Scuola di Dottorato
Università degli Studi di Milano-Bicocca

Department of

Economics, Management, and Statistics

Ph.D. program: **Economics, Statistics and Data Science** Cycle: **XXXVII**
Curriculum: **Statistics**

Distribution-free Outlier Detection

Name: **Chiara Gaia**
Surname: **Magnani**
Registration number: **849037**

Supervisor: Prof. **Aldo Solari**
Coordinator: Prof. **Matteo Manera**

Academic Year: 2024-2025

Abstract

This thesis addresses the problem of outlier detection, also referred to as anomaly detection or novelty detection. A flexible, distribution-free method is developed for collective outlier detection and enumeration, designed for situations in which the presence of outliers can be detected powerfully even though their precise identification may be challenging due to the sparsity, weakness, or elusiveness of their signals. This method builds upon recent developments in Conformal Inference and integrates classical ideas from other areas, including Multiple Testing and rank tests. The key innovation lies in developing a principled and effective approach for automatically choosing the most appropriate Machine Learning classifier and two-sample testing procedure for a given data set. The performance of the proposed method is investigated through extensive empirical demonstrations, including an analysis of the LHCO high-energy particle collision data set.

Sommario

Questa tesi affronta il problema della rilevazione di osservationi anomale, a cui spesso ci si riferisce anche con il nome di *outlier*. Viene proposto un metodo non parametrico in grado di rilevarne la presenza, quantificarne il numero ed eventualmente identificarle. Il metodo è progettato per situazioni in cui la presenza di osservazioni anomale può essere rilevata anche se la loro identificazione precisa può risultare difficile a causa della scarsità o dell'elusività di tali osservazioni. Questo metodo si basa su recenti sviluppi nel campo della *Conformal Inference* e integra idee classiche provenienti da altre aree, tra cui il *Multiple Testing* e i test basati sui ranghi. L'innovazione chiave consiste nello sviluppo di un approccio rigoroso ed efficace per la selezione automatica del classificatore di Machine Learning e della procedura di test a due campioni più appropriato per un determinato insieme di dati. Le prestazioni del metodo proposto sono valutate attraverso simulationi e analisi di dati reali, inclusa un'analisi dei dati sulle collisioni di particelle ad alta energia.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Outlier detection is a fundamental statistical problem with numerous applications, ranging from many areas of scientific research to fraud detection and security monitoring. In the age of artificial intelligence, its relevance is further growing due to concerns over the rise of potentially unrealistic synthetic data. The complexity of high-dimensional data has encouraged the use of sophisticated machine learning models for outlier detection, but these models often lack transparency and are prone to errors, complicating their reliability. This has spurred substantial interest in conformal inference (Vovk et al., 2005), which can provide principled statistical guarantees for any outlier detection algorithm under relatively mild assumptions.

While conformal inference is gaining momentum in both statistics and machine learning, its focus in the context of outlier detection has so far primarily been on individual-level outlier *identification*, where each data point is separately evaluated as a potential outlier. However, individual-level identification is not always feasible. In many applications, outliers may be too rare or weak to reach statistical significance (Donoho and Jin, 2015), or they may exhibit unremarkable behavior in isolation but reveal anomalous patterns when analyzed collectively (Feroze et al., 2021), such as showing under-dispersion relative to inliers.

To extend the applicability of conformal inference to these particularly challenging settings, this thesis introduces a novel approach for *collective outlier detection* that addresses two main challenges: (1) testing the global null hypothesis that a dataset—or a subset of

it—contains no outliers, and (2) estimating the number of outliers present. The proposed method solves these interrelated problems by leveraging in an innovative way powerful "black-box" machine learning algorithms and highly flexible, data-driven test statistics. This integrative approach is tailored to maximize power for the data at hand and offers reliable type-I error control under mild assumptions. Additionally, the proposed procedure is scalable to large data sets. It builds on recent advances in conformal inference, while also revisiting and integrating several classical concepts from other areas of statistics. This collective approach is demonstrated to succeed even in scenarios where individual outlier detection fails.

The proposed method is broadly applicable. Financial institutions, for instance, could use it to identify complex fraud schemes involving seemingly legitimate transactions. Similarly, cybersecurity systems could apply it to detect coordinated denial-of-service attacks (Ahmed and Mahmood, 2014). In high-energy physics, where collective outlier detection generally plays a central role (Vatanen et al., 2012), researchers could use the proposed procedure to sift through massive datasets from particle decay sensors in the search for new particles. These signals are typically rare and weak, making individual-level detection impractical, but they present a promising use case for the proposed method, as previewed in Figure 1.1.

1.2 Main Contributions of the Thesis

In this thesis, ACODE, an *Automatic Conformal Outlier Detection and Enumeration* method, is presented. ACODE leverages two-sample rank tests applied to univariate *conformity scores*, which can be generated by any model trained to distinguish outliers from inliers. By employing split-conformal inference, ACODE converts these scores into statistical tests based on *conformal p-values* controlling the type-I error rate. In addition, by utilizing the closed testing principle (Marcus et al., 1976), ACODE can construct simultaneous lower confidence bounds for the number of outliers within any subset of the test sample (Goeman and Solari, 2011), while also offering a global test for outlier detection as a byproduct of outlier enumeration.

While these high-level ideas are intuitive, the flexibility of the conformal inference frame-

work introduces substantial complexity and a great deal of implementation freedom. This leads to two key methodological questions addressed in this thesis regarding the choice of the most suitable classifier and rank test, along with other related methodological challenges. The main contributions are summarized below.

- **Automatic Selection of the Classifier.** The first question involves selecting the most effective classification algorithm for computing conformity scores. The trade-offs between existing approaches based on one-class classifiers (Bates et al., 2023) and positive-unlabeled learning via binary classification (Marandon et al., 2024) are examined. In particular, ACODE integrates AdaDetect with one-class classifiers, as suggested in Marandon et al. (2024) (see Sections 4.2.1 and 5.2). This integration provides valuable flexibility in practice, as the performance of different classifiers can vary substantially across scenarios (see Chapter 6 for details). Through this adaptive approach, ACODE retains the theoretical properties established in Marandon et al. (2024), while enhancing flexibility via data-driven classifier selection.

- **Combining Conformal p-values.** The second question concerns the choice of the rank test to determine statistical significance based on these scores. Focusing exclusively on rank tests is a natural choice for distribution-free inference that aligns with other works in the conformal inference literature (e.g., Bates et al., 2023; Kuchibhotla, 2021). This choice is further motivated by the fact that conformal p-values are precisely normalized ranks. Section 4.3 presents a comparison of existing approaches for combining conformal p-values in terms of admissibility and power, yielding several insightful observations. For example, the Simes test applied to conformal p-values is shown to be strictly dominated by the permutation Simes test. Additionally, the Wilcoxon-Mann-Whitney rank test can be expressed as the average of conformal p-values.

- **Adaptive Rank Test.** Central to the proposed method is the property of *locally most powerful rank* (LMPR) test, which characterizes the test that is most powerful in a neighborhood of the null hypothesis of no outliers among rank tests (see Section 4.3 for a detailed comparison). This property is particularly relevant for the goal of detecting and enumerating outliers when the signal is weak and sparse, namely in a

neighborhood of the null hypothesis. The proposed approach builds upon the Shiraishi test, an extension of the Wilcoxon-Mann-Whitney (WMW) sum-rank test (Wilcoxon, 1945; Mann and Whitney, 1947a), that Shiraishi (1985) proved to be LMPR for a two-component mixture alternative, but that depends on the distribution of (the scores of) the outliers. As this distribution is generally unknown, a principled, data-driven solution is presented in Section 5.2. ACODE implements an adaptive Shiraishi test based on the techniques described in Sections 5.3.3.1 and 5.3.3.2.

- **Asymptotic Null Distribution of the Shiraishi Test Statistic under Exchangeability.** The large sample asymptotic distribution of the Shiraishi test statistic is derived by Shiraishi (1985) under the assumption of i.i.d. conformity scores. The extension of this result to exchangeable conformity scores is given in Theorem 12. See Section 5.3.2 for details.

- **Asymptotic Relative Efficiency of the Shiraishi Test.** The comparison of the Shiraishi rank test with an *oracle* test that achieves maximal power by exploiting knowledge of the true distribution of conformity scores is presented in Section 5.3.4. Shiraishi (1985) showed that the local limiting power of the LMPR test relative to the most powerful test equals one minus the (limiting) proportion of test points in the combined sample. Additionally, while the optimal power of the oracle Neyman-Pearson test is not achievable through rank tests, the local optimality of the Shiraishi rank test is crucial, for detecting and enumerating outliers when the signal is weak and sparse—that is, when the alternative hypothesis lies in a neighborhood of the null hypothesis.

- **Equivalence of One-Step and Two-Step Oracle Optimal Procedures.** The main result in Section 5.3.4.1 shows that to detect the presence of signal in the test set, the oracle likelihood ratio test applied directly to the multivariate test observations is equivalent to a two-step oracle procedure that relies on the likelihood ratio test applied to univariate sufficient statistics, which serves as "oracle conformity scores". This characterization facilitates a more direct comparison to ACODE while highlighting a connection to the results in Marandon et al. (2024) on the optimal choice of non-conformity scores.

- **Shortcuts for Closed Testing with Adaptive Simes and Shiraishi Local Tests.**

  Novel computational shortcuts are proposed for efficient closed testing with two local tests: (*i*) the adaptive Simes test using the Storey estimator for the number of true nulls (Section 3.4.3), and (*ii*) the Shiraishi test with increasing outlier density (Section 3.4.6).

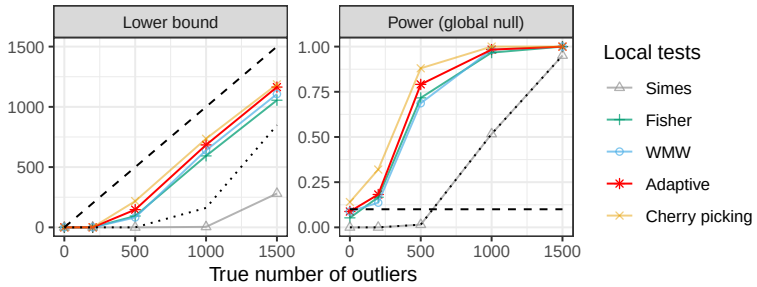

Figure 1.1: Preview of performance of ACODE on the LHCO data. Left: Median values over repeated experiments for a 90% lower confidence bound for the number of outliers. Right: power against the global null hypothesis of no outliers at the 10% level (horizontal line). The results are shown as a function of the true number of outliers in a test set of cardinality 10,000. ACODE utilizes a testing procedure that may be adaptively selected (red curve) or fixed (other solid curves). The "cherry-picking" (yellow curve) approach selects the combination of the classifier and testing procedure with the most appealing outcomes, resulting in inflated type-I error. Dotted curve: number of *individual* discoveries or power obtained by applying the Benjamini-Hochberg procedure (BH) to conformal p-values, controlling the false discovery rate below 10%.

1.2.1 Illustrative Application of ACODE to the LHCO Dataset

The effectiveness of this approach is showcased in Figure 1.1, analyzing data from the 2020 Large Hadron Collider Olympics (LHCO) (Kasieczka et al., 2021).

As detailed in Section 6.2, the analysis of the LHCO data aims to (1) detect outliers, or unusual collision events, within a test sample, and (2) establish a 90% lower confidence bound for the number of these outliers. The proposed method not only controls errors for both tasks but also provides insightful inferences. This illustrates ACODE's adaptability, as it selects the most suitable testing procedures in a data-driven manner.

Additionally, Figure 1.1 highlights the challenge of conducting an adaptive data-driven analysis without incurring selection bias. It compares ACODE's performance with that of a naive "cherry-picking" approach that greedily tests various classification algorithms and testing procedures, selecting the one with the most appealing outcomes. Unsurprisingly, such a heuristic can lead to inflated type-I errors, as selecting the winner among many candidates typically results in overly optimistic inferences, unless corrective measures are taken to mitigate selection bias. In contrast, ACODE is protected against selection bias while achieving similarly high power across all scenarios tested.

Figure 1.1 also highlights the potential advantages of an approach specifically designed for collective outlier detection over standard alternatives for individual-level discovery. The results show that ACODE leads to much more informative collective inferences compared to the Benjamini-Hochberg (BH) procedure for False Discovery Rate (FDR) control (Benjamini and Hochberg, 1995) applied to individual conformal p-values (Bates et al., 2023; Marandon et al., 2024). In fact, the BH procedure aims to identify outliers individually, and is generally not an effective solution for global testing or outlier enumeration.

1.2.2 Structure of the Thesis

The thesis is structured as follows. Chapter 2 reviews rank tests as tools for distribution-free inference. Two main examples are presented—the Wilcoxon-Mann-Whitney test (Wilcoxon, 1945; Mann and Whitney, 1947a) and the Shiraishi test (Shiraishi, 1985)—with a particular focus on their local optimality for specific alternatives.

Chapter 3 discusses Multiple Testing procedures designed to address the challenges arising in the selective inference framework. The following are presented: (*i*) methods for global testing with type-I error probability control (*ii*) multiple testing procedures for *family-wise error rate* control; (*iii*) the seminal Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) for controlling the false discovery rate; and, (*iv*) the Closed Testing principle (Marcus et al., 1976) for simultaneous inference with *false discovery proportion* control (Genovese and Wasserman, 2006; Goeman and Solari, 2011).

Chapter 4 reviews useful tools from Conformal Inference adopting the split conformal approach. These techniques provide distribution-free inference with finite-sample guarantees to perform outlier detection for multiple comparisons.

While Chapters 2-4 mostly provide a concise review of the existing literature and theoretical results on which the novel contributions are built, Chapters 5 and 6 present the original methodological and empirical contributions of this thesis. Chapter 5 develops the proposed method for collective outlier detection and enumeration and examines its theoretical properties. Chapter 6 reports extensive numerical experiments that evaluate the performance of the proposed method on both synthetic and real data.

Chapter 2

Rank Tests

2.1 Introduction

In this chapter, the basic definitions of rank statistics and some of their properties are reviewed, with a special emphasis on *local optimality* among rank tests and their asymptotic distributions, which will be fundamental for the theoretical and mathematical foundation of the novel contribution presented in Chapter 5. For more detailed treatments of this topic, see Cox and Hinkley (1979); Hájek et al. (1999); van der Vaart (2000); Lehmann and Romano (2005).

From an intuitive perspective, a rank statistic depends only on the relative ordering (ranks) of the observations. This means that such a rank statistic remains unchanged when actual values change as long as their ordering is preserved, but does differ when the relative order is modified. A rank test is a hypothesis test based on such a rank statistic.

Using ranks rather than actual values avoids distributional assumptions about the data, resulting in flexible inference procedures applicable to real-world scenarios where the underlying distribution is often unknown. However, this approach can result in information loss, potentially reducing test efficiency or power. Fortunately, the loss of efficiency is quite small when the best rank test is used (Section 6.1 Cox and Hinkley, 1979), and rank tests provide valuable distribution-free inference.

They are particularly relevant in comparative problems, such as two-sample tests, where two samples are given and the aim is to draw inferences about differences between their distributions.

Given a sample of real-valued random variables $W_1, \ldots, W_N$, for any $j \in [N]$ with $[N] := \{1, \ldots, N\}$ the rank of W_j is defined as

$$R_j := \sum_{i \in [N]} \mathbb{1}[W_i \leq W_j].$$

This definition is well-posed when the random variables $W_1, \ldots, W_N$ have continuous distribution functions, so that ties among observations occur with probability zero. In this case, the order statistic is denoted by $W_{(1)} < \ldots < W_{(N)}$ and the ranks satisfy the following equation

$$W_j = W_{(R_j)}, \qquad j \in [N].$$

If W_j is tied with some other observations, its rank is defined as the mean of all indices $i \in [N]$ such that $W_i = W_{(R_j)}$. Throughout this thesis, $W_1, \ldots, W_N$ are assumed to have a continuous distribution function. The following lemma, proved in Lemma 13.1 of van der Vaart (2000) and Section 3.1 of Hájek et al. (1999), summarizes some well-known results.

Lemma 1 (Lemma 13.1 in van der Vaart (2000))**.** *Assume that $W_1, \ldots, W_N$ is a random sample of independent and identically distributed observations drawn from a continuous distribution F with density f. The following results hold:*

(i) the ordered vector $W_{(1)} < \ldots < W_{(N)}$ and the rank vector $R_1, \ldots, R_N$ are independent;

(ii) the ordered vector $W_{(1)} < \ldots < W_{(N)}$ has density $N! \prod_{j \in [N]} f(w_j)$ on the region $w_1 < \ldots < w_1$ and 0 otherwise;

(iii) the rank vector $R_1, \ldots, R_N$ is uniformly distributed on the set of all $N!$ permutations of $[N]$.

From the definition of ranks and Lemma 1, two main characteristics that favor the use of rank tests can be derived. From point (iii) in Lemma 1, it is clear that any rank statistic maintains an identical distribution across all random vectors with independent and identically distributed (i.i.d.) observations. This property renders rank tests *distribution-free* within the class of i.i.d. models. Moreover, the ordering of observations remains unaffected by continuous and strictly increasing transformations of the data. This property

is called *invariance* under such transformations, which can be interpreted as coordinate changes, ensuring that rank tests produce consistent results regardless of the scale or units used to measure the underlying phenomenon.

2.1.1 The Wilcoxon-Mann-Whitney Statistic

As mentioned above, rank tests are particularly valuable for comparative problems, such as two-sample scenarios. In these situations, the aim is to make inferences about the distributions of two given samples, especially comparing aspects of their distributions, such as location parameters, or testing whether the samples are independent.

Consider two independent samples

$$X_1, \ldots, X_m \overset{iid}{\sim} F, \qquad Y_1, \ldots, Y_n \overset{iid}{\sim} Q \tag{2.1}$$

where F, Q are continuous cumulative distribution functions (c.d.f.'s) on $\mathbb{R}$. Let $R_{m+1}, \ldots, R_{m+n}$ denote the ranks of $Y_1, \ldots, Y_n$ in the pooled sample $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$. Consider the null hypothesis that the two samples are identically distributed, i.e.,

$$H : Q = F. \tag{2.2}$$

A well-known two-sample rank test for testing the null hypothesis in (2.2) is the *Wilcoxon rank-sum test* (Wilcoxon, 1945), which rejects the null hypothesis for large values of:

$$T^{\mathrm{W}} = \sum_{j \in [n]} R_{m+j}. \tag{2.3}$$

Tabulated values of the null distribution of the Wilcoxon rank-sum statistic can be found in Wilcoxon (1945) and Milton (1964).

By writing

$$R_{m+j} = \sum_{i \in [m]} \mathbb{1}[X_i \leq Y_j] + \sum_{j' \in [n]} \mathbb{1}[Y_{j'} \leq Y_j]$$

and noting that $\sum_{j \in [n]} \sum_{j' \in [n]} \mathbb{1}[Y_{j'} \leq Y_j] = n(n+1)/2$, it can be easily shown that the

Wilcoxon test is equivalent to the *Mann-Whitney U-statistic* (Mann and Whitney, 1947a):

$$T^{\mathrm{MW}} = \sum_{j\in[n]} \sum_{i\in[m]} \mathbb{1}[X_i \leq Y_j], \tag{2.4}$$

in the sense that $T^{\mathrm{W}} = T^{\mathrm{MW}} + n(n+1)/2$. Tabulated values of the null distribution of the Mann-Whitney statistic can be found in Mann and Whitney (1947a).

Hereafter, the statistic in (2.3), referred to as the *Wilcoxon-Mann-Whitney statistic* and denote as T^{WMW} due to its equivalence with the Mann-Whitney U-statistic, will be primarily used.

2.2 Rank Tests as Permutation Tests

Rank tests represent a particular class of *permutation tests* that rely on statistics computed from the ranks of observations rather than their exact values. The validity of permutation tests depends solely on the assumption of *exchangeability* among observations.

A random vector $W_1, \ldots, W_N$ with joint distribution is said to be *exchangeable* if, for every permutation $\sigma \in S_N$, where S_N is the set of all possible permutations of the index set $[N]$, it holds:

$$(W_1, \ldots, W_N) \overset{d}{=} (W_{\sigma(1)}, \ldots, W_{\sigma(N)}), \tag{2.5}$$

where $\overset{d}{=}$ denotes the equality in distribution. When $W_1, \ldots, W_N$ forms an exchangeable random vector, all its elements are identically distributed (though not necessarily independent), and all permutations of the elements are equally likely.

For exchangeable random vectors with discrete or continuous distributions, a simple characterization based on the symmetry of their probability functions can be provided. Suppose that the random vector $W_1, \ldots, W_N$ is either discrete or has joint density. Let p_W denote the probability mass function in the discrete case or the joint density function in the continuous case. The vector is exchangeable if and only if for all possible values $w_1, \ldots, w_N$ and for every permutation $\sigma \in S_N$:

$$p_W(w_1, \ldots, w_N) = p_W(w_{\sigma(1)}, \ldots, w_{\sigma(N)}).$$

Note that the term "exchangeable" can refer interchangeably to either a random vector or its joint distribution.

A permutation test is a hypothesis test to evaluates evidence against the null hypothesis that the vector $(W_1, \ldots, W_N)$ is exchangeable, using a test statistic $\tau(W_1, \ldots, W_N)$ and its permutation distribution. The test is fundamentally based on the permutation distribution of the statistic τ:

$$\Pr[\tau \leq \bar{\tau}] = \frac{\sum_{\sigma \in S_N} \mathbb{1}[\tau(W_{\sigma(1)}, \ldots, W_{\sigma(N)}) \leq \bar{\tau}]}{N!}. \tag{2.6}$$

Suppose the null hypothesis is rejected for large values of the statistic $\tau(W_1 \ldots, W_N)$. The permutation p-value is defined as:

$$p^{\text{perm}} = \frac{\sum_{\sigma \in S_N} \mathbb{1}[\tau(W_{\sigma(1)}, \ldots, W_{\sigma(N)}) \geq \tau(W_1, \ldots, W_N)]}{N!}. \tag{2.7}$$

Under the null hypothesis that the random vector $W_1, \ldots, W_N$ is exchangeable, the permutation p-value in (2.7) is valid for any test statistic that rejects the null hypothesis for large values. That is, under the assumption of exchangeability, for any $\alpha \in (0, 1)$ it holds:

$$\Pr[p^{\text{perm}} \leq \alpha] \leq \alpha. \tag{2.8}$$

Different proofs of this fundamental result are available (e.g., Hoeffding, 1952; Lehmann and Romano, 2005; Pesarin, 2015; Hemerik and Goeman, 2018).

As established in Hemerik and Goeman (2018), the fundamental assumption to prove that the permutation p-value is exact is their Condition 1. This condition essentially requires that the group of transformations can be partitioned into subsets of equal cardinality such that: (i) transformations within the same subset produce identical values of the test statistic, and (ii) transformations from different subsets produce different values of the test statistic. Under this condition, the permutation p-values are uniformly distributed over $\{1/N!, 2/N!, \ldots, 1\}$, so that the inequality in equation (2.8) becomes an equality

$$\Pr[p^{\text{perm}} \leq \alpha] = \alpha \tag{2.9}$$

for any $\alpha \in \{1/N!, 2/N!, \ldots, 1\}$. One important special case of Condition 1 occurs when the test statistic τ has no ties in its distribution. Under this assumption, their Condition 1 corresponds to partitioning the group of transformations into singletons. While this no-ties assumption ensures that permutation p-values are exact, the presence of ties causes the permutation p-values to be stochastically larger than the uniform distribution, i.e., the equality in (2.9) is not guaranteed for $\alpha \in \{1/N!, 2/N!, \ldots, 1\}$ and only the inequality in (2.8) holds. Any random variable which satisfies (2.8) is called *super-uniform* and tends to be conservative.

The computational cost of permutation p-values grows factorially with the number of observations in the sample, making exact computation prohibitive for large datasets. This has motivated the development of computationally more tractable approaches. Permutation p-values that are built on a smaller number $B < N!$ of permutations sampled uniformly at random:

$$p^{\text{perm}} = \frac{1 + \sum_{b \in [B]} \mathbb{1}[\tau(W_{\sigma_b(1)}, \ldots, W_{\sigma_b(N)}) \geq \tau(W_1, \ldots, W_N)]}{1 + B} \tag{2.10}$$

are proved to be valid, and exact under Condition 1 in Hemerik and Goeman (2018).

Hemerik and Goeman (2018) demonstrated that validity can be maintained even when using arbitrary subsets of the permutation group. This result removes the restrictive requirement that permutations must form a subgroup and allows practitioners to employ non-uniform sampling strategies, thereby broadening the range of computationally efficient permutation test designs. However, this raises the issue of how to choose a good subset of permutations—a question addressed by Koning and Hemerik (2024), who showed that permutation p-values remain valid when the set of considered permutations is restricted to a fixed subgroup of the full permutation group strategically chosen to obtain good power.

2.2.1 Finite-Sample Distribution of Rank Tests

The finite-sample distribution of a statistic can always be derived under the assumption that the vector of observations $W_1, \ldots, W_N$ is exchangeable, using the permutation distribution in (2.6). When dealing with rank statistics, their permutation distribution can be derived under the assumption that the observations $W_1, \ldots, W_N$ are i.i.d. This assumption ensures that the corresponding rank vector $R_1, \ldots, R_N$ is exchangeable, as proved in point (iii) of

Lemma 1.

Adapting the permutation distribution in (2.6) to the two-sample case, the permutation distribution of the Wilcoxon statistic is:

$$\Pr\left[T^{\mathrm{W}} = k\right] = \frac{\sum_{h \in \mathcal{C}_{m+n,n}} \mathbb{1}[T^{\mathrm{W}}\left(R_{m+h_1}, \ldots, R_{m+h_n}\right) = k]}{\binom{m+n}{n}}, \tag{2.11}$$

where $h = (h_1, \ldots, h_n)$ is an element of $\mathcal{C}_{m+n,n}$ which is the set of combinations of n elements from the set $[m+n]$.

Let $\pi_{n,m}(k)$ denote the numerator of the finite-sample distribution of the Wilcoxon statistic in (2.11). This can be expressed with the following recurrence formula (Brus, 1988). For any $m, n \in \mathbb{N}$ and for $k = n(n+1)/2, \ldots, n(n+2m+1)/2$:

$$\pi_{n,m}(k) = \pi_{n,m-1}(k) + \pi_{n-1,m}(k-n-m)$$

with initial conditions:

$$\pi_{n,0}(k) = \begin{cases} 1, & k = n(n+1)/2, \\ 0, & \text{otherwise}, \end{cases} \qquad \pi_{0,m}(k) = \begin{cases} 1, & k = 0, \\ 0, & \text{otherwise}, \end{cases}$$

and $\pi_{n,m}(k) = 0$ for $k < n(n+1)/2$.

