The Annals of Statistics
2023, Vol. 51, No. 1, 149–178
https://doi.org/10.1214/22-AOS2244

TESTING FOR OUTLIERS WITH CONFORMAL P-VALUES

BY STEPHEN BATES[1,a], EMMANUEL CANDÈS[2,b], LIHUA LEI[3,c], YANIV ROMANO[4,d] AND MATTEO SESIA[5,e]

[1]*Departments of Statistics and of EECS, University of California, Berkeley,* [a]*stephenbates@cs.berkeley.edu*
[2]*Departments of Statistics and Mathematics, Stanford University,* [b]*candes@stanford.edu*
[3]*Graduate School of Business, Stanford University,* [c]*lihualei@stanford.edu*
[4]*Departments of Electrical Engineering and of Computer Science, Israel Institute of Technology,* [d]*yromano@cs.technion.ac.il*
[5]*Department of Data Sciences and Operations, University of Southern California,* [e]*sesia@marshall.usc.edu*

This paper studies the construction of p-values for nonparametric outlier detection, from 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 general framework yielding 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 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. Further, 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 experiments on real and simulated data.

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 samples from P_X as *inliers*. This problem has applications in many domains, including medical diagnostics [76], spotting frauds or intrusions [58], forensic analysis [25], monitoring engineering systems for failures [75] and *out-of-distribution* detection in machine learning [32, 46, 47, 50]. A variety of machine-learning tools have been developed to address this task, which is sometimes called *one-class classification* [55, 60] because the data in $\mathcal{D}$ do not contain any outliers. However, such algorithms are often complex and their outputs are not directly covered by any precise statistical guarantees. Fortunately, conformal inference [82, 83] 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) [9]: the

Received April 2021; revised May 2022.
MSC2020 subject classifications. Primary 62G10, 62J15; secondary 62G15.
Key words and phrases. Conformal inference, out-of-distribution, false discovery rate, positive dependence.

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 statistical 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 [11, 19]. 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 (e.g.) 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 $F(\hat{s}(X_i))$ 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 $F(\hat{s}(X_i))$ 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 $\hat{F}_n$. In the following, we will discuss how to construct provably valid conformal p-values for a future observation X_{2n+1} by evaluating

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

where g is a suitable *adjustment function*, and the symbol $\circ$ denotes a composition; that is, $(f \circ g)(x) = f(g(x))$. 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.

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

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

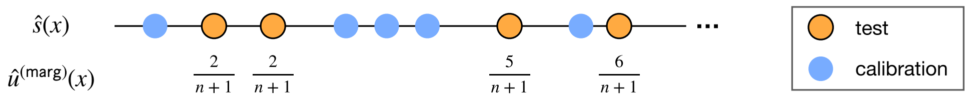

FIG. 1. *Visualization of the joint distribution of the conformal p-values. The distribution of $\hat{s}(x)$ is the same for calibration and inlier test points. The conformal p-value for each test point is the number of calibration points to its left, divided by the total number of calibration points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

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), that is,

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

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 $\{1/(n+1), 2/(n+1), \dots, 1\}$ if $X \sim P_X$ independently of the data in $\mathcal{D}^{\text{train}}$ [82, 83]. (If $\hat{s}(X)$ is not continuous, one can still verify that $\hat{u}^{(\text{marg})}(X)$ is superuniform in distribution.) However, this is not necessarily 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 anticonservative due to random fluctuations in the distribution of scores within $\mathcal{D}^{\text{cal}}$. Intuitively, this means the marginal p-values in (3) are only valid *on average* if data in $\mathcal{D}^{\text{cal}}$ are treated as random. Unfortunately, this guarantee may be too weak to be satisfactory for a practitioner who wants to compute p-values for a large number of test points but is constrained to working with a single calibration data set. Indeed, the numerical experiments presented in Section 5.2 will show that inferences based on marginal conformal p-values may be systematically invalid for a large fraction of practitioners working with "unlucky" calibration data sets.

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. 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 [24] for the global null that there are no outliers in $\mathcal{D}^{\text{test}}$, even if the calibration data in $\mathcal{D}^{\text{cal}}$ are treated as random, 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 *average* FDR control of the Benjamini–Hochberg (BH) procedure [9], even if the latter is applied with Storey's correction [72]. The behaviors of additional multiple testing procedures, such as the harmonic mean [93], Simes method [69] and Stouffer's method [73], applied to conformal p-values will be investigated empirically in Section 5.

In any case, regardless of whether the mutual dependence among marginal p-values theoretically invalidates the average inferences of a particular multiple-testing procedure, one may sometimes be interested in obtaining stronger guarantees conditional on the calibration data. 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. If the stakes are high, the researcher may need clear statistical guarantees about the output of such procedure (as opposed to blindly trusting a black-box model) and, therefore, 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 these data* $\mathcal{D}$. Instead, marginal validity suggests the system will work *on average* for different researchers starting from different data; of course, that may not be fully satisfactory for any one of them.

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

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

if $X_{2n+1} \sim P_X$, for any value of $\delta \in (0,1)$ prespecified 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. Further, 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 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, Section 4 will show 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 be helpful to interpret the output of black-box methods directly, without p-values. (However, as statisticians, we prefer the p-value approach because it is more versatile.) Further, we will show 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 and real data, in combination with different multiple testing procedures. These experiments will confirm empirically 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 nonparametric, in the sense that we leverage the information contained in a 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 [16, 30, 62, 92]. A wealth of data mining and machine-learning methods have been developed to address our nonparametric task [1, 2, 17, 37, 65]; these do not provide 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 [82, 83], which has been applied before in the context of outlier detection [13, 27, 29, 35, 45, 70]. 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 [23, 80, 81, 84, 86]. However, that guarantee no longer holds in the offline setting or beyond the global null. The most closely related work is that of [78], which extends conformal inference to provide a form of calibration-conditional coverage. That paper focused 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 [78]) not just for a prespecified 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, [5] 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 multiclass classification [3, 15, 31, 64] or in regression [18, 28, 36, 40, 63]. This paper is orthogonal, as our results can 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 [4, 38, 79], our paper focuses on sample splitting [48, 56]. The latter has the advantage of being computationally efficient, 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 [89], 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. Several works have explored the use of machine-learning and data hold-out methods for two-sample testing [26, 34, 39, 44, 52], reinforcing 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 [90, 91], Wald [87] and Tukey [77]. See [43] for a overview of the subject, [78] for a discussion of their connection with conformal inference, and [6, 57] 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 [96] with bootstrap rather than conformal inference methods.

2. Marginal conformal inference for outlier detection.

We begin by carefully studying the marginal validity of multiple tests based on split-conformal outlier detection p-values. The conformal p-values in (3) are marginally valid for the hypothesis that a test point follows the distribution P_X (see (2)), but they are not independent of each other when considering multiple test points. Consequently, they cannot be naively used to test a global null hypothesis that no points in a test set are outliers, with Fisher's combination test [24], for example. The failure of Fisher's test is caused by the particular dependence induced by the shared calibration data, although other procedures turn out to be robust to such dependence. In particular, we will prove conformal p-values are *positive regression dependent on a subset* (PRDS), which combined with the results of [11], implies the BH procedure will control the FDR.

2.1. *A negative result*: *Global testing with conformal p-values can fail.*

Fisher's combination test [24] 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 2.1. *Assume that $\hat{s}(X)$ is continuously distributed. Then, for any finite-valued function $G : [0, 1] \mapsto \mathbb{R}$, and for any pair of nulls (i, j),*

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

Motivated by Lemma 2.1 (see Section S1.1 in the Supplementary Material [7] 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 $|\mathcal{D}^{\mathrm{test}}|$ is proportional to $|\mathcal{D}^{\mathrm{cal}}|$.

THEOREM 2.2 (Type-I error of Fisher's combination test). *Assume $\hat{s}(X)$ is continuously distributed. Under the global null, if $m = \lfloor \gamma n \rfloor$ for some $\gamma \in (0, \infty)$, as n tends to infinity,*

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

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

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

The above asymptotic limits are independent of the distribution of $\hat{s}(X)$. In Section S1 of the Supplementary Material, we prove Theorem 2.2 holds for a broad class of combination tests based on $\sum_{i=1}^{n} G(\hat{u}^{(\mathrm{marg})}(X_{2n+i}))$, as long as $G(U)$ 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 $G(U) \sim \chi^2(2)$.

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

Corrections of Fisher's combination test are possible for some dependence structures. By Lemma 2.1, 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 Section S1.1 of the Supplementary Material 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 Section S1.2 of the Supplementary Material (Theorem S1) that a valid correction rejects the global null if

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

In Section S1.2 of the Supplementary Material, we also confirm the validity of (6) via Monte-Carlo simulations and show this is asymptotically equivalent to the correction proposed by [12, 42] 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 BH procedure, are known to be robust to a particular type of mutual p-value dependence called *positive regression dependent on a subset* (PRDS) [11].

DEFINITION 2.3. A random vector $X = (X_1, \dots, X_m)$ is PRDS on $I_0 \subset \{1, \dots, m\}$ if $\mathbb{P}[X \in A \mid X_i = x]$ is increasing in x for any $i \in I_0$ and any increasing set A.

In the multiple testing literature, X is often said to be PRDS if it is PRDS on the set of nulls. Above, for vectors a and b of equal dimension, we say $a \succeq b$ if all coordinates of a are no smaller than those of b, pairwise, and a set $A \subset \mathbb{R}^m$ is *increasing* if $a \in A$ and $b \succeq a$ implies $b \in A$. The PRDS property is a demanding form of positive dependence, which can be loosely interpreted as saying all pairwise correlations are positive. In view of the definition of marginal p-values in (3) and of Lemma 2.1, it should be intuitive that larger calibration scores 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 2.4 (Conformal p-values are PRDS). *Assume that $\hat{s}(X)$ is continuously distributed. Consider m test points $X_{2n+1}, \dots, X_{2n+m}$ such that the inliers are jointly independent of each other and of the data in $\mathcal{D}$. Then the marginal conformal p-values $(\hat{u}^{(\mathrm{marg})}(X_{2n+1}), \dots, \hat{u}^{(\mathrm{marg})}(X_{2n+m}))$ are PRDS on the set of inliers.*

If $\hat{s}(X)$ is not continuous, we can prove the PRDS property by modifying the p-value definition in (3); see Section S1.3 of the Supplementary Material. Theorem 2.4 implies that marginal conformal p-values can be used with the BH procedure to control the FDR for the null hypotheses

$$H_{0,i} : X_i \sim P_X, \quad i \in \{2n+1, \dots, 2n+m\},$$

although this guarantee only holds on average over random test *and calibration* data sets.

COROLLARY 2.5 (Benjamini and Yekutieli [11]). *In the setting of Theorem 2.4, the BH procedure applied at level $\alpha \in (0, 1)$ to $(\hat{u}^{(\mathrm{marg})}(X_{2n+1}), \dots, \hat{u}^{(\mathrm{marg})}(X_{2n+m}))$ 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 $\mathcal{H}_0 = \{i : H_{0,i}$ 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.

