\pi, N)$, where Bin denotes the Binomial distribution. While our exposition will be valid for general classifiers, we specifically target the use of the Random Forest (RF) classifier in this work. Random Forest is a powerful and flexible method developed by Breiman (2001), known to have a remarkably stable performance in applications (see e.g. the extensive work of Fernández-Delgado et al., 2014).

* Corresponding author. Address: ETH Zürich, HG G 10.1, Rämistrasse 101, 8092 Zürich.
E-mail address: jeffrey.naef@stat.math.ethz.ch (J. Näf).

This approach to testing was used in scientific applications, especially in the field of neuroscience. We refer to Kim et al. (2021) for an excellent overview of the literature. More recently, much additional work has been produced in this direction in the statistical literature, see e.g., Kim et al. (2021); Rosenblatt et al. (2021); Lopez-Paz and Oquab (2018); Borji (2019); Gagnon-Bartsch and Shem-Tov (2019); Kim et al. (2019); Cai et al. (2020). The closest relation to our work appears to be the recent work of Kim et al. (2021). Our first out-of-sample test in Section 2.1, though derived independently, is closely related to their test in Section 9.1. Moreover, Kim et al. (2021, Proposition 9.1) provide a consistency result for general classifiers under mild assumptions. We add to this discussion, by showing that under imbalance these assumptions nonetheless break down for the Bayes classifier, such that a test based on this classifier is not consistent. Kim et al. (2021) also provide a rule of thumb on when to use classification-based tests, as opposed to more fine-tuned statistical tests designed for a specific problem. We extend this discussion by adding a recommendation for when to use the RF-based test, as opposed to kernel-based tests, as for instance proposed in Gretton et al. (2012a), Gretton et al. (2012), Chwialkowski et al. (2015) and Jitkrittum et al. (2016). These tests are natural competitors to classification-based tests and our work indicates that:

1. If the differences between P , Q can be found in the marginal distributions, even sparsely so, the RF-based test tends to perform well. We demonstrate in Section 4.2 that the RF-based test succeeds in an example with marginal differences, which is difficult for kernel-based tests.
2. If the change is mostly found in the dependency structure, or copula, kernel tests like MMD may be preferable. As is demonstrated in Appendix B the RF-based test still has power, but less so than the kernel-based tests.

In addition, the Random Forest classifier brings two features to the two-sample testing problem: The out-of-bag (OOB) statistics and the variable importance measures. The former is used to increase sample efficiency, compared to a test based on a holdout sample, while the latter provides insights into the source of distributional differences.

Our work also shares similarities with Rosenblatt et al. (2021), Gagnon-Bartsch and Shem-Tov (2019) and Kim et al. (2019). The work of Gagnon-Bartsch and Shem-Tov (2019) focuses on the use of the in-sample classification error as a test statistic in the balanced case. Rosenblatt et al. (2021) focuses attention on the power of different classifier-based test statistics for specific alternatives. They also seem to be the first to propose the use of bootstrap-based classification tests. The work of Kim et al. (2019) presents a different approach based on regression and focuses on local testing, i.e. determining where the distributional difference appears.

The next two subsections list our contributions and demonstrate the advantages of our method with a small toy example. Section 2 introduces the two tests used, the first based on out-of-sample observations and the second on the OOB statistics. It closes with a theoretical insight into the consistency of classifier-based tests. Section 3 extends this theoretical insight into an asymptotic power analysis for a version of the OOB error-based test, using U-statistics theory. Finally, Section 4 discusses the role of the variable importance measure of the Random Forest and demonstrates the power of our tests with simulated as well as two real-world data sets.

1.1. Contributions

Our work differentiates itself from the existing literature in several aspects:

- The out-of-sample test based on the class-wise errors in Proposition 1, though similar to the one in Kim et al. (2021, Proposition 1), requires fewer assumptions to conserve the level asymptotically (though Kim et al. (2021) focus on a setting, where both the number of observations $N \rightarrow \infty$ as well as the dimension $d \rightarrow \infty$. In our work, d is assumed to be fixed).
- We show that no test based on the Bayes classifier is consistent for $\pi \neq 1/2$ in Lemma 1, but that a simple change in the classifier's "cutoff" restores consistency.
- We utilize the OOB error and variable importance measure in this context to both increase the power of the test and extract more meaning in practice. As shown in simulations, the increase in power with the OOB test is substantial.
- We analyze the asymptotic normality of an OOB error-based test statistic using U-statistics theory and use it to derive an expression for the approximate power of the test in Section 3.
- We provide empirical evidence in Section 4.2, and in Appendix B, that our test constitutes an important complementary method to powerful kernel-based tests, leading to improved performance in some traditionally difficult examples.
- Finally, we provide the R-package `hypoRF` available on CRAN, with an implementation of the method.

1.2. Motivational example

We consider a toy example to demonstrate the proposed methodology underlying the Random Forest classifier two-sample test. We choose P and Q to be five-dimensional multivariate Gaussian probability distributions. The covariance matrix of P is the identity and the distribution Q only differs from P in the last two components between which a positive correlation of 0.8 is imposed. The OOB statistics-based two-sample test correctly rejects with a p -value of 0.0099 (details are given in Section 2.2). Fig. 1 presents a visual summary of the test. The right plot displays the last two components of the sampled points. On the top left, the estimated means, by component and class, indicate that no distributional difference

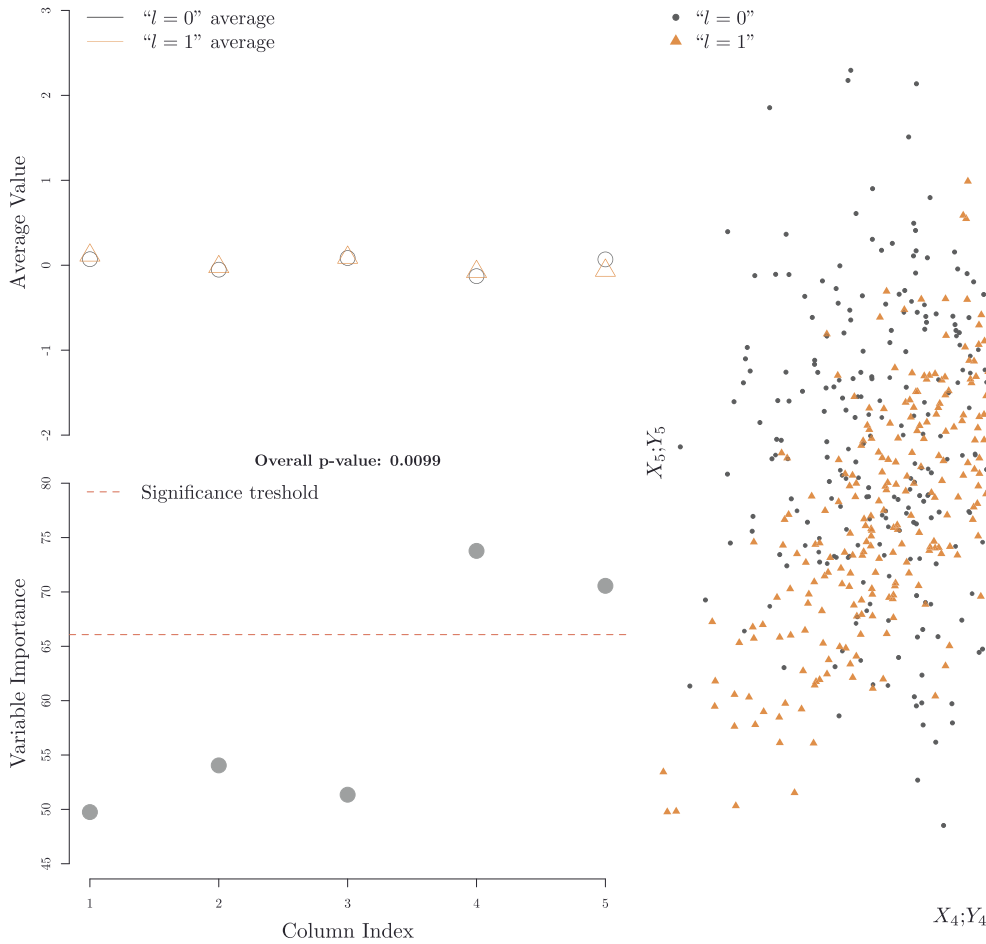


Fig. 1. (Intro) We sampled 300 observations from a $d = 5$ dimensional multivariate normal, with no correlation between the marginals. Likewise 300 observations were sampled from a multivariate normal, with the last two marginals having a correlation of 0.8. The Random Forest used 500 trees. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

is visible in the margins. The bottom left plot shows the variable importance measure for each component (as presented in Section 4.1). We can see that the last two components are picked-up as relevant variables, according to the threshold prescribed by the dotted red line.

Thus our method correctly rejects in this example and moreover delivers a hint as to which components might be responsible for the perceived difference in distribution.

2. Framework

Let $\mathbf{Z}_1, \dots, \mathbf{Z}_N$ be random vectors with values in $\mathcal{X} \subset \mathbb{R}^d$ and l_1, \dots, l_N corresponding labels in $\{0, 1\}$, collected in a dataset $D_N = \{(\mathbf{Z}_i, l_i)\}_{i=1}^N$ with

$$\mathbf{Z}_i \stackrel{iid}{\sim} (1 - \pi)P + \pi Q.$$

A sample \mathbf{Z}_i coming from the mixture component P (respectively Q) is labeled $l_i = 0$ (respectively $l_i = 1$). Let $\hat{g}(\mathbf{Z}) := g(\mathbf{Z}, D_{N_{train}})$ be a classifier trained on a subset $D_{N_{train}}$ of size $N_{train} < N$ of the observed data.

Given the setting above, we now present two tests based on the discriminative ability of \hat{g} . The first test uses an independent test set and is similar to the test proposed by Kim et al. (2021). The second test in Section 2.2 is entirely new and uses the OOB error to obtain its decision rule.

2.1. Out-of-sample test

Let $N_{test} = N - N_{train}$ be the number of test points. Moreover, $n_{0,r}$ is the number of observations coming from class 0, and $n_{1,r}$ the number of observations from class 1, for $r \in \{train, test\}$. We assume throughout the paper that $n_{0,r} \geq 1, n_{1,r} \geq 1$; otherwise we accept the null. If there is no difference in the distribution of the two groups, it clearly holds that

$$\mathbb{P}(\ell_i = 1 | \mathbf{Z}_i) = \mathbb{P}(\ell_i = 1) = \pi.$$

In other words, ℓ_i is independent of \mathbf{Z}_i . If $\pi = 1/2$, a test can be constructed by considering the overall out-of-sample classification error,

$$\hat{L}^{(\hat{g})} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \mathbb{I}\{\hat{g}(\mathbf{Z}_i) \neq \ell_i\},$$

which has $N_{test} \hat{L}^{(\hat{g})} \sim \text{Bin}(N_{test}, 1/2)$ under the null hypothesis. Here, $\mathbb{I}\{\hat{g}(\mathbf{Z}_i) \neq \ell_i\}$ takes the value 1 if $\hat{g}(\mathbf{Z}_i) \neq \ell_i$ and 0 otherwise. In an effort to extend this principle for general π , we instead use an approach based on the class-wise errors

$$\hat{L}_0^{(\hat{g})} = \frac{1}{n_{0,test}} \sum_{\{i:\ell_i=0\}} \mathbb{I}\{\hat{g}(\mathbf{Z}_i) \neq 0\}, \quad \hat{L}_1^{(\hat{g})} = \frac{1}{n_{1,test}} \sum_{\{i:\ell_i=1\}} \mathbb{I}\{\hat{g}(\mathbf{Z}_i) \neq 1\},$$

similar to Kim et al. (2021). Define, for $j \in \{0, 1\}$, the true class-wise loss for a given classifier \hat{g} as $L_j^{(\hat{g})} = \mathbb{P}(\hat{g}(\mathbf{Z}) \neq j | D_{N_{train}}, \ell = j)$. As shown in the proof of Proposition 1, conditioned on the training data and the number of observations from class $j \in \{0, 1\}$, $n_{j,test} \hat{L}_j^{(\hat{g})} | D_{N_{train}}, n_{j,test} \sim \text{Bin}(n_{j,test}, L_j^{(\hat{g})})$. The loss $L_j^{(\hat{g})}$ depends on the classifier and is generally not known, even under H_0 . However if $P = Q$, it holds that

$$\begin{aligned} L_0^{(\hat{g})} + L_1^{(\hat{g})} &= \mathbb{P}(\hat{g}(\mathbf{Z}) = 1 | D_{N_{train}}, \ell = 0) + \mathbb{P}(\hat{g}(\mathbf{Z}) = 0 | D_{N_{train}}, \ell = 1) \\ &= \mathbb{P}(\hat{g}(\mathbf{Z}) = 1 | D_{N_{train}}) + \mathbb{P}(\hat{g}(\mathbf{Z}) = 0 | D_{N_{train}}) \\ &= 1, \end{aligned}$$

where we used independence of ℓ and \mathbf{Z} when $P = Q$. As a side-note, this shows that $L_0^{(\hat{g})} + L_1^{(\hat{g})} = 1$ will be true, as soon as ℓ and $\hat{g}(\mathbf{Z})$ are independent. This follows if $P = Q$, but also if \hat{g} negates the dependence between ℓ and \mathbf{Z} , which essentially means it has no discriminating abilities.

Thus under H_0 , $L_0^{(\hat{g})} = 1 - L_1^{(\hat{g})}$. Define for $p \in [0, 1]$ the linear combination, $\hat{L}_p^{(\hat{g})} := (1 - p)\hat{L}_0^{(\hat{g})} + p\hat{L}_1^{(\hat{g})}$ and

$$\hat{\sigma}_c := 1/2 \sqrt{\frac{\hat{L}_0^{(\hat{g})}(1 - \hat{L}_0^{(\hat{g})})}{n_{0,test}} + \frac{\hat{L}_1^{(\hat{g})}(1 - \hat{L}_1^{(\hat{g})})}{n_{1,test}}}.$$

Let moreover,

$$\hat{g}(D_N) := (\hat{g}(\mathbf{Z}_1), \dots, \hat{g}(\mathbf{Z}_N)).$$

We are then able to formulate the following decision rule:

$$\delta_B(\hat{g}(D_{N_{test}})) := \mathbb{I}\left\{\hat{L}_{1/2}^{(\hat{g})} - 1/2 < \hat{\sigma}_c \Phi^{-1}(\alpha) + \epsilon_{N_{test}}\right\}, \tag{2}$$

where $\Phi^{-1}(\alpha)$ is the α quantile of the standard normal distribution and $\epsilon_{N_{test}}$ is a decreasing sequence of small non-random numbers. Then

Proposition 1. *There exists a sequence $\epsilon_{N_{test}}$, such that the decision rule in (2) conserves the level asymptotically, i.e.*

$$\limsup_{N_{test} \rightarrow \infty} \mathbb{P}(\delta_B(\hat{g}(D_{N_{test}})) = 1) \leq \alpha,$$

under $H_0 : P = Q$.

Proposition 1 is related to the first part of Proposition 9.1 in Kim et al. (2021). Note that we did not put any restrictions on how $L_0^{(\hat{g})}, L_1^{(\hat{g})}$ change individually and in particular, we made no assumption on how N_{train} behaves, as N_{test} goes to infinity. The reason for including the sequence ϵ_N is that, when N_{train} increases with N_{test} , boundary cases are possible, in which the variance $L_0^{(\hat{g})}(1 - L_0^{(\hat{g})}) + L_1^{(\hat{g})}(1 - L_1^{(\hat{g})})$ decreases as $1/N_{test}$ or faster, while still being nonzero for finite N . In this case the asymptotic normality of $(\hat{L}_{1/2}^{(\hat{g})} - 1/2)/\hat{\sigma}_c$ breaks down and it becomes increasingly difficult to control the behavior of the acceptance probability under the null. Adding ϵ_N makes it possible to circumvent this difficulty, albeit at the price of a potential loss in asymptotic power in these boundary cases. If N_{train} grows at the same rate as N_{test} , such boundary cases appear unlikely in practice. In fact, for a Random Forest classifier, it rather seems the classifier just outputs the majority class, such that $\hat{\sigma}_c = 0$ and $\hat{L}_{1/2}^{(\hat{g})} = 0, \hat{L}_{1/2}^{(\hat{g})} = 1$ or $\hat{L}_{1/2}^{(\hat{g})} = 1, \hat{L}_{1/2}^{(\hat{g})} = 0$. In this case the level is guaranteed, even if $\epsilon_N = 0$ for all N . We will in the following simply take $\epsilon_{N_{test}} = 0$ for the remainder of this paper. The test is summarized in Algorithm 1.

We briefly highlight the connection between the above decision rule and the one based on the overall classification error $\hat{L}^{(\hat{g})}$, in the case of $\pi = 1/2$ and $\epsilon_{N_{\text{test}}} = 0$. Since, for $\hat{\pi} = n_{1,\text{test}}/N_{\text{test}}$,

$$\hat{L}^{(\hat{g})} = (1 - \hat{\pi})\hat{L}_0^{(\hat{g})} + \hat{\pi}\hat{L}_1^{(\hat{g})} = \hat{L}_{\hat{\pi}}^{(\hat{g})}, \quad (3)$$

and $\hat{\pi} \rightarrow \pi = 1/2$ a.s., it holds that $|\hat{L}^{(\hat{g})} - \hat{L}_{1/2}^{(\hat{g})}| \rightarrow 0$, a.s. Consequently, the (unconditional) limiting distribution of $\hat{L}_{1/2}^{(\hat{g})}$ is the same as that of $\hat{L}^{(\hat{g})}$ or,

$$\frac{\sqrt{N_{\text{test}}} \left(\hat{L}_{1/2}^{(\hat{g})} - 1/2 \right)}{\sqrt{1/4}} \rightarrow N(0, 1),$$

under H_0 . In particular, the asymptotic variance of $\hat{L}_{1/2}^{(\hat{g})}$ under the null is the variance of $\hat{L}^{(\hat{g})}$ and thus one would expect the two tests to behave roughly the same for a large sample size, in the case of $\pi = 1/2$. However, as we demonstrate in Section 2.3, focusing on an equally weighted in-class loss, instead of the overall loss $\hat{L}^{(\hat{g})}$, can be beneficial when $\pi \neq 1/2$.

Algorithm 1 BinomialTest \leftarrow function(Z, ℓ, \dots)

Require: $\mathbf{Z} \in \mathbb{R}^{N \times d}$, $\ell \in \{0, 1\}^N$

- 1: $D_{N_{\text{train}}} \leftarrow (\ell_i, \mathbf{Z}_i)_{i=1}^{N_{\text{train}}}$ ▷ random separation of training data
- 2: Training of a classifier, $\hat{g}(\cdot)$ on $D_{N_{\text{train}}}$
- 3: $err_0 \leftarrow \frac{1}{n_{0,\text{test}}} \sum_{i=N_{\text{train}}+1}^N \mathbb{I}\{\ell_i = 0\} \mathbb{I}\{\hat{g}(\mathbf{Z}_i) \neq 0\}$
- 4: $err_1 \leftarrow \frac{1}{n_{1,\text{test}}} \sum_{i=N_{\text{train}}+1}^N \mathbb{I}\{\ell_i = 1\} \mathbb{I}\{\hat{g}(\mathbf{Z}_i) \neq 1\}$
- 5: $err_{1/2} \leftarrow \frac{1}{2} err_0 + \frac{1}{2} err_1$ ▷ calculating the out-of-sample classification error
- 6: $sig \leftarrow 1/2 \sqrt{err_0(1 - err_0)/n_{0,\text{test}} + err_1(1 - err_1)/n_{1,\text{test}}}$
- 7: **if** $sig > 0$ **then**
- 8: $pvalue \leftarrow \Phi \left(\frac{err_{1/2} - 1/2}{sig} \right)$
- 9: **else if** $sig == 0$ **then**
- 10: $pvalue \leftarrow \mathbb{I}\{err_{1/2} - 1/2 > 0\}$
- 11: **end if**
- 12: **return** $pvalue$

Naturally, the split in training and test set is not ideal. For finite sample sizes, one would like to have as many (test) samples as possible to detect differences. At the same time, it would be preferable to have the classifier trained on many data points. This in fact resembles a bias-variance trade-off, similar to what was described in Lopez-Paz and Oquab (2018): Let $g_{1/2}^*$ and $L_{\pi}^{(g_{1/2}^*)}$ be the Bayes classifier and Bayes error respectively, both defined in Section 2.3. For $\pi = 1/2$, there is a trade-off between the closeness of $L^{(\hat{g})}$ to $L_{\pi}^{(g_{1/2}^*)}$, which may be achieved through a large training set and the closeness of $\hat{L}^{(\hat{g})}$ to $L^{(\hat{g})}$, which is generally only true in large test sets.

2.2. Out-of-bag test

For the purpose of overcoming the arbitrary split in training and testing, Random Forest delivers an interesting tool: the OOB error introduced in Breiman (2001). Since each tree is built on a bootstrapped sample taken from D_N , approximately 1/3 of the trees will not use the i th observation (ℓ_i, \mathbf{Z}_i) . Thus we may use this ensemble of trees not containing observation i to obtain an estimate of the out-of-sample error for i . We slightly generalize this here, in assuming we have an ensemble learner g : That is, we assume to have iid copies of a random element ν, ν_1, \dots, ν_B , such that each $\hat{g}_{\nu_b}(\mathbf{Z}) := g(\mathbf{Z}, D_{N_{\text{train}}}, \nu_b)$ is a different classifier. We then consider the average

$$\hat{g}(\mathbf{Z}) := \frac{1}{B} \sum_{b=1}^B \hat{g}_{\nu_b}(\mathbf{Z}). \quad (4)$$

For $B \rightarrow \infty$, it holds that (a.s.) $\hat{g}(\mathbf{Z}) \rightarrow \mathbb{E}_{\nu}[\hat{g}_{\nu}(\mathbf{Z})]$. For Random Forest, ν usually represents the bootstrap sampling of observations and the sampling of variables to consider at each splitpoint for a given tree.

Let as before, $n_0 := \sum_{i=1}^N \mathbb{I}\{\ell_i = 0\}$ and $n_1 := \sum_{i=1}^N \mathbb{I}\{\ell_i = 1\}$, with $n_0 \geq 1, n_1 \geq 1$. We assume in the following that each $\hat{g}_{\nu_b}(\mathbf{Z})$ uses a bootstrapped sample from the original data, as Random Forest does. The class-wise OOB error of such an ensemble of learners trained on N observations is defined as

$$\mathcal{E}_0^{\text{ob}} = \frac{1}{n_0} \sum_{i=1}^N \mathbb{I}\{\ell_i = 0\} \mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq 0\},$$

$$\mathcal{E}_1^{oob} = \frac{1}{n_1} \sum_{i=1}^N \mathbb{I}\{\ell_i = 1\} \mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq 1\},$$

$$\mathcal{E}_p^{oob} = (1-p)\mathcal{E}_0^{oob} + p\mathcal{E}_1^{oob},$$

where \hat{g}_{-i} , represents the average over the ensemble of learners not containing the i^{th} observation for training.

