Outlier detection under False Omission Rate control

Adam Wawrzeńczyk[1][0000−0002−6202−7829]
and Jan Mielniczuk[1,2][0000−0003−2621−2303]

[1] Institute of Computer Science, Polish Academy of Sciences
Warsaw, Jana Kazimierza 5, 01-248, Poland
[2] Faculty of Mathematics and Information Science, Warsaw University of Technology
Warsaw, Koszykowa 75, 00-662 Poland
{adam.wawrzenczyk,jan.mielniczuk}@ipipan.waw.pl

Abstract. We argue that in many practical situations control of False Omission Rate (FOR) or Bayesian False Omission Rate (BFOR) is of primary importance. We develop and investigate such rule in the context of outlier detection, and propose its empirical formulation for practical use. We consider several score statistics used to detect outliers and study how well the introduced method controls FOR in practice. It is shown by analysis of several datasets that FOR control in contrast to FDR control is inherently tied to performance of the score statistic employed on both inlier and outlier data sets.

Keywords: False Omission Rate · Bayesian False Omission Rate · False Discovery Rate · outlier detection · fraud detection · one class classification · truncation rule.

1 Introduction

We consider the situation when a score statistic is learned on a random sample of regular observations (inliers) and used to detect out of distribution observations (outliers) with an objective to control the percentage of undetected outliers among the observations classified as inliers. Such a need arises in many practical situations: imagine a scrutiny of possibly fraudulent transactions for which one would like to detect all but a very small percent of frauds – such an approach accounts for the fact that trying to detect all frauds will require very stringent safety rules which would deter potential customers. Another example is development of a new test for a contagious disease (e.g. COVID-19), for which is vital to ensure that randomly chosen person will not pass it *if infected* with large probability (see [14]). In such situations it is much more important to control False Omission Rate (FOR, called also False Non-discovery Rate (FNR) [6]) than commonly used False Discovery Rate (FDR). FOR is defined as the expected value of False Omission Proportion i.e. proportion of undetected outliers among observations classified as inliers, whereas FDR is the expected proportion of inliers among observations deemed outliers. Obviously, FDR control in many situations

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_47

has evident advantages but we argue that in numerous cases FOR control – or its Bayesian analogue defined below – is of main interest, and procedures which ensure it are worth studying.

Our main objective here is to develop a rule which approximately controls FOR and to investigate its properties both theoretically and by means of analysis of real data sets. The rule developed here is derived analogously to Benjamini-Hochberg rule [2] using Frequentist Bayes approach (see e.g. [6], chapter 4). We also consider several methods scores for outlier detection and check how their choice influences control of FOR. Finally, we investigate ways of diminishing intrinsic variability of p-values due to the random split of the data set.

2 FOR control procedure

2.1 Preliminaries

Consider checking whether observations under study are outliers with the use of a specified score statistic $\hat{s}$ to test a null hypothesis $H_{0,i}$: i-th observation X_i is an inlier versus an alternative $H_{1,i}$: i-th observation is an outlier. We adopt throughout the convention that large values of $\hat{s}$ indicate outliers. It is known that when the cumulative distribution function (CDF) denoted by F of a test statistic is continuous, the distribution of the corresponding p-value equal to $1 - F(X_i)$, provided the null hypothesis is true, is uniform on $[0, 1]$. This is the fundamental property used to bound Family Wise Error Rate (FWER) defined as probability of falsely rejecting at least one null, or False Discovery Rate (FDR) defined below, which is more easily controlled. For a discussion of numerous solutions to the problem from which Benjamini-Hochberg (BH) procedure is the most commonly used, see e.g. [5]). In [7] analysis of behavior of FOR for BH procedure is given. However, construction of rules controlling FOR remains, to the best of our knowledge, largely untreated. Imagine now that we have a sample of n observations generated by mixture of distribution of inliers (occurring with probability π) and outliers (occurring with probability $1 - \pi$), and denote by $p_1, \ldots, p_n$ corresponding p-values of a test under consideration. Then we can write

$$p_i \sim \pi U + (1 - \pi) F_1, \quad i = 1, \ldots, n, \tag{1}$$

where U stands for the distribution function of the uniform distribution $U[0, 1]$: $U(t) = t$, F_1 is the cumulative distribution function of the p-values for outliers and „$\sim$" denotes „is distributed as". In the following we assume that mixing proportion π is known. This assumption is commonly met i.e. when prevalence of a certain disease can be precisely estimated based on independent data base.

We assume that n_0 observations are inliers (nulls) and $n_1 = n - n_0$ are outliers (non-nulls) and note that due to our mixture assumption (1) n_0 and n_1 are random and have Bernoulli distribution: $n_0 \sim Bin(n, \pi)$ and $n_1 \sim Bin(n, 1 - \pi)$. Consider a specific decision rule assigning each of n observations to inliers or outliers and denote by R the number of of rejected null hypotheses, by V the number of falsely rejected nulls and by Z the number of falsely not rejected

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_47

alternatives. Note that $Z = n_1 - (R - V)$ and let NR be the number of not rejected items. We will consider threshold rules such that for any $p_i \leq u$ the corresponding null hypothesis $H_{0,i}$ is rejected i.e. i-th element is considered an outlier. Threshold u is assumed here to be a fixed, predetermined point. We will write $NR(u)$ for NR to underline the dependence on u. Let

$$\begin{aligned} \mathrm{FOR} &= \mathbb{E}\left(\frac{Z}{NR(u)}\ \mathbb{I}\left\{NR(u) > 0\right\}\right), \\ \overline{\mathrm{FOR}} &= \mathbb{E}\left(\frac{NR(u) - n\pi(1-u)}{NR(u)}\ \mathbb{I}\left\{NR(u) > 0\right\}\right). \end{aligned} \tag{2}$$

FOR stands for False Omission Rate and $\overline{\mathrm{FOR}}$ is an estimable approximation of FOR obtained by replacing number of non-rejected nulls at the threshold u by its expected value $n\pi(1-u)$. Our aim is to construct a decision rule which approximately controls FOR i.e. such that for any given $\alpha \in (0,1)$ the inequality $\mathrm{FOR} \leq \alpha$ is satisfied.

