Multiple Outlier Testing with Conformal p-values

Stephen Bates*[1], Emmanuel Candès[2], Lihua Lei[3], Yaniv Romano[4], Matteo Sesia[5]

April 19, 2021

Abstract

This paper studies the construction of p-values for nonparametric outlier detection, taking a multiple-testing perspective. The goal is to test whether new independent samples belong to the same distribution as a reference data set or are outliers. We propose a solution based on conformal inference, a broadly applicable framework which yields p-values that are marginally valid but mutually dependent for different test points. We prove these p-values are positively dependent and enable exact false discovery rate control, although in a relatively weak marginal sense. We then introduce a new method to compute p-values that are both valid conditionally on the training data and independent of each other for different test points; this paves the way to stronger type-I error guarantees. Our results depart from classical conformal inference as we leverage concentration inequalities rather than combinatorial arguments to establish our finite-sample guarantees. Furthermore, our techniques also yield a uniform confidence bound for the false positive rate of any outlier detection algorithm, as a function of the threshold applied to its raw statistics. Finally, the relevance of our results is demonstrated by numerical experiments on real and simulated data.

Keywords — Conformal inference, out-of-distribution testing, false discovery rate, positive dependence.

1 Introduction

1.1 Problem statement and motivation

We consider an outlier detection problem in which one observes a data set $\mathcal{D} = \{X_i\}_{i=1}^{2n}$ containing $2n$ independent and identically distributed points $X_i \in \mathbb{R}^d$ drawn from an unknown distribution P_X (which may be continuous, discrete, or mixed). The goal is to test which among a new set of $n_{\text{test}} \geq 1$ independent observations $\mathcal{D}^{\text{test}} = \{X_{2n+i}\}_{i=1}^{n_{\text{test}}}$ are *outliers*, in the sense that they were not drawn from the same distribution P_X. By contrast, we refer to points drawn from P_X as *inliers*. This problem has applications in many domains, including medical diagnostics [1], spotting frauds or intrusions [2], forensic analysis [3], monitoring engineering systems for failures [4], and *out-of-distribution* detection in machine learning [5–8]. A variety of machine-learning tools have been developed to address this classification task, which is sometimes referred to as *one-class classification* [9, 10] because the data in $\mathcal{D}$ do not contain any outliers. However,

[*] Authors listed alphabetically.

[1] Departments of Statistics and of EECS, UC Berkeley.

[2] Departments of Statistics and of Mathematics, Stanford University.

[3] Department of Statistics, Stanford University.

[4] Departments of Electrical Engineering and of Computer Science, Technion—Israel Institute of Technology.

[5] Department of Data Sciences and Operations, University of Southern California.

such algorithms are often complex and their outputs are not directly covered by any precise statistical guarantees. Fortunately, conformal inference [11, 12] allows one to practically convert the output of any one-class classifier (if it is invariant to the ordering of the training observations) into a provably valid p-value for the null hypothesis $\mathcal{H}_{0,i} : X_i \sim P_X$, for any $X_i \in \mathcal{D}^{\text{test}}$.

In many applications, the number of outlier tests, n_{test}, is large and, therefore, it may be necessary to account for multiple comparisons to avoid making an excessive number of false discoveries. A meaningful error rate in this setting is the false discovery rate (FDR) [13]: the expected proportion of true inliers among the test points reported as outliers. For example, if a particular financial transaction is labeled by an automated system as likely to be fraudulent (i.e., unusual, or out-of-distribution compared to a data set of normal transactions), someone may then need to review it manually, and possibly contact the involved customer. Since these follow-up procedures have a cost, controlling the FDR may be a sensible solution to ensure resources are allocated efficiently. From a statistics perspective, multiple testing in this setting requires some care because classical conformal p-values corresponding to different values of $i > 2n$ are independent of each other only conditional on $\mathcal{D}$, although they are valid only marginally over $\mathcal{D}$. This situation is delicate because FDR control typically requires p-values that either are mutually independent or follow certain patterns of dependence [14, 15]. Similarly, global testing (i.e., aggregating evidence from multiple observations to test weaker batch-level hypotheses) may also require independent p-values. This paper addresses the above issues by carefully studying the theoretical properties of some standard multiple testing procedures applied to conformal p-values, and by developing new methods to compute p-values with stronger validity properties.

The conformal inference methods studied in this paper are statistical wrappers for one-class classifiers. The latter are algorithms trained on data clean of any outliers to compute a score function $\hat{s} : \mathbb{R}^d \to \mathbb{R}$ assigning a scalar value to any future data point, so that smaller (for example) values of $\hat{s}(X)$ provide evidence that X may be an outlier. By design, the classifier attempts to construct scores that separate outliers from inliers effectively, by learning from the data what inliers typically look like, and it may be based on sophisticated black-box models to maximize power. While often effective in practice, these machine-learning algorithms have the drawback of not offering any clear guarantees about the quality of their output. For example, they do not directly provide a null distribution for the classification scores $\hat{s}$ evaluated on true inliers, or any particular threshold to limit the rate of false positives. This is where conformal inference comes to help. After training $\hat{s}$ on a subset of the observations in $\mathcal{D}$, namely those in $\mathcal{D}^{\text{train}} = \{X_1, \dots, X_n\}$, the scores are evaluated on the remaining n hold-out samples in $\mathcal{D}^{\text{cal}} = \{X_{n+1}, \dots, X_{2n}\}$. (Note that $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{cal}}$ do not need to contain the same number of observations, although the current choice simplifies the notation without loss of generality). Let us assume, for simplicity, that $\hat{s}(X)$ has a continuous distribution if $X \sim P_X$ is independent of the data used to train $\hat{s}$, although this assumption could be relaxed at the cost of some additional technical details. Then, define F as the cumulative distribution function (CDF) of $\hat{s}(X)$. If we knew F, we could utilize $F(X_i)$ as an exact p-value for the null hypothesis $\mathcal{H}_{0,i} : X_i \sim P_X$, for any $X_i \in \mathcal{D}^{\text{test}}$, in the sense that $F(X_i)$ would be uniformly distributed if $\mathcal{H}_{0,i}$ is true. In practice, however, we do not have direct access to F because P_X is unknown and the machine-learning algorithm upon which $\hat{s}$ depends is assumed to be a black-box. Instead, we can evaluate the empirical CDF of $\hat{s}(X_i)$ for all $X_i \in \mathcal{D}^{\text{cal}}$, which we denote as $\hat{F}$. In the following, we will discuss how to construct provably valid conformal p-values for a future observation X_{2n+1} by evaluating

$$\hat{u}(X_{2n+1}) = \left(g \circ \hat{F} \circ \hat{s}\right)(X_{2n+1}), \tag{1}$$

where g is a suitable *adjustment function*, and the symbol $\circ$ denotes a composition; i.e., $(f \circ g)(x) = f(g(x))$. Note that, hereafter, we will treat the observations in $\mathcal{D}^{\text{train}}$ as fixed and focus on the randomness in the calibration ($\mathcal{D}^{\text{cal}}$) and test ($\mathcal{D}^{\text{test}}$) data, upon which conformal inferences are generally based.

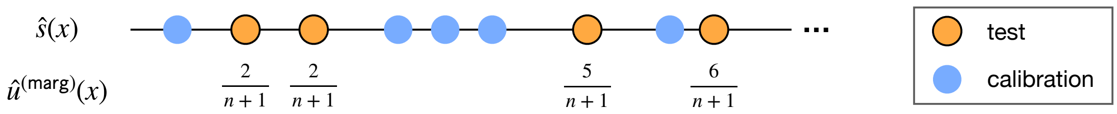

Figure 1: Visualization of the joint distribution of the conformal p-values. The distribution of $\hat{s}(x)$ is the same for calibration and inlier test points. The conformal p-value for each test point is the number of calibration points to its left, divided by the total number of calibration points plus one, as in (3).

1.2 Preview of contributions

In Section 2, we will focus on the classical conformal inference methods, which produce *marginally super-uniform* (conservative) p-values $\hat{u}^{(\text{marg})}(X_{2n+1})$ satisfying

$$\mathbb{P}\left[\hat{u}^{(\text{marg})}(X_{2n+1}) \leq t\right] \leq t, \tag{2}$$

for any $t \in (0,1)$, whenever X_{2n+1} is an inlier. We say these p-values are marginally valid because they depend on the calibration data in $\mathcal{D}^{\text{cal}}$, and both $\mathcal{D}^{\text{cal}}$ and X_{2n+1} are random in (2). In particular, the classical $\hat{u}^{(\text{marg})}$ is computed by applying the adjustment function $g^{(\text{marg})}(x) = (nx+1)/(n+1)$ to (1), i.e.,

$$\hat{u}^{(\text{marg})}(x) = \frac{1 + |\{i \in \mathcal{D}^{\text{cal}} : \hat{s}(X_i) \leq \hat{s}(x)\}|}{n+1}. \tag{3}$$

Note that (2) is implied by (3) because when $\hat{s}(X)$ follows a continuous distribution, $\hat{u}^{(\text{marg})}(X)$ is uniformly distributed on $\{1/(n+1), 2/(n+1), \ldots, 1\}$ if $X \sim P_X$ independently of the data in $\mathcal{D}^{\text{train}}$ [11, 12]. (If $\hat{s}(X)$ is not continuous, one can still verify that $\hat{u}^{(\text{marg})}(X)$ is super-uniform in distribution.) However, this is no longer true if one conditions on $\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{cal}}$, in which case $\hat{u}^{(\text{marg})}(X)$ may become anti-conservative. Furthermore, marginal p-values corresponding to different test points, $\{\hat{u}^{(\text{marg})}(X)\}_{X \in \mathcal{D}^{\text{test}}}$, are not mutually independent because they are all affected by $\mathcal{D}^{\text{cal}}$; see Figure 1 for a visualization of this dependence. This should be taken into account when adjusting for multiplicity in outlier detection applications because some common testing procedures are not generally valid for dependent p-values. For example, we will prove in Section 2 that the dependence among marginal p-values invalidates Fisher's combination test [16] for the global null that there are no outliers in $\mathcal{D}^{\text{test}}$, although this can be easily fixed by suitably adjusting the critical value. By contrast, we can prove the dependence between conformal p-values does not break the Benjamini-Hochberg procedure [13] for FDR control, even if the latter is applied with Storey's correction [17].

In any case, regardless of whether the mutual dependence among marginal p-values theoretically invalidates a particular multiple-testing procedure, one may sometimes be interested in obtaining stronger guarantees compared to the typical marginal validity of conformal p-values defined in (2). Consider for instance the following prototypical scenario. A researcher, or a company, acquires an expensive data set $\mathcal{D}$ containing clean examples of some variable X of interest, and wishes to leverage that information to construct a system to detect outliers in future test points, while avoiding an excess of false positives. Assuming the stakes in this application are sufficiently high, the researcher may need clear statistical guarantees about the output of such procedure (as opposed to blindly trusting a black-box model), and thus decides to employ conformal inference. Unfortunately, the marginal validity property in (2) tells us very little about how this outlier detection system may perform in the future for *this particular researcher relying on this particular data set* $\mathcal{D}$. Instead, marginal validity suggests the system will work *on average* for different researchers starting from different data sets; of course, that may not feel fully satisfactory for any one of them.

Therefore, we will construct in Section 3 conformal p-values satisfying a stronger property, which we call *calibration-conditional validity* (CCV). Formally, the novel p-values $\hat{u}^{(\text{ccv})}(x)$ will satisfy

$$\mathbb{P}\left[\mathbb{P}\left[\hat{u}^{(\text{ccv})}(X_{2n+1}) \leq t \mid \mathcal{D}\right] \leq t \;\text{ for all }\; t \in (0,1)\right] \geq 1 - \delta, \tag{4}$$

if $X_{2n+1} \sim P_X$, for any value of $\delta \in (0,1)$ pre-specified by the user. The crucial difference between (4) and (2) is that the latter intuitively guarantees the p-values are valid for at least a fraction $1-\delta$ of researchers;

this can give a precise measure of confidence to each one of them. Furthermore, calibration-conditional p-values have the advantage of making multiple testing straightforward. In fact, these p-values are still trivially independent of one another conditional on the calibration data, so their high-probability guarantee of validity will immediately extend to the output of any downstream multiple-testing procedure that assumes independence.

While most of this paper focuses on the validity of conformal p-values from a multiple-testing perspective, we will see in Section 4 that our high-probability results can also be utilized to construct a uniform upper confidence bound for the false positive rate of any machine-learning algorithm for outlier detection, as a function of the threshold applied to its raw output scores. This may help practitioners interpret the output of black-box methods directly, without necessarily operating in terms of p-values. (However, as statisticians, we prefer the p-value approach because it is more versatile.) Furthermore, our results can be easily leveraged to obtain predictive sets with stronger coverage guarantees compared to existing conformal methods.

Finally, in Section 5, we will compare the performance of marginal and calibration-conditional conformal p-values on simulated as well as real data, in combination with different multiple testing procedures. These numerical experiments will provide an empirical confirmation of our theoretical results, and also highlight how stronger guarantees sometimes come at the cost of lower power.

1.3 Related work

The outlier detection problem considered in this paper is fully non-parametric, in the sense that we leverage the information contained in an external clean data set, and nothing else, to infer whether a future test point may be an outlier. This is in contrast with the more classical problem of multivariate outlier detection within a single data set, leveraging modeling assumptions rather than clean external samples [18–21]. A wealth of data mining and machine-learning methods have been developed to address our non-parametric task [22–26]; these do not provide precise finite-sample guarantees on their own, but we can leverage them to compute scoring functions that powerfully separate outliers from inliers.

Our paper is based on conformal inference [11, 12], which has been applied before in the context of outlier detection [27–32]. However, previous works did not study the implications of marginal p-values on the validity of multiple outlier testing procedures, nor did they seek the conditional guarantees obtained here. Another line of work applied conformal inference to test the global null for streaming data [33–37]. However, the guarantee no longer holds in the offline setting or beyond the global null. The most closely related work is that of [38], which extends conformal inference to provide a form of calibration-conditional coverage. That paper focused explicitly on the prediction setting rather than on outlier detection, but is also directly relevant in our context, as discussed in Section 3.1. The main difference is that our novel high-probability bounds in Section 3 hold simultaneously for all possible coverage levels (in the language of [38]) not just for a pre-specified one—this feature being necessary to obtain conditionally valid p-values for multiple outlier testing.

Other works on conformal inference focused on different types of conditional coverage. For example, [39] studied the difficulty of computing valid conformal predictions (in a supervised setting) conditional on the features of a new test point, while we are interested in conditioning on the calibration data (in an outlier detection setting). Other works have focused on seeking approximate feature-conditional coverage in multi-class classification [40–43] or in regression problems [44–48]. This paper is orthogonal, in the sense that our results could be applied to strengthen their coverage guarantees by conditioning on the calibration data. It should be noted that, although conformal inference can be based on different data hold-out strategies [49–51], our paper focuses on sample splitting [52, 53]. The latter has the advantage of being the most computationally efficient option, and is necessary for us in theory because our high-probability bounds require the independence of the data points in addition to their exchangeability.

Further, the problem we consider is related to classical two-sample testing [54], although we take a different perspective. Two-sample testing compares two data sets to determine whether they were sampled from the same distribution, while our goal is to contrast many independent test points (or batches thereof) to the same reference set accounting for multiplicity. In any case, several recent works have explored the use of machine-learning and data hold-out methods for two-sample testing [55–59], which reinforces the connection with our work.

Finally, the duality between hypothesis testing and confidence intervals connects our conditionally calibrated p-values to the classical statistical topic of *tolerance regions*, which goes back to Wilks [60, 61], Wald [62], and Tukey [63]. See [64] for a overview of the subject, [38] for a discussion of their connection with conformal inference, and [65, 66] for modern examples using tolerance regions for predictive inference with neural networks. (Tolerance regions are predictive sets with a high-probability guarantee to contain the desired fraction of the population. For example, one can generate a tolerance region guaranteed to contain at least 80% of the population with probability 99%.) The construction of predictive intervals with (asymptotic) conditional validity in the aforementioned sense was also recently studied in [67] with bootstrap rather than conformal inference methods.

2 Marginal conformal inference for outlier detection

Before turning to calibration-conditional inferences, we carefully study the marginal validity of multiple tests based on split-conformal outlier detection p-values. The conformal p-values defined in (3) are marginally valid for the hypothesis that a single test point follows the distribution P_X, see (2), but they are not independent of each other when considering multiple test points. Consequently, we show they cannot be naively used to test a global null hypothesis that no points in a particular test set are outliers, with Fisher's combination test [16] for example. The failure of Fisher's test is caused by the particular dependence induced by the shared calibration data set, although other procedures turn out to be robust to such dependence. In particular, we then prove conformal p-values are *positive regression dependent on a subset* (PRDS), which combined with the results of [14], implies the Benjamini-Hochberg algorithm will control the FDR.

2.1 A negative result: global testing with conformal p-values can fail

Fisher's combination test [16] is a widely-used method to test the global null, in our case

$$H_0 : X_{2n+1}, \dots, X_{2n+m} \overset{\text{i.i.d.}}{\sim} P_X.$$

The idea is to aggregate the evidence from the individual tests, as follows. Given a p-value p_i for each null hypothesis i, Fisher's test rejects the global null at level α if

$$-2 \sum_{i=1}^{m} \log p_i \geq \chi^2(2m; 1-\alpha),$$

where $\chi^2(2m; 1-\alpha)$ is the $(1-\alpha)$-th quantile of the chi-square distribution with $2m$ degrees of freedom. This test is valid if the p-values stochastically dominate Unif([0, 1]) and are independent of each other. However, we prove in the following lemma that the standard (marginal) conformal p-values are positively correlated under arbitrary transformations, suggesting an inflation of the variance of the combination statistics.

Lemma 1. *Assume that $\hat{s}(X)$ is continuous. Then, for any function $G : [0, 1] \mapsto \mathbb{R}$, and for any pair of nulls (i, j),*

$$\text{Cor}\left[G(\hat{u}^{(\text{marg})}(X_{2n+i})), G(\hat{u}^{(\text{marg})}(X_{2n+j}))\right] = \frac{1}{n+2}.$$

Motivated by Lemma 1 (see Appendix A.1 for a detailed discussion), we obtain the following result which shows Fisher's combination test becomes invalid when applied to marginal conformal p-values. In particular, we characterize its type-I error in the asymptotic regime where $|\mathcal{D}^{\text{test}}|$ is proportional to $|\mathcal{D}^{\text{cal}}|$.

Theorem 1 (Type-I error of Fisher's combination test)**.** *Assume that $\hat{s}(X)$ is continuous. Then, under the global null, if $m = \lfloor \gamma n \rfloor$ for some $\gamma \in (0, \infty)$, as n tends to infinity,*

$$\mathbb{P}\left[-2 \sum_{i=1}^{m} \log\left[\hat{u}^{(\text{marg})}(X_{2n+i})\right] \geq \chi^2(2m; 1-\alpha)\right] \to \bar{\Phi}\left(\frac{z_{1-\alpha}}{\sqrt{1+\gamma}}\right),$$

where $z_{1-\alpha}$ and $\bar{\Phi}$ denote the $(1-\alpha)$-th quantile and tail function of the standard normal distribution, respectively. Furthermore, under the same asymptotic regime, for $W \sim N(0,1)$,

$$\mathbb{P}\left[-2\sum_{i=1}^{m}\log\left[\hat{u}^{(\mathrm{marg})}(X_{2n+i})\right] \geq \chi^2(2m; 1-\alpha) \mid \mathcal{D}\right] \xrightarrow{d} \bar{\Phi}(z_{1-\alpha} + \sqrt{\gamma}W). \tag{5}$$

Note that the above asymptotic limits are independent of the distribution of $\hat{s}(X)$. In Appendix A, we prove that Theorem 1 holds for a broad class of combination tests based on $\sum_{i=1}^{n} G(\hat{u}^{(\mathrm{marg})}(X_{2n+i}))$, provided that $G(U)$ has finite moments for $U \sim \mathrm{Unif}([0,1])$; Fisher's combination test is a special case with $G(u) = -2\log u$ and $G(U) \sim \chi^2(2)$.

Since $\gamma > 0$, the marginal type-I error is always larger than α whenever $\alpha < 0.5$. For illustration, consider $\alpha = 5\%$. When $\gamma = 3$, the marginal type-I error is as large as 20.5%; when $\gamma \to \infty$, the marginal type-I error is approaching 50%. Similarly, by (5), the 90-th percentile of the conditional type-I error converges to the 90-th percentile of $\bar{\Phi}(z_{1-q} + \sqrt{\gamma}W)$, which is $\bar{\Phi}(z_{0.95} + \sqrt{\gamma}z_{0.1})$. When $\gamma = 3$, the limit is 71.7%; when $\gamma \to \infty$, the limit is approaching 100%. This demonstrates the substantial adverse effect of dependence among marginal conformal p-values for Fisher's combination test.

Corrections of Fisher's combination test are possible for some dependence structures. By Lemma 1, the variance of the combination statistic is inflated by a factor $(1+\gamma)$ compared to that of the $\chi^2(2m; 1-\alpha)$ distribution (see Appendix A.1 for details). This yields an intuitive correction which divides the combination statistic by $\sqrt{1+\gamma}$. Surprisingly, this correction is asymptotically too conservative for marginal conformal p-values. We prove in Appendix A.2 (Theorem 5) that a valid correction rejects the global null if

$$\frac{-2\sum_{i=1}^{m}\log\left[\hat{u}^{(\mathrm{marg})}(X_{2n+i})\right] + 2(\sqrt{1+\gamma}-1)m}{\sqrt{1+\gamma}} \geq \chi^2(2m; 1-\alpha). \tag{6}$$

In Appendix A.2, we also confirm the validity of (6) via Monte-Carlo simulations and show this is asymptotically equivalent to the correction proposed by [68, 69] to address p-value dependence in more general contexts.

2.2 A positive result: conformal p-values are positively dependent

Certain multiple testing methods, such as the Benjamini-Hochberg procedure, are known to be robust to a particular type of mutual p-value dependence called *positive regression dependent on a subset* (PRDS) [14].