Unfortunately, the test statistic $\mathcal{E}_{1/2}^{oob}$ is difficult to handle; due to the complex dependency structure between the elements of the sum, it is not clear what the (asymptotic) distribution under the null is. For theoretical purposes, we consider in Section 3 a solution based on the concept of U-statistics. Here, we recommend using the OOB error together with a permutation test. See e.g., Good (1994) or Kim et al. (2021), who use it in conjunction with the out-of-sample error evaluated on a test set: We first calculate the class-wise OOB errors \mathcal{E}_0^{oob} , \mathcal{E}_1^{oob} and then reshuffle the labels K times to obtain K permutations, $\sigma_1, \dots, \sigma_K$ say. For each of these new datasets $(\mathbf{Z}_i, \ell_{\sigma_k(i)})_{i=1}^N$, $k \in \{1, \dots, K\}$, we calculate the OOB errors

$$\mathcal{E}_j^{oob,k} := \frac{1}{n_j} \sum_{i=1}^N \mathbb{I}\{\ell_{\sigma_k(i)} = j\} \mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_{\sigma_k(i)}\},$$

for $j \in \{0, 1\}$. Under H_0 , (ℓ_1, \dots, ℓ_N) and $(\mathbf{Z}_1, \dots, \mathbf{Z}_N)$ are independent and each $\mathcal{E}_{1/2}^{oob}$ is simply an iid draw from the distribution F of the random variable $\mathcal{E}_{1/2}^{oob} | (\mathbf{Z}_1, \dots, \mathbf{Z}_N)$. As such we can accurately approximate the α quantile $F^{-1}(\alpha)$ of said distribution by performing a large number of permutations and use the decision rule

$$\delta_{oob}(D_N) = \left\{ \mathcal{E}_{1/2}^{oob} \leq F^{-1}(\alpha) \right\}. \quad (5)$$

Thus, as in the decision in Equation (2), the rejection region depends on the data at hand. Nonetheless, the level will be conserved, as proven e.g. in Hemerik and Goeman (2018, Theorem 1).

Heuristically, this procedure will have power under the alternative, as in this case there is some dependence between (ℓ_1, \dots, ℓ_N) and $(\mathbf{Z}_1, \dots, \mathbf{Z}_N)$, formed by the difference in the distribution of the \mathbf{Z}_i . The OOB error $\mathcal{E}_{1/2}^{oob}$ will thus be different than those observed under permutations.

The whole procedure is described in Algorithm 2. We name this test ‘‘hypoRF’’.

Algorithm 2 hypoRF \leftarrow function(\mathbf{Z}, K, \dots)

Require: $\mathbf{Z} \in \mathbb{R}^{N \times d}$, $\ell \in \{0, 1\}^N$, K

```

1:  $D_N \leftarrow (\ell_i, \mathbf{Z}_i)_{i=1}^N$ 
2:  $n_j \leftarrow \sum_{i=1}^N \mathbb{I}\{\ell_{\sigma_k(i)} = j\}$ 
3: Training of an ensemble learner  $\hat{g}(\cdot)$  on  $D_N$ 
4:  $OOB_j \leftarrow \frac{1}{n_j} \sum_{i=1}^N \mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq j\} \mathbb{I}\{\ell_i = j\}$  ▷ calculating the OOB-error for  $j \in \{0, 1\}$ 
5:  $OOB_{1/2} \leftarrow 1/2(OOB_0 + OOB_1)$ 

6: for  $k$  in  $1:K$  do
7:  $D_N^k \leftarrow (\ell_{\sigma_k(i)}, \mathbf{Z}_i)_{i=1}^N$  ▷ reshuffle the label
8:  $OOB_j^k \leftarrow \frac{1}{n_j} \sum_{i=1}^N \mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq j\} \mathbb{I}\{\ell_{\sigma_k(i)} = j\}$ 
9:  $OOB_{1/2}^k \leftarrow 1/2(OOB_0^k + OOB_1^k)$  ▷ calculating the OOB-error
10: end for

11:  $mean \leftarrow \frac{1}{K} \sum_{k=1}^K OOB_{1/2}^k$ 
12:  $sig \leftarrow \sqrt{\frac{1}{K-1} \sum_{k=1}^K (OOB_{1/2}^k - mean)^2}$ 
13: if  $sig > 0$  then
14:  $pvalue \leftarrow \frac{1}{K+1} \left( \sum_{k=1}^K \mathbb{I}\{OOB_{1/2}^k < OOB_{1/2}\} + 1 \right)$ 
15: else if  $sig == 0$  then
16:  $pvalue \leftarrow \mathbb{I}\{OOB_{1/2} - mean > 0\}$ 
17: end if
18: return  $pvalue$ 

```

2.3. What classifier to use

The foregoing tests are valid for any classifier $g : \mathcal{X} \rightarrow \{0, 1\}$. In practice, most classifiers try to approximate the Bayes classifier: Let for p, q the densities of P, Q

$$\eta(\mathbf{z}) := \mathbb{E}[\ell | \mathbf{z}] = \frac{\pi q(\mathbf{z})}{\pi q(\mathbf{z}) + (1 - \pi)p(\mathbf{z})}, \quad (6)$$

then the Bayes classifier is given as $g_{1/2}^*(\mathbf{Z}) = \mathbb{I}\{\eta(\mathbf{Z}) > 1/2\}$, see e.g., Devroye et al. (1996). It is the classifier with minimal classification error, designated the Bayes error $L_{\pi}^{(g_{1/2}^*)} = \mathbb{P}(g_{1/2}^*(\mathbf{Z}) \neq \ell)$. Under H_0 , this Bayes error will be $\min(\pi, 1 - \pi)$.

An interesting question is whether $g_{1/2}^*$ leads to a consistent test in our framework. We first define consistency for a hypothesis test: Let Θ be the space of tuples of all distributions on \mathbb{R}^d , $\theta = (P, Q) \in \Theta$, $\Theta_0 = \{(P, Q) : P = Q\}$, $\Theta_1 = \{(P, Q) : P \neq Q\}$. Let $\delta : \mathcal{X}^N \rightarrow \{0, 1\}$ be a decision rule and $\phi(\theta) := \mathbb{E}_{\theta}[\delta]$. Following e.g., van der Vaart (1998) we call a test consistent at level α (for Θ_1), if $\limsup_N \sup_{\theta \in \Theta_0} \phi(\theta) \leq \alpha$ and for any $\theta \in \Theta_1$, $\liminf_N \phi(\theta) = 1$. For theoretical purposes, we extend this definition also to δ which depend on the unknown θ itself, for instance via the densities of P and Q respectively.

Under the assumption of equal class probabilities $\pi = 1/2$ the Bayes error has the property that,

$$L_{\pi}^{(g_{1/2}^*)} = L_{1/2}^{(g_{1/2}^*)} = 1/2(1 - TV(P, Q)), \tag{7}$$

where $TV(P, Q)$ is the total variation distance between P, Q : $TV(P, Q) = 2 \sup_A |P(A) - Q(A)|$, with the supremum taken over all Borel sets on \mathbb{R}^d . As TV defines a metric on the space of all probability measures on \mathbb{R}^d , it holds that $P = Q \iff TV(P, Q) = 0$. Consequently, as soon as there is any difference in P and Q , $TV(P, Q) > 0$ and $L_{\pi}^{(g_{1/2}^*)} < 1/2$. Thus we would expect a test based on $g_{1/2}^*$ to be consistent. More generally, Kim et al. (2021) prove that if the classifier \hat{g} is such that

$$\hat{L}_0^{(\hat{g})} = L_0 + o_{\mathbb{P}}(1), \hat{L}_1^{(\hat{g})} = L_1 + o_{\mathbb{P}}(1), \text{ for some } L_0, L_1 \in (0, 1) \text{ with } L_0 + L_1 = 1 - \varepsilon, \text{ for any } \varepsilon > 0, \tag{8}$$

then the decision rule in (2) is consistent.

Unfortunately, this assumption does not hold for $g_{1/2}^*$, if $\pi \neq 1/2$. In this case, simple counterexamples show that even when P, Q are different, it might still be that $L_0^{(g_{1/2}^*)} + L_1^{(g_{1/2}^*)} = 1$.

Lemma 1. Take $\mathcal{X} \subset \mathbb{R}$ and $\pi \neq 1/2$. Then no decision rule of the form, $\delta(D_N) = \delta(g_{1/2}^*(D_N))$ is consistent.

Thus even though we allow the classifier $g_{1/2}^*$ to depend for each $(P, Q) \in \Theta_1$ on the densities p of P and q of Q , we are not able to construct a consistent test. The problem appears to be that the Bayes classifier minimizes the overall classification loss, so that condition (8) cannot hold. In doing so, it focuses too much on the overrepresented class. Indeed, we might define the following alternative classifier: For given P, Q let g_{π}^* be the classifier that minimizes the error $L_{1/2}^{g_{\pi}^*}$, i.e. a classifier that solves the problem

$$\arg \min\{L_{1/2}^g : g : \mathcal{X} \rightarrow \{0, 1\} \text{ a classifier}\}. \tag{9}$$

It turns out that a slight variation to the Bayes classifier solves this problem:

Lemma 2. The classifier

$$g_{\pi}^*(\mathbf{z}) = \mathbb{I}\{\eta(\mathbf{z}) > \pi\}, \tag{10}$$

is a solution to (9). Moreover it holds that

$$1 - TV(P, Q) = L_0^{g_{\pi}^*} + L_1^{g_{\pi}^*}, \tag{11}$$

for any $\pi \in (0, 1)$.

Thus for this classifier a generalization of (7) holds for any $\pi \in (0, 1)$. In particular, it now yields a consistent test:

Corollary 1. The decision rule $\delta_B(g_{\pi}^*(D_N))$ in (2) is consistent for any $\pi \in (0, 1)$.

Since this theoretical classifier needs no training, the two testing approaches coincide with an evaluation of the classifier loss on the overall data D_N . While this analysis with theoretical classifiers is by no means sufficient for the much more complicated case of a classifier \hat{g} trained on data, it suggests that adapting the ‘‘cutoff’’ in a given classifier might improve consistency issues. Indeed, we use the classifier

$$\hat{g}(\mathbf{z}) = \mathbb{I}\{\hat{\eta}(\mathbf{z}) > \hat{\pi}\},$$

where $\hat{\pi}$ is an estimate of the prior probability based on the training data. As long as the latter is used (as opposed to the test data), the tests above are still valid.

3. Tests based on U-statistics

To avoid the splitting in training and test set, we introduced an OOB error-based test in Section 2.2. In this section, we discuss a potential framework to analyze a version of such a test theoretically. For $N_{train} \leq N$, let again, $n_{0,train} = \sum_{i=1}^{N_{train}} \mathbb{I}\{\ell_i = 0\}$ and $n_{1,train} = \sum_{i=1}^{N_{train}} \mathbb{I}\{\ell_i = 1\}$. Let $D_{N_{train}}^{-i}$ denote the data set without observation (ℓ_i, \mathbf{Z}_i) and define \hat{g} as in (4) for B learners. Then we consider the class-wise OOB error based on N_{train} observations:

$$\begin{aligned} h_{N_{train}}((\ell_1, \mathbf{Z}_1), \dots, (\ell_{N_{train}}, \mathbf{Z}_{N_{train}})) &:= \frac{1}{2} \left(\frac{1}{n_{0,train}} \sum_{i:\ell_i=0} \mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) = 1\} + \frac{1}{n_{1,train}} \sum_{i:\ell_i=1} \mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) = 0\} \right) \\ &= \frac{1}{2} \sum_{i=1}^{N_{train}} \varepsilon_i^{oob}, \end{aligned} \tag{12}$$

where

$$\varepsilon_i^{oob} := \mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i\} \left(\frac{1 - \ell_i}{n_{0,train}} + \frac{\ell_i}{n_{1,train}} \right),$$

for \hat{g}_{-i} trained on $D_{N_{train}}^{-i}$. Also recall that $L_j^{\hat{g}} = \mathbb{P}(\hat{g}(\mathbf{Z}) = j | D_{N_{train}}, \ell \neq j)$ for $j \in \{0, 1\}$ and $L_{1/2}^{\hat{g}} = 1/2(L_0^{\hat{g}} + L_1^{\hat{g}})$. We assume that the number of classifiers in the ensemble, $B \rightarrow \infty$, so that $\hat{g}(\mathbf{Z}) \rightarrow \mathbb{E}_v[\hat{g}_v(\mathbf{Z})]$, almost surely. We refer to the function $h_{N_{train}}$ as kernel of size N_{train} and define the incomplete U-Statistics,

$$\hat{U}_{N,K} := \frac{1}{K} \sum h_{N_{train}}((\mathbf{Z}_{i_1}, \ell_{i_1}), \dots, (\mathbf{Z}_{i_{N_{train}}}, \ell_{i_{N_{train}}})) \tag{13}$$

where the sum is taken over K randomly chosen subsets of size N_{train} - see e.g., Lee (1990), Fuchs et al. (2013), Mentch and Hooker (2016), Peng et al. (2019). We assume that K goes to infinity as N goes to infinity. Since we are only considering learners for which the i th sample point is not included, we may simply see \hat{g}_{-i} as the average of an infinite ensemble build on the dataset $D_{N_{train}}^{-i}$ only. Consequently, with the assumption of an infinite number of learners, the OOB error is “almost” unbiased for $\mathbb{E}[L_{1/2}^{\hat{g}}]$.

Lemma 3. $\mathbb{E}[h_{N_{train}}((\ell_1, \mathbf{Z}_{N_{train}}), \dots, (\ell_{N_{train}}, \mathbf{Z}_{N_{train}}))] = \mathbb{E}[L_{1/2}^{\hat{g}_{-i}}]$.

Here, $\mathbb{E}[L_{1/2}^{\hat{g}_{-i}}]$ refers to the expected value of the error based on the classifier trained on $N_{train} - 1$ data points. As such, it does not depend on i . This is essentially the same result as in Luntz and Brailovsky (1969) in the case of the leave-one-out error.

We are now able to show that $h_{N_{train}}$ in (12) is a symmetric function, unbiased for $\mathbb{E}[L_{1/2}^{\hat{g}_{-i}}]$:

Lemma 4. $h_{N_{train}}$ is a valid kernel for the expectation $\mathbb{E}[L_{1/2}^{\hat{g}_{-i}}]$.

Combining arguments from Mentch and Hooker (2016) and Wager and Athey (2017), we obtain the conditions for asymptotic normality listed in Theorem 1. Though both papers consider the asymptotic distribution of a Random Forest prediction at a fixed \mathbf{z} , the U-Statistics theory they develop can be used in our context as well. We also refer to Peng et al. (2019) and DiCiccio and Romano (2020), who already refined the results of Mentch and Hooker (2016) for asymptotic normality of a U-statistics with growing kernel size. Peng et al. (2019) in particular, derived a similar result to Theorem 1 independently from us. Let for random variables $\xi_1, \xi_2, \mathbb{V}(\xi_1), \text{Cov}(\xi_1, \xi_2)$ be the variance and covariance respectively and define for the following, for $c \in \{1, \dots, N_{train}\}$,

$$\zeta_{c, N_{train}} = \mathbb{V}(\mathbb{E}[h_{N_{train}}((\mathbf{Z}_1, \ell_1), \dots, (\mathbf{Z}_{N_{train}}, \ell_{N_{train}})) | (\mathbf{Z}_1, \ell_1), \dots, (\mathbf{Z}_c, \ell_c)]). \tag{14}$$

In particular, $\zeta_{1, N_{train}}$ and $\zeta_{N_{train}, N_{train}}$ will be of special interest. Lee (1990) provides an immediate important result:

Lemma 5. $N_{train} \zeta_{1, N_{train}} \leq \zeta_{N_{train}, N_{train}}$

Lemma 5, which is actually true for any U-statistics, shows that, whenever the second moment of the kernel $h_{N_{train}}$ exists, $\zeta_{1, N_{train}} = O(N_{train}^{-1})$. Then

Theorem 1. Assume that for $N \rightarrow \infty$, $N_{train} = N_{train}(N) \rightarrow \infty$ and $K = K(N) \rightarrow \infty$,

$$\lim_N \frac{KN_{train}^2}{N} \frac{\zeta_{1,N_{train}}}{\zeta_{N_{train},N_{train}}} = 0, \tag{15}$$

$$\lim_N \frac{\sqrt{KN_{train}}}{N} = 0. \tag{16}$$

Then,

$$\frac{\sqrt{K}(\hat{U}_{N,K} - \mathbb{E}[L_{1/2}^{(\hat{g}^{-1})}])}{\sqrt{\zeta_{N_{train},N_{train}}}} \xrightarrow{D} N(0, 1). \tag{17}$$

Condition (15) is hard to control in general, but with Lemma 5, it can be seen that choosing

$$\frac{KN_{train}}{N} \rightarrow 0, \tag{18}$$

is sufficient for both (15) and (16). If $K = \log(N_{train})^{1+d}$, this corresponds to the condition $\log(N_{train})^{1+d}N_{train}/N \rightarrow 0$ required by Wager and Athey (2017). In the context of Random Forest, Theorem 1 essentially proves that the OOB error of a prediction function that is bounded, is asymptotically normal if the number of trees is “high” and if K forests are trained on subsamples such that (15) and (16) are true. Since the OOB error with infinite learners is essentially the leave-one-out error in the context of cross-validation, this also means that a test of the cross-validation error could be derived under much weaker assumptions as for instance in Fuchs et al. (2013). The key reason for the generality of the result, as was also realized by Peng et al. (2019), is that K should be chosen small relative to N . This introduces additional variance, such that conditions on $\zeta_{1,N_{train}}$ usually required in such results, see e.g., DiCiccio and Romano (2020), can be replaced by (18). This has an additional computational advantage, but it may come at the price of reduced power, as will be seen in Corollary 2.

Mentch and Hooker (2016, Section 3) also provide a consistent estimate for $\zeta_{c,N_{train}}$, denoted $\hat{\zeta}_{c,N_{train}}$, for any $c \in \{1, \dots, N_{train}\}$. As its population counterpart, this estimator is also bounded by 1 for all c and N_{train} in our case. Thus if for a classifier (15) and (16) are true, the decision rule

$$\delta(\hat{g}(D_N)) = \mathbb{I} \left\{ \frac{\sqrt{K}(\hat{U}_{N,K} - 1/2)}{\sqrt{\hat{\zeta}_{N_{train},N_{train}}}} < \Phi^{-1}(\alpha) \right\}, \tag{19}$$

constitutes a valid test. To illustrate Theorem 1, Fig. 2 displays the simulated distribution of

$$Z = \frac{\sqrt{K}(\hat{U}_{N,K} - \mathbb{E}[L_{1/2}^{(\hat{g}^{-i})}])}{\sqrt{\hat{\zeta}_{N_{train},N_{train}}}}, \tag{20}$$

for $P = N(\boldsymbol{\mu}_1, I_{10 \times 10})$ and $Q = N(\boldsymbol{\mu}_2, I_{10 \times 10})$, with $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = 0.4/\sqrt{10} \cdot \mathbf{1}$. We simulated $S = 500$ replications using $N = 6000$, $K = \lceil 2 * \log(N) \rceil = 17$ and $N_{train} = \lceil N/(K * \log(\log(N))) \rceil = 163$.

With this at hand, we can construct another test:

Corollary 2. Assume the conditions of Theorem 1 hold true and that $\hat{\zeta}_{N_{train},N_{train}}/\zeta_{N_{train},N_{train}} \xrightarrow{P} 1$. Then the decision rule in (19) conserves the level asymptotically and has approximate power

$$\Phi \left(\Phi^{-1}(\alpha) + \sqrt{\frac{K}{\zeta_{N_{train},N_{train}}}} (1/2 - \mathbb{E}[L_{1/2}^{(\hat{g}^{-i})}]) \right). \tag{21}$$

The test has thus power going to one, as soon as

$$\limsup_N \mathbb{E}[L_{1/2}^{(\hat{g}^{-i})}] < 1/2. \tag{22}$$

Condition (22) mirrors condition (A9) in Kim et al. (2021), in that it asks for a better than chance prediction in expectation. Crucially, Corollary 2 also illustrates the downside of the weak assumptions used in Theorem 1: The power is dependent on \sqrt{K} , as well as the accuracy of the trained classifier through $\mathbb{E}[L_{1/2}^{(\hat{g}^{-i})}]$. Since our theory requires that K is of small order compared to N , we lose power, at least theoretically. In practice, it appears from simulations with Random Forest that $\zeta_{N_{train},N_{train}}$ decreases to zero and roughly behaves like $1/N_{train}$. From the asymptotic power expression above, it can be seen that this would offset the small order K . Nonetheless, the test of Corollary 2 appears less powerful than the Binomial and

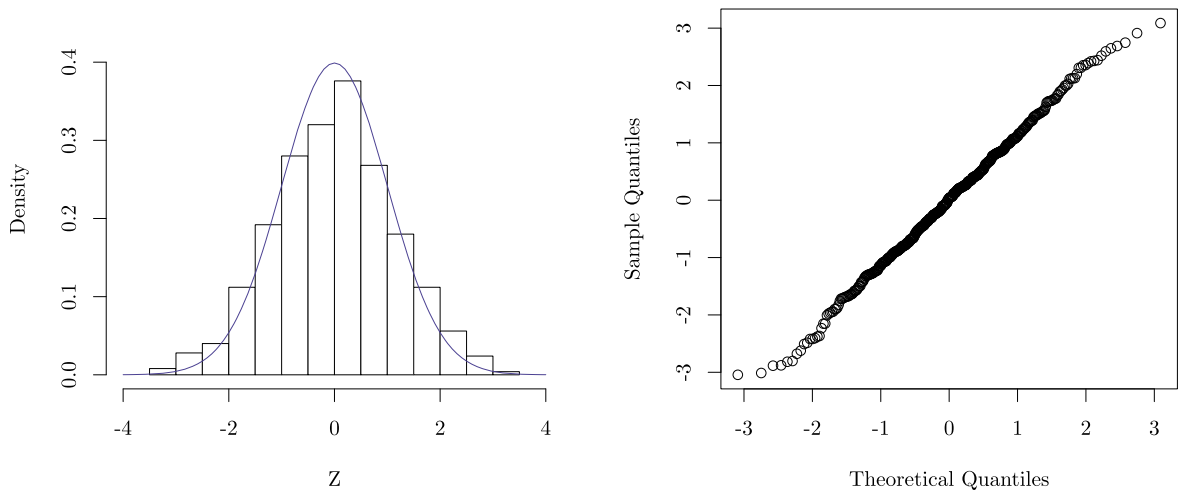


Fig. 2. Illustration of the asymptotic normality of the OOB error based test-statistic for the Random Forest classifier. In this example, $P = N(\mathbf{0}, I_{10 \times 10})$ and $Q = N(0.4/\sqrt{10} \cdot \mathbf{1}, I_{10 \times 10})$, an $N = 6000$, $K = 2 \lceil \log(N) \rceil = 17$ were chosen over 500 replications.

hypoRF test. In the example of Fig. 2, plugging the estimate of $\mathbb{E}[L_{1/2}^{(\hat{g}-i)}]$ obtained from the 500 repetitions into (21) and averaging, we obtain an expected power of 0.63. The actual power, i.e. the fraction of rejected tests over the 500 repetitions, is given as 0.61. The Binomial test with Random Forest on the other hand, reaches a power of 1. This illustrates that the test derived in this section still lags behind the test that uses sample-splitting. Nonetheless, modern U -statistics theory gives powerful theoretical tools to construct OOB-error based tests with tractable asymptotic power.