In the traditional setting one aims at controlling False Discovery Rate (FDR) at the level α, where FDR is defined as

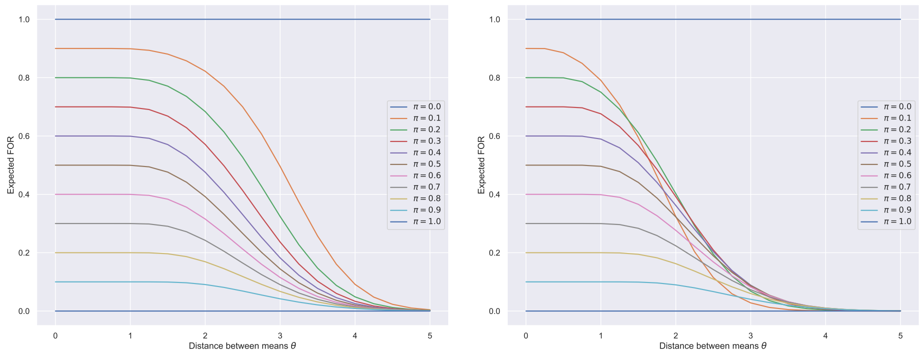
$$\mathrm{FDR} = \mathbb{E}\left(\frac{V}{R}\ \mathbb{I}\left\{R > 0\right\}\right). \tag{3}$$

We note that although it might appear at the first sight that controlling FOR defined in (2) is analogous to controlling FDR, this is not the case as the roles of inliers and outliers are not exchangeable. The difference is due to differences in distributions of p-values for false signals and false non-signals. Namely, we assume that the distribution of inliers'score $\hat{s}$ is known, and it follows that for a threshold u, the distribution of the p-value corresponding to false signal, i.e. inlier smaller than u is given by the uniform distribution on $[0, u]$, whereas for the false non-signal it pertains to unknown distribution F_1 and equals $F_1(s)/(1 - F_1(u))$ for $s \in [u, 1]$. As F_1 is unknown, in contrast to known (i.e. uniform) distribution of p-values for the inliers, the problem of control of FOR is considerably harder than the control of FDR. We would like the distribution of F_1 to be concentrated close to 0, but this may vary depending on the quality of the score function in general and its performance on the studied dataset in particular (see Figure 5).

We also note that similarly to testing (where decrease of level of significance leads to smaller values of power), when FDR is controlled at the level α, FOR is uncontrolled and can attain any level less than proportion of outliers $1 - \pi$.

Example 1. Assume that distribution of the score statistic $\hat{s}$ for inliers is given by the standard normal distribution $N(0,1)$ and outliers by $N(\theta, 1)$, where $\theta > 0$. We reject the null for large values of s. Then straightforward calculation show that the distribution function of p-value $(1 - \Phi)(s)$ for an outlier is given by $F_1(s) = 1 - \Phi\left(-\Phi^{-1}(s) - \theta\right) = \Phi\left(\Phi^{-1}(s) + \theta\right)$, where Φ is CDF of $N(0,1)$. Using the formula (8) below, the values of FOR at threshold u^*_{FDR} corresponding to Benjamini-Hochberg procedure [2] or its modified version with Storey's correction [13] can be calculated, and are shown in Figure 1. The figure shows that

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_47



(a)

(b)

Fig. 1: Values of FOR when FDR is controlled against the mean distance θ between inliers and outliers (see text) (a) FDR $\leq \pi\alpha$ (Benjamini-Hochberg procedure) (b) FDR $\leq \alpha$ (Benjamini-Hochberg procedure with Storey's correction, [13]), $\alpha = 0.05$.

for small θ (when the inliers and outliers become less separated and threshold u^*_{FDR} becomes smaller), the value of FOR gets larger and approaches proportion $1 - \pi$ of inliers in the mixture.

We also introduce Bayesian False Omission Rate (BFOR)

$$\mathrm{BFOR} = \frac{\mathbb{E}\, Z}{\mathbb{E}(NR(u))}, \tag{4}$$

following the analogous treatment of False Discovery Rates (see e.g. [7] and [6], section 2.2). Efron [6] argues that from Bayesian view point, control of Bayesian False Discovery Rate – the quantity defined analogously to Eq. (4), but for false discoveries – is of main interest.

2.2 Control of FOR: theoretical results

We prove in Theorem 1 below that all introduced quantities are approximately equal for large sample sizes, namely

$$\mathrm{FOR} \approx \mathrm{BFOR}, \qquad \overline{\mathrm{FOR}} \approx \mathrm{BFOR}.$$

Let $G(t) = \pi U(t) + (1-\pi)F_1(t)$ be a mixture distribution of p-values. We have

Theorem 1. *(i) Assume that considered decision rule rejects all null hypotheses with corresponding p-values smaller or equal u such that* $0 < G(u) < 1$. *Then*

$$FOR = BFOR \times (1 - (1 - G(u))^n) < BFOR.$$

(ii) For the decision rule defined as in (i) we have

$$\overline{FOR} \leq BFOR + o\left(\frac{1}{n}\right).$$

Proof. (i) We will use shorthand $p \in O$ („p" standing for p-value and „O" standing for „Outliers") meaning that p-value corresponds to an outlier. The proof follows by noting that $P(p \in O|p > u)$ equals

$$\frac{P(p > u|p \in O)P(p \in O)}{P(p > u)} = \frac{(1 - F_1(u))(1 - \pi)}{1 - G(u)} = \text{BFOR}. \tag{5}$$

Denote BFOR $= \gamma(u)$. Thus, given $NR(u)$,

$$Z|NR(u) \sim Bin(NR(u), \gamma(u)),$$

For $NR \neq 0$, using the formula for the expected value of the binomial, we have that

$$\mathbb{E}\left(\frac{Z}{NR(u)}\middle| NR(u)\right) = \frac{\mathbb{E}(Z|NR(u))}{NR(u)} = \frac{NR(u)\gamma(u)}{NR(u)} = \gamma(u) = \text{BFOR}$$

and thus

$$\begin{aligned} \text{FOR} &= \sum_{i>0} \mathbb{E}\left(\frac{Z}{NR(u)}\middle| NR(u) = i\right) P(NR(u) = i) \\ &= \sum_{i>0} \gamma(u) P(NR(u) = i) = \gamma(u) P(NR(u) \neq 0), \end{aligned}$$

which implies (i) as $NR(u) \sim Bin(n, 1 - G(u))$.