Definition 1 (PRDS)**.** *A random vector $X = (X_1, \ldots, X_m)$ is PRDS if for any $i \in \{1, \ldots, m\}$ and any increasing set D, the probability $\mathbb{P}[X \in D \mid X_i = x]$ is increasing in x.*

Above, for vectors a and b of equal dimension, we say $a \succeq b$ if every coordinate of a is no smaller than the corresponding coordinate of b, and a set $D \subset \mathbb{R}^m$ is *increasing* if $a \in D$ and $b \succeq a$ implies $b \in D$. The PRDS property is a demanding form of positive dependence which can be interpreted, loosely speaking, as saying all pairwise correlations are positive. In view of the definition of marginal p-values in (3) and the result in Lemma 1, it should be intuitive that larger scores in the calibration set make the p-values for all test points simultaneously smaller, and vice-versa. This idea is formalized by the following result proving marginal conformal p-values are PRDS.

Theorem 2 (Conformal p-values are PRDS)**.** *Assume that $\hat{s}(X)$ is continuous. Consider m test points $X_{2n+1}, \ldots, X_{2n+m}$ such that the first $m' \leq m$ of them are inliers, jointly independent of each other and of the data in $\mathcal{D}$. Then, the marginal conformal p-values $(\hat{u}^{(\mathrm{marg})}(X_{2n+1}), \ldots, \hat{u}^{(\mathrm{marg})}(X_{2n+m'}))$ are PRDS.*

When $\hat{s}(X)$ is not continuous, we can also prove the PRDS property by modifying the definition (3) of marginal conformal p-values; see Appendix A.3 for details. It follows from Theorem 2 that marginal conformal p-values can be used to control the FDR with the Benjamini-Hochberg procedure for the null hypotheses

$$H_{0,i} : X_i \sim P_X, \qquad i \in \{2n+1, \ldots, 2n+m\}.$$

Corollary 1 (Benjamini and Yekutieli [14])**.** *In the setting of Theorem 2, the Benjamini-Hochberg procedure applied at level $\alpha \in (0,1)$ to $(\hat{u}^{(\text{marg})}(X_{2n+1}), \dots, \hat{u}^{(\text{marg})}(X_{2n+m}))$ controls the FDR at level $\pi_0\alpha$, where π_0 is the proportion of true nulls. That is,*

$$\mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{\max\{1, |\mathcal{R}|\}}\right] \leq \pi_0\alpha \leq \alpha, \tag{7}$$

where $\mathcal{H}_0 = \{i : H_{0,i} \text{ holds}\} \subseteq \{2n+1, \dots, 2n+m\}$ is the subset of true inliers in the test set, and $\mathcal{R} \subseteq \{2n+1, \dots, 2n+m\}$ is the subset of test points reported as likely outliers.

This proves the FDR can be controlled, although only on average over the calibration data because the above expectation is taken over both $\mathcal{D}$ and the future test points. While such marginal guarantee may be satisfactory for someone carrying out several independent applications, individual practitioners committed to a single calibration data set may prefer stronger results.

2.3 A positive result: Storey's correction does not break FDR control

When the proportion of nulls is much smaller than 1, as it may be the case in many out-of-distribution detection problems, the Benjamini-Hochberg procedure is conservative, as shown in Corollary 1. If π_0 is known, a simple remedy is to replace the target FDR level with α/π_0. However, π_0 is rarely known in practice and hence it needs to be estimated. Given p-values p_i for all null hypotheses, it was proposed by Storey et al. in [17, 70] to estimate π_0 as

$$\hat{\pi}_0 = \frac{1 + \sum_{i=1}^{m} I(p_i > \lambda)}{m(1-\lambda)},$$

and then to apply the Benjamini-Hochberg procedure at level $\alpha/\hat{\pi}_0$; see Appendix A.4 for details. If the null p-values are super-uniform in the sense of (2), mutually independent, and independent of the non-null p-values, this provably controls the FDR in finite samples [17]. However, unlike in its standard version, the Benjamini-Hochberg procedure with Storey's correction may fail to control the FDR if the p-values are PRDS; see Section 6.3 of [71].

Surprisingly, we show below that the positive correlation (Lemma 1) among the marginal conformal p-values does not break the FDR control at all. The proof of Theorem 3 rests on a novel FDR bound for the Benjamini-Hochberg procedure with Storey's correction applied to PRDS p-values (Theorem 6 in Appendix A.4).

Theorem 3. *Set $\lambda = K/(n+1)$ for any integer K. Assume $\hat{s}(X)$ is continuous. In the setting of Corollary 1, the Benjamini-Hochberg procedure with Storey's correction applied at level $\alpha \in (0,1)$ to the marginal p-values $(\hat{u}^{(\text{marg})}(X_{2n+1}), \dots, \hat{u}^{(\text{marg})}(X_{2n+m}))$ controls the FDR at level α. That is,*

$$\mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{\max\{1, |\mathcal{R}|\}}\right] \leq \alpha. \tag{8}$$

3 Calibration-conditional conformal p-values

3.1 Warm up: analyzing the false positive rate

Having noted that conformal inferences hold in theory only marginally over the calibration data, the first question one may ask is: how bad can these inferences be conditional on a particular calibration set? We will address this question by developing high-probability bounds for the conditional deviation from uniformity of marginal p-values, starting here from the simplest case of pointwise bounds. The purpose of a pointwise bound is to control the probability that a null p-value (corresponding to a true inlier) is smaller than α, conditional on $\mathcal{D}$, for some *fixed* threshold $\alpha \in (0,1)$. In other words, we wish to understand the conditional false positives rate (FPR) corresponding to the threshold α,

$$\text{FPR}(\alpha; \mathcal{D}) := \mathbb{P}\left[\hat{u}^{(\text{marg})}(X_{2n+1}) \leq \alpha \mid \mathcal{D}\right], \tag{9}$$

beyond what we know from the marginal guarantee in (2), which is $\mathbb{E}[\text{FPR}(\alpha; \mathcal{D})] \leq \alpha$. The quantity in (9) can be studied precisely with existing results due to [38]. We revisit this topic here because it serves as an intuitive introduction to the more involved high-probability bounds that we will propose later.

Looking at the definition of $\hat{u}^{(\text{marg})}(X)$ in (3), we see that, if $\hat{s}(X)$ has a continuous distribution,

$$\text{FPR}(\alpha; \mathcal{D}) = F\left(\hat{F}^{-1}\left(\frac{(n+1)\alpha}{n}\right)\right),$$

where F and $\hat{F}$ are, respectively, the true and empirical (evaluated on the calibration data) CDF of $\hat{s}(X)$. Therefore, the deviation of $\text{FPR}(\alpha; \mathcal{D})$ (a random variable depending on $\mathcal{D}$) from α depends on the quality of $\hat{F}^{-1}((n+1)\alpha/n)$ as an approximation of $F^{-1}(\alpha)$, which can be understood through classical results for the order statistics of uniform variables.

Proposition 1 (Pointwise FPR of marginal conformal p-values, adapted from [38])**.** *Let* $\ell = \lfloor (n+1)\alpha \rfloor$. *If* $\hat{s}(X)$ *is continuous,* $\text{FPR}(\alpha; \mathcal{D})$ *follows a* $\text{Beta}(\ell, n+1-\ell)$ *distribution.*

Figure 2 visualizes the FPR distribution from Proposition 1 for different values of the calibration set size. This shows precisely how a smaller $\mathcal{D}^{\text{cal}}$ makes marginal p-values more conservative on average, but also more likely to be overly liberal on occasion. For example, we can see there is a non-negligible probability that $\text{FPR}(0.1; \mathcal{D}) > 0.15$ with 100 calibration points, whereas it seems very unlikely that $\text{FPR}(0.1; \mathcal{D}) > 0.12$ with 1600 calibration points. However, it is still quite possible that $\text{FPR}(0.01; \mathcal{D}) > 0.015$ even with 1600 calibration points. In general, Proposition 1 implies the coefficient of variation (relative spread) of the FPR is approximately proportional to $(|\mathcal{D}^{\text{cal}}|\alpha)^{-1/2}$. While this result is informative, it is limited for our purposes because it provides only a pointwise bound—it takes α as fixed—whereas uniform bounds are needed to construct conditionally valid p-values that can be safely used with any multiple-testing procedure, as discussed in the next section.

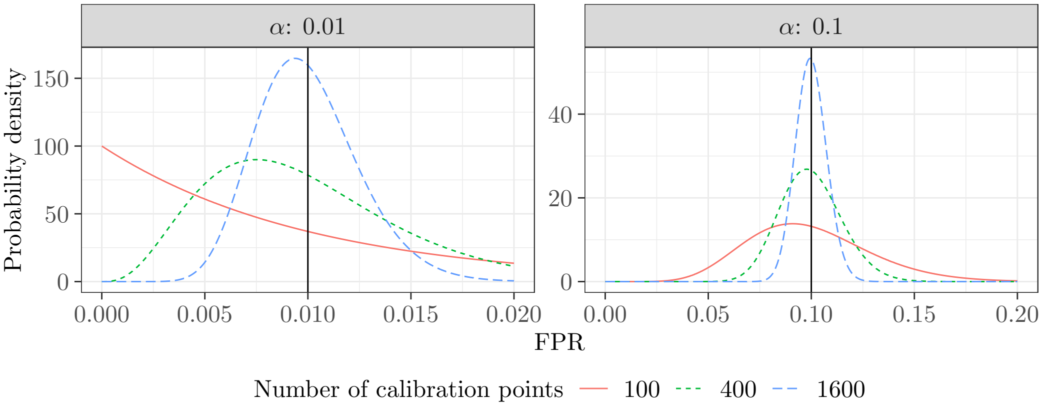

Figure 2: Distribution of the false positive rate obtained by thresholding marginal conformal p-values at levels $\alpha = 0.01$ and $\alpha = 0.1$, as a function of the number of calibration points.

Before presenting our novel high-probability bounds, let us pause for a moment to emphasize that Proposition 1 is interesting beyond the scope of outlier detection. In fact, this result clarifies the issue of how to best choose the number of calibration data points in general applications of split-conformal inference; see [72] for an empirical demonstration of this issue in a regression context, for example. In fact, while we chose to present the result in Proposition 1 in terms of FPR because this paper focuses on outlier detection, it is immediate to recast it in terms of predictive sets by leveraging the mirror-image symmetry property of the beta distribution, i.e., $1 - X \sim \text{Beta}(b, a)$ if $X \sim \text{Beta}(a, b)$. (See Section 4.2 for more details about the connection to predictive sets.)

3.2 A generic strategy to adjust marginal conformal p-values

Proposition 1 implies marginal conformal p-values may be anti-conservative conditional on $\mathcal{D}$. Therefore, in the language of (1), our goal is to find an adjustment function leading to conditionally valid p-values, i.e., satisfying (4). The following theorem suggests a generic strategy through a simultaneous upper confidence bound for order statistics.

Theorem 4 (Conditional p-value adjustment)**.** *Let* $U_1, \ldots, U_n \overset{\text{i.i.d.}}{\sim} \text{Unif}([0,1])$*, with order statistics* $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(n)}$*, and fix any* $\delta \in (0,1)$*. Suppose* $0 \leq b_1 \leq b_2 \leq \ldots \leq b_n \leq 1$ *are* n *reals such that*

$$\mathbb{P}\left[U_{(1)} \leq b_1, \ldots, U_{(n)} \leq b_n\right] \geq 1 - \delta. \tag{10}$$

Let also $b_0 = 0, b_{n+1} = 1$*, and* $h : [0,1] \mapsto [0,1]$ *be a piece-wise constant function such that*

$$h(t) = b_{\lceil (n+1)t \rceil}, \ t \in [0,1].$$

Then, $\hat{u}^{(\text{ccv})} = h \circ \hat{u}^{(\text{marg})}$ *satisfies* (4)*, i.e.,* $\hat{u}^{(\text{ccv})}(X_{2n+1})$ *is a calibration-conditional valid p-value.*

Figure 3 illustrates the idea of Theorem 4. Here, we set $n = 500$ and generate 100 independent realizations of the order statistics $(U_{(1)}, \ldots, U_{(n)})$. Each of the 100 blue curves corresponds to a sample path, plotted against the normalized index i/n. The black curve tracks the theoretical mean of $(U_{(1)}, \ldots, U_{(n)})$, and the red curve corresponds to a particular a sequence of b_i values derived from the generalized Simes inequality (detailed in the next subsection) for $\delta = 0.1$. We observe relatively few sample paths cross the red curve, and all crossings occur at small indices. This suggests the upper confidence bounds provided by Theorem 4 can be especially tight for lower indices of the order statistics, which is essential to obtain reasonably powerful CCV p-values for outlier detection. Of course, calibration-conditional validity still necessarily comes at some power cost. For example, a marginal p-value of $\hat{u}^{(\text{marg})}(X) = 25/(n+1) \approx 0.05$ results in a CCV p-value of $h(25/(n+1)) = b_{25} \approx 0.075$ in this case.

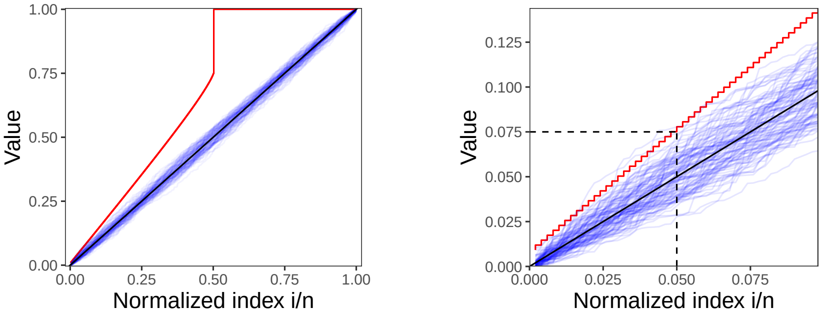

Figure 3: Illustration of Theorem 4. The red curve gives the sequence derived by generalized Simes inequality (Proposition 2) with $k = n/2 = 250$. The right panel zooms in on small indices.

3.3 Generalized Simes Inequality

The larger p-values typically do not matter in multiple testing problems, as it is the small ones that determine which hypotheses are rejected. Therefore, to maximize power, we would like the b_i values in Theorem 4 to be as small as possible for low indices i, while we may be satisfied with letting $b_i = 1$ for large i. The generalized Simes inequality yields a desirable class of $(b_1, \ldots, b_n)$ sequences with this property.

Proposition 2 (Generalized Simes Inequality, from Equation (3.5) in [73]). *For any positive integer $k \leq n$, the uniform bound* (10) *in Theorem 4 holds with*

$$b_{n+1-i} = 1 - \delta^{1/k} \left(\frac{i \cdots (i-k+1)}{n \cdots (n-k+1)} \right)^{1/k}, \quad i = 1, \ldots, n.$$

The original motivation of [73] was to compute thresholds for step-up procedure to achieve k-FWER control; there, the parameter k was set to be a small integer. Here, we exploit Proposition 2 differently, choosing $k = n/2$ so that the b_i values with lower indices i are as small as possible while those with larger indices i may be uninformative (note that $b_{n-k+2} = \ldots = b_n = 1$). In particular, our choice corresponds to

$$b_1 = 1 - \delta^{2/n} = 1 - \exp\left\{ -\frac{2\log(1/\delta)}{n} \right\} \approx \frac{2\log(1/\delta)}{n}.$$

Therefore, the smallest possible marginal p-value equal to $1/(n+1)$ would be mapped to $h(1/(n+1)) \approx 2\log(10)/n = 4.61/n$, if $\delta = 0.1$, for example, since $\hat{u}^{(\text{ccv})}(X) = h(\hat{u}^{(\text{marg})}(X))$. If $n = 10000$, then $h(1/(n+1)) < 0.0005$, which is larger than the marginal p-value, but much smaller than what one would obtain from other standard uniform bounds. For example, the DKWM inequality [74, 75] would imply a result similar to that of Proposition 2 but with $b_i = \min\{(i/n) + \sqrt{\log(2/\delta)/2n}, 1\}$; this would map the smallest possible marginal p-value to $1/(n+1) + \sqrt{\log(2/\delta)/2n} > 0.1$, in the above example. The comparison between the generalized Simes inequality and the DKWM inequality is expanded Appendix B, where we also consider an additional uniform bound based on the linear-boundary crossing probability for the empirical CDF [76]. This comparison confirms the generalized Simes inequality yields the most powerful adjustment for our multiple testing purposes. In practice, we find that $k = n/2$ works well, as motivated empirically in Appendix C. (Note that larger values of k would lower further the smallest possible adjusted p-value, but at the cost of raising other small p-values).

4 Extensions beyond conformal p-values

4.1 Simultaneous confidence bounds for the false positive rate

Some practitioners may be accustomed to thinking about outlier detection in terms of FPR—the probability of incorrectly reporting as outlier any true inlier—rather than p-values. In particular, they may wonder what the FPR can be if they report X_{2n+1} as likely to be an outlier whenever the classification score $\hat{s}(X_{2n+1})$ (computed by some black-box outlier detection algorithm) is above a threshold t, as a function of t, so that they may choose a posteriori which value of t to adopt. This question is closely related to the problem of constructing CCV p-values, so our method provides an answer. In fact, the next result shows Theorem 4 also yields a simultaneous upper confidence bound for the CDF.

Proposition 3. *Let F denote the true CDF of some distribution from which n i.i.d. samples, $Z_1, \ldots, Z_n$, are drawn, and denote by $\hat{F}_n$ the corresponding empirical CDF. With the same notation as in Theorem 4,*

$$\mathbb{P}\left[F(z) \leq h(\hat{F}_n(z)), \ \forall z \in \mathbb{R} \right] \geq 1 - \delta. \tag{11}$$

Applying Proposition 3 to the CDF of the scores $\hat{s}$ computed by any one-class classification algorithm provides a uniform upper confidence bound for its FPR, namely $\text{FPR}(t) := \mathbb{P}\left[\hat{s}(X_{2n+1}) \leq t\right]$, as a function of the detection threshold t. In words, this guarantees that reporting as outliers an observation with black-box score equal to z is likely (with probability at least $1-\delta$) to result in a FPR no greater than $h(\hat{F}_n(z))$, where $\hat{F}_n(z)$ is the empirical CDF of the analogous scores computed on a calibration data set of size n. Figure 4 shows a practical example of this upper bound based on the empirical distribution of scores evaluated on 1000 calibration points, with $\delta = 0.1$ and $k = n/2$ (the exact details of this example are the same as those of the numerical experiments presented later in Section 5.2). For instance, this plot informs us that reporting as outliers future samples with scores below -0.5 is likely to result in an FPR below 0.025.

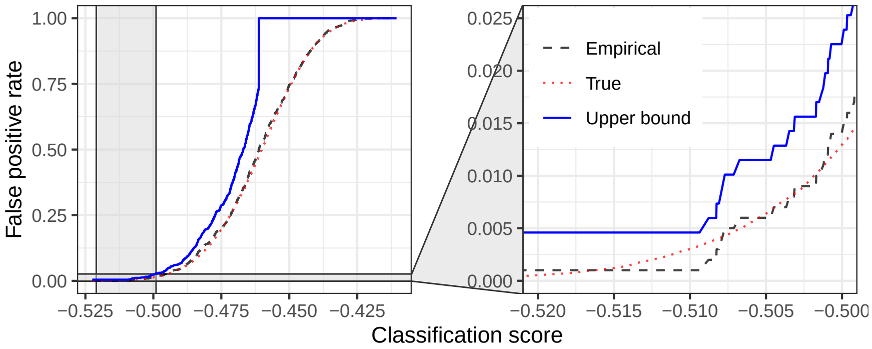

Figure 4: FPR calibration curve for an isolation forest one-class classifier on simulated data, as a function of the reporting threshold for the classification scores. The upper bound (solid blue) is guaranteed to lie above the true FPR curve (dotted red) with probability at least 90%. The dashed black curve corresponds to the empirical FPR. The panel on the right zooms in on small values (likely outliers).

Note that the construction of a uniform confidence band for an unknown CDF is a widely studied problem. For example, the DKWM inequality [74, 75] implies the bound in (11) with $h(z) = \min\{z + \sqrt{\log(2/\delta)/2n}, 1\}$. However, the DKWM bound is tightest at $z = 1/2$ and loose near 0, which would limit the power to detect outliers. Therefore, it is preferable for our purposes to have a function $h(z)$ that is as close as possible to the identity for small values of z, as discussed earlier in Section 3.3.

4.2 Simultaneously-valid prediction sets

Lastly, CCV p-values can be easily re-purposed to strengthen the marginal guarantees generally obtainable for conformal predictions. In particular, for each $\alpha \in (0, 1)$, one can define a predictive set

$$\hat{\mathcal{C}}^{\alpha} := \{x : \hat{u}^{(\text{ccv})}(x) > \alpha\}. \tag{12}$$

These sets are simultaneously valid for all α, conditional on the calibration data. That is, they satisfy

$$\mathbb{P}\left[\mathbb{P}\left[X_{2n+1} \in \hat{\mathcal{C}}^{\alpha} \mid \mathcal{D}\right] \geq 1 - \alpha \text{ for all } \alpha \in (0,1)\right] \geq 1 - \delta. \tag{13}$$

In words, if we use CCV p-values to construct prediction sets, the probability that a new observation falls within $\hat{\mathcal{C}}^{\alpha}$ is at least $1 - \alpha$, simultaneously for all $\alpha \in (0, 1)$ with high probability. This is stronger than the usual conformal guarantee, as the latter holds marginally over $\mathcal{D}$ and only for a single pre-specified α.

5 Numerical experiments

5.1 Setup

The following experiments are designed to simulate a world in which our methods are independently applied by J practitioners. Each practitioner $j \in [J]$ has an independent data set $\mathcal{D}_j$ (to train and calibrate the method), and L test sets $\mathcal{D}_{j,l}^{\text{test}}$ (to compute p-values and evaluate performance), each corresponding to different possible future scenarios $l \in [L]$. The data sets contain $2n$ observations each ($|\mathcal{D}_j| = 2n$), and the test sets contain n_{test} observations each ($|\mathcal{D}_{j,l}^{\text{test}}| = n_{\text{test}}$). Imagine that, from the practitioner's present point of view, the data set $\mathcal{D}_j$ is fixed but the test set is random, so that $\mathcal{D}_{j,l}^{\text{test}}$ represents the test set for practitioner j under future scenario l. Then, as discussed in Section 1.2, practitioner j is most interested in the FDR (or other measures of type-I errors, alternatively) conditional on $\mathcal{D}_j$, i.e., in the random variable

$$\text{cFDR}(\mathcal{D}_j) := \mathbb{E}\left[\text{FDP}(\mathcal{D}^{\text{test}}; \mathcal{D}_j) \mid \mathcal{D}_j\right],$$

where $\mathrm{FDP}(\mathcal{D}^{\text{test}}; \mathcal{D}_j)$ is the proportion of inliers among the test points reported as outliers, based on the procedure calibrated on $\mathcal{D}_j$. This motivates the definition of the following performance measures. For any $j \in [J]$, we compute

$$\widehat{\mathrm{cFDR}}(\mathcal{D}_j) := \frac{1}{L}\sum_{l=1}^{L} \mathrm{FDP}(\mathcal{D}^{\text{test}}_{j,l}; \mathcal{D}_j), \qquad \widehat{\mathrm{cPower}}(\mathcal{D}_j) := \frac{1}{L}\sum_{l=1}^{L} \mathrm{Power}(\mathcal{D}^{\text{test}}_{j,l}; \mathcal{D}_j), \tag{14}$$

where $\mathrm{Power}(\mathcal{D}^{\text{test}}_{j,l}; \mathcal{D}_j)$ is the proportion of outliers in $\mathcal{D}^{\text{test}}_{j,l}$ correctly identified as such by practitioner j.