4. Application

In this section, we first describe the proposed significance threshold for the variable importance measure and apply the hypoRF test to simulated and real application cases. In the simulation section, we will compare the hypoRF to recent kernel-based tests by investigating the power of a selected scenario. A more extensive simulation study is given in Appendix B. In Section 4.3, two real data sets from biology and finance are considered.

4.1. Variable importance measure

Variable importance measures in the context of Random Forest are practical tools introduced by Breiman (2001). As a by-product of the hypoRF test of Section 2.2, we obtain a significance threshold for such a given variable importance measure: For each permutation, we record the maximum variable importance measure I_σ over all variables, thus approximating the distribution of I_σ under H_0 . The estimated $1 - \alpha$ quantile of this distribution will then be used as the significance threshold. Every variable with an importance measure above this threshold will be called significant. This should serve as an additional hint, as to which components a rejection decision might originate from. We will use in all instances the “Gini” importance measure or “Mean Decrease Impurity”, see e.g., Biau and Scornet (2016, Section 5).

Obtaining p -values for the variable importance measure by permuting the response vector was developed much earlier in Altmann et al. (2010) and further developed in Janitza et al. (2018). As we are not directly interested in p -values for each variable, our approach differs slightly and is more in the spirit of the Westfall-Young permutation approach, see e.g., Westfall et al. (1995). Since we use a permutation approach already to define the decision rule of the hypoRF test, the significance threshold for the variable importance arises without any additional cost.

Fig. 1 in Section 1.2 demonstrates that in this example the Random Forest is able to correctly identify the effect of the last two components. This appears remarkable, as there is only a change in dependence, but no marginal change. On the other hand, one could imagine a situation, where no significant variable may be identified, but the test overall still rejects. This is illustrated in Fig. 3. In this example, instead of endowing only the last two components with correlations, we introduced correlations of 0.4 between all variables when changing from P to Q . Again the hypoRF test manages to differentiate between the two distributions. However this time, no significant variables can be identified. This seems sensible, as the source of change is divided equally between the different components in this example. Any situation could also be a mixture of the above extreme examples: There could be one or several significant variables, but the test still rejects, even after removing them. Section 4.3 will show real-world examples in which some variables can be identified to be significant in the above sense.

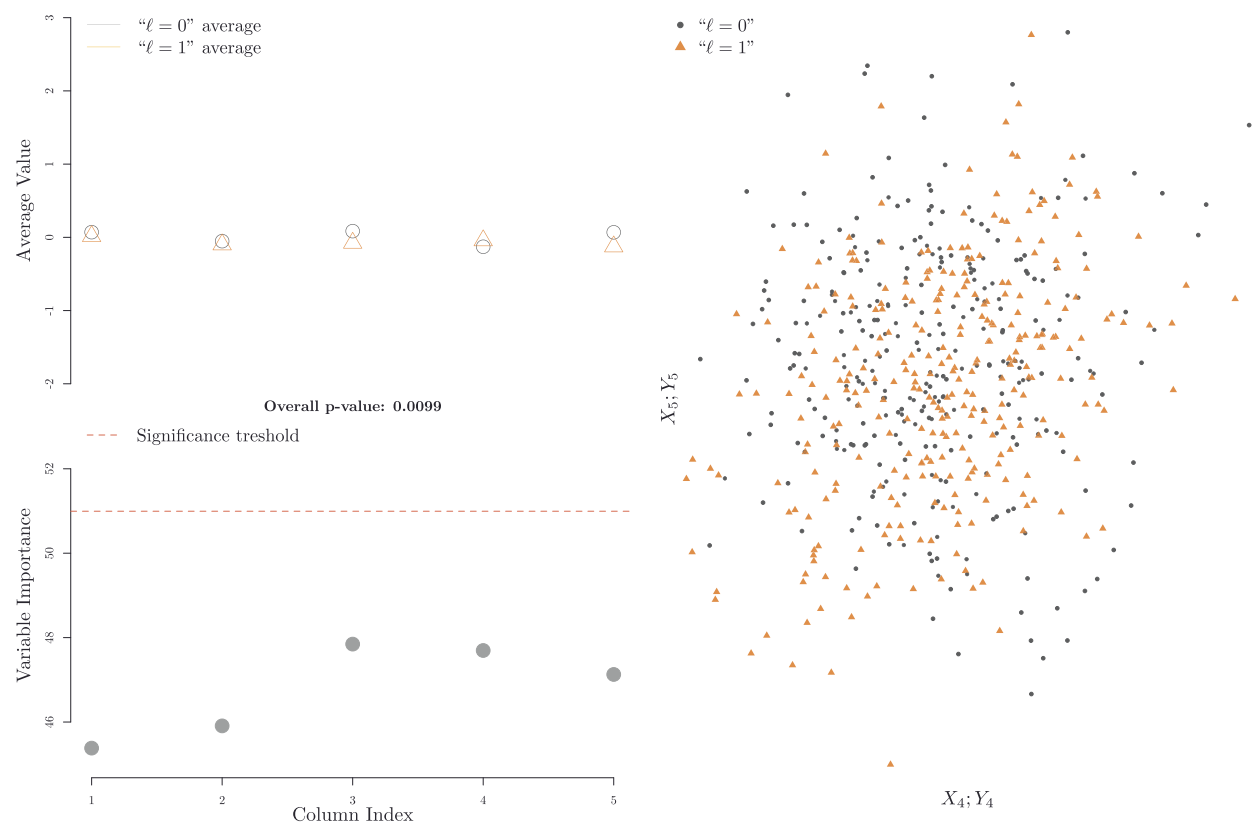


Fig. 3. (Application) We sampled 300 observations from a $d = 5$ dimensional multivariate normal, with no correlation between the marginals. Likewise 300 observations were sampled from a multivariate normal, where the pairwise correlation between the columns is 0.4. The Random Forest used 500 trees. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

4.2. Simulation

In what follows, we will demonstrate the power of the proposed tests through simulation, and compare it with 3 kernel methods and a recently proposed Random Forest test based on the classification probability. To this end, we will use both the first version of the test, as described in Algorithm 1 (“Binomial” test), and the refined version in Algorithm 2 (“hypoRF” test). For the latter, as mentioned in Section 2.2, we will use $K = 100$ permutations. For the Binomial test described in Algorithm 1 we decided to set $N_{train} = N_{test}$, as taking half of the data as training and the other half as a test set seems to be a sensible solution a priori. To conduct our simulations we will use the R-package “hypoRF” developed by the authors, which consists of the “hypoRF” function including the two proposed tests. For each pair of samples, we run all tests and save the decisions. The estimated power is then the fraction of rejected among the S tests.

The 3 kernel-based tests include the “quadratic time MMD” (Gretton et al., 2012a) using a permutation approach to approximate the H_0 distribution (“MMDboot”), its optimized version “MMD-full”, as well as the “ME” test with optimized locations, “ME-full” (Jitkrittum et al., 2016). The original idea of the “MMD-full” was formulated in Gretton et al. (2012), however they subsequently used a linear version of the MMD. We instead use the approach of Jitkrittum et al. (2016), which uses the optimization procedure of Gretton et al. (2012) together with the quadratic MMD from Gretton et al. (2012a). A Python implementation of these methods is available from the link provided in Jitkrittum et al. (2016) (<https://github.com/wittawatj/interpretable-test>). Among these tests, it seems the MMDboot is still somewhat of a gold-standard, with newer methods, such as those presented in Gretton et al. (2012), Chwialkowski et al. (2015) and Jitkrittum et al. (2016), more focused on developing more efficient versions of the test that are nearly as good. Nonetheless, the new methods often end up being surprisingly competitive or even better in some situations, as recently demonstrated in Jitkrittum et al. (2016). Thus we chose to include MMD-full, ME-full as well. For all tests, we use a Gaussian kernel, which is a standard and reasonable choice if no a priori knowledge about the optimal kernel is available. The Gaussian kernel requires a bandwidth parameter σ , which is tuned in MMD-full and ME-full based on training data. For MMDboot we use the “median heuristic”, as described in Gretton et al. (2012a, Section 8), which takes σ to be the median (Euclidean) distance between the elements in $(\mathbf{Z}_i)_{i=1}^{2n}$.

Finally, we consider the method of Cai et al. (2020), which is a test based on the classification probability of Random Forest. We would like to emphasize that their first publication on arXiv appeared more than 6 months after our first

upload on arXiv. As such, we do not view them as a direct competitor. Nonetheless, it seems interesting to compare their performance to the one of hypoRF, as they use a permutation approach based on the *in-sample* probability estimates.

We would like to stress that we did not use any tuning for the parameters of the RF-based tests, just as we did not use any tuning for MMDboot. As such, comparing the MMD/ME-full to the other methods might not be entirely fair. On the other hand, our chosen sample size might be too small for the optimized versions to work at full capacity. In particular, all optimized tests suffer from a similar drawback as our Binomial test: The tuning of the method takes up half of the available data. While Jitkrittum et al. (2016) find that ME-full outperforms the MMD, they only observe settings where the latter also uses half of the data to tune its kernel, as proposed in Gretton et al. (2012). In our terminology, they only compare ME-full to MMD-full, instead of MMDboot. It seems unclear a priori what happens if we instead employ the median heuristic for the MMD and let it use all of the available data, as in Gretton et al. (2012a). It should also be said that both optimization and testing of the ME-full scale linearly in N , making its performance below all the more impressive. On the other hand, the optimization depends on some hyperparameters common in gradient-based optimization, such as step size taken in the gradient step, the maximum number of iterations, etc. As this optimization is rather complicated for large d , some parameter choices sometimes lead to a longer runtime of the ME than the calculation-intensive hypoRF and CPT-RF. In general, it seems both runtime and performance of ME-full are in practice highly dependent on the chosen hyperparameters; we tried 3 different sets of parameters based on the code in <https://github.com/wittawatj/interpretable-test> with very different power results. The setting used in this simulation study is the exact same as that used in their simulation study.

As discussed in Ramdas et al. (2015), changing the parameters of our experiments (for instance the dimension d) should be done in a way that leaves the Kullback-Leibler (KL) Divergence constant. When varying the dimension d we generally follow this suggestion, though in our case, this is not as imminent; whatever unconscious advantage we might give our testing procedure is also inherent in the competing methods. Finally, also note that, while our methods would be in principle applicable to arbitrary classifiers, we did not compare our proposed tests with tests based on other classifiers, such as those used in Lopez-Paz and Oquab (2018). Rather, we believe the choice of classifiers for binary classification is a more general problem and should be studied separately, as for example done extensively in Fernández-Delgado et al. (2014). The only exception to this, is our use of an LDA classifier-based test for the example of a Gaussian mean-shift in B.1.

Where not otherwise stated, we use for the following experiments: $N = 600$ observations, 300 per class, $d = 200$ dimensions, $K = 100$ permutations and 600 trees for the RF-based tests. In some examples, we additionally study a sparse case, where the intended change in distribution appears only in $c < d$ components. Throughout, notation such as

$$P = \sum_{t=1}^T \omega_t N(\boldsymbol{\mu}_t, \Sigma_t),$$

with $\omega_t \geq 0$, $\sum_{t=1}^T \omega_t = 1$, $\boldsymbol{\mu}_t \in \mathbb{R}^d$, $\Sigma_t \in \mathbb{R}^{d \times d}$ means P is a discrete mixture of T d -valued Gaussians. Moreover, if P_1, \dots, P_d are distributions on \mathbb{R} , we will denote by

$$P = \prod_{j=1}^d P_j,$$

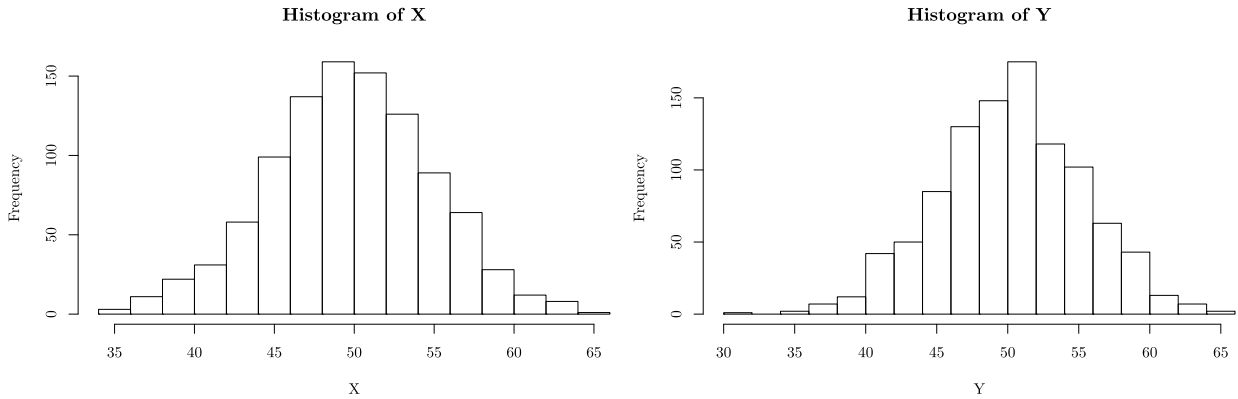
their product measure on \mathbb{R}^d . In other words, in this case, we simply take all the components of \mathbf{X} to be independent.

The prime example which we present here in the main text is rather challenging. Let $P = N(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu}$ set to $50 \cdot \mathbf{1}$ and $\Sigma = 25 \cdot I_{d \times d}$. For the alternative, we consider the mixture

$$Q = \lambda H_c + (1 - \lambda)P,$$

$\lambda \in [0, 1]$, and H_c some distribution on \mathbb{R}^d . This is a “contamination” of P by H_c with λ determining the contamination strength. Here, we take H_c to be another independent $(d - c)$ -variate Gaussian together with c components that are in turn independent Binomial(100, 0.5) distributed. We thereby choose parameters such that the Binomial components in H_c have the same mean and variance as the Gaussian components and such that differentiating between Binomial and Gaussian is known to be difficult. Fig. 4 displays two realizations of a Gaussian and Binomial component respectively. We take $d = 200$ and c to be 10% of 200, or $c = 20$.

This problem is difficult; the Binomial and Gaussian components can hardly be differentiated by eye, the contamination level varies and the contamination is only detectable in c out of d components. Moreover, the combination of discrete and continuous components means the optimal kernel choice might not be clear, even with full information. Thus even for 300 observations for each class, no test displays any power until we reach a contamination level of 0.5. However, for higher contamination levels, Fig. 5 clearly displays the superiority of the RF-based tests: None of the kernel tests appear to significantly rise over the level of 5%. On the other hand, the two proposed tests slowly grow from around 0.05 to almost 0.4 in the case of the hypoRF test. Interestingly, while relatively close at first, the difference in power between the Binomial test and the hypoRF grows and is starkest for $\lambda = 1$, again demonstrating the benefit of using the OOB error as a test statistic. Although slightly worse than the hypoRF, the CPT-RF is also clearly beating the Binomial test, highlighting the benefit of using the permutation approach with (in-sample) classification probabilities.



(a) Binomial(100, 0.5) distribution

(b) $N(50, 25)$ distribution

Fig. 4. (Contamination) Illustration of the difference in marginals in the c columns of H_c .

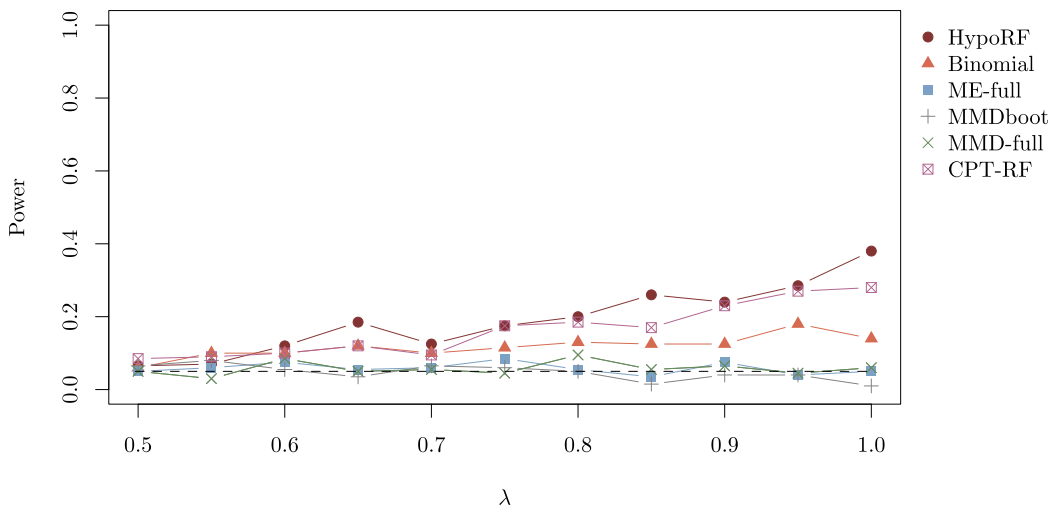


Fig. 5. (Contamination) A point in the figure represents a simulation of size $S = 200$ for a specific test and a $\lambda \in (0.5, 0.55, \dots, 1)$. Each of the $S = 200$ simulation runs we sampled 300 observations from the contaminated distribution with $\lambda \in (0.5, 0.55, \dots, 1)$ and $c = 20$. Likewise 300 observations were sampled from $d = 200$ independent standard normal distributions. The Random Forest used 600 trees and a minimal node size to consider a random split of 4.

Finally, we consider the case $d = c$, so that H_c simply consists of d independent Binomial distributions. The result is displayed in Fig. 6 and all RF-based tests are now extremely strong, while the kernel tests fail to detect any signal.

More simulation examples can be found in Appendix B.

4.3. Real data

As a first application, we consider a high-dimensional microarray data set from Ramey (2016). The data set is about breast cancer, originally provided by Gravier et al. (2010). They examined 168 patients with 2905 gene expressions, each over a five-year period. The 111 patients with no metastasis of small node-negative breast carcinoma after diagnosis were labeled as “good”, and the 57 patients with early metastasis were labeled as “poor”.

The application of the hypoRF to the two groups is summarized in Fig. 7. The test detects a clear difference between the groups “good” and “poor” with “8p23”, “8p21” and “3q25” being the most important (and significant) genes. There seems to be a high correlation between the genes that are located close to each other (especially within the same chromosome). This has the effect that the Random Forest takes a more or less arbitrary choice at a split point between those highly correlated genes. This in turn is reflected in the variable importance measure. For this reason, one should be careful when interpreting the variable importance measure on a gene level. It appears that chromosomes 8 and 3 play an important role in distinguishing the two groups. This finding is in line with Gravier et al. (2010, Figure 2, p. 1129).

In the second example, we are interested in the relative importance of financial risk factors (asset-specific characteristics). We claim that a financial risk factor has explanatory power if it contributes significantly to the classification of individual

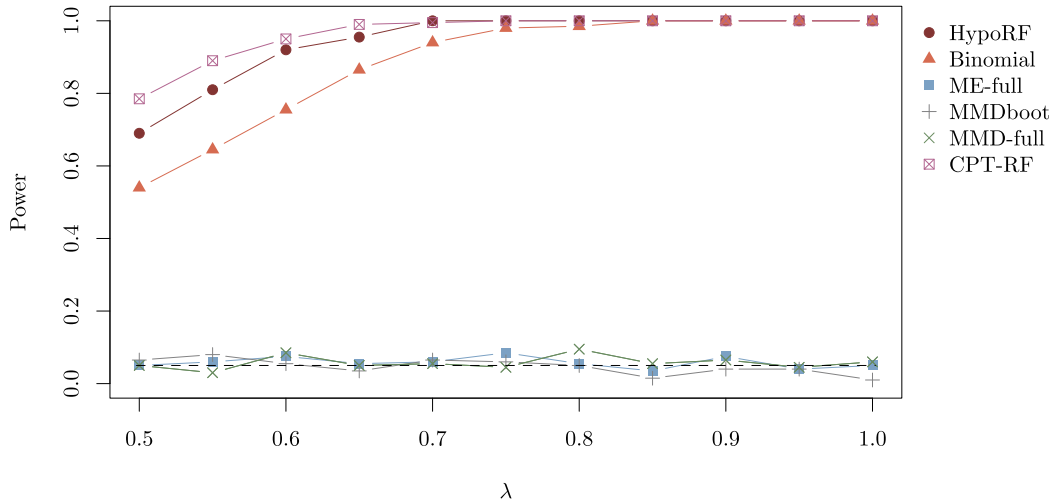


Fig. 6. (Contamination) A point in the figure represents a simulation of size $S = 200$ for a specific test and a $\lambda \in (0.5, 0.55, \dots, 1)$. Each of the $S = 200$ simulation runs we sampled 300 observations from the contaminated distribution with $\lambda \in (0.5, 0.55, \dots, 1)$ and $d = c$. Likewise 300 observations were sampled from $d = 200$ independent standard normal distributions. The Random Forest used 600 trees and a minimal node size to consider a random split of 4.

stock returns above or below the overall median. We use monthly stock return data from the Center for Research in Security Prices (CRSP). Our sample period starts in January 1977 and ends in December 2016, totaling 40 years. Additionally, we obtain the 94 stock-level predictive characteristics used by Gu et al. (2020) from Dacheng Xiu’s webpage - see, <http://dachxiu.chicagobooth.edu>. Between 1977 and 2016 we only use stocks for which we have a full return history. This leads to 501 stocks with 94 stock-specific characteristics. The group “positive” contains stocks and time points for which the return was above the overall median and vice versa for the “negative” group. The two groups are balanced and contain more than 120,000 observations each.