Since the Mann-Whitney U-statistic is equal to the Wilcoxon statistic up to the additive constant $n(n+1)/2$, the finite-sample distribution of the Mann-Whitney statistic is equal to that of the Wilcoxon statistic, with $k = 0, \ldots, mn$. It is worth noting that the recurrence formula of $\pi_{n,m}$, and consequently the distribution of the Wilcoxon-Mann-Whitney test, depends only on the sample sizes n and m.

2.3 Local Optimality among Invariant Tests

In this section, a theoretical justification for the Wilcoxon–Mann–Whitney statistic is provided by showing that it is locally most powerful among invariant tests under a specific alternative. Following the same method, also the *Shiraishi statistic* (Shiraishi, 1985) is derived. It is a generalization of the Wilcoxon-Mann-Whitney statistic. This establishes

their optimality in detecting small deviations from the null hypothesis under specific class of alternatives, and explains their effectiveness in practice.

Let us consider the two-sample problem:

$$X_1, \ldots, X_m \overset{iid}{\sim} F, \qquad Y_1, \ldots, Y_n \overset{iid}{\sim} Q, \tag{2.12}$$

where F is a continuous c.d.f's with density f. Consider the null hypothesis that the two samples are i.i.d. from the same distribution, i.e.,

$$H : Q = F$$

against the alternative hypothesis that Q is a function of F distinct from F, i.e.,

$$K : Q = h(F)$$

where h is a differentiable function on $[0, 1]$ with derivative h'. Following the Neyman-Pearson lemma, the optimal rank test rejects the null hypothesis for large values of

$$\frac{\Pr[R_Y = r_Y \ ; \ K]}{\Pr[R_Y = r_Y \ ; \ H]} \tag{2.13}$$

where $R_Y = (R_{m+1}, \ldots, R_{m+n})$ is the rank vector of $(Y_1, \ldots, Y_n)$ with respect to the pooled sample $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ and $r_Y = (r_{m+1}, \ldots, r_{m+n})$ is one of its realizations.

Since the rank vector R_Y is uniformly distributed under the null distribution over the set of combinations of n elements from the set $[m + n]$, the likelihood ratio (2.13) is equal to

$$\binom{m+n}{n} \Pr[R_Y = r_Y \ ; \ K].$$

Hoeffding (1948) and Lehmann (1953) establish the joint distribution of the ranks $R_{m+1}, \ldots, R_{m+n}$ of the test observations under the alternative hypothesis K as follows. Using the notation $\{R_Y = r_Y\}$ to indicate the set of vectors $(Y_1, \ldots, Y_n)$ such that their

rank vector R_Y is equal to r_Y, it follows:

$$\begin{aligned}
\Pr[R_Y = r_Y; K] &= \int_{\{R_Y = r_Y\}} \prod_{j\in[n]} h'(F(y_j)) f(y_j)\, dy_1 \cdots\, dy_n \\
&= \frac{1}{\binom{m+n}{n}} \int_{\{R_Y = r_Y\}} \prod_{j\in[n]} h'(F(y_j)) \binom{m+n}{n} \prod_{j\in[n]} f(y_j)\, dy_1 \cdots\, dy_n \\
&= \frac{1}{\binom{m+n}{n}} \mathbb{E}_F \left[\prod_{j\in[n]} h'(F(Y_j)) \;\middle|\; R_Y = r_Y \right] \\
&= \frac{1}{\binom{m+n}{n}} \mathbb{E}_F \left[\prod_{j\in[n]} h'(F(Y_{(r_{m+j})}) \right].
\end{aligned}$$

Under the null distribution, the random variable $F(Y_{(r_{m+j})})$ is the r_{m+j}th ordered statistic in a vector of N uniform random variables on $[0,1]$. Let $U_N^{(r_{m+j})}$ denote it. Hence, the probability of the test ranks under the alternative can be written as:

$$\Pr[R_Y = r_Y; K] = \frac{1}{\binom{m+n}{n}} \mathbb{E} \left[\prod_{j\in[n]} h'(U_N^{(r_{m+j})}) \right],$$

and the locally most powerful test rejects the null hypothesis for large values of:

$$\mathbb{E} \left[\prod_{j\in[n]} h'(U_N^{(r_{m+j})}) \right].$$

Assuming that the function h depends on a real parameter θ and that the null hypothesis $H : Q = F$ is equivalent to test $H : \theta = 0$, the locally most powerful test among rank tests rejects the null hypothesis for large values of

$$\frac{d}{d\theta} \mathbb{E} \left[\prod_{j\in[n]} h'(U_N^{(r_{m+j})}) \right],$$

evaluated in $\theta = 0$. In the following sections, the local most powerful test for specific choices of h is derived.

2.3.1 Location-Shift Alternative

The location shift model can be obtained from (2.12) by setting

$$h(F(y)) = F(y - \theta) \tag{2.14}$$

for some $\theta > 0$, which can also be expressed in terms of $u = F(y)$ as:

$$h(u) = F(F^{-1}(u) - \theta), \qquad u \in [0, 1] \tag{2.15}$$

where $y = F^{-1}(u)$.

Computing the derivative $h'(u) = f(F^{-1}(u) - \theta)/f(F^{-1}(u))$ of h, assume that the density f is differentiable with respect to θ and denote by $\dot{f}$ its derivative with respect to θ. As a consequence, the derivative of the expectation with respect to θ becomes:

$$\begin{aligned}\frac{d}{d\theta}\mathbb{E}\left[\prod_{j\in[n]} h'(U_N^{(r_{m+j})})\right] &= \frac{d}{d\theta}\mathbb{E}\left[\prod_{j\in[n]} \frac{f\left(F^{-1}\left(U_N^{(r_{m+j})}\right) - \theta\right)}{f\left(F^{-1}\left(U_N^{(r_{m+j})}\right)\right)}\right] \\ &= \mathbb{E}\left[\sum_{j\in[n]} \frac{-\dot{f}\left(F\left(U_N^{(r_{m+j})}\right) - \theta\right)}{f\left(F^{-1}\left(U_N^{(r_{m+j})}\right)\right)} \prod_{k\neq j, k\in[n]} \frac{f\left(F^{-1}\left(U_N^{(r_{m+k})}\right) - \theta\right)}{f\left(F^{-1}\left(U_N^{(r_{m+k})}\right)\right)}\right],\end{aligned}$$

which evaluated in $\theta = 0$ is equal to:

$$\sum_{j\in[n]} \mathbb{E}\left[\frac{-\dot{f}\left(F\left(U_N^{(r_{m+j})}\right)\right)}{f\left(F^{-1}\left(U_N^{(r_{m+j})}\right)\right)}\right]. \tag{2.16}$$

Now, consider the logistic distribution with c.d.f. and its density

$$F(w) = \frac{1}{1 + e^w}, \qquad f(w) = \frac{e^w}{(1 + e^w)^2}$$

for $w \in \mathbb{R}$. The inverse of the c.d.f. with $v = F(w)$ and the derivative of the density are respectively:

$$F^{-1}(v) = \ln \frac{v}{1 - v}, \qquad \dot{f}(w) = \frac{e^{-w}(e^{-w} - 1)}{(1 + e^{-w})^3}.$$

Substituting these expressions into equation (2.16), the LMPI test that rejects the null hypothesis $H : \theta = 0$ for large values of the statistic in (2.16) becomes the Wilcoxon-Mann-Whitney test, as proved in Cox and Hinkley (1979).

2.3.2 Shiraishi's Alternatives

Consider the alternative distribution given by

$$h(F(y)) = (1-\theta)F(y) + \theta G(F(y)) \tag{2.17}$$

where $\theta > 0$ and G is a continuous c.d.f. with density g. By setting $y = F^{-1}(u)$, it follows:

$$h(u) = (1-\theta)u + \theta G(u), \qquad u \in [0,1], \tag{2.18}$$

which has derivative $h'(u) = (1-\theta) + \theta g(u)$. To derive the test statistic, the derivative of the expectation with respect to θ is computed:

$$\begin{aligned}\frac{d}{d\theta}\mathbb{E}\left[\prod_{j\in[n]} h'(U_N^{(r_{m+j})})\right] &= \frac{d}{d\theta}\mathbb{E}\left[\prod_{j\in[n]}\left[(1-\theta)+\theta g\left(U_N^{(r_{m+j})}\right)\right]\right] \\ &= \mathbb{E}\left[\sum_{j\in[n]}\left[-1+g\left(U_N^{(r_{m+j})}\right)\right]\prod_{k\neq j,k\in[n]}\left[(1-\theta)+\theta g\left(U_N^{(r_{m+j})}\right)\right]\right]\end{aligned}$$

which evaluated in $\theta = 0$ is equal to:

$$-n + \sum_{j\in[n]}\mathbb{E}\left[g\left(U_N^{(r_{m+j})}\right)\right]. \tag{2.19}$$

As a consequence, when considering the model

$$X_1,\ldots,X_m \overset{iid}{\sim} F, \qquad Y_1,\ldots,Y_n \overset{iid}{\sim} (1-\theta)F + \theta G(F) \tag{2.20}$$

with $\theta \in [0,1]$, corresponding to the choice of h in (2.17), the LMPI test rejects the null hypothesis $H: \theta = 0$ for large values of the statistic:

$$T^g = \sum_{j\in[n]} \mathbb{E}\left[g(U_N^{(R_{m+j})})\right], \tag{2.21}$$

where $U_N^{(R_{m+j})}$ is the R_{m+j}th ordered statistic in a vector of $N = m+n$ standard uniform random variables. This statistic is called the *Shiraishi statistic* (Shiraishi, 1985) and the optimality result is stated in the following theorem that is proved in Shiraishi (1985).

Theorem 1 (Shiraishi (1985))**.** *Assume the model in* (2.20) *has continuous c.d.f.'s F and G with density g. Then, the LMPI test rejects $H: \theta = 0$ for large values of:*

$$T^{\mathrm{g}} = \sum_{j\in[n]} \mathbb{E}[g(U_N^{(R_{m+j})})].$$

Note that, for very small values of θ, the model in (2.17) is the nonparametric generalization of the parametric rare/weak effects model that motivates the Higher Criticism test of Donoho and Jin (2004, 2015) (see Section 3.4.2.4).

Analogous to the Wilcoxon-Mann-Whitney statistic, the finite-sample distribution of the Shiraishi test statistic can be derived using the same recurrence method, as in Section 2.2.1. However, the numerator $\pi_{n,m}(k)$ depends on the density g, so the recurrence formula varies with the specific choice of the g.

The approach in Section 2.2.1 is adopted to derive the permutation distribution. For any $u \in \mathbb{R}$, the permutation distribution of the Shiraishi statistic is given by:

$$\Pr\left[T^g = u\right] = \frac{\sum_{h\in\mathcal{C}_{m+n,n}} \mathbb{1}[T^g\left(R_{m+h_1},\ldots,R_{m+h_n}\right) = u]}{\binom{m+n}{n}}, \tag{2.22}$$

where $h = (h_1,\ldots,h_n)$ is an element of $C_{m+n,n}$ which is the set of combinations of n elements from the set $[m+n]$ and the numerator depends on g.

2.3.2.1 Lehmann's Alternatives

Lehmann (1953) introduces the *Lehmann's alternatives*, a family of non-parametric alter-

natives that is a special case of the Shiraishi alternatives corresponding to

$$G(F) = F^k,$$

for integers $k > 1$. This model assumes that the outliers have weak signals, as for any $k > 1$, F^k is the distribution of the maximum of k independent random variables distributed as F. Under this alternative, Y is stochastically larger than X with $\Pr[X < Y] = k/(k+1)$. Lehmann demonstrates that for $k = 2$, the Wilcoxon-Mann-Whitney test is locally most powerful invariant test under the mixture model:

$$Y_1, \ldots, Y_n \overset{iid}{\sim} (1-\theta)F + \theta F^2. \tag{2.23}$$

This result can be obtained as a particular case of the more general result proved for the Shiraishi test, setting $G(F) = F^2$ in equation (2.17).

More generally, under the Lehmann's alternative $G(F) = F^k$, for integers $k \geq 2$, the associated density function is $g(u) = ku^{k-1}\mathbb{1}[u \in [0,1]]$ and the test statistics in (2.21) reduces to:

$$T^{\text{WMW, k}} = \frac{k}{(N+1)\cdot \ldots \cdot (N+k-1)} \sum_{j \in [n]} \prod_{l=0}^{k-2} (R_{m+j} + l), \tag{2.24}$$

which are the locally most powerful rank tests against the class of Lehmann's alternatives. While the Wilcoxon-Mann-Whitney tests do not require stochastically larger alternative distributions for validity, this property ensures that the tests are unbiased (Lehmann, 1951).

2.4 Asymptotic Null Distribution of Linear Rank Statistics

Knowing the null distribution of test statistics is essential for computing critical values and performing hypothesis tests. However, finite-sample distributions are often unavailable in closed form, especially in a distribution-free settings where the underlying population distributions are unknown. In such situations, one can either use the tabulated values of the statistic's null distribution for small sample sizes or leverage the asymptotic distribution of the test statistic.

Hájek et al. (1999) and van der Vaart (2000) provide a general result about the asymptotic normal distribution of *linear rank statistics*.

Given a vector $W_1, \ldots, W_N$ of i.i.d. random variables with ranks $R_1, \ldots, R_N$, a linear rank statistic can be expressed as

$$T^{\mathrm{L}} = \sum_{i \in [N]} c_i a_i(R_i), \qquad c_i, a_i \in \mathbb{R} \tag{2.25}$$

where $(c_i)_{i \in [N]}$ are called the coefficients and $(a_i)_{i \in [N]}$ are called the scores, with a_i being a function of the rank R_i for each $i \in [N]$. The expectation and variance of a linear rank statistic of the form (2.25) are, respectively,

$$\mathbb{E}\left[T^{\mathrm{L}}\right] = N\bar{c}\bar{a}, \qquad \mathrm{Var}\left(T^{\mathrm{L}}\right) = \frac{1}{N-1} \sum_{i \in [N]} (c_i - \bar{c})^2 \sum_{i \in [N]} (a_i - \bar{a})^2 \tag{2.26}$$

with $\bar{c} = N^{-1} \sum_{i \in [N]} c_i$ and $\bar{a} = N^{-1} \sum_{i \in [N]} a_i$.

To establish the asymptotic behavior, assume that the scores are generated by a function $\phi : [0, 1] \to \mathbb{R}$ of the form:

$$a_i = \mathbb{E}\left[\phi(U_N^{(i)})\right] \tag{2.27}$$

with $U_N^{(i)}$ the ith ordered statistic in a vector of N standard uniform random variables, or of the form:

$$a_i = \phi\left(\frac{i}{N+1}\right). \tag{2.28}$$

Note that when ϕ is the identity function, the two definitions (2.27) and (2.28) coincide. The coefficients are required to satisfy the following conditions:

$$\frac{\max_{i \in [N]} (c_i - \bar{c})^2}{\sum_{i \in [N]} (c_i - \bar{c})^2} \underset{N \to \infty}{\longrightarrow} 0 \tag{2.29}$$

Then, the asymptotic behavior of the linear rank statistics in (2.25) is formalized in the following theorem. Throughout, the notation $N \to \infty$ will be used to indicate $\min\{m, n\} \to \infty$.

Theorem 2 (Corollary 13.8, van der Vaart (2000))**.** *Consider a vector $W_1, \ldots, W_N$ of i.i.d. random variables with ranks $R_1, \ldots, R_N$ and a test statistic of the form in* (2.25)*,*

with coefficients c_i satisfying the condition in (2.29) *and scores as in* (2.27) *where ϕ is a measurable, nonconstant function such that*

$$\int_0^1 \phi^2(u)\, du < \infty. \tag{2.30}$$

Then, as $N \to \infty$, the standardized rank statistic

$$\frac{T^{\mathrm{L}} - \mathbb{E}[T^{\mathrm{L}}]}{\sqrt{\mathrm{Var}(T^{\mathrm{L}})}}$$

converges in distribution to the standard normal distribution, where the expectation and variance are given in (2.26). *The result also holds when considering scores as in* (2.28), *where ϕ is a continuous function almost everywhere, that is nonconstant and satisfies:*

$$N^{-1} \sum_{i \in [N]} \phi^2 \left(\frac{i}{N+1} \right) \underset{N \to \infty}{\longrightarrow} \int_0^1 \phi^2(u)\, du. \tag{2.31}$$

It is worth noting that asymptotic null distribution of a linear rank statistic generally depends on ϕ. In the case of the Wilcoxon-Mann-Whitney test, when ϕ is the identity function multiplied by $N+1$ in equation (2.28), the asymptotic null distribution depends only on the sample sizes n and m, as already mentioned in Section 2.2.1.

2.4.1 Asymptotic Null Distributions of the Wilcoxon-Mann-Whitney and Shiraishi Statistics

Let us consider the Wilcoxon-Mann-Whitney statistic in equation (2.3) and the Shiraishi statistic in equation (2.21). The pooled sample $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ is denoted by $W_1, \ldots, W_N$, with $N = m+n$. Both the Wilcoxon-Mann-Whitney and the Shiraishi statistic can be expressed as linear statistics by choosing appropriate coefficients and scores. The Wilcoxon-Mann-Whitney statistic is retrieved by using the coefficients:

$$c_i = \begin{cases} 0, & i \in \{1, \ldots, m\} \\ N+1, & i \in \{m+1, \ldots, N\} \end{cases} \tag{2.32}$$

and the following scores in the form of (2.28):

$$a_i(R_i) = \phi\left(\frac{R_i}{N+1}\right) = \frac{R_i}{N+1}, \qquad i \in [N]. \tag{2.33}$$

The Shiraishi statistic can be obtained with the following coefficients:

$$c_i = \begin{cases} 0, & i \in \{1, \ldots, m\} \\ 1, & i \in \{m+1, \ldots, N\} \end{cases} \tag{2.34}$$

and the following scores in the form of (2.27):

$$a_i(R_i) = \mathbb{E}\left[g(U_N^{(R_i)})\right], \qquad i \in [N]. \tag{2.35}$$

While the asymptotic behavior under the null hypothesis can be formally derived using Corollary 13.8 in van der Vaart (2000), these asymptotic properties are already well-established in the statistical literature. Mann and Whitney (1947a) proved that the standardized Mann-Whitney U-statistic

$$\frac{T^{\mathrm{MW}} - mn/2}{\sqrt{mn(N+1)/12}}$$

has asymptotically standard normal distribution as $m, n \to \infty$. As a consequence, it is easy to verify that the standardized Wilcoxon statistic

$$\frac{T^{\mathrm{W}} - n(N+1)/2}{\sqrt{mn(N+1)/12}}$$

has asymptotically standard normal distribution as $N \to \infty$.

In his paper, Shiraishi (1985) proved that under the assumptions that the density g is bounded and that $\delta := \lim_{N\to\infty} n/N$ exists with $0 < \delta < 1$, the standardized Shiraishi statistic

$$\frac{\sqrt{N(N-1)}(T^g - n\mu_N)}{\sqrt{nm\sum_{h\in[N]}(\mathbb{E}[g(U_N^{(h)})] - \mu_N)^2}} \tag{2.36}$$

with $\mu_N = N^{-1}\sum_{h\in[N]}\mathbb{E}[g(U_N^{(h)})]$ converges in distribution under the null hypothesis that

$H : \theta = 0$ to a standard normal random variable as $N \to \infty$.

It is worth noting that the requirement that the density g is bounded is stronger than the assumption in Corollary 13.8 in van der Vaart (2000), which only requires:

$$\int |g(u)|^2 \, du < \infty. \tag{2.37}$$

Indeed, $g(u) = 2/(3u^{1/3})$ is a density on $[0, 1]$, that satisfies (2.37) but that is unbounded in a (right) neighborhood of 0. This stronger assumption is necessary to prove the asymptotic normality of the Shiraishi statistic under a specific class of local alternatives, as described in Section 2.5.

To derive the asymptotic behavior of the Wilcoxon-Mann-Whitney and Shiraishi statistics from the Corollary 13.8 in van der Vaart (2000), the assumptions on the coefficients c_i and on the scores a_i need to be satisfied.

For the Wilcoxon-Mann-Whitney statistic, the function $\phi(u) = u, \ u \in [0, 1]$ used in (2.33) is continuous and non-constant, and it satisfies (2.31). This can be verified by observing that:

$$N^{-1} \sum_{i \in [N]} \left(\frac{i}{N+1} \right)^2 = N^{-1}(N+1)^{-2} \sum_{i \in [N]} i^2 = \frac{N(N+1)(2N+1)}{6N(N+1)^2} \underset{N \to \infty}{\longrightarrow} \frac{1}{3}$$

and

$$\int_0^1 u^2 \, du = \frac{1}{3}.$$

The coefficients in (2.32) satisfy the condition in (2.29). Their mean is $\bar{c} = n(N+1)$ and

$$\max_{i \in [N]} (c_i - \bar{c})^2 = n^2(N+1)^2,$$

while

$$\begin{aligned} \sum_{i \in [N]} (c_i - \bar{c})^2 &= mn^2(N+1)^2 + n(n-1)^2(N+1)^2 \\ &= (N+1)^2 n[mn + (n-1)^2]. \end{aligned}$$

Therefore, the required ratio converges to zero:

$$\frac{\max_{i\in[N]}(c_i-\bar{c})^2}{\sum_{i\in[N]}(c_i-\bar{c})^2}=\frac{n^2(N+1)^2}{(N+1)^2n[mn+(n-1)^2]}=\frac{n}{(m+1)n-1}\underset{N\to+\infty}{\longrightarrow}0.$$

For the Shiraishi statistic, the function $\phi(u)=g(u),\ u\in[0,1]$ used in (2.33) is measurable since it is a density function. Two additional conditions on g are required: (i) g is non-constant (excluding only the uniform distribution), and (ii) $\int_0^1 g^2(u)\ du<\infty$. Both conditions require explicit verification depending on the chosen density function. The coefficients for this case have $\bar{c}=n/N$ and

$$\max_{i\in[N]}(c_i-\bar{c})^2=\left(\frac{n}{N}\right)^2,\qquad\sum_{i\in[N]}(c_i-\bar{c})^2m\left(\frac{n}{N}\right)^2+n\left(1-\frac{n}{N}\right)^2=\frac{mn}{N}.$$

Therefore, the required ratio converges to zero:

$$\frac{\max_{i\in[N]}(c_i-\bar{c})^2}{\sum_{i\in[N]}(c_i-\bar{c})^2}=\frac{n^2}{N^2}\cdot\frac{N}{mn}=\frac{n}{mN}\underset{N\to+\infty}{\longrightarrow}0.$$

2.5 Asymptotic Distribution of the Shiraishi statistics under Local Alternatives

In this section, the asymptotic behavior of the Wilcoxon-Mann-Whitney and Shiraishi statistics is studied in a neighborhood of the null hypothesis. Hence, the following class of alternatives close to the null hypothesis is considered:

$$K_N:\theta=\bar{\theta}_N>0,\qquad\bar{\theta}_N=\bar{\theta}/\sqrt{N},\tag{2.38}$$

with $\bar{\theta}>0$. Under this sequence of alternatives, Shiraishi (1985) establishes the asymptotic distribution of the Shiraishi statistic under the assumption that g is bounded, as stated in the following theorem.

Theorem 3 (Shiraishi (1985))**.** *Consider the model in* (2.20) *and the statistic* T^g *defined in* (2.21) *for testing* $H:\theta=0$ *against the sequence of alternatives given in* (2.38)*. Assume that the density* g *is bounded and* $\lim_{N\to\infty}n/N=\delta$ *with* $0<\delta<1$*. Then, the*

standardized Shiraishi statistic in (2.36) *has asymptotically normal distribution with mean* $\bar{\theta}\sqrt{\text{Var}(g(U))\delta(1-\delta)}$ *and variance* 1*, where* U *is uniformly distributed on* $(0,1)$*. The null distribution is obtained by setting* $\bar{\theta}=0$*.*

The asymptotic distribution of the Wilcoxon-Mann-Whitney statistic under the local alternatives in equation (2.38) can be derived from the previous theorem by setting $g(u) = (N+1)u$ with $u \in [0,1]$. With this choice of g the Shiraishi statistic in equation (2.21) reduces to the Wilcoxon-Mann-Whitney statistic in equation (2.3). This is due to the fact that $U_N^{(R_{m+j})}$ is distributed as a Beta$(R_{m+j}, N - R_{m+j} + 1)$ with expected value equal to $R_{m+j}/(N+1)$.

An important observation must be made regarding the validity of the Shiraishi test. Consider the model in equation (2.20), where g represents the density of the c.d.f. G in the two-component mixture model. Crucially, the Shiraishi test defined in equation (2.21) remains valid for any choice of function g, regardless of whether it corresponds to the true density in the two-component mixture model. This validity holds because both the Shiraishi statistic, its distribution and critical values depend on the chosen function g. Consequently, even if the Shiraishi statistic is constructing using a function g that does not match the density of the c.d.f. G, the test will maintain its validity, although its power may be affected. This result is established in Theorem 4, which is a generalization of Theorem 3 and is proved in Shiraishi (1985). To state the theorem, a necessary condition is required. Consider a linear rank statistic of the form:

$$T_* = \sum_{j\in[n]} a_N(R_{m+j}), \tag{2.39}$$

with score function $a_N : \mathbb{N} \to \mathbb{R}$. The following conditions are imposed on the scores.

Condition 1. There exists a square integrable function $\psi : [0,1] \to \mathbb{R}$ with $\int_0^1 \left[\psi(u) - \int_0^1 \psi(v)\, dv\right]^2 du > 0$ such that:

$$\lim_{N\to\infty} \int_0^1 \left[a_N(1 + \lfloor uN \rfloor) - \psi(u)\right]^2 = 0,$$

where $\lfloor u \rfloor$ denotes the largest integer not exceeding u.