Our experiments will demonstrate that the proposed simultaneous calibration method leads to sufficiently small $\widehat{\mathrm{cFDR}}(\mathcal{D}_j)$ for the desired fraction of practitioners, while the traditional point-wise calibration generally only leads to small values of the marginal FDR, namely $\widehat{\mathrm{mFDR}} := \frac{1}{J}\sum_{j=1}^{J} \widehat{\mathrm{cFDR}}(\mathcal{D}_j)$.

5.2 Outlier detection on simulated data

5.2.1 Data description

We begin to investigate the empirical performance of different methods for calibrating conformal p-values on synthetic data. The data are generated by sampling each data point $X_i \in \mathbb{R}^{50}$ from a multivariate Gaussian mixture model P_X^a, such that $X_i = \sqrt{a}\, V_i + W_i$, for some constant $a \geq 1$ and appropriate random vectors $V_i, W_i \in \mathbb{R}^{50}$. Here, V_i has independent standard Gaussian components, and each coordinate of W_i is independent and uniformly distributed on a discrete set $\mathcal{W} \subseteq \mathbb{R}^{50}$ with cardinality $|\mathcal{W}| = 50$. The vectors in $\mathcal{W}$ are sampled independently from the uniform distribution on $[-3, 3]^{50}$, before the beginning of our experiments, and then held constant thereafter. (Therefore, each coordinate of W_i is uniformly distributed on $[-3, 3]$, but it is not the case that the different W_i's are independent and identically distributed on $[-3, 3]^{50}$; instead, the fixed set $\mathcal{W}$ makes this a mixture model.)

The data sets $\mathcal{D}_j$ are sampled from P_X^a with $a = 1$ and $n = 1000$. The total $2n$ observations in each $\mathcal{D}_j$ are further divided into $n_{\text{train}} = 1000$ observations used to fit a one-class SVM classifier scoring function $\hat{s}$ (implemented in the Python package `scikit-learn` [77]), and $n_{\text{cal}} = 1000$ observations used to calibrate the conformal p-values, as in (1), leading to a valid p-value $\hat{u}(X_{n+1}) \in [0, 1]$ for any new data point X_{n+1}. The total number of data sets is $J = 100$, each of which is associated with $L = 100$ test sets. A random subset of the observations in each test set $\mathcal{D}^{\text{test}}_{j,l}$ is sampled from P_X^a with $a = 1$, while the others are outliers, in the sense that they are sampled from P_X^a with $a > 1$, as specified below.

5.2.2 Individual outlier detection

First, we focus on a data generating model under which 90% of the $n_{\text{test}} = 1000$ observations in each $\mathcal{D}^{\text{test}}_{j,l}$ are sampled from P_X^a with $a = 1$, and we seek to identify the remaining 10% of outliers. For this purpose, we calibrate a conformal p-value for all observations in $\mathcal{D}^{\text{test}}_{j,l}$, and then we apply the Benjamini-Hochberg procedure at some nominal FDR level α to account for the multiple comparisons, with and without Storey's correction based on the estimated null proportion. In the following, we apply our conditional calibration method with the parameters $\delta = 0.1$ and $k = n_{\text{cal}}/2$ (see below for comments about the choice of k).

Figure 5 shows the distribution of $\widehat{\mathrm{cFDR}}(\mathcal{D}_s)$ and $\widehat{\mathrm{cPower}}(\mathcal{D}_s)$, corresponding to $\alpha = 0.1$, for different values of the signal strength a (recall that here $a = 1$ corresponds to no signal), when the Benjamini-Hochberg procedure is utilized to account for the multiple comparisons. The results confirm the calibration-conditional p-values control the conditional FDR for at least 90% of practitioners, while the marginal p-values do not. In fact, marginal p-values only control the conditional FDR if the number of samples in the calibration data set is very large; see Figure A3, Appendix C. Furthermore, we note that both methods control the marginal FDR, as also predicted by our theoretical results. Figure A4 presents the results obtained by applying Storeys' correction to the Benjamini-Hochberg procedure, while Figure A5 summarizes additional experiments in which our conditional calibration method is applied with $\delta = 0.25$. Finally, Figure A6 visualizes the effect of different values of the Simes parameter k on our calibration-conditional p-value, showing that $k = n_{\text{cal}}/2$ works relatively well, although the performance does not appear to be extremely sensitive to this choice.

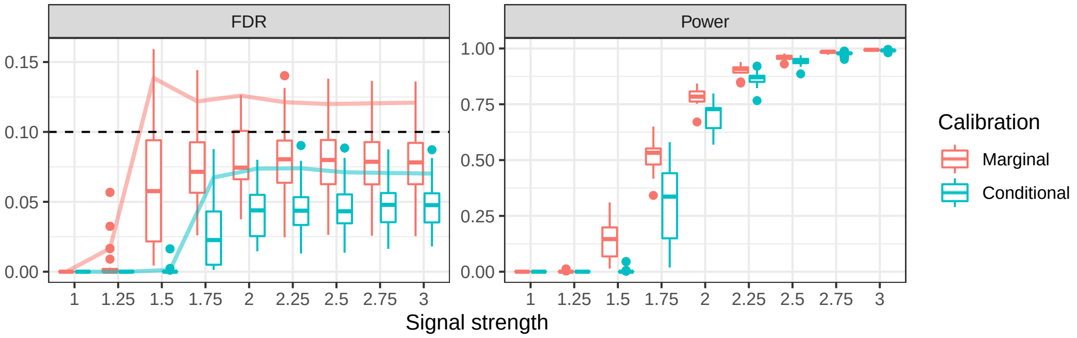

Figure 5: Performance of different methods for calibrating conformal p-values in a simulated outlier detection problem, as a function of the signal strength. The box plots visualize the distribution of FDR and power, as defined in (14), conditional on 100 independent data sets. The solid curves indicate the 90-th quantile of the conditional FDR distribution. The nominal FDR 0.1, and the conditional method is applied with $\delta = 0.1$.

5.2.3 Batch outlier detection

We now consider the global testing problem of detecting whether a batch of new observations contains any outliers. For this purpose, we follow the same approach as before, with the only difference that the $n_{\text{test}} = 1000$ observations in each test set are now sub-divided into 100 batches of size 10. The 10 calibrated p-values in each batch are combined with Fisher's method to test the batch-specific global null. Then, the Benjamini-Hochberg procedure with Storey's correction is applied to control the FDR over all batches. This simulation is designed such that 90% of the batches contain no outliers (i.e., all samples are drawn from P_X^a with $a = 1$), while 50% of the samples in the remaining batches are outliers (i.e., they are drawn from P_X^a with $a = 2$). Of course, batched testing is less informative than the precise identification of outliers discussed in the previous section, but the advantage now is that we can achieve higher power. Figure 6 shows that, even though this problem is relatively easy (the power is almost equal to 1), the use of marginal p-values may still lead to a conditional FDR that is noticeably higher than expected for many researchers. By contrast, simultaneous calibration appears to be conservative for all of them, without much sacrifice in power.

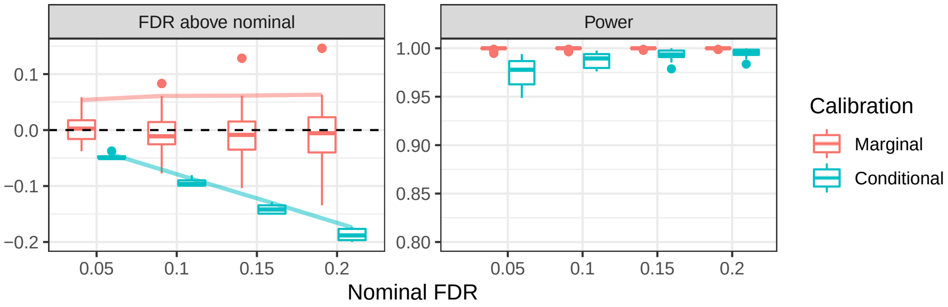

Figure 6: Performance of different methods for calibrating conformal p-values in a simulated outlier batch detection problem, as a function of the nominal FDR level. The excess FDR is defined as the difference between the empirical FDR and the nominal FDR. Note that both methods achieve power close to one in this example. Other details are as in Figure 5.

Finally, we study the effect of the batch size on the performance of different calibration methods under the global null hypothesis (i.e., when there are no outliers in the test set). As before, the p-values in each batch are combined with Fisher's method and the global null is rejected if the resulting p-value is smaller than 0.1. As before, the experiment is repeated for 100 independent data sets and 1000 test sets. Figure 7 shows that marginal p-values do not lead to valid inferences, especially if the batch size is large. By contrast, the calibration-conditional method always remains valid.

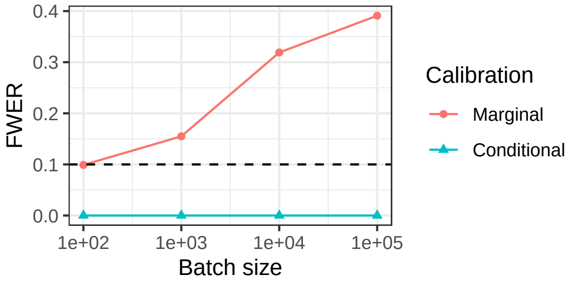

Figure 7: Family-wise error rate (FWER) in a simulated outlier batch detection problem under the global null hypothesis, using different calibration methods for the conformal p-values. The results are shown as a function of the batch size. The global null is rejected if the Fisher's combined p-value is below 0.1, which means the nominal FWER is 10% (horizontal dashed line).

5.3 Outlier detection on real data

5.3.1 Data description

Table 1: Summary of the benchmark data sets for outlier detection utilized in our applications.

	ALOI [78, 79]	Cover [80]	Credit card [81]	KDDCup99 [78, 82]	Mammography [83]	Digits [84]	Shuttle [85]
Features d	27	10	30	40	6	16	9
Inliers n_{inliers}	283301	286048	284315	47913	10923	6714	45586
Outliers n_{outliers}	1508	2747	492	200	260	156	3511

We turn to study the performance of the calibration schemes from Section 5.2 on several benchmark data sets for outlier detection, summarized in Table 1. As before, the Simes simultaneous calibration is applied with $\delta = 0.1$ and $k = n_{\text{cal}}/2$. We utilize an isolation forest [86] machine-learning algorithms $\hat{s}$ as the base method for detecting anomalies, available in the Python `sklearn` package. We rely on the default hyper-parameters, except for the 'contamination' parameter which we set equal to 0.1. Additional experiments based on one-class SVM and Local Outlier Factor (LOF) algorithms are presented in Appendix C (Tables A2–A3).

5.3.2 Individual outlier detection

Here, we follow the experimental setup of Section 5.2.2. The difference is that we need to construct multiple training, calibration, and test sets by randomly splitting the n_{inlier} inlier examples into three disjoint subsets of size n_{train}, n_{cal} and n_{test}, respectively. A total of $n_{\text{inlier}}/2$ data points is used for training and calibration, i.e., $n_{\text{train}} + n_{\text{cal}} = n_{\text{inlier}}/2$ with $n_{\text{cal}} = \min\{2000, n_{\text{train}}/2\}$, while outlier examples are only included in the test sets. For each training/calibration data subset, we sample 100 test sets of size $n_{\text{test}} = \min\{2000, n_{\text{train}}/3\}$. Each test set contains 90% of randomly chosen inliers, and 10% of outliers. It should be noted that, in contrast to the simulated experiments of Section 5.2.2 in which the data were effectively infinitely abundant, there is some overlap between the samples in different test sets.

Figure 8 compares the performance of marginal and simultaneously calibrated p-values on the credit card data set [81], as a function of the nominal FDR level. Here, the Benjamini-Hochberg procedure is applied with Storey's correction. Note that the proposed Simes simultaneous calibration leads to FDR control for at least 90% of simulated practitioners, as expected. This stands in contrast with the marginal calibration approach, which controls the FDR only marginally.

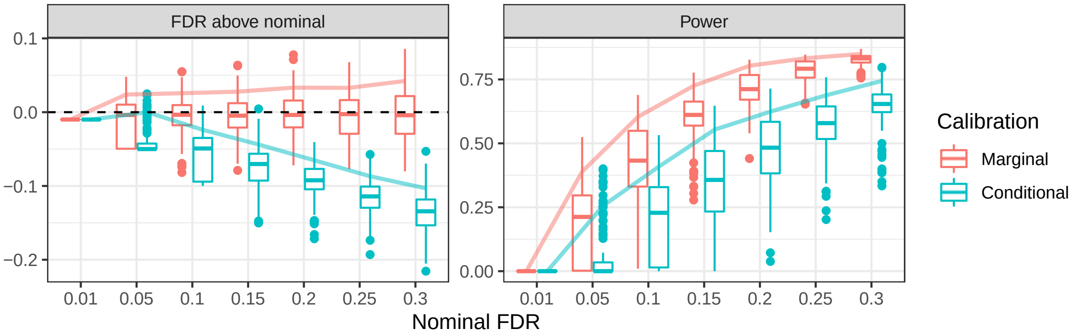

Figure 8: Outlier detection performance on credit card fraud data. Conformal p-values based on an isolation forest model are calibrated using different methods. The Benjamini-Hochberg procedure with Storey's correction is then applied to control the FDR over the set of test points. Other details are as in Figure A4.

Consistent conclusion can be drawn from Table 2, which compares the two calibration procedures on all benchmark data sets at the nominal FDR level of 0.2. Additional results corresponding to different outlier detection algorithms (one-class SVM and LOF) can be found in Table A1, Appendix C.2. In all cases, we adopt the `sklearn` default parameters. Finally, Table A2 summarizes the performance of different calibration and detection methods across all data sets when the Benjamini-Hochberg procedure is applied without Storey's correction.

Table 2: Outlier detection performance on different data sets, using alternative methods for calibrating conformal p-values. The FDR and power diagnostics are defined conditional on the training and calibration data, as defined in Section 5.1. The nominal marginal FDR level is 0.2. Empirical FDR values larger than the nominal level are colored in orange; values at least one standard deviation above it are colored in red.

	FDR				Power			
	Mean		90th percentile		Mean		90-th quantile	
Dataset	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.
ALOI	0.025	0	0.02	0	0	0	0	0
Cover	0.08	0.013	0.277	0.049	0.008	0.002	0.03	0.004
Credit card	0.197	0.106	0.233	0.135	0.712	0.469	0.803	0.624
KDDCup99	0.196	0.105	0.234	0.135	0.755	0.62	0.825	0.713
Mammography	0.18	0.031	0.282	0.112	0.167	0.036	0.342	0.155
Digits	0.177	0.029	0.27	0.116	0.347	0.056	0.603	0.213
Shuttle	0.196	0.107	0.234	0.138	0.981	0.976	0.984	0.981

5.3.3 Batch outlier detection

We now focus on global testing for outlier batch detection, similarly to Section 5.2.3. The available data are divided into training, calibration, and test sets according to the same scheme as in Section 5.3.2; the only difference is that the size of the test sets is now equal to 1000, so as to follow as closely as possible the same experimental protocol as in Section 5.2.3.

Figure 9 compares the performance of the different calibration methods as a function of the nominal FDR level. The p-values in each batch are combined with Fisher's method, and then the Benjamini-Hochberg procedure is applied with Storey's correction. Again, we observe that simultaneous calibration is required to ensure the conditional FDR is controlled in at least 90% of the applications, although it involves some power loss. Both calibration methods control the marginal FDR.

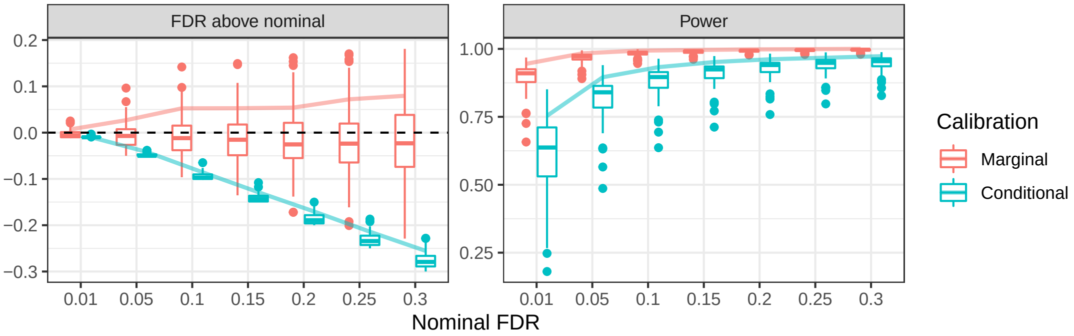

Figure 9: Outlier batch detection performance on credit card fraud data. Conformal p-values are computed based on an isolation forest model and calibrated using different methods. Other details are as in Figure 6.

Table 3 summarizes the performance of the two alternative calibration methods on all data sets. Here, the nominal FDR level is 0.1 and the Benjamini-Hochberg procedure is applied with the Storey correction. Again, the results show that the Simes method controls the conditional FDR 90% of the time, although at some cost in power, while the marginal calibration method does not. See Table A3, Appendix C.2 for additional results that, in addition to the isolation forest, include also the one-class SVM and LOF algorithms for outlier detection. Finally, Table A4 summarizes performance of the different methods on all data sets when the Benjamini-Hochberg procedure is applied without the Storey correction.

Table 3: Outlier batch detection performance on different data sets, using alternative methods for calibrating conformal p-values. The nominal FDR level is 0.1. Other details are as in Table 2.

	FDR				Power			
	Mean		90-th quantile		Mean		90-th quantile	
Data set	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.
ALOI	0.072	0.003	0.178	0	0.002	0	0.004	0
Cover	0.092	0.006	0.183	0.01	0.18	0.017	0.359	0.034
Credit card	0.092	0.005	0.153	0.014	0.983	0.885	0.993	0.933
KDDCup99	0.088	0.005	0.129	0.013	0.999	0.979	1	0.994
Mammography	0.072	0.004	0.126	0.016	0.61	0.21	0.765	0.361
Digits	0.09	0.005	0.148	0.014	0.97	0.626	0.999	0.836
Shuttle	0.087	0.006	0.137	0.013	1	1	1	1

6 Discussion

This paper has studied the multiple testing problem for outlier detection using conformal p-values. Conformal p-values provide a natural approach to outlier detection (when clean training data are available) with the advantage of being able to leverage any black-box machine-learning tool, producing fully non-parametric inferences that are provably valid in finite samples and require no modeling beyond the i.i.d. assumption. Of course, a possible limitation (or perhaps strength, depending on the viewpoint) of conformal inference is that its agnosticism prevents very confident statements, as conformal p-values can never be smaller than

$1/(n+1)$, where n is the number of clean data points available for calibration. Therefore, this solution may not be as powerful as likelihood-based approaches, especially if the signals are strong but sparse. However, it does seem preferable if clean data are available but accurate models are not.

Whenever the conformal framework is appropriate for a particular outlier detection application, the problem of multiple testing considered in this paper is likely to be relevant, as it often the case that possible outliers are to be detected among many possible inlier test points, and reporting an excess of false discoveries would be undesirable. Our work brings attention to the delicacy of such task, showing that the mutual dependence of conformal p-values breaks certain methods (e.g., Fisher's combination test) and makes the validity of others (e.g., the Benjamini-Hochberg procedure) not obvious. In particular, we find our PRDS result interesting because this property is well-known as a theoretical assumption for FDR control, but it is typically difficult to verify in practical applications [14, 15].

Our methodological contribution is a technique based on high-probability bounds to compute calibration-conditional conformal p-values that are mutually independent and can thus be directly trusted in any multiple testing procedure. Our bounds are stronger than those in the previous conformal inference literature because they are simultaneous in nature and, consequently, they can also be useful for practitioners to tune a posteriori the significance threshold for machine-learning statistics above which to report their discoveries. Unsurprisingly, our simulations demonstrate that calibration-conditional inferences are less powerful on average than marginal conformal inferences; therefore, the additional comfort of their stronger guarantees should be weighted against the potential loss of some interesting findings. Nonetheless, we prefer to leave such considerations to practitioners on a case-by-case basis, as our objective here was simply to explain the theoretical properties and general relative advantages of different statistical methods.

Finally, this work opens new directions for future research. For example, focusing on split-conformal p-values, we did not study other hold-out approaches, such as the jackknife+ [50] or bootstrap sampling [51], that may practically yield higher power, although they are also more computationally expensive. A separate line of research may focus on relaxing the i.i.d. assumption to improve power in a multiple testing setting with structured outliers [87]. In fact, our theory naturally allows for some degree of dependence among the test points, as long as the inliers are independent of each other and of the outliers. Furthermore, we mentioned but did not explore the possible connection between our multiple outlier testing problem (especially regarding our results on Fisher's combination method) and classical two-sample testing. Finally, the high-probability bounds developed here may prove useful for purposes other than the calibration of conformal p-values; for instance, we already discussed a straightforward extension to obtain simultaneously valid prediction sets, but other possible applications may involve predictive distributions [88] and functionals thereof [89], or the comparison of different machine-learning algorithms in terms of estimated generalization error [90, 91], for example.

Software availability

A software implementation of the methods described in this paper is available online, in the form of a Python package, at `https://github.com/msesia/conditional-conformal-pvalues.git`, along with usage examples and notebooks to reproduce our numerical experiments.

Acknowledgements

S.B. gratefully acknowledges the support of the Ric Weiland fellowship. E.C. was supported by Office of Naval Research grant N00014-20-12157, by the National Science Foundation grants OAC 1934578 and DMS 2032014, by the Army Research Office (ARO) under grant W911NF-17-1-0304, and by the Simons Foundation under award 814641. L.L. gratefully acknowledges the support of the National Science Foundation grants OAC 1934578, the Discovery Innovation Fund for Biomedical Data Sciences, and the NIH grant R01MH113078. Y.R. thanks the Career Advancement Fellowship, Technion, for providing research support.