The application of the hypoRF test on the two groups is summarized in Fig. 8. The ordering of the different risk factors is in line with the findings in Gu et al. (2020, Figure 5, p. 34), 1-month momentum being the most important characteristic.

One could argue that stocks that are at time point t close to the overall median are more or less randomly assigned to one of the two groups. Hence, a possible option is to only assign a stock and time point to a certain group if the return is above (below) a certain threshold, i.e., overall median $\pm \epsilon$. However, we observed that the result is robust for different values of ϵ .

5. Discussion

We discussed in this paper two easy-to-use and powerful tests based on Random Forest and empirically demonstrated their efficacy. We presented some consistency and power results and showed a way of adapting the Bayes classifier to obtain a consistent test. This adaptation consisted simply of changing the “cutoff” of the classifier. Especially the test based on the OOB statistics (hypoRF) proved to be powerful and additionally delivered a way to assess the significance of individual variables. This was demonstrated in applications using medical and financial data.

After our first publication on arXiv, Cai et al. (2020) developed an approach based on a smooth transformation of the in-sample probabilities. Interestingly, experiments using their approach with OOB probability estimates, as a hybrid of their and our methodology, delivered promising results. Investigating this further could lead to a further improvement in power for RF-based tests.

Appendix A. Proofs

A.1. Proofs to Section 2

Proposition 2 (Restatement of Proposition 1). *The decision rule in (2) conserves the level asymptotically, i.e.*

$$\limsup_{N_{test} \rightarrow \infty} \mathbb{P}(\delta_B(\hat{g}(D_{N_{test}})) = 1) \leq \alpha,$$

under $H_0 : P = Q$.

Proof. Let $\mathcal{H}_N = \{D_{N_{train}}, n_{1, test}\}$ and assume $P = Q$. Note that, $n_{1, test}, n_{0, test}$ contain the same probabilistic information, so it does not matter which we condition on. We first prove that,

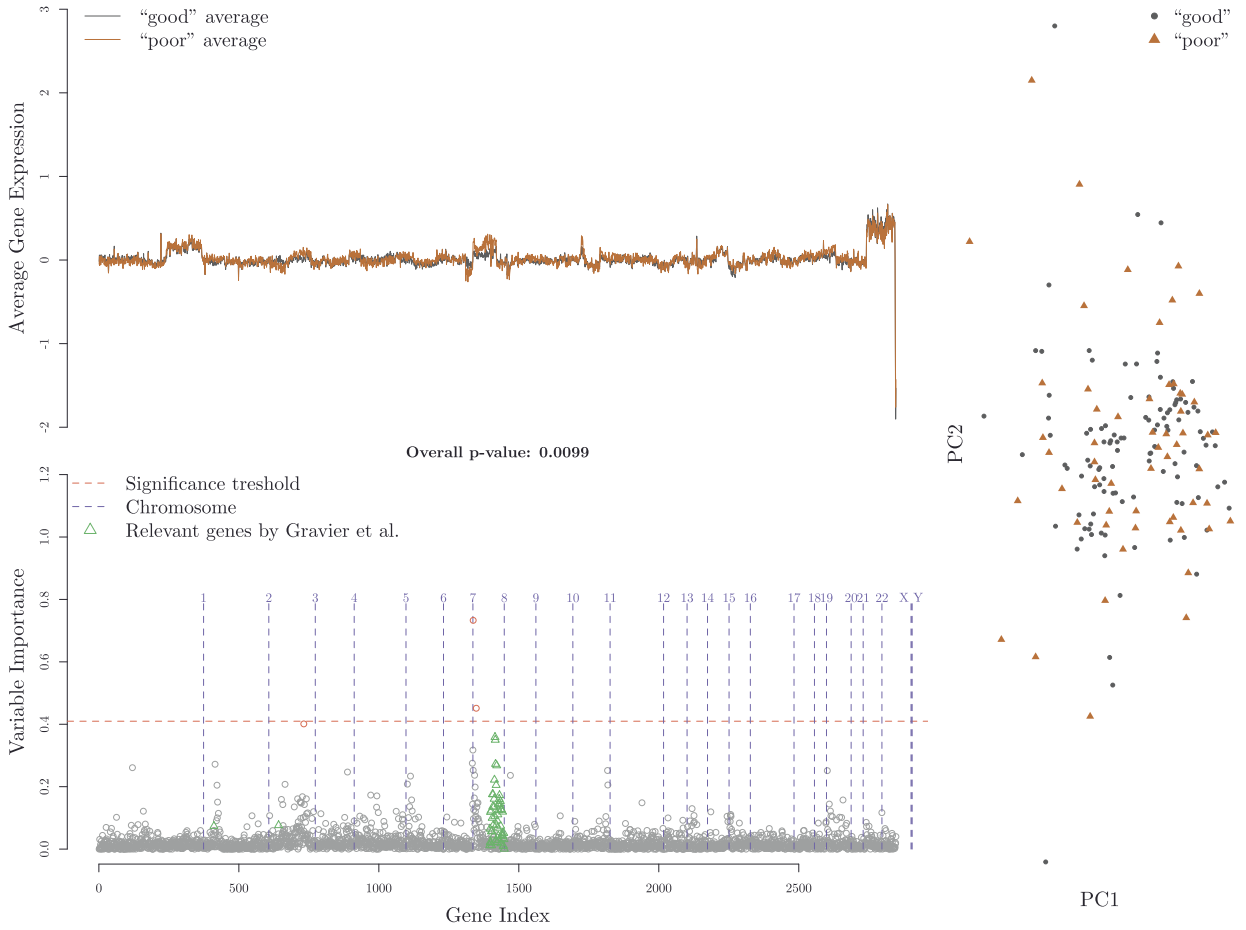


Fig. 7. (Genes) The variable importance (gene importance) combined with the average gene expression is illustrated. The test rejects the null hypothesis that the two groups “good” and “poor” come from the same distribution with a p -value of 0.0099. The 3 most important genes are “8p23”, “8p21” and “3q25” (marked in red). The green triangles represent the important genes reported by Gravier et al. (2010). Additionally, the plot of the first two principal components highlights the fact that there seem to be no obvious clusters. Note: only 15% of the total variance is explained by the first 2 principal components. The Random Forest used 1000 trees and a minimal node size to consider a random split of 4. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

$$n_{j,test} \hat{L}_j^{(\hat{g})} | \mathcal{H}_N \sim \text{Bin}(n_{j,test}, L_j^{(\hat{g})}), \tag{A.1}$$

for $j \in \{0, 1\}$ and $\hat{L}_0^{(\hat{g})}, \hat{L}_1^{(\hat{g})}$ are conditionally independent given $D_{N_{train}}, n_{1,test}$. To prove (A.1) first note that by exchangeability (due to iid sampling),

$$\sum_{i:\ell_i=j} \mathbb{I}\{\hat{g}(\mathbf{Z}_i) \neq \ell_i\} \stackrel{D}{=} \sum_{i=1}^{n_{j,test}} \mathbb{I}\{\hat{g}(\mathbf{Z}_i) \neq j\},$$

$j \in \{0, 1\}$. Conditional on \mathcal{H}_N , the above is a sum of $n_{j,test}$ iid, elements $\mathbb{I}\{\hat{g}(\mathbf{Z}_i) \neq j\}$, with

$$\mathbb{I}\{\hat{g}(\mathbf{Z}_i) \neq j\} | \mathcal{H}_N \sim \text{Bin}(1, \mathbb{P}(\hat{g}(\mathbf{Z}_i) \neq j | \mathcal{H}_N)).$$

Finally, since the event $\hat{g}(\mathbf{Z}_i) \neq j$ is independent of $n_{j,test}$,

$$\begin{aligned} \mathbb{P}(\hat{g}(\mathbf{Z}_i) \neq j | \mathcal{H}_N) &= \mathbb{P}(\hat{g}(\mathbf{Z}_i) \neq j | D_{N_{train}}) \\ &= \mathbb{P}(\hat{g}(\mathbf{Z}_i) \neq j | D_{N_{train}}, \ell_i = j) \\ &= L_j^{(\hat{g})}, \end{aligned}$$

where we again used the independence of ℓ_i and \mathbf{Z}_i under H_0 .

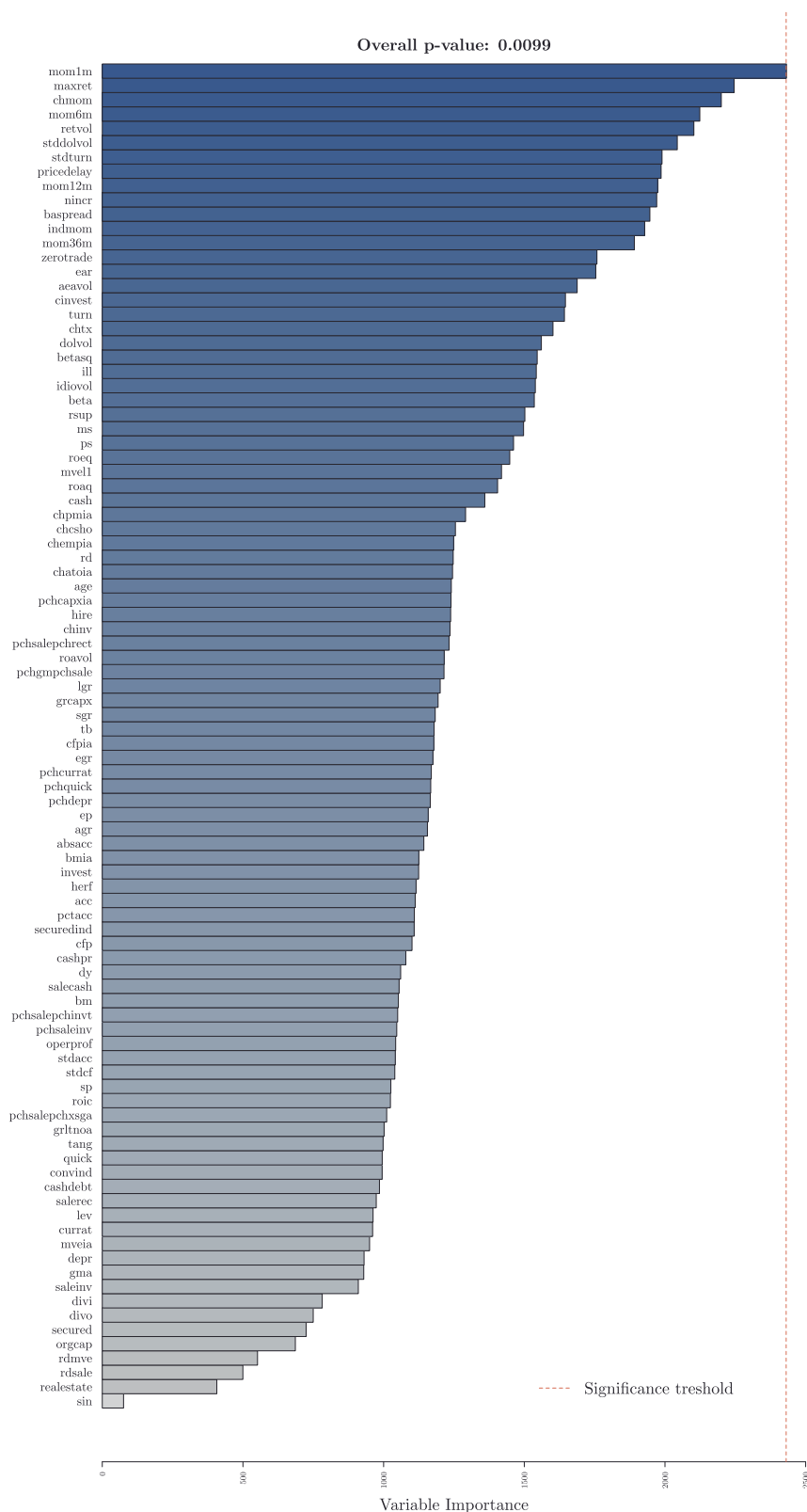


Fig. 8. (Riskfactors) The sorted variable importance of the 94 stock-specific characteristics are illustrated. More information on the 94 characteristics are listed in Table C.1 of Appendix C. The Test rejects with a p -value of almost zero. Nevertheless, the only significant characteristic is the 1-month momentum.

Let $\tilde{\sigma}_c^2 := L_0^{(\hat{g})}(1 - L_0^{(\hat{g})}) + L_1^{(\hat{g})}(1 - L_1^{(\hat{g})})$ and recall that

$$\hat{\sigma}_c^2 = \frac{1}{4} \left(\frac{\hat{L}_0^{(\hat{g})}(1 - \hat{L}_0^{(\hat{g})})}{n_{0,test}} + \frac{\hat{L}_1^{(\hat{g})}(1 - \hat{L}_1^{(\hat{g})})}{n_{1,test}} \right).$$

We additionally define, $\hat{\sigma}_j^2 = \hat{L}_j^{(\hat{g})}(1 - \hat{L}_j^{(\hat{g})})/n_{j,test}$, $j \in \{0, 1\}$. Moreover, set for all N_{test} :

$$\epsilon_{N_{test}} := \epsilon \cdot \frac{1}{N_{test}^\nu},$$

for some $\epsilon > 0$ and $\nu \in (1/2, 1)$. Note that we assume $N_{test} \rightarrow \infty$, while N_{train} might also increase to infinity at any rate, or stay constant. Let for the following

$$E := \left\{ \frac{n_{1,test}}{N_{test}} \rightarrow \pi \right\}.$$

Then $\mathbb{P}(E) = 1$, as $\frac{n_{1,test}}{N_{test}} \rightarrow \pi$ a.s.

First assume for a realized sequence of $D_{N_{train}}$, $N_{test}\tilde{\sigma}_c^2 \rightarrow \infty$ holds. Then for a realized sequence of $n_{1,test}$, with the property that $n_{1,test}/N_{test} \rightarrow \pi$ (i.e. on E), it holds that

$$\limsup_{N_{test} \rightarrow \infty} \mathbb{P}(\delta_B(D_N) = 1 | \mathcal{H}_N) \leq \Phi(\Phi^{-1}(\alpha)) = \alpha.$$

Indeed in this case, conditional on \mathcal{H}_N ,

$$\frac{\hat{L}_{1/2}^{(\hat{g})} - 1/2}{\hat{\sigma}_c} \xrightarrow{D} N(0, 1). \tag{A.2}$$

This is essentially a consequence of the Lindeberg-Feller Central Limit Theorem, but we provide the exact steps now. The key is the following decomposition:

$$\frac{\hat{L}_{1/2}^{(\hat{g})} - 1/2}{\hat{\sigma}_c} = \frac{\hat{\sigma}_0}{2\hat{\sigma}_c} \frac{(\hat{L}_0^{(\hat{g})} - L_0^{(\hat{g})})}{\hat{\sigma}_0} + \frac{\hat{\sigma}_1}{2\hat{\sigma}_c} \frac{(\hat{L}_1^{(\hat{g})} - L_1^{(\hat{g})})}{\hat{\sigma}_1}, \tag{A.3}$$

and

$$\left(\frac{\hat{\sigma}_0}{2\hat{\sigma}_c} \right)^2 + \left(\frac{\hat{\sigma}_1}{2\hat{\sigma}_c} \right)^2 = \frac{\hat{\sigma}_0^2 + \hat{\sigma}_1^2}{4\hat{\sigma}_c^2} = 1, \tag{A.4}$$

by the definitions of $\hat{\sigma}_0$, $\hat{\sigma}_1$, $\hat{\sigma}_c$. If $N_{test}L_0^{(\hat{g})}(1 - L_0^{(\hat{g})}) \rightarrow \infty$ and $N_{test}L_1^{(\hat{g})}(1 - L_1^{(\hat{g})}) \rightarrow \infty$, then it follows directly from the Central Limit Theorem that, conditional on \mathcal{H}_N ,

$$\frac{(\hat{L}_0^{(\hat{g})} - L_0^{(\hat{g})})}{\hat{\sigma}_0} \xrightarrow{D} N(0, 1) \text{ and } \frac{(\hat{L}_1^{(\hat{g})} - L_1^{(\hat{g})})}{\hat{\sigma}_1} \xrightarrow{D} N(0, 1). \tag{A.5}$$

Thus in this case, (A.2) follows immediately from (A.5) combined with (A.3) and (A.4). If only $N_{test}L_1^{(\hat{g})}(1 - L_1^{(\hat{g})}) \rightarrow \infty$, while $N_{test}L_0^{(\hat{g})}(1 - L_0^{(\hat{g})}) \rightarrow \infty$ is not true, then only the asymptotic normality of \hat{L}_1 in (A.5) holds. In this case, the variance is driven by \hat{L}_1 , while the variation in \hat{L}_0 is negligible. More formally, using that under the null $L_1^{(\hat{g})} = 1 - L_0^{(\hat{g})}$, we may write

$$\frac{\hat{L}_{1/2}^{(\hat{g})} - 1/2}{\hat{\sigma}_c} = \left(\frac{\sqrt{n_{1,test}}(\hat{L}_0^{(\hat{g})} - L_0^{(\hat{g})})}{\sqrt{L_1^{(\hat{g})}(1 - L_1^{(\hat{g})})}} + \frac{\sqrt{n_{1,test}}(\hat{L}_1^{(\hat{g})} - L_1^{(\hat{g})})}{\sqrt{L_1^{(\hat{g})}(1 - L_1^{(\hat{g})})}} \right) \frac{\sqrt{L_1^{(\hat{g})}(1 - L_1^{(\hat{g})})}}{\sqrt{n_{1,test}2\hat{\sigma}_c}}. \tag{A.6}$$

Then,

$$\frac{\sqrt{L_1^{(\hat{g})}(1 - L_1^{(\hat{g})})}}{\sqrt{n_{1,test}2\hat{\sigma}_c}} \xrightarrow{p} 1. \tag{A.7}$$

Moreover, for all $\delta > 0$,

$$\begin{aligned} \mathbb{P} \left(\frac{\sqrt{n_{1,\text{test}}}}{\sqrt{L_1^{(\hat{g})}(1-L_1^{(\hat{g})})}} \left| \hat{L}_0^{(\hat{g})} - L_0^{(\hat{g})} \right| > \delta \mid \mathcal{H}_N \right) &\leq \frac{n_{1,\text{test}}}{\delta^2 L_1^{(\hat{g})}(1-L_1^{(\hat{g})})} \mathbb{V}(\hat{L}_0^{(\hat{g})} \mid \mathcal{H}_N) \\ &= \frac{n_{1,\text{test}}}{\delta^2 L_1^{(\hat{g})}(1-L_1^{(\hat{g})})} \frac{L_0^{(\hat{g})}(1-L_0^{(\hat{g})})}{n_{0,\text{test}}} \\ &\approx \frac{L_0^{(\hat{g})}(1-L_0^{(\hat{g})})}{L_1^{(\hat{g})}(1-L_1^{(\hat{g})})}, \end{aligned}$$

on E . Since $N_{\text{test}}L_1^{(\hat{g})}(1-L_1^{(\hat{g})}) \rightarrow \infty$ is still true, this means that

$$\frac{L_0^{(\hat{g})}(1-L_0^{(\hat{g})})}{L_1^{(\hat{g})}(1-L_1^{(\hat{g})})} = \frac{N_{\text{test}}L_0^{(\hat{g})}(1-L_0^{(\hat{g})})}{N_{\text{test}}L_1^{(\hat{g})}(1-L_1^{(\hat{g})})} \rightarrow 0,$$

on E and thus,

$$\frac{\sqrt{n_{1,\text{test}}}(\hat{L}_0^{(\hat{g})} - L_0^{(\hat{g})})}{\sqrt{L_1^{(\hat{g})}(1-L_1^{(\hat{g})})}} \xrightarrow{p} 0. \tag{A.8}$$

Combining the asymptotic normality of $\hat{L}_1^{(\hat{g})}$ as in (A.5), (A.6), (A.7) and (A.8), (A.2) remains true. The same argument can be made analogously if $N_{\text{test}}L_0^{(\hat{g})}(1-L_0^{(\hat{g})}) \rightarrow \infty$, but $N_{\text{test}}L_1^{(\hat{g})}(1-L_1^{(\hat{g})}) \rightarrow \infty$ does not hold. Finally note that $\epsilon_{N_{\text{test}}}$ is of too small order to make a difference in this case, since $\sqrt{N_{\text{test}}}\epsilon_{N_{\text{test}}} \rightarrow 0$.

Now assume that $N_{\text{train}}, D_{N_{\text{train}}}$ are such that $\liminf N_{\text{test}}\hat{\sigma}_c^2 \rightarrow \infty$ does not hold. In this case, using again Markov's inequality,

$$N_{\text{test}}(\hat{L}_{1/2} - 1/2) = O_{\mathbb{P}}(1),$$

conditionally on \mathcal{H}_N , i.e. $\lim_{M \rightarrow \infty} \limsup_{N_{\text{test}}} \mathbb{P}(N_{\text{test}}(\hat{L}_{1/2} - 1/2) > M \mid \mathcal{H}_N) = 0$. Thus,

$$\mathbb{P}(\delta_B(D_N) = 1 \mid \mathcal{H}_N) \leq \mathbb{P}(N_{\text{test}}(\hat{L}_{1/2} - 1/2) > \epsilon \cdot N_{\text{test}}^{1-\nu} \mid \mathcal{H}_N) \rightarrow 0,$$

as $\epsilon \cdot N_{\text{test}}^{1-\nu} \rightarrow \infty$.

Thus we have shown that for a realized sequence of $D_{N_{\text{train}}}, n_{1,\text{test}}$, with the property that $n_{1,\text{test}}/N_{\text{test}} \rightarrow \pi$, it holds that

$$\limsup_{N_{\text{test}} \rightarrow \infty} \mathbb{P}(\delta_B(D_N) = 1 \mid \mathcal{H}_N) \leq \alpha.$$

On the other hand,

$$\begin{aligned} \limsup_{N_{\text{test}} \rightarrow \infty} \mathbb{P}(\delta_B(D_N) = 1) &= \limsup_{N_{\text{test}} \rightarrow \infty} \mathbb{E}[\mathbb{P}(\delta_B(D_N) = 1 \mid \mathcal{H}_N) \mathbb{I}_E] \\ &\leq \mathbb{E} \left[\limsup_{N_{\text{test}} \rightarrow \infty} \mathbb{P}(\delta_B(D_N) = 1 \mid \mathcal{H}_N) \mathbb{I}_E \right] \\ &\leq \alpha. \quad \square \end{aligned}$$

Lemma 6 (Restatement of Lemma 1). Take $\mathcal{X} \subset \mathbb{R}$ and $\pi \neq 1/2$. Then no decision rule of the form, $\delta(D_N) = \delta(g_{1/2}^*(D_N))$ is consistent.