Theorem 4 (Shiraishi (1985))**.** *Consider the model in* (2.20) *and assume that the density* g *is bounded and* $\lim_{N\to\infty} n/N = \delta$ *with* $0 < \delta < 1$*. Consider the linear rank statistic in* (2.39) *for testing* $H : \theta = 0$ *against the sequence of alternatives given in* (2.38)*. Then, under Condition 1 the standardized statistic*

$$\frac{\sqrt{N(N-1)}(T_* - n\bar{a}_N)}{\sqrt{mn\sum_{i\in[N]}(a_N(i) - \bar{a}_N)^2}}$$

with $\bar{a}_N = \sum_{i\in[N]} a_N(i)/N$ *has asymptotically normal distribution with mean*

$$\frac{\bar{\theta}\sqrt{\delta(1-\delta)}\mathrm{Cov}(\psi(U), g(U))}{\sqrt{\mathrm{Var}(\psi(U))}}$$

and variance 1*, where* U *is uniformly distributed on* $(0, 1)$*. The null distribution is obtained by setting* $\bar{\theta} = 0$*.*

2.5.1 Asymptotic Power Function against Local Alternatives

Consider the model in equation (2.20) and denote the densities of the c.d.f.'s F and G by f and g, respectively. For the validity of the asymptotic null distribution of the Shiraishi statistic in Corollary 13.8 of van der Vaart (2000), the density g needs to satisfy the condition in equation (2.37). Under these conditions, Shiraishi (1985) establishes that the power of the Shiraishi test against the sequence of local alternatives in equation (2.38) is:

$$\mathrm{Pow}(T^g; \alpha, \delta, \bar{\theta}, g) = 1 - \Phi(z_\alpha - \bar{\theta}[\delta(1-\delta)\mathrm{Var}\{g(U)\}]^{1/2}) \tag{2.40}$$

where Φ and z_α are the c.d.f. and the $(1-\alpha)$-quantile of the standard normal distribution, respectively, U is a standard uniform random variable and $\delta = \lim_{N\to\infty} n/N$, with $N = n + m$ being the total sample size. Setting $\bar{\theta} = 0$ retrieves the power function under the null hypothesis. Note that condition (2.37) ensures that the variance of $g(U)$ is finite.

The power function of the Shiraishi test increases with $\bar{\theta}[\mathrm{Var}(g(U))\delta(1-\delta)]^{1/2}$, reaching its maximum value when $\delta = 1/2$. Therefore, to maximize the asymptotic power of the Shiraishi test, the two sample sizes m and n must grow at the same asymptotic rate, which for fixed N requires $m = n$.

By setting the density function g equal to $g^{\mathrm{WMW}}(u) = (N+1)u$ in equation (2.40), the power function of the Wilcoxon-Mann-Whitney test is:

$$\mathrm{Pow}(T^{\mathrm{WMW}}; \alpha, \delta, \bar{\theta}, g^{\mathrm{WMW}}) = 1 - \Phi(z_\alpha - (N+1)\bar{\theta}[\delta(1-\delta)1/12]^{1/2}). \tag{2.41}$$

Section 5.3.4.2 provides a detailed comparison between the Shiraishi statistic and an (idealized) *oracle Neyman-Pearson test* in terms of their local power functions. This analysis offers further theoretical support for ACODE by examining the optimality properties of the rank tests that it employs.

2.6 Asymptotic Relative Efficiency under Local Alternatives

When the sample size is large, any reasonable test should easily distinguish between the null and alternative hypotheses, especially when they differ substantially. Therefore, comparing tests with large samples requires more nuanced approaches. One option is to compare tests locally—i.e., for alternatives close enough to the null hypothesis to make the testing problem more challenging. Specifically, the null hypothesis $H : \theta = 0$ is tested against the sequence of local alternatives in equation (2.38).

A classic method for comparing two tests locally is evaluating their Asymptotic Relative Efficiency (ARE), first introduced by Pitman (1949). This quantity measures the ratio of sample sizes required by two tests to achieve the same power against a sequence of local alternatives, while maintaining type-I error control. Definitions and results are presented for the broader class of tests based on *locally uniformly asymptotically normal statistics* (not necessarily rank statistics), including specific results for the Wilcoxon-Mann-Whitney and the Shiraishi tests.

Consider a sample $W_1, \ldots, W_N$ with joint c.d.f. $Q_{N,\theta}$ with $\theta \geq 0$, and the null hypothesis $H : \theta = 0$ against the local alternatives in equation (2.38). To define the asymptotic relative efficiency, consider two different test statistics $\psi_{1,N}$ and $\psi_{2,N}$ built on N observations, with respective power functions $\beta_{h,N}(\theta)$, $h = 1, 2$. Fix a significance level $\alpha \in (0, 1)$ and a constant $\gamma \in (\alpha, 1)$, and define N_h as the minimal number of observations satisfying:

$$\beta_{h,N_h}(0) \leq \alpha, \qquad \beta_{h,N_h}(\bar{\theta}_N) \geq \gamma. \tag{2.42}$$

The asymptotic relative efficiency of the tests that reject the null hypothesis for large values of $\psi_{h,N}$ is defined as:

$$\text{ARE}\left(\psi_{1,N}, \psi_{2,N}\right) = \lim_{N\to\infty} \frac{N_2}{N_1}.$$

When $\text{ARE}\left(\psi_{1,N}, \psi_{2,N}\right) > 1$, the first sequence of tests achieves the conditions in (2.42) with fewer observations, indicating superior efficiency.

In order to establish an alternative formulation for the asymptotic relative efficiency, based on the asymptotic local power, the two tests $\psi_{h,N}$, $h = 1, 2$ are required to be *locally uniformly asymptotically normal*. That is, for all sequences $\bar{\theta}_N$ they satisfy:

$$\frac{\sqrt{N}(\psi_{h,N} - \mu_i(\bar{\theta}_N))}{\sigma_h(\bar{\theta}_N)} \xrightarrow[N\to\infty]{d} N(0,1). \tag{2.43}$$

Assuming that μ_h is differentiable at zero and σ_h is continuous at zero, with $\mu_h'(0) > 0$ and $\sigma_h(0) > 0$, the asymptotic power of the test that reject the null hypothesis for large values of $\psi_{h,N}$ can be written as:

$$\pi_{h,N}(\bar{\theta}_N) = 1 - \Phi\left(z_\alpha + o(1) - \sqrt{N}\theta_N \frac{\mu_h'(0)}{\sigma_h(0)}(1 + o(1))\right) + o(1), \tag{2.44}$$

where z_α is the $(1-\alpha)$-quantile of the standard normal distribution. Under the assumptions (2.43) and (2.44), the ARE of the two test statistics can be expressed as stated in the following theorem.

Theorem 5 (Theorem 14.19 in van der Vaart (2000))**.** *Consider the statistical models* $\{Q_{N,\theta} : \theta \geq 0\}$ *such that for any fixed* N *the total variation distance* $||Q_{N,\bar{\theta}} - Q_{N,0}||_{TV} \to 0$ *as* $\bar{\theta} \to 0$, $\bar{\theta} > 0$. *Let* $\psi_{1,N}$ *and* $\psi_{2,N}$ *be two sequences of statistics that satisfy* (2.43) *for every sequence* $\bar{\theta}_N$ *decreasing monotonically to zero, and functions* μ_h *and* σ_h *such that* μ_h *is differentiable at zero and* σ_h *is continuous at zero, with* $\mu_h'(0) > 0$ *and* $\sigma_h(0) > 0$. *Then, the ARE of the tests that rejects the null hypothesis* $H : \theta = 0$ *for large values of* $\psi_{i,N}$, $i = 1, 2$ *is equal to*

$$\left(\frac{\mu_1'(0)/\sigma_1(0)}{\mu_2'(0)/\sigma_2(0)}\right)^2,$$

for every sequence of alternatives $\bar{\theta}_N$ *decreasing monotonically to zero, independently of* $\alpha > 0$ *and* $\gamma \in (\alpha, 1)$.

The following corollary states that in the setting of Theorem 5 two tests have the same asymptotic power if their asymptotic relative efficiency is equal to 1.

Corollary 1. *Under the same assumptions of Theorem 5, assume that the* $\mathrm{ARE}(\psi_{1,N}, \psi_{2,N})$ *equals 1. Then,* $\psi_{1,N}$ *and* $\psi_{2,N}$ *have the same asymptotic power functions, i.e.,* $\pi_{1,N}(\bar{\theta}_N) = \pi_{2,N}(\bar{\theta}_N)$ *for all sequences* $\bar{\theta}_N = \bar{\theta}/\sqrt{N}$, $\bar{\theta} > 0$.

Proof. If $\mathrm{ARE}(\psi_{1,N}, \psi_{2,N}) = 1$, by Theorem 5 it follows that $\mu_1'(0)/\sigma_1(0) = \mu_2'(0)/\sigma_2(0)$, and from Equation (2.44) the thesis follows. □

2.6.1 Performance of the Wilcoxon-Mann-Whitney Test via Asymptotic Relative Efficiency

The results in this section are known from classical literature on rank-based tests. Key contributions include Pitman (1949); Hodges and Lehmann (1956); Chernoff and Savage (1958); Hodges and Lehmann (1961), which compared the Wilcoxon-Mann-Whitney test with other other nonparametric and parametric alternatives.

Consider the location shift model introduced in Section 2.3.1. For the distribution $F(\cdot - \theta)$, test the null hypothesis $H : \theta = 0$ against the alternative $K : \theta > 0$. For this problem, the Wilcoxon-Mann-Whitney test is compared with the following alternative tests: the normal scores test Fisher and Yates (1938) and the t-test Student (1908).

The normal scores is a rank-based test rejects the null hypothesis for large values of:

$$T^{\mathrm{NS}} = \sum_{j \in [n]} \mathbb{E}\left[Z_N^{(R_{m+j})}\right], \tag{2.45}$$

where R_{m+j} is the rank of Y_j in the pooled vector $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ for any $j \in [n]$ and $Z_N^{(R_{m+j})}$ is the R_{m+j}th order statistic from $N = m + n$ independent standard normal random variables.

Under the assumption of equal variances between the two samples, the t-test test rejects the null hypothesis for large values of:

$$t = \frac{\left|\bar{X} - \bar{Y}\right|}{\sigma_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \tag{2.46}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, respectively, and σ_p is the pooled sample standard deviation of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$:

$$\sigma_p = \sqrt{\frac{(m-1)\sigma_X^2 + (n-1)\sigma_Y^2}{m+n-2}}, \tag{2.47}$$

where σ_X^2 and σ_Y^2 are the sample variances:

$$\sigma_X^2 = \frac{\sum_{i\in[m]}(X_i - \bar{X})^2}{m-1}, \qquad \sigma_Y^2 = \frac{\sum_{j\in[n]}(Y_j - \bar{Y})^2}{n-1}.$$

From the Neyman-Pearson lemma, the t-test is the uniformly most powerful test when both samples follow normal distributions with equal variances.

Comparing the Wilcoxon-Mann-Whitney test with the t-test and the normal scores test, Pitman (1949) established that when F has density f on $\mathbb{R}$ and variance σ^2, the asymptotic relative efficiency of the Wilcoxon-Mann-Whitney test with respect to the t-test is given by:

$$\operatorname{ARE}\left(T^{\mathrm{WMW}}, t\right) = 12\sigma^2 \left(\int_{\mathbb{R}} f^2(w)\, dw\right)^2. \tag{2.48}$$

This general formula reveals several important properties of the Wilcoxon-Mann-Whitney test's performance. For instance, it is clear that $\operatorname{ARE}\left(T^{\mathrm{WMW}}, t\right)$ can be arbitrarily large, indicating that the Wilcoxon-Mann-Whitney test can be substantially more efficient than the t-test for certain distributions. This occurs when the density f does not belong to:

$$L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{L}) = \left\{ f : \mathbb{R} \to \mathbb{R} : \int_{\mathbb{R}} f^2(w) dw < +\infty \right\},$$

which is the space of square-integrable functions on $\mathbb{R}$ with respect to the Lebesgue measure $\mathcal{L}$ and the Borelian σ-algebra $\mathcal{B}(\mathbb{R})$. Moreover, Hodges and Lehmann (1956) proved that $\operatorname{ARE}\left(T^{\mathrm{WMW}}, t\right) \geq 0.864$ for any distribution F. This lower bound ensures that the Wilcoxon-Mann-Whitney test cannot perform drastically worse than the t-test, even in unfavorable scenarios.

When F is the normal distribution—the setting where the t-test is uniformly most powerful—Pitman (1949) showed that $\operatorname{ARE}\left(T^{\mathrm{WMW}}, t\right) = 3/\pi \approx 0.955$. This result demonstrates that even when the t-test is optimal, the Wilcoxon-Mann-Whitney test retains

approximately 95.5% of its efficiency.

Table 2.1 presents asymptotic relative efficiency values for various specific distributions (Lehmann, 2009). These results consistently show that the Wilcoxon-Mann-Whitney test is preferable to the parametric t-test across different scenarios, with the two tests achieving equivalent performance only for the uniform distribution.

F	Logistic	Double exponential	Exponential	Normal	Uniform
ARE $\left(T^{\mathrm{WMW}}, t\right)$	1.097	1.5	3.0	$3/\pi$	1.0

Table 2.1: Asymptotic relative efficiency values of the Wilcoxon-Mann-Whitney test with respect to the t-test for some specific distributions (Lehmann, 2009).

The relative performance of the Wilcoxon-Mann-Whitney test versus the normal scores test is summarized in Table 2.2. The results in Table 2.2 reveal a clear pattern: the normal scores test is generally preferred for distributions with light tails, while the Wilcoxon-Mann-Whitney test becomes more efficient when dealing with heavy-tailed distributions.

F	Logistic	Double exponential	Cauchy	Normal	Uniform
ARE $\left(T^{\mathrm{WMW}}, T^{\mathrm{NS}}\right)$	1.05	1.18	1.41	0.955	0

Table 2.2: Asymptotic relative efficiency values of the Wilcoxon-Mann-Whitney test with respect to the normal scores test for some specific distributions (Lehmann, 2009).

Moreover, Hodges and Lehmann (1961) established that $\mathrm{ARE}(T^{\mathrm{WMW}}, T^{\mathrm{NS}}) \leq 6/\pi \approx 1.91$ for any distribution F. This upper bound ensures that the Wilcoxon-Mann-Whitney test cannot infinitely outperform the normal scores test, providing a limit on the relative efficiency advantage.

Finally, it is worth noting the superior general performance of the normal scores test compared to the parametric t-test. Chernoff and Savage (1958) proved that $\mathrm{ARE}(T^{\mathrm{NS}}, t) \geq 1$ for any distribution F, with equality achieved only when F is the normal distribution. This result demonstrates that the normal scores test always outperforms or matches the t-test in terms of asymptotic efficiency, making it a uniformly superior nonparametric alternative to the t-test.

These results provide clear guidance for test selection. Under normality assumptions, the t-test remains the optimal choice, as expected from the Neyman-Pearson lemma. However, when normality cannot be assumed, distribution-free rank tests are preferable. Specifically, for heavy-tailed distributions (such as Cauchy or double exponential), the Wilcoxon-Mann-Whitney test offers superior efficiency due to its robustness against outliers. For light-tailed distributions (with tails similar to or lighter than the normal distribution), the normal scores test provides better performance while maintaining nonparametric flexibility.

On another note, while the Wilcoxon-Mann-Whitney test may not always be the most efficient based on the asymptotic relative efficiency values, it provides the practical advantage of intuitive interpretation, making it accessible for practitioners without extensive statistical training (Lehmann, 2009).

2.6.2 Performance of the Shiraishi Test via Asymptotic Relative Efficiency

This section discusses the asymptotic relative efficiency of the Shiraishi test—the locally most powerful rank test under the model (2.20)—with respect to the global most powerful likelihood ratio test given by the Neyman-Pearson lemma:

$$\rho(Y_1, \ldots, Y_n) = \frac{\prod_{j\in[n]} \left[(1-\bar{\theta})f(Y_j) + \bar{\theta}g(F(Y_j))f(Y_j)\right]}{\prod_{j\in[n]} f(Y_j)}. \tag{2.49}$$

Shiraishi (1985) proved that the asymptotic relative efficiency between the Shiraishi test and the Neyman-Pearson test is arbitrarily close to 1 as the proportion n/N of test observations becomes small.

Theorem 6 (Shiraishi (1985))**.** *Consider the model* (2.20) *with bounded outlier density* g *and assume* $0 < \delta = \lim_{N\to\infty} n/N < 1$*. Then, the likelihood ratio test* ρ *in* (2.49) *and the Shiraishi test* T^g *in* (2.21) *satisfy the locally uniform asymptotical normality* (2.43) *and the* $\mathrm{ARE}(T^g, \rho)$ *equals* $1 - \delta$*.*

The quantity δ represents the efficiency loss from using a rank-based test instead of the most powerful parametric test. This loss is often acceptable given that rank tests remain applicable even when the underlying distributions are unknown, although at the cost of

discarding information from the exact observation values.

Note that while the Shiraishi test can never achieve the same Pitman efficiency as the Neyman-Pearson test as $\text{ARE}(T^g, \rho) < 1$, it can approach the Neyman-Pearson test's efficiency arbitrarily closely for arbitrarily small values of δ. Another interesting observation regards the power of the Shiraishi test which is maximum for $\delta = 1/2$, as discussed in Section (2.5.1). This creates a trade-off: the value of δ that maximizes asymptotic power does not maximize efficiency which requires $\delta \to 0$.

Chapter 3

Multiple Testing

3.1 Introduction and Classic Setup

This chapter discusses the challenge of multiplicity correction. When conducting statistical analyses involving multiple comparisons, a correction for multiplicity needs to be applied in order to control the type-I error rate at the desired significance level. In the classic setup multiple hypotheses $H_1, \ldots, H_n$ are tested using valid p-values that for any $u \in (0, 1)$ ensure:

$$\Pr[p_j \leq u \; ; \; H_j] \leq u, \qquad \forall \, j \in [n].$$

However, performing multiple tests with more than one true null yields an inflation of type-I errors. To see this, let $I_0 \subseteq [n]$ denote the (unknown) subset of indices corresponding to true null hypotheses with cardinality $|I_0| = n_0$. Under the assumption that $n_0 > 1$, the expected number of type-I errors is greater than the prespecified significance level α:

$$\sum_{j \in I_0} \Pr[p_j \leq \alpha] = n_0 \alpha > \alpha, \qquad \forall n_0 > 1. \tag{3.1}$$

For this reason, the need for multiplicity correction arises.

3.2 Error Rates

Let $I_1 = [n] \setminus I_0$ denote the subset of indices corresponding to true null hypotheses, and $n_1 = n - n_0$ the number of false hypotheses. Denote as $R \subseteq [n]$ the index set of the rejected

hypotheses (also called *discoveries*). For a fixed significance level $\alpha \in (0,1)$, the rejection set is obtained by rejecting the null hypotheses corresponding to p-values smaller than a suitable critical value c:

$$R = \{j \in [n] : p_j \leq c\}.$$

Denote as $R \cap I_0$ the index set of false discoveries.

Depending on the objective of the inference, different error measures are appropriate. One possibility is controlling the quantiles of the *false discovery proportion* (FDP), defined as the proportion of type-I errors among all rejections:

$$\mathrm{FDP}(R) = \frac{|R \cap I_0|}{|R| \vee 1}, \tag{3.2}$$

where $\vee$ denotes the maximum between two numbers.

A more stringent approach aims to control the *family-wise error rate* (FWER), defined as the probability of making even one false rejection:

$$\mathrm{FWER}(R) = \Pr[|R \cap I_0| > 0]. \tag{3.3}$$

Note that it is equivalent to control the probability $\Pr[\mathrm{FDP}(R) > 0]$.

A more flexible approach aims to control the the *false discovery rate* (FDR), defined as the expectation of the false discovery proportion:

$$\mathrm{FDR}(R) = \mathbb{E}[\mathrm{FDP}(R)]. \tag{3.4}$$

A multiple testing procedure achieves FWER control at level $\alpha \in (0,1)$ if $\mathrm{FWER}(R) \leq \alpha$. Similarly, a procedure achieves FDR control at level α if $\mathrm{FDR}(R) \leq \alpha$. Note that FWER controlling procedures automatically ensure FDR control. To see this, note that $0 \leq \mathrm{FDP} \leq 1$ and $\mathrm{FDP}(R) \leq \mathbb{1}[\mathrm{FDP}(R) > 0]$. Taking the expectations on both sides, we have that

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}(R)] \leq \Pr[\mathrm{FDP}(R) > 0] = \mathrm{FWER}, \tag{3.5}$$

proving that FDR control is less stringent than FWER control. This reduced stringency translates into improved statistical power, particularly when testing a large number of

hypotheses. Power gains are further amplified when the proportion of false null hypotheses increases.

Note that if $n_0 = n$, $\text{FDP}(R) = \mathbb{1}[\text{FDP}(R) > 0]$ and the FDP is distributed as a Bernoulli random variable, which implies $\text{FDP}(R) = \text{FWER}(R)$. When $n = 1$, the equality $\text{FDP}(R) = \text{FWER}(R)$ still holds, and both reduce to the type-I error probability in Section 3.1.

3.3 FDR Controlling Methods

3.3.1 The Benjamini-Hochberg Procedure

Benjamini and Hochberg (1995) propose a practical method for FDR control, known as Benjamini-Hochberg (BH) procedure and detailed in Algorithm 1.

Algorithm 1: Benjamini-Hochberg procedure

Input: p-values $p_1, \ldots, p_n$; significance level $\alpha \in (0, 1)$.

Sort the p-values in ascending order: $p_{(1)} \leq \cdots \leq p_{(n)}$;

if $p_{(j)} > j\alpha/n$ *for all* $j \in [n]$ **then**

    Reject no hypotheses;

else

    Compute $j^* = \max\left\{j \in [n] : p_{(j)} \leq j\alpha/n\right\}$;

    Compute the index set R of all the hypotheses corresponding to $\{p_{(j)}\}_{j \leq j^*}$;

Output: Index set R corresponding to the rejected hypotheses.

The theoretical foundation of the BH procedure has been strengthened through subsequent research addressing various dependence structures among test statistics. Benjamini and Hochberg (1995) originally proved that the BH procedure controls the false discovery rate at level $n_0\alpha/n$ under independence of p-values corresponding to true nulls for any configuration of the false nulls.

This result was later extended by Benjamini and Yekutieli (2001) to the case where p-values exhibit *Positive Regression Dependence on the Subset* (PRDS) of true nulls I_0. To define this property, recall that a subset $D \subseteq [0, 1]^n$ is called *increasing* if whenever $p \in D$ and $q \geq p$, then $q \in D$.

Definition 1 (PRDS property)**.** A set of p-values $p_1, \ldots, p_n$ satisfies the PRDS property on the subset $I \subseteq [n]$ if for any increasing set $D \subseteq [0,1]^n$ and for each index $j \in I$, the probability

$$\Pr[(p_1, \ldots, p_n) \in D \mid p_j \leq t]$$

is non-decreasing in $t \in (0, 1]$.

Furthermore, Benjamini and Yekutieli (2001) demonstrated that the Benjamini-Hochberg procedure maintains FDR control at level $n_0\alpha/n$ under arbitrary dependence structures by replacing the significance level α with the more conservative threshold:

$$\frac{\alpha}{\left(\sum_{j\in[n]} 1/j\right)}.$$

3.3.2 Adaptive Benjamini-Hochberg Procedures

A limitation of the standard Benjamini-Hochberg procedure stems from the fact that it controls the FDR at level $n_0\alpha/n$ rather than at the desired level α. This leads to increasing conservativeness as the number of false nulls grows, potentially resulting in substantial power loss. To address this limitation, researchers have developed *adaptive* procedures that incorporate an estimate of the proportion of true null hypotheses (e.g., Benjamini and Hochberg, 2000; Storey, 2002; Storey et al., 2004; Benjamini et al., 2006; Blanchard and Roquain, 2009; Solari and Goeman, 2017; Magnani and Solari, 2021).

The rationale behind adaptive methods is that if the proportion of true nulls $\pi_0 = n_0/n$ can be accurately estimated using an estimator $\hat{\pi}_0$, then applying the BH procedure with the adjusted significance level $\tilde{\alpha} = \alpha/\hat{\pi}_0$ yields FDR control at the target level α:

$$\mathrm{FDR}(R_{\mathrm{BH}}(\tilde{\alpha})) \leq \pi_0\tilde{\alpha} = \alpha\pi_0/\hat{\pi}_0,$$

where $R_{\mathrm{BH}}(\tilde{\alpha})$ is the rejection set of the BH procedure at level $\tilde{\alpha}$. Indeed, when $\hat{\pi}_0$ provides a good approximation of π_0, the bound $\alpha\pi_0/\hat{\pi}_0$ approaches α, thereby recovering the power lost due to conservativeness. These adaptive strategies can similarly be applied to enhance the power of other conservative procedures, such as the Bonferroni method. The most widely used estimator in adaptive FDR procedures is the Storey estimator (Spjotvoll, 1972;

Storey, 2002; Storey et al., 2004):

$$\hat{\pi}_0 = \frac{1 + \sum_{j \in [n]} \mathbb{1}[p_j > \lambda]}{(1-\lambda)n},$$

where $\lambda \in (0, 1)$ is a tuning parameter commonly set to $1/2$ (Storey and Tibshirani, 2003) or to the significance level α (Blanchard and Roquain, 2009). Storey and Tibshirani (2003) proved that the adaptive Benjamini-Hochberg procedure with Storey estimator ensures finite FDR control under the assumption of independence among p-values. An alternative approach, proposed by Magnani and Solari (2021), employs the Hommel estimator (Hommel, 1988) for the number of true null hypotheses, defined as:

$$\hat{n}_0 = \max \left\{ j \in [n] : p_{n-j+i} > \frac{i\alpha}{j} \text{ for } i \in [j] \right\}. \tag{3.6}$$

3.4 FWER and FDP Controlling Methods: the Closed Testing Principle

This section introduces multiple testing procedures for FWER and FDP control based on the closed testing principle (Marcus et al., 1976). Sonnemann (1982) established closed testing procedure as a general framework encompassing all FWER-controlling methods as particular instances. More recently, Goeman and Solari (2011) developed closed testing procedure for FDP control, and Goeman et al. (2021) showed that any FDP-controlling method either corresponds to a closed testing procedure or is dominated by one. In a recent work by Xu et al. (2025), the closure principle has been generalized to encompass a broader class of multiple testing procedures—including the FDR-controlling methods leveraging the use of e-values.

In general, the closed testing principle facilitates the task of constructing a multiple testing procedure by reducing it to the simpler task of choosing appropriate *local tests* for all the intersection hypotheses.

The key advantage of closed testing procedure is its ability to support simultaneous post-hoc inference. They provide confidence upper bounds for the false discovery proportion that are valid regardless of how researchers select their data. This simultaneous control ensures

that the false discovery proportion remains controlled even when researchers choose specific subsets *post-hoc*, i.e., based on patterns they observe in the data, without requiring them to pre-specify their selection criteria. FDP-controlling procedures maintain statistical validity while allowing researchers the flexibility to explore their data and make inferences on subsets they identify during the analysis, rather than restricting them to predetermined subsets of the data.

3.4.1 The Closed Testing Procedure

Consider the elementary hypotheses $H_1, \ldots, H_n$. For any subset $S \subseteq [n]$ of selected hypotheses, define $\tau(S) = |I_0 \cap S|$ as the number of true nulls within S, where $|\cdot|$ denotes set cardinality. The false discovery proportion in any non-empty subset $S \subseteq [n]$ is then given by $\pi(S) = \tau(S)/|S|$, representing the fraction of true nulls among the selected hypotheses.