(ii) Observe that for $X \sim Bin(n, p)$ we have

$$\begin{aligned} \mathbb{E}\left(\frac{1}{X}\mathbb{I}\{X > 0\}\right) &\geq \mathbb{E}\left(\frac{1}{X+1}\right) - P(X = 0) \\ &= \frac{1}{p(n+1)}[1 - (1-p)^{n+1}] - (1-p)^n, \end{aligned} \tag{6}$$

where the inequality follows from $\mathbb{E}\,\frac{1}{X+1}\,\mathbb{I}\{X = 0\} = P(X = 0)$ and the final equality above is proved in [4].

DOI: 10.1007/978-3-031-36024-4_47

Thus for $X = NR(u) \sim Bin(n, 1 - G(u))$:

$$\begin{aligned}
&\mathbb{E}\left(\frac{NR(u) - n\pi(1-u)}{NR(u)}\mathbb{I}\{NR(u) > 0\}\right)\\
&\quad = P(NR(u) > 0) - n\pi(1-u) \times \mathbb{E}\left(\frac{1}{NR(u)}\mathbb{I}\{NR(u) > 0\}\right)\\
&\quad \leq 1 - G(u)^n - n\pi(1-u)\times\\
&\qquad \times\left[\frac{1}{(n+1)(1-G(u))}(1 - G(u)^{n+1}) - G(u)^n\right] \qquad (7)\\
&\quad =1 - \frac{n\pi(1-u)}{(n+1)(1-G(u))} - G(u)^n(1 - n\pi(1-u) -\\
&\qquad - G(u)n\pi(1-u))\\
&\quad =1 - \frac{\pi(1-u)}{1-G(u)} + o\left(\frac{1}{n}\right) = \text{BFOR} + o\left(\frac{1}{n}\right),
\end{aligned}$$

where inequality follows from (6). The two first terms in penultimate expression above are equal to BFOR $+ o(n^{-1})$ due to $n^{-1} - (n+1)^{-1} = (n(n+1))^{-1} = o(n^{-1})$ and all remaining terms are also $o(n^{-1})$. ■

We note that it follows from the proof that both $Z \sim Bin\,(n, (1-\pi)(1-F_1(u))$ and $NR(u) \sim Bin(n, 1 - G(u))$ are binomially distributed and thus we have

$$\frac{\mathbb{E}(Z)}{\mathbb{E}(NR(u))} = \frac{(1-\pi)(1-F_1(u))}{1-G(u)} = \frac{1 - G(u) - \pi(1-u)}{1-G(u)} = P(p \in O|p > u). \qquad (8)$$

Note that we assume in Theorem 1 that threshold u does not depend on data. We conjecture that in a general case, when threshold will be data-dependent, FOR, BFOR and $\overline{\text{FOR}}$ are also approximately equivalent.

Replacing FOR in the condition FOR $= \alpha$ by its approximation BFOR one obtains the following equality

$$\frac{1 - G(u) - \pi(1-u)}{1 - G(u)} = \alpha, \qquad (9)$$

or equivalently

$$1 - G(u) = \frac{\pi}{1-\alpha} \times (1-u). \qquad (10)$$

Theorem 2. *Solution $u^* \in (0, 1)$ of (10) exists and is unique provided that (i) $G(\cdot)$ is strictly concave and (ii) $G'(1) \geq \pi/(1-\alpha)$.*

Proof. Indeed, the condition (ii) is equivalent to the condition that the derivative of $1-G(u)$ at 1 is not larger than the derivative of the line $(\pi/1-\alpha)\times(1-u)$ at 1. As $1-G(u)$ is strictly convex it is enough to check that $1-G(0) = 1 \geq \pi/(1-\alpha)$. But this follows from (ii) since $1 > G'(1)$ as density $g(s) = G'(s)$ is strictly decreasing in view of strict concavity and $\int_0^1 g(s)\,ds = 1$. Uniqueness of the solution is due to the strict concavity of G. ■

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_47

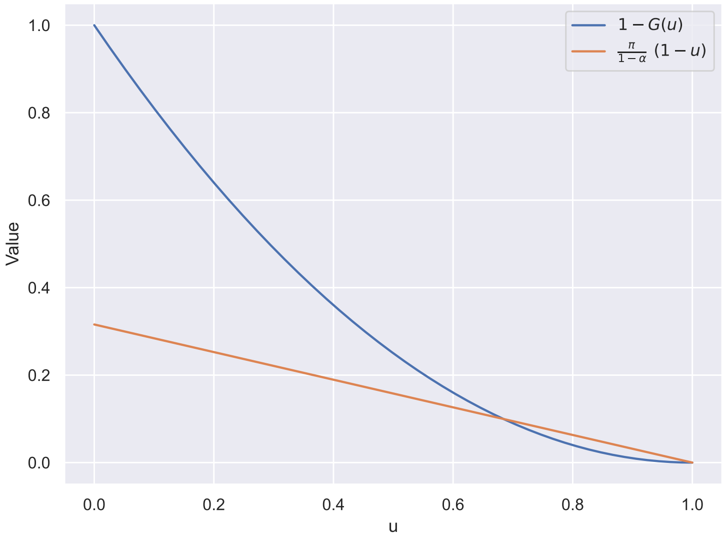


Fig. 2: Illustration of equation (10). Convex curve $1 - G(u)$ starts at 0 above the value of the line $(\pi/(1 - \alpha) \times (1 - u)$ and intersects it at a point u^*.

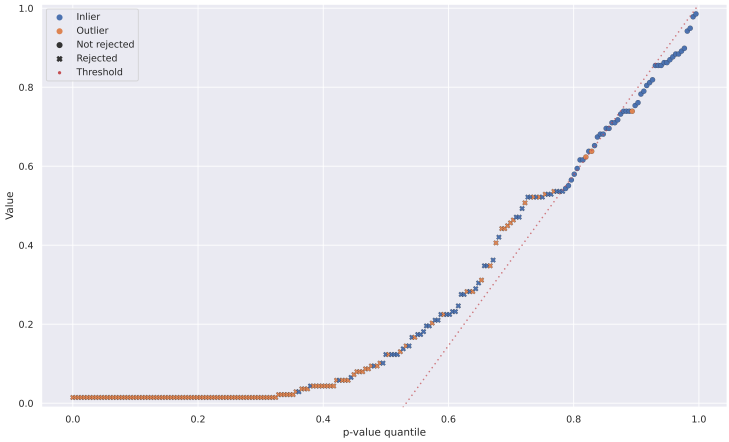


Fig. 3: Illustration of rule (11). The index of the smallest p-value which down-crosses the line $1 - (1 - \alpha)/\pi \times (1 - u)$ corresponds to the threshold in (11).