Proof. We first show that if $\pi \neq \frac{1}{2}$, one can construct $(P, Q) \in \Theta_1$ that the Bayes classifier is not able to differentiate. Consider $\pi > 1/2$, $d = 1$ and Q being the uniform distribution on $(0, 1)$, with density $q(z) = \mathbb{I}\{z \in (0, 1)\}$. We write $q = \mathbb{I}(0, 1)$ for short. P is a mixture of Q and another uniform on $R \subset (0, 1)$, so that

$$p = (1 - \alpha)\mathbb{I}(0, 1) + \alpha \frac{\mathbb{I}R}{|R|}.$$

Giving Q a label of 1 and P a label of 0 when observing $(1 - \pi)P + \pi Q$, and taking $|R| = 1/2$, the Bayes classifier is then given as $g_{1/2}^*(z) = \mathbb{I}\{\eta(z) > 1/2\}$, where

$$\eta(z) := \begin{cases} \pi / (\pi + (1 - \pi)(1 + \alpha)), & \text{if } z \in R \\ \pi / (\pi + (1 - \pi)(1 - \alpha)), & \text{if } z \notin R \end{cases}.$$

Simple algebra shows that for any $0 < \alpha < \min(\pi / (1 - \pi) - 1, 1)$, $\eta(z) > 1/2$ and thus $g_{1/2}^*(z) = 1$ for all $z \in (0, 1)$. In particular, $L_0^{(g_{1/2}^*)} = 1$ and $L_1^{(g_{1/2}^*)} = 0$, such that $L_0^{(g_{1/2}^*)} + L_1^{(g_{1/2}^*)} = 1$ and $L_\pi^{(g_{1/2}^*)} = 1 - \pi = \min(\pi, 1 - \pi)$.

On the other hand, for any $\theta_0 \in \Theta_0$, $\eta(z) = \mathbb{E}[\ell|z] = \pi > 1/2$, which shows that $g_{1/2}^*(z) = 1$ for all z . Consequently, for $\theta_1 = (P, Q)$ in the above example and $\theta_0 \in \Theta_0$ arbitrary, it holds that

$$\mathbb{E}_{\theta_0}[f(g_{1/2}^*(D_N))] = \mathbb{E}_{\theta_1}[f(g_{1/2}^*(D_N))],$$

for any bounded measurable function $f : \{0, 1\}^N \rightarrow \mathbb{R}$. In particular, since the test conserves the level by assumption, $\phi(\theta_1) = \phi(\theta_0) \leq \alpha$ and the test has no power.

□

Lemma 7 (Restatement of Lemma 2). *The classifier*

$$g_\pi^*(\mathbf{z}) = \mathbb{I} \{ \eta(\mathbf{z}) > \pi \}, \tag{A.9}$$

is a solution to (9). Moreover it holds that

$$1 - TV(P, Q) = L_0^{g_\pi^*} + L_1^{g_\pi^*}, \tag{A.10}$$

for any $\pi \in (0, 1)$.

Proof. If Relation (A.10) is true, it immediately follows that g_π^* is a solution to (9). Indeed, let $h_\#P$ be the push-forward measure of P through a measurable function $h : \mathcal{X} \rightarrow \mathbb{R}$. Taking $h = g$, for an arbitrary classifier g , it holds that

$$\begin{aligned} 1 - (L_0^{g_\pi^*} + L_1^{g_\pi^*}) &= TV(P, Q) \\ &\geq P(g(\mathbf{X}) = 0) - Q(g(\mathbf{Y}) = 0) \\ &= \mathbb{P}(g(\mathbf{Z}) = 0 | \ell = 0) - \mathbb{P}(g(\mathbf{Z}) = 0 | \ell = 1) \\ &= 1 - (L_0^g + L_1^g), \end{aligned}$$

where the first inequality follows, because $\{\mathbf{x} : g(\mathbf{x}) = 0\}$ and $\{\mathbf{y} : g(\mathbf{y}) = 0\}$ are two Borel sets on \mathcal{X} . Consequently, it also holds for any classifier g that

$$L_{1/2}^g = \frac{1}{2}(L_0^g + L_1^g) \geq \frac{1}{2}(L_0^{g_\pi^*} + L_1^{g_\pi^*}) = L_{1/2}^{g_\pi^*}.$$

It remains to prove (A.10): It is well-known that (one of) the sets attaining the maximum in the definition of $TV(P, Q)$ is given by $A^* := \{\mathbf{z} : q(\mathbf{z}) \leq p(\mathbf{z})\}$. It is possible to rewrite A^* :

$$\begin{aligned} A^* &= \left\{ \mathbf{z} : \frac{\pi q(\mathbf{z})}{(1 - \pi)p(\mathbf{z}) + \pi q(\mathbf{z})} \leq \frac{\pi}{1 - \pi} \frac{(1 - \pi)p(\mathbf{z})}{(1 - \pi)p(\mathbf{z}) + \pi q(\mathbf{z})} \right\} \\ &= \left\{ \mathbf{z} : \eta(\mathbf{z}) \leq \frac{\pi}{1 - \pi} (1 - \eta(\mathbf{z})) \right\} \\ &= \{\mathbf{z} : \eta(\mathbf{z}) \leq \pi\}. \end{aligned}$$

Thus

$$\begin{aligned} TV(P, Q) &= P(A^*) - Q(A^*) = \mathbb{P}(\eta(\mathbf{Z}) \leq \pi | \ell = 0) - \mathbb{P}(\eta(\mathbf{Z}) \leq \pi | \ell = 1) \\ &= 1 - \mathbb{P}(\eta(\mathbf{Z}) > \pi | \ell = 0) - \mathbb{P}(\eta(\mathbf{Z}) \leq \pi | \ell = 1) \\ &= 1 - (\mathbb{P}(\eta(\mathbf{Z}) > \pi | \ell = 0) + \mathbb{P}(\eta(\mathbf{Z}) \leq \pi | \ell = 1)) \\ &= 1 - (L_0^{g_\pi^*} + L_1^{g_\pi^*}). \quad \square \end{aligned}$$

Corollary 3 (Restatement of Corollary 1). *The decision rule $\delta_B(g_\pi^*(D_N))$ in (2) is consistent for any $\pi \in (0, 1)$.*

Proof. We restate here the decision rule in (2) for completeness,

$$\delta_B(g_{\pi}^*(D_N)) = \mathbb{I} \left\{ \hat{L}_{1/2}^{(g_{\pi}^*)} - 1/2 < \hat{\sigma}_c \Phi^{-1}(\alpha) + \epsilon_N \right\},$$

since $N_{test} = N$.

First we show that the decision rule conserves the level, for $\epsilon_N = 0$ for all N . Since, for any P, Q , $P = Q$, $\eta(z) = \pi$, $\hat{L}_0^{(g_{\pi}^*)} = 0$ and $\hat{L}_1^{(g_{\pi}^*)} = 1$ a.s., so that for all $\theta_0 \in \Theta_0$ and any sample size,

$$\phi(\theta_0) = \mathbb{P}_{\theta_0}(\hat{L}_{1/2}^{(g_{\pi}^*)} < 1/2) = 0.$$

Thus in particular $\sup_{\Theta_0} \phi(\theta_0) = 0 \leq \alpha$.

Assume $\theta \in \Theta_1$, so that $TV(P, Q) > 0$. We assume first that also $TV(P, Q) < 1$. Since now the classifier itself does not need to be estimated, it holds that

$$N_j \hat{L}_j^{(g_{\pi}^*)} | N_j \sim \text{Bin}(N_j, L_j^{(g_{\pi}^*)}).$$

Since $1 > TV(P, Q) > 0$, $0 < L_0^{(g_{\pi}^*)} + L_1^{(g_{\pi}^*)} < 1$, so that $N L_j^{(g_{\pi}^*)} (1 - L_j^{(g_{\pi}^*)}) \rightarrow \infty$ for $j = 0$ and $j = 1$. Conditional on any sequence of N_0, N_1 , such that $N_0 \rightarrow \infty$ and $N_1 \rightarrow \infty$, as $N \rightarrow \infty$,

$$\sqrt{N_0}(\hat{L}_0^{(g_{\pi}^*)} - L_0^{(g_{\pi}^*)}) \xrightarrow{D} N(0, L_0^{(g_{\pi}^*)}(1 - L_0^{(g_{\pi}^*)})) \text{ and } \sqrt{N_1}(\hat{L}_1^{(g_{\pi}^*)} - L_1^{(g_{\pi}^*)}) \xrightarrow{D} N(0, L_1^{(g_{\pi}^*)}(1 - L_1^{(g_{\pi}^*)})),$$

and since $\hat{L}_0^{(g_{\pi}^*)}, \hat{L}_1^{(g_{\pi}^*)}$ are conditionally independent, it holds that

$$\frac{\hat{L}_{1/2}^{(g_{\pi}^*)} - L_{1/2}^{(g_{\pi}^*)}}{1/2 \sqrt{\frac{\hat{L}_0^{(g_{\pi}^*)}(1 - \hat{L}_0^{(g_{\pi}^*)})}{N_0} + \frac{\hat{L}_1^{(g_{\pi}^*)}(1 - \hat{L}_1^{(g_{\pi}^*)})}{N_1}}} = \frac{\hat{L}_{1/2}^{(g_{\pi}^*)} - L_{1/2}^{(g_{\pi}^*)}}{\hat{\sigma}_c} \xrightarrow{D} N(0, 1),$$

as in Proposition 1. Consequently,

$$\mathbb{P} \left(\frac{\hat{L}_{1/2}^{(g_{\pi}^*)} - 1/2}{\hat{\sigma}_c} < \Phi^{-1}(\alpha) | N_0 \right) = \mathbb{P} \left(\frac{\hat{L}_{1/2}^{(g_{\pi}^*)} - L_{1/2}^{(g_{\pi}^*)}}{\hat{\sigma}_c} < \Phi^{-1}(\alpha) - \frac{L_{1/2}^{(g_{\pi}^*)} - 1/2}{\hat{\sigma}_c} | N_0 \right).$$

Now for any realized sequence of N_0, N_1 such that $N_0 \rightarrow \infty$ and $N_1 \rightarrow \infty$, as $N \rightarrow \infty$, this probability goes to 1, since $L_{1/2}^{(g_{\pi}^*)} - 1/2 < 0$ and $\hat{\sigma}_c \rightarrow 0$. Since $N_1/N \rightarrow \pi$, a.s., and $N_0 = N - N_1$, this will be true for almost all sequences. Thus applying dominated convergence to the above conditional result, one sees that

$$\mathbb{P} \left(\frac{\hat{L}_{1/2}^{(g_{\pi}^*)} - 1/2}{\hat{\sigma}_c} < \Phi^{-1}(\alpha) \right) \rightarrow 1.$$

If $TV(P, Q) = 1$ on the other hand, $L_{1/2}^{(g_{\pi}^*)} = 0$ and $\hat{\sigma}_c = 0$ a.s. and trivially the rejection probability becomes

$$\mathbb{P}(\hat{L}_{1/2}^{(g_{\pi}^*)} < 1/2) = 1. \quad \square$$

A.2. Proofs to Section 3

Lemma 8 (Restatement of Lemma 3). $\mathbb{E}[h_{N_{train}}((\ell_1, \mathbf{Z}_{N_{train}}), \dots, (\ell_{N_{train}}, \mathbf{Z}_{N_{train}}))] = \mathbb{E}[L_{1/2}^{(\hat{g}_{-i})}]$.

Proof. First we note that

$$\mathbb{E}[L_{1/2}^{(\hat{g}_{-i})}] = \frac{1}{2} \left(\mathbb{P}(\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i | \ell_i = 1) + \mathbb{P}(\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i | \ell_i = 0) \right).$$

Let $B(i) \leq B$ be the number of classifiers in the ensemble, not containing observation i . Since we assume that each classifier in the ensemble receives a bootstrapped version of $D_{N_{train}}$, there is a probability $p > 0$, that any given classifier \hat{g}_{v_b} will not contain observation i . Since this bootstrapping is done independently for each classifier, we have that $B(i) \sim \text{Bin}(p, B)$. Thus as $B \rightarrow \infty$, also $B(i) \rightarrow \infty$ a.s. and thus $\hat{g}_{-i}(\mathbf{Z}) = \mathbb{E}_v[\hat{g}_v(D_{N_{train}}^{-i})(\mathbf{Z})]$, or

$$\begin{aligned} \mathbb{E}[\epsilon_i^{ob}] &= \mathbb{E}[\mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i\} \left(\frac{1 - \ell_i}{n_{0,train}} + \frac{\ell_i}{n_{1,train}} \right)] \\ &= \mathbb{E} \left[\mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i\} \frac{1 - \ell_i}{n_{0,train}} \right] + \mathbb{E} \left[\mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i\} \frac{\ell_i}{n_{1,train}} \right]. \end{aligned}$$

Now, since $\ell_i = \mathbb{I}\{\ell_i = 1\}$, it holds that

$$\begin{aligned} \mathbb{E} \left[\mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i\} \frac{\ell_i}{n_{1,train}} \right] &= \mathbb{E} \left[\frac{1}{n_{1,train}} \cdot \mathbb{P}(\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i | \ell_i = 1, n_{1,train}) \right] \\ &= \mathbb{E} \left[\frac{\mathbb{P}(\ell_i = 1 | n_{1,train})}{n_{1,train}} \cdot \mathbb{P}(\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i, | n_{1,train}, \ell_i = 1) \right] \\ &= \mathbb{E} \left[\frac{\mathbb{P}(\ell_i = 1 | n_{1,train})}{n_{1,train}} \right] \cdot \mathbb{P}(\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i | \ell_i = 1), \end{aligned} \tag{A.11}$$

since the event $\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i$ is independent of $n_{1,train}$ given the event $\ell_i = 1$. Finally,

$$\begin{aligned} \mathbb{E} \left[\frac{\mathbb{P}(\ell_i = 1 | n_{1,train})}{n_{1,train}} \right] &= \frac{1}{N_{train}} \mathbb{E} \left[\sum_{i=1}^{N_{train}} \frac{\mathbb{P}(\ell_i = 1 | n_{1,train})}{n_{1,train}} \right] \\ &= \frac{1}{N_{train}} \mathbb{E} \left[\mathbb{E} \left[\frac{1}{n_{1,train}} \sum_{i=1}^{N_{train}} \ell_i \middle| n_{1,train} \right] \right] \\ &= \frac{1}{N_{train}}. \end{aligned} \tag{A.12}$$

Combining (A.11) and (A.12), we obtain:

$$\mathbb{E} \left[\mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i\} \frac{\ell_i}{n_{1,train}} \right] = \frac{1}{N_{train}} \mathbb{P}(\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i, | \ell_i = 1).$$

Similarly,

$$\mathbb{E} \left[\mathbb{I}\{\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i\} \frac{1 - \ell_i}{n_{0,train}} \right] = \frac{1}{N_{train}} \mathbb{P}(\hat{g}_{-i}(\mathbf{Z}_i) \neq \ell_i, | \ell_i = 0).$$

Thus indeed,

$$\mathbb{E}[h_{N_{train}}((\ell_1, \mathbf{Z}_{N_{train}}), \dots, (\ell_{N_{train}}, \mathbf{Z}_{N_{train}}))] = N_{train} \mathbb{E}[\varepsilon_1^{oob}] = \mathbb{E}[L_{1/2}^{\hat{g}_{-i}}]. \quad \square$$

Lemma 9. $h_{N_{train}}$ is a valid kernel for the expectation $\mathbb{E}[L_{1/2}^{\hat{g}_{-i}}]$.

Proof. Unbiasedness was proven above. Symmetry follows, since for any two permutations σ_1, σ_2 , there exists i, j such that $\sigma_1(j) = \sigma_2(i) := u$, and thus

$$\begin{aligned} \varepsilon_{\sigma_1(i)}^{oob} &= \mathbb{E}[\mathbb{I}\{g(\mathbf{Z}_{\sigma_1(i)}, D_{N_{train}}^{-\sigma_1(i)}, \theta) \neq \ell_{\sigma_1(i)}\} \left(\frac{1 - \ell_{\sigma_1(i)}}{n_{0,train}} + \frac{\ell_{\sigma_1(i)}}{n_{1,train}} \right) | D_{N_{train}}^{\sigma_1}] \\ &= \mathbb{E}[\mathbb{I}\{g(\mathbf{Z}_u, D_{N_{train}}^{-u}, \theta) \neq \ell_u\} \left(\frac{1 - \ell_u}{n_{0,train}} + \frac{\ell_u}{n_{1,train}} \right) | D_{N_{train}}^{\sigma_1}] \\ &= \mathbb{E}[\mathbb{I}\{g(\mathbf{Z}_u, D_{N_{train}}^{-u}, \theta) \neq \ell_u\} \left(\frac{1 - \ell_u}{n_{0,train}} + \frac{\ell_u}{n_{1,train}} \right) | D_{N_{train}}^{\sigma_2}] \\ &= \varepsilon_{\sigma_2(j)}^{oob}, \end{aligned}$$

where $D_{N_{train}}^{\sigma_s} = (\mathbf{Z}_{\sigma_s(1)}, \ell_{\sigma_s(1)}), \dots, (\mathbf{Z}_{\sigma_s(N_{train})}, \ell_{\sigma_s(N_{train})})$, $s \in \{1, 2\}$. But that means the sum in (12) does not change. \square

We also need a well-known auxiliary result:

Lemma 10. Let $(\xi_N)_N, \xi$ be an arbitrary sequence of random variables. If every subsequence has a subsequence such that $\xi_{N(k(l))} \xrightarrow{D} \xi$, then $\xi_N \xrightarrow{D} \xi$.

Theorem 2 (Restatement of Theorem 1). Assume that for $N \rightarrow \infty$, $N_{train} = N_{train}(N) \rightarrow \infty$ and $K = K(N) \rightarrow \infty$,

$$\lim_N \frac{KN_{train}^2}{N} \frac{\zeta_{1,N_{train}}}{\zeta_{N_{train},N_{train}}} = 0, \tag{A.13}$$

$$\lim_N \frac{\sqrt{KN_{train}}}{N} = 0. \tag{A.14}$$

Then,

$$\frac{\sqrt{K}(\hat{U}_{N,K} - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}])}{\sqrt{\zeta_{N_{train},N_{train}}}} \xrightarrow{D} N(0, 1). \tag{A.15}$$

Proof. Let for the following $\xi_i = (\mathbf{Z}_i, \ell_i)$ for brevity and consider the complete U-statistics

$$\hat{U}_N := \frac{1}{\binom{N}{N_{train}}} \sum h_{N_{train}}(\xi_{i_1}, \dots, \xi_{i_{N_{train}}}), \tag{A.16}$$

where the sum is taken over all $\binom{N}{N_{train}}$ possible subsets of size $N_{train} \leq N$ from $\{1, \dots, N\}$. From the ‘‘H-Decomposition’’, see e.g., Lee (1990), the variance of \hat{U}_N can be bounded as,

$$\begin{aligned} \mathbb{V}(\hat{U}_N) &\leq \frac{N_{train}^2}{N} \zeta_{1,N_{train}} + \frac{N_{train}^2}{N^2} \mathbb{V}(h) \\ &\leq \frac{N_{train}^2}{N} \zeta_{1,N_{train}} + \frac{N_{train}^2}{N^2} \zeta_{N_{train},N_{train}}, \end{aligned}$$

see also Wager and Athey (2017, Lemma 7). Thus it holds for all $\varepsilon > 0$ that

$$\begin{aligned} \mathbb{P}\left(\frac{\sqrt{K}|\hat{U}_N - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}]|}{\sqrt{\zeta_{N_{train},N_{train}}}} > \varepsilon\right) &\leq \frac{K\mathbb{V}(\hat{U}_N)}{\varepsilon^2 \zeta_{N_{train},N_{train}}} \\ &= \frac{1}{\varepsilon^2} \left(\frac{KN_{train}^2}{N} \frac{\zeta_{1,N_{train}}}{\zeta_{N_{train},N_{train}}} + \frac{KN_{train}^2}{N^2} \right) \\ &\rightarrow 0, \end{aligned}$$

by (A.13) and (A.14).