REMARK. The BH procedure applied to the marginal conformal p-values is equivalent to the semisupervised BH procedure proposed by [53] (posted on arXiv 2 months after our paper), which was first studied by [88] and later generalized by [95] and [61]. These works employ a martingale-based technique to prove FDR control without relying on the PRDS property. Theorem 3.1 in [53] also proves a lower bound showing the FDR is almost $\pi_0\alpha$.

2.3. *A positive result: Storey's correction does not break FDR control.* When the proportion of nulls is much smaller than one, as it may be the case in many out-of-distribution detection problems, the BH procedure is conservative, as shown in Corollary 2.5. 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 [71, 72] 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 BH procedure at level $\alpha/\hat{\pi}_0$; see Section S1.4 of the Supplementary Material for details. If the null p-values are superuniform (2), mutually independent and independent of the nonnull p-values, this provably controls the FDR in finite samples [72]. However, unlike in its standard version, the BH procedure with Storey's correction is not generally guaranteed to control the FDR if the p-values are PRDS; see Section 6.3 of [10].

Surprisingly, we show below that the positive correlation (Lemma 2.1) among the marginal conformal p-values does not break the FDR control at all. The proof of Theorem 2.6 rests on a novel FDR bound for the BH procedure with Storey's correction applied to any type of superuniform p-values that are PRDS and almost surely bounded from below by a constant; see Theorem S2 in Section S1.4 of the Supplementary Material [7]. This result is not limited to conformal p-values and may also be useful for other multiple testing problems, such as those involving permutation p-values.

THEOREM 2.6 (Storey's BH with conformal p-values controls the FDR). *Set $\lambda = K/(n+1)$ for any integer K. Assume $\hat{s}(X)$ is continuously distributed. In the setting of Corollary 2.5, the BH procedure with Storey's correction applied to marginal conformal p-values $(\hat{u}^{(\mathrm{marg})}(X_{2n+i}))_{i=1}^{m}$ controls the FDR at the nominal level.*

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 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 below α, conditional on $\mathcal{D}$, for a *fixed* threshold $\alpha \in (0,1)$. In other words, we wish to understand the conditional false positives rate (FPR) corresponding to the threshold α,

$$\mathrm{FPR}(\alpha; \mathcal{D}) := \mathbb{P}\big[\hat{u}^{(\mathrm{marg})}(X_{2n+1}) \leq \alpha \mid \mathcal{D}\big], \tag{8}$$

beyond what we know from (2), which is $\mathbb{E}[\mathrm{FPR}(\alpha;\mathcal{D})] \leq \alpha$. The quantity in (8) was already studied precisely by [78]. We revisit this topic here because it serves as an intuitive introduction to the more involved novel high-probability bounds presented later.

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

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

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

PROPOSITION 3.1 (Pointwise FPR of marginal conformal p-values, from [78]). *Let $\ell = \lfloor (n+1)\alpha \rfloor$. If $\hat{s}(X)$ is continuously distributed*, $\mathrm{FPR}(\alpha;\mathcal{D}) \sim \mathrm{BETA}(\ell, n+1-\ell)$.

Figure 2 visualizes the FPR distribution from Proposition 3.1 for different calibration set sizes. This shows precisely how a smaller $\mathcal{D}^{\mathrm{cal}}$ makes marginal p-values more conservative on average, but also more likely to be overly liberal on occasion. For example, there is a non-negligible probability that $\mathrm{FPR}(0.1;\mathcal{D}) > 0.15$ with 100 calibration points, whereas it seems very unlikely that $\mathrm{FPR}(0.1;\mathcal{D}) > 0.12$ with 1600 calibration points. However, it is still quite

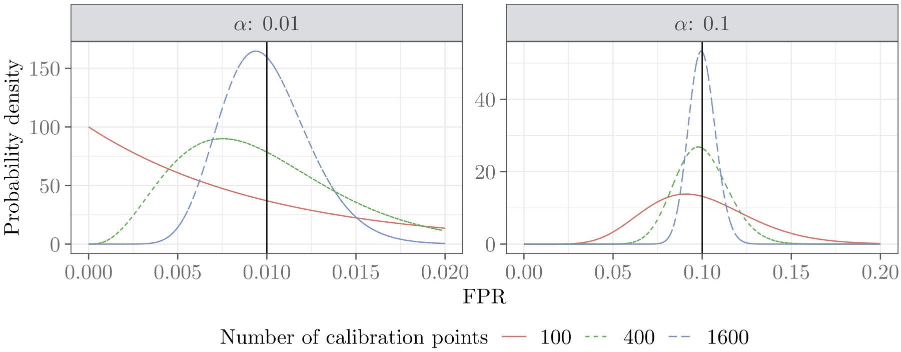

FIG. 2. *Distribution of the false positive rate obtained by thresholding marginal conformal p-values at levels* $\alpha = 0.01$ *and* $\alpha = 0.1$, *as a function of the number of calibration points.*

possible that $\mathrm{FPR}(0.01; \mathcal{D}) > 0.015$ even with 1600 calibration points. In general, Proposition 3.1 implies the coefficient of variation (relative spread) of the FPR is approximately proportional to $(|\mathcal{D}^{\mathrm{cal}}|\alpha)^{-1/2}$. While this result is informative and it is broadly relevant to the issue of how to best choose the number of calibration data points for split-conformal inference [68], it is limited for our purposes. In fact, 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.

3.2. *A generic strategy to adjust marginal conformal p-values*. Proposition 3.1 implies marginal conformal p-values may be anticonservative conditional on $\mathcal{D}$. Therefore, in the language of (1), our goal is to find an adjustment function leading to conditionally valid p-values, that is, satisfying (4). The following theorem suggests a generic strategy through a simultaneous upper confidence bound for order statistics.

THEOREM 3.2 (Conditional p-value adjustment). *Let* $U_1, \dots, U_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0, 1])$, *with order statistics* $U_{(1)} \le U_{(2)} \le \cdots \le U_{(n)}$, *and fix any* $\delta \in (0, 1)$. *Suppose* $0 \le b_1 \le b_2 \le \cdots \le b_n \le 1$ *are n reals such that*

$$\mathbb{P}[U_{(1)} \le b_1, \dots, U_{(n)} \le b_n] \ge 1 - \delta. \tag{9}$$

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

$$h(t) = b_{\lceil (n+1)t \rceil}, \quad t \in [0, 1]. \tag{10}$$

Then $\hat{u}^{(\mathrm{ccv})}(X_{2n+1})$ *is a calibration-conditional valid p-value.*

Figure 3 illustrates the idea of Theorem 3.2. Here, we set $n = 1000$ and generate 100 independent realizations of the order statistics $(U_{(1)}, \dots, U_{(n)})$. 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 $(U_{(1)}, \dots, U_{(n)})$, while the orange and yellow curves correspond to two particular sequences of b_i values derived from the generalized Simes inequality for $\delta = 0.1$ and the DKWM [21, 54] inequality, detailed in the next subsection. We observe relatively few paths cross the orange curve, and all crossings occur at small indices. This suggests the upper confidence bounds provided by Theorem 3.2 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 $\hat{u}^{(\mathrm{marg})}(X) = 25/(n + 1) \approx 0.025$ results in a CCV p-value of $h(25/(n + 1)) = b_{25} \approx 0.0377$ in this case.

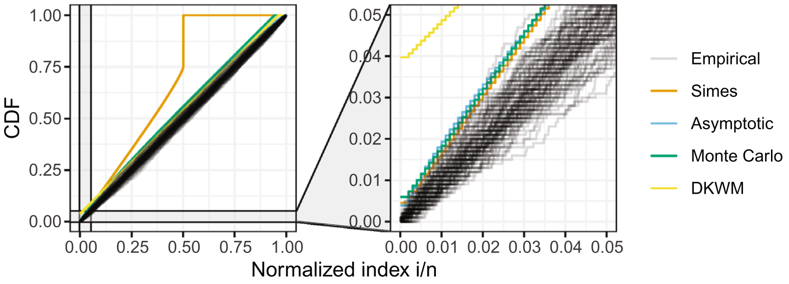

FIG. 3. *Illustration of Theorem 3.2 with $n = 1000$ and $\delta = 0.1$. The orange and yellow curves give the sequences derived by the generalized Simes inequality with $k = 500$ and the DKWM inequality, respectively. The blue and green curves (very close to each other) give the corresponding sequences obtained with the asymptotic and Monte Carlo adjustments described below. The right panel zooms in on small indices.*

3.3. *Simes adjustment of marginal conformal p-values.* Large p-values typically do not matter in multiple testing problems, as it is the small ones that determine the rejections. Therefore, to maximize power, we would like the b_i values in Theorem 3.2 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 $(b_1, \dots, b_n)$ sequences with this property.

PROPOSITION 3.3 (Generalized Simes inequality, from equation (3.5) in [66]). *For any positive integer $k \le n$, the uniform bound (9) in Theorem 3.2 holds with*

$$b^{\mathrm{s}}_{n+1-i} = 1 - \delta^{1/k}\left(\frac{i \cdots (i-k+1)}{n \cdots (n-k+1)}\right)^{1/k}, \quad i = 1, \dots, n. \tag{11}$$

The original motivation of [66] 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 3.3 differently, choosing $k = n/2$ so that the b^{s}_i values with lower indices i are as small as possible while those with larger indices i may be uninformative (note that $b^{\mathrm{s}}_{n-k+2} = \dots = b^{\mathrm{s}}_n = 1$). In particular, our choice corresponds to

$$b^{\mathrm{s}}_1 = 1 - \delta^{2/n} = 1 - \exp\left\{-\frac{2\log(1/\delta)}{n}\right\} \approx \frac{2\log(1/\delta)}{n}.$$

Therefore, the smallest possible marginal p-value would be mapped to $h(1/(n+1)) \approx 2\log(10)/n = 4.61/n$, if $\delta = 0.1$, for example, since $\hat{u}^{(\mathrm{ccv})}(X) = h(\hat{u}^{(\mathrm{marg})}(X))$. If $n = 1000$, then $h(1/(n+1)) \approx 0.0046$, 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 [21, 54] would imply a result similar to that of Proposition 3.3 but with

$$b^{\mathrm{d}}_i = \min\{(i/n) + \sqrt{\log(2/\delta)/2n}, 1\}; \tag{12}$$

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 in Section S3 of the Supplementary Material [7], where we also consider an additional uniform bound based on the linear-boundary crossing probability for the empirical CDF [20]. 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 Section S4 of the Supplementary Material [7]. (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).

3.4. *Asymptotic adjustment of marginal conformal p-values.* The Simes adjustment with $k = n/2$ leads to p-values satisfying (4) exactly; however, this causes the smallest possible marginal conformal p-values to be inflated by a factor of order $1/n$, and larger ones may be inflated even more. A natural question is whether this approach is efficient or whether more powerful alternatives may be available to achieve (4). We begin to address this matter by comparing the Simes adjustment to an alternative *asymptotic* approach that provides a natural benchmark; this solution will be valid in the limit of large n but does not guarantee (4) exactly in finite samples. Recall Donsker's theorem, the classical result from empirical process theory stating that, in the large-n limit, the rescaled difference between the true and the empirical CDFs of the calibration scores respectively F and $\hat{F}_n$, converges in distribution to a standard Brownian Bridge. Precisely, $\sqrt{n}(\hat{F}_n - F) \xrightarrow{d} \mathbb{G}$, where $\mathbb{G}$ is the Gaussian process on [0, 1] with mean zero and covariance $\mathbb{E}[\mathbb{G}(t_1)\mathbb{G}(t_2)] = t_1 \wedge t_2 - t_1 t_2$, for all $t_1, t_2 \in [0, 1]$. This result suggests the following *asymptotic adjustment* of marginal conformal p-values.