DOI: 10.1007/978-3-031-36024-4_47

Theoretical solution is shown in Figure 1. Thus we know that the truncation level u^* for such that BFOR $= \mathbb{E}(Z)/\mathbb{E}(NR(u^*)) = \alpha$ exists under above conditions. Note that the assumption that $G(\cdot)$ is strictly concave (or, equivalently, that $g(\cdot)$ is strictly decreasing) is natural in the considered context. Namely, it implies in the view of (1) that density f_1 of p-value distribution F_1 for outliers is strictly decreasing, and, consequently, it is more likely to obtain smaller p-values for outliers than larger ones.

2.3 FOR control: empirical rule

Now we consider solution to the empirical counterpart of (10). Note that due to (8) BFOR is easily estimated, and we obtain the following rule: for a given $\alpha \in (0,1)$ find p-value $p_{(i^*)}$ such that

$$p_{(i^*)} = \begin{cases} \min_{p_{(i)}} B & \text{if } B \neq \emptyset; \\ 1 & \text{otherwise,} \end{cases}$$
$$\text{where } B = \left\{ p_{(i)} : p_{(i)} \leq 1 - \left(1 - \frac{i}{n}\right) \frac{1-\alpha}{\pi} \right\} \tag{11}$$

and „accept" (treat as inliers) all p-values strictly larger than $p_{(i^*)}$, where $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$. This follows by plugging in empirical distribution $G_n(t) = \#\{i : p_i \leq t\}/n$ of all p-values for G in (10) and noting that $G_n\left(p_{(i)}\right) = \frac{i}{n}$. Note also that the threshold in (11) equals 1 for $i = n$ and is approximately equal to $1 - (1-\alpha)/\pi \leq 0$ (implied by condition (ii) of Theorem 1) for $i = 1$. Thus i^* is an index of the ordered p-value corresponding to the first moment when ordered p-values down-cross (cross from above to below) the line $1 - (1-u) \times (1-\alpha)/\pi$. This empirical rule is analogous (in a symmetric way) to Benjamini-Hochberg threshold construction for FDR control: starting from the largest p-values (as those are of interest when controlling not-rejected examples; this is an mirror image of Benjamini-Hochberg procedure starting from the smallest p-values) we look for the last (i.e. the smallest) index where FOR is still controlled, and use it as a threshold separating inliers from outliers.

2.4 Construction of *p*-values

We now discuss the framework in which p-values appearing in (11) are defined (*Multisplit* procedure). Note that as we do not know CDF of score statistic $\hat{s}$ for inliers, we can not compute p-value directly as $1 - F(X)$ and F needs to be estimated. We thus consider a sample $\mathcal{D} = \{X_1, \ldots, X_{2n}\}$ of size $2n$ consisting of inliers which will be split into training $\mathcal{D}^{train}$ and calibration $\mathcal{D}^{cal}$ samples consisting of n observation each. Moreover, let $\hat{s}$ will be a real-valued score statistic constructed to distinguish inliers from outliers. We adopt the convention that large values of $\hat{s}$ indicate a possible outlier. We consider the empirical distribution of $\hat{s}(X_i)$ for $X_i \in \mathcal{D}^{cal}$ as approximation of F and define p-value $\hat{p} = \hat{p}(X)$ of X as

$$\hat{p} = \frac{\#\left\{X_i \in \mathcal{D}^{cal} : \hat{s}(X_i) \geq \hat{s}(X)\right\} + 1}{n+1}. \tag{12}$$

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_47

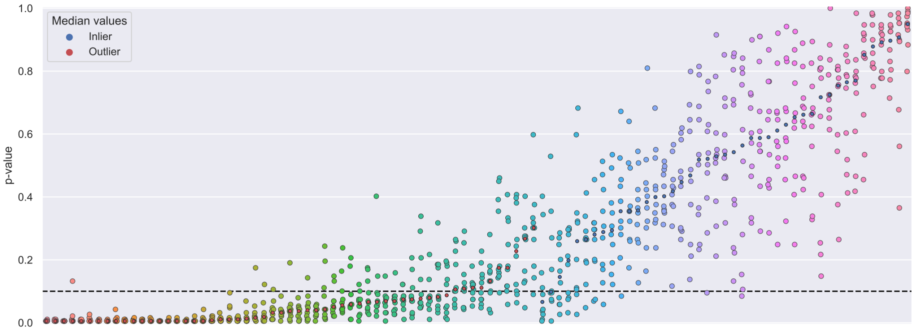


Fig. 4: p-value lottery for A^3 classifier on 100 first examples from *Tic-tac-toe* dataset; examples are sorted according to their class and median p-value. p-values for 10 random training-calibration splits are very unstable – median range width is 0.31, and maximal difference between minimal and maximal p-value for one of the examples exceeded 0.83. FOR control based a single split would not be reliable.

Consider the sample $S_1, \dots, S_{n+1}$ consisting of observations $\hat{s}(X_i)$ for $X_i \in \mathcal{D}^{cal}$ augmented by $S = \hat{s}(X)$. When S corresponds to an inlier, observations $S_1, \dots, S_{n+1}$ are equi-distributed and it follows that for continuous $\hat{s}(X)$, $\hat{p}(X)$ is uniformly distributed on $\{1/(n+1), \dots, n/(n+1), 1\}$ given $\mathcal{D}^{cal}$, and thus $P(\hat{p}(X) \leq t|\mathcal{D}^{cal}) \leq t$ (see e.g. [1]) – this means that distribution of $p(X)$ is super-uniform.