For a fixed significance level $\alpha \in (0, 1)$, the method of Goeman and Solari (2011) provides $(1-\alpha)$-confidence upper bounds $\pi_\alpha(S)$ that hold simultaneously across all possible subsets $S \subseteq [n]$:

$$\Pr\left[\pi(S) \leq \pi_\alpha(S) \text{ for all } S \subseteq [n]\right] \geq 1 - \alpha. \tag{3.7}$$

Since $\tau(S) = \pi(S) \cdot |S|$, these bounds are equivalent to confidence upper bounds $t_\alpha(S)$ on the number of false discoveries:

$$\Pr\left[\tau(S) \leq t_\alpha(S) \text{ for all } S \subseteq [n]\right] \geq 1 - \alpha. \tag{3.8}$$

This framework naturally extends to provide lower bounds on discoveries. Let $d(S) = |S| - \tau(S)$ denote the number of true discoveries in the subset S. From equation (3.8), simultaneous $(1-\alpha)$-confidence lower bounds $d_\alpha(S) = |S| - t_\alpha(S)$ can be derived for the number of true discoveries:

$$\Pr\left[d(S) \geq d_\alpha(S) \text{ for all } S \subseteq [n]\right] \geq 1 - \alpha. \tag{3.9}$$

Equivalently, this provides a lower bound on the proportion of true discoveries:

$$\Pr\left[d(S)/|S| \geq d_\alpha(S)/|S| \text{ for all } S \subseteq [n]\right] \geq 1 - \alpha. \tag{3.10}$$

The confidence bounds are built leveraging the closed testing principle (Marcus et al., 1976). For any $S \subseteq [n]$ define the intersection hypothesis H_S as the intersection of all the elementary hypotheses whose indices are contained in S, that is:

$$H_S := \bigcap_{j \in S} H_j, \qquad S \subseteq [n]. \tag{3.11}$$

Closed testing is defined by its *local tests*, which are used to test each intersection hypothesis. Let $\phi_\alpha(S) \in \{0, 1\}$ denote the local test for H_S, $S \subseteq [n]$, with 1 indicating rejection at level α. A major computational challenge of closed testing procedure is that they require performing up to $2^n - 1$ tests, which grows exponentially with n. To address this issue, various computational *shortcuts* have been proposed (e.g., Goeman and Solari, 2011; Goeman et al., 2019; Dobriban, 2020; Tian et al., 2023) and are discussed in Sections 3.4.3 and 3.4.6.

After conducting all the possible intersection tests with the local tests, the closed testing procedure corrects for multiplicity by rejecting H_S if and only if for all $J \supseteq S$ the hypothesis H_J is rejected by the local test $\phi_\alpha(J)$. That is, the corrected test for H_S is defined as:

$$\bar{\phi}_\alpha(S) := \min_{J \supseteq S} \{\phi_\alpha(J)\}, \tag{3.12}$$

and rejects H_S if and only if $\bar{\phi}_\alpha(S)$ equals 1. A closed testing procedure is *consonant* if the local tests for any $S \subseteq [n]$ are chosen in such a way that the rejection of H_S implies the rejection of H_j for at least one $j \in S$.

Under the assumption that the probability of rejecting H_{I_0} at the significance level α is at most α, i.e.:

$$\Pr[\phi_\alpha(I_0) = 1 \ ; \ H_{I_0}] \leq \alpha, \tag{3.13}$$

the closed testing method (detailed in Algorithm 2) provides the confidence lower bound for the number of true discoveries $d(S)$ in S:

$$d_\alpha(S) = \min_{K \subseteq S} \left\{ |S \setminus K| \ : \ \bar{\phi}_\alpha(K) = 0 \right\}. \tag{3.14}$$

The following result, due to Goeman and Solari (2011), guarantees that closed testing

Algorithm 2: Closed testing blueprint

Input: Individual hypotheses H_j, for all $j \in [n]$; a method ϕ for carrying out *local* tests $\phi_\alpha(J)$ of H_J, for any $J \subseteq [n]$; significance level $\alpha \in (0, 1)$.

1 For each $J \subseteq [n]$, test H_J at level α; let $\phi_\alpha(J) \in \{0, 1\}$ denote the rejection indicator.

2 For each $J \subseteq [n]$, adjust the local test by setting $\bar{\phi}_\alpha(J) := \min\{\phi_\alpha(K) : K \supseteq J\}$.

3 For any $S \subseteq [n]$, the lower bound for the number of true discoveries in S is:

$$d_\alpha(S) := \min_{K \subseteq S} \{|S \setminus K| : \bar{\phi}_\alpha(K) = 0\}. \tag{3.15}$$

Output: A $(1 - \alpha)$ lower bound $d_\alpha(S)$ for the number of outliers in any $S \subseteq [n]$.

leads to simultaneously valid confidence bounds as long as the local tests are valid.

Proposition 1 (Goeman and Solari (2011))**.** *If Algorithm 2 is applied using a valid local test ϕ satisfying $\mathbb{P}[\phi_\alpha(I_0) = 1 \ ; \ H_{I_0}] \leq \alpha$, then the output lower bounds $d(S)$ satisfy:*

$$\mathbb{P}\left[d(S) \leq |I_1 \cap S| \text{ for all } S \subseteq [n]\right] \geq 1 - \alpha. \tag{3.16}$$

The selection of appropriate local tests involves balancing statistical power with computational feasibility. Common choices are the Bonferroni method and Simes test. However, the Simes test offers clear advantages: it is strictly more powerful than the Bonferroni method under the PRDS property, and linear-time computational shortcuts are available (Goeman et al., 2019). Therefore, the Simes test is typically the preferred choice. See Section 3.4.4 for a comparison with the Benjamini-Hochberg procedure.

Beyond these standard choices, closed testing can also be efficiently implemented with local tests that combine p-values through monotonic transformations, such as Fisher's combination method (Tian et al., 2023). Building on this flexibility, Chapter 5 proposes a novel approach that employs closed testing with local rank-based tests. The key advantage of this approach is its ability to provide distribution-free inference. However, choosing the most suitable local test is challenging, as highlighted in Sections 4.3. This motivates the adoption of a data-driven approach for selecting the local test (see Chapter 5).

3.4.2 Classic Local Tests

This section provides an overview of well-established local tests commonly used in the literature. Table 3.1 provides a summary of these tests and the assumptions required to ensure FWER and FDP control.

Local Test	**Assumption**	**Closed Testing**	
		FWER control	**FDP control**
Bonferroni	None	Holm (1979)	Implied by FWER (consonant)
Simes test (Simes, 1986)	PRDN[†]	Hommel (1988)	Goeman et al. (2019)
Fisher combination method	Independence	Dobriban (2020)	Tian et al. (2023)
Higher Criticism (Donoho and Jin, 2004)	Independence	Implied by FDP	Goeman et al. (2021)
Adaptive Simes test with Storey estimator	Independence	Implied by FDP	Heller and Solari (2023)
Adaptive Simes test with Storey estimator	Conformal p-values	Implied by FDP	Magnani et al. (2024)

[†] A set of p-values is said to satisfy the *Positive Regression Dependence within Nulls* (PRDN) if the set of null p-values obey the PRDS property on I_0 (Su, 2018).

Table 3.1: Summary of local tests for the classic setup.

3.4.2.1 Bonferroni Correction

The *Bonferroni test* rejects the null hypothesis H_S at level $\alpha \in (0, 1)$ if the statistic

$$T_S^{\text{Bonf}} := \min_{j \in S} |S| p_j \tag{3.17}$$

is less than or equal to α. The Bonferroni correction ensures type-I error probability control with valid p-values:

$$\Pr\left[\min_{j\in S}|S|p_j \leq \alpha \ ; \ H_S\right] = \Pr\left[\bigcup_{j\in S}\left\{p_j \leq \frac{\alpha}{|S|}\right\} \ ; \ H_S\right] \leq \sum_{j\in S}\Pr\left[p_j \leq \frac{\alpha}{|S|} \ ; \ H_{|S|}\right] \leq \alpha.$$

This bound holds under any type of dependence among p-values, making the method particularly appealing for its robustness and flexibility. However, the Bonferroni correction is known to be conservative, particularly when the number of tests is large, often resulting in reduced statistical power. The conservativeness becomes less pronounced under stronger distributional assumptions. Specifically, when the p-values are independent and identically distributed as a standard uniform random variable under the null hypothesis, the type-I error probability control becomes nearly exact:

$$\Pr\left[\min_{j\in S}|S|p_j \leq \alpha \ ; \ H_S\right] = 1-\Pr\left[\bigcap_{j\in S}\left\{p_j \geq \frac{\alpha}{|S|}\right\} \ ; \ H_S\right] = 1-\left(1-\frac{\alpha}{|S|}\right)^{|S|} \underset{|S|\to\infty}{\longrightarrow} 1-e^{-\alpha},$$

where $1-e^{-\alpha}$ is very close to the significance level for small values of α. Indeed, as $\alpha \to 0$, the first-order Taylor expansion gives $e^{-\alpha} = 1-\alpha+o(\alpha)$. This means that for values of α near 0, we have $1-e^{-\alpha} \approx 1-(1-\alpha) = \alpha$.

Because the rejection of the H_S depends solely on the smallest p-value, Bonferroni method is effective when evidence strongly contradicts a few elementary hypotheses, but lacks power to reject the null when the evidence against the elementary hypotheses is mild.

Bonferroni method as local test in closed testing yields the Holm (1979) procedure for FWER control.

3.4.2.2 Simes Test

In order to mitigate the conservativeness of Bonferroni method, *Simes test* (Simes, 1986) rejects the H_S at level $\alpha \in (0,1)$ if the Simes statistic:

$$T_S^{\text{Simes}} := \min_{j\in\{1,\ldots,|S|\}} \frac{|S|}{j} p_{(j:S)} \tag{3.18}$$

is less than or equal to α, where $p_{(j:S)}$ is the $j$the ordered p-value in the sequence $\{p_j : j \in S\}$. This means that the H_S is rejected if any ordered p-values $p_{(j:S)}$ is smaller than $\alpha j/|S|$, performing well with a few strong effects. The rejection region of Bonferroni method is contained within the rejection region of Simes test, making the latter more powerful than the former. Simes test control the type-I error probability under Simes inequality (Simes, 1986):

$$\Pr\left[\bigcup_{j\in S}\left\{p_{(j:S)} \leq \frac{i\alpha}{|S|}\right\}\right] \leq \alpha. \tag{3.19}$$

This inequality holds when the p-values are i.i.d. (in which case it is actually an equality, and the test achieves exact type-I error probability control) or under certain positive dependence conditions (Sarkar and Chang, 1997; Sarkar, 1998), such as the PRDS property. When the p-values satisfy the PRDS property under H_S, i.e., for $I = S$, Simes inequality in equation (3.19) is guaranteed to hold. Simes test as local test in closed testing yields the Hommel (1988) procedure for FWER control.

3.4.2.3 Fisher Combination Method

In settings with weak signals, the Bonferroni and Simes tests may lack power, and alternative approaches become valuable. The *Fisher combination method*, which rejects the null hypothesis for large values of

$$T_S^{\text{Fisher}} := -2\sum_{j\in S}\log(p_j), \tag{3.20}$$

is specifically designed for detecting many weak signals. The Fisher combination method requires that the p-values be i.i.d. as a standard uniform random variable under the global null, under which the test statistic follows a chi-squared distribution $\chi^2_{2|S|}$ with $2|S|$ degrees of freedom, ensuring type-I error probability control.

3.4.2.4 Higher Criticism

Donoho and Jin (2004, 2015) introduced Higher Criticism as another powerful approach for the many-weak-signals regime. They consider a parametric model with independent

observations where under the null hypothesis H_j, $j \in S$ the observation Y_j is a standard normal random variable, while under the alternative hypothesis K_j the observation Y_j is a normal random variable with unit variance and mean equal to $\mu_j > 0$.

This framework can be formulated as a global testing problem where the alternative distribution is a mixture model:

$$H_S : Y_j \overset{iid}{\sim} N(0,1), \quad \forall j \in S, \qquad K_S : \; Y_j \overset{iid}{\sim} (1-\theta)N(0,1) + \theta N(\mu,1), \quad \forall j \in S, \quad (3.21)$$

where $\mu > 0$ represents the common signal strength of and $\theta \in (0,1)$ denotes the proportion of non-null observations. The global testing problem is equivalent to:

This setting characterizes the *rare/weak effects* model, where the observations that departs from the null distribution are sparse (i.e., θ is small), and individually undetectable (i.e., μ is small). To establish theoretical *detection boundaries*, let both parameters θ and μ scale with the sample size s. Specifically, consider the asymptotic regime:

$$\begin{aligned} \theta_s &= s^{-\beta}, \qquad \beta \in (1/2, 1)\,, \\ \mu_s &= \sqrt{2r\log(s)}, \qquad r \in (0,1), \end{aligned} \quad (3.22)$$

where β governs the signal sparsity and r controls the signal amplitude.

The choice of scaling in (3.22) ensures that individual effects remain undetectable. For any $r \in (0,1)$, the signal amplitude μ_s is smaller than the expectation of the maximum $Z := \max_{j \in S} Y_j$ under the null, which satisfies:

$$\mathbb{E}[Z] = \int_{-\infty}^{+\infty} z\phi(z)\Phi(z)^{s-1}\, dz = \sqrt{2\log(s)} + o\left(\log(s)^{-1/2}\right).$$

This confirms that the individual effects are indeed hard to detect in model (3.21) for any $r \in (0,1)$. When both r and β are fixed and known, the distributions under the null and the alternative hypothesis are completely specified, making the Neyman-Pearson likelihood ratio test optimal.

Nevertheless, as emphasized by Donoho and Jin (2004), there is a threshold effects for the likelihood ratio test: as the sample size n grows, the sum of the type-I and type-II errors tends to 0 or to 1 depending on whether the signal amplitude r exceeds the detection

boundary, given by the following function of the sparsity parameter β:

$$\rho^*(\beta) = \begin{cases} \beta - \frac{1}{2}, & \beta \in \left(\frac{1}{2}, \frac{3}{4}\right], \\ \left(1 - \sqrt{1-\beta}\right)^2, & \beta \left(\frac{3}{4}, 1\right). \end{cases} \tag{3.23}$$

Hence, if $r > \rho^*(\beta)$ the likelihood ratio test can asymptotically distinguish between the global null and alternative; otherwise, when $r < \rho^*(\beta)$, even the optimal likelihood ratio test fails to asymptotically separate the null and alternative hypotheses.

However, in practice, the parameters θ_s and μ_s are typically unknown, rendering the Neyman-Pearson test inapplicable. To address this limitation, Donoho and Jin (2004) proposed the *Higher Criticism* statistic, building on foundational work by Tukey (1976, 1989). This method serves as a practical alternative that matches the power of the optimal test in the detectable region $\left\{(\beta, r) : 1/2 < \beta < 1, r > \rho^*(\beta)\right\}$, without requiring complete specification of the data distributions, as needed for the Neyman-Pearson test.

The Higher Criticism test rejects the global null at the significance level $\alpha \in (0, 1)$ when the statistic:

$$T^{\mathrm{HC}} = \max_{1 \le j \le \alpha_0 s} \frac{\sqrt{s}(i/s - p_{(j:S)})}{p_{(j:S)}(1 - p_{(j:S)})} \tag{3.24}$$

exceeds a critical value $h(s, \alpha)$ satisfying $\Pr[T^{\mathrm{HC}} > h(s, \alpha)\ ;\ H] \le \alpha$. In equation (3.24), $\alpha_0 \in (0, 1)$ is a tuning parameter commonly set equal to $1/2$, and $p_{(j:S)}$ denotes the jth order statistic among the p-values $p_j = 1 - \Phi(Y_j)$, $j \in S$, where Φ is the standard normal c.d.f.

From Theorem 1.1 in Donoho and Jin (2004), where it is proved that the standardized test statistic $T^{\mathrm{HC}}/\sqrt{2\log(\log s)}$ converges in probability to 1, it follows that the growth rate of the critical value $h(s, \alpha)$ is:

$$h(s, \alpha) \approx \sqrt{2\log(\log s)}.$$

Considering a sequence of significance levels α_n such that $\alpha_s \to 0$ *slowly enough* to ensure that

$$h(s, \alpha_s) \approx \sqrt{2\log(\log s)}(1 + o(1)),$$

Donoho and Jin (2004) established in their Theorem 1.2 that Higher Criticism achieves full asymptotic power in the same region as the Neyman-Pearson test, that is $\{(\beta, r) : 1/2 < \beta < 1, r > \rho^*(\beta)\}$. In the complementary region $\{(\beta, r) : 1/2 < \beta < 1, r < \rho^*(\beta)\}$, no test can reliably distinguish between hypotheses, since the most powerful Neyman-Pearson test has no power.

To demonstrate the optimality of Higher Criticism in the rare/weak effects regime, Donoho and Jin (2004) also compared it with other standard methods, including the Bonferroni test and Fisher combination method. They showed that these alternative approaches have detection boundaries that are higher (i.e., worse) than Higher Criticism, implying that they require stronger signals for reliable detection. Specifically, the detection boundary for the Bonferroni test is:

$$\rho_{\text{Bonf}}(\beta) = (1 - \sqrt{1-\beta})^2, \qquad \beta \in (1/2, 1).$$

This reveals that $\rho^*(\beta) = \rho_{\text{Bonf}}(\beta)$ for $\beta \in (3/4, 1)$ and $\rho^*(\beta) < \rho_{\text{Bonf}}(\beta)$ for $\beta \in (1/2, 3/4]$, demonstrating Higher Criticism's superior performance in the sparse/weak signal regime. Higher Criticism achieves equal optimality with Bonferroni for very sparse weak signals ($\beta \in (3/4, 1)$) and superior performance for moderately sparse weak signals ($\beta \in (1/2, 3/4)$). Remarkably, the same detection boundary characterizes the Benjamini-Hochberg procedure. Donoho and Jin (2004) proved that the Benjamini-Hochberg procedure has the same detection boundary as the Bonferroni method, i.e., $\rho_{\text{Bonf}}(\beta) = \rho_{\text{BH}}(\beta)$, for any $\beta \in (1/2, 1)$, indicating that both methods for type-I error probability control and approaches for FDR control face identical fundamental limitations in this setting.

Further supporting the superiority of Higher Criticism, Donoho and Jin (2004) proved that the Fisher combination method cannot asymptotically detect any signal in this regime, as stated in the following result.

Theorem 7 (Theorem 1.5 in Donoho and Jin (2004))**.** *Consider the model in* (3.21) *with parameters as in* (3.22)*. Then, asymptotically the Fisher test statistic is unable to separate the null hypothesis from the alternative in the testing problem* (3.21)*.*

3.4.3 Computational Shortcuts with Classic Local Tests

As established in previous sections, the closed testing procedure serve dual purposes: they are primarily employed for FDP control but also provide a valuable framework for FWER control. Both applications face a common computational challenge: closed testing procedure requires evaluating exponentially many tests, which hinders its practical feasibility. This limitation is particularly pronounced when computing lower bounds for FDP control.

To address this computational burden, researchers have developed various shortcuts for both FWER and FDP control. It is important to note that these computational techniques are tailored to specific local tests, and no general shortcuts for the closed testing procedure exist. Shortcuts fall into two categories: *exact* shortcuts that yield identical rejections to the full closed testing procedure, and *approximate* or *conservative* shortcuts that may produce fewer rejections than the full procedure.

Table 3.2 summarizes shortcuts for different classic local tests, which are discussed in the subsequent sections.

Local Test	**Validity**	**Error Rate**	**Exact**	**Approximate**
Simes test	PRDN	FWER	Hommel (1988) (quadratic time)	Hochberg (1988) (linear time on sorted p-values)
	PRDN	FWER	Meijer et al. (2019) (linear time on sorted p-values)	
	PRDN	FDP	Goeman et al. (2019)	Goeman and Solari (2011)
	Permutation	FDP		Hemerik et al. (2019)
Adaptive Simes test with Storey estimator	Independent or conformal p-values	FDP	Magnani et al. (2024)	

Table 3.2: Summary of shortcuts of the closed testing procedure for FWER and FDP control with classic local tests.

3.4.3.1 Shortcuts for FWER and FDP with Simes Local Test

This section briefly reviews the existing shortcut for FWER and FDP control via closed testing with Simes local test.

The Hommel (1988) procedure provides the first shortcut for FWER control, offering exact inference with quadratic time complexity $\mathcal{O}(n^2)$ in the number of hypotheses. Hochberg (1988) introduced an approximate shortcut that ensures FWER control. This method achieves faster $\mathcal{O}(n \log(n))$ computation time at the cost of reduced statistical power compared to Hommel's method. Subsequently, Meijer et al. (2019) proposed another exact shortcut for FWER control, achieving linear complexity $\mathcal{O}(n)$ after an initial sorting step for the p-values.

For FDP control, Goeman and Solari (2011) presented an approximate shortcut for FDP confidence bounds with quadratic time complexity $\mathcal{O}(|S|^2)$ for any subset $S \subseteq [n]$. Goeman et al. (2019) improved upon this approach by introducing an exact shortcut that computes $d_{\text{Simes}}(S)$ in linearitmic time $\mathcal{O}(|S| \log |S|)$. The novel ACODE method relies on this shortcut, which is detailed in Algorithm 3.

Algorithm 3: Shortcut for closed testing with Simes test (Goeman et al., 2019)

Input: p-values $p_1, \ldots, p_n$; significance level $\alpha \in (0, 1)$.

1 Sort the p-values $p_{(1)} \leq \ldots \leq p_{(n)}$;

2 Compute $h = \max\{0 \leq k \leq n : p_{(n-k+j)} > j\alpha/k \text{ for } j = 1, \ldots, k\}$;

3 Compute $d_{\text{Simes}} = d_{\text{Simes}}([n]) = n - h$;

4 **for *any*** $S \subset [n]$ **do**

5 ⌊ compute $d_{\text{Simes}}(S) = \min\{0 \leq k \leq |S| : p_{(k+j:S)} > j\alpha/h \text{ for } j = 1, \ldots, |S| - k\}$;

Output: A simultaneous $(1 - \alpha)$-confidence lower bound $d_{\text{Simes}}(S)$ for the number of outliers in S, for any $S \subseteq [n]$.

Line 2 of Algorithm 3 gives h, which is a $(1-\alpha)$-confidence upper bound for the number of true inliers $|I_0|$ in the overall test set. Then $d_{\text{Simes}} = n - h$ is a $(1 - \alpha)$-confidence lower bound for the number of false hypotheses (cfr. number of true outliers in the test sample) $|I_1|$.

3.4.3.2 Shortcuts for FDP control with Storey-Simes Local Tests

As discussed in Section 3.3.2, the BH procedure guarantees that expected proportion of inliers among the discoveries is bounded by $\alpha\pi_0$, where $\pi_0 = |I_0|/n$ is the unknown proportion of inliers in the test sample. When π_0 is expected to be not close to 1, $\alpha\pi_0$ falls below the target level α, making the BH procedure slightly too conservative. This motivates adjusting the level into $\alpha/\hat{\pi}_0$ with $\hat{\pi}_0$ an estimate of π_0, resulting in a π_0-adaptive version of the BH algorithm.

The idea of the adaptive BH procedure extends to the Simes test, leading to a π_0-adaptive version of Simes test for H_S that incorporates the proportion of true null hypotheses in S. The test is formulated as:

$$\phi_S^{\text{Storey}-\text{Simes}} = \mathbb{1}\left\{\min_{k\in\{1,\ldots,|S|\}}\{|S|p_{(k:S)}/k\} \leq \alpha/\hat{\pi}_0^S\right\}, \tag{3.25}$$

where

$$\hat{\pi}_0^S \quad = \quad \frac{1+\sum_{j\in S}\mathbb{1}\{p_j > \lambda\}}{|S|(1-\lambda)} \tag{3.26}$$

and $\lambda = h/(|S|+1)$ for any pre-specified integer h. This estimator is the Schweder-Spjøtvoll or Storey's estimator 1.1 (Schweder and Spjøtvoll, 1982; Storey, 2002; Storey et al., 2004), introduced in Section 3.3.2. The test $\phi_S^{\text{Storey}-\text{Simes}}$ was considered in Bogomolov (2023) and Heller and Solari (2023). The shortcut given in Algorithm 4 allows exact calculation of $d_{\text{Storey}-\text{Simes}}(S)$ in cubic time in $|S|$.

Algorithm 4: Shortcut for closed testing with adaptive Simes local tests.

Input : p-values $p_1, \ldots, p_n$; index set $S \subseteq [n]$; significance level $\alpha \in (0,1)$; tuning parameter $\lambda \in (0,1)$.

Sort the p-values: $p_{(1)} \leq \ldots \leq p_{(n)}$;
Initialize $d_{\text{Storey}-\text{Simes}}(S) \leftarrow 0$;
for *each* $i \in \{1, \ldots, |S|\}$ **do**
    for *each* $j \in \{0, \ldots, n - |S| + i - 1\}$ **do**
        $l \leftarrow |S| - i + j + 1$;
        $(q_1, \ldots, q_l) \leftarrow (p_{(i:S)}, \ldots, p_{(|S|:S)}, p_{(n-|S|-j+1:S^c)}, \ldots, p_{(n-|S|:S^c)})$;
        if $\min_{j \in [l]} \left(l \frac{q_j}{j} \right) > \alpha \left(\frac{l(1-\lambda)}{1 + \sum_{j \in [l]} \mathbb{1}\{q_j > \lambda\}} \right)$, **then**
            return $d_{\text{Storey}-\text{Simes}}(S)$.
    $d_{\text{Storey}-\text{Simes}}(S) = d_{\text{Storey}-\text{Simes}}(S) + 1$;

Output : $(1-\alpha)$-confidence lower bound $d_{\text{Storey}-\text{Simes}}(S)$ for the number of true discoveries in S

3.4.4 Closed Testing with Simes Local Tests vs BH

This section compares closed testing with the Simes local test and the Benjamini-Hochberg. This comparison is motivated by the fact that the Simes test and the Benjamini-Hochberg procedure share the same critical values, suggesting potential relationships between the two methods.

Let ϕ denote a level-α test for the global null hypothesis $H = H_{[n]}$. Closed testing provides the following quantities:

- $\mathbb{1}\{\phi = 1\}$ to indicate outlier detection;
- d to provide outlier enumeration;
- $|D|$ to quantify outlier discoveries.