References

[1] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. "Novelty detection for the identification of masses in mammograms". In: *1995 Fourth International Conference on Artificial Neural Networks*. IET. 1995, pp. 442–447.

[2] A. Patcha and J.-M. Park. "An overview of anomaly detection techniques: Existing solutions and latest technological trends". In: *Computer networks* 51.12 (2007), pp. 3448–3470.

[3] F. Fortunato, L. Anderlucci, and A. Montanari. "One-class classification with application to forensic analysis". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 69.5 (2020), pp. 1227–1249.

[4] L. Tarassenko, D. A. Clifton, P. R. Bannister, S. King, and D. King. "Novelty Detection". In: *Encyclopedia of Structural Health Monitoring*. American Cancer Society, 2009.

[5] D. Hendrycks and K. Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *arXiv preprint arXiv:1610.02136* (2016).

[6] S. Liang, Y. Li, and R. Srikant. "Enhancing the reliability of out-of-distribution image detection in neural networks". In: *arXiv preprint arXiv:1706.02690* (2017).

[7] K. Lee, K. Lee, H. Lee, and J. Shin. "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks". In: *NeurIPS*. 2018.

[8] K. Lee, H. Lee, K. Lee, and J. Shin. "Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples". In: *International Conference on Learning Representations*. 2018.

[9] M. M. Moya, M. W. Koch, and L. D. Hostetler. "One-class classifier networks for target recognition applications". In: *NASA STI/Recon Technical Report N* 93 (1993), p. 24043.

[10] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. "A review of novelty detection". In: *Signal Processing* 99 (2014), pp. 215–249.

[11] V. Vovk, A. Gammerman, and C. Saunders. "Machine-learning applications of algorithmic randomness". In: *International Conference on Machine Learning*. 1999, pp. 444–453.

[12] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[13] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[14] Y. Benjamini and D. Yekutieli. "The control of the false discovery rate in multiple testing under dependency". In: *Annals of Statistics* (2001), pp. 1165–1188.

[15] S. Clarke, P. Hall, et al. "Robustness of multiple testing procedures against dependence". In: *Annals of Statistics* 37.1 (2009), pp. 332–358.

[16] R. Fisher. *Statistical methods for research workers*. Oliver & Boyd (Edinburgh), 1925.

[17] J. D. Storey, J. E. Taylor, and D. Siegmund. "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.1 (2004), pp. 187–205.

[18] S. S. Wilks. "Multivariate statistical outliers". In: *Sankhyā: The Indian Journal of Statistics, Series A* (1963), pp. 407–426.

[19] D. M. Hawkins. *Identification of outliers.* Vol. 11. Springer, 1980.

[20] M. Riani, A. C. Atkinson, and A. Cerioli. "Finding an unknown number of multivariate outliers". In: *Journal of the Royal Statistical Society: series B (statistical methodology)* 71.2 (2009), pp. 447–466.

[21] A. Cerioli. "Multivariate outlier detection with high-breakdown estimators". In: *Journal of the American Statistical Association* 105.489 (2010), pp. 147–156.

[22] S. S. Khan and M. G. Madden. "One-class classification: taxonomy of study and review of techniques". In: *The Knowledge Engineering Review* 29.3 (2014), pp. 345–374.

[23] S. Agrawal and J. Agrawal. "Survey on anomaly detection using data mining techniques". In: *Procedia Computer Science* 60 (2015), pp. 708–713.

[24] C. C. Aggarwal. "Outlier analysis". In: *Data mining.* Springer. 2015, pp. 237–263.

[25] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli. "Adversarially learned one-class classifier for novelty detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2018, pp. 3379–3388.

[26] R. Chalapathy and S. Chawla. "Deep learning for anomaly detection: A survey". In: *preprint at arXiv:1901.03407* (2019).

[27] R. Laxhammar and G. Falkman. "Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories". In: *Annals of Mathematics and Artificial Intelligence* 74.1-2 (2015), pp. 67–94.

[28] J. Smith, I. Nouretdinov, R. Craddock, C. Offer, and A. Gammerman. "Conformal anomaly detection of trajectories with a multi-class hierarchy". In: *International symposium on statistical learning and data sciences.* Springer. 2015, pp. 281–290.

[29] V. Ishimtsev, A. Bernstein, E. Burnaev, and I. Nazarov. "Conformal k-NN Anomaly Detector for Univariate Data Streams". In: *Conformal and Probabilistic Prediction and Applications.* PMLR. 2017, pp. 213–227.

[30] L. Guan and R. Tibshirani. "Prediction and outlier detection in classification problems". In: *arXiv preprint arXiv:1905.04396* (2019).

[31] F. Cai and X. Koutsoukos. "Real-time Out-of-distribution Detection in Learning-Enabled Cyber-Physical Systems". In: *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS).* IEEE. 2020, pp. 174–183.

[32] M. Haroush, T. Frostig, R. Heller, and D. Soudry. "Statistical Testing for Efficient Out of Distribution Detection in Deep Neural Networks". In: *arXiv preprint arXiv:2102.12967* (2021).

[33] V. Vovk, I. Nouretdinov, and A. Gammerman. "Testing Exchangeability On-Line." In: Jan. 2003, pp. 768–775.

[34] V. Fedorova, A. Gammerman, I. Nouretdinov, and V. Vovk. "Plug-in martingales for testing exchangeability on-line". In: *arXiv preprint arXiv:1204.3251* (2012).

[35] V. Vovk. "Testing randomness". In: *arXiv preprint arXiv:1906.09256* (2019).

[36] V. Vovk. "Testing for concept shift online". In: *arXiv preprint arXiv:2012.14246* (2020).

[37] V. Vovk, I. Petej, I. Nouretdinov, E. Ahlberg, L. Carlsson, and A. Gammerman. "Retrain or not retrain: Conformal test martingales for change-point detection". In: *arXiv preprint arXiv:2102.10439* (2021).

[38] V. Vovk. “Conditional Validity of Inductive Conformal Predictors”. In: *Proceedings of the Asian Conference on Machine Learning.* Vol. 25. 2012, pp. 475–490.

[39] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. “The limits of distribution-free conditional predictive inference”. In: *preprint at arXiv:1903.04684* (2019).

[40] Y. Hechtlinger, B. Póczos, and L. Wasserman. *Cautious Deep Learning.* arXiv:1805.09460. 2018.

[41] Y. Romano, M. Sesia, and E. J. Candès. “Classification with Valid and Adaptive Coverage”. In: *Advances in Neural Information Processing Systems* 33 (2020).

[42] M. Cauchois, S. Gupta, and J. Duchi. “Knowing what you know: valid confidence sets in multiclass and multilabel prediction”. In: *preprint at arXiv:2004.10181* (2020).

[43] A. N. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan. “Uncertainty Sets for Image Classifiers using Conformal Prediction”. In: *preprint at arXiv:2009.14193* (2020).

[44] Y. Romano, E. Patterson, and E. Candès. “Conformalized Quantile Regression”. In: *Advances in Neural Information Processing Systems 32.* 2019, pp. 3543–3553.

[45] R. Izbicki, G. Shimizu, and R. Stern. “Flexible distribution-free conditional predictive bands using density estimators”. In: *International Conference on Artificial Intelligence and Statistics.* PMLR. 2020, pp. 3068–3077.

[46] V. Chernozhukov, K. Wüthrich, and Y. Zhu. “Distributional conformal prediction”. In: *preprint at arXiv:1909.07889* (2019).

[47] D. Kivaranovic, K. D. Johnson, and H. Leeb. “Adaptive, Distribution-Free Prediction Intervals for Deep Networks”. In: *International Conference on Artificial Intelligence and Statistics.* PMLR. 2020, pp. 4346–4356.

[48] C. Gupta, A. K. Kuchibhotla, and A. K. Ramdas. “Nested conformal prediction and quantile out-of-bag ensemble methods”. In: *arXiv preprint arXiv:1910.10562* (2019).

[49] V. Vovk. “Cross-conformal predictors”. In: *Annals of Mathematics and Artificial Intelligence* 74.1-2 (2015), pp. 9–28.

[50] R. F. Barber, E. J. Candès, A. Ramdas, R. J. Tibshirani, et al. “Predictive inference with the jackknife+”. In: *Annals of Statistics* 49.1 (2021), pp. 486–507.

[51] B. Kim, C. Xu, and R. Foygel Barber. “Predictive inference is free with the jackknife+-after-bootstrap”. In: *Advances in Neural Information Processing Systems* 33 (2020).

[52] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. “Inductive Confidence Machines for Regression”. In: *Machine Learning: European Conference on Machine Learning ECML 2002.* 2002, pp. 345–356.

[53] J. Lei, A. Rinaldo, and L. Wasserman. “A Conformal Prediction Approach to Explore Functional Data”. In: *Annals of Mathematics and Artificial Intelligence* 74 (Feb. 2013).

[54] F. Wilcoxon. “Individual comparisons by ranking methods”. In: *Breakthroughs in statistics.* Springer, 1992, pp. 196–202.

[55] J. Friedman. *On multivariate goodness-of-fit and two-sample testing.* Tech. rep. No. SLAC-PUB-10325. Stanford Linear Accelerator Center, Menlo Park, CA (US), 2004.

[56] D. Lopez-Paz and M. Oquab. “Revisiting classifier two-sample tests”. In: *International Conference on Learning Representations.* 2017.

[57] A. K. Kuchibhotla. “Exchangeability, Conformal Prediction, and Rank Tests”. In: *arXiv preprint arXiv:2005.06095* (2020).

[58] X. Hu and J. Lei. “A Distribution-Free Test of Covariate Shift Using Conformal Prediction”. In: *arXiv preprint arXiv:2010.07147* (2020).

[59] I. Kim, A. Ramdas, A. Singh, L. Wasserman, et al. “Classification accuracy as a proxy for two-sample testing”. In: *Annals of Statistics* 49.1 (2021), pp. 411–434.

[60] S. S. Wilks. “Determination of Sample Sizes for Setting Tolerance Limits”. In: *Ann. Math. Statist.* 12.1 (Mar. 1941), pp. 91–96.

[61] S. S. Wilks. “Statistical Prediction with Special Reference to the Problem of Tolerance Limits”. In: *Ann. Math. Statist.* 13.4 (Dec. 1942), pp. 400–409.

[62] A. Wald. “An Extension of Wilks’ Method for Setting Tolerance Limits”. In: *Ann. Math. Statist.* 14.1 (Mar. 1943), pp. 45–55.

[63] J. W. Tukey. “Non-Parametric Estimation II. Statistically Equivalent Blocks and Tolerance Regions–The Continuous Case”. In: *Ann. Math. Statist.* 18.4 (Dec. 1947), pp. 529–539.

[64] K. Krishnamoorthy and T. Mathew. *Statistical Tolerance Regions: Theory, Applications, and Computation.* Wiley Series in Probability and Statistics. Wiley, 2009.

[65] S. Park, O. Bastani, N. Matni, and I. Lee. “PAC Confidence Sets for Deep Neural Networks via Calibrated Prediction”. In: *International Conference on Learning Representations.* 2020.

[66] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan. “Distribution-Free, Risk-Controlling Prediction Sets”. In: *arXiv preprint* (2021). arXiv:2101.02703.

[67] Y. Zhang and D. N. Politis. “Bootstrap prediction intervals with asymptotic conditional validity and unconditional guarantees”. In: *arXiv preprint arXiv:2005.09145* (2020).

[68] M. B. Brown. “400: A method for combining non-independent, one-sided tests of significance”. In: *Biometrics* (1975), pp. 987–992.

[69] J. T. Kost and M. P. McDermott. “Combining dependent P-values”. In: *Statistics & Probability Letters* 60.2 (2002), pp. 183–190.

[70] J. D. Storey. “A direct approach to false discovery rates”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498.

[71] Y. Benjamini, A. M. Krieger, and D. Yekutieli. “Adaptive linear step-up procedures that control the false discovery rate”. In: *Biometrika* 93.3 (2006), pp. 491–507.

[72] M. Sesia and E. J. Candès. “A comparison of some conformal quantile regression methods”. In: *Stat* 9.1 (2020).

[73] S. K. Sarkar et al. “Generalizing Simes’ test and Hochberg’s stepup procedure”. In: *Annals of Statistics* 36.1 (2008), pp. 337–363.

[74] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. “Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator”. In: *Ann. Math. Stat.* (1956), pp. 642–669.

[75] P. Massart. “The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality”. In: *Annals of Probability* (1990), pp. 1269–1283.

[76] A. Dempster. “Generalized D_n^+ Statistics”. In: *Ann. Math. Stat.* 30.2 (1959), pp. 593–597.

[77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[78] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study". In: *Data mining and knowledge discovery* 30.4 (2016), pp. 891–927.

[79] *Amsterdam Library of Object Images (ALOI) Data Set.* `https://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI/literature/ALOI`. Not normalized, without duplicates. Accessed: January, 2021.

[80] *Covertype Data Set.* `http://odds.cs.stonybrook.edu/forestcovercovertype-dataset`. Accessed: January, 2021.

[81] *Credit Card Fraud Detection Data Set.* `https://www.kaggle.com/mlg-ulb/creditcardfraud`. Accessed: January, 2021.

[82] *KDD Cup 1999 Data Set.* `https://www.kaggle.com/mlg-ulb/creditcardfraud`. Not normalized, without duplicates, categorial attributes removed. Accessed: January, 2021.

[83] *Mammography Data Set.* `http://odds.cs.stonybrook.edu/mammography-dataset/`. Accessed: January, 2021.

[84] *Pen-Based Recognition of Handwritten Digits Data Set.* `http://odds.cs.stonybrook.edu/pendigits-dataset`. Accessed: January, 2021.

[85] *Statlog (Shuttle) Data Set.* `http://odds.cs.stonybrook.edu/shuttle-dataset`. Accessed: January, 2021.

[86] F. T. Liu, K. M. Ting, and Z.-H. Zhou. "Isolation forest". In: *2008 eighth ieee international conference on data mining*. IEEE. 2008, pp. 413–422.

[87] A. Li and R. F. Barber. "Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.1 (2019), pp. 45–74.

[88] V. Vovk, I. Nouretdinov, V. Manokhin, and A. Gammerman. "Cross-conformal predictive distributions". In: *Conformal and Probabilistic Prediction and Applications*. PMLR. 2018, pp. 37–51.

[89] W. Wisniewski, D. Lindsay, and S. Lindsay. "Application of conformal prediction interval estimations to market makers' net positions". In: *Conformal and Probabilistic Prediction and Applications*. PMLR. 2020, pp. 285–301.

[90] M. J. Holland. "Making learning more transparent using conformalized performance prediction". In: *arXiv preprint arXiv:2007.04486* (2020).

[91] P. Bayle, A. Bayle, L. Mackey, and L. Janson. "Cross-validation confidence intervals for test error". In: *Advances in Neural Information Processing Systems* 33 (2020).

[92] T. Lipták. "On the combination of independent tests". In: *Magyar Tud Akad Mat Kutato Int Kozl* 3 (1958), pp. 171–197.

[93] W. Van Zwet and J. Oosterhoff. "On the combination of independent test statistics". In: *Ann. Math. Stat.* 38.3 (1967), pp. 659–680.

[94] V. Vovk and R. Wang. "Combining p-values via averaging". In: *Biometrika* 107.4 (2020), pp. 791–808.

[95] V. Petrov. “Sums of Independent Random Variables”. In: *Yu. V. Prokhorov. V. StatuleviCius (Eds.)* (1975).

[96] P. Moran. “The random division of an interval”. In: *Supplement to the Journal of the Royal Statistical Society* 9.1 (1947), pp. 92–98.

[97] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A first course in order statistics.* SIAM, 2008.

[98] J. Durbin. *Distribution Theory for Tests Based on Sample Distribution Function.* Vol. 9. SIAM, 1973.

[99] V. Kotel'Nikova and E. Chmaladze. “On computing the probability of an empirical process not crossing a curvilinear boundary”. In: *Theory of probability & its applications* 27.3 (1983), pp. 640–648.

[100] D. Siegmund. “Boundary crossing probabilities and statistical applications”. In: *Annals of Statistics* (1986), pp. 361–404.

A Technical proofs

A.1 Correlation structure of null marginal conformal p-values

For notational convenience, we write p_i instead of $\hat{u}^{(\mathrm{marg})}(X_{2n+i})$. When $X_{2n+1}, \ldots, X_{2n+m}$ are all inliers which are drawn from P_X, the conformal p-values $p_1, \ldots, p_m$ are exchangeable. Lemma 1 suggests that the variance of the combination statistic with any transformation $G(\cdot)$ is $(1+\gamma)$ times as large as that when the p-values are i.i.d.. In fact, when $G(\cdot)$ is square-integrable, under the global null,

$$\begin{aligned}\mathrm{Var}\left[\sum_{i=1}^{m} G(p_i)\right] &= m\mathrm{Var}\left[G(p_1)\right] + m(m-1)\mathrm{Cov}\left[G(p_1), G(p_2)\right] \\ &= \left(m + \frac{m(m-1)}{n+2}\right)\mathrm{Var}\left[G(p_1)\right] \\ &\approx (1+\gamma)m\mathrm{Var}\left[G(p_1)\right].\end{aligned}$$

Proof of Lemma 1. Without loss of generality, assume $i = 1$ and $j = 2$. Let $(R_1, \ldots, R_n, R_{n+1}, R_{n+2})$ be the rank of $(S_1, \ldots, S_{n+2}) \overset{d}{=} (\hat{s}(X_{n+1}), \ldots, \hat{s}(X_{2n}), \hat{s}(X_{2n+1}), \hat{s}(X_{2n+2}))$ in the ascending order. Then $S_1, \ldots, S_{n+2}$ are i.i.d. draws from a non-atomic distribution, $(R_1, \ldots, R_{n+2})$ are mutually distinct almost surely and for any permutation $\pi : \{1, \ldots, n+2\} \mapsto \{1, \ldots, n+2\}$,

$$(S_{\pi(1)}, \ldots, S_{\pi(n+2)}) \overset{d}{=} (S_1, \ldots, S_{n+2}).$$

Therefore,

$$(R_1, \ldots, R_{n+2}) \sim \mathrm{Unif}(\{1, \ldots, n+2\}).$$

By definition,

$$p_1 = \frac{1}{n+1}\sum_{i=1}^{n+1} I(S_i \leq S_{n+1}), \quad R_{n+1} = \sum_{i=1}^{n+2} I(S_i \leq S_{n+1}).$$

Thus,

$$p_1 = \frac{R_{n+1} - I(S_{n+2} \leq S_{n+1})}{n+1}.$$

Similarly,

$$p_2 = \frac{R_{n+2} - I(S_{n+1} \leq S_{n+2})}{n+1}.$$

For any $j \in \{1, \dots, n+1\}$,

$$\mathbb{P}\left[p_1 = p_2 = \frac{j}{n+1}\right] = \mathbb{P}\left[R_{n+1} = j+1, R_{n+2} = j\right] + \mathbb{P}\left[R_{n+1} = j, R_{n+2} = j+1\right]$$
$$= 2\mathbb{P}\left[R_{n+1} = j+1, R_{n+2} = j\right] = \frac{2}{(n+2)(n+1)}.$$

For any $1 \leq j < k \leq n+1$,

$$\mathbb{P}\left[p_1 = \frac{j}{n+1}, p_2 = \frac{k}{n+1}\right] = \mathbb{P}\left[R_{n+1} = j, R_{n+2} = k+1\right] = \frac{1}{(n+2)(n+1)}.$$

By symmetry,

$$\mathbb{P}\left[p_1 = \frac{k}{n+1}, p_2 = \frac{j}{n+1}\right] = \frac{1}{(n+2)(n+1)}.$$

As a result,

$$\mathbb{E}[G(p_1)G(p_2)] = \frac{2}{(n+2)(n+1)} \sum_{j=1}^{n+1} G^2\left(\frac{j}{n+1}\right) + \frac{1}{(n+2)(n+1)} \sum_{j \neq k} G\left(\frac{j}{n+1}\right) G\left(\frac{k}{n+1}\right)$$
$$= \frac{1}{(n+2)(n+1)} \sum_{j=1}^{n+1} G^2\left(\frac{j}{n+1}\right) + \frac{1}{(n+2)(n+1)} \left\{\sum_{j=1}^{n+1} G\left(\frac{j}{n+1}\right)\right\}^2.$$

On the other hand, since p_1 is uniformly distributed on $\{1/(n+1), 2/(n+1), \dots, 1\}$,

$$\mathbb{E}[G(p_1)] = \frac{1}{n+1} \sum_{j=1}^{n+1} G\left(\frac{j}{n+1}\right), \quad \mathbb{E}[G^2(p_1)] = \frac{1}{n+1} \sum_{j=1}^{n+1} G^2\left(\frac{j}{n+1}\right).$$

Note that $\mathbb{E}[G^2(p_1)] < \infty$ since $G(i/(n+1)) \in \mathbb{R}$. As a result,

$$\mathrm{Cov}\left[G(p_1), G(p_2)\right] = \frac{1}{(n+2)(n+1)} \sum_{j=1}^{n+1} G^2\left(\frac{j}{n+1}\right) - \frac{1}{(n+2)(n+1)^2} \left\{\sum_{j=1}^{n+1} G\left(\frac{j}{n+1}\right)\right\}^2$$
$$= \frac{1}{n+2} \left\{\mathbb{E}[G^2(p_1)] - (\mathbb{E}[G(p_1)])^2\right\} = \frac{1}{n+2} \mathrm{Var}[G(p_1)].$$

Therefore,

$$\mathrm{Cor}\left[G(p_1), G(p_2)\right] = \frac{\mathrm{Cov}\left[G(p_1), G(p_2)\right]}{\sqrt{\mathrm{Var}[G(p_1)]\mathrm{Var}[G(p_2)]}} = \frac{1}{n+2}.$$

□

A.2 Failure of type-I error control with combination tests

We state a theorem for general (adjusted) combination tests which reject the global null if

$$\sum_{i=1}^{m} G(\hat{u}^{(\mathrm{marg})}(Z_{2n+i})) \geq \xi c_{1-\alpha}(G),$$

where $\xi > 0$ is a pre-specified constant and

$$c_{1-\alpha}(G) \triangleq \mathrm{Quantile}_{1-\alpha}\left(\sum_{i=1}^{m} G(U_i)\right), \quad U_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0,1]).$$

Theorem 5. *Assume $\hat{s}(X)$ is continuous and $G(\cdot) : [0,1] \mapsto \mathbb{R}$ is a non-constant function satisfying*

(i) $\int_0^1 G^{2+\eta}(u)du < \infty$;

(ii) $\left|\frac{1}{n+1}\sum_{j=1}^{n+1} G^k\left(j/(n+1)\right) - \int_0^1 G^k(u)du\right| = o(1/\sqrt{n})$, *for* $k \in \{1,2\}$;

(iii) $\max_{j\in\{1,\ldots,n+1\}} G(j/(n+1)) = o(\sqrt{n})$.