As a starting point, note that $\sup_{t\in[0,1]} |\mathbb{G}(t)|$ follows the Kolmogorov distribution [41], whose $1 - \delta$ quantile, namely q_δ^{K}, can be computed. Therefore, a simple way of constructing approximately valid conditional conformal p-values would be to add $q_\delta^{\mathrm{K}}/\sqrt{n}$ to the marginal p-values. Unfortunately, this naive solution would suffer from the same limitation of the DKWM approach mentioned in the previous section: it is a correction of constant size which is not very attractive for multiple testing because it is extremely conservative for small p-values of order $1/n$. Instead, a more useful solution is suggested by the adaptive bound of [22], which proved that the empirical process $\hat{V}_n(t)$ defined as

$$\hat{V}_n(t) = \sqrt{n}\frac{F(t) - \hat{F}_n(t)}{\sqrt{\hat{F}_n(t)[1 - \hat{F}_n(t)]}}, \quad t \in [0, 1],$$

satisfies $\lim_{n\to\infty} \mathbb{P}[\sup_{t\in[0,1]} \hat{V}_n(t) \le c_n(\delta)] \ge 1 - \delta$, where $c_n(\delta)$ is defined as

$$c_n(\delta) := \frac{-\log[-\log(1-\delta)] + 2\log\log n + (1/2)\log\log\log n - (1/2)\log\pi}{\sqrt{2\log\log n}}.$$

This yields a straightforward asymptotic simultaneous upper confidence bound for $F(t)$ and, in light of Theorem 3.2, it suggests the following approximately valid adjustment:

$$\hat{u}^{(\text{a-ccv})} = h^{\mathrm{a}} \circ \hat{u}^{(\text{marg})}, \tag{13}$$

where h^{a} is the piecewise constant function on [0, 1] defined such that $h^{\mathrm{a}}(t) = b^{\mathrm{a}}_{\lceil (n+1)t \rceil}$, for $t \in [0, 1]$, with $b^{\mathrm{a}}_0 = 0$, $b^{\mathrm{a}}_{n+1} = 1$, and

$$b_i^{\mathrm{a}} = \min\left\{\frac{i}{n} + c_n(\delta)\frac{\sqrt{i(n-i)}}{n\sqrt{n}}, 1\right\}, \quad i = 1, \dots, n. \tag{14}$$

In Section S2.1.1 of the Supplementary Material [7], we will show that $b_1^{\mathrm{a}} \le b_2^{\mathrm{a}} \le \cdots \le b_n^{\mathrm{a}}$, as required by Theorem 3.2. See Figure 3 for a visualization of the simultaneous CDF bound corresponding to this adjustment function. The smallest possible marginal p-value is mapped by this function to $h^{\mathrm{a}}(1/(n+1)) \approx (1 + c_n(\delta))/n$. For example, if $\delta = 0.1$ and $n = 1000$, this is approximately $4.09/n \approx 0.0041$, which is very similar to the corresponding constant 0.0046 obtained with the Simes adjustment. However, $\hat{u}^{(\text{a-ccv})}$ has the advantage of being reasonably tight for all p-values, not just the smallest ones, and thus it will generally allow for higher power compared to the Simes adjustment when n is large.

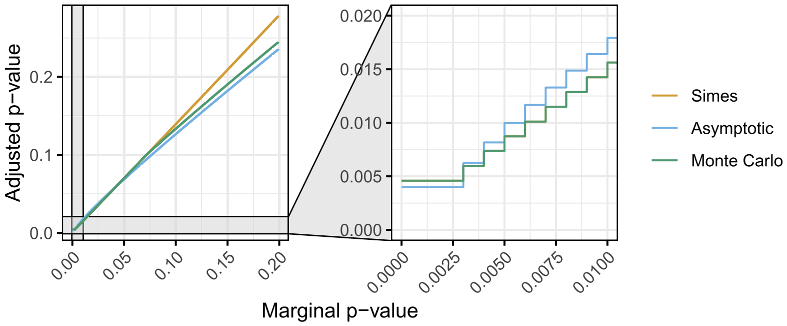

FIG. 4. *Comparison of different adjustment functions, with $n = 1000$ and $\delta = 0.1$. In the zoomed-in panel on the right-hand side, the Simes (orange) and Monte Carlo (green) curves cannot be distinguished.*

3.5. *Monte Carlo adjustment of marginal conformal p-values.* Although the Simes adjustment is more conservative than the asymptotic one in the limit of large n, it has two distinct advantages in finite samples. First, it leads to p-values satisfying (4) exactly, with no asymptotic approximations. Second, the peculiar shape of its uniform empirical CDF envelope allows it to apply smaller corrections to relatively low p-values, possibly yielding higher power in multiple-testing applications; see Figure 4 for an illustration. These observations motivate the development of the following new type of adjustment function, which is based on *Monte Carlo* rather than analytical calculations and is designed to combine the strengths of the two aforementioned approaches. In particular, the Monte Carlo solution proposed here is based on a uniform empirical CDF bound that is (a) theoretically valid in finite samples and (b) whose shape mimics that of the Simes approach for very small p-values while tracking the asymptotic envelope relatively closely for larger ones; see Figure 4 for a preview.

Having fixed any n and δ, denote by $h^{\mathrm{s}} : [0, 1] \to [0, 1]$ the Simes piecewise constant function obtained by combining (10) with (11), using $k = n/2$. Recall that this satisfies (9) exactly. Let also $h^{\mathrm{a},\hat{\delta}} : [0, 1] \to [0, 1]$ denote the asymptotic piecewise constant function obtained by combining (10) with (14), after replacing the prespecific parameter δ with a variable $\hat{\delta}$, which can take any values in $(0, 1)$. Note that it will be useful to keep the dependence of this function on $\hat{\delta}$ explicit. Recall that $h^{\mathrm{a},\hat{\delta}}$ satisfies (9) approximately if n is large and $\hat{\delta} = \delta$. Next, define a new piecewise constant function $h^{\mathrm{m},\hat{\delta}} : [0, 1] \to [0, 1]$ as

$$h^{\mathrm{m},\hat{\delta}}(t) = \min\{h^{\mathrm{s}}(t), h^{\mathrm{a},\hat{\delta}}(t)\}, \quad t \in [0, 1]. \tag{15}$$

Note that this function can be conveniently written in the form of (10) with a suitable choice of $b_1, \dots, b_n$. Now, the goal is to find the smallest possible $\hat{\delta}$, as a function of n and δ, such that the $b_1, \dots, b_n$ sequence corresponding to the function $h^{\mathrm{m},\hat{\delta}}$ defined in (15) satisfies (9). The problem can be solved with a bisection search for $\hat{\delta}$ on $(0, 1)$, approximating the probability in (9) through a simple Monte Carlo simulation—it suffices to generate a sufficiently large number of independent random samples of size n from a uniform distribution. A feasible solution always exists because $h^{\mathrm{m},\hat{\delta}}$ reduces to h^{s} as $\hat{\delta} \to 1$, and h^{s} satisfies (9). This Monte Carlo simulation is not computationally expensive for reasonable values of n, as long as δ is not too small; for example, it takes a few seconds on a personal computer to obtain a very accurate estimate of $\hat{\delta}$ with $\delta = 0.1$ and n as large as $10,000$. Of course, if n is extremely large, the Monte Carlo simulation is not even needed, as in that case one could just rely directly on the asymptotic adjustment. See Figure 3 for a visualization of the simultaneous CDF bound corresponding to this adjustment function.

While the Monte Carlo adjustment approaches the asymptotic one in the limit of large n, it may lead to more powerful p-values for multiple testing if n is small. In fact, the Simes function $h^{\mathrm{s}}(t)$ can be lower than the asymptotic $h^{\mathrm{a},\delta}(t)$ for values of t very close to 0, and $h^{\mathrm{m},\hat{\delta}}(t)$ inherits this ability of preserving very small p-values relatively intact, as shown in Figure 4. At the same time, as it will be demonstrated shortly, the Monte Carlo adjustment tends to be more powerful than the Simes adjustment when testing a single hypothesis, or when dealing with many nonnull hypotheses, because $h^{\mathrm{a},\delta}(t)$ is lower than $h^{\mathrm{s}}(t)$ for moderately small values of t; see the left-hand side panel of Figure 4. Additional figures in Section S3 of the Supplementary Material [7] show that this relative advantage grows even larger as n increases.

The Monte Carlo adjustment applied in this paper and implemented in the accompanying software package involves an additional modification to the expression in (15), whose discussion has been postponed until now to simplify the explanation. In practice, $h^{\mathrm{m},\hat{\delta}}(t)$ is defined as in (15) only for $t \leq 1/2$; then, for $t > 1/2$, the function is extended it as a tangent straight line because there would be little point in tightening the CDF envelope above $1/2$, as that region involves p-values unlikely to be rejected anyway. The advantage of this approach is that it decreases the boundary crossing probability of the empirical CDF for all $t > 1/2$ compared to the asymptotic solution, allowing a slightly more liberal adjustment for the more interesting p-values below 1/2; see Figure S4 in Section S3 of the Supplementary Material [7].

3.6. *Power analyses of conformal p-value adjustments.* As marginal p-values are smaller than calibration-conditional p-values, the latter tend to involve some loss of power, while the former are not always valid, depending on the multiple testing procedure utilized. In this section, we would like to study the power gap between the marginal and calibration-conditional approaches within settings in which both types of conformal p-values lead to valid tests. However, traditional power analyses require stronger modeling assumptions (i.e., the distributions of inliers and outliers) and the specification of additional algorithmic details (i.e., the form of the conformity score functions) compared to the framework followed in this paper; in fact, conformal p-values are extremely flexible and can be applied in fully nonparametric settings with any conformity score function. We overcome this hurdle by analyzing the *effective level* of a test applied to calibration-conditional p-values as a proxy for a power analysis. More precisely, a test at level α applied to calibration-conditional p-values is generally equivalent to an analogous test at level α' applied to marginal p-values, for some $\alpha' < \alpha$. Comparing α to α' gives a measure of the loss in power incurred by calibration-conditional p-values that is specific to a particular testing procedure, but requires no assumptions about either the machine learning model utilized to compute conformity scores or the inlier and outlier distributions. Thus, α' is studied below for different testing procedures.