Outlier enumeration is implied by outlier identification (by using $|D|$) and, in turn, implies outlier detection (by using $\mathbb{1}\{d > 0\}$). These quantities can be applied to any subset S of

the test sample, obtaining $\mathbb{1}\{\phi_S = 1\}$, $d(S)$, and $|D \cap S|$. Simultaneous confidence bounds $d(S)$ obtained by closed testing satisfy the following relations:

$$\mathbb{1}\{\phi_S = 1\} \geq \mathbb{1}\{d(S) > 0\}, \qquad d(S) \geq |D \cap S|. \tag{3.27}$$

Specifically, the index set of the discoveries (cfr. localized outliers) with closed testing with Simes local test is given by

$$D_{\text{Simes}} = \{j \in [n] : d_{\text{Simes}}(\{j\}) = 1\} = \{j \in [n] : p_j \leq \alpha/h\},$$

where h (given at line 2 of Algorithm 3) is the $(1-\alpha)$-confidence upper bound for the number of true hypotheses $n_0 = |I_0|$.

All hypotheses H_j with index $j \in D_{\text{Simes}}$ are rejected by closed testing while controlling the familywise error rate at level-α (Hommel, 1988): $\mathbb{P}(|D_{\text{Simes}} \cap I_0| > 0) \leq \alpha$.

The BH algorithm applied to $p_1, \ldots, p_n$ at level α returns the index set of the discoveries:

$$D_{\text{BH}} \quad = \quad \{j \in [n] : p_j \leq \alpha d_{\text{BH}}/n\} \tag{3.28}$$

where

$$d_{\text{BH}} \quad = \quad \max\left\{k \in \{0, \ldots, n\} : \sum_{j=1}^{n} \mathbb{1}\{p_j \leq \alpha k/n\} \geq k\right\}. \tag{3.29}$$

is the number of discoveries made by the BH procedure. To distinguish between discoveries with FWER and FDR guarantees, discoveries made by closed testing will be referred to as FWER discoveries, while those made by an FDR controlling procedure will be called FDR discoveries.

A simple observation is that the event where the closed testing procedure with Simes local test and the BH procedure make at least one non-trivial statement is identical:

$$\mathbb{1}\{d_{\text{Simes}} > 0\} = \mathbb{1}\{d_{\text{BH}} > 0\}. \tag{3.30}$$

Consequently, the two methods have the same weak FWER control and the same power for

outlier detection. This is clearly illustrated in the right panel of Figure 1.1, where the solid gray line (representing the power of ACODE with the Simes local test) and the dotted gray line (representing the power of the Benjamini–Hochberg procedure) completely overlap. In particular, their agreement under the global null hypothesis (i.e., when there are 0 outliers in the test set) reflects the fact that both methods control the FWER at the same level. Furthermore, Goeman et al. (2019) showed that the confidence bound d_{Simes} is between the number of Hommel-FWER discoveries and the number of BH-FDR discoveries, i.e.

$$|D_{\text{Simes}}| \leq d_{\text{Simes}} \leq |D_{\text{BH}}|. \tag{3.31}$$

Finally, note that the BH discoveries do not provide the true discovery guarantee, i.e.,

$$\mathbb{P}(|D_{\text{BH}} \cap S| \leq |I_1 \cap S| \text{ for all } S) \tag{3.32}$$

may be less than $1 - \alpha$.

3.4.5 One-Way ANOVA Setup: Many-to-One Comparisons

This section considers the *many-to-one comparison problem*, where the objective is to compare the distributions of observations from several groups against that of a benchmark group. To address this problem, a closed testing procedure with appropriate local tests can be employed. The relevant local tests are summarized in Table 3.3 and described in detail in the following sections.

3.4.5.1 Parametric Gaussian Model

Suppose to have a sample of m observations $X_1, \ldots, X_m$ drawn independently from a normal distribution with mean μ and variance σ^2. Additionally, suppose to have n observations Y_j, $j \in [n]$ drawn independently from a normal distribution with same variance σ^2 but potentially different means μ_j, $j \in [n]$. For each $j \in [n]$ consider the null hypothesis

Local Test	**Assumption**	**Closed Testing**	
		FWER control	**FDP control**
Max t-test	Homoscedastic Normal Distribution	Dunnett (1955) (step down procedure)	Implied by FWER
WMW test (Wilcoxon, 1945)	Distribution-free	Implied by FDP	Magnani et al. (2024)
Shiraishi test (Shiraishi, 1985)	Distribution-free with G_j stochastically larger than $U(0,1)$	Implied by FDP	Magnani et al. (2024)

Table 3.3: Summary of local tests for the one-way ANOVA setup.

$H_j : \mu_j = \mu$ against the alternative $K_j : \mu_j > \mu$. The test is performed using the statistic:

$$T_j = \frac{Y_j - \bar{X}}{\hat{\sigma}_p \sqrt{1 + 1/m}}$$

where $\bar{X}$ is the sample mean of $X_1, \ldots, X_m$ and $\hat{\sigma}_p$ is the pooled standard deviation estimator of the common variance σ. This statistic follows a Student t distribution with $m-1$ degrees of freedom.

The *single-step Dunnett test* rejects the null hypothesis H_j a level α if T_j exceeds the $(1-\alpha)$-quantile of the distribution of $\max\{T_1, \ldots, T_n\}$, ensuring the family-wise error rate control at level α. Note that $(T_1, \ldots, T_n)$ has n-variate Student t distribution with a correlation matrix having off-diagonal elements equal to $1/(1+m)$.

An improved version of Dunnett's procedure applies the closed testing principle using the single-step Dunnett test $T_S = \max_{j \in S}\{T_j\}$ as local test for the intersection hypothesis $H_S = \cap_{j \in S} H_j$, $S \subseteq [n]$. This procedure is referred to as *step-down Dunnett* (Bretz et al., 2016).

3.4.5.2 Nonparametric Model

Steel (1959) presents a multiple comparison rank sum procedure for comparing multiple treatments against a control in a one-way classification with equal numbers of observations under nonparametric assumptions.

Assume $m = n$ and that the observations $X_1, \ldots, X_m$ are drawn independently from a distribution F. For each $j \in [n]$, the observation Y_j is drawn independently from a distribution Q_j. For each $j \in [n]$, consider the null hypothesis $H_j : Q_j = F$ against the alternative $K_j : Q_j \succeq F$, where $\succeq$ denotes the stochastic ordering. Under these conditions, if we assume that $Q_j = G_j(F)$ for some G_j distribution on $[0, 1]$ stochastically larger than the standard uniform distribution, we can use the Shiraishi test (and the Wilcoxon-Mann-Whitney test as a special case) as local test in the closed testing procedure.

3.4.6 Computational Shortcuts with Local Tests for Many-to-One Comparisons

This section details the shortcuts for the local tests in the one-way ANOVA setup, which are summarizes in Table 3.4

Local Test	**Error Rate**	**Exact**
Monotone, symmetric and separable tests	FDP	Tian et al. (2023)
Shiraishi test with increasing outlier density	FDP	Magnani et al. (2024)
Symmetric, Monotone	FWER	Dobriban et al. (2015)

Table 3.4: Summary of shortcuts of the closed testing procedure for FWER and FDP control in the one-way ANOVA setup.

3.4.6.1 Shortcut for FDP control with Monotone, Symmetric, and Separable Tests

Tian et al. (2023) introduced an exact shortcut for FDP control via closed testing with local tests that satisfy three key properties: *monotonicity*, *symmetry* and *separability*. These properties are satisfied by classical tests such as the Mann-Whitney test and Fisher's combination test. Tests meeting these criteria can be expressed as the sum of test scores (e.g., p-values or other test statistics) corresponding to the elementary hypothesis contained in the set of interest $S \subseteq [n]$. The formal mathematical expression is provided in equation (4)

of Tian et al. (2023)), while precise definitions of these properties are detailed in their Appendix B in Tian et al. (2023). Algorithm 1 in Tian et al. (2023) achieves linear-time computation $\mathcal{O}(|S|)$ of the lower bound $d(S)$ for any $S \subseteq [n]$, following an initial preprocessing step that computes and sorts the test statistics.

3.4.6.2 Shortcut for FDP control with Shiraishi Local Tests

To enable the calculation of $d_{\text{Shiraishi}}(S)$, Magnani et al. (2024) develop a shortcut detailed in Algorithm 5. This algorithm computes $d_{\text{Shiraishi}}(S)$ in quadratic time for a given increasing density function g, after sorting the p-values.

Let $a^l = (a^l_1, \ldots, a^l_{m+l})$ be a vector where the kth element is $a^l_k = \mathbb{E}[g(U^{(k)}_{m+l})]$. Since g is increasing, the elements of the vector are ordered such that $a^l_1 \leq \ldots \leq a^l_{m+l}$. The score vector a^l can be estimated with the desired accuracy using Monte Carlo simulation. The critical value $c^G_\alpha(m, l)$ corresponds to the $(1-\alpha)$-quantile of the normal distribution with the following mean and variance:

$$\mu_{m+l} = \frac{l}{m+l} \sum_{k \in [m+l]} a^l_k, \qquad \sigma^2_{m+l} = \frac{ml}{(m+l)(m+l-1)} \sum_{k \in [m+l]} \left(a^l_k - \frac{\mu_{m+l}}{l} \right)^2.$$

Algorithm 5: Shortcut for closed testing with Shiraishi local tests.

Input: Calibration scores $(X_1, \ldots, X_m)$; test scores $(Y_1, \ldots, Y_n)$; an increasing density function g; significance level $\alpha \in (0, 1)$.

Compute $(R_{m+1}, \ldots, R_{m+n})$, the ranks of $(Y_1, \ldots, Y_n)$ within $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$;

for $h = n, \ldots, 1$ **do**

  if $\sum_{j=1}^{h} \mathbb{E}[g(U_{m+h}^{(R_{m+j})})] < c_\alpha^G(m, h)$ **then**

    return h;

$d_{\text{Shiraishi}} = d_{\text{Shiraishi}}([n]) = n - h$;

for ***any*** $S \subset [n]$ **do**

  for $i = 1, \ldots, |S|$ **do**

    for $j = 0, \ldots, \max(h - |S| + i - 1, 0)$ **do**

      $l \leftarrow |S| - i + j + 1$;

      $(W_1, \ldots, W_l) \leftarrow (Y_{(1:S)}, \ldots, Y_{(|S|-i+1:S)}, Y_{(1:S^c)}, \ldots, Y_{(j:S^c)})$;

      Compute $(R_{m+1}, \ldots, R_{m+l})$, the ranks of $(W_1, \ldots, W_l)$ within $(X_1, \ldots, X_m, W_1, \ldots, W_l)$;

      if $\sum_{j=1}^{l} \mathbb{E}[g(U_{m+l}^{(R_{m+j})})] < c_\alpha^G(m, l)$, **then**

        return $d_{\text{Shiraishi}}(S)$.

    $d_{\text{Shiraishi}}(S) = d_{\text{Shiraishi}}(S) + 1$;

Output: A $(1 - \alpha)$-lower bound $d_{\text{Shiraishi}}(S)$ for the number of outliers in S, for any $S \subseteq [n]$.

Chapter 4

Conformal p-values

4.1 Introduction

Conformal Inference (Gammerman et al., 1998; Vovk et al., 1999; Saunders et al., 1999) provides a general framework for equipping machine learning algorithms with valid inferential guarantees and uncertainty quantification, which is especially important given that these algorithms often lack interpretability due to their extreme complexity.

Conformal Inference operates without making assumptions about the underlying model and requires only minimal assumptions about the data distribution. It is a distribution-free approach, offering data-driven uncertainty quantification that is valid in finite samples, relying on the sole assumption of data *exchangeability*.

The most common goal in conformal inference is testing the exchangeability of a test point with the reference sequence, that can also be interpreted as testing whether the test point represents an outlier relative to the reference sequence.

Throughout this chapter split conformal techniques are considered to perform outlier detection for multiple comparisons leveraging conformal p-values.

4.2 Multiple Outlier Detection via Conformal p-values

The crucial idea in Conformal Inference is to measure the *conformity* of a test point to a reference dataset—that is, how well a test point aligns with the reference sample. To this end, a *conformity score function* is adopted, with larger values indicating greater disagreement. This serves as the central component in constructing *conformal p-values*.

The score function is often based on a model that needs to be trained using available observations. This thesis focuses on the *split conformal* framework, which requires data-splitting to train the score function. With this approach the available data are split into two samples: the training set is used to train the score function, and the calibration set is used for building the conformal p-values and making inference.

To formalize this, consider a reference data set $Z_1, \ldots, Z_m \in \mathbb{R}^d$ comprising m observations each with dimensions $d \geq 1$, where d may be large compared to m. Assume these are independent and identically distributed random samples from some unknown distribution P_0 on $\mathbb{R}^d$. Consider also a test set $Z_{m+1}, \ldots, Z_{m+n}$ and, for each $j \in [n]$, assume $Z_{m+j} \in \mathbb{R}^d$ is independently sampled from some unknown distribution P_j, which may or may not be equal to P_0. That is,

$$Z_1, \ldots, Z_m \overset{\text{i.i.d.}}{\sim} P_0, \qquad\qquad Z_{m+j} \overset{\text{ind.}}{\sim} P_j, \ \forall j \in [n]. \tag{4.1}$$

The data points sampled from P_0 will be referred to as *inliers* and those from any $P_j \neq P_0$ as *outliers*. For each $j \in [m]$, define the null hypothesis $H_j : P_j = P_0$. Let $I_0 := \{j \in [n] : P_j = P_0\}$ denote the subset of inliers in the test set. Similarly, let $I_1 = [n] \setminus I_0$ denote the subset of outliers in the test set.

Let $\hat{s} : \mathbb{R}^d \to \mathbb{R}$ denote a score function trained on a training sample $\mathcal{D}_{\text{train}}$. In general, $\hat{s}$ may possibly depend also on the unordered collection of observations $\{Z_1, \ldots, Z_m, Z_{m+1}, \ldots, Z_{m+n}\}$ (see Section 4.2.1 for further details). Taking this into consideration, the calibration scores and the test scores are denoted, respectively, as:

$$X_i = \hat{s}(Z_i; \mathcal{D}_{\text{train}}, \{Z_1, \ldots, Z_{n+m}\}), \qquad Y_j = \hat{s}(Z_{m+j}; \mathcal{D}_{\text{train}}, \{Z_1, \ldots, Z_{n+m}\}), \tag{4.2}$$

for any $i \in [m]$ and $j \in [n]$. For any $j \in [n]$, the conformity p-value for the test point Z_{m+j} is defined as:

$$p_j = \frac{1 + \sum_{i \in [m]} \mathbb{1}\{X_i \geq Y_j\}}{m + 1}. \tag{4.3}$$

Under the assumption that the score vector $(X_1, \ldots, X_m, Y_j)$ is exchangeable and has no ties, this is a permutation p-value (Angelopoulos et al., 2025). It is exact for $\alpha \in \{1/(m+1)!, 2/(m+1)!, \ldots, 1\}$ under Condition 1 in Hemerik and Goeman (2018). When this

condition does not hold, the conformal p-value is not too conservative for large values of m.

Assume now that the calibration points $Z_1, \ldots, Z_m$ are i.i.d. with marginal distribution P_0. For each $j \in [n]$, test the null hypothesis

$$H_j : Z_{m+j} \sim P_0, \tag{4.4}$$

using the conformal p-value defined in equation (4.3). Rejecting the null hypothesis indicates that the jth test point does not come from the reference distribution P_0.

4.2.1 Score Exchangeability with One-Class and Binary Classifiers

The conformal literature has primarily considered two types of models for constructing score functions in outlier detection: one-class classifiers (Bates et al., 2023) and binary classifiers (Marandon et al., 2024).

One-class classifiers build models using only a single class of data and construct scores that distinguish inliers from outliers by learning from the data what inliers typically look like. Common examples include isolation forests and support vector machines. In contrast, binary classifiers such as deep neural networks, random forests, and AdaBoost use labeled data from both classes to train the model to distinguish inliers from outliers. As a consequence, when using binary classifiers, the conformity score function depends on both the training dataset and the test set—the latter being the only sample that may contain outliers.

Using an independent set of inliers as training set, one-class classifiers when applied to i.i.d. observations preserve the i.i.d. structure, yielding i.i.d. conformity scores. In contrast, binary classifiers introduce dependence among the conformity scores, because the data-adaptive score function depends on both the calibration and test data too. To preserve the exchangeability structure (although not independence)—which is crucial for the validity of conformal p-values—Marandon et al. (2024) established that the binary score function must satisfy an invariance property when applied to exchangeable observations. Specifically, Lemma 3.2 in Marandon et al. (2024) shows that the scores in equation (4.2) obtained via binary classifiers applied to exchangeable observations yield exchangeable scores, provided that the score function satisfies the following invariant property: for any permutation π of

$[m+n]$,

$$\hat{s}(z; \mathcal{D}_{\text{train}}, \{Z_1, \dots Z_{m+n}\}) = \hat{s}(z; \mathcal{D}_{\text{train}}, \{Z_{\pi(1)}, \dots, Z_{\pi(m+n)}\}). \tag{4.5}$$

This theoretical property translates into the practical method of training the binary classifier on an augmented test set that combines the original test observations with inliers drawn from the calibration set. This "dilution" technique is designed to ensure that the binary classifier treats calibration and test inliers as exchangeable.

In light of these results, it is important to note that the definition of the conformity scores in equation (4.2), combined with the data generating model defined in (4.1), implies that the scores $(X_1, \dots, X_m, (Y_j)_{j \in I_0})$ are exchangeable (Marandon et al., 2024). Furthermore, it is easy to see that $(X_1, \dots, X_m, (Y_j)_{j \in I_0})$ are mutually independent if the function $\hat{s}$ does not depend on $\{Z_1, \dots, Z_{n+m}\}$.

4.2.2 PRDS Conformal p-values with One-Class and Binary Classifier

Under the assumption of independent observations, Bates et al. (2023) prove that the conformal p-values constructed from scores obtained from one-class classifiers are PRDS on the subset $I_0 \subseteq [n]$ corresponding to the true nulls. This result is formalized in Theorem 2.4 in Bates et al. (2023) and is stated below.

Theorem 8 (Theorem 2.4 in Bates et al. (2023))**.** *Assume that the score function $\hat{s}(Z)$ is continuously distributed. Consider n test points $Z_{m+1}, \dots, Z_{m+n}$ such that the inliers are jointly independent of each other and of the data in $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{cal}}$. Then, the conformal p-values $p_1, \dots, p_n$ are PRDS on the set of inliers.*

As a consequence, the Benjamini-Hochberg procedure applied at level $\alpha \in (0, 1)$ to conformal p-values obtained via one-class classifiers controls the FDR at level α (Corollary 2.5 in Bates et al., 2023). Bates et al. (2023) prove that also the adaptive BH procedure with the Storey estimator for the number of true nulls ensures FDR control at level α.

Theorem 9 (Theorem 2.6 in Bates et al. (2023))**.** *Set $\lambda = K/(m+1)$ for any integer K. Assume $\hat{s}$ is continuously distributed. In the setting of Theorem 8, the Benjamini-Hochberg procedure with Storey estimator for the number of true null applied at level $\alpha \in (0, 1)$ to the conformal p-values $p_1, \dots, p_n$ controls the FDR at level α.*

Subsequently, Marandon et al. (2024) establish that also conformal p-values obtained via binary classifiers are PRDS under the assumption that the scores

$$(X_1, \dots, X_m, (Y_j)_{I_0}) \tag{4.6}$$

are exchangeable with no ties almost surely.

Theorem 10 (Theorem 3.3 in Marandon et al. (2024))**.** *Assume that the scores defined in* (4.2) *have no ties almost surely and that the inlier scores in* (4.6) *are exchangeable. Then, the conformal p-values $p_1, \dots, p_n$ are PRDS on the set of inliers and the null p-values are superuniform.*

As a consequence, the Benjamini-Hochberg procedure applied at level $\alpha \in (0, 1)$ to conformal p-values obtained via binary classifiers controls the FDR at level α (Theorem 3.4 in Marandon et al., 2024).

4.3 Local Testing with Conformal p-values

This section aims to emphasize the challenge of selecting the best local test for the closed testing procedure. Several tests from Chapters 2 and 3 are revisited as local tests within the closed testing procedure, adopting a conformal perspective. The notation is adapted to emphasize the dependence on the specific intersection hypothesis H_S under consideration, whereas in previous chapters the notation reflected only the global null hypothesis.

An overview of the existing approaches for combining p-values is presented in Section 4.3.1. The methods are compared in terms of admissibility and power, with a summary provided in Table 4.1. Specifically, the Simes test (and its adaptive variant using the Storey estimator for the number of true nulls) and the Fisher combination method applied to conformal p-values are examined. In addition, the Wilcoxon-Mann-Whitney test is also considered, and it is shown that it can be expressed as the average of conformal p-values.

This analysis strengthens the rationale for adopting a data-driven approach, described in Chapter 5.

Method	Validity	Precision	Power	References
Simes[†]	PRDS	Exact if $\alpha(m+1)/\lvert S\rvert$ is an integer	Rare/Strong effects	Simes (1986), Sarkar (2008)
Fisher	Asymptotic[§] (Permutations)	Conservative (Exact)	Dense/Weak effects	Fisher (1925), Bates et al. (2023)
WMW	Permutations[‡]	Exact[¶]	Rare/Weak effects	Wilcoxon (1945), Mann and Whitney (1947b)

[†] A variant of the Simes' test incorporating Storey's estimator for the proportion of nulls has also been proven to be valid and may lead to improvement if the proportion of true nulls is not close to 1. See Appendix 4.3.2 for more details and the proof of the exactness of the Simes' test when $\alpha(m+1)/|S|$ is an integer.

[§] The asymptotic approximation proposed by Bates et al. (2023) works well even with moderate values of m and $|S|$. Empirical observations suggest that permutations can sometimes enhance power, as the asymptotic approximation tends to be somewhat conservative when both m and $|S|$ are small. Therefore, for very small samples, one may want to embed Fisher's combination statistic within a permutation test, which increases the computational cost but ensures exact finite-sample validity under minimal exchangeability assumptions. See Appendix 4.3.4 for a more detailed review.

[‡] Permutation-based methods offer exact finite-sample validity as long as the inlier scores in $(X_1, \ldots, X_m, (Y_j)_{j\in I_0})$ are exchangeable.

[¶] The critical value for the WMW test may be calculated either exactly, in small samples, or through accurate asymptotic approximations. See Appendix 4.3.5 for further details.

Table 4.1: Standard approaches for testing H_S.

4.3.1 Existing Approaches for Combining p-values

The combination of p-values dates back to the 1930s with Fisher, Tippett, and Pearson, originally under an independence assumption (see Owen, 2009, for a review).

Combination methods that do not assume independence fall into three main categories: those valid for any type of dependence, which tend to be conservative in practice (Vovk and Wang, 2020); those designed for specific types of dependence, such as the PRDS assumption (Chen et al., 2023); and those based on permutations (see Stoepker et al., 2024, for a recent example involving the Higher Criticism statistic).

Combination methods with universal validity are *inadmissible* for conformal p-values as they are outperformed by the Simes method. Indeed, Vovk et al. (2022) demonstrated that universally valid methods are dominated by the Simes method (Simes, 1986), which is applicable to PRDS p-values (Sarkar, 2008). However, beyond scenarios featuring rare and strong signals, even the Simes method is often surpassed by alternative approaches.

As proved in Section 4.3.3, the Simes method applied to conformal p-values is inad-

missible because it is dominated by the permutation-based Simes method. However, the difference between the two is minimal unless m and $|S|$ are small, which makes the Simes method appealing for computational reasons.

In general, even if permutation tests are applicable, it is unclear which test statistic should be used in practice. Birnbaum (1954) demonstrated that any combination method is optimal against some alternative. For instance, Fisher's method, recently refined by Bates et al. (2023) to accommodate the positive dependencies of conformal p-values (see Section 4.3.4), is well-suited for scenarios with numerous weak signals (Heard and Rubin-Delanchy, 2018).

Chapter 2 discusses an elegant optimality result given by Lehmann (1953), which shows that the WMW test is the locally most powerful rank test for the global null hypothesis $\theta = 0$ under a nonparametric alternative in which the test scores Y_j follow the mixture distribution $(1-\theta)F + \theta F^2$. Here, $\theta \in [0, 1]$ is the proportion of outliers and F denotes the cumulative distribution function of the calibration scores X_i.

As demonstrated in Section 4.3.5, the Mann-Whitney test statistic for testing H_S can be formulated as a combination of conformal p-values: $T_S^{\mathrm{MW}} = (m+1)\sum_{j\in S}(1-p_j)$. Since T_S^{MW} is a monotonic function of $\bar{p}_S = |S|^{-1}\sum_{j\in S} p_j$, the test statistic can be simplified to the average of conformal p-values $\bar{p}_S$, leading to rejection of H_S for small values of $\bar{p}_S$.

The *average of the p-values*, also known as Edgington's method (Edgington, 1972), has been studied by Rüschendorf (1982) and Meng (1994). They show that twice the average of the p-values is a valid p-value for arbitrary dependence. Recently, Choi and Kim (2023) demonstrated that the "twice the average" rule cannot be improved even under the assumption of exchangeability of the p-values.

In contrast, the Wilcoxon-Mann-Whitney test, when utilizing the average of the conformal p-values as the test statistic, rejects H_S if $\bar{p}_S \leq \alpha^{\mathrm{Perm}}$, where the critical value α^{Perm} is determined by taking the α quantile of the permutation distribution of $\bar{p}_S$, ensuring exact validity. This critical value α^{Perm} may be calculated either exactly in small samples or through asymptotic approximations. A well-known asymptotic approach when m and $|S|$ are both large is that of Hoeffding (1948), which leads to a much more powerful test compared to the combination approach for arbitrary dependence.

These considerations motivate the choice of the data-driven approach for selecting the

local test in the closed testing procedure.

4.3.2 The Simes Test for Conformal p-values

This section extends the discussion of the Simes test from Chapter 3 by examining its admissibility and introducing notation better suited for testing intersection hypotheses within the closed testing framework.

The Simes combining function defined in equation (3.18) is symmetric because it remains unchanged under any permutation of $(p_j)_{j\in S}$.

The Simes test (Simes, 1986) rejects H_S at level $\alpha \in (0,1)$ if and only if $\phi_S^{\text{Simes}} = 1$, where

$$\phi_S^{\text{Simes}} = \mathbb{1}\left\{T_S^{\text{Simes}} \leq \alpha\right\}. \tag{4.7}$$

The PRDS property of the conformal p-values in (4.3) implies that the Simes test applied to the conformal p-values $(p_j)_{j\in S}$ is a valid level-α test for the intersection null hypothesis H_S (Sarkar, 2008): for any given $S \subseteq [n]$, $\mathbb{P}(\phi_S^{\text{Simes}} = 1 | H_S) \leq \alpha$.