Then, under the global null, if $m = \lfloor \gamma n \rfloor$ *for some* $\gamma \in (0,\infty)$*, as* $n \to \infty$*,*

$$\mathbb{P}\left[\sum_{i=1}^{m} G(\hat{u}^{(\mathrm{marg})}(X_{2n+i})) \geq \xi c_{1-\alpha}(G) - m(\xi - 1)\int_0^1 G(u)du\right] \to \bar{\Phi}\left(\frac{\xi z_{1-\alpha}}{\sqrt{1+\gamma}}\right), \tag{15}$$

where $z_{1-\alpha}$ *and* $\bar{\Phi}$ *denote the* $(1-\alpha)$*-th quantile and the tail function of the standard normal distribution, respectively. Furthermore, under the same asymptotic regime, for* $W \sim N(0,1)$*,*

$$\mathbb{P}\left[\sum_{i=1}^{m} G(\hat{u}^{(\mathrm{marg})}(X_{2n+i})) \geq \xi c_{1-\alpha}(G) - m(\xi - 1)\int_0^1 G(u)du \mid \mathcal{D}\right] \xrightarrow{d} \bar{\Phi}(\xi z_{1-\alpha} + \sqrt{\gamma}W). \tag{16}$$

Remark 1. *For Fisher's combination test,* $G(u) = -2\log u$*. Since* $G(U) \sim \chi^2(2)$*, condition (i) is clearly satisfied. To verify (ii), we note that* $G(u)$ *is decreasing and* $|G'(u)| = 2/u$ *is decreasing. Thus, for* $u \in [(j-1)/(n+1), j/(n+1)]$*, for* $k \in \{1,2\}$*,*

$$0 \leq G^k(u) - G^k\left(\frac{j}{n+1}\right) \leq \frac{k}{n+1}G^{k-1}\left(\frac{j}{n+1}\right)G'\left(\frac{j}{n+1}\right) \leq \frac{8\log(n+1)}{j}.$$

As a result,

$$\left|\frac{1}{n+1}\sum_{j=1}^{n+1} G^k\left(j/(n+1)\right) - \int_0^1 G^k(u)du\right| \leq \sum_{j=1}^{n+1}\left|\frac{1}{n+1}G^k\left(\frac{j}{n+1}\right) - \int_{(j-1)/(n+1)}^{j/(n+1)} G^k(u)du\right|$$

$$\leq \sum_{j=1}^{n+1}\int_{(j-1)/(n+1)}^{j/(n+1)} \left|G^k\left(j/(n+1)\right) - G^k(u)\right|du \leq \frac{1}{n+1}\sum_{j=1}^{n+1}\frac{8\log(n+1)}{j} = O\left(\frac{\log^2 n}{n}\right).$$

Thus, (ii) is proved. Finally, (iii) is satisfied because $G(j/(n+1)) \leq G(1/(n+1)) = O(\log n)$*. Therefore, Theorem 1 is a special case of Theorem 5 with* $\xi = 1$*. In general, it is easy to verify (i)–(iii) for various other combination functions [92–94].*

Remark 2. *By* (15)*, the limiting marginal type-I error is* α *when* $\xi = \sqrt{1+\gamma}$*. This implies* (6) *by noting that* $\int_0^1(-2\log u)du = 2$*. By* (16)*, since the random variable* $\bar{\Phi}(\xi z_{1-\alpha} + \sqrt{\gamma}W)$ *has a positive density everywhere, the* $(1-\delta)$*-th quantile of the conditional type-I error converges to the* $(1-\delta)$*-th quantile of* $\bar{\Phi}(\xi z_{1-\alpha} + \sqrt{\gamma}W)$*, which is* $\bar{\Phi}(\xi z_{1-\alpha} - \sqrt{\gamma}z_{1-\delta})$*. Thus, the conditional type-I error is controlled at level* α *with probability at least* $1-\delta$ *asymptotically if* $\xi = 1 + \sqrt{\gamma}z_{1-\delta}/z_{1-\alpha}$*.*

Remark 3. *To confirm our theory, we run Monte-Carlo simulations with* $n = 10^5$ *and* $\gamma \in \{2^{-3}, 2^{-2}, \ldots, 2^3\}$*, estimating the average type-I error across* 10^4 *samples. Since* $\hat{s}(X)$ *is continuous, we can assume that* $\hat{s}(X) \sim \mathrm{Unif}([0,1])$ *without loss of generality, as we will show in the proof. Figure A1 presents the simulated and asymptotic type-I errors for both the unadjusted (*$\xi = 1$*) and adjusted (*$\xi = \sqrt{1+\gamma}$*) Fisher's combination test given by* (6)*.*

Remark 4. *If the* p_i*'s are dependent, [68] and [69] approximate the null distribution by a rescaled chi-square distribution* $c\chi^2(f)$*, where* c *and* f *are chosen to match the mean and variance of the Fisher's combination statistic* S_{Fisher}*. Specifically,*

$$c = \frac{\mathrm{Var}[S_{\mathrm{Fisher}}]}{2\mathbb{E}[S_{\mathrm{Fisher}}]}, \quad f = \frac{2\mathbb{E}[S_{\mathrm{Fisher}}]^2}{\mathrm{Var}[S_{\mathrm{Fisher}}]}.$$

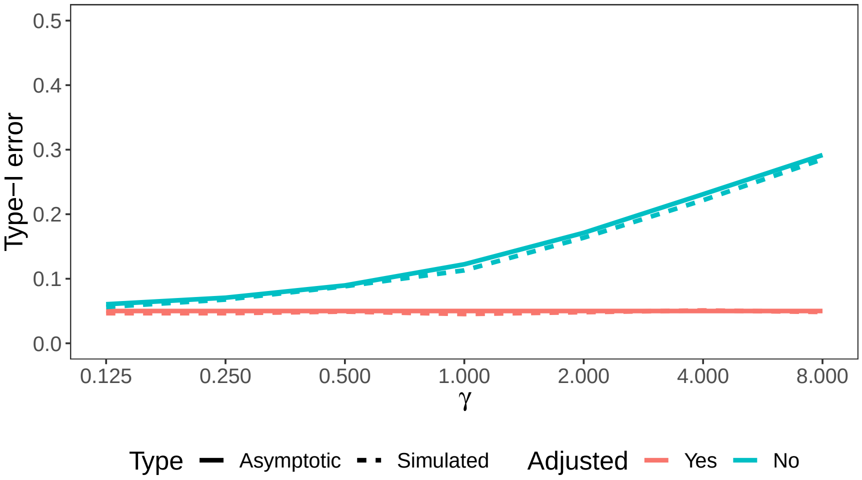

Figure A1: Type-I errors of unadjusted and adjusted Fisher's combination test.

In our case, it is easy to see that

$$\mathbb{E}[S_{\text{Fisher}}] \approx 2m, \quad \text{Var}[S_{\text{Fisher}}] \approx 4m(1+\gamma).$$

As a result, the null distribution is approximated by $(1+\gamma)\chi^2(2m/(1+\gamma))$*. The central limit theorem implies that* $\chi^2(f) \approx N(f, 2f)$ *when* f *is large. Thus, the critical value for this approximation is*

$$(1+\gamma)\chi^2(2m/(1+\gamma); 1-\alpha) \approx (1+\gamma)\left(\frac{2m}{1+\gamma} + \sqrt{\frac{2m}{1+\gamma}}z_{1-\alpha}\right) = 2m + \sqrt{2m(1+\gamma)}z_{1-\alpha}.$$

Similarly, the critical value for our correction (6) *is*

$$\sqrt{1+\gamma}\chi^2(2m; 1-\alpha) - 2(\sqrt{1+\gamma}-1)m \approx \sqrt{1+\gamma}(2m+\sqrt{2m}z_{1-\alpha}) - 2(\sqrt{1+\gamma}-1)m \approx 2m + \sqrt{2m(1+\gamma)}z_{1-\alpha}.$$

Therefore, both corrections are asymptotically equivalent.

To prove Theorem 5, we start by stating two lemmas. The first lemma is a general Berry-Esseen bound for sums of independent (but not necessarily identically distributed) random variables with potentially infinite third moments.

Lemma 2. *[[95], p. 112, Theorem 5] Let* $X_1, X_2, \ldots, X_n$ *be independent random variables such that* $\mathbb{E}[X_j] = 0$*, for all* j*. Assume also* $\mathbb{E}[X_j^2 g(X_j)] < \infty$ *for some function* g *that is non-negative, even, and non-decreasing in the interval* $x > 0$*, with* $x/g(x)$ *being non-decreasing for* $x > 0$*. Write* $B_n = \sum_j \text{Var}[X_j]$*. Then,*

$$d_K\left(\mathcal{L}\left(\frac{1}{\sqrt{B_n}}\sum_{j=1}^n X_j\right), N(0,1)\right) \le \frac{A}{B_n g(\sqrt{B_n})}\sum_{j=1}^n \mathbb{E}\left[X_j^2 g(X_j)\right],$$

where A *is a universal constant,* $\mathcal{L}(\cdot)$ *denotes the probability law,* d_K *denotes the Kolmogorov-Smirnov distance (i.e., the* ℓ_∞*-norm of the difference of CDFs)*

The second lemma is a well-known representation of the spacing between consecutive order statistics.

Lemma 3 (From [96]; see also Section 4 of [97])**.** *Let* $U_1, \ldots, U_n \overset{i.i.d.}{\sim} \text{Unif}([0,1])$ *and* $U_{(1)} \le U_{(2)} \le \ldots \le U_{(n)}$ *be their order statistics. Then*

$$(U_{(1)} - U_{(0)}, \ldots, U_{(n+1)} - U_{(n)}) \overset{d}{=} \left(\frac{V_1}{\sum_{k=1}^{n+1} V_k}, \ldots, \frac{V_{n+1}}{\sum_{k=1}^{n+1} V_k}\right),$$

where $U_{(0)} = 0, U_{(n+1)} = 1$*, and* $V_1, \ldots, V_{n+1} \overset{\text{i.i.d.}}{\sim} \text{Exp}(1)$*.*

Proof of Theorem 5. We first prove the limiting conditional type-I error (16). For convenience, we write p_i instead of $\hat{u}^{(\mathrm{marg})}(X_{2n+i})$ and S_j instead of $\hat{s}(X_{n+j})$. Since $\hat{s}(X)$ is continuous,

$$p_i = \frac{1 + |\{j \in \mathcal{D}^{\mathrm{cal}} : S_j \le \hat{s}(X_{2n+i})\}|}{n+1} = \frac{1 + |\{j \in \mathcal{D}^{\mathrm{cal}} : F_S(S_j) \le F_S(\hat{s}(X_{2n+i}))\}|}{n+1}$$

where F_S denotes the CDF of $\hat{s}(X)$ conditional on $\mathcal{D}$. As a result, we can assume $\hat{s}(X) \sim \mathrm{Unif}([0,1])$ without loss of generality. Conditional on $\mathcal{D}$, $p_1, \ldots, p_m$ are i.i.d. random variables with

$$\mathbb{P}\left[p_i = \frac{j}{n+1} \mid \mathcal{D}\right] = S_{(j)} - S_{(j-1)}, \qquad j = 1, \ldots, n+1,$$

where $S_{(1)} < S_{(2)} < \ldots < S_{(n)}$ denote the order statistics of $(S_1, \ldots, S_n)$, and $S_{(0)} = 0, S_{(n+1)} = 1$. By Lemma 3, we can reformulate the distribution of p_i conditional on $\mathcal{D}$ as

$$\mathbb{P}\left[p_i = \frac{j}{n+1} \mid \mathcal{D}\right] = \frac{V_j}{\sum_{k=1}^{n+1} V_k}, \qquad j = 1, \ldots, n+1. \tag{17}$$

As a result, for $k \in \{1, 2\}$,

$$\mathbb{E}\left[G^k(p_i) \mid \mathcal{D}\right] = \frac{\sum_{j=1}^{n+1} G^k\left(\frac{j}{n+1}\right) V_j}{\sum_{j=1}^{n+1} V_j} = \frac{(n+1)^{-1} \sum_{j=1}^{n+1} G^k\left(\frac{j}{n+1}\right) V_j}{(n+1)^{-1} \sum_{j=1}^{n+1} V_j}. \tag{18}$$

By the strong law of large number,

$$\frac{1}{n+1} \sum_{j=1}^{n+1} V_j \overset{\text{a.s.}}{\to} \mathbb{E}[V_1] = 1. \tag{19}$$

Let $g_n = \max_{j \in \{1, \ldots, n+1\}} G(j/(n+1))$. Since $V_1, \ldots, V_{n+1}$ are independent,

$$\begin{aligned} \mathrm{Var}\left[\frac{1}{n+1} \sum_{j=1}^{n+1} G^k\left(\frac{j}{n+1}\right) V_j\right] &= \sum_{j=1}^{n+1} \frac{1}{(n+1)^2} \mathbb{E}\left[G^{2k}\left(\frac{j}{n+1}\right)(V_j - 1)^2\right] \\ = \frac{1}{(n+1)^2} \sum_{j=1}^{n+1} G^{2k}\left(\frac{j}{n+1}\right) &\le \frac{g_n^{2k-2}}{(n+1)} \frac{1}{n+1} \sum_{j=1}^{n+1} G^2\left(\frac{j}{n+1}\right). \end{aligned} \tag{20}$$

By condition (ii),

$$\left|\frac{1}{n+1} \sum_{j=1}^{n+1} G^2\left(\frac{j}{n+1}\right) - \int_0^1 G^2(u)du\right| = o(1),$$

and thus

$$\frac{1}{n+1} \sum_{j=1}^{n+1} G^2\left(\frac{j}{n+1}\right) = O(1).$$

By condition (iii), $g_n = o(\sqrt{n})$. Together with (20), we obtain that for $k \in \{1, 2\}$,

$$\mathrm{Var}\left[\frac{1}{n+1} \sum_{j=1}^{n+1} G^k\left(\frac{j}{n+1}\right) V_j\right] = o(1).$$

By Chebyshev's inequality,

$$\frac{1}{n+1} \sum_{i=1}^{n} G^k\left(\frac{j}{n+1}\right)(V_j - 1) = o_P(1).$$

Applying the condition (ii) again, we arrive at

$$\frac{1}{n+1} \sum_{i=1}^{n} G^k\left(\frac{j}{n+1}\right) V_j - \int_0^1 G^k(u)du = o_P(1).$$

By (18),

$$\mathbb{E}\left[G^k(p_i) \mid \mathcal{D}\right] - \int_0^1 G^k(u)du = o_P(1), \qquad k \in \{1,2\}. \tag{21}$$

Let a_n be a deterministic sequence such that $a_n < 1/2$, and $U \sim \text{Unif}([0,1])$. Let also $\mathcal{E}_n$ be the event that $\mathcal{D}$ is such that

$$\mathcal{E}_n = \left\{\mathcal{D} : \frac{\text{Var}[G(p_i) \mid \mathcal{D}]}{\text{Var}[G(U)]} \in [1-a_n, 1+a_n]\right\}. \tag{22}$$

Since G is a non-constant function, $\text{Var}[G(U)] > 0$. By (21), we can choose $a_n = o(1)$ such that

$$\mathbb{P}\left[\mathcal{E}_n^c\right] = o(1).$$

Let

$$W_m = \frac{\sum_{i=1}^m \{G(p_i) - \mathbb{E}[G(p_i) \mid \mathcal{D}]\}}{\sqrt{m\text{Var}[G(p_i) \mid \mathcal{D}]}}.$$

By Lemma 2 with $g(x) = x$,

$$d_K\left(\mathcal{L}\left(W_m \mid \mathcal{D}\right), N(0,1)\right) \le \frac{A}{\sqrt{m}} \frac{\mathbb{E}\left[|G(p_i) - \mathbb{E}[G(p_i) \mid \mathcal{D}]|^3\right]}{\text{Var}[G(p_i) \mid \mathcal{D}]^{3/2}},$$

where A is a universal constant. Since $G(p_i) \le g_n$ almost surely, by condition (iii),

$$\mathbb{E}\left[|G(p_i) - \mathbb{E}[G(p_i) \mid \mathcal{D}]|^3\right] \le 2g_n\text{Var}[G(p_i) \mid \mathcal{D}].$$

Thus,

$$d_K\left(\mathcal{L}\left(W_m \mid \mathcal{D}\right), N(0,1)\right) \le \frac{2A}{\sqrt{m}} \frac{g_n}{\text{Var}\left[G(p_i) \mid \mathcal{D}\right]^{1/2}}.$$

On the event $\mathcal{E}_n$, the condition (iii) and that $n = O(m)$ imply that

$$d_K\left(\mathcal{L}\left(W_m \mid \mathcal{D}\right), N(0,1)\right) \le \frac{4Ag_n}{\sqrt{m\text{Var}[G(U)]}} = o(1).$$

Since the Kolmogorov distance is invariant under rescalings, we have

$$d_K\left(\mathcal{L}\left(\sqrt{\frac{\text{Var}[G(p_i) \mid \mathcal{D}]}{\text{Var}[G(U)]}}W_m \mid \mathcal{D}\right), N\left(0, \frac{\text{Var}[G(p_i) \mid \mathcal{D}]}{\text{Var}[G(U)]}\right)\right) = o(1).$$

Since $\text{Var}[G(p_i) \mid \mathcal{D}]/\text{Var}[G(U)] \in [1-a_n, 1+a_n] \to 1$,

$$d_K\left(N\left(0, \frac{\text{Var}[G(p_i) \mid \mathcal{D}]}{\text{Var}[G(U)]}\right), N(0,1)\right) = o(1).$$

Let

$$K_n \triangleq d_K\left(\mathcal{L}\left(\sqrt{\frac{\text{Var}[G(p_i) \mid \mathcal{D}]}{\text{Var}[G(U)]}}W_m \mid \mathcal{D}\right), N(0,1)\right). \tag{23}$$

The above arguments show that $K_n = o(1)$ on the event $\mathcal{E}_n$.

On the other hand, let

$$c_m = \frac{c_{1-\alpha}(G) - m\mathbb{E}[G(U)]}{\sqrt{m\text{Var}[G(U)]}},$$

and

$$\tilde{W}_n = \frac{\sqrt{n+1}(\mathbb{E}[G(p_i) \mid \mathcal{D}] - \mathbb{E}[G(U)])}{\sqrt{\text{Var}[G(U)]}}.$$

Then

$$\mathbb{P}\left[\sum_{i=1}^{m} G(p_i) \geq \xi c_{1-\alpha}(G) - m(\xi-1)\mathbb{E}[G(U)] \mid \mathcal{D}\right] = \mathbb{P}\left[\sqrt{\frac{\mathrm{Var}[G(p_i) \mid \mathcal{D}]}{\mathrm{Var}[G(U)]}} W_m + \sqrt{\frac{m}{n+1}}\tilde{W}_n \geq \xi c_m \mid \mathcal{D}\right].$$

By (23),

$$\left|\mathbb{P}\left[\sqrt{\frac{\mathrm{Var}[G(p_i) \mid \mathcal{D}]}{\mathrm{Var}[G(U)]}} W_m + \sqrt{\frac{m}{n+1}}\tilde{W}_n \geq \xi c_m \mid \mathcal{D}\right] - \bar{\Phi}\left(\xi c_m - \sqrt{\frac{m}{n+1}}\tilde{W}_n\right)\right| \leq K_n.$$

Since $K_n = o(1)$ on $\mathcal{E}_n$ and $\mathbb{P}[\mathcal{E}_n^c] = o(1)$, we obtain that

$$\left|\mathbb{P}\left[\sum_{i=1}^{m} G(p_i) \geq c_{1-\alpha}(G) \mid \mathcal{D}\right] - \bar{\Phi}\left(\xi c_m - \sqrt{\frac{m}{n+1}}\tilde{W}_n\right)\right| = o_P(1). \tag{24}$$

Since $\bar{\Phi}$ is a continuous function and $m/n \to \gamma$, to prove (16), it remains to prove

$$c_m \xrightarrow{p} z_{1-\alpha}, \quad \tilde{W}_n \xrightarrow{d} N(0,1). \tag{25}$$

Without loss of generality, we assume that $\eta \leq 1$ in the condition (i). By Lemma 2 with $g(x) = x^\eta$, which clearly fulfills the criteria, we have that

$$d_K\left(\frac{\sum_{j=1}^{m} G(U_i) - \mathbb{E}[G(U)]}{\sqrt{m\mathrm{Var}[G(U)]}}, N(0,1)\right) \leq \frac{A}{m^{\eta/2}} \frac{\mathbb{E}[|G(U) - \mathbb{E}[G(U)]|^{2+\eta}]}{\mathrm{Var}[G(U)]^{1+\eta/2}} = o(1). \tag{26}$$

By definition, c_m is the $(1-\alpha)$-th quantile of $\left(\sum_{j=1}^{m} G(U_i) - \mathbb{E}[G(U)]\right)/\sqrt{m\mathrm{Var}[G(U)]}$. By (26),

$$|\bar{\Phi}(c_m) - \alpha| = |\bar{\Phi}(c_m) - \bar{\Phi}(z_{1-\alpha})| = o(1).$$

Since $\bar{\Phi}'(z_{1-\alpha}) > 0$, it implies the first part of (25).