We now use the idea of Lee (1990, Lemma A) to prove (17): As in Mentch and Hooker (2016), we denote by $\mathcal{S}_{N,N_{train}} = \{S_j : j = 1, \dots, \binom{N}{N_{train}}\}$ the set of all possible subsamples of size N_{train} sampled without replacement. Let $M_{N,N_{train}} = (M_{S_1}, \dots, M_{S_{\binom{N}{N_{train}}}})$ be the number of times each subsample appears when sampling K times. Then $M_{N,N_{train}} | \xi_1, \xi_2, \dots$ is multinomial distributed. Thus

$$\begin{aligned} \frac{\sqrt{K}(\hat{U}_{N,K} - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}])}{\sqrt{\zeta_{N_{train},N_{train}}}} &\stackrel{D}{=} \sqrt{K}^{-1} \left(\sum_{i=1}^{\binom{N}{N_{train}}} M_{S_i} (h_{N_{train}}(S_i) - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}]) \right) / \sqrt{\zeta_{N_{train},N_{train}}} \\ &\stackrel{D}{=} \frac{1}{\sqrt{\zeta_{N_{train},N_{train}}} \sqrt{K}} \left(\sum_{i=1}^{\binom{N}{N_{train}}} \frac{K}{\binom{N}{N_{train}}} (h_{N_{train}}(S_i) - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}]) \right) + \\ &\quad \frac{1}{\sqrt{\zeta_{N_{train},N_{train}}} \sqrt{K}} \left(\sum_{i=1}^{\binom{N}{N_{train}}} (M_{S_i} - \frac{K}{\binom{N}{N_{train}}}) (h_{N_{train}}(S_i) - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}]) \right) \\ &\stackrel{D}{=} \frac{\sqrt{K}(\hat{U}_N - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}])}{\sqrt{\zeta_{N_{train},N_{train}}}} + \\ &\quad \sqrt{K} \left(\frac{1}{K} \sum_{i=1}^{\binom{N}{N_{train}}} (M_{S_i} - \frac{K}{\binom{N}{N_{train}}}) \frac{(h_{N_{train}}(S_i) - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}])}{\sqrt{\zeta_{N_{train},N_{train}}}} \right). \end{aligned} \tag{A.17}$$

Let $a_i = (h_{N_{train}}(S_i) - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}]) / \sqrt{\zeta_{N_{train}, N_{train}}}$, as in Lee (1990). Then

$$\hat{U}_{N,2} = \binom{N}{N_{train}}^{-1} \sum_{i=1}^{(N, N_{train})} a_i^2,$$

is again a U-statistics with $\mathbb{E}[\hat{U}_{N,2}] = 1$ and

$$\begin{aligned} \mathbb{P}(|\hat{U}_{N,2} - 1| > \varepsilon) &\leq \frac{1}{\varepsilon^2} \frac{N_{train}^2}{N} \mathbb{V}(\mathbb{E}[(h_{N_{train}}(S_i) - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}])^2 | \xi_1]) + \frac{N_{train}^2}{N^2} \\ &= o\left(\frac{N_{train}}{N}\right) \\ &= o(K^{-1/2}), \end{aligned}$$

using Lemma 5. Thus, $\hat{U}_{N,2} \xrightarrow{P} 1$ and this will be true for any given subsequence as well. Similarly,

$$\binom{N}{N_{train}}^{-1} \sum_{i=1}^{(N, N_{train})} a_i \leq \frac{\sqrt{K}(\hat{U}_N - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}])}{\sqrt{\zeta_{N_{train}, N_{train}}}} \xrightarrow{P} 0.$$

For each given subsequence we can thus choose a further subsequence such that $\hat{U}_{N,2} \xrightarrow{a.s.} 1$, as well as $\sqrt{K}(\hat{U}_N - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}]) / \sqrt{\zeta_{N_{train}, N_{train}}} \xrightarrow{a.s.} 0$. Then it follows from (A.17) and the same characteristic function arguments as in Lee (1990, Lemma A) that,

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathbb{E}[\exp(it\sqrt{K}(\hat{U}_{N,K} - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}]) / \sqrt{\zeta_{N_{train}, N_{train}}})] &= \\ \lim_{N \rightarrow \infty} \mathbb{E}\left[\exp\left(it\sqrt{K}(\hat{U}_N - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}]) / \sqrt{\zeta_{N_{train}, N_{train}}}\right)\right] \cdot \exp\left(-\frac{t^2}{2}\right) &= \\ = \exp\left(-\frac{t^2}{2}\right), \end{aligned}$$

where we suppressed the dependence on the chosen subsequence. Thus the subsequence converges in distribution to $N(0, 1)$ and by Lemma 10, so does the overall sequence. \square

Corollary 4 (Restatement of Corollary 2). Assume the conditions of Theorem 1 hold true and that $\hat{\zeta}_{N_{train}, N_{train}} / \zeta_{N_{train}, N_{train}} \xrightarrow{P} 1$. Then the decision rule in (19) conserves the level asymptotically and has approximate power

$$\Phi\left(\Phi^{-1}(\alpha) + \sqrt{\frac{K}{\zeta_{N_{train}, N_{train}}}}(1/2 - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}])\right). \tag{A.18}$$

Proof. From Theorem 1 and the assumption that $\hat{\zeta}_{N_{train}, N_{train}}$ is a consistent estimator, it follows that

$$\frac{\sqrt{K}(\hat{U}_{N,K} - \mathbb{E}[L_{1/2}^{\hat{g}^{-1}}])}{\sqrt{\hat{\zeta}_{N_{train}, N_{train}}}} \xrightarrow{D} N(0, 1).$$

In particular, under H_0 , as $\mathbb{E}[L_{1/2}^{\hat{g}^{-1}}] = 1/2$:

$$\frac{\sqrt{K}(\hat{U}_{N,K} - 1/2)}{\sqrt{\hat{\zeta}_{N_{train}, N_{train}}}} \xrightarrow{D} N(0, 1),$$

so that the decision rule (19) attains the right level asymptotically. Moreover, under the alternative, for $t^* := \Phi^{-1}(\alpha)$,

$$\mathbb{P}\left(\frac{\sqrt{K}(\hat{U}_{N,K} - 1/2)}{\sqrt{\hat{\zeta}_{N_{train}, N_{train}}}} < t^*\right)$$

$$\begin{aligned}
 &= \mathbb{P} \left(\frac{\sqrt{K}(\hat{U}_{N,K} - \mathbb{E}[L_{1/2}^{\hat{g}^{-i}}])}{\sqrt{\hat{\zeta}_{N_{train}, N_{train}}}} < t^* - \frac{\sqrt{K}(\mathbb{E}[L_{1/2}^{\hat{g}^{-i}}] - 1/2)}{\sqrt{\hat{\zeta}_{N_{train}, N_{train}}}} \right) \\
 &= \Phi \left(t^* + \frac{\sqrt{K}(1/2 - \mathbb{E}[L_{1/2}^{\hat{g}^{-i}}])}{\sqrt{\hat{\zeta}_{N_{train}, N_{train}}}} \right) + o_{\mathbb{P}}(1). \quad \square
 \end{aligned}$$

Appendix B. Further simulations

Additional simulation examples can be found in the next three subsections.

B.1. Gaussian mean shift

The classical and most prominent example of two-sample testing is the detection of a mean-shifts between two Gaussians. That is, we assume $\mathbb{P}_X = N(\boldsymbol{\mu}_1, I_{d \times d})$ and $\mathbb{P}_Y = N(\boldsymbol{\mu}_2, I_{d \times d})$ so that the testing problem reduces to

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ vs } H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

We will implement this by simply taking $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + (\delta/\sqrt{d}) \cdot \mathbf{1}$, for some $\delta \in \mathbb{R}$.

It appears clear that our test should not be the first to choose here. For d much smaller than n , the optimal test would be given by Hotelling’s test (Hotelling, 1931). For d approaching and even superseding n , the MMD with a Gaussian kernel, or an LDA classifier as in Kim et al. (2021), might be the logical next choice. For this reason, we also included the LDA classifier in this example. For all the other examples, the simulated power of the LDA two-sample test is always no better than the level - as expected. Allowing the trees in the forest to grow fully, i.e., setting the minimum node size to a low number like 1, one observes a type of overfitting of the Random Forest. Thus we would expect our test to be beaten at least by MMDboot. Surprisingly this does not happen: As can be seen in Fig. B.1, all the RF-based tests display an impressive amount of power, where our hypoRF test is the strongest in all the provided mean shift scenarios. The Binomial test is even stronger than MMDboot and LDA, which seems surprising given the known strong performance of the MMD and LDA in this situation. The hypoRF test on the other hand towers above all others, together with MMD-full. In fact, the hypoRF and Binomial test almost appear to give respectively an upper and lower bound for the MMD-full in this example. Aside from the impressive power of our tests, it is also interesting to note the difference between MMD-full and MMDboot. While this seems not surprising, given that MMD-full is essentially the optimized version of MMDboot, we will see in subsequent examples that their power ranking is often reversed.

To make the example more interesting, one might ask what happens if the mean shift is not present in all of the d components, but only in $c < d$ of them? This was noted to be a difficult problem in Chwialkowski et al. (2015). We therefore study a “sparse” case $c = 2$ (1% out of $d = 200$) and a “moderately sparse” case $c = 20$ (10% out of $d = 200$), now considering $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + (\delta/\sqrt{c}) \cdot \mathbf{1}$. Note that there is some advantage here, as we now scale δ only by a factor of $\sqrt{c} < \sqrt{d}$. Thus, if a test is able to detect the sparse changes well, it should display a higher power than before. Indeed as seen in Fig. B.2, the performance of the kernel tests is remarkably stable (given the randomness inherent in the simulation), when changing from $c = d = 200$ to $c = 20$ to $c = 2$. On the other hand, the performance of the RF-based tests appears to increase. Thus the odds only shift in favor of our tests and the test of Cai et al. (2020): For $c = 20$ the optimized MMD, MMD-full, is still competitive, though MMDboot, ME-full, and LDA fall further behind. While the hypoRF, the CPT-RF and the fully optimized MMD test reach a power of close to 1, the remaining kernel tests and LDA stay below 0.7. The Binomial test, on the other hand, displays almost the same performance as MMD-full, ending with a power of a bit over 0.8. Its performance is amplified in the sparse case, in which the Binomial, CPT-RF and hypoRF test beat the other tests by a large margin. The power of both tests quickly increases from around 0.05 to 1, as δ passes from 0.2 to 1. While the performance of the Binomial test is impressive, the hypoRF test manages to pick up the nuanced changes even faster, at times almost doubling the power of the Binomial test. Though the price to pay for this is a much higher computational effort.

It should be said that both the sparse and moderately sparse case here are tailor-made for a RF-based classifier; not only are the changes only appearing in a few components, but they appear marginally and are thus easy to detect in the splitting process of the trees. Nonetheless, it seems surprising how strong the tests perform. We will now turn to more complex examples, where changes in the marginals alone are not as easy, or even impossible to detect.

B.2. Changing the dependency structure

The previous example focused only on cases where the changes in distribution can be observed marginally. For these examples, it would in principle be enough to compare the marginal distributions to detect the difference between Q and P . An interesting class of problems arises when we instead leave the marginal distribution unchanged but change the dependency structure when moving from P to Q . We will hereafter study two examples; the first one concerning a simple change from a multivariate Gaussian with independent components to one with nonzero correlation. The second one again

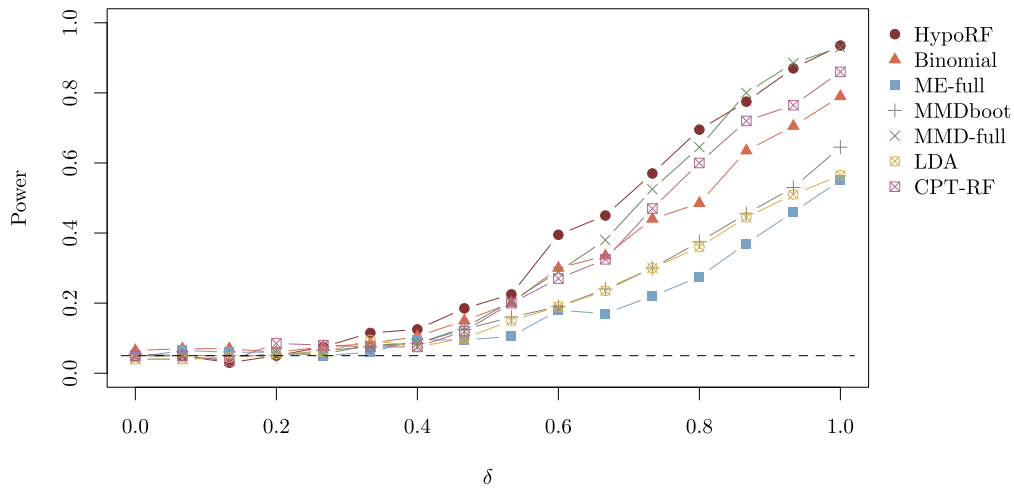


Fig. B.1. (Mean Shift) A point in the figure represents a simulation of size $S = 200$ for a specific test and a $\delta \in (0, 0.0667, 0.1334, 0.2, \dots, 1)$. Each of the $S = 200$ simulation runs we sampled 300 observations from a $d = 200$ dimensional multivariate normal distribution with a mean shift of $\frac{\delta}{\sqrt{d}}$ and likewise $n = 300$ observations from $d = 200$ independent standard normal distributions. The Random Forest used 600 trees and a minimal node size to consider a random split of 4.

takes P to have independent Gaussian components, but induces a more complex dependence structure on Q , via a t -copula. Thus for what follows, we set $P = N(0, I_{d \times d})$.

First, consider $Q = N(0, \Sigma)$, where Σ is some positive definite correlation matrix. As for any d there are potentially $d(d - 1)/2$ unique correlation coefficients in this matrix, the number of possible specifications is enormous even for small d . For simplicity, we only consider a single correlation number ρ , which we either use (I) in all $d(d - 1)/2$ or (II) in only $c < d(d - 1)/2$ cases.

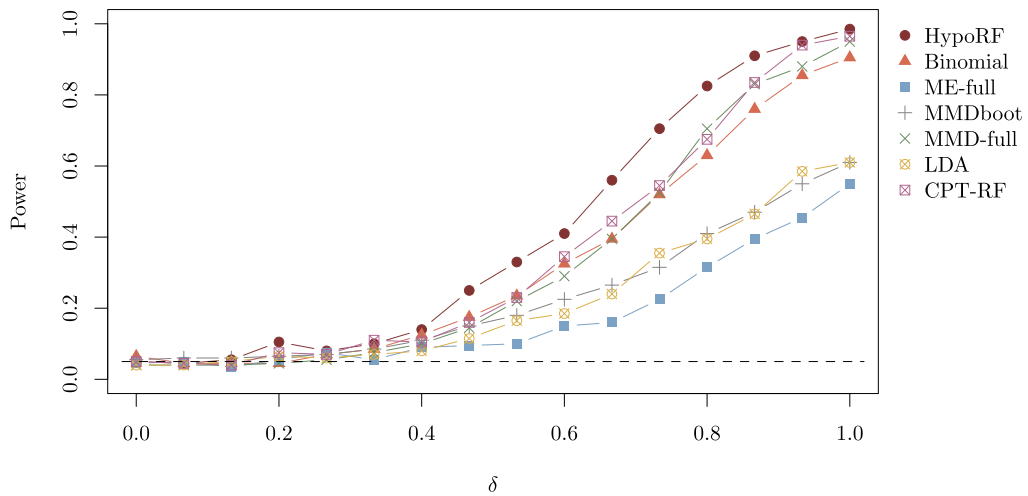
Fig. B.3 displays the result of case (I). Now the superiority of our hypoRF test is challenged, though it manages to at least hold its own against MMD-full and ME-full. The roles of MMD-full and MMD are also reversed, the latter now displaying a much higher power, that in fact dwarfs the power of all other tests. MMD-full displays together with the Binomial test the smallest amount of power, both apparently suffering from the decrease in sample size. ME-full on the other hand, which suffers the same drawback, manages to have a strong performance, on par with the hypoRF. This is all the more impressive, keeping in mind that the ME is a test that scales linearly in N . Case (II) can be seen in Fig. B.4. Again the resulting “sparsity” is beneficial for our test, with the hypoRF now being on par with the powerful MMD test, and with ME-full only slightly above the Binomial test.

In the second example, we study a change in dependence, which is more interesting than the simple change of the covariance matrix. In particular, Q is now given by a distribution that has standard Gaussian marginals bound together by a t -copula, see e.g., Demarta and McNeil (2005) or McNeil et al. (2015, Chapter 5). While the density and cdf of the resulting distribution Q are relatively complicated, it is simple and insightful to simulate from this distribution, as described in Demarta and McNeil (2005): Let $x \mapsto t_\nu(x)$ denote the cdf of a univariate t -distribution with ν degrees of freedom, and $T_\nu(R)$ the multivariate t -distribution with dispersion matrix R and ν degrees of freedom. We first simulate from a multivariate t -distribution with dispersion matrix R and degrees of freedom ν , to obtain $\mathbf{T} \sim T_\nu(R)$. In the second step, simply set $\mathbf{Y} := (\Phi^{-1}(t_\nu(T_1)), \dots, \Phi^{-1}(t_\nu(T_p)))^T$. We denote $Q = T_\Phi(\nu, R)$. What kind of dependency structure does \mathbf{Y} have? It is well known that $\mathbf{T} \sim t_\nu(R)$ has

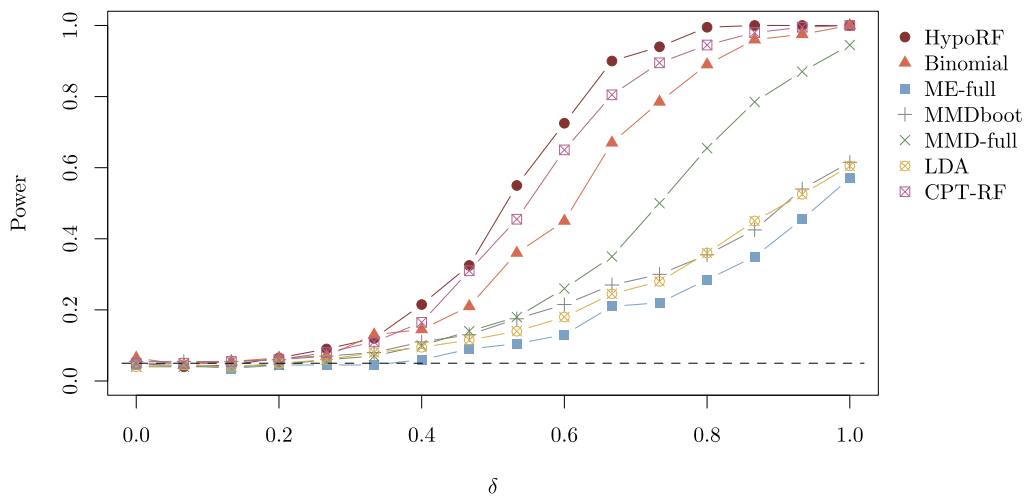
$$\mathbf{T} \stackrel{D}{=} G^{-1/2} \mathbf{N},$$

with $\mathbf{N} \sim N(0, R)$ and $G \sim \text{Gamma}(\nu/2, \nu/2)$ independent of \mathbf{N} . As such, the dependence induced in \mathbf{T} , and therefore in Q , is dictated through the mutual latent random variable G . It persists, even if $R = I_{d \times d}$ and induces more complex dependencies than mere correlation. These dependencies are moreover stronger, the smaller ν , though this effect is hard to quantify. One reason this dependency structure is particularly interesting in our case is that it spans more than two columns, contrary to correlation which is an inherent bivariate property. We again study the case (I) with all d components tied together by the t -copula, and (II) only the first $c = 20 < d$ components having a t -copula dependency, while the remaining $d - c = 180$ columns are again independent $N(0, 1)$.

The results for case (I) are shown in Fig. B.5. Now our tests, together with ME-full cannot compete with CPT-RF, MMD and MMD-full. However for the ME-full, this again depends on the chosen hyperparameters, for some settings ME-full was as good as MMD-full. Though there appears to be no clear way how to determine this. Both MMD-based tests manage to stay at almost one, even for $\nu = 8$, which seems to be an extremely impressive feat. The CPT-RF test falls behind the two MMD-based tests, but has still an impressively high power, compared to our hypoRF test. Our best test, on the other hand,



(a) $c = 20$, moderately sparse case.



(b) $c = 2$ sparse case.

Fig. B.2. (Mean Shift) A point in the figures represents a simulation of size $S = 200$ for a specific test and a $\delta \in (0, 0.125, 0.25, \dots, 1)$. Each of the $S = 200$ simulation runs we sampled $n = 300$ observations from a $d = 200$ dimensional multivariate Gaussian distribution, where c columns have a shift in mean of $\frac{\delta}{\sqrt{c}}$ and likewise $n = 300$ observations from $d = 200$ independent standard normal distributions. The Random Forest used 600 trees and a minimal node size to consider a random split of 4.

loses power quickly for $\nu > 4$, while the Binomial test does so even for $\nu > 2$. The results for case (II) shown in Fig. B.6, are similarly insightful. Given the difficulty of this problem, it is not surprising that almost all of the tests fail to have any power for $\nu > 3$. The exception is once again the MMD, performing incredibly strong up to $\nu = 5$. The performance of MMDboot is not only interesting in that it beats our tests, but also in how it beats all other kernel approaches in the same way. In particular, MMD-full stands no chance, which again is likely, in part, due to the reduced sample size the MMDboot has available for testing. Though hard to generalize, it appears from this analysis that a complex, rather weak dependence, is a job best done by the plain MMDboot.

B.3. Multivariate Blob

A well-known difficult example is the “Gaussian Blob”, an example where “the main data variation does not reflect the difference between P and Q ” (Gretton et al., 2012), see e.g., Gretton et al. (2012) and Jitkrittum et al. (2016). We study here the following generalization of this idea: Let $T \in \mathbb{N}$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_t)_{t=1}^T$, $\boldsymbol{\mu}_t \in \mathbb{R}^d$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_t)_{t=1}^T$, with $\boldsymbol{\Sigma}_t$ a positive definite $d \times d$ matrix. We consider the mixture

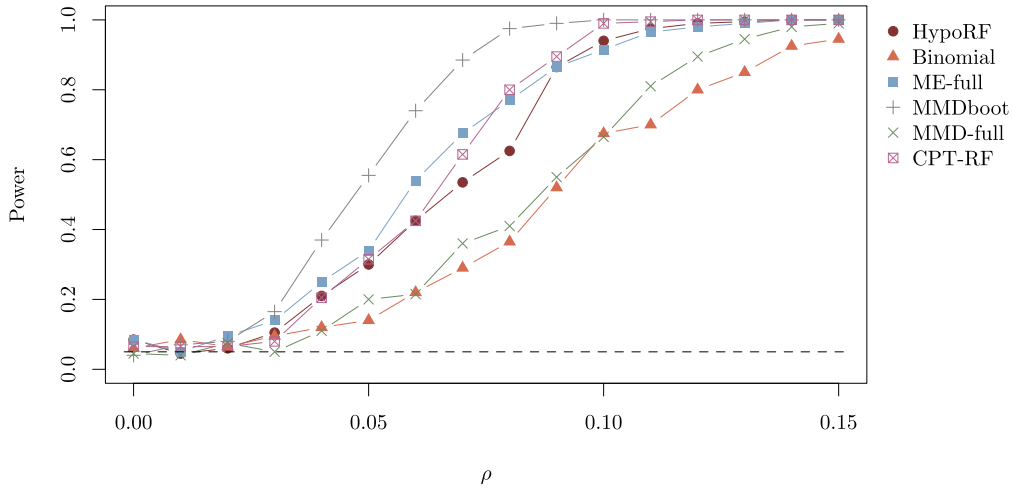


Fig. B.3. (Dependency) A point in the figure represents a simulation of size $S = 200$ for a specific test and a $\rho \in (0, 0.01, 0.02, \dots, 0.15)$. Each of the $S = 200$ simulation runs we sampled 300 observations from a $d = 60$ dimensional multivariate normal distribution with $\rho \in (0, 0.01, 0.02, \dots, 0.15)$, representing Q . Likewise 300 observations were sampled from a $d = 60$ dimensional multivariate normal distribution using $\rho = 0$, representing P . The Random Forest used 600 trees and a minimal node size to consider a random split of 4.