As the definition above depends on the training-calibration split we initially considered several versions of $\hat{p}$:

- p_{single}: one-split version defined above,
- p_{med}: median of $p_1, \dots, p_k$ when $p_1, \dots, p_k$ are p-values based on k random splits,
- $p_{2med} = 2 \times p_{med}$.

Definitions of p_{med} and p_{2med} are based on analogous proposals in variable selection and their purpose is to decrease variability incurred due to the random split; the phenomenon named p-value lottery (see [11]). Its occurrence is confirmed by Figure 4 which shows substantial variability of p-values depending on a random split for A^3 classifier. The distribution of $p_{single}(X)$ is super-uniform and the the same is also true for p_{2med}, with the proof being analogous to that of Theorem 11.1 in [3]. p-value p_{med} is also considered as in practice p_{2med} is too conservative and thus inflates FDR in consequence. As our experiments confirm this, we focus on p_{med} only in the following.

3 Experimental setting

We tested the proposed FOR control procedure as described in Section 2.3. We consider four different score functions to obtain the outlier scores:

DOI: 10.1007/978-3-031-36024-4_47

Table 1: Dataset summary

Dataset	**Samples**	**Features**	**Inlier rate π**	**Dataset**	**Samples**	**Features**	**Inlier rate π**
Abalone	4177	8	0.34	*Madelon*	2600	500	0.50
Arrhythmia	452	279	0.54	*Musk*	6598	166	0.85
Banknote-auth	1372	4	0.56	*Optdigits*	5620	64	0.20
Breast-w	699	9	0.66	*Pendigits*	10992	16	0.20
Dermatology	366	34	0.31	*Satimage*	6430	36	0.24
Diabetes	768	8	0.65	*Segment*	2310	19	0.29
Fertility	100	9	0.88	*Seismic-bumps*	210	7	0.33
Gas-drift	13910	128	0.51	*Semeion*	1593	256	0.20
Glass	214	9	0.68	*Sonar*	208	60	0.47
Haberman	306	3	0.74	*Spambase*	4601	57	0.61
Heart-statlog	270	13	0.56	*Tic-tac-toe*	958	9	0.65
Ionosphere	351	34	0.64	*Vehicle*	846	18	0.26
Isolet	7797	617	0.27	*Waveform-5000*	5000	40	0.34
Jm1	10885	21	0.81	*Wdbc*	569	30	0.63
Kc1	2109	21	0.85	*Yeast*	1484	8	0.16

– Isolation Forest [9] (abbreviated to *IForest*),
– Activation Anomaly Analysis A^3 [12] (neural network based model),
– Mahalanobis distance [10] based score (abbreviated as *Mahalanobis*),
– Empirical Cumulative distribution based Outlier Detection *ECOD* [8], as well as its variant applying ECOD to PCA-transformed data (abbreviated as *ECOD+PCA*).

For each score function, the p-values are obtained from scores using *Multisplit* procedure (section 2.4), and control procedures (e.g. FOR control procedure) were applied; number of random splits in Multisplit procedure was set as $k = 10$. Each experiment used 60% of the inliers for training+calibration, and the remaining inliers and all of the outliers as the test set. For each test case, we repeated the entire process (starting from the training+calibration / test split) 20 times. For the control level we used $\alpha = 0.1$, which is a common value considered in literature. Code implementing the FOR control procedure, all tested methods and experiments is available publicly on GitHub[3].

Tests were conducted on 30 datasets constructed from real-world classification data. One of the classes (or several relatively similar ones) was selected as the inlier class, while other classes were considered as outliers. Basic summary of the datasets is presented in Table 1. For details on dataset construction, as well as their visualizations, we refer to the GitHub repository[4].

4 Results

Table 2 aggregates FOR mean values (and their standard errors) for the proposed FOR control procedure. FOR is controlled by at least one method on 20 datasets, but there are only 3 datasets where the same holds true for all methods at once. Mean FOR value was below 2α for at least one classifier in 29 out of 30 cases (except *Yeast* dataset). Even though A^3 controlled FOR on the largest number

[3] https://github.com/wawrzenczyka/FOR-CTL
[4] https://github.com/wawrzenczyka/FOR-CTL-datasets

Table 2: FOR values and standard errors under FOR control on tested datasets, for level $\alpha = 0.1$. Magenta „✓” denotes FOR $\leq \alpha$; black „✓” denote weaker FOR $\leq 2\alpha$.