To prove the second part of (25), we recall (18) with $k = 1$ that

$$\tilde{W}_n = \frac{(n+1)^{-1/2}\sum_{j=1}^{n+1}\left\{G\left(\frac{j}{n+1}\right) - \mathbb{E}[G(U)]\right\} V_j}{\sqrt{\mathrm{Var}[G(U)]}\left(\sum_{j=1}^{n+1} V_j\right)/(n+1)}.$$

Set $X_j = (n+1)^{-1/2}\left\{G\left(\frac{j}{n+1}\right) - \mathbb{E}[G(U)]\right\}(V_j - 1)$ and $g(x) = x$ in Lemma 2. Then

$$B_n = \frac{1}{n+1}\sum_{j=1}^{n+1}\left\{G\left(\frac{j}{n+1}\right) - \mathbb{E}[G(U)]\right\}^2.$$

By the condition (ii), we have that

$$B_n = \frac{1}{n+1}\sum_{j=1}^{n} G^2\left(\frac{j}{n+1}\right) - \frac{2\mathbb{E}[G(U)]}{n+1}\sum_{j=1}^{n} G\left(\frac{j}{n+1}\right) + (\mathbb{E}[G(U)])^2 \to \mathrm{Var}[G(U)]. \tag{27}$$

By the condition (i), (iii) and (27),

$$\begin{aligned}\sum_{j=1}^{n+1}\mathbb{E}|X_j|^3 &\leq \frac{1}{(n+1)^{3/2}}\sum_{j=1}^{n+1}\left|G\left(\frac{j}{n+1}\right) - \mathbb{E}[G(U)]\right|^3 \\ &\leq \frac{g_n + \mathbb{E}[G(U)]}{\sqrt{n+1}}\frac{1}{n+1}\sum_{j=1}^{n+1}\left(G\left(\frac{j}{n+1}\right) - \mathbb{E}[G(U)]\right)^2 \\ &= \frac{g_n + \mathbb{E}[G(U)]}{\sqrt{n+1}} B_n = o(1).\end{aligned}$$

Let

$$\tilde{W}_n' = \frac{1}{\sqrt{B_n}} \sum_{j=1}^{n+1} X_j = \frac{1}{\sqrt{(n+1)B_n}} \sum_{j=1}^{n+1} \left\{ G\left(\frac{j}{n+1}\right) - \mathbb{E}[G(U)] \right\} (V_j - 1).$$

Then Lemma 2 implies that

$$d_K\left(\tilde{W}_n', N(0,1)\right) \le \frac{A\sum_{j=1}^{n+1} \mathbb{E}|X_j|^3}{B_n^{3/2}} = o(1). \tag{28}$$

By definition,

$$\tilde{W}_n = \left(\tilde{W}_n' + \frac{1}{\sqrt{(n+1)B_n}} \sum_{j=1}^{n+1} \left\{ G\left(\frac{j}{n+1}\right) - \mathbb{E}[G(U)] \right\} \right) \sqrt{\frac{B_n}{\mathrm{Var}[G(U)]}} \frac{1}{\left(\sum_{j=1}^{n} V_j\right)/(n+1)}.$$

The condition (ii) with $k = 1$ implies that

$$\frac{1}{\sqrt{(n+1)}} \sum_{j=1}^{n+1} \left\{ G\left(\frac{j}{n+1}\right) - \mathbb{E}[G(U)] \right\} = o(1). \tag{29}$$

By (19), (27), (28), (29) and Slutsky's Lemma, we prove the second part of (25). Therefore, the limiting conditional type-I error (16) is proved.

Next, we prove the limiting marginal type-I error (15). Since

$$\mathbb{P}\left[\sum_{i=1}^{m} G(p_i) \ge \xi c_{1-\alpha}(G) - m(\xi - 1)\mathbb{E}[G(U)] \mid \mathcal{D} \right]$$

is bounded almost surely, the convergence in distribution implies the convergence in expectation. Therefore,

$$\mathbb{P}\left[\sum_{i=1}^{m} G(p_i) \ge \xi c_{1-\alpha}(G) - m(\xi - 1)\mathbb{E}[G(U)] \right] \to \mathbb{E}[\bar{\Phi}(\xi z_{1-\alpha} + \sqrt{\gamma} W)].$$

Let W' be an independent copy of W. Then

$$\bar{\Phi}(\xi z_{1-\alpha} + \sqrt{\gamma} W) = \mathbb{P}\left[W' \ge \xi z_{1-\alpha} + \sqrt{\gamma} W \mid W \right].$$

As a result,

$$\mathbb{E}[\bar{\Phi}(\xi z_{1-\alpha} + \sqrt{\gamma} W)] = \mathbb{P}\left[W' \ge \xi z_{1-\alpha} - \sqrt{\gamma} W \right] = \mathbb{P}\left[W' - \sqrt{\gamma} W \ge \xi z_{1-\alpha} \right].$$

The proof is completed by the fact that $W' - \sqrt{\gamma} W \sim N(0, 1 + \gamma)$.

□

A.3 Conformal p-values are PRDS

Proof of Theorem 2. Let $Z = (S_{(1)}, \ldots, S_{(n)})$ be the order statistics of $(\hat{s}(X_i))_{i \in \{n+1, \ldots, 2n\}}$, the conformal scores evaluated on the calibration set. Let $Y = (p_1, \ldots, p_m)$ be the conformal p-values evaluated on the test set (i.e., $p_j = \hat{u}^{(\mathrm{marg})}(X_{2n+j})$). Then,

$$\begin{aligned}\mathbb{P}\left[Y \in D \mid Y_i = y \right] &= \int \mathbb{P}\left[Y \in D \mid Z = z \right] \mathbb{P}\left[Z = z \mid Y_i = y \right] dz \\ &= \mathbb{E}_{Z \mid Y_i = y}\left[\mathbb{P}\left[Y \in D \mid Z \right] \right].\end{aligned}$$

With this representation, the conclusion will be implied by the following two lemmas.

Lemma 4. *For a non-decreasing set D and vectors z, z' such that $z \preceq z'$, then*

$$\mathbb{P}\left[Y \in D \mid Z = z\right] \geq \mathbb{P}\left[Y \in D \mid Z = z'\right].$$

Lemma 5. *For $y \geq y'$, there exists $Z_1 \sim Z \mid Y_i = y$ and $Z_2 \sim Z \mid Y_i = y'$ such that $\mathbb{P}\left[Z_1 \preceq Z_2\right] = 1$.*

In words, Lemma 4 states that the conformal p-values increase as the conformal scores on the calibration set decrease, while Lemma 5 states that a larger conformal p-value indicates the calibration conformal scores are smaller. The proof follows easily from these. Take any $y \geq y'$ and let Z_1 and Z_2 be as in the statement of Lemma 5. Then,

$$\begin{aligned}\mathbb{P}\left[Y \in D \mid Y_i = y\right] &= \mathbb{E}_{Z_1}\left[\mathbb{P}\left[Y \in D \mid Z = Z_1\right]\right] \\ &\geq \mathbb{E}_{Z_2}\left[\mathbb{P}\left[Y \in D \mid Z = Z_2\right]\right] \\ &= \mathbb{P}\left[Y \in D \mid Y_i = y'\right].\end{aligned}$$

The inequality follows from Lemma 4 and the fact that $\mathbb{P}\left[Z_1 \preceq Z_2\right] = 1$, which comes from Lemma 5. □

Lemma 4 follows immediately from the definition of marginal conformal p-values in (3). Lemma 5 is proved below.

Proof of Lemma 5, continuous case. As in the proof of Theorem 5, since $\hat{s}(X)$ is continuous, we can assume without loss of generality that the scores S_i follow the uniform distribution on $[0, 1]$. Let $S'_{(1)} \leq S'_{(2)} \leq \ldots \leq S'_{(n+1)}$ be the order statistics of $(\hat{s}(X_{n+1}), \ldots, \hat{s}(X_{2n+1}))$ and R_{2n+1} be the rank of $\hat{s}(X_{2n+1})$ among these. By definition,

$$\left\{(S_{(1)}, \ldots, S_{(n)}) \mid R_{2n+1} = k, S'_{(1)}, \ldots, S'_{(n+1)}\right\} = (S'_{(1)}, \ldots, S'_{(k-1)}, S'_{(k+1)}, \ldots, S'_{(n+1)}).$$

Since $\hat{s}(X)$ is continuous, R_{2n+1} is independent of $(S'_{(1)}, S'_{(2)}, \ldots, S'_{(n+1)})$. As a result, for any positive integer $k \leq n+1$,

$$\left\{(S_{(1)}, \ldots, S_{(n)}) \mid R_{2n+1} = k\right\} \overset{d}{=} (S'_{(1)}, \ldots, S'_{(k-1)}, S'_{(k+1)}, \ldots, S'_{(n+1)}).$$

The right-hand-side is clearly entry-wise non-increasing in k. Since $p_1 = R_{2n+1}/(n+1)$, Lemma 5 is proved for $i = 1$. The same proof carries over to other indices i.

□

Extension to non-continuous scores. When $\hat{s}(X)$ has atoms, the set of conformity scores $\{\hat{s}(X_i) : i \in \mathcal{D}^{\text{cal}}\}$ have ties with non-zero probability. In this case, we replace the marginal conformal p-value (2) by a randomized version, i.e.,

$$p_j = \frac{|\{i \in \mathcal{D}^{\text{cal}} : \hat{s}(X_i) < \hat{s}(X_{2n+j})\}| + \lceil(1 + |\{i \in \mathcal{D}^{\text{cal}} : \hat{s}(X_i) = \hat{s}(X_{2n+j})\}|)U_j\rceil}{n+1}, \tag{30}$$

where $U_1, U_2, \ldots$ are i.i.d. random variables drawn from Unif$([0, 1])$ which are independent of the data. Note that (30) is identical to (2) almost surely when $\hat{s}(X)$ is continuous. Now we prove that the marginal conformal p-values defined in (30) satisfy the PRDS property.

Proposition 4 (Theorem 2 for the non-continuous case)**.** *Consider the setting of Theorem 2, but where $\hat{s}(\cdot)$ is not assumed to be continuous. Define the randomized marginal p-values as in* (30)*. Then, the marginal conformal p-values $(p_1, \ldots, p_m)$ are PRDS.*

The proof follows as above, once we verify Lemma 4 and Lemma 5 in the more general setting.

Proof of Lemma 4, general case. Let $U = (U_1, \ldots, U_m)$. By definition, U is independent of (Y, Z), and thus

$$\mathbb{P}[Y \in D \mid Z = z] = \mathbb{P}[Y \in D \mid Z = z, U], \quad \text{a.s..}$$

Let $p_j(x; z, u)$ denote the mapping from (X_{2n+j}, Z, U) to p_j. Then

$$p_j(x; z, u) = \frac{m_<(x; z) + \lceil \{1 + m_=(x; z)\}u \rceil}{n + 1},$$

where

$$m_<(x, z) = \sum_{i=1}^{n} I(z_i < x), \quad m_=(x, z) = \sum_{i=1}^{n} I(z_i = x).$$

If $z \preceq z'$,

$$m_<(x, z) \geq m_<(x, z'), \quad m_<(x, z) + m_=(x, z) \geq m_<(x, z') + m_=(x, z'). \tag{31}$$

We claim that the mapping $p_j(x; z, u)$ is non-increasing in z for every x and u. Equivalently, we will show that for any x and $u \in [0, 1]$,

$$m_<(x, z) + \lceil \{1 + m_=(x, z)\}u \rceil \geq m_<(x, z') + \lceil \{1 + m_=(x, z')\}u \rceil. \tag{32}$$

We consider three cases.

Case 1: if $m_<(x, z) = m_<(x, z')$, (31) implies that $m_=(x, z) \geq m_=(x, z')$. Thus, (32) is obviously true.

Case 2: if $m_<(x, z) + m_=(x, z) = m_<(x, z') + m_=(x, z')$, let $a = 1 + m_=(x, z)$ and $b = m_<(x, z) - m_<(x, z')$. Then (32) is equivalent to

$$b \geq \lceil (a + b)u \rceil - \lceil au \rceil.$$

This can be proved using the fact that $\lceil (a + b)u \rceil \leq \lceil au \rceil + \lceil bu \rceil$.

Case 3: if $m_<(x, z) > m_<(x, z')$ and $m_<(x, z) + m_=(x, z) > m_<(x, z') + m_=(x, z')$, then $m_<(x, z) \geq m_<(x, z') + 1$ and $m_<(x, z) + m_=(x, z) \geq m_<(x, z') + m_=(x, z') + 1$ since $m_<(x, z), m_<(x, z'), m_=(x, z)$, and $m_=(x, z')$ are all integers. Then

$$\begin{aligned} m_<(x, z) + \lceil \{1 + m_=(x, z)\}u \rceil &\geq m_<(x, z) + \{1 + m_=(x, z)\}u \\ &= m_<(x, z)(1 - u) + \{1 + m_=(x, z) + m_<(x, z)\}u \\ &\geq \{1 + m_<(x, z')\}(1 - u) + \{2 + m_=(x, z') + m_<(x, z')\}u \\ &= m_<(x, z') + \{1 + m_=(x, z')\}u + 1 \\ &\geq m_<(x, z') + \lceil \{1 + m_=(x, z')\}u \rceil. \end{aligned}$$

Therefore, (32) is proved. As a result, the mapping from $(X_{2n+1}, \ldots, X_{2n+m}, Z, U)$ to Y is entry-wise non-increasing in Z given $(X_{2n+j}, \ldots, X_{2n+m}, U)$. Since $\{X_{2n+j} : j = 1, \ldots, m\}$, Z, and U are mutually independent, we arrive at

$$\mathbb{P}[Y \in D \mid Z = z, U] \geq \mathbb{P}[Y \in D \mid Z = z', U], \quad \text{a.s..}$$

The independence between U and Z implies that $(U \mid Z = z) \overset{d}{=} (U \mid Z = z')$. Lemma 4 then follows from the above inequality. □

Proof of Lemma 5, general case. Let $R_{2n+j} = (n + 1)p_j$. Note that R_{2n+j} can be interpreted as the rank with ties broken randomly. As in the proof for the continuous case, we first prove that

$$\left\{(S_{(1)}, \ldots, S_{(n)}) \mid R_{2n+1} = k, S'_{(1)}, \ldots, S'_{(n+1)}\right\} = (S'_{(1)}, \ldots, S'_{(k-1)}, S'_{(k+1)}, \ldots, S'_{(n+1)}). \tag{33}$$

Let $k_- = \max\{\ell : S'_{(\ell)} < S_{2n+1}\}$ and $k_+ = \min\{\ell : S'_{(\ell)} > S_{2n+1}\}$. Then $S'_\ell = S_{2n+1}$ for any $k_- < \ell < k_+$. Since there exists at least one ℓ with $S'_{(\ell)} = S_{2n+1}$, i.e., the index corresponding to S_{2n+1}, we have $k_+ - k_- \geq 2$. By definition,

$$1 + |\{i \in \mathcal{D}^{\text{cal}} : \hat{s}(X_i) = \hat{s}(X_{2n+j})\}| = |\{i \in \mathcal{D}^{\text{cal}} \cup \{2n+1\} : \hat{s}(X_i) = \hat{s}(X_{2n+j})\}| = k_+ - k_- - 1.$$

As a result,

$$k = k_- + \lceil (k_+ - k_- - 1)U_1 \rceil \in (k_-, k_+).$$

Therefore, $\hat{s}(X_{2n+1}) = S'_{(k)}$ and (33) is proved.

It remains to prove that R_{2n+1} is independent of $(S'_{(1)}, S'_{(2)}, \ldots, S'_{(n+1)})$. For any non-decreasing sequence $a_1 \leq \ldots \leq a_{n+1}$, let $1 = n_0 < n_1 < \ldots < n_m = n+1$ be integers such that

$$a_{n_{j-1}} = \ldots = a_{n_j - 1} < a_{n_j}, \quad j = 1, \ldots, m-1, \quad a_{n_{m-1}-1} < a_{n_{m-1}} = \ldots = a_{n_m}$$

Let $\pi : \{1, \ldots, n+1\} \mapsto \{1, \ldots, n+1\}$ be a uniform random permutation. Since $X_{n+1}, \ldots, X_{2n+1}$ are i.i.d., Conditioning on the event that,

$$\left\{ (\hat{s}(X_{n+1}), \ldots, \hat{s}(X_{2n+1})) \mid (S'_{(1)}, \ldots, S'_{(n+1)}) = (a_1, \ldots, a_{n+1}) \right\} \overset{d}{=} \left(a_{\pi(1)}, \ldots, a_{\pi(n+1)}\right).$$

For any $j = 1, \ldots, m-1$, if $\pi(n+1) \in [n_{j-1}, n_j)$,

$$|\{i : a_{\pi(i)} = a_{\pi(n+1)}\}| = n_j - n_{j-1}, \quad |\{i : a_{\pi(i)} < a_{\pi(n+1)}\}| = n_{j-1} - 1,$$

and thus,

$$R_{2n+1} = n_{j-1} - 1 + \lceil (n_j - n_{j-1})U_j \rceil.$$

Similarly, if $\pi(n+1) \in [n_{m-1}, n_m]$,

$$R_{2n+1} = n_{m-1} - 1 + \lceil (n_m - n_{m-1} + 1)U_j \rceil.$$

For any k, let $j_k = \max\{j : n_j \leq k\}$, and $\mathcal{I}_k$ be the set $\{n_{j_k - 1}, \ldots, n_{j_k} - 1\}$ if $j_k < m$ and $\{n_{j_k - 1}, \ldots, n_{j_k}\}$ otherwise. Then

$$\begin{aligned}
&\mathbb{P}(R_{2n+1} = k \mid (S'_{(1)}, \ldots, S'_{(n+1)}) = (a_1, \ldots, a_{n+1})) \\
&= \mathbb{P}\left(\pi(n+1) \in \mathcal{I}_k, U_1 \in \left(\frac{k - n_{j_k - 1}}{|\mathcal{I}_k|}, \frac{k + 1 - n_{j_k - 1}}{|\mathcal{I}_k|}\right]\right) \\
&= \mathbb{P}\left(\pi(n+1) \in \mathcal{I}_k\right) \mathbb{P}\left(U_1 \in \left(\frac{k - n_{j_k - 1}}{|\mathcal{I}_k|}, \frac{k + 1 - n_{j_k - 1}}{|\mathcal{I}_k|}\right]\right) \\
&= \frac{|\mathcal{I}_k|}{n+1} \frac{1}{|\mathcal{I}_k|} = \frac{1}{n+1}.
\end{aligned}$$

Therefore, R_{2n+1} is independent of $(S'_{(1)}, \ldots, S'_{(n+1)})$. The proof of Lemma 5 is then completed. □

A.4 Storey's correction does not break FDR control

Given a p-value p_i for the i-th null hypothesis, let $p_{(1)} \leq \ldots \leq p_{(m)}$ be the ordered statistics. Given a target FDR level α and a scalar $\lambda \in (0, 1)$, the rejection set of the Storey-BH procedure is

$$\mathcal{R} = \left\{ i : p_i \leq \frac{\alpha R}{m \hat{\pi}_0}, p_i < \lambda \right\},$$

where

$$\hat{\pi}_0 = \frac{1 + \sum_{i=1}^m I(p_i \geq \lambda)}{m(1-\lambda)} \triangleq \frac{1 + A}{m(1-\lambda)}$$

and

$$R = \max\left\{r : p_{(r)} \le \frac{\alpha r}{m\hat{\pi}_0}, p_{(r)} < \lambda\right\}.$$

The parameter λ is often chosen as 0.5 or α or $1-\alpha$.

We start with a novel FDR bound for this procedure applied to PRDS p-values.

Theorem 6. *Assume that $(p_1, \ldots, p_n)$ is PRDS and each null p-value is super-uniform with an almost sure lower bound $p_{\min} \in [0,1]$. Then*

$$\mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{\max\{1, |\mathcal{R}|\}}\right] \le \alpha(1-\lambda)\sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\frac{1}{1+A} \mid p_i \le p_*\right],$$

where

$$p_* = \max\left\{\frac{\alpha(1-\lambda)}{m}, p_{\min}\right\}.$$

Proof. Let

$$V_i = I(H_i \text{ is rejected}) \le I\left(p_i \le \alpha(1-\lambda)\frac{R}{1+A}\right).$$

Then

$$\begin{aligned}\mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{\max\{1, |\mathcal{R}|\}}\right] &= \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\frac{V_i}{R \vee 1}\right] = \sum_{i \in \mathcal{H}_0}\sum_{r=1}^{m} \frac{1}{r}\mathbb{P}\left(p_i \le \alpha(1-\lambda)\frac{r}{1+A}, R = r\right) \\ &= \sum_{i \in \mathcal{H}_0}\sum_{r=1}^{m}\sum_{a=1}^{m} \frac{1}{r}\mathbb{P}\left(p_i \le \alpha(1-\lambda)\frac{r}{1+a}, R = r, A = a\right).\end{aligned}$$

Let $r_0(a) = \max\{1, \lceil (1+a)p_{\min}/(1-\lambda)\alpha \rceil\}$. By definition, the summand for a given a is non-zero only if $r \ge r_0(a)$. Thus,

$$\begin{aligned}\mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{\max\{1, |\mathcal{R}|\}}\right] &= \sum_{i \in \mathcal{H}_0}\sum_{a=1}^{m}\sum_{r=r_0(a)}^{m} \frac{1}{r}\mathbb{P}\left(p_i \le \alpha(1-\lambda)\frac{r}{1+a}\right)\mathbb{P}\left(R = r, A = a \mid p_i \le \alpha(1-\lambda)\frac{r}{1+a}\right) \\ &\overset{(i)}{\le} \sum_{i \in \mathcal{H}_0}\sum_{a=1}^{m}\sum_{r=r_0(a)}^{m} \frac{1}{r}\cdot\alpha(1-\lambda)\frac{r}{1+a}\mathbb{P}\left(R = r, A = a \mid p_i \le \alpha(1-\lambda)\frac{r}{1+a}\right) \\ &= \alpha(1-\lambda)\sum_{i \in \mathcal{H}_0}\sum_{a=1}^{m}\sum_{r=r_0(a)}^{m} \frac{1}{1+a}\mathbb{P}\left(R = r, A = a \mid p_i \le \alpha(1-\lambda)\frac{r}{1+a}\right),\end{aligned}$$

where (i) uses the super-uniformity of the null p-value. Let $\mathcal{T}$ denote the set of all possible values that $r/(1+a)$ can take such that $\mathbb{P}(p_i \le \alpha(1-\lambda)r/(1+a)) > 0$, i.e.

$$\mathcal{T} = \left\{\frac{r}{1+a} : a \in \{1, \ldots, m\}, r \in \{r_0(a), \ldots, m\}, a + r \le m\right\}.$$

Clearly, $\mathcal{T}$ is a finite set. Let $t_1 \le t_2 \le \ldots \le t_M$ denote the values of $\mathcal{T}$. It is easy to see that

$$\alpha(1-\lambda)t_1 \ge \max\left\{p_{\min}, \frac{\alpha(1-\lambda)}{m}\right\} = p_*. \tag{34}$$

Then

$$
\begin{aligned}
\mathbb{E}\left[\frac{|\mathcal{R}\cap\mathcal{H}_0|}{\max\{1,|\mathcal{R}|\}}\right] &\le \alpha(1-\lambda)\sum_{i\in\mathcal{H}_0}\sum_{j=1}^{M}\sum_{a=1}^{m}\frac{1}{1+a}\mathbb{P}\left(R=(1+a)t_j, A=a \mid p_i\le\alpha(1-\lambda)t_j\right)\\
&= \alpha(1-\lambda)\sum_{i\in\mathcal{H}_0}\sum_{j=1}^{M}\mathbb{E}\left[\frac{I\{R=(1+A)t_j\}}{1+A}\mid p_i\le\alpha(1-\lambda)t_j\right]\\
&= \alpha(1-\lambda)\sum_{i\in\mathcal{H}_0}\sum_{j=1}^{M}\left\{\mathbb{E}\left[H_j(p)\mid p_i\le\alpha(1-\lambda)t_j\right]-\mathbb{E}\left[H_{j+1}(p)\mid p_i\le\alpha(1-\lambda)t_j\right]\right\}\\
&= \alpha(1-\lambda)\sum_{i\in\mathcal{H}_0}\left\{\mathbb{E}\left[H_1(p)\mid p_i\le\alpha(1-\lambda)t_1\right]\right.\\
&\quad\left.-\sum_{j=1}^{M-1}\left(\mathbb{E}\left[H_{j+1}(p)\mid p_i\le\alpha(1-\lambda)t_j\right]-\mathbb{E}\left[H_{j+1}(p)\mid p_i\le\alpha(1-\lambda)t_{j+1}\right]\right)\right\},
\end{aligned}
$$

where

$$
H_j(p)=\frac{I\{R\ge(1+A)t_j\}}{1+A},\quad H_{M+1}(p)=0.
$$

Since A is an increasing function of p and R is a decreasing function of p, $H_j(p)$ is decreasing in p. The PRDS property implies that for any $j=1,\dots,M-1$,

$$
\mathbb{E}\left[H_{j+1}(p)\mid p_i\le\alpha(1-\lambda)t_j\right]-\mathbb{E}\left[H_{j+1}(p)\mid p_i\le\alpha(1-\lambda)t_{j+1}\right]\ge 0.
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\left[\frac{|\mathcal{R}\cap\mathcal{H}_0|}{\max\{1,|\mathcal{R}|\}}\right] &\le \alpha(1-\lambda)\sum_{i\in\mathcal{H}_0}\mathbb{E}\left[H_1(p)\mid p_i\le\alpha(1-\lambda)t_1\right]\\
&\le \alpha(1-\lambda)\sum_{i\in\mathcal{H}_0}\mathbb{E}\left[\frac{1}{1+A}\mid p_i\le\alpha(1-\lambda)t_1\right]\\
&\le \alpha(1-\lambda)\sum_{i\in\mathcal{H}_0}\mathbb{E}\left[\frac{1}{1+A}\mid p_i\le p_*\right],
\end{aligned}
$$

where the last step follows from (34), the PRDS property, and the fact that $p\mapsto 1/(1+A)$ is decreasing. □

To prove Theorem 3, we present an additional lemma.