A level-α test ϕ_S is said to *dominate* another level-α test ψ_S if $\phi_S \geq \psi_S$ and the inequality is strict on a set in which the p-values lie with strictly positive probability. A test is said to be *admissible* if it is not strictly dominated by any test.

Theorem 3.1 in Vovk et al. (2022) demonstrates that the Simes test dominates all tests valid under arbitrary dependence of the p-values and relying on symmetric combining functions.

4.3.3 The Permutation-Based Simes Test

This section introduces the permutation Simes test and show that it dominates the classic Simes test.

The Simes test for conformal p-values is itself inadmissible, as it is dominated by the permutation-based Simes method.

The theory of permutation tests and related references can be found in the books Cox and Hinkley (1979), Lehmann and Romano (2005), Pesarin (2001), Pesarin and Salmaso

(2010), Pesarin and Salmaso (2025) as well as in the works of Hemerik and Goeman (2018, 2021).

Write $(W_1, \ldots, W_{m+|S|})$ for the combined sample $(X_1, \ldots, X_m, (Y_j)_{j\in S})$. Consider the permutation group Π whose elements are permutations of the set of integers $[m + |S|]$. Choose a test statistic $T = T(W_1, \ldots, W_{m+|S|})$ for which small values are to be regarded as evidence against H_S. For any $\pi \in \Pi$, let T_π denote the value of T applied to the permuted vector $(W_{\pi(1)}, \ldots, W_{\pi(m+|S|)})$). For $\alpha \in (0, 1)$, the permutation critical value is defined by:

$$\alpha^{\text{Perm}} = \max\left(a \in \{0, (T_\pi)_{\pi\in\Pi}\} : \sum_{\pi\in\Pi} \mathbb{1}\{T_\pi \leq a\} \leq \alpha|\Pi|\right). \tag{4.8}$$

Then, the permutation-based Simes test with $T = T_S^{\text{Simes}}$ in (3.18) rejects H_S at level α if and only if $\phi_S^{\text{SimesPerm}} = 1$, where

$$\phi_S^{\text{SimesPerm}} = \mathbb{1}\left\{T_S^{\text{Simes}} \leq \alpha^{\text{Perm}}\right\}. \tag{4.9}$$

A permutation test is *exactly valid*, i.e. $\mathbb{P}[\phi_S^{\text{Perm}} = 1 \ ; \ H_S] = \alpha$, for any choice of test statistic under the sole assumption that the scores $(X_1, \ldots, X_m, (Y_i)_{i\in I_0})$ are exchangeable. However, achieving exact validity for all values of α is generally not feasible due to the discreteness of the permutation distribution. Nevertheless, it is possible to ensure exactness through a mathematical artifice using a randomized critical region.

The computational cost of a permutation test may be $\mathcal{O}((m+|S|)!)$, which is prohibitive if m or $|S|$ are even moderately large. The typical solutions involve restricting the set of permutations, either by utilizing a fixed subgroup or a random subset of Π (chosen independently of the conformity scores) with an added trivial identity permutation; e.g., see Theorems 1 and 2 in Hemerik and Goeman (2018).

The precise computation of α^{Perm} for the Simes statistic requires $\binom{n+|S|}{|S|}$ permutations, representing the number of ways, disregarding order, that $|S|$ test units can be chosen from among $n + |S|$ units. However, this α^{Perm} does not depend on the scores $(W_1, \ldots, W_{m+|S|})$ but only on m, $|S|$ and α, making it possible to tabulate. This is because the permutation-based Simes test is a *rank* test. To see this, consider the conformal p-value corresponding

to the permuted vector $(W_{\pi(1)}, \ldots, W_{\pi(m+|S|)})$), that is:

$$p_j^\pi \;=\; \frac{1}{(m+1)}\left(1+\sum_{k=1}^{m}\mathbb{1}\{W_{\pi(k)} \geq W_{\pi(m+j)}\}\right).$$

This yields p_j in (4.3) for the identity permutation. Note that p_j^π does not depend on the scores $(W_1, \ldots, W_{m+|S|})$, but only on their ranks.

The next proposition demonstrates that the Simes test based on conformal p-values is inadmissible, as it is strictly dominated by the permutation-based Simes test. Moreover, the two tests coincide when $\alpha(m+1)/|S|$ is an integer.

Proposition 2. *The α-level test $\phi_S^{\text{SimesPerm}}$ in (4.9) dominates the α-level test ϕ_S^{Simes} in (4.7). The domination is strict for at least one combination of α, m, n and $S \subseteq [n]$. Moreover, it holds that*

$$\frac{|S|}{m+1}\left\lfloor \alpha\frac{m+1}{|S|}\right\rfloor \leq \mathbb{P}[\phi_S^{\text{Simes}}=1\,;\; H_S] \leq \mathbb{P}[\phi_S^{\text{SimesPerm}}=1\,;\; H_S] \leq \alpha, \tag{4.10}$$

and if $\alpha(m+1)/|S|$ is an integer, then all inequalities in (4.10) becomes equalities, i.e. the Simes test based on conformal p-values is of exact size α.

Proof. If $\alpha^{\text{Perm}} > \alpha$, then $\phi_S^{\text{SimesPerm}} \geq \phi_S^{\text{Simes}}$.
If $\alpha^{\text{Perm}} < \alpha$, then $\phi_S^{\text{SimesPerm}} \leq \phi_S^{\text{Simes}}$. The goal is to show that if $\alpha^{\text{Perm}} < \alpha$, then $\phi_S^{\text{SimesPerm}} = \phi_S^{\text{Perm}}$ for all $(p_j)_{j\in S}$. In order to derive a contradiction, suppose $\phi_S^{\text{Simes}} > \phi_S^{\text{PermSimes}}$ for some $(\tilde{p}_j)_{j\in S}$. The corresponding test statistic $\tilde{T}_S^{\text{Simes}}$ must be $\alpha^{\text{Perm}} < \tilde{T}_S^{\text{Simes}} \leq \alpha$. Since $\tilde{T}_S^{\text{Simes}} = T_{\tilde{\pi}}$ for some $\tilde{\pi} \in \Pi$, it follows that $|\Pi|^{-1}\sum_{\pi\in\Pi}\mathbb{1}\{T_\pi \leq \alpha\} > \alpha$ which results in a contradiction because $\mathbb{P}[\phi_S^{\text{Simes}}=1\,;\; H_S] \leq \alpha$.
If $\alpha^{\text{Perm}} = \alpha$, then $\phi_S^{\text{Simes}} = \phi_S^{\text{PermSimes}}$. By Corollary 3.5 in Marandon et al. (2024), if $\alpha(m+1)/|S|$ is an integer, then all inequalities in (1) becomes equalities. □

Table 4.2 below presents the size of the tests ϕ^{Simes} and $\phi^{\text{SimesPerm}}$, along with the critical value α^{Perm}, as a function of m, with $\alpha = 0.1$ and $n = 3$.

m	9	14	19	24	29	34	39	44	49	54
$\mathbb{P}(\phi_S^{\text{Simes}} = 1 \mid H_S)$	0.000	0.022	0.013	0.009	0.100	0.086	0.075	0.072	0.064	0.058
$\mathbb{P}(\phi_S^{\text{SimesPerm}} = 1 \mid H_S)$	0.055	0.025	0.014	0.009	0.100	0.097	0.083	0.072	0.065	0.058
α^{Perm}	0.200	0.133	0.100	0.080	0.100	0.171	0.150	0.111	0.100	0.091

Table 4.2: Size of the tests ϕ^{Simes} and $\phi^{\text{SimesPerm}}$, along with the critical value α^{Perm}, as a function of m, with $\alpha = 0.1$ and $n = 3$.

4.3.4 The Fisher Combination Test for Conformal p-values

An alternative classical approach for testing H_S based on conformal p-values $(p_j)_{j \in S}$ is provided by the Fisher combination method (Fisher, 1925), discussed in Chapter 3. This method originally assumed independence but was later refined by Bates et al. (2023) to accommodate the PRDS exhibited by conformal p-values.

Assuming that the scores $S_1, \ldots, S_m, S_{m+1}, \ldots, S_{m+n}$ are continuously distributed, Bates et al. (2023) prove that for any function $G : [0,1] \to \mathbb{R}$ such that $G(U)$ has finite moments with U a standard uniform random variable, the correlation between any pair of transformed null p-values $G(p_j)$ and $G(p_h)$ is equal to $1/(m+2)$ for any $j, h \in [n_0]$. That is,

$$\text{Cor}\,(p_j, p_h) = \frac{1}{m+2}.$$

Fisher combination method is a special case with $G(u) = -2\log(u)$ and $G(U)$ following a Chi-squared distribution with 2 degree of freedom. Based on this result, considering the global null $H_{[n]}$ they show that the variance of the Fisher combination method applied to null conformal p-values is inflated by a factor $1+(n-1)/(m+2)$ compared to the variance of the Fisher combination test applied to p-values that are i.i.d. Indeed, for any finite-valued function $G : [0,1] \to \mathbb{R}$ and assuming that all the hypotheses are true:

$$\begin{aligned}\text{Var}\left(\sum_{j \in [n]} G(p_j)\right) &= \left(1 + \frac{n-1}{m+2}\right) n\text{Var}(G(p_1)) \\ &\underset{m,n \to \infty}{\approx} (1+\gamma) n\text{Var}(G(p_1)),\end{aligned}$$

provided that $\gamma = \lim_{m,n\to\infty} n/m \in (0, +\infty)$.

Adapting the notation to a generic intersection hypothesis H_S, the corrected Fisher

combination test rejects H_S at level α if and only if $\phi_S^{\text{Fisher}} = 1$, where

$$\phi_S^{\text{Fisher}} = \mathbb{1}\left\{-2\sum_{j\in S}\log(p_j) > c_\alpha^{\chi^2}(|S|)\sqrt{1+|S|/m} - 2|S|\left(\sqrt{1+|S|/m}-1\right)\right\}. \quad (4.11)$$

Bates et al. (2023) proved that this test is asymptotically valid in the limit of large $|S|$ and m, with both sample sizes growing at the same rate. This differs from the classical Fisher test (Fisher, 1925), which assumes the p-values are mutually independent and can be recovered from (4.11) by letting $m \to \infty$ while holding $|S|$ fixed.

4.3.5 The Wilcoxon-Mann-Whitney Rank Test

The Wilcoxon-Mann-Whitney test, presented in Chapter 3, is one of the most well-known rank tests. Here, the notation is adapted to better accommodate testing intersection hypotheses within the closed testing framework. A result from Edgington (1972) that expresses the Wilcoxon-Mann-Whitney statistic as a function of the mean of conformal p-values is reviewed.

For any $S \subseteq [n]$ and any $j \in S$, let R_{m+j} denote the rank of Y_j among $(X_1, \ldots, X_m, (Y_j)_{j\in S})$. The dependence of R_{m+j} on S is left implicit to simplify the notation, since this does not create ambiguity. This classical two-sample Wilcoxon (Wilcoxon, 1945) or Mann-Whitney (Mann and Whitney, 1947b) test rejects H_S for large values of the sum of ranks in the test sub-sample S

$$T_S^{\text{W}} = \sum_{j\in S} R_{m+j}, \quad (4.12)$$

or for large values of the U-statistic

$$T_S^{\text{MW}} = \sum_{j\in S}\sum_{i=1}^{m} \mathbb{1}\{X_i < Y_j\}. \quad (4.13)$$

The Mann Whitney U-statistic can be expressed as $T_S^{\text{MW}} = T_S^{\text{W}} + |S|(|S|+1)/2$, differing only by a constant term. The two formulations result in the same test: the Wilcoxon-

Mann-Whitney test. The WMW level-α test is given by

$$\phi_S^{\mathrm{WMW}} = \mathbb{1}\left\{T_S^{\mathrm{W}} \geq c_\alpha^{\mathrm{W}}(m, |S|)\right\} = \mathbb{1}\left\{T_S^{\mathrm{MW}} \geq c_\alpha^{\mathrm{MW}}(m, |S|)\right\}. \tag{4.14}$$

The critical values $c_\alpha^{\mathrm{W}}(m, |S|)$ and $c_\alpha^{\mathrm{MW}}(m, |S|)$ are the $(1-\alpha)$ quantiles of the permutation distribution of T_S^{W} and T_S^{MW}, respectively. Since the two test statistics result in the same test, it will be denoted by T_S^{MWW} and the critical value by $c_\alpha^{\mathrm{WMW}}(m, |S|)$, consistently with Chapter 2.

For small samples, the permutation null distribution of T_S^{WMW} can be found either via recursion (Mann and Whitney, 1947b) or direct permutations. The test ϕ_S^{WMW} is of exact size α, i.e. $\mathbb{P}(\phi_S^{\mathrm{WMW}} = 1|H_S) = \alpha$, for $\alpha \in \Lambda = \{a_r, r \in [m|S|]\}$, where $a_r = \mathbb{P}(T_S^{\mathrm{WMW}} \geq r|H_S)$; for $\alpha \in (0,1) \setminus \Lambda$, it is conservative, i.e. $\mathbb{P}(\phi_S^{\mathrm{WMW}} = 1|H_S) < \alpha$.

A well-known approach when m and $|S|$ are both large is that of Hoeffding (1948), which is based on an application of the Central Limit Theorem for U-statistics. Under the null hypothesis H_S, for large m and $|S|$, the Mann Whitney statistic T_S^{MW} is approximately normally distributed:

$$T_S^{\mathrm{MW}} \approx N\left(\frac{|S|m}{2}, \frac{m|S|(m+|S|+1)}{12}\right). \tag{4.15}$$

Other asymptotic approximations include the Edgeworth expansion of Fix and Hodges (1955) and the uniform approximation of Buckle et al. (1969).

It is also interesting to note that the Mann-Whitney statistic T_S^{MW} reduces to $T_j^{\mathrm{MW}} = \sum_{i=1}^m \mathbb{1}\{X_i < Y_j\}$ if $S = \{j\}$, for any $j \in [n]$. The statistic T_j^{MW} can be equivalently expressed as a function of the conformal p-value p_j in (4.3):

$$T_j^{\mathrm{MW}} = \sum_{i=1}^m \mathbb{1}\{X_i < Y_j\} = m - \sum_{i=1}^m \mathbb{1}\{X_i \geq Y_j\} = (m+1)(1-p_j) \tag{4.16}$$

When $S = \{j\}$, the WMW two-sample test, with the 1st sample being $X_1, \ldots, X_n$ and the second "sample" being just Y_j, rejects H_j if the rank of Y_j in the sequence $(X_1, \ldots, X_n, Y_j)$ is large or, equivalently, if the conformal p-value p_j in (4.3) is small. Therefore, the WMW test provides a particularly intuitive bridge between the modern framework of conformal

inference and the classical world of two-sample rank tests (Kuchibhotla, 2021).

Another interesting connection, formally stated in the next proposition, is that the Mann-Whitney test for H_S can be defined by the average of conformal p-values test statistic, also known as Edgington's method (Edgington, 1972).

Proposition 3. *The Mann-Whitney test statistic T_S^{MW} can be equivalently expressed as a function of the conformal p-values p_j in* (4.3)*:*

$$T_S^{MW} = (m+1)\sum_{j\in S}(1-p_j).$$

Then the WMW test can be defined by the average of conformal p-values test statistic, also known as Edgington's method (Edgington, 1972):

$$\bar{p}_S = \frac{1}{|S|}\sum_{j\in S} p_j.$$

Proof. Note that The Mann-Whitney test statistic T_S^{MW} is the sum of the individual contributions T_j^{MW} with $j \in S$, i.e. $T_S^{\mathrm{MW}} = \sum_{j\in S} T_j^{\mathrm{MW}}$. Then $T_S^{\mathrm{MW}} = (m+1)\sum_{j\in S}(1-p_j)$ follows from (4.16). Since T_S^{MW} is a monotonic function of $\bar{p}_S = |S|^{-1}\sum_{j\in S} p_j$, the test statistic can be simplified to the average of conformal p-values $\bar{p}_S$, leading to rejection of H_S for small values of $\bar{p}_S$. □

Chapter 5

ACODE: Automatic Conformal Outlier Detection and Enumeration

5.1 Introduction

This chapter details ACODE, a novel method for collective outlier detection that was introduced in Chapter 1. Sections 5.1.1 and 5.1.2 provide a brief literature review on outlier detection in conformal inference and state the problem that ACODE addresses.

Section 5.2 describes the method ACODE for estimating the number of outliers via a *conformalized* closed testing procedure, detailing how the automatic selection of the classifier and local test is incorporated into this novel approach.

Section 5.3 discusses local tests for outlier detection within closed testing. Section 4.3.1 provides a brief and intuitive summary of the comparison among existing approaches for combining p-values, detailed in Section 4.3. Section 5.3.2 revisits the Shiraishi test from Chapter 2 with notation adapted to emphasize dependence on H_S, and extends its asymptotic normality from independence to exchangeability of scores. Section 5.3.3 details the implementation of the adaptive Shiraishi test using the "test set dilution" technique from Chapter 4. The adaptive statistic is proved to retain its asymptotically normal null distribution under mild conditions on the density estimator, and propose an intuitive estimation method for the outlier density. Section 5.3.4 compares the Shiraishi test with an (idealized) oracle Neyman-Pearson test in terms of power.

5.1.1 Research Background

Conformal inference, introduced in Chapter 4, is an active research topic, broadly seeking reliable uncertainty estimation for the predictions output by black-box machine learning models. In the context of outlier detection (Laxhammar and Falkman, 2015; Guan and Tibshirani, 2022), prior works focused on individual-level identification, often aiming to control the FDR (Bates et al., 2023; Marandon et al., 2024). By contrast, the focus of this thesis is on collective detection and enumeration. The proposed method tends to lead to more informative results when dealing with weak or sparse signals that are difficult to localize. This approach is also connected to a broader literature on multiple testing and large-scale inference.

Recent advancements in large-scale inference have led to methods for signal detection and estimation of the proportion of non-null effects (Cai and Sun, 2017). Donoho and Jin (2004) proposed an elegant solution for signal detection using Tukey's Higher Criticism statistic. Meinshausen and Rice (2006) extended the Higher Criticism method to provide a lower confidence bound for the proportion of non-null effects. Goeman et al. (2021) further advanced this approach by embedding Higher Criticism into a closed testing procedure (Marcus et al., 1976), providing simultaneous bounds for the false discovery proportion (FDP). However, the FDP bounds derived from Higher Criticism are valid only under the assumption of independent p-values, which does not hold in the conformal inference setting.

Since the works by Genovese and Wasserman (2006) and Goeman and Solari (2011), there has been significant growth in methods providing simultaneous FDP bounds (Goeman et al., 2019; Blanchard et al., 2020; Katsevich and Ramdas, 2020; Tian et al., 2023; Heller and Solari, 2023). These have been utilized across various domains, including neuroimaging (Rosenblatt et al., 2018; Goeman et al., 2023) and genomics (Ebrahimpoor et al., 2020), as well as in permutation-based (Hemerik et al., 2019; Andreella et al., 2023; Blain et al., 2022; Vesely et al., 2023) and knockoff-based (Li et al., 2024) approaches.

In this contribution, conformal inference is integrated with closed testing to construct simultaneous lower confidence bounds for the number of outliers in any subset of the test set. A central challenge is selecting an effective local testing procedure suited to the data at hand. Recent findings show that a closed testing procedure is *admissible* if and only

if all its *local tests* are admissible (Goeman et al., 2021), underscoring the importance of designing a closed testing approach with admissible local tests.

A typical concern with closed testing is the computational cost, generally exponential. However, efficient polynomial-time *shortcuts* are often available (Goeman et al., 2019; Dobriban, 2020; Tian et al., 2023), as discussed in Section 3.4.3, making the proposed method feasible and scalable.

5.1.2 Problem Statement

Prior work on conformal inference for outlier detection (Bates et al., 2023; Marandon et al., 2024) has primarily focused on testing individual-level hypotheses H_j for each $j \in [n]$, with the goal of controlling the FDR. However, as discussed in Section 5.1.1, there are many scenarios where the power to reject these individual-level hypotheses H_j is very low, highlighting the need for an alternative approach.

In this chapter, two key tasks related to collective outlier detection and enumeration are addressed. The first task, *outlier detection*, involves testing the potentially easier-to-reject intersection hypothesis $H_S := \cap_{j \in S} H_j$, which posits that a subset $S \subseteq [n]$ of test points contains no outliers. The second task, *outlier enumeration*, focuses on estimating the number of outliers within a subset $S \subseteq [n]$, specifically by providing a lower confidence bound, without necessarily identifying them. For both tasks, cases where S is either fixed or data-driven are considered. In the latter scenario, the closed testing principle (Marcus et al., 1976) is utilized to obtain simultaneous inferences that remain valid for all possible subsets S. This approach gives practitioners flexible tools for interactive data analysis.

To deal with the possibly high-dimensional nature of the data and avoid parametric assumptions about their distribution, the aforementioned tasks will be tackled by using the conformal inference framework described in Chapter 4 adopting the setting and the model described in Chapter 4.2.

5.2 Estimating the Number of Outliers via Conformalized Closed Testing

This section presents Algorithm 6, which summarizes the closed testing method based on local tests that take as input the calibration and test scores to either built the ranks or the conformal p-values associated to the test observations. Therefore, Algorithm 6 outputs a simultaneous lower bound $d(S)$ guaranteed to satisfy (3.16).

Valid conformity scores can be obtained through various methods. A common approach involves using a one-class classifier trained on an independent set of inliers (Bates et al., 2023). Alternatively, one can employ a binary classifier via positive-unlabeled learning (Marandon et al., 2024). Another option arises when labeled outliers are available; here, the scores can be generated using a binary classifier as discussed in Liang et al. (2024). For simplicity, only scenarios where all labeled data consist of inliers are examined. The case where the reference set might be contaminated with unlabeled or labeled outliers represents an important area for future research as emphasized in Bashari et al. (2025), and the strategy proposed by Liang et al. (2024) would provide a valuable starting point.

Algorithm 6: Conformal Outlier Detection and Enumeration

Input: Inlier data $\mathcal{D}^{\text{train}} = \{Z_1^{\text{train}}, \ldots, Z_{m_{\text{train}}}^{\text{train}}\}$ and $D^{\text{cal}} = \{Z_1, \ldots, Z_m\}$.
Test data $D^{\text{test}} = \{Z_{m+1}, \ldots, Z_{m+n}\}$. Significance level $\alpha \in (0, 1)$.
Machine learning algorithm $\mathcal{A}$ for one-class or binary classification.
Chosen local testing method ϕ; e.g., Simes, WMW, etc.

if $\mathcal{A}$ *is a one-class classification algorithm* **then**
  Train $\mathcal{A}$ using the data in $(Z_1^{\text{train}}, \ldots, Z_{m_{\text{train}}}^{\text{train}})$.
else if $\mathcal{A}$ *is a binary classification algorithm* **then**
  Train $\mathcal{A}$ using $(Z_1^{\text{train}}, \ldots, Z_{m_{\text{train}}}^{\text{train}})$ and $\{Z_1, \ldots, Z_m, Z_{m+1}, \ldots, Z_{m+n}\}$.
Apply $\mathcal{A}$ to evaluate the calibration and test scores $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_n)$, and compute the ranks and the conformal p-values as in equation (4.3).
For any desired $S \subseteq [n]$, compute $d(S)$ using Algorithm 2 (ref. Section 3.4.3) and the local testing method ϕ applied to the ranks or the conformal p-values.

Output: A $(1 - \alpha)$ lower bound $d(S)$ for the number of outliers in any $S \subseteq [n]$.

Proposition 1 in Chapter 3 implies Algorithm 6 produces valid inferences as long as the local test ϕ_{I_0} is valid, and the proposed method is precisely designed to achieve this (ref. Section 5.3). While the lower bound $d(S)$ output by Algorithm 6 is *simultaneously* valid for all possible subsets $S \subseteq [n]$, in the sense of (3.16), this does not mean Algorithm 6 needs to explicitly output $d(S)$ for all $S \subseteq [n]$. On the contrary, one would typically apply Algorithm 6 focusing on a particular (but possibly data-driven) choice of S, as demonstrated in Section 6. In any case, the components of Algorithm 6 involving the model training and the computation of the conformity scores only need to be applied once, irrespective of S.

It is important to note that the flexibility of Algorithm 6, which can accommodate a variety of classifiers and testing procedures, introduces some challenges. In particular, different classification algorithms and testing procedure may result in significant performance variation across different data sets, and it is unclear a priori how to maximize power. Unfortunately, leaving too much latitude to practitioners is not always desirable, as it may inadvertently encourage "cherry picking" behaviors that, as previously illustrated in Figure 1.1, can result in invalid inferences. This issue motivates the extension introduced in the next section. This extension will channel the flexibility of Algorithm 6 into a principled, automatic method that often achieves competitively high power in practice while adding a layer of protection against the risks of human-driven selection bias.

5.2.1 Data-Driven Tuning

Algorithm 6 is extended to leverage a potentially diverse suite of classifiers and local testing procedures. Algorithm 7 is proposed to approximately maximize power subject to type-I error control. The algorithm is inspired by similar approaches proposed by Liang et al. (2024) and Marandon et al. (2024) in the context of individual outlier identification under FDR control. The idea is to embed the data-driven tuning step within an additional layer of sample splitting. As detailed below, this is crucial for maintaining the validity of local tests after selecting the most effective classifier and local testing method.

The initial tuning module of Algorithm 7 sees the calibration and test data only through the lenses of the unordered collection $\{Z_1, \ldots, Z_m\} \cup \{Z_{m+1}, \ldots, Z_{m+n}\}$. Consequently, when Algorithm 6 is applied again in the inference step of Algorithm 7, using $(Z_1, \ldots, Z_m)$ as a calibration set and $(Z_{m+1}, \ldots, Z_{m+n})$ as a test set, the scores $(Z_1, \ldots, Z_{m+n})$ corre-

Algorithm 7: Automatic Conformal Outlier Detection and Enumeration

Input: Inlier data $\mathcal{D}^{\text{train}} = \{Z_1^{\text{train}}, \ldots, Z_{m_{\text{train}}}^{\text{train}}\}$, $D^{\text{cal}} = \{Z_1, \ldots, Z_m\}$, $\mathcal{D}^{\text{tune}} = \{Z_1^{\text{tune}}, \ldots, Z_{m_{\text{tune}}}^{\text{tune}}\}$.
Test data $D^{\text{test}} = \{Z_{m+1}, \ldots, Z_{m+n}\}$, Significance level $\alpha \in (0, 1)$.
A list of algorithms $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(K)}$ for one-class or binary classification.
A list of local testing methods $\phi^{(1)}, \ldots, \phi^{(L)}$; e.g., Simes, WMW, etc.