3.6.1. *Testing a single hypothesis.* Suppose a marginal conformal p-value $\hat{u}^{(\mathrm{marg})}(X_{2n+1})$ for a single test point X_{2n+1} is available, and we wish to test whether X_{2n+1} is an outlier. The level-α test based on the marginal p-value rejects when $\hat{u}^{(\mathrm{marg})}(X_{2n+1}) \leq \alpha$. We will compare this to a test based on a calibration-conditional p-value. That is, we take the marginal p-value and adjust it with a generic piecewise constant function $h : [0, 1] \to [0, 1]$ in the form of (10). Then we reject the null if $h \circ \hat{u}^{(\mathrm{marg})} \leq \alpha$, or, equivalently, if

$$\hat{u}^{(\mathrm{marg})} \leq i^*(\alpha; h)/(n+1),$$

where $i^*(\alpha; h) = \max\{i \in \{1, \dots, n\} : b_i \leq \alpha\}$ and $b_1, \dots, b_n$ indicate the step positions defining h in (10). As $\hat{u}^{(\mathrm{marg})}$ is uniformly distributed, $i^*(\alpha; h)/(n+1)$ is the effective level of the analogous marginal test.

With the asymptotic adjustment h^{a}, the threshold for the calibration-conditional test can be calculated by solving a quadratic equation, and the solution in the large-n limit is

$$\frac{i^*(\alpha; h^{\mathrm{a}})}{n+1} = O\left(\frac{\alpha}{1 + c_n^2(\delta)/n}\right) = \alpha\left[1 - O\left(\frac{\log\log n}{n}\right)\right],$$

because

$$i^*(\alpha; h^{\mathrm{a}}) = \left\lfloor \frac{c_n^2(\delta)n + 2n^2\alpha - c_n(\delta)n\sqrt{c_n^2(\delta) + 4n\alpha - 4n\alpha^2}}{2[c_n^2(\delta) + n]} \right\rfloor.$$

In words, the cost in power of the asymptotic p-value adjustment from Section 3.4 can be understood by noting that the significance threshold α is effectively decreased by a factor of order $(\log\log n)/n$. Similarly, the effective α-level with the DKWM adjustment h^{d}, given by (12), is $\alpha - O(1/\sqrt{n})$. By contrast, for the Simes adjustment, we can show the effective α-level is strictly below α when $k = \lceil \zeta n \rceil$ for some $\zeta > 0$. In fact, using the concavity of the mapping $a(x) = \log(1 - 1/x)$, Jensen's inequality implies

$$(16) \qquad b_i^{\mathrm{s}} = 1 - \delta^{1/k} e^{\frac{1}{k}\sum_{\ell=n-k+1}^{n} a(\frac{\ell}{i-1})} \geq 1 - e^{a(\frac{n-k/2+1/2}{i-1})} = \frac{i-1}{n-k/2+1/2}.$$

As a result,

$$(17) \qquad \frac{i^*(a; h^{\mathrm{s}})}{n+1} \leq \alpha(1 - \zeta/2) + o(1).$$

In this sense, the asymptotic and DKWM adjustment are nearly as efficient as the marginal test for a single hypothesis, though the former is more powerful, while the Simes adjustment is asymptotically inefficient.

Analogous threshold calculations for the Monte Carlo adjustments in the same setting cannot be performed analytically because $i^*(\alpha; h^{\mathrm{m},\hat{\delta}})$ does not have a simple expression for the sequences b corresponding to those functions h. However, these analyses are easy to carry out numerically. Figure 5(a) summarizes these power analyses by comparing the effective significance levels obtained with the alternative adjustment functions, as a function of n. The results show the Monte Carlo adjustment behaves very similar to the efficient asymptotic solution in the limit of large n, but it can be even more powerful when the sample size is small thanks to the shape of its CDF envelope, which reduces the inflation of smaller p-values. The Simes adjustment behaves similar to the Monte Carlo one when the sample size is small, but

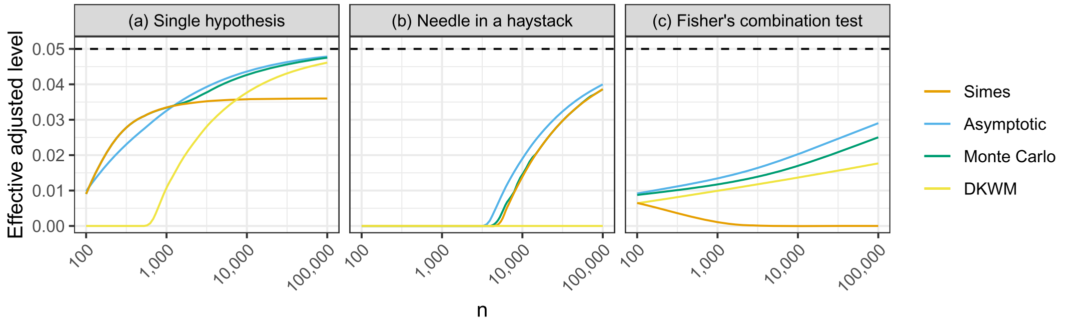

FIG. 5. *Power analysis of different adjustments for marginal conformal p-values under* 3 *alternative settings. The effective level resulting from the p-value adjustment for a test at nominal level* $\alpha = 0.05$ (*dashed horizontal line*) *is plotted as a function of the number of calibration samples*, *assuming the number of test points m grows as* $\sqrt{n}$. (*a*) *Testing a single hypothesis.* (*b*) *FWER control with a single strong signal* (*here the values for DKWM are all equal to* 0). (*c*) *Testing a global null with Fisher's combination test.*

it is not efficient in the large-n limit. In that case, the effective significance level for testing a single hypothesis does not converge at all to the nominal level α in the large-n limit. Finally, the DKWM adjustment is extremely conservative unless n is very large.

3.6.2. *Needle in a haystack.* Consider a multiple testing problem in which there are m possible outliers to be tested: the first test point, X_{2n+1}, is an outlier (a false null hypothesis), while the remaining $m-1$ points, $X_{2n+2}, \dots, X_{2n+m}$, are inliers (true nulls). The goal is to identify the outlier, controlling the familywise error rate below α. To further simplify the problem, imagine the signal strength for the true outlier is so high that the marginal conformal p-value for this point takes its minimal value with probability one:

$$\hat{u}_{2n+1}^{(\text{marg})} = \frac{1}{n+1}.$$

Then we reject the null if the adjusted p-value for the outlier is below the Bonferroni level:

$$h \circ \hat{u}_{2n+1}^{(\text{marg})} \leq \frac{\alpha}{m}. \tag{18}$$

In the case of the asymptotic adjustment function, the rejection event can be written as

$$\left\{h^{\text{a}} \circ \hat{u}_{2n+1}^{(\text{marg})} \leq \frac{\alpha}{m}\right\} \iff \left\{\hat{u}_{2n+1}^{(\text{marg})} + \frac{1}{n(n+1)} + c_n(\delta)\frac{\sqrt{n-1}}{n\sqrt{n}} \leq \frac{\alpha}{m}\right\}.$$

Thus, the calibration-conditional test at level α is equivalent to the marginal test at level $(\alpha + \Delta\alpha)/m$, where

$$\Delta\alpha = -\frac{m}{n}\left(\frac{1}{n+1} + c_n(\delta)\sqrt{\frac{n-1}{n}}\right) = -\frac{m}{n}\sqrt{2\log\log n}\big(1 + o(1)\big).$$

Here, the calibration-conditional and marginal tests only differ by a $\sqrt{\log\log n}$ factor.

In the case of the Simes adjustment with $k = n/2$, the rejection event is

$$\left\{h^{\text{s}} \circ \hat{u}_{2n+1}^{(\text{marg})} \leq \frac{\alpha}{m}\right\} \iff \left\{\hat{u}_{2n+1}^{(\text{marg})} + \frac{2\log(1/\delta)}{n}\big(1 + o(1)\big) - \frac{1}{n+1} \leq \frac{\alpha}{m}\right\},$$

which implies the equivalent level for the test is $(\alpha + \Delta\alpha)/m$, with

$$\Delta\alpha = -\frac{m}{n}\big(2\log(1/\delta) - 1 + o(1)\big).$$

Similarly, for the DKWM adjustment, it is easy to see that

$$\Delta\alpha = -\frac{m}{\sqrt{n}}\left(\sqrt{\frac{\log(2/\delta)}{2}} + o(1)\right).$$

Therefore, in the large-n limit, the Simes adjustment is even more powerful than the asymptotic correction for this problem because it does not involve the slightly suboptimal $\sqrt{\log\log n}$ factor. Unsurprisingly, the additive inflation by the DKWM adjustment results in a much larger power loss. Although the Monte Carlo method is not as amenable to analytical calculations, it is easy to verify numerically that its power is almost the same as that of the asymptotic correction in this setting; see Figure 5(b). Interestingly, the numerical power analysis in Figure 5(b) shows the asymptotic adjustment tends to be more powerful than the Simes adjustment for this problem, although it is slightly less powerful in the large-n limit. In fact, $\sqrt{\log\log n} < (2\log(1/\delta) - 1)$ unless n is extremely large or δ is extremely small. Note that m is proportional to $\sqrt{n}$ in Figure 5, which explains why the effective level with the DKWM adjustment remains stuck at zero as n grows.

3.6.3. *Fisher's combination test of the global null.* Consider a multiple testing problem in which there are m test data points $X_{2n+1}, \dots, X_{2n+m}$ and none of them are outliers. The goal is to test the global null by applying Fisher's combination test to conformal p-values modified by an adjustment function h, for different choices of the latter. Intuitively, the effective α-level of this test will depend on the expected value of Fisher's combination statistic under the null—a smaller $\mathbb{E}_{H_0}[-\log(h \circ \hat{u}^{(\mathrm{marg})})]$ yields a more conservative test. Therefore, we begin by deriving this quantity analytically for the asymptotic, DKWM and Simes adjustments; see Section S2 of the Supplementary Material [7] for further details.

THEOREM 3.4 (Expected value of Fisher's combination statistic with conformal p-values). *Fix any $\delta > 0$ and $\zeta > 0$, and let $n \to \infty$. Then, under the global null hypothesis*:

(a) $\mathbb{E}_{H_0}[-\log(h^{\mathrm{a}} \circ \hat{u}^{(\mathrm{marg})})] = 1 - (\pi/2)c_n(\delta)/\sqrt{n} + O((\log n)(\log\log n)/n)$.