Lemma 6. *[Lemma 1 from [71]] If* $Y\sim\mathrm{Binom}(k-1,p)$*, then* $\mathbb{E}[1/(1+Y)]\le 1/kp$*.*

Proof of Theorem 3. As in the proof of Theorem 5, since $\hat{s}(X)$ is continuous, we can assume $\hat{s}(X)\sim\mathrm{Unif}([0,1])$ without loss of generality. We write p_i instead of $\hat{u}^{(\mathrm{marg})}(X_{2n+i})$ and S_j instead of $\hat{s}(X_{n+j})$. Then

$$
p_j=\frac{1+\sum_{i=1}^{n}I(S_i\le S_{n+j})}{n+1}.
$$

Then $p_j\ge 1/(n+1)$ almost surely. Let $m_0=|\mathcal{H}_0|$ and we assume that $\mathcal{H}_0=\{1,\dots,m_0\}$ without loss of generality. Since $p=(p_1,\dots,p_m)$ are PRDS and exchangeable, Theorem 4 implies that

$$
\mathbb{E}\left[\frac{|\mathcal{R}\cap\mathcal{H}_0|}{\max\{1,|\mathcal{R}|\}}\right]\le\alpha(1-\lambda)m_0\mathbb{E}\left[\frac{1}{1+A}\mid p_1\le\max\left\{\frac{1}{n+1},\frac{\alpha(1-\lambda)}{m}\right\}\right].
$$

Since $1/(1+A)$ is decreasing in p, using the PRDS property again, we have

$$
\mathbb{E}\left[\frac{|\mathcal{R}\cap\mathcal{H}_0|}{\max\{1,|\mathcal{R}|\}}\right]\le\alpha(1-\lambda)m_0\mathbb{E}\left[\frac{1}{1+A}\mid p_1\le\frac{1}{n+1}\right]=\alpha(1-\lambda)m_0\mathbb{E}\left[\frac{1}{1+A}\mid p_1=\frac{1}{n+1}\right]. \tag{35}
$$

Let $A_0 = \sum_{j=2}^{m_0} I(p_j \geq \lambda)$. Then

$$\mathbb{E}\left[\frac{1}{1+A} \mid p_1 = \frac{1}{n+1}\right] \leq \mathbb{E}\left[\frac{1}{1+A_0} \mid p_1 = \frac{1}{n+1}\right].$$

Let $S_{(1)} \leq S_{(2)} \leq \ldots \leq S_{(n+1)}$ denote the order statistics of $S_1, \ldots, S_{n+1}$ and R_{n+1} denote the rank of S_{n+1}. Since $S_1 \sim \mathrm{Unif}([0,1])$, there is no tie almost surely.

Now we compute

$$\mathbb{E}\left[\frac{1}{1+A_0} \mid p_1 = \frac{1}{n+1}, S_{(1)}, \ldots, S_{(n+1)}\right] = \mathbb{E}\left[\frac{1}{1+A_0} \mid R_{n+1} = 1, S_{(1)}, \ldots, S_{(n+1)}\right]. \tag{36}$$

By definition,

$$p_2, \ldots, p_{m_0} \mid S_1, \ldots, S_{n+1} \overset{i.i.d.}{\sim} \frac{1 + \sum_{j=1}^{n} I(S_j \leq U)}{n+1}$$

where $U \sim \mathrm{Unif}([0,1])$. Note that there is a bijection between $(S_1, \ldots, S_{n+1})$ and $(S_{(1)}, \ldots, S_{(n+1)}, R_1, \ldots, R_{n+1})$ for vectors without ties. The above distributional equivalence can be rewritten as

$$p_2, \ldots, p_{m_0} \mid R_1, \ldots, R_{n+1}, S_{(1)}, \ldots, S_{(n+1)} \overset{i.i.d.}{\sim} \frac{1 + \sum_{j=1}^{n+1} I(S_{(j)} \leq U) - I(S_{(R_{n+1})} \leq U)}{n+1}.$$

Since the RHS does not depend on $(R_1, \ldots, R_n)$, $(p_2, \ldots, p_{m_0})$ is independent of $(R_1, \ldots, R_n)$ conditional on $(R_{n+1}, S_{(1)}, \ldots, S_{(n+1)})$. As a result,

$$p_2, \ldots, p_{m_0} \mid R_{n+1} = 1, S_{(1)}, \ldots, S_{(n+1)} \overset{i.i.d.}{\sim} \frac{1 + \sum_{j=2}^{n+1} I(S_{(j)} \leq U)}{n+1}.$$

Recall $K = (n+1)\lambda \in \mathbb{Z}$. Then

$$\begin{aligned}
\mathbb{P}\left(p_2 \geq \lambda \mid R_{n+1} = 1, S_{(1)}, \ldots, S_{(n+1)}\right) &= \mathbb{P}\left(\sum_{j=2}^{n+1} I(S_{(j)} \leq U) \geq K - 1 \mid S_{(2)}, \ldots, S_{(n+1)}\right) \\
&= \mathbb{P}\left(U \geq S_{(K)} \mid S_{(2)}, \ldots, S_{(n+1)}\right) \\
&= 1 - S_{(K)}.
\end{aligned}$$

Therefore,

$$I(p_2 \geq \lambda), \ldots, I(p_{m_0} \geq \lambda) \mid R_{n+1} = 1, S_{(1)}, \ldots, S_{(n+1)} \overset{i.i.d.}{\sim} \mathrm{Ber}\left(1 - S_{(K)}\right).$$

This implies that

$$A_0 \mid R_{n+1} = 1, S_{(1)}, \ldots, S_{(n+1)} \sim \mathrm{Binom}\left(m_0 - 1, 1 - S_{(K)}\right).$$

By Lemma 6,

$$\mathbb{E}\left[\frac{1}{1+A_0} \mid R_{n+1} = 1, S_{(1)}, \ldots, S_{(n+1)}\right] \leq \frac{1}{m_0\{1 - S_{(K)}\}}.$$

Since R_{n+1} is independent of $(S_{(1)}, \ldots, S_{(n+1)})$,

$$\mathbb{E}\left[\frac{1}{1+A_0} \mid R_{n+1} = 1\right] \leq \mathbb{E}\left[\frac{1}{m_0\{1 - S_{(K)}\}}\right]. \tag{37}$$

By symmetry and the property of order statistics,

$$1 - S_{(K)} \overset{d}{=} S_{(n+2-K)} \sim \mathrm{Beta}(n + 2 - K, K).$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1-S_{(K)}}\right] &= \int_0^1 \frac{1}{x}\frac{\Gamma(n+2)}{\Gamma(n+2-K)\Gamma(K)}x^{n+1-K}(1-x)^{K-1}dx \\
&= \int_0^1 \frac{\Gamma(n+2)}{\Gamma(n+2-K)\Gamma(K)}x^{n-K}(1-x)^{K-1}dx \\
&= \frac{\Gamma(n+2)\Gamma(n+1-K)}{\Gamma(n+2-K)\Gamma(n+1)} \\
&= \frac{n+1}{n+1-K} = \frac{1}{1-\lambda}. \qquad (38)
\end{aligned}$$

Putting (35), (37) and (38) together, we prove the result. □

A.5 Conditional p-value adjustment

Proof of Theorem 4. Let $S_i = \hat{s}(X_{n+i})$ for $i = 1, \ldots, n$ with $F^-(t) = \mathbb{P}[S_i < t \mid \mathcal{D}^{\mathrm{train}}]$, and $S_{(1)} \le S_{(2)} \le \ldots S_{(n)}$ be the order statistics. Then it is easy to see that

$$(F^-(S_{(1)}), \ldots, F^-(S_{(n)})) \preceq (U_{(1)}, \ldots, U_{(n)}),$$

where $\preceq$ denotes the entry-wise stochastic dominance in the sense that $(A_1, \ldots, A_n) \preceq (B_1, \ldots, B_n)$ iff

$$\mathbb{P}[A_1 \le z_1, \ldots, A_n \le z_n] \ge \mathbb{P}[B_1 \le z_1, \ldots, B_n \le z_n], \quad \forall (z_1, \ldots, z_n) \in \mathbb{R}^n.$$

When F is continuous, the equality in distribution holds. Let $\mathcal{E}_n$ denote the event on which $F^-(S_{(i)}) \le b_i$ for all $i = 1, \ldots, n$. Then

$$\mathbb{P}[\mathcal{E}_n] \ge 1 - \delta.$$

Now we prove the following claim, which directly yields the theorem:

$$\mathbb{P}\left[\hat{u}^{(\mathrm{ccv})}(X_{2n+1}) \le t \mid \mathcal{D}\right] \le t, \quad \forall t \in [0,1], \quad \text{if } \mathcal{D} \in \mathcal{E}_n. \qquad (39)$$

Note that the image of $\hat{u}^{(\mathrm{ccv})}$ is $\{b_1, \ldots, b_n, 1\}$, it remains to prove (39) with $t \in \{b_1, \ldots, b_n, 1\}$. When $t = 1$, it clearly holds. When $t = b_i$,

$$\hat{u}^{(\mathrm{ccv})}(X_{2n+1}) \le b_i \Longleftrightarrow \hat{u}^{(\mathrm{marg})}(X_{2n+1}) \le \frac{i}{n+1} \Longleftrightarrow \hat{s}(X_{2n+1}) < S_{(i)}.$$

Thus,

$$\mathbb{P}\left[\hat{u}^{(\mathrm{ccv})}(X_{2n+1}) \le b_i \mid \mathcal{D}\right] = \mathbb{P}\left[\hat{s}(X_{2n+1}) < S_{(i)} \mid \mathcal{D}\right] = F^-(S_{(i)}).$$

By definition of $\mathcal{E}_n$, (39) holds for all $t \in \{b_1, \ldots, b_n\}$.

□

A.6 Simultaneous confidence bounds for the false positive rate

Proof of Proposition 3. Note that $h(i/n) = b_{\lceil i+i/n \rceil} = b_{i+1}$ where we let $b_{n+1} = 1$ for convenience. Then, the event that $F(Z_{(i)}) \le h((i-1)/n) = b_i$ for all $i \in \{1, \ldots, n\}$ occurs with probability at least $1 - \delta$, where $Z_{(1)} \le \ldots \le Z_{(n)}$ are the order statistics. Under this event, for any $z \in [Z_{(i-1)}, Z_{(i)})$, where we let $Z_{(0)} = \infty$ and $Z_{(n+1)} = \infty$ for convenience, $\hat{F}_n(z) = (i-1)/n$ and thus

$$F(z) \le F(Z_{(i)}) \le b_i = h\left(\hat{F}_n(z)\right).$$

On the other hand, if $h : [0,1] \to [0,1]$ is a function such that $h(\hat{F}_n(z))$ is a uniform upper confidence band of F for any CDF F, then (10) holds with $b_i = h(i/n)$. □

B Numerical comparisons of different adjustment functions

In addition to the adjustment functions derived from the generalized Simes inequality and the DKWM inequality, we consider here another class of simultaneous bounds based on the so-called *boundary crossing probability* [76, 98–100]—the probability that $F(z)$ ever crosses $h(\hat{F}_n(z))$ for a fixed function $h(\cdot)$. This probability is generally difficult to compute analytically, but the special case of a linear $h(\cdot)$ is an exception. Assuming that F is the CDF of Unif([0, 1]), let $\hat{F}_n(z)$ is the empirical CDF of $S_1, \dots, S_n \overset{\text{i.i.d.}}{\sim} \text{Unif}([0,1])$. Then, [76] proved that

$$\mathbb{P}\left[\hat{F}_n(z) \leq b + \frac{1-b}{1-a}z,\ \forall z \in (0,1)\right] = 1 - \Delta_{\text{Dempster}}(a, b; n),$$

for any $a, b \in (0, 1)$, where

$$\Delta_{\text{Dempster}}(a, b; n) := a \sum_{j=0}^{\lfloor n(1-b) \rfloor} \frac{n!}{j!(n-j)!} \left(a + \frac{1-a}{1-b}\frac{j}{n}\right)^{j-1} \left(1 - a - \frac{1-a}{1-b}\frac{j}{n}\right)^{n-j}. \tag{40}$$

If we replace S_i with $1 - S_i$, then $\hat{F}_n(z)$ becomes $1 - \hat{F}_n(1-z)$. Further, replacing z by $1 - z$ leads to

$$\mathbb{P}\left[z \leq \frac{1-a}{1-b}\hat{F}_n(z) + a,\ \forall z \in (0,1)\right] = 1 - \Delta_{\text{Dempster}}(a, b; n). \tag{41}$$

For any pair (a, b) with $\Delta_{\text{Dempster}}(a, b; n) = \delta$, we obtain a function $h(z) = a + (1-a)z/(1-b)$ satisfying (11), which yields the following sequence satisfying (10):

$$b_i = a + \frac{1-a}{1-b}\frac{i}{n}.$$

Given any a, it is easy to compute the corresponding b such that $\Delta_{\text{Dempster}}(a, b; n) = \delta$ via a binary search.

Note that this leads to adjusted p-values that cannot be lower than $b_1 = a + (1-a)/(1-b)n$. To ensure a fair comparison with the method based on the generalized Simes inequality, we choose a via another binary search such that the resulting b_1 matches that given by the Simes inequality for a particular value of k. If there exists no value of a yielding the same b_1 as the Simes method, we set a as to minimize b_1. Figure A2 compares the adjustment functions yielded by the generalized Simes inequality, the DKWM inequality, and the Dempster exact linear-boundary crossing probability with $k \in \{n/4, n/2\}$ and $n \in \{300, 1000, 3000, 10000\}$ for small marginal p-values within $[0, 0.05]$. It is clear that the Simes adjustment function is the best option in most scenarios, except when $n = 10000$ and $\hat{u}^{(\text{marg})}(X) > 0.03$, in which case the DKWM bound is tighter. Nonetheless, for the purpose of multiple testing, we would rarely expect p-values above 0.03 to be significant.

C Numerical outlier detection experiments

C.1 Outlier detection on simulated data

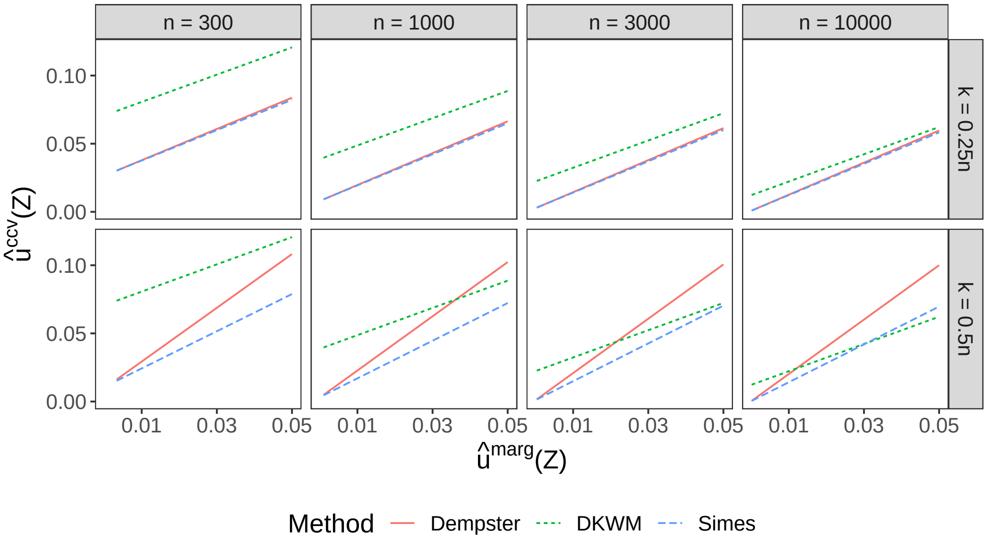

Figure A2: Comparison of different adjustment functions.

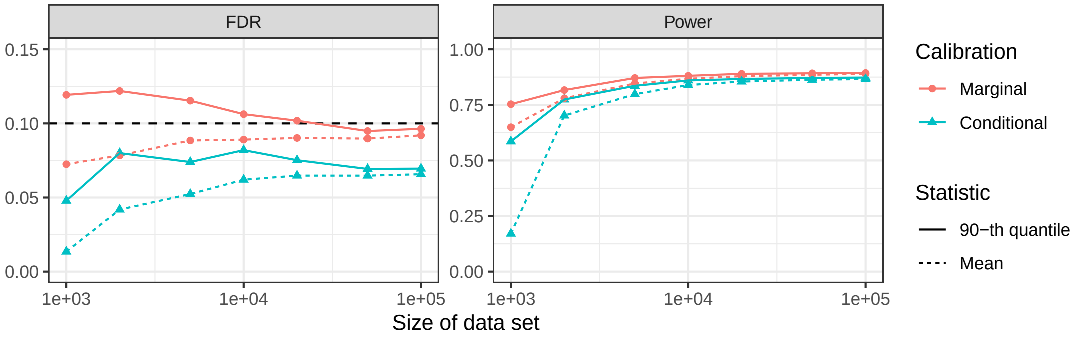

Figure A3: FDR and power in a simulated outlier detection problem as a function of the number of samples in the data set (half of which are utilized for calibration). Other details are as in Figure 5.

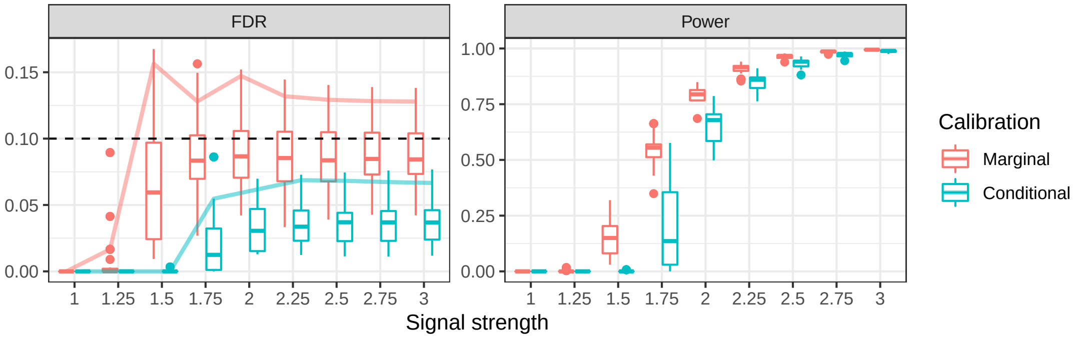

Figure A4: FDR and power in a simulated outlier detection problem, using the Benjamini-Hochberg procedure with Storey's correction. Other details are as in Figure 5.