Define $\tilde{\mathcal{D}}^{\text{test}} = \{Z_1, \ldots, Z_m\} \cup \{Z_{m+1}, \ldots, Z_{m+n}\}$.
for ***any*** $S \subseteq [n]$ **do**
  for *each* $k \in [K]$ **do**
    for *each* $l \in [L]$ **do**
      Compute $d_{k,l}(S)$ by applying Algorithm 6 based on $\mathcal{A}^{(k)}$ and $\phi^{(l)}$, using $\mathcal{D}^{\text{train}}$ for training, $\mathcal{D}^{\text{tune}}$ for calibration, and $\tilde{\mathcal{D}}^{\text{test}}$ as the test set.
      Compute $\hat{p}_{k,l}(S)$, the p-value for H_S based on $\mathcal{A}^{(k)}$ and $\phi^{(l)}$.
  Find $(\hat{k}(S), \hat{l}(S)) = \arg\max_{k \in [K], l \in [L]} d_{k,l}(S)$; if not unique, choose $(\hat{k}(S), \hat{l}(S)) = \arg\min_{k \in [K], l \in [L]} \hat{p}_{k,l}(S)$.
  Compute $d(S)$ by applying Algorithm 6 based on $\mathcal{A}^{(\hat{k}(S))}$ and $\phi^{(\hat{l}(S))}$, using $\mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{tune}}$ for training, $\mathcal{D}^{\text{cal}}$ for calibration, and $\mathcal{D}^{\text{test}}$ as the test set.

Output: A $(1 - \alpha)$ lower bound $d(S)$ for the number of outliers in any $S \subseteq [n]$.

sponding to inliers are still exchangeable conditional on the selected classifier and testing procedure. Thus, Algorithm 7 can often achieve high power without selection bias, as previewed in Figure 1.1 and confirmed in Section 6. Formally, Algorithm 7 also outputs a simultaneous lower bound $d(S)$ guaranteed to satisfy (3.16), as long as the local tests are valid.

There is an interesting distinction between exchangeability and independence in the context of Algorithm 7. Recall from Sections 4.2.1 and 5.1.2 that the model in (4.1) and the approach described in (4.2) lead to i.i.d. scores for the inliers in the case of one-class classification, and only exchangeable scores in the case of binary classification (Marandon et al., 2024).

Algorithm 7 changes this picture a little. When conditioning on the selected classifier and closed testing procedure, the scores of the inliers among $(Z_1, \ldots, Z_{m+n})$, although still exchangeable, can no longer be independent, irrespective of whether they were initially obtained through one-class or binary classification. This is because such conditioning is similar to (but weaker than) conditioning on the unordered collection $\{Z_1, \ldots, Z_m\} \cup$

$\{Z_{m+1}, \ldots, Z_{m+n}\}$, an operation that transforms the distribution of $(Z_1, \ldots, Z_{m+n})$ into a finite-population distribution.

Importantly, the inference procedure proposed remains valid without assuming independence, relying only on the weaker condition of exchangeability among inliers. As shown in Proposition 2.2 of Gazin et al. (2024), the joint distribution of null conformal p-values is identical under exchangeability and i.i.d. conditions. Furthermore, Theorem 12 shows that the asymptotic distribution of the Shiraishi statistic is unaffected by this distinction.

5.3 Local Tests for Outlier Detection

Consider the null hypothesis, denoted as H_S, that a *fixed* subset $S \subseteq [n]$ of test points contains no outliers. As outlined in Section 4.3, there are many tests available for this hypothesis, each with unique strengths and weaknesses. Given the limited scope of existing optimality results and the unfeasibility of predicting which method will perform best on a specific data set, ACODE is designed to integrate a *toolbox* of such testing procedures, including standard aggregation methods of conformal p-values, such as the Simes test and the Fisher combination method (discussed in Chapter 4), and the rank tests of Shiraishi and Wilcoxon-Mann-Whitney. It then selects the most effective approach in a principled data-driven manner, as explained in Section 5.2.1.

Both in the case of standard aggregation methods of conformal p-values and in the case of rank tests, the local tests are built based on the conformity scores. When using binary classifiers to extract the scores, the resulting scores are no longer independent and retain only exchangeability and this might change the distribution for the considered tests, hinfering the validity of these tests.

5.3.1 The Simes Test and Fisher Combination Method under Exchangeability

This section presents a result in Gazin et al. (2024) that proves that the joint distribution of the conformal p-values remains the same whether the inlier scores are i.i.d and or merely exchangeable. Their distribution is detailed in Proposition 2.2 in Gazin et al. (2024).

Theorem 11 (Proposition 2.2 in Gazin et al. (2024))**.** *Assume that the scores* $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ *defined in* (4.2) *are exchangeable and have no ties almost surely. Then, the conformal p-values* $p_1, \ldots, p_n$ *have joint distribution* $P_{m,n}$ *which is defined as:*

$$\begin{cases} (p_1, \ldots, p_n \mid U) \overset{iid}{\sim} P^U, \\ U = (U_1, \ldots, U_m) \overset{iid}{\sim} U([0,1]), \end{cases} \tag{5.1}$$

where P^U *is the discrete distribution on* $[0,1]$ *such that* $P^U(K/(m+1)) = U_{(K)} - U_{K-1}$ *for any* $K \in [m+1]$.

As a consequence, the distribution of the Simes test and the Fisher combination method remains the same whether the conformal p-values are built on scores from one-class classifiers or binary classifiers. This result is crucial for the validity of ACODE when these methods are applied.

5.3.2 The Shirashi Test under Exchangeability

This section proves the validity of the Shiraishi test under the sole assumption of score exchangeability.

Recall the model from equation (2.21) in Chapter 2, where the Shiraishi statistic was first introduced. As a starting point, let us assume the outlier distribution G is known. In practice, however, the test discussed in this section can be implemented using a data-driven estimate $\hat{G}$ of G, as detailed in Section 5.3.3. Importantly, this test is proved to be valid when using $\hat{G}$ in place of G, provided that $\hat{G}$ is estimated in a manner that preserves the exchangeability between the calibration and test scores.

Under this model, the test based on the Shiraishi statistic T^g in equation (2.21) rejects $H : \theta = 0$ at level $\alpha \in (0,1)$ if T^{g} is larger than a suitable critical value $c_\alpha^{\mathrm{g}}(m,n)$, discussed later. The indicator of this rejection event is:

$$\phi^{\mathrm{g}} = \mathbb{1}\left\{T^{\mathrm{g}} > c_\alpha^{\mathrm{g}}(m,n)\right\}. \tag{5.2}$$

The critical value $c_\alpha^{\mathrm{g}}(m,n)$ in (5.2) can be obtained assuming only the exchangeability of the scores either through a Monte Carlo simulation of the permutation distribution—

feasible for small sample sizes—or via an asymptotic normal approximation (Shiraishi, 1985) for large sample sizes under the additional assumption of independence. This result of Shiraishi (1985) is discussed in Section 2.5.

In order to overcome the strong assumption of independence, the following theorem extends the result of Shiraishi (1985) relying solely on the exchangeability of the pooled score vector. Accordingly, the null hypothesis

$$H' : (X_1, \ldots, X_m, Y_1, \ldots, Y_n) \text{ is exchangeable,}$$

which is implied by $H : \theta = 0$ in the setting of Theorem 1, is considered in the next theorem. This result holds without requiring assumptions about the score model.

Theorem 12. *Assume $g : [0,1] \to \mathbb{R}$ is bounded and $\lim_{N\to\infty} n/N = \delta$, with $0 < \delta < 1$. Under the null hypothesis H', the rescaled Shiraishi test statistic*

$$\frac{\sqrt{N(N-1)}(T^{\mathrm{g}} - n\mu_N)}{\sqrt{nm \sum_{h\in[N]} (\mathbb{E}[g(U_N^{(h)})] - \mu_N)^2}} \tag{5.3}$$

with $\mu_N = N^{-1} \sum_{h\in[N]} \mathbb{E}[g(U_N^{(h)})]$ converges in distribution to the standard normal as $m, n \to \infty$.

Proof. Assume the notation from Theorem 1, and denote the pooled score vector $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ by $(W_1, \ldots, W_N)$. By Theorem 2 in Kuchibhotla (2021), under the null hypothesis H', the vector of ranks $(\mathrm{rank}(W_1), \ldots, \mathrm{rank}(W_N))$ is exchangeable and uniformly distributed over all permutations of $[N]$. Specifically, for any permutation $(\pi(1), \ldots, \pi(N))$ of $[N]$, it holds:

$$\mathbb{P}(\mathrm{rank}(W_1) = \pi(1), \ldots, \mathrm{rank}(W_N) = \pi(N) \; ; \; H') = \frac{1}{N!}.$$

This implies that the distribution of the rank vector is distribution-free under H', meaning it does not depend on the specific distribution of $(W_1, \ldots, W_N)$. In particular, this holds when $W_1, \ldots, W_N$ are i.i.d.

As a result, the permutation null distribution of the test statistic $T^{\mathrm{g}} = T^{\mathrm{g}}(\mathrm{rank}(W_1), \ldots, \mathrm{rank}(W_N))$, including its $(1-\alpha)$-quantile $c_\alpha^{\mathrm{g}}(m, n)$, depends only on the function g and the sample sizes

m and n, not on the underlying distribution of the W_i's. The result follows by assuming that $(W_1, \ldots, W_N)$ are i.i.d. with continuous c.d.f. F, and applying Theorem 1 from Shiraishi (1985).

□

As a consequence, for any $S \subseteq [n]$ with cardinality s, the test statistic

$$T_S^{\mathrm{g}} = \sum_{j \in S} \mathbb{E}[g(U_{m+s}^{(R_S^{m+j})})], \tag{5.4}$$

where R_S^{m+j} is the rank of Y_j among $(X_1, \ldots, X_m, (Y_j)_{j \in S})$ and $U_{m+s}^{(R_S^{m+j})}$ is the R_S^{m+j}-th order statistic in a sample of Uniform random variables of size $m + s$, is asymptotically normal with mean

$$\mu_{m+s} = \frac{s}{m+s} \sum_{h \in [m+s]} \mathbb{E}[g(U_{m+s}^{(h)})]$$

and variance

$$\sigma_{m+s}^2 = \frac{sm}{m + s(m+s-1)} \sum_{h \in [m+s]} \left(\mathbb{E}[g(U_{m+s}^{(h)})] - \frac{\mu_{m+s}}{s} \right)^2 .$$

The oracle procedure studied in Theorem 12 can be translated into a practical test by replacing the unknown alternative density g with a suitable empirical estimate $\hat{g}$, as explained in Section 5.3.3. As shown in Chapter 6, in some scenarios of interest this data-driven approximation of the oracle Shiraishi test can achieve much higher power compared to other local testing methods, such as Simes', Fisher's, and the WMW test. In any case, the true strength of ACODE lies in its flexibility, as it does not depend on any single testing procedure but can instead dynamically select the most effective approach based on the data at hand. The following section describes the "test set dilution" technique, introduced in Chapter 4, which is crucial for the implementation of the adaptive Shiraishi test and for the validity of ACODE while automatically selecting the best classifier and local test.

5.3.3 Implementation of the Adaptive Shiraishi Statistic

This section details how to empirically estimate the outlier distribution G and its density function g, which is essential for implementing a data-driven version of Shirashi's local testing procedure. Shirashi's optimality result is based on the following mixture model:

$$\begin{aligned} X_1, \ldots, X_m &\overset{\text{i.i.d.}}{\sim} F, \\ Y_1, \ldots, Y_n &\overset{\text{i.i.d.}}{\sim} (1-\theta)F + \theta G(F), \end{aligned} \tag{5.5}$$

where $\theta \in [0,1]$ represents the proportion of outliers, and G is a distribution function on $[0,1]$. In the following, a practical and effective approach is described for computing an estimate $\hat{G}$ of G using the data in $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$, while ensuring that an approximately valid and approximately LMPI test can still be performed using Shirashi's method conditional on $\hat{G}$.

5.3.3.1 Diluting the Test Set for Conditional Exchangeability

Section 4.2.1 introduces the "test set dilution" technique and discusses the theoretical reasons why it enables the use of test observations while maintaining score exchangeability. While this technique also serves to select the classifier and local test, here the focus is on its application to outlier density estimation.

The key idea, inspired by approaches proposed in different contexts by Marandon et al. (2024) and Liang et al. (2024), is to randomly split the calibration sample $(X_1, \ldots, X_m)$ into two disjoint subsets, namely $(X_1, \ldots, X_{m_1})$ and $(X_{m_1+1}, \ldots, X_m)$, for some choice of m_1 such that $1 < m_1 < m$, like $m_1 = \lceil m/2 \rceil$. Intuitively, $(X_1, \ldots, X_{m_1})$ is used as a smaller calibration set and $(X_{m_1+1}, \ldots, X_m, Y_1, \ldots, Y_n)$ as a "diluted" test set when fitting G.

More precisely, the estimation of G and its density function g is carried out by fitting the following mixture model:

$$\begin{aligned} X_1, \ldots, X_{m_1} &\overset{\text{i.i.d.}}{\sim} F, \\ X_{m_1+1}, \ldots, X_m, Y_1, \ldots, Y_n &\overset{\text{i.i.d.}}{\sim} (1-\theta_1)F + \theta_1 G(F), \end{aligned} \tag{5.6}$$

where $\theta_1 = (\theta n + m_1)/(n + m_1)$ and G is the same outlier distribution function as in equa-

tion (5.5). Any approach may be applied to fit this mixture model, as long as it is invariant to the ordering of the data in $(X_{m_1+1}, \ldots, X_m, Y_1, \ldots, Y_n)$. Otherwise, it is sufficient to permute the scores of the diluted test set before estimating g. The motivation for this method is that it ensures the inlier scores among $(Y_1, \ldots, Y_n)$ remain exchangeable with the smaller calibration set $(X_{m_1+1}, \ldots, X_m)$ conditional on the estimated $\hat{G}$ (equivalently, $\hat{g}$). Therefore, any downstream conformal inferences that utilizes $(X_{m_1+1}, \ldots, X_m)$ as a reference (or calibration) set are still valid. This is akin to the idea underlying AdaDetect, the individual outlier identification method proposed by Marandon et al. (2024).

A subtle point worth highlighting is that the asymptotic critical values for Shirashi's test outlined in Section 5.3.2 were derived under the assumption that all inliers are i.i.d. (5.5), whereas conditional on $\hat{G}$ (equivalently, $\hat{g}$) the inliers among $(Y_1, \ldots, Y_n)$ are only exchangeable with $(X_{m_1+1}, \ldots, X_m)$ but not independent. However, Theorem 12 ensures the validity of the proposed approach under the sole assumption of exchangeability.

5.3.3.2 A Mixture Modeling Approach

One well-established approach in the literature for estimating one component of a two-component mixture model is that proposed by Patra and Sen (2016), who also studied the mathematical properties of their estimator. The procedure is detailed as follows.

1. Estimate the inlier distribution F from $X_1, \ldots, X_{m_1}$ with the empirical c.d.f. $\hat{F}$.

2. Apply an inverse CDF transform based on the fitted $\hat{F}$ distribution to $(X_{m_1+1}, \ldots, X_m, Y_1, \ldots, Y_n)$, making the inlier distribution approximately uniform on $[0, 1]$. Let $(\tilde{X}_{m_1+1}, \ldots, \tilde{X}_m, \tilde{Y}_1, \ldots, \tilde{Y}_n)$ denote the transformed test set.

3. Fit the following model:

$$\tilde{X}_{m_1+1}, \ldots, \tilde{X}_m, \tilde{Y}_1, \ldots, \tilde{Y}_n \overset{\text{i.i.d.}}{\sim} (1 - \theta_1)\text{Uniform}(0, 1) + \theta_1 G, \tag{5.7}$$

   using the estimators for θ_1 and G proposed by Patra and Sen (2016), obtaining $\hat{\theta}_1$ and $\hat{G}$, respectively. From $\hat{G}$ derive the outlier density estimator $\hat{g}$ with the method in Patra and Sen (2016).

4. Output the estimated outlier distribution $\hat{G}$ and its density $\hat{g}$.

Additionally, a simple and intuitive approach that has proven effective in practice is presented here.

1. Rescale the data, $X_1, \ldots, X_m, Y_1, \ldots, Y_n$, to be between 0 and 1, if they are not already scaled as such (many standard classifiers output scores in this range by default).

2. Fit a Beta(b_1, b_2) distribution to $X_1, \ldots, X_{m_1}$ via maximum likelihood. To avoid numerical instabilities from values too close to 0 or 1, threshold all data points between 0.001 and 0.999. Let $\hat{b}_1$ and $\hat{b}_2$ be the estimated parameters.

3. Apply an inverse CDF transform based on the fitted Beta$(\hat{b}_1, \hat{b}_2)$ distribution to $(X_{m_1+1}, \ldots, X_m, Y_1, \ldots, Y_n)$, making the inlier distribution approximately uniform on $[0, 1]$. Let $(\tilde{X}_{m_1+1}, \ldots, \tilde{X}_m, \tilde{Y}_1, \ldots, \tilde{Y}_n)$ denote the transformed test set.

4. Fit the mixture model in equation (5.7) via maximum likelihood, where G is approximated by a Beta(b_1', b_2') distribution. The estimated parameters are denoted as $\hat{b}_1'$ and $\hat{b}_2'$.

5. Output the estimated outlier distribution: $\hat{G} \approx \text{Beta}(\hat{b}_1', \hat{b}_2')$ and its density $\hat{g}$.

The invariance of this estimation procedure to the ordering of $(\tilde{X}_{m_1+1}, \ldots, \tilde{X}_m, \tilde{Y}_1, \ldots, \tilde{Y}_n)$ ensures the validity the Shiraishi test regardless of how closely $\hat{G}$ approximates the true G. Moreover, as demonstrated in the numerical experiments in Chapter 6, this estimation method performs well enough in practice to provide this testing method with a noticeable advantage over standard approaches, such as the WMW test, particularly in scenarios where the oracle Shiraishi test has a significant edge.

5.3.3.3 Enabling Closed Testing Shortcuts for Monotone Statistics

When applied within a closed testing framework, Shirashi's testing procedure can become computationally expensive without an appropriate shortcut. Section 3.4.3 discusses a fast and exact shortcut, which relies on the monotonicity of the outlier probability density—specifically, the derivative of the function $\hat{G}$ estimated above. This motivates the need for an estimation procedure that guarantees a monotone outlier density.

Although more sophisticated approaches have been studied to address this issue (e.g., Patra and Sen, 2016), a simple post-hoc solution that works well for the purposes of this thesis is to approximate the probability density function of the Beta$(\hat{b}'_1, \hat{b}'_2)$ distribution, obtained as described in the previous section, with a monotone increasing function using isotonic regression on a finite grid of values. Since it may not be clear in advance whether the best approximation of the outlier density is increasing or decreasing, the direction of monotonicity is adaptively chosen by comparing the residual sum of squares for the two corresponding isotonic regression models.

While this method is somewhat heuristic, it is more reliable in practice than other, more sophisticated approaches. Nonetheless, it is likely that this approach could be further refined with some additional effort.

5.3.4 Theoretical Study of the Power for the Local Testing Problem

This section provides additional theoretical support for the proposed approach by comparing it to an idealized *Neyman-Pearson oracle* procedure that achieves maximal power by leveraging (unrealistic) knowledge of the true data distribution. For simplicity, consider a fixed local hypothesis of the form $H_S := \bigcap_{j \in S} H_j$, which asserts that a subset $S \subseteq [n]$ contains no outliers, and without loss of generality restrict the attention to the case $S = [n]$, which corresponds to the global testing problem.

It is important to note that optimality of a local test does not necessarily imply optimality for the full closed testing procedure, since optimality in multiple testing typically requires additional simplifying assumptions (Spjotvoll, 1972; Heller and Rosset, 2020; Rosset et al., 2022; Heller et al., 2023; Dobriban et al., 2015; Sun and Cai, 2007; Storey, 2007; Westfall et al., 1998). Nonetheless, studying the optimal local test is an intuitive and valuable way to justify the proposed approach theoretically. Extending the theoretical study presented in this section to the closed testing setting is an interesting direction for future work.

Recall from Section 5.1.2 that the proposed method tests the global hypothesis $H_{[n]} := \bigcap_{j=1}^{n} H_j$ via two main steps. The first step, described in Section 5.1.2, computes scalar-valued conformity scores $X_1, \ldots, X_m, Y_1, \ldots, Y_n \in \mathbb{R}$ for the calibration and test points by applying a data-driven scoring function $\hat{s} : \mathbb{R}^d \to \mathbb{R}$ to the observations $Z_1, \ldots, Z_{m+n} \in \mathbb{R}^d$.

These scores provide a low-dimensional summary of the data used in the second step to test $H_{[n]}$ using local standard p-value aggregation methods or other two-sample testing procedures, including those discussed in Sections 4.3 and 5.3.

Theorem 13 shows that the oracle Neyman-Pearson test also follows this two-step structure, applying a likelihood ratio test to scalar-valued sufficient statistics that may be interpreted as conformity scores obtained via an ideal scoring function s^*. Moreover, this oracle scoring function is proved to be equivalent to the optimal scoring function studied by Marandon et al. (2024) in the context of individual outlier identification under FDR control. This connection supports the use of binary classification models to learn a powerful scoring function $\hat{s}$, as also suggested by Marandon et al. (2024).

5.3.4.1 A Two-Step Oracle Procedure

In this section, assume that the n test data points $Z_{m+1}, \ldots, Z_{m+n} \in \mathbb{R}^d$ are i.i.d. samples from a mixture distribution:

$$Z_{m+1}, \ldots, Z_{m+n} \overset{\text{i.i.d.}}{\sim} (1-\pi)P_0 + \pi \bar{P}_1, \tag{5.8}$$

where $\pi \in [0,1]$ is the expected proportion of outliers, and P_0 is the inlier distribution from which the calibration points $Z_1, \ldots, Z_m$ are drawn i.i.d., as in equation (4.1).

This mixture model can be reconciled with the setup in Section 5.1.2 by expressing the outlier component $\bar{P}_1$ as a weighted average of the individual outlier distributions P_j in equation (4.1): $\bar{P}_1 := \sum_{j\in[n]} a_j P_j$, where the weights $a_j \geq 0$ satisfy $\sum_{j\in[n]} a_j = 1$. With this choice, the model in (5.8) closely resembles that in (4.1), although the two are not exactly equivalent. The model in (5.8) simplifies the theoretical analysis in this section but is not required for type-I error control of the proposed method.

In addition, assume that P_0 and each P_j are continuous distributions with densities p_0 and p_j, respectively, for all $j \in [n]$. Consequently, the outlier component $\bar{P}_1$ is also a continuous distribution with density

$$\bar{p}_1 = \sum_{j\in[n]} a_j p_j. \tag{5.9}$$

Under this setup, consider the problem of testing whether all test points are inliers drawn from the distribution P_0 against the alternative that the proportion of outliers is strictly positive equal to $\bar{\pi} > 0$. This global null hypothesis is denoted as $\tilde{H} : \pi = 0$, with the corresponding alternative hypothesis $\tilde{K} : \pi = \bar{\pi} > 0$.

By the Neyman–Pearson lemma, the most powerful test for this problem is the likelihood ratio test applied to the multivariate test observations $Z_{m+1}, \ldots, Z_{m+n}$. Clearly, this test is not directly usable in practice, as the likelihood ratio statistic depends on knowing both the inlier and outlier distributions.

Recall that ACODE follows the two-step structure described previously. A key consideration in this framework is that making inference on outliers in the test score sample may not be equivalent to making inference on outliers in the original test sample. This distinction arises because the scoring function can mask the underlying signal. For example, when $\hat{s}(Z_{m+j}) = 1$ for any $j \in [n]$ the outlier signal present in the raw test sample becomes undetectable in the corresponding score sample. To formalize this, let $\theta \in [0, 1]$ be the expected proportion of outliers in the test score sample. The global testing problem in the two-step setup is then defined as: $H : \theta = 0$, against $K : \theta = \bar{\theta} > 0$.

The following result shows that the likelihood ratio test for $\tilde{H} : \pi = 0$ versus $\tilde{K} : \pi = \bar{\pi} > 0$ applied to the multivariate test observations is equivalent to a two-step oracle procedure that tests $H : \theta = 0$ versus $K : \theta = \bar{\theta} > 0$, using the likelihood ratio test applied to univariate sufficient statistics, which serves as "oracle non-conformity scores". This characterization is useful because it makes the comparison to ACODE more direct, and highlights a connection to the results in Marandon et al. (2024) on the optimal choice of non-conformity scores.

Theorem 13. *The optimal likelihood ratio test for $\tilde{H} : \pi = 0$ versus $\tilde{K} : \pi = \bar{\pi} > 0$ applied to the test observations $Z_{m+1}, \ldots, Z_{m+n}$ distributed as in equation (5.8) is equivalent to the optimal likelihood ratio test for $H : \theta = 0$ versus $K : \theta = \bar{\theta} > 0$, where θ is the proportion of outliers in the score vector $Y_1^*, \ldots, Y_n^*$ with $Y_j^* := r(Z_{m+j})$, $j \in [n]$ and scoring function*

$$r(z) := \frac{\bar{p}_1(z)}{p_0(z)}, \tag{5.10}$$

with $\bar{p}_1(z)$ *defined in* (5.9)*. The test on* θ *rejects the null hypothesis for large values of*

$$\rho^{\text{scores}}(Y_1^*, \ldots, Y_n^*) = \frac{\prod_{j\in[n]} \left[(1-\bar{\theta}) f_0(Y_j^*) + \bar{\theta}\bar{f}_1(Y_j^*)\right]}{\prod_{j\in[n]} f_0(Y_j^*)}, \tag{5.11}$$

where f_0 *is the density of* Y_j^* *when* Z_{m+j} *is drawn from the inlier distribution* P_0*, and* $\bar{f}_1 := \sum_{j\in[n]} b_j f_j$ *with* $b_j > 0$*,* $\sum_{j\in[n]} b_j = 1$*, and* f_j *is the density of* Y_j^* *when* Z_{m+j} *is drawn from the outlier distribution* P_j*.*

Following Theorem 13, define the optimal scoring function as $s^*(z) = r(z)$ from equation (5.10) and the optimal testing procedure as ρ^{scores} from equation (5.11). Both the optimal scoring function and testing procedure are idealized oracle tools that cannot be directly applied in practice, since they require knowledge of both the inlier and outlier distributions.

The scoring function in (5.10) is essentially the approach proposed by Marandon et al. (2024), adapted to the context of this thesis where the focus is on testing the global problem rather than identifying individual outliers. The key difference lies in the testing objectives: Marandon et al. (2024) test whether each individual observation comes from a two-component mixture of inlier and outlier populations versus being a pure outlier. In contrast, the global null hypothesis of no outliers in the test set is tested against the alternative that at least one observation is an outlier.