(b) $\mathbb{E}_{H_0}[-\log(h^{\mathrm{d}} \circ \hat{u}^{(\mathrm{marg})})] = 1 - b_n(\delta)\log(e/b_n(\delta)) + O(\log n/n)$, *where $b_n(\delta)$ is defined as $b_n(\delta) = \sqrt{\log(2/\delta)/(2n)}$.*

(c) *If $k = \lceil \zeta n \rceil$, $\mathbb{E}_{H_0}[-\log(h^{\mathrm{s}} \circ \hat{u}^{(\mathrm{marg})})] \le 1 - \zeta - (1-\zeta)\log(1-\zeta) + O(\log n/n)$.*

All three adjustments above yield conservative tests because, in the limit of large n, $\mathbb{E}_{H_0}[\sum_{i=1}^{m} -2\log(h \circ \hat{u}^{(\mathrm{marg})}(X_{2n+i}))] < 2m$. The gap is $O(m\sqrt{\log\log n}/\sqrt{n})$ for the asymptotic adjustment (the most efficient one in this case), $O(m \log n/\sqrt{n})$ for the DKWM adjustment, and $O(m)$ for the Simes adjustment (the least efficient one in this case). In Section S2 of the Supplementary Material [7], we compute the effective α-level for each adjustment under different regimes. As those derivations are lengthy, we summarize the results below:

- For the asymptotic adjustment, the effective α is $\alpha(1 + o(1))$ if $m = o(n/\log\log n)$, and $O(1/\log^c n)$ for some constant c when $m = \gamma n$ for some $\gamma \in (0, 1)$.
- For the DKWM adjustment, the effective α is $\alpha(1 + o(1))$ if $m = o(n/\log^2 n)$, and $\exp\{-O(\log^2 n)\}$ when $m = \gamma n$ for some $\gamma \in (0, 1)$.
- For Simes, the effective α is $\exp\{-O(\min\{m, n\}/\log n)\}$ if $m/\log n \to \infty$.

In Figure 5(c), we compare the effective α-levels computed numerically with $m = \sqrt{n}$, including also the theoretically intractable Monte Carlo adjustment. These results confirm the Simes method becomes extremely conservative for large n, as its effective level tends to 0 instead of α. By contrast, the Monte Carlo adjustment yields approximately the same effective significance threshold as the asymptotic method.

Finally, it is interesting to compare these power analyses for calibration-conditional p-values with the exact adjustment of Fisher's combination test from Theorem 2.2. If $m = \gamma n$ for some $\gamma \in (0, 1)$, it follows from (5) that Fisher's combination test applied to marginal conformal p-values is valid at level α, conditional on the calibration data, if its nominal significance level is lowered by a factor that depends on δ—the proportion of calibration data sets for which the test is allowed to be invalid—but remains constant with respect to n. By contrast, applying Fisher's combination test to calibration-conditional p-values results in an effective level α that at best *decreases* as $1/\mathrm{polylog}(n)$, for the asymptotic adjustment. Therefore, calibration-conditional p-values are not always optimal with Fisher's combination test, at least not compared to the ad-hoc correction of the latter presented in Theorem 2.2, if $m = \gamma n$, but they have the advantage of flexibility. In fact, calibration-conditional p-values can be utilized by any multiple testing algorithm, including, for example, the BH procedure.

3.6.4. *Testing multiple hypotheses by the BH procedure.* Consider a multiple testing problem in which there are m test data points $X_{2n+1}, \dots, X_{2n+m}$ and the goal is to detect outliers with FDR control. If the BH procedure is applied to the adjusted p-values, all hypotheses with $h \circ \hat{u}^{(\mathrm{marg})}(X_{2n+i}) \le \alpha R(\alpha; h)/m$ are rejected, where

$$R(\alpha; h) = \max\left\{r \in \{0, 1, \dots, m\} : \#\left\{i : h \circ \hat{u}^{(\mathrm{marg})}(X_{2n+i}) \le \frac{r\alpha}{m}\right\} \ge r\right\}.$$

As a benchmark, we consider the number of rejections obtained with the marginal p-values:

$$R_{\mathrm{marg}}(\alpha) = \max\left\{r \in \{0, 1, \dots, m\} : \#\left\{i : \hat{u}^{(\mathrm{marg})}(X_{2n+i}) \le \frac{r\alpha}{m}\right\} \ge r\right\}.$$

In the case of the asymptotic adjustment,

$$\frac{h^{\mathrm{a}}(i/(n+1))}{i/(n+1)} \le \frac{n+1}{n}\left\{1 + c_n(\delta)\sqrt{\frac{n-i}{ni}}\right\}. \tag{19}$$

This quantity is decreasing in i, implying that

$$\max_i \frac{h^{\mathrm{a}}(i/(n+1))}{i/(n+1)} \le \frac{n+1}{n}\left\{1 + c_n(\delta)\sqrt{\frac{n-1}{n}}\right\} = \sqrt{2\log\log n} + o(1).$$

Therefore, all hypotheses rejected by the BH procedure applied to marginal p-values at a lower level $\alpha/(\sqrt{2\log\log n} + o(1))$ are also rejected by the BH procedure applied to adjusted p-values, implying the effective FDR level for h^{a} is at least $\alpha/(\sqrt{2\log\log n} + o(1))$. If $\sqrt{2\log\log n} << \log m$, this is more powerful than the Benjamini–Yekutieli procedure [11], whose effective FDR level is $\alpha/(\log m + O(1))$. Further, the ratio given by (19) is $1 + o(1)$ if $i/\log\log n \to \infty$, implying that, in the limit of $R_{\mathrm{marg}}(\alpha)/\log\log n \to \infty$, all marginal rejections are also rejected by the BH procedure applied to adjusted p-values with the target FDR level $\alpha(1 + o(1))$. In summary, the cost of the asymptotic adjustment never exceeds $\sqrt{2\log\log n} + o(1)$, and it is negligible if the number of rejections made by the marginal BH procedure grows faster than $\log\log n$.

In the case of the DKWM adjustment, the maximal ratio between the adjusted and marginal p-values is $O(\sqrt{n})$, though it becomes $1 + o(1)$ when $i/\sqrt{n} \to 0$. Thus, unless the marginal BH procedure can reject many more than $\sqrt{n}$ hypotheses, the power cost of the DKWM adjustment will be much higher than that of the asymptotic adjustment.

In the case of the Simes adjustment, we can show that, if $k = \lceil \zeta n \rceil$ for some $\zeta \in (0, 1)$, the ratio between the adjusted and marginal p-values is bounded by a constant that depends on δ and ζ. Analogous to (16), the concavity of $a(x)$ implies

$$b_i^{\mathrm{s}} \le 1 - \delta^{1/k} e^{\frac{1}{2}(a(\frac{n}{i-1}) + a(\frac{n-k+1}{i-1}))} = 1 - \delta^{1/k}\sqrt{\left(1 - \frac{i-1}{n}\right)\left(1 - \frac{i-1}{n-k+1}\right)}.$$

Since $k = \lceil \zeta n \rceil$, $\sqrt{(1 - (i-1)/n)(1 - (i-1)/(n-k+1))} = 1 - (2-\zeta)i/(2n(1-\zeta)) + o(1/n)$, and $\delta^{1/k} = e^{-\log(1/\delta)/k} = 1 - \log(1/\delta)/(\zeta n) + o(1/n)$; above, all $o(1/n)$ terms are uniform over i. Then

$$b_i^{\mathrm{s}} \le \frac{\log(1/\delta)}{\zeta n} + \frac{2-\zeta}{2(1-\zeta)}\frac{i}{n} + o\left(\frac{1}{n}\right),$$

and for any i,

$$\frac{h^{\mathrm{s}}(i/(n+1))}{i/(n+1)} \le \frac{\log(1/\delta)}{\zeta} + \frac{2-\zeta}{2(1-\zeta)} + o\left(\frac{1}{n}\right).$$

Thus, the power cost of the Simes adjustment does not grow with n, which is more appealing compared to the asymptotic adjustment in the worst case. However, (17) indicates the cost is never negligible even if $R_{\text{marg}}(\alpha)$ is large, consistent with the behavior of the Simes adjustment observed in Section 3.6.1 for the case of a single hypothesis tested without multiplicity corrections. Thus, the asymptotic adjustment (and the substantially similar Monte Carlo approach) can be expected to be more powerful in practical applications involving FDR control, as long as a reasonably large number of discoveries is expected.

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 below 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 3.2 also yields a simultaneous upper confidence bound for the CDF.

PROPOSITION 4.1 (Simultaneous confidence bounds for the FPR). *Let F denote the true CDF of some distribution from which n i.i.d. samples, $Z_1, \dots, Z_n$, are drawn, and denote by $\hat{F}_n$ the corresponding empirical CDF. With the same notation as in Theorem 3.2,*

$$\mathbb{P}\big[F(z) \le h\big(\hat{F}_n(z)\big), \forall z \in \mathbb{R}\big] \ge 1 - \delta. \tag{20}$$

Applying Proposition 4.1 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 $\text{FPR}(t) := \mathbb{P}[\hat{s}(X_{2n+1}) \le t]$, as a function of the detection threshold t. In other 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 6 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.

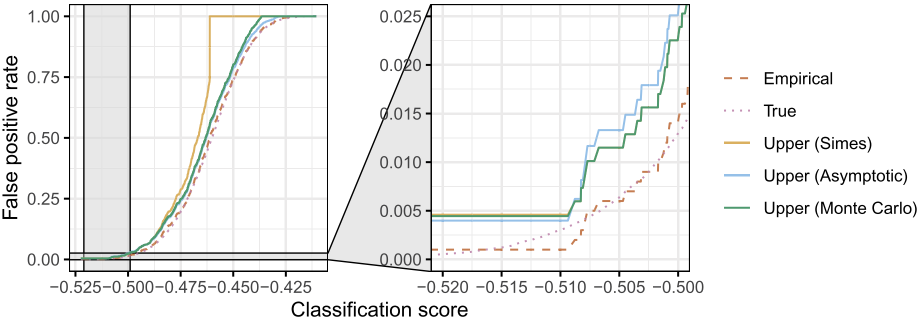

FIG. 6. *FPR calibration curves obtained with different adjustment methods for an isolation forest one-class classifier on simulated data, as a function of the reporting threshold for the classification scores. Each upper bound (solid) is guaranteed to lie above the true FPR curve (dotted) with probability 90%. The dashed 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 [21, 54] implies the bound in (20) 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 repurposed to strengthen the marginal guarantees generally obtainable for conformal predictions. In particular, for each $\alpha \in (0, 1)$, one can define a predictive set

$$\hat{\mathcal{C}}^{\alpha} := \{x : \hat{u}^{(\mathrm{ccv})}(x) > \alpha\}. \tag{21}$$

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

$$\mathbb{P}[\mathbb{P}[X_{2n+1} \in \hat{\mathcal{C}}^{\alpha} \mid \mathcal{D}] \geq 1 - \alpha \text{ for all } \alpha \in (0, 1)] \geq 1 - \delta. \tag{22}$$

In other words, if we use CCV p-values to construct prediction sets, the probability that a new test observation falls within $\hat{\mathcal{C}}^{\alpha}$ is at least $1 - \alpha$, simultaneously for all $\alpha \in (0, 1)$ with high probability over the random calibration data in $\mathcal{D}$. This is stronger than the usual conformal guarantee, as the latter holds marginally over $\mathcal{D}$ and only for a single prespecified 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 $\mathcal{D}_{j,l}^{\mathrm{test}}$ (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 ($|\mathcal{D}_{j,l}^{\mathrm{test}}| = n_{\mathrm{test}}$). Imagine that, from the practitioner's point of view, $\mathcal{D}_j$ is fixed but the test set is random, so that $\mathcal{D}_{j,l}^{\mathrm{test}}$ 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 in other measures of type-I errors) conditional on $\mathcal{D}_j$, that is, in the random variable