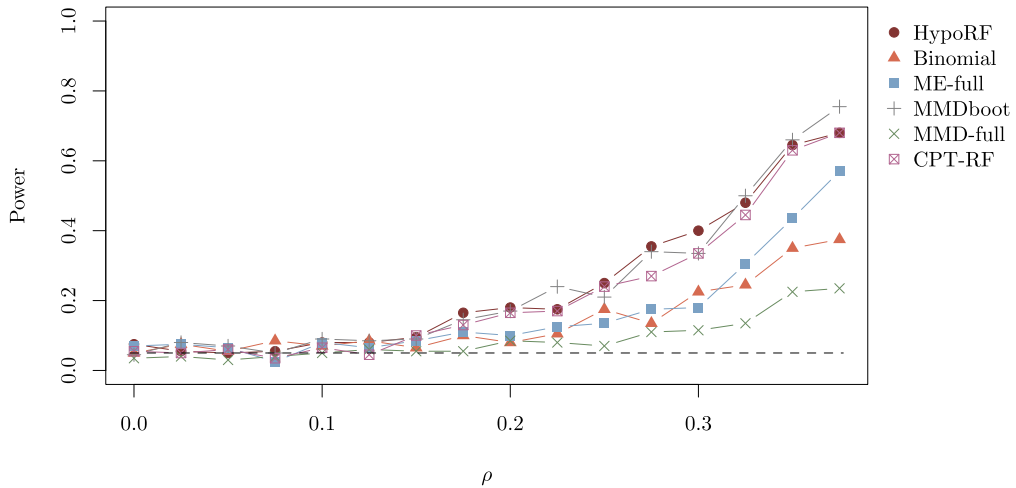


Fig. B.4. (Dependency) A point in the figure represents a simulation of size $S = 200$ for a specific test and a $\rho \in (0, 0.025, 0.05, \dots, 0.375)$. Each of the $S = 200$ simulation runs we sampled 300 observations from a $d = 10$ dimensional multivariate normal distribution with $c = 4$ values in the correlation matrix equal to $\rho \in (0, 0.025, 0.05, \dots, 0.375)$, representing Q . Likewise 300 observations were sampled from a multivariate normal distribution using $\rho = 0$, representing P . The Random Forest used 600 trees and a minimal node size to consider a random split of 4.

$$N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \sum_{t=1}^T \frac{1}{T} N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t).$$

For $\boldsymbol{\mu}$, we will always use a baseline vector of size d , we say, and include in $\boldsymbol{\mu}$ all possible enumerations of choosing d elements from $w \in \mathbb{R}^d$ with replacement. This gives a total number of $T = c^d$ possibilities and each $\boldsymbol{\mu}_t \in \mathbb{R}^d$ is one possible such enumeration. For example, if $c = d = 2$ and $w = (1, 2)$ then we may set $\boldsymbol{\mu}_1 = (1, 1)$, $\boldsymbol{\mu}_2 = (2, 2)$, $\boldsymbol{\mu}_3 = (1, 2)$, $\boldsymbol{\mu}_4 = (2, 1)$. We will refer to each element of this mixture as a “Blob” and study two experiments where we change the covariance matrices $\boldsymbol{\Sigma}_t$ of the blobs when changing from P to Q , i.e.,

$$P = N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_X), \quad Q = N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y).$$

Obviously it quickly gets infeasible to simulate from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, as with increasing d the number of blobs explodes. Though, as shown below, this difficulty can be circumvented when $\boldsymbol{\Sigma}_t$ is diagonal for all t . The example also considerably worsens the curse of dimensionality, as even for small d the numbers of observations in each Blob is likely to be very small. Thus for 300 observations, we have a rather difficult example at hand.

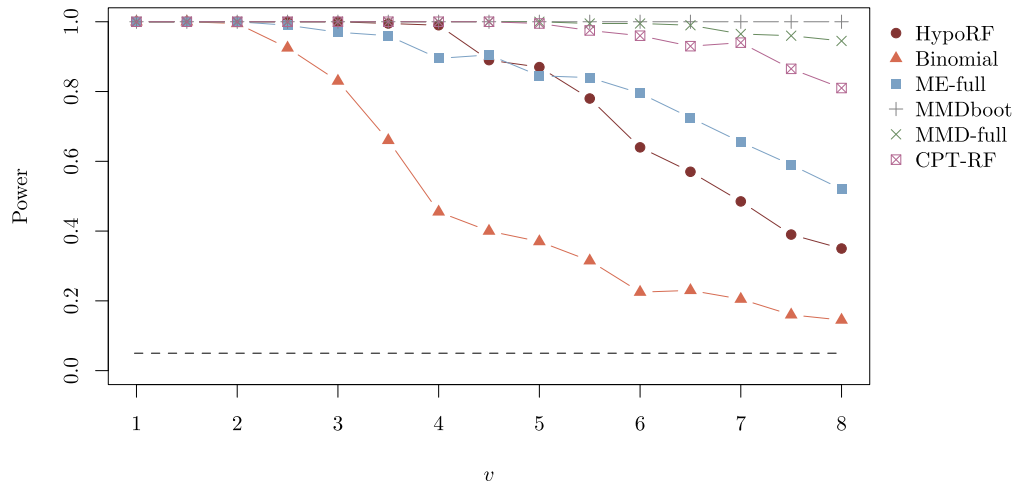


Fig. B.5. (Dependency) A point in the figure represents a simulation of size $S = 200$ for a specific test and a $v \in (1, 1.5, \dots, 8)$. Each of the $S = 200$ simulation runs we sampled 300 observations from the Student-t Copula with $R = I_{d \times d}$, $v \in (1, 1.5, \dots, 8)$ and $d = 60$ standard normally distributed margins and likewise 300 observations from the multivariate normal. The Random Forest used 600 trees and a minimal node size to consider a random split of 4.

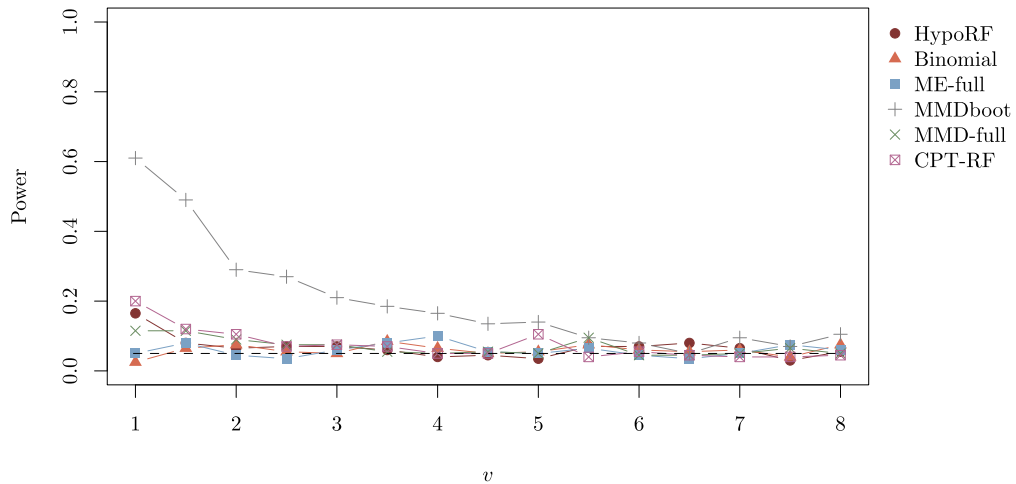


Fig. B.6. (Dependency) A point in the figure represents a simulation of size $S = 200$ for a specific test and a $v \in (1, 1.5, \dots, 8)$. Each of the $S = 200$ simulation runs we sampled 300 observations from a $d - c = 180$ dimensional multivariate Gaussian distribution and a $d = 20$ dimensional Student-t Copula with $R = I_{d \times d}$, $v \in (1, 1.5, \dots, 8)$ and standard normally distributed margins, representing Q . Likewise 300 observations were sampled from a multivariate normal distribution, representing P . The Random Forest used 600 trees and a minimal node size to consider a random split of 4.

Table B.1

(Blob) Power for different N , d and number of Blobs. Each power was calculated with a simulation of size $S = 500$ for a specific test.

N	d	Blobs	ME-full	MMD	MMD-full	Binomial	hypoRF
600	2	2 ²	0.056	0.054	0.072	0.204	0.306
600	2	3 ²	0.064	0.048	0.070	0.070	0.190
600	3	2 ³	0.052	0.040	0.060	0.088	0.116
600	3	3 ³	0.056	0.060	0.060	0.064	0.084

We will subsequently study two experiments. The first one takes $w = (1, 2, 3)$, $\Sigma_{1,X} = \Sigma_{2,X} = \dots = \Sigma_{t,X} = I_{d \times d}$ and $\Sigma_{1,Y} = \Sigma_{2,Y} = \dots = \Sigma_{t,Y} = \Sigma$ to be a correlation matrix with nonzero elements on the off-diagonal. In particular, we generate Σ randomly at the beginning of the S trials for a given d , such that (1) it is a positive definite correlation matrix and (2) it has a ratio of minimal to maximal eigenvalue of at most $1 - 1/\sqrt{d}$. For $d = 2$, this corresponds to the original Blob example as in Gretton et al. (2012), albeit with a less strict bound on the eigenvalue ratio. The resulting distribution for $d = 1$ and $d = 2$ is plotted in Fig. B.7.

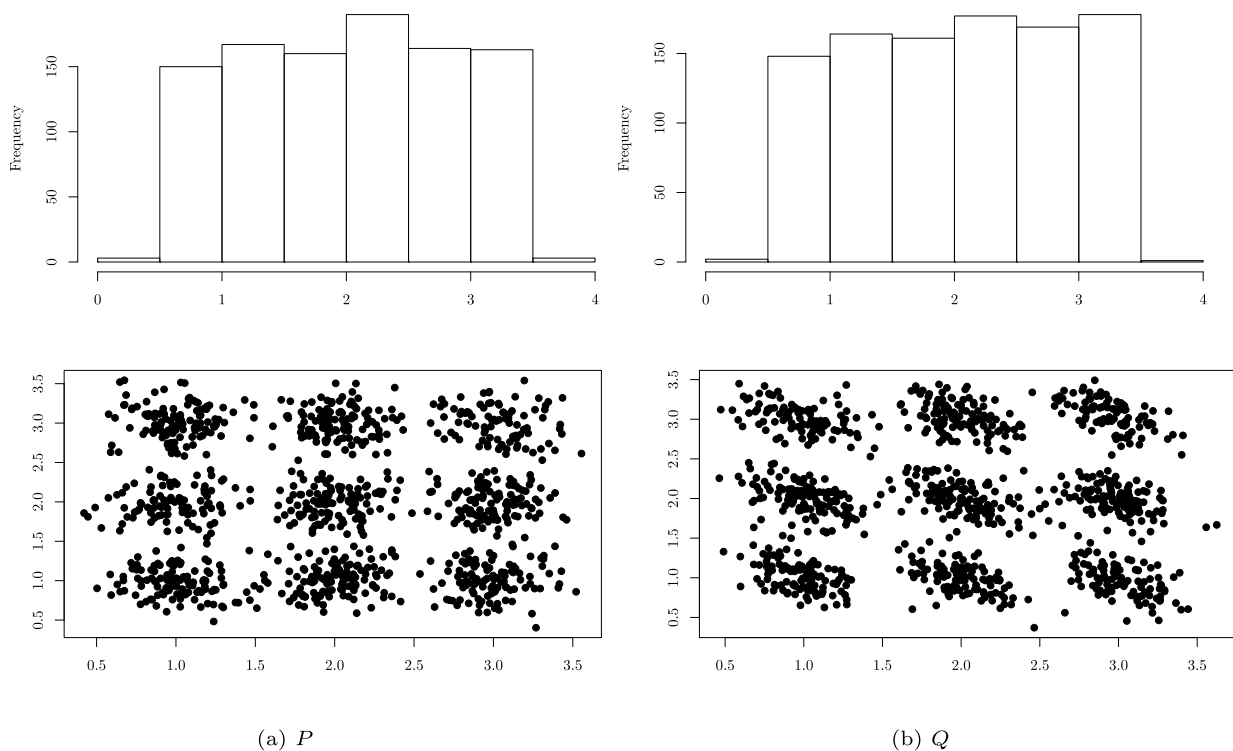


Fig. B.7. (Blob) Illustration of the original Blob example. Below: Illustration for $d = 2$. Above: First marginals of P and Q respectively.

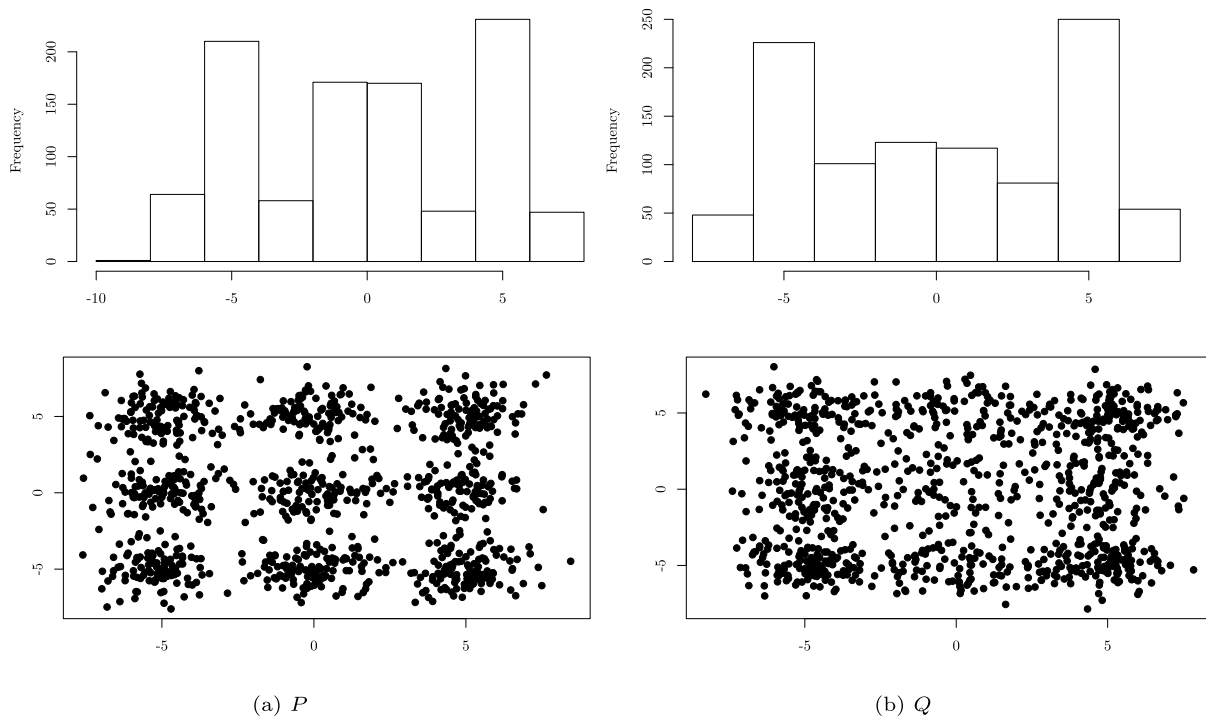


Fig. B.8. (Blob) Illustration of the second Blob example. Below: Illustration for $d = 2$. Above: First marginals of P and Q respectively.

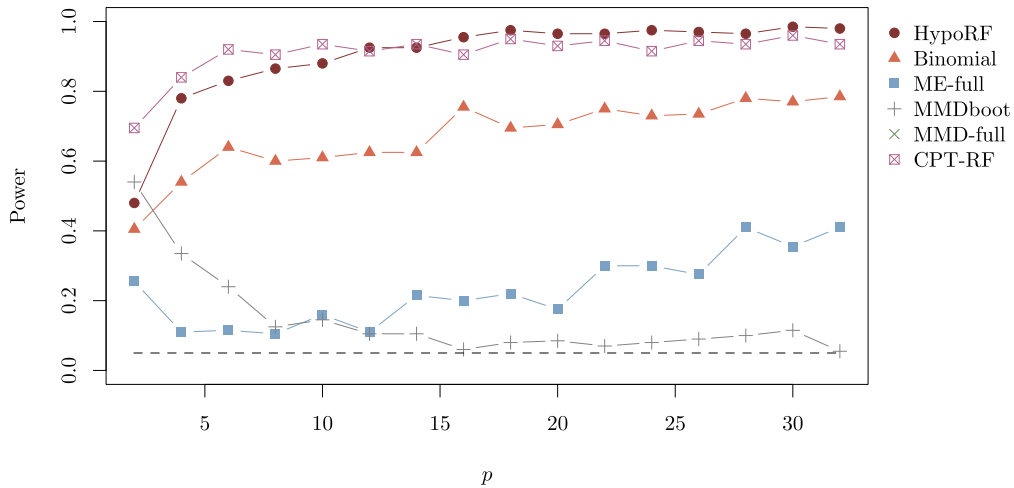


Fig. B.9. (Blob) A point in the figure represents a simulation of size $S = 200$ for a specific test and a $d \in (2, 4, 6, 8, 10, 20, 40, 80, 120, 200)$. Each of the $S = 200$ simulation runs we sampled 300 observations from $N(\mu, \Sigma_X)$ and likewise 300 observations from $N(\mu, \Sigma_Y)$. The Random Forest used 600 trees and a minimal node size to consider a random split of 4.

Table B.1 displays the result of the experiment with our usual set-up and a variation of $d = 2, 3$ and the number of blobs being 2^d and 3^d . Surprisingly our hypoRF test is the only one displaying notable power throughout the example. MMD and MMD-full are not able to detect any difference between the distribution with this sample size. Interestingly, the ME which we would have expected to work well in this example is also only at the level. However, this again depends on the specification chosen for the hyperparameters of the optimization. For another parametrization, we obtained a power of 0.116 for $d = 2$, $blobs = 2^2$ and 0.082 for $d = 2$ and $blobs = 3^2$, all other values being on the level.

The second experiment takes $w = (-5, 0, 5)$ and for all t , $\Sigma_{t,X}$, $\Sigma_{t,Y}$ to be diagonal and generated similarly to μ . That is, we take $\Sigma_{t,X} = \text{diag}(\sigma_{t,X}^2)$, where each $\sigma_{t,X}$ is a vector including d draws with replacement from a base vector $v_X \in \mathbb{R}^d$, and analogously with $\Sigma_{t,Y}$. In this case, it is possible to rewrite P and Q , as

$$P = \prod_{j=1}^d P_X \text{ and } Q = \prod_{j=1}^d P_Y,$$

with

$$P_X = \frac{1}{3}N(w_1, v_{1,X}^2) + \frac{1}{3}N(w_2, v_{2,X}^2) + \frac{1}{3}N(w_3, v_{3,X}^2),$$

and

$$P_Y = \frac{1}{3}N(w_1, v_{1,Y}^2) + \frac{1}{3}N(w_2, v_{2,Y}^2) + \frac{1}{3}N(w_3, v_{3,Y}^2).$$

As such, it is feasible to simulate from P and Q , even for large d , by simply simulating d times from P_X and P_Y . We consider $w = (-5, 0, 5)$ and the standard deviations

$$\begin{aligned} (v_{1,X}, v_{2,X}, v_{3,X}) &= (1, 1, 1), \\ (v_{1,Y}, v_{2,Y}, v_{3,Y}) &= (1, 2, 1). \end{aligned}$$

The change between the distributions is subtle even in notation; only the standard deviation of the middle mixture component is changed from 1 to 2. This has the effect that the middle component gets spread out more, causing it to melt into the other two. The resulting distribution for $d = 1$ and $d = 2$ is plotted in Fig. B.8. Unsurprisingly, P looks quite similar as in Fig. B.7. The marginal plots ($d = 1$) appear to be very different, though this is only an effect of having centers $(-5, 0, 5)$ instead of $(1, 2, 3)$. On the other hand, while not clearly visible, it can be seen that the different blobs of Q display different behavior in variance; every Blob in positions $(2, 1)$, $(2, 2)$, $(2, 3)$, $(1, 2)$, $(3, 2)$ on the 3×3 grid has its variance increased.

The results of the simulations are seen in Fig. B.9. The Binomial, CPT-RF and hypoRF test display a power quickly increasing with dimensions, regardless of the decreasing number of observations in each Blob. This also holds true, to a smaller degree, for the ME-full, which due to its location optimization appears to be able to adapt to the problem structure. However, its power considerably lacks behind the RF-based tests. In contrast, the behavior of the MMD-based tests quickly deteriorates as the number of samples per Blob decreases. Indeed from a kernel perspective, all points have more or less the same distance from each other, whether they are coming from P or Q . Thus the extreme power of the MMD to detect

“joint” changes in the structure of the data (i.e., dependency changes) cements its downfall here, as it is unable to detect the marginal difference.

This example might appear rather strange; it has a flavor of a mathematical counterexample, simple or even nonsensical on the outset, but proving an important point: While the differences between P and Q are obvious to the naked eye if only one marginal each is plotted with a histogram, the example manages to completely fool the kernel tests (under a Gaussian kernel at least). As such it is not only a demonstration of the merits of our test but also a way of fooling general kernel tests. It might be interesting to find real-world applications, where such data structure is likely.

Appendix C. Financial riskfactors

Table C.1

(Riskfactors) This table lists the 94 financial characteristics we use in Section 4.3. We obtain the characteristics used by Gu et al. (2020) from Dacheng Xiu's webpage; see <http://dachxiu.chicagobooth.edu>. Note that the data is collected in Green et al. (2017).