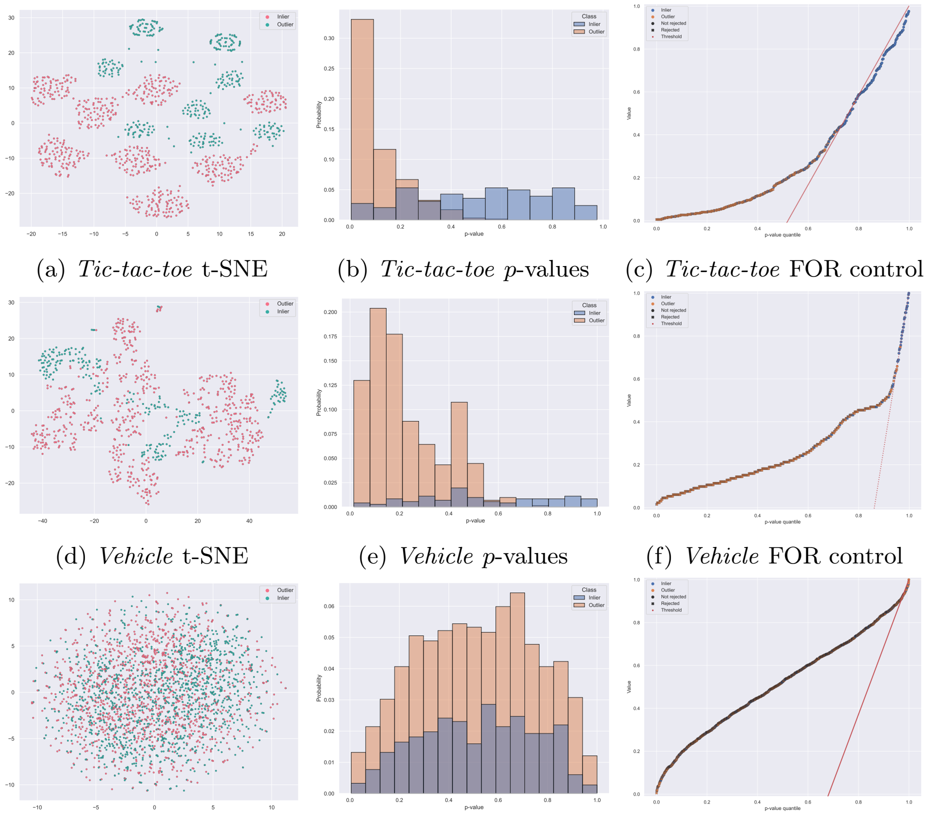
Dataset	**IForest**	$\boldsymbol{A^3}$	**Mahalanobis**	**ECOD**	**ECOD + PCA**
Musk	0.000 ± 0.000 ✓✓	0.137 ± 0.005 ✓	0.415 ± 0.040	0.000 ± 0.000 ✓✓	0.135 ± 0.011 ✓
Seismic-bumps	0.061 ± 0.018 ✓✓	0.093 ± 0.028 ✓✓	0.056 ± 0.018 ✓✓	0.048 ± 0.016 ✓✓	0.099 ± 0.025 ✓✓
Ionosphere	0.067 ± 0.014 ✓✓	0.107 ± 0.016 ✓	0.082 ± 0.011 ✓✓	0.081 ± 0.016 ✓✓	0.140 ± 0.009 ✓
Tic-tac-toe	0.075 ± 0.008 ✓✓	0.061 ± 0.004 ✓✓	0.140 ± 0.023 ✓	0.117 ± 0.012 ✓	0.272 ± 0.030
Breast-w	0.076 ± 0.005 ✓✓	0.083 ± 0.003 ✓✓	0.086 ± 0.004 ✓✓	0.057 ± 0.005 ✓✓	0.095 ± 0.004 ✓✓
Isolet	0.078 ± 0.006 ✓✓	0.047 ± 0.014 ✓✓	0.282 ± 0.006	0.090 ± 0.009 ✓✓	0.363 ± 0.004
Dermatology	0.078 ± 0.013 ✓✓	0.037 ± 0.006 ✓✓	0.049 ± 0.014 ✓✓	0.052 ± 0.008 ✓✓	0.206 ± 0.022
Semeion	0.078 ± 0.014 ✓✓	0.074 ± 0.012 ✓✓	0.049 ± 0.012 ✓✓	0.118 ± 0.016 ✓	0.000 ± 0.000 ✓✓
Banknote-auth	0.079 ± 0.009 ✓✓	0.086 ± 0.029 ✓✓	0.078 ± 0.003 ✓✓	0.069 ± 0.012 ✓✓	0.085 ± 0.006 ✓✓
Pendigits	0.091 ± 0.005 ✓✓	0.091 ± 0.004 ✓✓	0.108 ± 0.006 ✓	0.100 ± 0.012 ✓✓	0.099 ± 0.010 ✓✓
Satimage	0.094 ± 0.004 ✓✓	0.000 ± 0.000 ✓✓	0.084 ± 0.005 ✓✓	0.129 ± 0.010 ✓	0.105 ± 0.007 ✓
Segment	0.094 ± 0.007 ✓✓	0.317 ± 0.082	0.109 ± 0.009 ✓	0.085 ± 0.005 ✓✓	0.151 ± 0.017 ✓
Kc1	0.094 ± 0.008 ✓✓	0.091 ± 0.008 ✓✓	0.089 ± 0.009 ✓✓	0.170 ± 0.022 ✓	0.129 ± 0.007 ✓
Wdbc	0.097 ± 0.009 ✓✓	0.088 ± 0.010 ✓✓	0.100 ± 0.006 ✓✓	0.097 ± 0.010 ✓✓	0.137 ± 0.007 ✓
Optdigits	0.101 ± 0.010 ✓	0.074 ± 0.011 ✓✓	0.083 ± 0.008 ✓✓	0.160 ± 0.036 ✓	0.188 ± 0.012 ✓
Gas-drift	0.114 ± 0.010 ✓	0.000 ± 0.000 ✓✓	0.146 ± 0.011 ✓	0.088 ± 0.014 ✓✓	0.143 ± 0.013 ✓
Spambase	0.122 ± 0.021 ✓	0.139 ± 0.008 ✓	0.140 ± 0.017 ✓	0.217 ± 0.056	0.173 ± 0.025 ✓
Vehicle	0.127 ± 0.024 ✓	0.000 ± 0.000 ✓✓	0.073 ± 0.009 ✓✓	0.160 ± 0.045 ✓	0.162 ± 0.014 ✓
Glass	0.147 ± 0.030 ✓	0.163 ± 0.025 ✓	0.121 ± 0.019 ✓	0.186 ± 0.039 ✓	0.156 ± 0.018 ✓
Heart-statlog	0.153 ± 0.020 ✓	0.277 ± 0.035	0.170 ± 0.021 ✓	0.140 ± 0.026 ✓	0.246 ± 0.050
Diabetes	0.179 ± 0.017 ✓	0.151 ± 0.021 ✓	0.240 ± 0.030	0.271 ± 0.079	0.133 ± 0.023 ✓
Waveform-5000	0.204 ± 0.020	0.264 ± 0.076	0.161 ± 0.041 ✓	0.142 ± 0.024 ✓	0.355 ± 0.046
Abalone	0.206 ± 0.008	0.093 ± 0.014 ✓✓	0.170 ± 0.015 ✓	0.302 ± 0.030	0.167 ± 0.017 ✓
Arrhythmia	0.214 ± 0.037	0.187 ± 0.076 ✓	0.197 ± 0.030 ✓	0.300 ± 0.058	0.292 ± 0.017
Fertility	0.219 ± 0.024	0.121 ± 0.016 ✓	0.218 ± 0.024	0.229 ± 0.036	0.223 ± 0.022
Yeast	0.228 ± 0.069	0.299 ± 0.088	0.226 ± 0.078	0.301 ± 0.084	0.251 ± 0.092
Jm1	0.237 ± 0.008	0.000 ± 0.000 ✓✓	0.189 ± 0.022 ✓	0.312 ± 0.027	0.242 ± 0.020
Haberman	0.293 ± 0.035	0.472 ± 0.057	0.366 ± 0.058	0.153 ± 0.033 ✓	0.259 ± 0.039
Sonar	0.431 ± 0.078	0.125 ± 0.047 ✓	0.252 ± 0.052	0.175 ± 0.083 ✓	0.506 ± 0.030
Madelon	0.748 ± 0.013	0.495 ± 0.010	0.722 ± 0.008	0.100 ± 0.069 ✓✓	0.150 ± 0.082 ✓

of datasets, we will concentrate on *IForest* due to the higher consistency of its results. *IForest* managed to control FOR $\leq \alpha$ in 19 cases, $FOR \leq 2\alpha$ in 7 additional ones, and failed to keep FOR below 2α on the remaining 9 datasets.