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

where $\mathrm{FDP}(\mathcal{D}^{\mathrm{test}}; \mathcal{D}_j)$ 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

$$\begin{aligned} \widehat{\mathrm{cFDR}}(\mathcal{D}_j) &:= \frac{1}{L}\sum_{l=1}^{L} \mathrm{FDP}(\mathcal{D}_{j,l}^{\mathrm{test}}; \mathcal{D}_j), \\ \widehat{\mathrm{cPower}}(\mathcal{D}_j) &:= \frac{1}{L}\sum_{l=1}^{L} \mathrm{Power}(\mathcal{D}_{j,l}^{\mathrm{test}}; \mathcal{D}_j), \end{aligned} \tag{23}$$

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

Our experiments will demonstrate simultaneous calibration leads to sufficiently small $\widehat{\mathrm{cFDR}}(\mathcal{D}_j)$ for the desired fraction of practitioners, while the traditional pointwise calibration 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 calibration methods 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` [59]), 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 $\mathcal{D}_{j,l}^{\text{test}}$ 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 $\mathcal{D}_{j,l}^{\text{test}}$ are sampled from P_X^a with $a = 1$, and we seek the remaining 10% of outliers. For this purpose, we calibrate a conformal p-value for all observations in $\mathcal{D}_{j,l}^{\text{test}}$, and then we apply the BH 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 7 shows the distribution of $\widehat{\text{cFDR}}(\mathcal{D}_s)$ and $\widehat{\text{cPower}}(\mathcal{D}_s)$, corresponding to $\alpha = 0.1$, for different values of the signal strength a (recall that $a = 1$ corresponds to no signal), when the BH 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

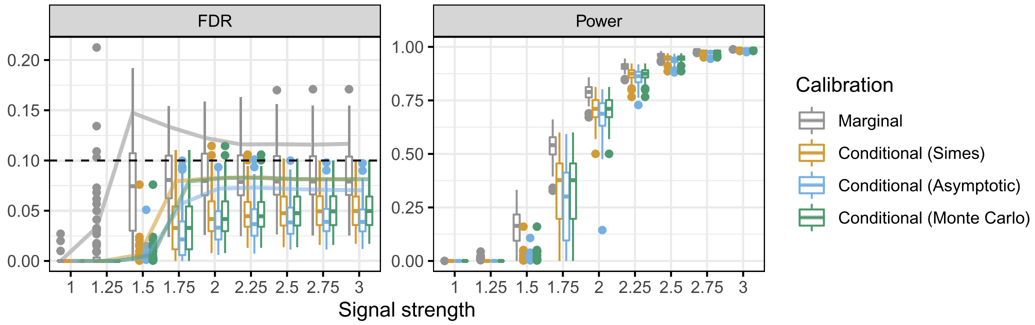

FIG. 7. *Performance of different methods for calibrating conformal p-values in a simulated outlier detection problem, as a function of the signal strength. The box plots visualize the distribution of FDR and power, as defined in* (23), *conditional on* 100 *independent data sets. The solid curves indicate the* 90*th quantile of the conditional FDR distribution. The nominal FDR* 0.1, *and the conditional method is applied with* $\delta = 0.1$.

FDR if the number of samples in the calibration data set is very large; see Figure S5 in Section S4 of the Supplementary Material [7]. Among the three conditional calibration alternatives considered here, the Monte Carlo and Simes methods yield slightly higher power than the asymptotic approximation. All methods control the marginal FDR, as predicted by our theoretical results. Figure S7 presents the results obtained by applying Storeys' correction to the BH procedure, while Figure S8 summarizes additional experiments in which the conditional calibration is applied with $\delta = 0.25$. Finally, Figure S9 visualizes the effect of different values of the k on the conditional p-values calibrated with the Simes method, showing that $k = n_{\text{cal}}/2$ works relatively well, although the performance does not appear to be extremely sensitive to this choice.

5.2.3. *Batch outlier detection.* We now consider the global testing problem of detecting whether a batch of new observations contains any outliers. We follow the same approach as before, but the $n_{\text{test}} = 1000$ observations in each test set are now subdivided 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 BH procedure with Storey's correction is applied to control the FDR over all batches. By design, 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 $a = 1.75$). Of course, batched testing is less informative than the precise identification of outliers, but the advantage now is that we can achieve higher power. Figure 8 shows that, even though this problem is relatively easy (the power is almost 1), marginal p-values may lead to a conditional FDR that is higher than expected for many researchers. By contrast, simultaneous calibration is conservative for the vast majority of them, without much power loss. Among the conditional adjustments, the Monte Carlo method and the asymptotic approximation yield higher power in this setting.

Next, we study the effect of the batch size, 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. The experiment is repeated for 100 independent data sets and 1000 test sets. Figure 9 shows that marginal p-values do not lead to valid inferences, especially if the batch size is large. By contrast, the tests based on calibration-conditional p-values always remain valid.

Finally, Figure S11 compares the performances of alternative global testing methods for combining the p-values in each batch, in the same experiments as in Figure 8. The combinations considered are the harmonic mean with equal weights [93], Simes' [69] and Stouffer's

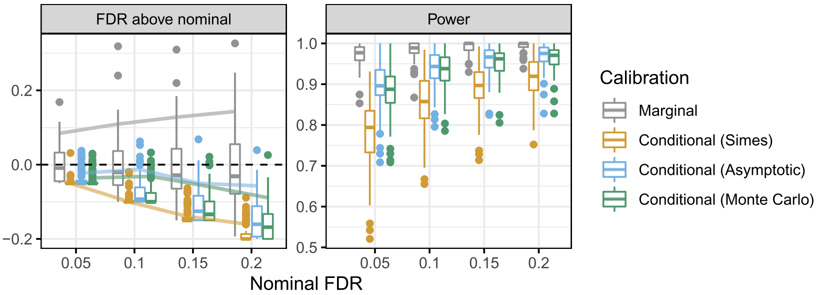

FIG. 8. *Performance of different methods for calibrating conformal p-values in a simulated outlier batch detection problem, as a function of the nominal FDR level. The excess FDR is defined as the difference between the empirical FDR and the nominal FDR. Other details are as in Figure 7.*

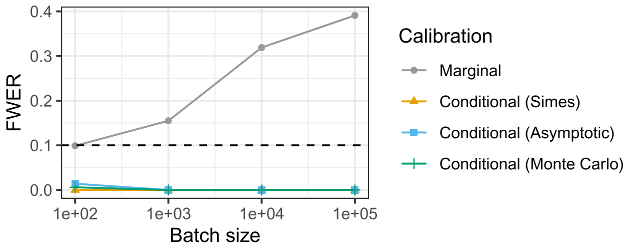

FIG. 9. *Familywise error rate* (*FWER*) *in a simulated outlier batch detection problem under the global null hypothesis*, *using different calibration methods for the conformal p-values*. *The results are shown as a function of the batch size*. *The global null is rejected if the Fisher's combined p-value is below* 0.1, *which means the nominal FWER is* 10% (*horizontal dashed line*).

[73] p-values. The harmonic mean and Simes' p-values yield no discoveries, as those methods are designed to have power if the signals are few and strong (e.g., a single outlier per batch), while each nonnull batch here contains several outliers and marginal conformal p-values must be above $1/(n+1)$. Fisher's p-values appear to be more powerful than Stouffer's in these experiments, even if the former are simultaneously adjusted with our Monte Carlo method and the latter are not. Unlike Fisher's combination test, not all global testing methods may become invalid on average when applied to positively dependent p-values. For example, the harmonic mean [93] and Simes's p-values are robust to positive dependencies [67], and Stouffer's combination p-value can also be modified to account for known dependencies [74]. Yet our p-value adjustments remain useful even with those combination tests because they enable valid inferences conditional on the calibration data; see Figure S11.

5.3. *Outlier detection on real data.*

5.3.1. *Data description.* 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. The conditional p-values are calibrated with $\delta = 0.1$ using the Monte Carlo method, which is valid in finite samples and have demonstrated in the previous sections to be more powerful than the alternatives. We utilize an isolation forest [51] machine-learning algorithms $\hat{s}$ as the base method for detecting anomalies, available in the Python `sklearn` package. We rely on the default hyperparameters, 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 Section S4 of the Supplementary Material (Tables S2–S3).

5.3.2. *Individual outlier detection.* Here, we follow the 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

TABLE 1
Summary of the data sets for outlier detection utilized in our applications

	ALOI [14, 97]	Cover [98]	Credit card [99]	KDDCup99 [14, 100]	Mammography [101]	Digits [102]	Shuttle [103]
Features d	27	10	30	40	6	16	9
Inliers n_{inliers}	283,301	286,048	284,315	47,913	10,923	6714	45,586
Outliers n_{outliers}	1508	2747	492	200	260	156	3511

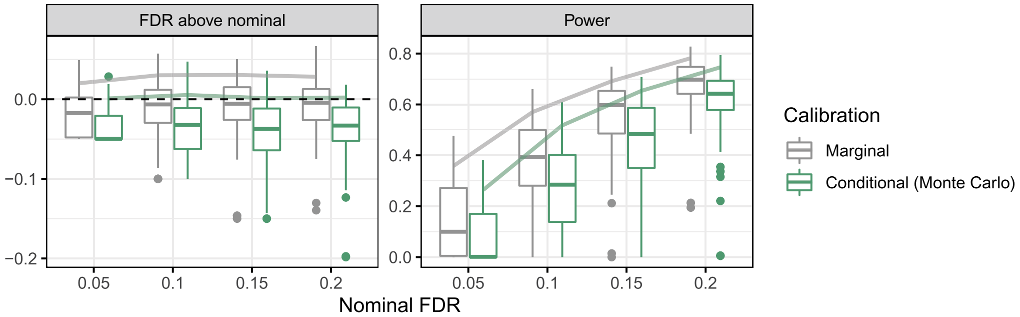

FIG. 10. *Outlier detection performance on credit card fraud data. The Benjamini–Hochberg procedure with Storey's correction is applied to conformal p-values, based on an isolation forest model and calibrated with different methods. The results are shown as a function of the nominal FDR level. Other details are as in Figure 7.*

n_{test}, respectively. In total, $n_{\text{inlier}}/2$ data points are used for training and calibration, that is, $n_{\text{train}} + n_{\text{cal}} = n_{\text{inlier}}/2$ 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 containing 90% of randomly chosen inliers, and 10% of outliers. In contrast to the experiments of Section 5.2.2 in which the data were effectively infinitely abundant, here we have some overlap between the samples in different test sets.

Figure 10 compares the performance of marginal and simultaneously calibrated p-values on the credit card data [99], as a function of the nominal FDR level. Here, the BH procedure is applied with Storey's correction. The Monte Carlo simultaneous calibration leads to FDR control for at least 90% of simulated practitioners, as expected. Consistent conclusion can be drawn from Table 2, which reports on all other data sets. Additional results corresponding to different outlier detection algorithms (one-class SVM and LOF) can be found in Table S1, Section S4.2. In all cases, we adopt the `sklearn` default parameters. Finally, Table S2 summarizes the performance of different calibration and detection methods across all data sets when the BH procedure is applied without Storey's correction.