No.	Acronym	Firm characteristic	Frequency	Literature
1	absacc	Absolute accruals	Annual	Bandyopadhyay et al. (2010)
2	acc	Working capital accruals	Annual	Sloan (1996)
3	aeavol	Abnormal earnings announcement volume	Quarterly	Lerman et al. (2008)
4	age	Years since first Compustat coverage	Annual	Jiang et al. (2005)
5	agr	Asset growth	Annual	Cooper et al. (2008)
6	baspread	Bid-ask spread	Monthly	Amihud and Mendelson (1989)
7	beta	Beta	Monthly	Fama and MacBeth (1973)
8	betasq	Beta squared	Monthly	Fama and MacBeth (1973)
9	bm	Book-to-market	Annual	Rosenberg et al. (1985)
10	bmia	Industry-adjusted book-to-market	Annual	Asness et al. (2000)
11	cash	Cash holdings	Quarterly	Palazzo (2012)
12	cashdebt	Cash flow to debt	Annual	Ou and Penman (1989)
13	cashpr	Cash productivity	Annual	Chandrashekar and Rao (2009)
14	cfp	Cash flow to price ratio	Annual	Desai et al. (2004)
15	cfpia	Industry-adjusted cash flow to price ratio	Annual	Asness et al. (2000)
16	chatoia	Industry-adjusted change in asset turnover	Annual	Soliman (2008)
17	chcsho	Change in shares outstanding	Annual	Pontiff and Woodgate (2008)
18	chempia	Industry-adjusted change in employees	Annual	Asness et al. (2000)
19	chinv	Change in inventory	Annual	Thomas and Zhang (2002)
20	chmom	Change in 6-month momentum	Monthly	Gettleman and Marks (2006)
21	chpmia	Industry-adjusted change in profit margin	Annual	Soliman (2008)
22	chtx	Change in tax expense	Quarterly	Thomas and Zhang (2011)
23	cinvest	Corporate investment	Quarterly	Titman et al. (2004)
24	convind	Convertible debt indicator	Annual	Valta (2016)
25	currat	Current ratio	Annual	Ou and Penman (1989)
26	depr	Depreciation / PP&E	Annual	Holthausen and Larcker (1992)
27	divi	Dividend initiation	Annual	Michaely et al. (1995)
28	divo	Dividend omission	Annual	Michaely et al. (1995)
29	dolvol	Dollar trading volume	Monthly	Chordia et al. (2001)
30	dy	Dividend to price	Annual	Litzenberger and Ramaswamy (1982)
31	ear	Earnings announcement return	Quarterly	Kishore et al. (2008)
32	egr	Growth in common shareholder equity	Annual	Richardson et al. (2005)
33	ep	Earnings to price	Annual	Basu (1977)
34	gma	Gross profitability	Annual	Novy-Marx (2013)
35	grcapx	Growth in capital expenditures	Annual	Anderson and Garcia-Feijóo (2006)
36	grltnoa	Growth in long term net operating assets	Annual	Fairfield et al. (2003)
37	herf	Industry sales concentration	Annual	Hou and Robinson (2006)
38	hire	Employee growth rate	Annual	Belo et al. (2014)
39	idiovol	Idiosyncratic return volatility	Monthly	Ali et al. (2003)
40	ill	Illiquidity	Monthly	Amihud (2002)
41	indmom	Industry momentum	Monthly	Moskowitz and Grinblatt (1999)
42	invest	Capital expenditures and inventory	Annual	Moskowitz and Grinblatt (2010)
43	lev	Leverage	Annual	Bhandari (1988)
44	lgr	Growth in long-term debt	Annual	Richardson et al. (2005)
45	maxret	Maximum daily return	Monthly	Bali et al. (2011)
46	mom12m	12-month momentum	Monthly	Jegadeesh and Titman (1993)
47	mom1m	1-month momentum	Monthly	Jegadeesh and Titman (1993)
48	mom36m	36-month momentum	Monthly	Jegadeesh and Titman (1993)
49	mom6m	6-month momentum	Monthly	Jegadeesh and Titman (1993)
50	ms	Financial statement score	Quarterly	Mohanram (2005)
51	mvel1	Size	Monthly	Banz (1981)
52	mveia	Industry-adjusted size	Annual	Asness et al. (2000)
53	nincr	Number of earnings increases	Quarterly	Barth et al. (1999)

(continued on next page)

Table C.1 (continued)

No.	Acronym	Firm characteristic	Frequency	Literature
54	operprof	Operating profitability	Annual	Fama and French (2015)
55	orgcap	Organizational capital	Annual	Eisfeldt and Papanikolaou (2013)
56	pchcapxia	Industry adjusted change in capital exp.	Annual	Abarbanell and Bushee (1998)
57	pchcurrat	Change in current ratio	Annual	Ou and Penman (1989)
58	pchdepr	Change in depreciation	Annual	Holthausen and Larcker (1992)
59	pchgmpchsale	Change in gross margin - change in sales	Annual	Abarbanell and Bushee (1998)
60	pchquick	Change in quick ratio	Annual	Ou and Penman (1989)
61	pchsalepchinv	Change in sales - change in inventory	Annual	Abarbanell and Bushee (1998)
62	pchsalepchrect	Change in sales - change in A/R	Annual	Abarbanell and Bushee (1998)
63	pchsalepchxsga	Change in sales - change in SG&A	Annual	Abarbanell and Bushee (1998)
64	ppchsaleinv	Change sales-to-inventory	Annual	Ou and Penman (1989)
65	pctacc	Percent accruals	Annual	Hafzalla et al. (2011)
66	pricedelay	Price delay	Monthly	Hou and Moskowitz (2005)
67	ps	Financial statements score	Annual	Piotroski (2000)
68	quick	Quick ratio	Annual	Ou and Penman (1989)
69	rd	R&D increase	Annual	Eberhart et al. (2004)
70	rdmve	R&D to market capitalization	Annual	Guo et al. (2006)
71	rdsale	R&D to sales	Annual	Guo et al. (2006)
72	realestate	Real estate holdings	Annual	Tuzel (2010)
73	retvol	Return volatility	Monthly	Ang et al. (2006)
74	roaq	Return on assets	Quarterly	Balakrishnan et al. (2010)
75	roavol	Earnings volatility	Quarterly	Francis et al. (2004)
76	roeq	Return on equity	Quarterly	Hou et al. (2015)
77	roic	Return on invested capital	Annual	Brown and Rowe (2007)
78	rsup	Revenue surprise	Quarterly	Kama (2009)
79	salecash	Sales to cash	Annual	Ou and Penman (1989)
80	saleinv	Sales to inventory	Annual	Ou and Penman (1989)
81	salerec	Sales to receivables	Annual	Ou and Penman (1989)
82	secured	Secured debt	Annual	Valta (2016)
83	securedind	Secured debt indicator	Annual	Valta (2016)
84	sgr	Sales growth	Annual	Lakonishok et al. (1994)
85	sin	Sin stocks	Annual	Hong and Kacperczyk (2009)
86	sp	Sales to price	Annual	Barbee et al. (1996)
87	stddolvol	Volatility of liquidity (dollar trading volume)	Monthly	Chordia et al. (2001)
88	stdturn	Volatility of liquidity (share turnover)	Monthly	Chordia et al. (2001)
89	stdacc	Accrual volatility	Quarterly	Bandyopadhyay et al. (2010)
90	stdcf	Cash flow volatility	Quarterly	Huang (2009)
91	tang	Debt capacity/firm tangibility	Annual	Almeida and Campello (2007)
92	tb	Tax income to book income	Annual	Lev and Nissim (2004)
93	turn	Share turnover	Monthly	Datar et al. (1998)
94	zerotrade	Zero trading days	Monthly	Liu (2006)

References

- Abarbanell, J., Bushee, B., 1998. Abnormal returns to a fundamental analysis strategy. *Account. Rev.* 73 (1), 19–45.
- Ali, A., Hwang, L., Trombley, M., 2003. Arbitrage risk and the book-to-market anomaly. *J. Financ. Econ.* 69 (2), 355–373.
- Almeida, H., Campello, M., 2007. Financial constraints, asset tangibility, and corporate investment. *Rev. Financ. Stud.* 20 (5), 1429–1460.
- Altmann, A., Tološi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics* 26 (10), 1340–1347. <https://doi.org/10.1093/bioinformatics/btq134>. <https://academic.oup.com/bioinformatics/article-pdf/26/10/1340/16892402/btq134.pdf>.
- Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. *J. Financ. Mark.* 5 (1), 31–56.
- Amihud, Y., Mendelson, H., 1989. The effects of beta, bid-ask spread, residual risk, and size on stock returns. *J. Finance* 44 (2), 479–486.
- Anderson, C., Garcia-Feijóo, L., 2006. Empirical evidence on capital investment, growth options, and security returns. *J. Finance* 61 (1), 171–194.
- Ang, A., Hodrick, R.J., Xing, Y., Zhang, X., 2006. The cross-section of volatility and expected returns. *J. Finance* 61 (1), 259–299.
- Asness, C., Porter, B., Stevens, R., 2000. Predicting stock returns using industry-relative firm characteristics. Working paper.
- Balakrishnan, K., Bartov, E., Faurel, L., 2010. Post loss/profit announcement drift. *J. Account. Econ.* 50 (1), 20–41.
- Bali, T.G., Cakici, N., Whitelaw, R.F., 2011. Mxing out: stocks as lotteries and the cross-section of expected returns. *J. Financ. Econ.* 99 (2), 427–446.
- Bandyopadhyay, S.P., Huang, A.G., Wirjanto, T.S., 2010. The accrual volatility anomaly. Working paper. School of Accounting and Finance, University of Waterloo.
- Banz, R.W., 1981. The relationship between return and market value of common stocks. *J. Financ. Econ.* 9 (1), 3–18.
- Barbee, W., Mukherji, S., Raines, G., 1996. Do sales-price and debt-equity explain stock returns better than book-market and firm size? *Financ. Anal. J.* 52 (2), 56–60.
- Barth, M., Elliott, J., Finn, M., 1999. Market rewards associated with patterns of increasing earnings. *J. Account. Res.* 37 (2), 387–413.
- Basu, S., 1977. Investment performance of common stocks in relation to their price-earnings ratios: a test of the efficient market hypothesis. *J. Finance* 32 (3), 663–682.
- Belo, F., Lin, X., Bazdresch, S., 2014. Labor hiring, investment, and stock return predictability in the cross section. *J. Polit. Econ.* 122 (1), 129–177.
- Bhandari, L.C., 1988. Debt/equity ratio and expected common stock returns: empirical evidence. *J. Finance* 43 (2), 507–528.
- Biau, G., Scornet, E., 2016. A random forest guided tour. *Test* 25 (2), 197–227.
- Borji, A., 2019. Pros and cons of GAN evaluation measures. *Comput. Vis. Image Underst.* 179, 41–65. <https://doi.org/10.1016/j.cviu.2018.10.009>. <http://www.sciencedirect.com/science/article/pii/S1077314218304272>.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45 (1), 5–32. <https://doi.org/10.1023/A:1010933404324>.

- Brown, D., Rowe, B., 2007. The productivity premium in equity returns. Working paper.
- Cai, H., Goggin, B., Jiang, Q., 2020. Two-sample test based on classification probability. *Stat. Anal. Data Min. ASA Data Sci. J.* 13 (1), 5–13. <https://doi.org/10.1002/sam.11438>. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11438>.
- Chandrashekar, S., Rao, R.K., 2009. The productivity of corporate cash holdings and the cross-section of expected stock returns. *McCombs Research Paper Series No. FIN-03-09*.
- Chordia, T., Subrahmanyam, A., Anshuman, V.R., 2001. Trading activity and expected stock returns. *J. Financ. Econ.* 59 (1), 3–32.
- Chwialkowski, K.P., Ramdas, A., Sejdinovic, D., Gretton, A., 2015. Fast two-sample testing with analytic representations of probability measures. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates, Inc., pp. 1981–1989. <http://papers.nips.cc/paper/5685-fast-two-sample-testing-with-analytic-representations-of-probability-measures.pdf>.
- Cooper, M.J., Gulen, H., Schill, M.J., 2008. Asset growth and the cross-section of stock returns. *J. Finance* 63 (4), 1609–1651.
- Datar, V.T., Naik, N.Y., Radcliffe, R., 1998. Liquidity and stock returns: an alternative test. *J. Financ. Mark.* 1 (2), 203–219.
- Demarta, S., McNeil, A.J., 2005. The t copula and related copulas. *Int. Stat. Rev.* 73 (1), 111–129.
- Desai, H., Rajgopal, S., Venkatachalam, M., 2004. Value-glamour and accruals mispricing: one anomaly or two? *Account. Rev.* 79 (2), 355–385.
- Devroye, L., Györfi, L., Lugosi, G., 1996. *A Probabilistic Theory of Pattern Recognition*. Springer.
- DiCiccio, C., Romano, J.P., 2020. CLT for U-Statistics with Growing Dimension. Tech. Rep. Stanford University, Department of Statistics. <https://statistics.stanford.edu/sites/g/files/sbiybj6031/f/2020-01rev.pdf>.
- Eberhart, A.C., Maxwell, W.F., Siddique, A.R., 2004. An examination of long-term abnormal stock returns and operating performance following R&D increases. *J. Finance* 59 (2), 623–650.
- Eisfeldt, A., Papanikolaou, D., 2013. Organization capital and the cross-section of expected returns. *J. Account. Res.* 68 (4), 1365–1406.
- Fairfield, P., Whisenant, S., Yohn, L., 2003. Accrued earnings and growth: implications for future profitability and market mispricing. *Account. Rev.* 78 (1), 353–371.
- Fama, E., MacBeth, J., 1973. Risk, return, and equilibrium: empirical tests. *J. Polit. Econ.* 81 (3), 607–636.
- Fama, E.F., French, K.R., 2015. A five factor asset pricing model. *J. Financ. Econ.* 116 (1), 1–22.
- Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* 15, 3133–3181. <http://jmlr.org/papers/v15/delgado14a.html>.
- Francis, J., LaFond, R., Olsson, P., Schipper, K., 2004. Costs of equity and earnings attributes. *Account. Rev.* 79 (4), 967–1010.
- Friedman, J., 2004. On multivariate goodness-of-fit and two-sample testing. Stanford Linear Accelerator Center, Menlo Park, CA (US).
- Fuchs, M., Hornung, R., Bin, R.D., Boulesteix, A.-L., 2013. A U-statistic estimator for the variance of resampling-based error estimators. <http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-17654-2>.
- Gagnon-Bartsch, J., Shem-Tov, Y., 2019. The classification permutation test: a flexible approach to testing for covariate imbalance in observational studies. *Ann. Appl. Stat.* 13 (3), 1464–1483. <https://doi.org/10.1214/19-AOAS1241>.
- Gettleman, E., Marks, J.M., 2006. Acceleration strategies. SSRN Working Paper Series.
- Good, P., 1994. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer Series in Statistics. Springer, New York, NY.
- Gravier, Eleonore, Pierron, G., Vincent-Salomon, A., Gruel, N., Raynal, V., Savignoni, A., De Rycke, Y., Pierga, J.-Y., Lucchesi, C., Reyat, F., Fourquet, A., Roman-Roman, S., Radvanyi, F., Sastre-Garau, X., Asselain, B., Delattre, O., 2010. A prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes Chromosomes Cancer* 49 (12), 1125.
- Green, J., Hand, J., Zhang, F., 2017. The characteristics that provide independent information about average US monthly stock returns. *Rev. Financ. Stud.* 30, 4389–4436.
- Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., 2012a. A kernel two-sample test. *J. Mach. Learn. Res.* 13 (1), 723–773. <http://dl.acm.org/citation.cfm?id=2503308.2188410>.
- Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., Sriperumbudur, B.K., 2012. Optimal kernel choice for large-scale two-sample tests. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., pp. 1205–1213. <http://papers.nips.cc/paper/4727-optimal-kernel-choice-for-large-scale-two-sample-tests.pdf>.
- Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. *Rev. Financ. Stud.* 33 (5), 2223–2273. <https://doi.org/10.1093/rfs/hhaa009>. <https://academic.oup.com/rfs/advance-article-pdf/doi/10.1093/rfs/hhaa009/32892601/hhaa009.pdf>.
- Guo, R., Lev, B., Shi, C., 2006. Explaining the short- and long-term IPO anomalies in the us by R&D. *J. Bus. Finance Account.* 33 (3–4), 550–579.
- Hafzalla, N., Lundholm, R., Matthew Van Winkle, E., 2011. Percent accruals. *Account. Rev.* 86 (1), 209–236.
- Hemerik, J., Goeman, J., 2018. Exact testing with random permutations. *Test (Madrid, Spain)* 27 (4), 811–825. <https://doi.org/10.1007/s11749-017-0571-1>. PMID: 30930620. <https://pubmed.ncbi.nlm.nih.gov/30930620>.
- Holthausen, R., Larcker, D., 1992. The prediction of stock returns using financial statement information. *J. Account. Econ.* 15, 373–411.
- Hong, H., Kacperczyk, M., 2009. The price of sin: the effects of social norms on markets. *J. Financ. Econ.* 93, 15–36.
- Hotelling, H., 1931. The generalization of student's ratio. *Ann. Math. Stat.* 2 (3), 360–378. <https://doi.org/10.1214/aoms/1177732979>.
- Hou, K., Moskowitz, T., 2005. Market frictions, price delay, and the cross-section of expected returns. *Rev. Financ. Stud.* 18 (3), 981–1020.
- Hou, K., Robinson, D., 2006. Industry concentration and average stock returns. *J. Finance* 61 (4), 1927–1956.
- Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: an investment approach. *Rev. Financ. Stud.* 28 (3), 650–705.
- Huang, A.G., 2009. The cross section of cashflow volatility and expected stock returns. *J. Empir. Finance* 16 (3), 409–429.
- Janitzka, S., Celik, E., Boulesteix, A.-L., 2018. A computationally fast variable importance test for random forests for high-dimensional data. *Adv. Data Anal. Classif.* 12 (4), 885–915. <https://doi.org/10.1007/s11634-016-0276-4>.
- Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market efficiency. *J. Finance* 48 (1), 65–91.
- Jiang, G., Lee, C., Zhang, Y., 2005. Information uncertainty and expected returns. *Rev. Acc. Stud.* 10, 185–221.
- Jitkrittum, W., Szabó, Z., Chwialkowski, K.P., Gretton, A., 2016. Interpretable distribution features with maximum testing power. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., pp. 181–189. <http://papers.nips.cc/paper/6148-interpretable-distribution-features-with-maximum-testing-power.pdf>.
- Kama, I., 2009. On the market reaction to revenue and earnings surprises. *J. Bank. Finance* 36, 31–50.
- Kim, I., Lee, A.B., Lei, J., 2019. Global and local two-sample tests via regression. *Electron. J. Stat.* 13 (2), 5253–5305. <https://doi.org/10.1214/19-EJS1648>.
- Kim, I., Ramdas, A., Singh, A., Wasserman, L., 2021. Classification accuracy as a proxy for two-sample testing. *Ann. Stat.* 49 (1), 411–434. <https://doi.org/10.1214/20-AOS1962>.
- Kishore, R., Brandt, M., Santa-Clara, P., Venkatachalam, M., 2008. Earnings announcements are full of surprises. Working paper.
- Lakonishok, J., Shleifer, A., Vishny, R.W., 1994. Contrarian investment, extrapolation, and risk. *J. Finance* 49 (5), 1541–1578.
- Lee, A.J., 1990. *U-Statistics: Theory and Practice, Statistics: A Series of Textbooks and Monographs*. CRC Press, New York.
- Lerman, A., Livnat, J., Mendenhall, R.R., 2008. The high-volume return premium and post-earnings announcement drift. Available at SSRN 1122463.
- Lev, B., Nissim, D., 2004. Taxable income, future earnings, and equity values. *Account. Rev.* 79 (4), 1039–1074.
- Litzenberger, R., Ramaswamy, K., 1982. The effects of dividends on common stock prices tax effects or information effects? *J. Finance* 37 (2), 429–443.
- Liu, W., 2006. A liquidity-augmented capital asset pricing model. *J. Financ. Econ.* 82 (3), 631–671.

- Lopez-Paz, D., Oquab, M., 2018. Revisiting classifier two-sample tests. arXiv:1610.06545.
- Luntz, A.C., Brailovsky, V.L., 1969. On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica* 3.
- McNeil, A.J., Frey, R., Embrechts, P., 2015. *Quantitative Risk Management: Concepts, Techniques, and Tools*, revised edition. Princeton University Press, Princeton.
- Mentch, L., Hooker, G., 2016. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* 17 (1), 841–881. <http://dl.acm.org/citation.cfm?id=2946645.2946671>.
- Michaely, R., Thaler, R., Womack, K., 1995. Price reactions to dividend initiations and omissions: overreaction or drift? *J. Finance* 50 (2), 573–608.
- Mohanram, P., 2005. Separating winners from losers among lowbook-to-market stocks using financial statement analysis. *Rev. Acc. Stud.* 10, 133–170.
- Moskowitz, T., Grinblatt, M., 1999. Do industries explain momentum? *J. Finance* 54 (4), 1249–1290.
- Moskowitz, T., Grinblatt, M., 2010. A better three-factor model that explains more anomalies. *J. Finance* 65 (2), 563–594.
- Novy-Marx, R., 2013. The other side of value: good growth and the Gross profitability premium. *J. Financ. Econ.* 108 (1), 1–28.
- Ou, J., Penman, S., 1989. Financial statement analysis and the prediction of stock returns. *J. Account. Econ.* 11 (4), 295–329.
- Palazzo, B., 2012. Cash holdings, risk, and expected returns. *J. Financ. Econ.* 104 (1), 162–185.
- Peng, W., Coleman, T., Mentch, L., 2019. Asymptotic distributions and rates of convergence for random forests via generalized U-statistics. arXiv:1905.10651.
- Piotroski, J.D., 2000. Value investing: the use of historical financial statement information to separate winners from losers. *J. Account. Res.*, 1–41.
- Pontiff, J., Woodgate, A., 2008. Share issuance and cross-sectional returns. *J. Finance* 63 (2), 921–945.
- Ramdas, A., Reddi, S.J., Póczos, B., Singh, A., Wasserman, L.A., 2015. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In: AAAI. AAAI Press, pp. 3571–3577.
- Ramey, J., 2016. *datamicroarray: a collection of small-sample, high-dimensional microarray data sets to assess machine-learning algorithms and models*. <https://github.com/ramhiser/datamicroarray>.
- Richardson, S.A., Sloan, R.G., Soliman, M.T., Tuna, I., 2005. Accrual reliability, earnings persistence and stock prices. *J. Account. Econ.* 39 (3), 437–485.
- Rosenberg, B., Reid, K., Lanstein, R., 1985. Persuasive evidence of market inefficiency. *J. Portf. Manag.* 11 (3), 9–16.
- Rosenblatt, J., Gilron, R., Mukamel, R., 2021. Better-than-chance classification for signal detection. *Biostatistics* 22 (2), 365–380. <https://doi.org/10.1093/biostatistics/kxz035>. (Oxford, England).
- Sloan, R., 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? (Digest summary). *Account. Rev.* 71 (3), 289–315.
- Soliman, M.T., 2008. The use of DuPont analysis by market participants. *Account. Rev.* 83 (3), 823–853.
- Thomas, J., Zhang, F.X., 2011. Tax expense momentum. *J. Account. Res.* 49 (3), 791–821.
- Thomas, J.K., Zhang, H., 2002. Inventory changes and future returns. *Rev. Acc. Stud.* 7 (2–3), 163–187.
- Titman, S., Wei, K.J., Xie, F., 2004. Capital investments and stock returns. *J. Financ. Quant. Anal.* 39 (04), 677–700.
- Tuzel, S., 2010. Corporate real estate holdings and the cross-section of stock returns. *Rev. Financ. Stud.* 23 (6), 2268–2302.
- Valta, P., 2016. Strategic default, debt structure, and stock returns. *J. Financ. Quant. Anal.* 51 (1), 1–33.
- van der Vaart, A., 1998. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Wager, S., Athey, S., 2017. Estimation and inference of heterogeneous treatment effects using random forests. arXiv:1510.04342.
- Westfall, P., Young, S., Kohne, S., Pigeot, I., 1995. Resampling-based multiple testing. Examples and methods for p-value adjustment. *Comput. Stat. Data Anal.*, 235.