Figure 5 illustrates 2-dimensional t-SNE representation of the data (1st column), distributions of the obtained p-values (2nd column) and FOR control procedure visualization (3rd column) for three of the datasets. *Tic-tac-toe* is an example of an easy dataset: we can see that the data forms distinct, separate clusters (Fig. 5a) and therefore there is a clear difference in inlier and outlier p-value distributions (Fig. 5b), as well as nearly perfectly uniform inlier distribution – *IForest* captured the inlier distribution really well. In that case, FOR control procedure has no issues capturing the clean portion of inliers (Fig. 5c) with the occasional outlier samples allowed by the α parameter.

Vehicle dataset in the second row is of medium difficulty: inlier and outlier samples (Fig. 5d) are a lot more difficult to separate. Note that the outlier p-value distribution (Fig. 5e) is shifted right, towards higher values, and inlier distribution is not as regular as in previous case. Though this example is significantly harder, we can see that in this particular case FOR control (Fig. 5f) divided the examples perfectly – though multiple outlier samples are extremely close to the threshold and might be incorrectly undetected with a small variations in their p-values. That leads to FOR for this dataset in Table 2 being slightly higher than desired.

DOI: 10.1007/978-3-031-36024-4_47



(a) *Tic-tac-toe* t-SNE (b) *Tic-tac-toe* p-values (c) *Tic-tac-toe* FOR control

(d) *Vehicle* t-SNE (e) *Vehicle* p-values (f) *Vehicle* FOR control

(g) *Madelon* t-SNE (h) *Madelon* p-values (i) *Madelon* FOR control

Fig. 5: FOR control for datasets with varying difficulty, based on Isolation Forest scores. Red (green) dots correspond to inliers (outliers).

Madelon dataset, on the other hand, was selected as an hard problem example. T-SNE visualization (Fig. 5g) does not capture any visible outlier characteristics, which suggests that the relationships in the data are complex. As a consequence, p-value distributions obtained from *IForest* scores (Fig. 5h) are extremely similar between outliers and inliers; note that the inlier p-value density is slightly bell shaped and thus deviates from the uniform, moreover the outliers p-value density does not decrease. As the proposed procedure assumes those properties, their lack has a profound impact on the FOR control (Fig. 5i). Lower than expected (when uniformity holds) number of inlier examples with high p-values causes outliers with high p-values to take their place; this results in the dramatic omission of dominant part of outlier samples, which results in a very high FOR value, as presented in Tab. 2. We note that deviation from uniformity for inlier p-values may be due to the fact that for this synthetic dataset inliers are not generated from one distribution.

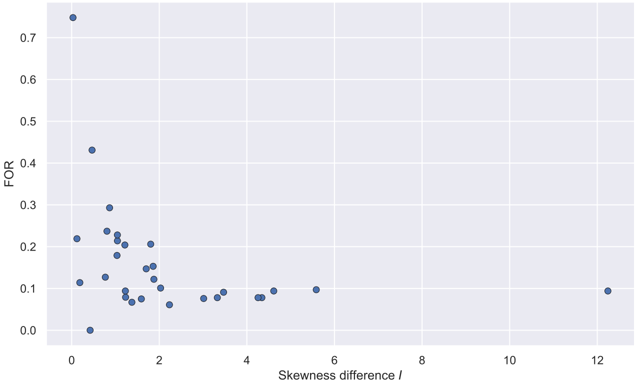


Fig. 6: Mean FOR values versus mean skewness difference I between outlier and inlier p-values distributions; each dot on the plot corresponds to one dataset.

Figure 5 suggests that obtaining high quality scores (and, as a result, reliable p-values) from the classifier is fundamental in order to ensure a good FOR control. That makes p-value distribution properties worth inspecting. In particular, we explored the effect of difference in skewnesses I of outlier and inlier p-value distribution, given by formula $I = Skew_{OUT} - Skew_{IN}$, on the empirical FOR value. We expect inlier distribution to be uniform (so $Skew_{IN}$ should be close to 0), whereas probability mass for the outlier distribution should be concentrated on small p-values (resulting in large positive values of $Skew_{OUT}$), which should mean that for a p-value distributions satisfying the imposed assumptions, I should be both positive and relatively large. Indeed, as illustrated in the Figure 6, datasets where I is large are also the ones where the proposed procedure works really well; on the other hand, when I falls below 2, FOR control becomes unreliable. This emphasizes the dependency of FOR control on outlier score quality – we can control FOR only if outlier scores make that possible.

Figure 7 visualizes relationship between FOR and FDR when FOR is controlled on all sets. Observe that good control of FOR doesn't imply low FDR value (and vice versa, see Figure 1). Moreover, this holds irrespectively of the scoring method – even when a given method controls FOR at a given level, its FDR might remain high.