5.3.3. *Batch outlier detection.* We now focus on global testing for outlier batch detection, similarly to Section 5.2.3. The data are divided into training, calibration and test sets according to the same scheme as in Section 5.3.2, but the size of the test sets is now 1000, following as closely as possible the experimental protocol of Section 5.2.3.

TABLE 2
Outlier detection performance using alternative methods for calibrating conformal p-values. The FDR and power are defined conditional on the training and calibration data, as in Section 5.1. The nominal marginal FDR level is 0.2. *Empirical FDR values above* 0.2 *are in orange*; *those one standard deviation above it are in red*

	FDR				Power			
	Mean		90th percentile		Mean		90-th quantile	
Dataset	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.	Marg.	Cond.
ALOI	0.025	0.001	0.048	0	0	0	0	0
Cover	0.099	0.044	0.297	0.148	0.012	0.006	0.038	0.02
Credit card	0.191	0.162	0.228	0.202	0.679	0.611	0.782	0.746
KDDCup99	0.194	0.131	0.23	0.168	0.754	0.684	0.825	0.753
Mammography	0.187	0.056	0.286	0.17	0.176	0.059	0.337	0.22
Digits	0.202	0.052	0.266	0.173	0.417	0.096	0.629	0.355
Shuttle	0.196	0.163	0.228	0.198	0.981	0.98	0.984	0.983

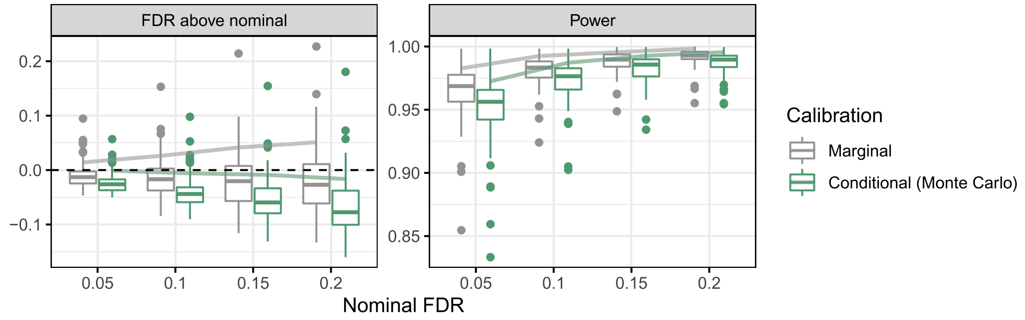

FIG. 11. *Outlier batch detection performance on credit card fraud data. Conformal p-values are computed based on an isolation forest model and calibrated using different methods. Other details are as in Figure 8.*

Figure 11 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 the BH procedure is applied with Storey's correction. Again, simultaneous calibration turns out to be needed for conditional FDR control in at least 90% of the applications, although it involves some power loss. Both calibration methods control the marginal FDR.

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 BH procedure is applied with the Storey correction. Again, the results show that the Monte Carlo method controls the conditional FDR 90% of the time, although at some cost in power, while the marginal calibration method does not. See Table S3 in Section S4.2 of the Supplementary Material for additional results that, in addition to the isolation forest, include also the one-class SVM and LOF algorithms for outlier detection. Finally, Table S4 summarizes performance of the different methods on all data sets when the BH procedure is applied without the Storey correction.

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 nonparametric 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.

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	0.07	0.016	0.2	0.081	0.001	0	0.004	0.002
Cover	0.08	0.05	0.158	0.12	0.184	0.132	0.333	0.243
Credit card	0.086	0.059	0.126	0.094	0.981	0.973	0.992	0.987
KDDCup99	0.091	0.044	0.145	0.08	1	0.998	1	1
Mammography	0.069	0.014	0.116	0.03	0.599	0.334	0.742	0.521
Digits	0.084	0.016	0.141	0.033	0.968	0.814	0.995	0.926
Shuttle	0.094	0.047	0.142	0.087	1	1	1	1

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 possible outliers are often 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 BH 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 [11, 19].

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+ [4] or bootstrap sampling [38], 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 [49]. In fact, our theory only requires the calibration and test inliers to be exchangeable and mutually independent, while the outliers in the test data may have dependencies with one another. Further, we mentioned but did not explore a connection between our multiple outlier testing problem (especially regarding our results on Fisher's combination method) and classical two-sample testing. Finally, our high-probability bounds may be useful beyond 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 [85] and functionals thereof [94], or the comparison of different machine-learning algorithms in terms of estimated generalization error [8, 33], for example.

Software availability. A Python software package implementing the methods described in this paper is available in the Supplementary Material and online at https://github.com/msesia/conditional-conformal-pvalues.git. This package also includes usage examples and notebooks to reproduce our numerical experiments.

Acknowledgments. We are grateful to the anonymous referees and Associate Editor for their helpful comments and suggestions.

Authors listed alphabetically.

Funding. S. Bates was supported by a Ric Weiland fellowship. E. Candès 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. Lei 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. Romano was supported by the Israel Science Foundation (grant no. 729/21) and by the Career Advancement Fellowship of the Technion.

SUPPLEMENTARY MATERIAL

Supplementary article (DOI: 10.1214/22-AOS2244SUPPA; .pdf). The supplementary article [7] contains the omitted mathematical proofs, as well as additional tables and figures related to numerical experiments.

Computer code (DOI: 10.1214/22-AOS2244SUPPB; .zip). A Python software package implementing the methods described in this paper is available in the Supplementary Material. This package also includes usage examples and notebooks to reproduce our numerical experiments.

REFERENCES

[1] AGGARWAL, C. C. (2015). Outlier analysis. In *Data Mining* 237–263. Springer, Berlin.

[2] AGRAWAL, S. and AGRAWAL, J. (2015). Survey on anomaly detection using data mining techniques. *Proc. Comput. Sci.* **60** 708–713.

[3] ANGELOPOULOS, A. N., BATES, S., MALIK, J. and JORDAN, M. I. (2020). Uncertainty sets for image classifiers using conformal prediction. Preprint. Available at arXiv:2009.14193.

[4] BARBER, R. F., CANDÈS, E. J., RAMDAS, A. and TIBSHIRANI, R. J. (2021). Predictive inference with the jackknife+. *Ann. Statist.* **49** 486–507. MR4206687 https://doi.org/10.1214/20-AOS1965

[5] BARBER, R. F., CANDÈS, E. J., RAMDAS, A. and TIBSHIRANI, R. J. (2021). The limits of distribution-free conditional predictive inference. *Inf. Inference* **10** 455–482. MR4270755 https://doi.org/10.1093/imaiai/iaaa017

[6] BATES, S., ANGELOPOULOS, A., LEI, L., MALIK, J. and JORDAN, M. (2021). Distribution-free, risk-controlling prediction sets. *J. ACM* **68** 43. MR4402354 https://doi.org/10.1145/3478535

[7] BATES, S., CANDÈS, E., LEI, L., ROMANO, Y. and SESIA, M. (2023). Supplement to "Testing for outliers with conformal p-values." https://doi.org/10.1214/22-AOS2244SUPPA, https://doi.org/10.1214/22-AOS2244SUPPB

[8] BAYLE, P., BAYLE, A., MACKEY, L. and JANSON, L. (2020). Cross-validation confidence intervals for test error. *Adv. Neural Inf. Process. Syst.* **33**.

[9] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

[10] BENJAMINI, Y., KRIEGER, A. M. and YEKUTIELI, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93** 491–507. MR2261438 https://doi.org/10.1093/biomet/93.3.491

[11] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

[12] BROWN, M. B. (1975). 400: A method for combining non-independent, one-sided tests of significance. *Biometrics* 987–992.

[13] CAI, F. and KOUTSOUKOS, X. (2020). Real-time out-of-distribution detection in learning-enabled cyber-physical systems. In 2020 *ACM/IEEE* 11*th International Conference on Cyber-Physical Systems* (*ICCPS*) 174–183. IEEE, Los Alamitos, CA.

[14] CAMPOS, G. O., ZIMEK, A., SANDER, J., CAMPELLO, R. J. G. B., MICENKOVÁ, B., SCHUBERT, E., ASSENT, I. and HOULE, M. E. (2016). On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Min. Knowl. Discov.* **30** 891–927. MR3513246 https://doi.org/10.1007/s10618-015-0444-8

[15] CAUCHOIS, M., GUPTA, S. and DUCHI, J. C. (2021). Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *J. Mach. Learn. Res.* **22** 81. MR4253774

[16] CERIOLI, A. (2010). Multivariate outlier detection with high-breakdown estimators. *J. Amer. Statist. Assoc.* **105** 147–156. MR2757198 https://doi.org/10.1198/jasa.2009.tm09147

[17] CHALAPATHY, R. and CHAWLA, S. (2019). Deep learning for anomaly detection: A survey. Preprint. Available at arXiv:1901.03407.

[18] CHERNOZHUKOV, V., WÜTHRICH, K. and ZHU, Y. (2021). Distributional conformal prediction. *Proc. Natl. Acad. Sci. USA* **118** e2107794118. MR4389990 https://doi.org/10.1073/pnas.2107794118

[19] CLARKE, S. and HALL, P. (2009). Robustness of multiple testing procedures against dependence. *Ann. Statist.* **37** 332–358. MR2488354 https://doi.org/10.1214/07-AOS557

[20] DEMPSTER, A. P. (1959). Generalized D_n^+ statistics. *Ann. Math. Stat.* **30** 593–597. MR0107322 https://doi.org/10.1214/aoms/1177706275

[21] DVORETZKY, A., KIEFER, J. and WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Stat.* **27** 642–669. MR0083864 https://doi.org/10.1214/aoms/1177728174

[22] EICKER, F. (1979). The asymptotic distribution of the suprema of the standardized empirical processes. *Ann. Statist.* **7** 116–138. MR0515688

[23] FEDOROVA, V., GAMMERMAN, A., NOURETDINOV, I. and VOVK, V. (2012). Plug-in martingales for testing exchangeability on-line. In *Proceedings of the* 29th *International Coference on International Conference on Machine Learning*. *ICML*'12 923–930. Omnipress, Madison, WI.

[24] FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[25] FORTUNATO, F., ANDERLUCC, L. and MONTANARI, A. (2020). One-class classification with application to forensic analysis. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 1227–1249. MR4166864 https://doi.org/10.1111/rssc.12438

[26] FRIEDMAN, J. (2004). On multivariate goodness-of-fit and two-sample testing Technical Report No. SLAC-PUB-10325. Stanford Linear Accelerator Center, Menlo Park, CA.

[27] GUAN, L. and TIBSHIRANI, R. (2022). Prediction and outlier detection in classification problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 524–546. MR4412996 https://doi.org/10.1111/rssb.12443

[28] GUPTA, C., KUCHIBHOTLA, A. K. and RAMDAS, A. K. (2021). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognit.* 108496.

[29] HAROUSH, M., FROSTIG, T., HELLER, R. and SOUDRY, D. (2021). Statistical testing for efficient out of distribution detection in deep neural networks. ArXiv preprint. Available at arXiv:2102.12967.