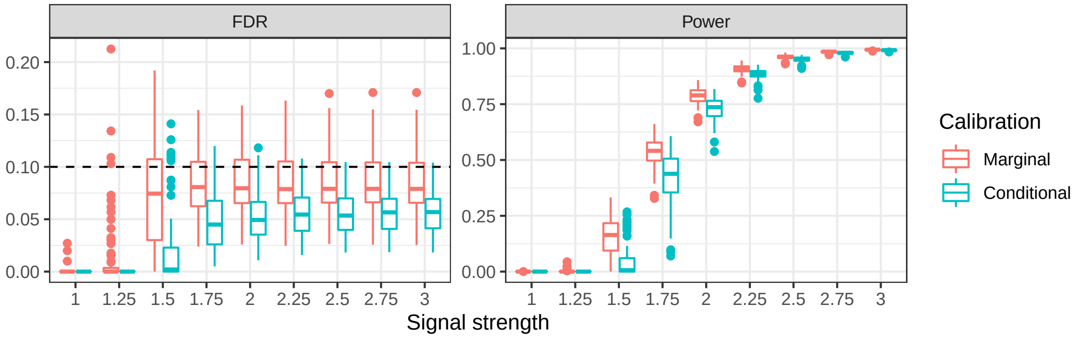

Figure A5: FDR and power in a simulated outlier detection problem, as a function of the signal strength. The conditional calibration method is applied with $\delta = 0.25$ instead of $\delta = 0.05$. Other details are as in Figure 5.

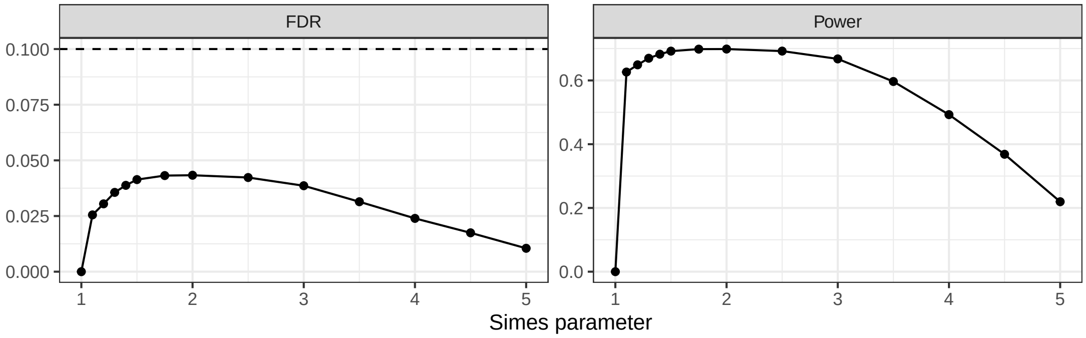

Figure A6: Performance of simultaneously calibrated conformal p-values as a function of the Simes parameter n/k. The signal strength is equal to 2. Other details are as in Figure 5.

C.2 Outlier detection on real data

Table A1: Outlier detection performance on real data, using different data sets, machine learning models, and nominal FDR levels. Other details are as in Table 2.

		FDR				Power			
		Mean		90th percentile		Mean		90-th quantile	
Model	Nominal	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.
ALOI									
IForest	0.05	0	0	0	0	0	0	0	0
	0.10	0.001	0	0	0	0	0	0	0
	0.20	0.025	0	0.02	0	0	0	0	0
LOF	0.05	0	0	0	0	0	0	0	0
	0.10	0.005	0	0	0	0	0	0	0
	0.20	0.056	0.003	0.176	0	0.002	0	0.003	0
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0.005	0	0.012	0	0	0	0.001	0
	0.20	0.069	0.001	0.228	0	0.003	0	0.01	0
Cover									
IForest	0.05	0	0	0	0	0	0	0	0
	0.10	0.011	0	0.032	0	0.002	0	0.003	0
	0.20	0.08	0.013	0.277	0.049	0.008	0.002	0.03	0.004
LOF	0.05	0.05	0.026	0.069	0.041	0.949	0.91	0.968	0.943
	0.10	0.1	0.056	0.126	0.075	0.973	0.955	0.98	0.969
	0.20	0.198	0.111	0.23	0.138	0.987	0.976	0.991	0.982
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
Credit card									
IForest	0.05	0.037	0.012	0.074	0.05	0.185	0.062	0.389	0.256
	0.10	0.095	0.042	0.126	0.076	0.426	0.207	0.603	0.409
	0.20	0.197	0.106	0.233	0.135	0.712	0.469	0.803	0.624
LOF	0.05	0.001	0	0	0	0	0	0	0
	0.10	0.018	0.001	0.022	0	0.001	0	0	0
	0.20	0.087	0.021	0.278	0.031	0.007	0.002	0.022	0.001
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
KDDCup99									
IForest	0.05	0.04	0.013	0.074	0.036	0.378	0.208	0.515	0.44
	0.10	0.095	0.043	0.125	0.077	0.594	0.397	0.703	0.524
	0.20	0.196	0.105	0.234	0.135	0.755	0.62	0.825	0.713
LOF	0.05	0.001	0	0	0	0	0	0	0
	0.10	0.038	0	0.159	0	0.012	0	0.055	0
	0.20	0.141	0.037	0.27	0.158	0.039	0.011	0.07	0.055

(Continued on Next Page...)

Table A1: Outlier detection performance on real data, using different data sets, machine learning models, and nominal FDR levels. Other details are as in Table 2. *(continued)*

		FDR				Power			
		Mean		90th percentile		Mean		90-th quantile	
Model	Nominal	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
Mammography									
IForest	0.05	0.011	0	0.051	0	0.012	0	0.035	0
	0.10	0.067	0	0.163	0	0.061	0	0.184	0
	0.20	0.18	0.031	0.282	0.112	0.167	0.036	0.342	0.155
LOF	0.05	0.003	0	0	0	0.001	0	0	0
	0.10	0.032	0	0.169	0	0.023	0	0.112	0
	0.20	0.167	0.018	0.272	0.061	0.195	0.017	0.316	0.036
SVM	0.05	0.011	0	0.031	0	0.004	0	0.015	0
	0.10	0.075	0	0.196	0	0.042	0	0.086	0
	0.20	0.186	0.003	0.256	0.002	0.169	0.001	0.267	0.001
Digits									
IForest	0.05	0.006	0	0.006	0	0.007	0	0.005	0
	0.10	0.049	0.002	0.159	0	0.073	0.003	0.245	0
	0.20	0.177	0.029	0.27	0.116	0.347	0.056	0.603	0.213
LOF	0.05	0.01	0	0.045	0	0.038	0.005	0.149	0
	0.10	0.06	0.003	0.142	0.001	0.282	0.017	0.775	0.005
	0.20	0.191	0.059	0.245	0.144	0.821	0.297	0.984	0.795
SVM	0.05	0.001	0	0	0	0	0	0	0
	0.10	0.045	0	0.152	0	0.018	0	0.048	0
	0.20	0.167	0.005	0.253	0.007	0.24	0.004	0.475	0.002
Shuttle									
IForest	0.05	0.048	0.023	0.068	0.035	0.946	0.872	0.977	0.97
	0.10	0.097	0.052	0.13	0.071	0.975	0.953	0.981	0.977
	0.20	0.196	0.107	0.234	0.138	0.981	0.976	0.984	0.981
LOF	0.05	0.051	0.026	0.07	0.044	0.991	0.857	0.998	0.988
	0.10	0.099	0.055	0.125	0.075	0.999	0.992	1	0.999
	0.20	0.197	0.109	0.236	0.137	1	0.999	1	1
SVM	0.05	0.047	0.007	0.067	0.034	0.904	0.152	0.998	0.91
	0.10	0.101	0.052	0.12	0.072	0.999	0.953	1	0.998
	0.20	0.202	0.112	0.232	0.13	1	1	1	1

Table A2: Outlier detection performance on real data, using different data sets, machine learning models, and nominal FDR levels. Other details are as in Table A1.

		FDR				Power			
		Mean		90th percentile		Mean		90-th quantile	
Model	Nominal	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.
ALOI									
IForest	0.05	0	0	0	0	0	0	0	0
	0.10	0.001	0	0	0	0	0	0	0
	0.20	0.027	0.003	0.03	0	0	0	0	0
LOF	0.05	0	0	0	0	0	0	0	0
	0.10	0.005	0	0	0	0	0	0	0
	0.20	0.054	0.007	0.175	0	0.002	0	0.003	0
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0.006	0	0.015	0	0	0	0.001	0
	0.20	0.083	0.013	0.263	0.031	0.004	0.001	0.012	0.002
Cover									
IForest	0.05	0	0	0	0	0	0	0	0
	0.10	0.008	0	0.031	0	0.001	0	0.002	0
	0.20	0.073	0.021	0.244	0.085	0.007	0.003	0.025	0.009
LOF	0.05	0.045	0.031	0.063	0.049	0.943	0.922	0.965	0.955
	0.10	0.09	0.065	0.115	0.085	0.971	0.961	0.978	0.971
	0.20	0.179	0.129	0.209	0.156	0.985	0.979	0.989	0.984
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
Credit card									
IForest	0.05	0.032	0.016	0.068	0.057	0.164	0.085	0.344	0.297
	0.10	0.084	0.051	0.114	0.085	0.384	0.248	0.58	0.452
	0.20	0.178	0.126	0.211	0.154	0.678	0.539	0.775	0.667
LOF	0.05	0.001	0	0	0	0	0	0	0
	0.10	0.017	0.003	0.013	0	0.001	0	0	0
	0.20	0.085	0.03	0.274	0.052	0.006	0.002	0.022	0.003
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
KDDCup99									
IForest	0.05	0.034	0.018	0.067	0.047	0.355	0.238	0.501	0.455
	0.10	0.086	0.055	0.116	0.088	0.561	0.453	0.687	0.617
	0.20	0.176	0.123	0.212	0.156	0.738	0.666	0.813	0.735
LOF	0.05	0.001	0	0	0	0	0	0	0
	0.10	0.037	0.004	0.159	0.004	0.012	0.001	0.055	0
	0.20	0.138	0.051	0.264	0.177	0.038	0.015	0.069	0.055

(Continued on Next Page...)

Table A2: Outlier detection performance on real data, using different data sets, machine learning models, and nominal FDR levels. Other details are as in Table A1. *(continued)*

		FDR				Power			
		Mean		90th percentile		Mean		90-th quantile	
Model	Nominal	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
Mammography									
IForest	0.05	0.008	0	0.024	0	0.009	0	0.014	0
	0.10	0.056	0.002	0.155	0.001	0.05	0.003	0.16	0.002
	0.20	0.161	0.053	0.252	0.165	0.147	0.059	0.315	0.217
LOF	0.05	0.001	0	0	0	0	0	0	0
	0.10	0.029	0	0.154	0	0.019	0	0.057	0
	0.20	0.141	0.035	0.259	0.178	0.161	0.036	0.307	0.164
SVM	0.05	0.006	0	0.009	0	0.002	0	0.004	0
	0.10	0.065	0	0.185	0	0.036	0	0.068	0
	0.20	0.17	0.053	0.245	0.173	0.146	0.033	0.244	0.091
Digits									
IForest	0.05	0.005	0	0.002	0	0.006	0	0.001	0
	0.10	0.042	0.003	0.143	0	0.057	0.005	0.181	0
	0.20	0.159	0.05	0.257	0.172	0.296	0.093	0.541	0.368
LOF	0.05	0.009	0	0.029	0	0.029	0.005	0.111	0
	0.10	0.046	0.007	0.129	0.017	0.209	0.034	0.687	0.066
	0.20	0.173	0.086	0.226	0.162	0.77	0.429	0.975	0.861
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0.042	0	0.149	0	0.015	0	0.047	0
	0.20	0.145	0.02	0.242	0.052	0.168	0.018	0.417	0.048
Shuttle									
IForest	0.05	0.043	0.028	0.061	0.043	0.939	0.891	0.976	0.973
	0.10	0.088	0.061	0.117	0.087	0.973	0.963	0.98	0.978
	0.20	0.176	0.124	0.209	0.158	0.981	0.978	0.983	0.982
LOF	0.05	0.045	0.032	0.065	0.049	0.988	0.934	0.998	0.992
	0.10	0.089	0.065	0.111	0.09	0.998	0.995	1	0.999
	0.20	0.178	0.127	0.211	0.156	1	0.999	1	1
SVM	0.05	0.039	0.015	0.061	0.043	0.827	0.358	0.997	0.993
	0.10	0.091	0.063	0.111	0.081	0.999	0.997	1	0.999
	0.20	0.183	0.13	0.216	0.156	1	1	1	1

Table A3: Outlier batch detection performance on real data, using Storey's correction to control the FDR. Other details are as in Table 3.

		FDR				Power			
		Mean		90th percentile		Mean		90-th quantile	
Model	Nominal	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.
ALOI									
IForest	0.05	0.035	0.003	0.1	0	0.001	0	0.002	0
	0.10	0.071	0.003	0.164	0	0.001	0	0.004	0
	0.20	0.141	0.014	0.278	0.058	0.004	0	0.008	0.001
LOF	0.05	0.034	0	0.091	0	0.023	0.003	0.039	0.007
	0.10	0.082	0.003	0.161	0	0.047	0.005	0.071	0.012
	0.20	0.185	0.009	0.284	0.032	0.106	0.011	0.153	0.02
SVM	0.05	0.032	0.004	0.097	0	0.003	0	0.007	0.002
	0.10	0.064	0.006	0.173	0	0.006	0.001	0.01	0.003
	0.20	0.152	0.013	0.289	0.08	0.012	0.002	0.022	0.005
Cover									
IForest	0.05	0.036	0.006	0.088	0.003	0.086	0.013	0.183	0.029
	0.10	0.08	0.008	0.158	0.035	0.163	0.025	0.328	0.048
	0.20	0.173	0.016	0.25	0.066	0.301	0.049	0.536	0.106
LOF	0.05	0.035	0.004	0.059	0.012	1	1	1	1
	0.10	0.074	0.01	0.115	0.022	1	1	1	1
	0.20	0.163	0.021	0.225	0.039	1	1	1	1
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
Credit card									
IForest	0.05	0.039	0.004	0.066	0.011	0.967	0.863	0.983	0.917
	0.10	0.083	0.009	0.123	0.021	0.982	0.916	0.993	0.953
	0.20	0.168	0.023	0.238	0.045	0.992	0.951	0.997	0.973
LOF	0.05	0.034	0.005	0.109	0	0.03	0.005	0.047	0.01
	0.10	0.071	0.009	0.153	0.005	0.055	0.009	0.09	0.017
	0.20	0.158	0.017	0.248	0.088	0.109	0.016	0.155	0.029
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
KDDCup99									
IForest	0.05	0.035	0.004	0.062	0.01	0.998	0.971	1	0.989
	0.10	0.077	0.009	0.11	0.021	0.999	0.988	1	0.997
	0.20	0.167	0.019	0.215	0.037	1	0.996	1	1
LOF	0.05	0.032	0.003	0.09	0	0.061	0.013	0.087	0.024
	0.10	0.072	0.005	0.14	0.011	0.103	0.022	0.144	0.036
	0.20	0.16	0.014	0.258	0.069	0.178	0.036	0.236	0.052

(Continued on Next Page...)

Table A3: Outlier batch detection performance on real data, using Storey's correction to control the FDR. Other details are as in Table 3. *(continued)*

		FDR				Power			
		Mean		90th percentile		Mean		90-th quantile	
Model	Nominal	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
Mammography									
IForest	0.05	0.036	0.003	0.072	0.009	0.489	0.177	0.667	0.311
	0.10	0.073	0.007	0.116	0.022	0.615	0.266	0.767	0.433
	0.20	0.143	0.018	0.22	0.045	0.743	0.384	0.856	0.564
LOF	0.05	0.031	0.002	0.065	0.006	0.448	0.14	0.571	0.234
	0.10	0.066	0.005	0.118	0.017	0.58	0.228	0.699	0.35
	0.20	0.135	0.015	0.209	0.039	0.713	0.352	0.8	0.485
SVM	0.05	0.011	0.001	0.031	0	0.377	0.095	0.458	0.144
	0.10	0.024	0.002	0.054	0.003	0.492	0.157	0.568	0.221
	0.20	0.053	0.005	0.097	0.018	0.613	0.248	0.688	0.324
Digits									
IForest	0.05	0.04	0.003	0.074	0.01	0.924	0.56	0.988	0.783
	0.10	0.079	0.008	0.127	0.019	0.968	0.728	0.997	0.903
	0.20	0.161	0.02	0.234	0.042	0.99	0.86	1	0.962
LOF	0.05	0.042	0.004	0.08	0.012	0.999	0.945	1	0.999
	0.10	0.087	0.009	0.14	0.021	1	0.984	1	1
	0.20	0.178	0.022	0.258	0.045	1	0.997	1	1
SVM	0.05	0.041	0.004	0.079	0.017	0.803	0.367	0.88	0.511
	0.10	0.086	0.009	0.145	0.028	0.889	0.528	0.942	0.677
	0.20	0.179	0.019	0.265	0.039	0.95	0.691	0.978	0.808
Shuttle									
IForest	0.05	0.036	0.004	0.062	0.01	1	1	1	1
	0.10	0.077	0.009	0.116	0.019	1	1	1	1
	0.20	0.163	0.021	0.222	0.036	1	1	1	1
LOF	0.05	0.041	0.003	0.066	0.012	1	1	1	1
	0.10	0.085	0.009	0.128	0.023	1	1	1	1
	0.20	0.172	0.023	0.236	0.043	1	1	1	1
SVM	0.05	0.031	0.003	0.055	0.009	1	1	1	1
	0.10	0.067	0.008	0.107	0.018	1	1	1	1
	0.20	0.151	0.018	0.21	0.034	1	1	1	1

Table A4: Outlier batch detection performance on real data, using different data sets, machine learning models, and nominal FDR levels. Other details are as in Table A3.

		FDR				Power			
		Mean		90th percentile		Mean		90-th quantile	
Model	Nominal	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.
ALOI									
IForest	0.05	0.037	0.004	0.1	0	0.001	0	0.002	0
	0.10	0.09	0.01	0.188	0.028	0.001	0	0.004	0
	0.20	0.173	0.017	0.316	0.09	0.003	0	0.007	0.001
LOF	0.05	0.034	0.002	0.098	0	0.021	0.003	0.038	0.007
	0.10	0.07	0.006	0.185	0.004	0.042	0.005	0.071	0.013
	0.20	0.153	0.014	0.294	0.082	0.095	0.01	0.143	0.019
SVM	0.05	0.036	0.002	0.092	0	0.003	0	0.006	0.001
	0.10	0.068	0.007	0.175	0.008	0.006	0.001	0.012	0.002
	0.20	0.156	0.012	0.283	0.074	0.013	0.001	0.023	0.004
Cover									
IForest	0.05	0.041	0.002	0.107	0	0.09	0.013	0.159	0.029
	0.10	0.074	0.01	0.15	0.04	0.162	0.027	0.273	0.06
	0.20	0.151	0.022	0.242	0.078	0.292	0.051	0.485	0.093
LOF	0.05	0.04	0.004	0.07	0.011	1	1	1	1
	0.10	0.083	0.009	0.121	0.023	1	1	1	1
	0.20	0.173	0.023	0.234	0.045	1	1	1	1
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
Credit card									
IForest	0.05	0.043	0.005	0.07	0.014	0.966	0.862	0.986	0.923
	0.10	0.087	0.012	0.133	0.027	0.983	0.914	0.995	0.957
	0.20	0.179	0.026	0.256	0.049	0.992	0.949	0.999	0.977
LOF	0.05	0.029	0.001	0.093	0	0.028	0.005	0.044	0.011
	0.10	0.06	0.004	0.134	0.005	0.051	0.009	0.08	0.016
	0.20	0.14	0.012	0.243	0.06	0.101	0.016	0.15	0.026
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
KDDCup99									
IForest	0.05	0.037	0.004	0.061	0.012	0.998	0.972	1	0.991
	0.10	0.079	0.009	0.117	0.02	0.999	0.988	1	0.997
	0.20	0.166	0.021	0.232	0.038	1	0.996	1	1
LOF	0.05	0.036	0.001	0.102	0	0.062	0.011	0.094	0.021
	0.10	0.076	0.004	0.153	0.01	0.104	0.02	0.154	0.034
	0.20	0.161	0.012	0.264	0.053	0.18	0.035	0.261	0.059

(Continued on Next Page...)

Table A4: Outlier batch detection performance on real data, using different data sets, machine learning models, and nominal FDR levels. Other details are as in Table A3. *(continued)*

		FDR				Power			
		Mean		90th percentile		Mean		90-th quantile	
Model	Nominal	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.
SVM	0.05	0	0	0	0	0	0	0	0
	0.10	0	0	0	0	0	0	0	0
	0.20	0	0	0	0	0	0	0	0
Mammography									
IForest	0.05	0.031	0.004	0.057	0.01	0.472	0.146	0.611	0.256
	0.10	0.065	0.006	0.11	0.023	0.601	0.234	0.726	0.36
	0.20	0.134	0.014	0.197	0.038	0.732	0.352	0.825	0.49
LOF	0.05	0.033	0.004	0.061	0.009	0.434	0.127	0.552	0.216
	0.10	0.067	0.008	0.12	0.023	0.571	0.212	0.683	0.328
	0.20	0.138	0.016	0.204	0.037	0.707	0.331	0.802	0.447
SVM	0.05	0.012	0.001	0.03	0	0.389	0.095	0.484	0.137
	0.10	0.027	0.002	0.054	0.002	0.506	0.158	0.595	0.22
	0.20	0.06	0.004	0.098	0.016	0.627	0.248	0.709	0.323
Digits									
IForest	0.05	0.035	0.002	0.061	0.008	0.918	0.523	0.979	0.773
	0.10	0.075	0.007	0.119	0.017	0.966	0.7	0.997	0.872
	0.20	0.163	0.017	0.253	0.038	0.989	0.841	1	0.951
LOF	0.05	0.04	0.002	0.072	0.008	0.999	0.941	1	0.998
	0.10	0.083	0.006	0.127	0.018	1	0.983	1	1
	0.20	0.169	0.017	0.241	0.039	1	0.996	1	1
SVM	0.05	0.037	0.002	0.063	0.009	0.807	0.347	0.886	0.487
	0.10	0.082	0.007	0.121	0.019	0.894	0.517	0.94	0.659
	0.20	0.169	0.015	0.234	0.028	0.951	0.686	0.977	0.797
Shuttle									
IForest	0.05	0.042	0.004	0.069	0.012	1	1	1	1
	0.10	0.086	0.011	0.127	0.023	1	1	1	1
	0.20	0.176	0.025	0.244	0.045	1	1	1	1
LOF	0.05	0.039	0.004	0.063	0.013	1	1	1	1
	0.10	0.079	0.009	0.118	0.023	1	1	1	1
	0.20	0.164	0.022	0.226	0.041	1	1	1	1
SVM	0.05	0.032	0.003	0.061	0.01	1	1	1	1
	0.10	0.069	0.008	0.111	0.02	1	1	1	1
	0.20	0.153	0.018	0.22	0.038	1	1	1	1