5 Conclusions

In the paper we propose the first (to our knowledge) empirical procedure allowing for FOR control in the outlier detection scenario, which is vital in many real-life scenarios. Our approach is mathematically justified and accounts for prior research on the related control algorithms, such as Benjamini-Hochberg procedure for FDR control and its empirical Bayes underpinning. It is important to note that the FOR control problem is substantially harder than its FDR counterpart,

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_47

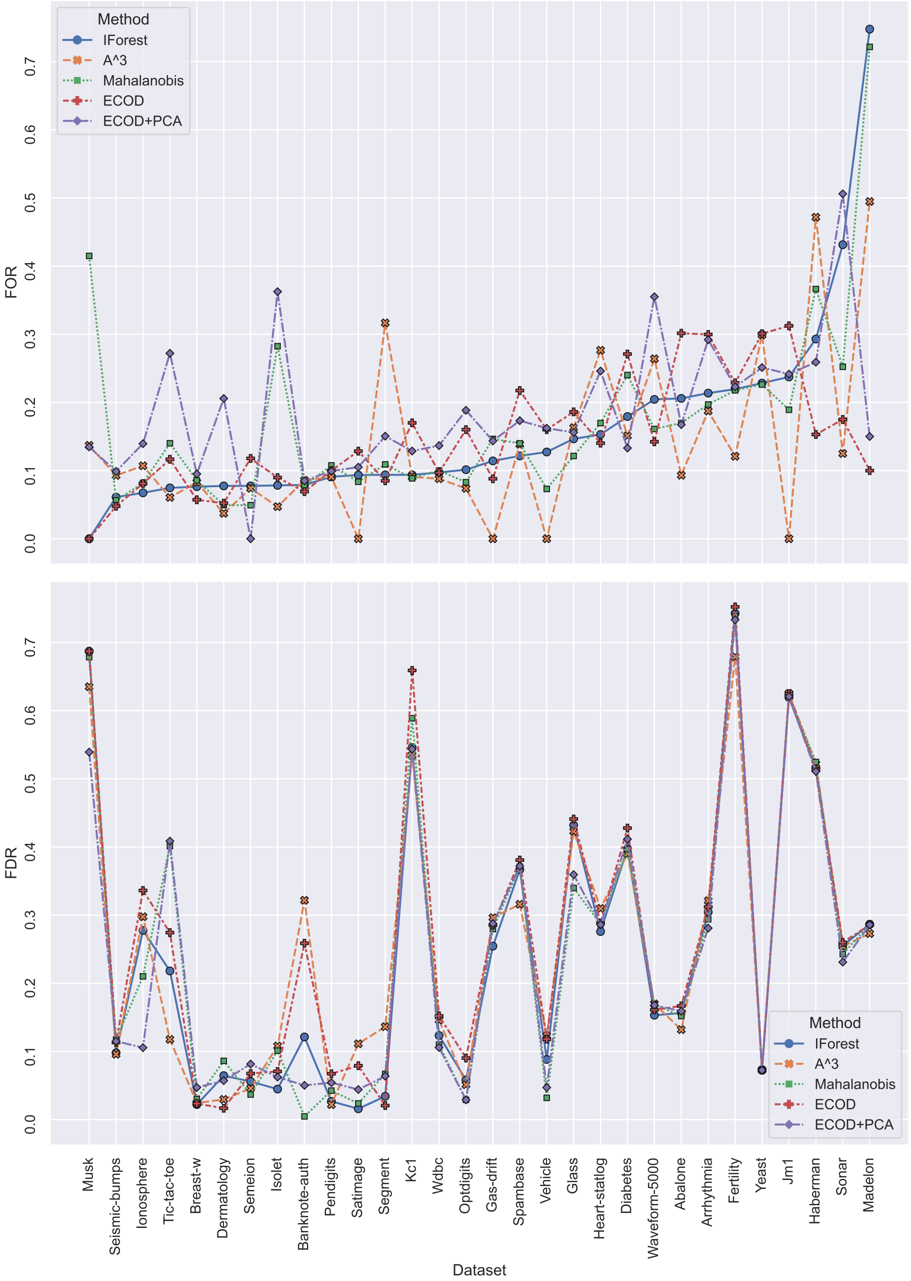


Fig. 7: FOR values for all methods and data sets (upper panel) with the corresponding values of FDR (lower panel).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_47

due to threshold calculation requiring outlier distribution properties. The experiments presented in the paper prove method's capability of controlling FOR as long as good quality p-values are provided to the algorithm. That ties into the most significant limitation of the described procedure – the dependence on the outlier scores supplied by the external methods makes their imperfections transfer to the researched task. FOR control procedure is sensitive to breaking its assumptions – this is most evident when skewness difference between outlier and inlier p-value distribution is low, which results in outliers replacing a portion of missing inlier distribution, which in turn causes their uncontrolled omission. Futher research on the topic might include handling those scenarios by making FOR control procedure robust to closeness in the inlier and outlier distribution, as well as consideration of estimators of π in the threshold rule such as Storey's estimator [13].

References

1. Bates, S., Candès, E., Lei, L., Romano, Y., Sesia, M.: Testing for outliers with conformal p-values. Annals of Statistics, to appear (2023)
2. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) **57**, 289–300 (1995)
3. Bühlmann, P., van de Geer, S.: Statistics for High-Dimensional Data. Springer (2011)
4. Chao, M.T., Strawderman, W.E.: Negative moments of positive random variables. Journal of the American Statistical Association **67**, 429–431 (1972)
5. Dudoit, S., van der Laan, M.J.: Multiple Testing Procedures with Applications to Genomics. Springer (Jan 2008)
6. Efron, B.: Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics Monographs, Cambridge University Press (2010)
7. Genovese, C., Wasserman, L.: Operating characteristics and extensions of false discovery rate procedures. Journal of the Royal Statistical Society: Series B (Methodological) **64**(3), 499–517 (2002)
8. Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., Chen, G.H.: Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. In: IEEE Transactions on Knowledge and Data Engineering (2022)
9. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. pp. 413–422 (2008)
10. Liu, R.Y., Parelius, J.M., Singh, K.: Multivariate analysis by data depth: descriptive statistics, graphics and inference. The Annals of Statistics **27**(3), 783–858 (1999)
11. Meinshausen, N., Meier, L., Bühlmann, P.: P-values for high-dimensional regression. Journal of the American Statistical Association **104**, 1671–1681 (Nov 2008)
12. Sperl, P., Schulze, J.P., Böttinger, K.: Activation anomaly analysis. In: Machine Learning and Knowledge Discovery in Databases, pp. 69–84. Springer (2021)
13. Storey, J.: Direct approach to false discovery rates. Journal of the Royal Statistical Society. Series B (Methodological) **64**, 479–498 (2002)
14. Takahashi, H., Ichinose, N., Yasusei, O.: False-negative rate of sars-cov-2 rt-pcr tests and its relationship to test timing and illness severity. IdCases **28** (2022)