[30] HAWKINS, D. M. (1980). *Identification of Outliers*. *Monographs on Applied Probability and Statistics*. CRC Press, London. MR0584791

[31] HECHTLINGER, Y., PÓCZOS, B. and WASSERMAN, L. (2018). Cautious deep learning. Available at arXiv:1805.09460.

[32] HENDRYCKS, D. and GIMPEL, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proc. Int. Conf. Learn. Represent.*

[33] HOLLAND, M. J. (2020). Making learning more transparent using conformalized performance prediction. ArXiv preprint. Available at arXiv:2007.04486.

[34] HU, X. and LEI, J. (2020). A distribution-free test of covariate shift using conformal prediction. ArXiv preprint. Available at arXiv:2010.07147.

[35] ISHIMTSEV, V., BERNSTEIN, A., BURNAEV, E. and NAZAROV, I. (2017). Conformal k-NN anomaly detector for univariate data streams. In *Conformal and Probabilistic Prediction and Applications* 213–227. PMLR, Stockholm, Sweden.

[36] IZBICKI, R., SHIMIZU, G. and STERN, R. (2020). Flexible distribution-free conditional predictive bands using density estimators. In *International Conference on Artificial Intelligence and Statistics* 3068–3077. PMLR, Online.

[37] KHAN, S. S. and MADDEN, M. G. (2014). One-class classification: Taxonomy of study and review of techniques. *Knowl. Eng. Rev.* **29** 345–374.

[38] KIM, B., XU, C. and FOYGEL BARBER, R. (2020). Predictive inference is free with the jackknife+-after-bootstrap. *Adv. Neural Inf. Process. Syst.* **33**.

[39] KIM, I., RAMDAS, A., SINGH, A. and WASSERMAN, L. (2021). Classification accuracy as a proxy for two-sample testing. *Ann. Statist.* **49** 411–434. MR4206684 https://doi.org/10.1214/20-AOS1962

[40] KIVARANOVIC, D., JOHNSON, K. D. and LEEB, H. (2020). Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics* 4346–4356. PMLR, Online.

[41] KOLMOGOROV, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Inst. Ital. Attuari, Giorn.* **4** 83–91.

[42] KOST, J. T. and MCDERMOTT, M. P. (2002). Combining dependent p-values. *Statist. Probab. Lett.* **60** 183–190. MR1945440 https://doi.org/10.1016/S0167-7152(02)00310-3

[43] KRISHNAMOORTHY, K. and MATHEW, T. (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation. Wiley Series in Probability and Statistics.* Wiley, Hoboken, NJ. MR2500599 https://doi.org/10.1002/9780470473900

[44] KUCHIBHOTLA, A. K. (2020). Exchangeability, conformal prediction, and rank tests. ArXiv preprint. Available at arXiv:2005.06095.

[45] LAXHAMMAR, R. and FALKMAN, G. (2015). Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Ann. Math. Artif. Intell.* **74** 67–94. MR3353897 https://doi.org/10.1007/s10472-013-9381-7

[46] LEE, K., LEE, H., LEE, K. and SHIN, J. (2018). Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*.

[47] LEE, K., LEE, K., LEE, H. and SHIN, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*.

[48] LEI, J., RINALDO, A. and WASSERMAN, L. (2015). A conformal prediction approach to explore functional data. *Ann. Math. Artif. Intell.* **74** 29–43. MR3353895 https://doi.org/10.1007/s10472-013-9366-6

[49] LI, A. and BARBER, R. F. (2019). Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 45–74. MR3904779

[50] LIANG, S., LI, Y. and SRIKANT, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. ArXiv preprint. Available at arXiv:1706.02690.

[51] LIU, F. T., TING, K. M. and ZHOU, Z.-H. (2008). Isolation forest. In 2008 *Eighth IEEE International Conference on Data Mining* 413–422.

[52] LOPEZ-PAZ, D. and OQUAB, M. (2017). Revisiting classifier two-sample tests. In *International Conference on Learning Representations*.

[53] MARY, D. and ROQUAIN, E. (2022). Semi-supervised multiple testing. *Electron. J. Stat.* **16** 4926–4981. MR4490412 https://doi.org/10.1214/22-ejs2050

[54] MASSART, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Ann. Probab.* **18** 1269–1283. MR1062069

[55] MOYA, M. M., KOCH, M. W. and HOSTETLER, L. D. (1993). One-class classifier networks for target recognition applications. *NASA STI/Recon Technical Report N* **93** 24043.

[56] PAPADOPOULOS, H., PROEDROU, K., VOVK, V. and GAMMERMAN, A. (2002). Inductive confidence machines for regression. In *Machine Learning: ECML* 2002. *Lecture Notes in Computer Science* **2430** 345–356. Springer, Berlin. MR2050303 https://doi.org/10.1007/3-540-36755-1_29

[57] PARK, S., BASTANI, O., MATNI, N. and LEE, I. (2020). PAC confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*.

[58] PATCHA, A. and PARK, J.-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Netw.* **51** 3448–3470.

[59] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A. et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12** 2825–2830. MR2854348

[60] PIMENTEL, M. A., CLIFTON, D. A., CLIFTON, L. and TARASSENKO, L. (2014). A review of novelty detection. *Signal Process.* **99** 215–249.

[61] RAVA, B., SUN, W., JAMES, G. M. and TONG, X. (2021). A burden shared is a burden halved: A fairness-adjusted approach to classification. ArXiv preprint. Available at arXiv:2110.05720.

[62] RIANI, M., ATKINSON, A. C. and CERIOLI, A. (2009). Finding an unknown number of multivariate outliers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 447–466. MR2649609 https://doi.org/10.1111/j.1467-9868.2008.00692.x

[63] ROMANO, Y., PATTERSON, E. and CANDÈS, E. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems* 32 3543–3553.

[64] ROMANO, Y., SESIA, M. and CANDÈS, E. J. (2020). Classification with valid and adaptive coverage. *Adv. Neural Inf. Process. Syst.* **33**.

[65] SABOKROU, M., KHALOOEI, M., FATHY, M. and ADELI, E. (2018). Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3379–3388.

[66] SARKAR, S. K. (2008). Generalizing Simes' test and Hochberg's stepup procedure. *Ann. Statist.* **36** 337–363. MR2387974 https://doi.org/10.1214/009053607000000550

[67] SARKAR, S. K. and CHANG, C.-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Statist. Assoc.* **92** 1601–1608. MR1615269 https://doi.org/10.2307/2965431

[68] SESIA, M. and CANDÈS, E. J. (2020). A comparison of some conformal quantile regression methods. *Stat* **9** e261. MR4104217 https://doi.org/10.1002/sta4.261

[69] SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754. MR0897872 https://doi.org/10.1093/biomet/73.3.751

[70] SMITH, J., NOURETDINOV, I., CRADDOCK, R., OFFER, C. and GAMMERMAN, A. (2015). Conformal anomaly detection of trajectories with a multi-class hierarchy. In *International Symposium on Statistical Learning and Data Sciences* 281–290. Springer, Berlin.

[71] STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. MR1924302 https://doi.org/10.1111/1467-9868.00346

[72] STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 187–205. MR2035766 https://doi.org/10.1111/j.1467-9868.2004.00439.x

[73] STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A. and WILLIAMS JR, R. M. (1949). The American soldier: Adjustment during army life. (Studies in social psychology in world war II), vol. 1.

[74] STRUBE, M. J. (1985). Combining and comparing significance levels from nonindependent hypothesis tests. *Psychol. Bull.* **97** 334.

[75] TARASSENKO, L., CLIFTON, D. A., BANNISTER, P. R., KING, S. and KING, D. (2009). Novelty detection. In *Encyclopedia of Structural Health Monitoring* Wiley, Hoboken, NJ.

[76] TARASSENKO, L., HAYTON, P., CERNEAZ, N. and BRADY, M. (1995). Novelty detection for the identification of masses in mammograms. In 1995 *Fourth International Conference on Artificial Neural Networks* 442–447. IET, Cambridge, UK.

[77] TUKEY, J. W. (1947). Non-parametric estimation. II. Statistically equivalent blocks and tolerance regions–the continuous case. *Ann. Math. Stat.* **18** 529–539. MR0023033 https://doi.org/10.1214/aoms/1177730343

[78] VOVK, V. (2012). Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning* **25** 475–490.

[79] VOVK, V. (2015). Cross-conformal predictors. *Ann. Math. Artif. Intell.* **74** 9–28. MR3353894 https://doi.org/10.1007/s10472-013-9368-4

[80] VOVK, V. (2020). Testing for concept shift online. ArXiv preprint. Available at arXiv:2012.14246.

[81] VOVK, V. (2021). Testing randomness online. *Statist. Sci.* **36** 595–611. MR4323055 https://doi.org/10.1214/20-sts817

[82] VOVK, V., GAMMERMAN, A. and SAUNDERS, C. (1999). Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning* 444–453.

[83] VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York. MR2161220

[84] VOVK, V., NOURETDINOV, I. and GAMMERMAN, A. (2003). Testing exchangeability on-line. 768–775.

[85] VOVK, V., NOURETDINOV, I., MANOKHIN, V. and GAMMERMAN, A. (2018). Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications* 37–51. PMLR, Maastricht, The Netherlands.

[86] VOVK, V., PETEJ, I., NOURETDINOV, I., AHLBERG, E., CARLSSON, L. and GAMMERMAN, A. (2021). Retrain or not retrain: Conformal test martingales for change-point detection. In *Conformal and Probabilistic Prediction and Applications* 191–210. PMLR, Online.

[87] WALD, A. (1943). An extension of Wilks' method for setting tolerance limits. *Ann. Math. Stat.* **14** 45–55. MR0007965 https://doi.org/10.1214/aoms/1177731491

[88] WEINSTEIN, A., BARBER, R. and CANDES, E. (2017). A power and prediction analysis for knockoffs with lasso statistics. ArXiv preprint. Available at arXiv:1712.06465.

[89] WILCOXON, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in Statistics* 196–202. Springer, Berlin.

[90] WILKS, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Ann. Math. Stat.* **12** 91–96. MR0004451 https://doi.org/10.1214/aoms/1177731788

[91] WILKS, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *Ann. Math. Stat.* **13** 400–409. MR0007592 https://doi.org/10.1214/aoms/1177731537

[92] WILKS, S. S. (1963). Multivariate statistical outliers. *Sankhya, Ser. A* **25** 407–426. MR0173334

[93] WILSON, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proc. Natl. Acad. Sci. USA* **116** 1195–1200. MR3904688 https://doi.org/10.1073/pnas.1814092116

[94] WISNIEWSKI, W., LINDSAY, D. and LINDSAY, S. (2020). Application of conformal prediction interval estimations to market makers' net positions. In *Conformal and Probabilistic Prediction and Applications* 285–301. PMLR, Online.

[95] YANG, C.-Y., LEI, L., HO, N. and FITHIAN, W. (2021). BONuS: Multiple multivariate testing with a data-adaptive test statistic. ArXiv preprint. Available at arXiv:2106.15743.

[96] ZHANG, Y. and POLITIS, D. N. (2022). Bootstrap prediction intervals with asymptotic conditional validity and unconditional guarantees. *Inf. Inference*.

[97] 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.

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

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

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

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

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

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