This connection with the work of Marandon et al. (2024) supports the use of binary classification models to learn a powerful scoring function $\hat{s}$ in practice, as suggested by the theoretical results proving that fitting binary classifiers approximates well the optimal scoring function. However, since one-class classifiers can outperform binary classifiers, particularly when the signal is weak and sparse, they suggest integrating one-class classifiers into the AdaDetect toolbox. ACODE implements this suggestion by integrating AdaDetect with one-class classifiers. By leveraging both approaches, ACODE retains the theoretical properties established in Marandon et al. (2024) while enhancing robustness through adaptive, data-driven classifier selection. This approach enables ACODE to effectively approximate the optimal score function, as demonstrated in the numerical experiments presented in Chapter 6.

5.3.4.2 Rank-Based Local Optimality versus Neyman-Pearson Optimality

Here, the discussion on local optimality of the Shiraishi test is extended by comparing its local power with that of the Neyman-Pearson test defined in equation 5.11.

The model in Theorem 1 is retrieved for the optimal scores:

$$X_1^*, \ldots, X_m^* \overset{i.i.d.}{\sim} F, \qquad Y_1^*, \ldots, Y_n^* \overset{i.i.d.}{\sim} (1-\theta)F + \theta G(F), \tag{5.12}$$

where G has bounded density g. Note that model (5.12) is slightly more restrictive than model in Theorem 13, as it excludes cases where the outlier density is unbounded. Nevertheless, it still encompasses a broad class of models. With the aim of comparing the power of the Shiraishi test and the optimal Neyman-Pearson test in (5.11) in a neighborhood of the null hypothesis of no outliers, consider the sequence of testing problems with local simple alternative:

$$H : \theta = 0 \qquad \text{vs} \qquad K_N : \theta = \bar{\theta}_N > 0, \tag{5.13}$$

where $\bar{\theta}_N = \bar{\theta}/\sqrt{N}$, $\bar{\theta} > 0$ and $N = m + n$. In this setup, from Shiraishi (1985) the power function of the Neyman-Pearson test and of the Shiraishi test at significant level $\alpha \in (0, 1)$ are, respectively,

$$\text{Pow}(\rho^{\text{scores}}; \alpha, \delta, \bar{\theta}, g) = 1 - \Phi(z_\alpha - \bar{\theta}[\delta \text{Var}\{g(U)\}]^{1/2}) \tag{5.14}$$

and

$$\text{Pow}(T^g; \alpha, \delta, \bar{\theta}, g) = 1 - \Phi(z_\alpha - \bar{\theta}[\delta(1-\delta)\text{Var}\{g(U)\}]^{1/2}), \tag{5.15}$$

where Φ and z_α are, respectively, the c.d.f. and the $(1-\alpha)$-quantile of the standard normal distribution, U is a standard uniform random variable and $\delta = \lim_{N\to\infty} n/N$, with $N = n + m$ the total sample size. The power functions under the null hypothesis of no outliers can be retrieved by allowing $\bar{\theta} = 0$.

The top panel in Figure 5.1 shows the power of the two tests as a function of δ for fixed α, θ and g. The Neyman-Pearson test dominates the Shiraishi test uniformly in δ, with the two tests coinciding only in a neighborhood of the null hypothesis. While the Neyman-Pearson power increases monotonically as the number of test observations increases, the Shiraishi

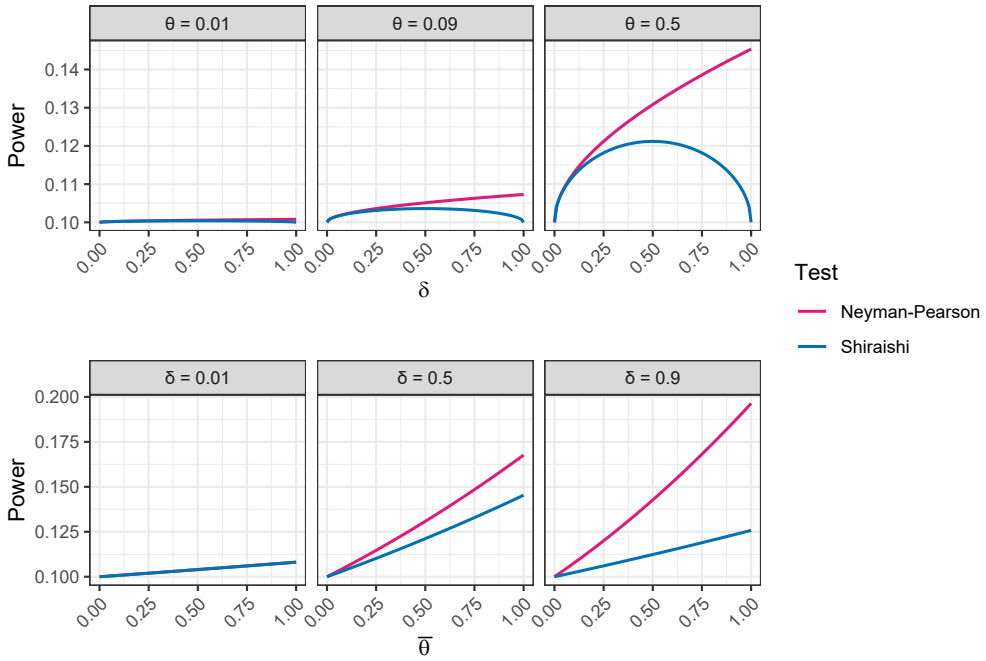

Figure 5.1: Power functions of the Neyman-Pearson and Shiraishi test. Top: power as a function of the sample size ratio δ for fixed α, $\bar{\theta}$ and g. Bottom: power as a function of $\bar{\theta}$ for fixed α, δ and g. In all panels, the significance level $\alpha = 0.1$ and the outlier density g is the Beta distribution with both parameters equal to 2.

power reaches its maximum at $\delta = 1/2$. This suggests that the Shiraishi test performs best when the test and calibration sample sizes are balanced. The oracle Neyman-Pearson test, by contrast, does not depend on the calibration sample; it relies only on the test set and its power increases with the test sample size.

The bottom panel in Figure 5.1 shows the power of the two tests as a function of $\bar{\theta}$ for fixed α, δ and g. As expected, both tests exhibit increasing power as $\bar{\theta}$ increases, with the Neyman-Pearson oracle test uniformly dominating the Shiraishi test. Equations (5.14) and (5.15) show that the two power functions coincide only near the null hypothesis for any value of δ. Additionally, they converge for any value of $\bar{\theta}$ when δ is small, as demonstrated in the left panel ($\delta = 0.01$).

The Neyman-Pearson test is expected to uniformly dominate the Shiraishi test in $\bar{\theta}$ since the latter is a rank test and exploits only the information in the ranking of test scores

rather than the score values themselves. This loss of information results in lower power and can be quantified by the (Pitman) Asymptotic Relative Efficiency between the two tests (ref. Section 2.6).

Chapter 6

Empirical Results

6.1 Experiments with Synthetic Data

ACODE is applied on synthetic data from a distribution inspired by Liang et al. (2024) and Bates et al. (2023). Each observation $Z_i \in \mathbb{R}^{1000}$ is sampled from a multivariate Gaussian mixture P_Z^a, such that $Z_i = \sqrt{a}\, V_i + W_i$, for some constant $a \geq 0$ and vectors $V_i, W_i \in \mathbb{R}^{1000}$. The inliers correspond to $a = 1$ and the outliers to $a = 0.7$. The elements of V_i are standard Gaussian, while each element of W_i is independent and uniformly distributed on a discrete set $\mathcal{W} \subseteq \mathbb{R}^{1000}$ with $|\mathcal{W}| = 1000$. The vectors in $\mathcal{W}$ are independently sampled from Uniform$([-3, 3]^{1000})$ prior to the first experiment. The numbers of inliers in the training, calibration, and tuning sets utilized by ACODE are 1000, 750, and 250, respectively.

ACODE is applied using 6 classification algorithms and 5 local testing procedures. The algorithms include 3 one-class classifiers (isolation forest, support vector machine, and "local outlier factor" nearest neighbors) and 3 binary classifiers (deep neural network, random forest, and AdaBoost), all implemented in the Python package *scikit-learn*. The one-class classifiers compute out-of-sample scores, e.g., as in Bates et al. (2023), while the binary classifiers compute in-sample scores with the approach of Marandon et al. (2024).

The testing procedures leveraged by ACODE are: Simes' test with and without Storey's correction, Fisher's combination method, the WMW test, and the Shiraishi test described in Section 5.3.2, applied using $G(F) = F^3$. Recall that this choice of G is optimal against Lehmann's alternative with $k = 3$, which may or may not fit the data well. An implementation of the Shiraishi test based on a data-driven estimate of G will be considered

later.

6.1.1 Global Outlier Enumeration

The analysis begins by constructing 90% lower confidence bounds for the total number of outliers in a test set of size 1000, varying the proportion of outliers as a control parameter. Figure 6.1 summarizes median lower bounds produced by ACODE over 100 independent experiments, separately for different classifiers and local testing procedures. The left and center panels compare the performance of ACODE applied using only the 3 one-class or binary classifiers, respectively, while the right panel corresponds to ACODE automatically selecting a classifier from the full suite of all 6 options.

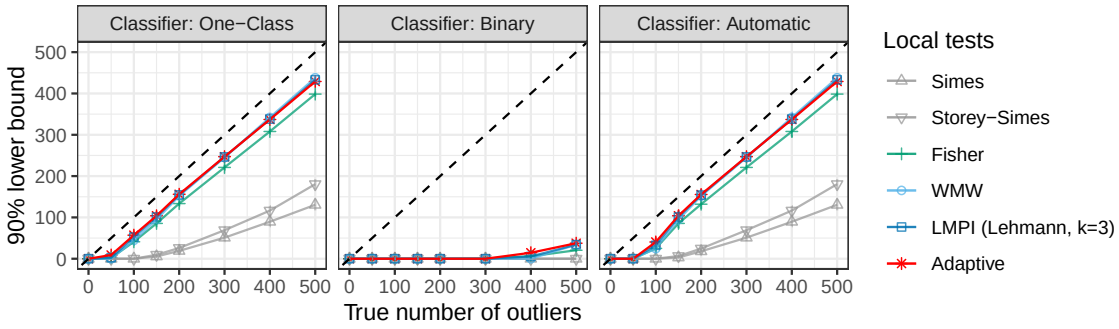

Figure 6.1: Median values for a 90% lower confidence bound on the number of outliers in a test set, computed by ACODE on synthetic data based on different classifiers and local testing procedures. The results are shown as a function of the true number of outliers within a test set of size 1000. The most adaptive version of ACODE can automatically select an effective classifier and local testing procedure in a data-driven way.

The findings highlight that one-class classifiers generally outperform binary classifiers on these data, consistently with the observations of Liang et al. (2024), although binary classifiers can be advantageous in other contexts, as shown in Section 6.3.1. Figure 6.1 also distinguishes between the distinct performances of the 5 local testing procedures (shown in different colors), and the scenario in which ACODE selects one procedure adaptively. For these data, the WMW test yields the most informative (highest) lower confidence bounds, and ACODE's performance closely approximates that of an *ideal oracle* that knows the optimal combination of classifier and testing procedure.

Figure ?? presents the relative frequencies of the selected testing procedures across different classifier groups, plotted against the number of outliers in the test set. The synthetic data were generated using the same setup as in Figure 6.1. Consistent with the observations from Figure 6.1, the WMW test is the most frequently selected method, while the Simes method is chosen least often—and is never selected when using the Storey estimator for the number of inliers. Overall, the relative frequencies of the selected methods appear similar across the three classifier groups, with no substantial differences observed.

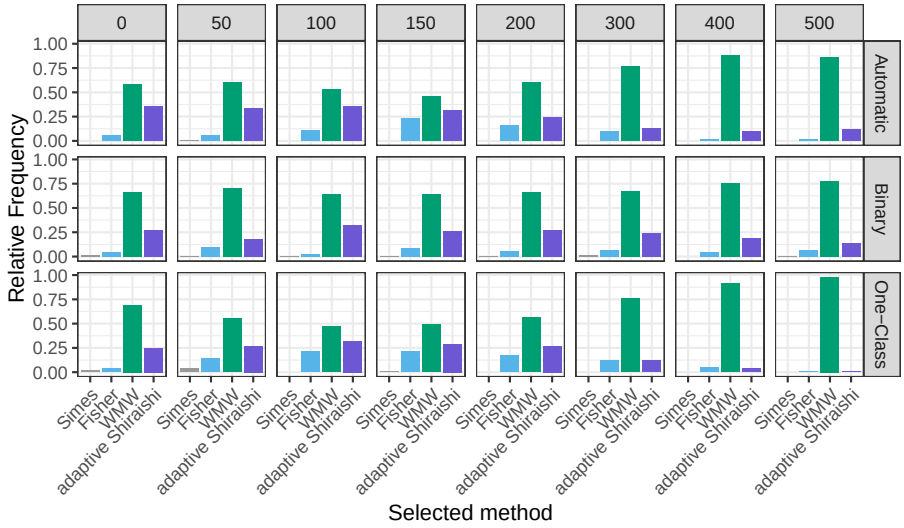

Figure 6.2: Relative frequencies of selected testing procedures for different classifier groups as a function of the number of outliers in a test set of 1000 samples. The "Automatic" group corresponds to classifiers chosen automatically from both binary and one-class classifiers, while for the "Binary" and "One-Class" groups the classifier is chosen automatically only from binary and one-class classifiers, respectively. The results are obtained by applying ACODE on synthetic data generated as in Figure 6.1.

6.1.2 Selective Outlier Enumeration

Figure 6.3 reports on related experiments in which the goal is to construct a 90% lower confidence bounds for the number of outliers within a data-driven subset of test points, selected as those with the largest conformity scores. To facilitate the interpretation of these experiments, ACODE is applied using a fixed classification algorithm (a one-class support vector machine), not a suite of 6 different algorithms. This ensures the test subsets are

always selected based on comparable conformity scores in all repetitions of the experiments. These results further demonstrate how the efficacy of different local testing procedures can vary in different situations. Simes' method performs better with very small selected sets, while the Shiraishi test with $G(F) = F^3$ excels in cases with moderately large selected sets. Once more, ACODE can approximately maximize power by autonomously identifying the most effective local testing procedure for each scenario.

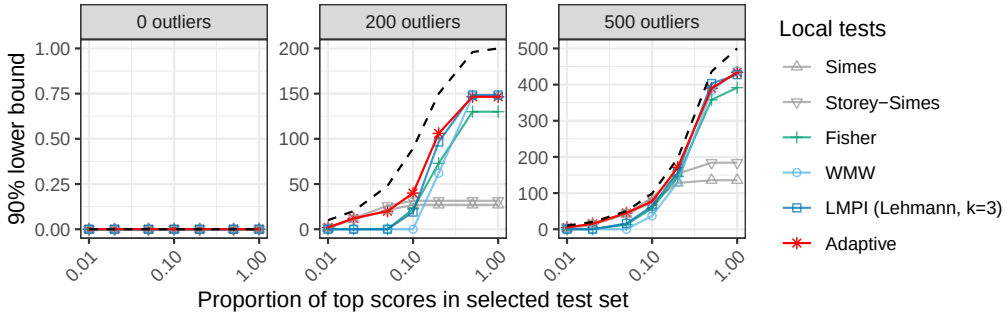

Figure 6.3: Median values for a 90% lower confidence bound on the number of outliers within an adaptively selected subset of 1000 test points, in experiments similar to those of Figure 6.1. The results are shown as a function of the proportion of selected test points and of the total number of outliers in the test set. The dashed curve corresponds to the true number of outliers in this selected set. In these experiments, ACODE is applied using a one-class support vector classifier to compute the conformity scores.

6.1.3 Hunting for Adversarial Anomalies

In this section, ACODE is applied leveraging an additional local testing procedure, utilizing a more flexible implementation of the LMPR approach outlined in Section 5.3.2. This approach enables ACODE to even detect elusive anomalies hidden by an adversary.

Inlier data are sampled from a standard multivariate normal distribution with 100 independent components. The outliers are generated by an "adversary" agent through the following process. The adversary first trains a one-class support vector machine on a separate dataset consisting of 1,000 inlier points. Then, to generate each outlier, it randomly produces three independent inliers and selects the one closest to the decision boundary of the support vector machine. This approach results in outliers that are difficult to detect on an individual basis but can be identified collectively, as their conformity scores tend to be

under-dispersed compared to those of the inliers.

Without prior knowledge of the distribution of outlier scores, ACODE is applied as in previous experiments, using the same toolbox of six machine learning classifiers, but including a sixth local testing procedure. This procedure is a practical approximation of the Shiraishi test described in Section 5.3.2, using an empirical estimate $\hat{G}$ of G. This estimate is obtained by fitting a mixture of Beta distributions, while preserving the exchangeability between calibration and test scores, as detailed in Section 5.3.3.

The results reported in Figure 6.4 demonstrate that ACODE effectively detects these elusive adversarial outliers, even though the first five local testing procedures considered become ineffective in this scenario. For additional experiments that offer a more detailed view of the performance of the empirical approximation of the Shiraishi local testing procedure in different settings, please refer to Figures 6.11–6.12 in Section 6.3.1.

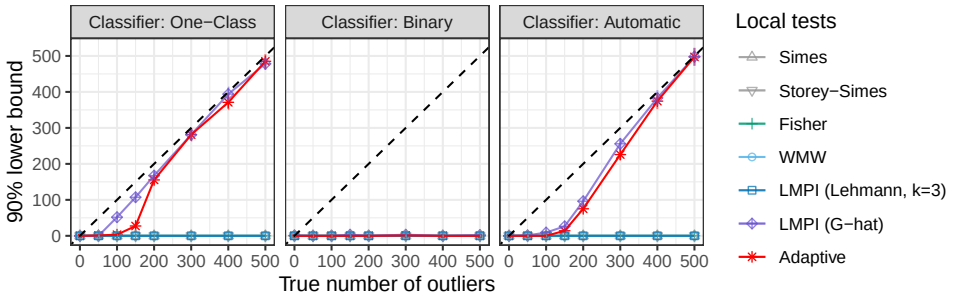

Figure 6.4: Empirical 90-th quantile for a 90% lower confidence bound on the number of outliers in a test set, computed by ACODE on synthetic data with adverserially hidden outliers exhibiting underdispersed conformity scores. Most local testing procedures cannot detect these outliers, but a data-driven approximation of the Shiraishi test enables ACODE to achieve high power. Other details are as in Figure 6.1.

6.2 Experiments with High-Energy Particle Collision Data

ACODE is now applied on data from the 2020 LHCO contest (Kasieczka et al., 2021), revisiting the example previewed in Figure 1.1. Participants in this contest were asked to estimate the number of interesting "signal" particle collision events within a given data set by contrasting it with a "reference" data exclusively comprised of "background" events generated by a physics simulation software. While the LHCO challenge is closely related to

the problem of collective outlier detection, the analysis in this section overlooks some of the physics intricacies and emphasizes instead the statistical aspects of the proposed method.

The analysis is restricted to a subset of the LHCO data containing 1,100,000 observations of 14 high-level features constructed based on subject domain knowledge. Each observation was labeled as either a background (inlier) or a *2-prong signal* (outlier) event, where the latter refers to hadron collision resulting in the emission of two charged particles. The goals are to detect and enumerate the outliers within a randomly chosen unlabeled test set, having access to a reference set of independent and identically distributed inliers.

In each experiment, a test set of 10,000 events is randomly drawn, among which the proportion of outliers is a control parameter varied between 0 and 0.15. In addition, three disjoint subsets of inliers are randomly sampled, to be used for training (cardinality 10,000), calibration (cardinality 1000), and tuning (cardinality 1000), respectively. ACODE is then applied as described in Section 6.1, leveraging the same suites of 6 classification algorithms and 5 local testing procedures. The additional local testing procedure described in Section 6.1.3 (the data-driven approximation of the Shiraishi oracle) is omitted here, for simplicity. This omission is due to its somewhat higher computational cost of its closed testing shortcut and the lack of clear advantages observed with these data.

Figure 1.1 compares the performance of different implementations of ACODE, in terms of power and the total number of detected outliers, as a function of the true number of outliers. See Figure 6.13 and Table 6.1 in Section 6.3.1 for a more comprehensive view of these results. The latter detail the performance of ACODE separately applied based on each of the 5 alternative local testing procedures considered here— to prevent overcrowding, only a subset of these local procedures were previously displayed in Figure 1.1.

Note that these results also include comparisons of ACODE's performance to that of a naive "cherry-picking" heuristic version of the proposed method, which does not provide valid inferences, as well as to that of the BH procedure applied to conformal p-values. The latter is an effective approach for individual outlier *identification* under FDR control (Bates et al., 2023; Marandon et al., 2024), but it is not designed to *estimate* the number of outliers.

Overall, ACODE yields more informative inferences compared to individual-level outlier identification, especially when leveraging the WMW test or Fisher's combination method. Further, ACODE achieves near-oracle performance with respect to the selections of classifier

and local testing procedure, without incurring in selection bias.

Figure 6.5 describes related results from experiments in which ACODE is applied to construct a 90% lower confidence bound for the number of outliers in a data-driven subset of test points, selected as those with the largest conformity scores. Similar to Figure 6.3, in these experiments the scores are computed by a single classifier, AdaBoost, to facilitates the interpretation of the findings. The results confirm the Shiraishi test for Lehmann's alternative with $k = 3$ tends to lead to more informative lower bounds compared to other local testing procedures, including the WMW test. Moreover, the adaptive version of ACODE again approximately maximizes power by selecting the testing procedure automatically. Additional results, with qualitatively similar conclusions, are presented in Section 6.3.2.

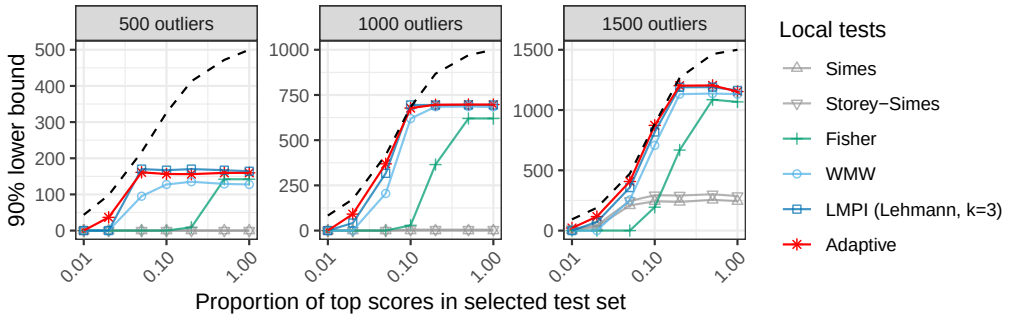

Figure 6.5: Median values over repeated experiments for a 90% lower confidence bound on the number of outliers within an adaptively selected subset of 1000 test points, in numerical experiments otherwise similar to those of Figure 6.1. The results are shown as a function of the proportion of selected test points and of the total number of outliers in the test set. The dashed curve corresponds to the true number of outliers in this selected set.

6.3 Additional Empirical Results

6.3.1 Numerical Experiments with Synthetic Data

Figures 6.6 and 6.7 describe findings similar to those in Figures 6.1 and 6.3, respectively. The distinction is that now the data are simulated from a binomial model borrowed from Liang et al. (2024), for which binary classifiers are more powerful than one-class classifiers. Figures 6.8 and 6.9 summarize the 90-th empirical quantiles of the lower confidence bounds for the numbers of outliers presented in Figures 6.1 and 6.6, respectively. Figure 6.10 delves into experiments related to those depicted in Figures 6.6 and 6.7, but utilizing synthetic data from a mixture model that bridges between the distributions considered above. These results highlight ACODE's flexibility, which selects an effective classifier and local testing procedure for each specific case.

Figures 6.11 and 6.12 offer additional insights into the performance of the Shiraishi local testing procedure, outlined in Section 5.3.2, when applied in different settings. These results highlight its practical advantages in situations where the outlier scores are either underdispersed or overdispersed relative to the inlier scores.

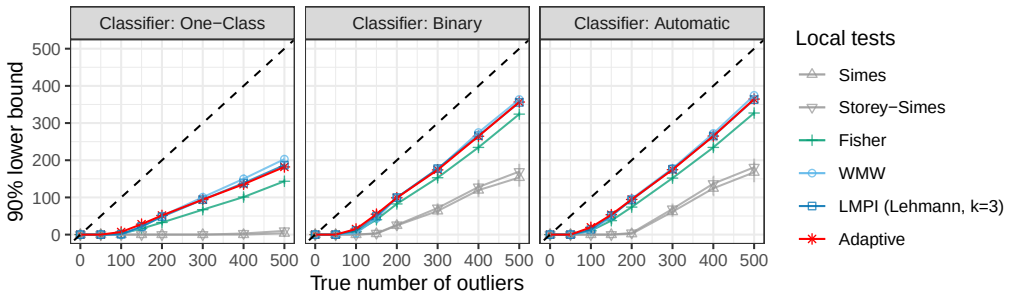

Figure 6.6: Median values for a 90% lower confidence bound on the total number of outliers in a test set, computed by ACODE on synthetic data based on different classifiers and local testing procedures. The results are shown as a function of the true number of outliers within a test set of size 1000. In these experiments, the synthetic data are generated from a binomial model borrowed from Liang et al. (2024).

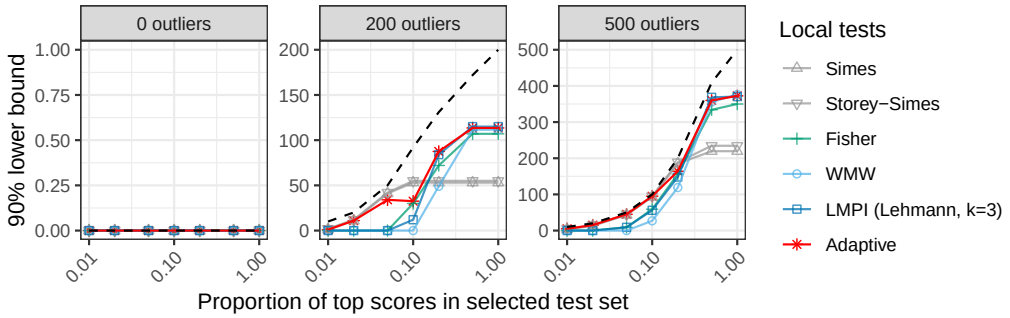

Figure 6.7: Lower confidence bounds for the number of outliers within an adaptively selected subset of 1000 test points, in numerical experiments otherwise similar to those of Figure 6.1. The results are shown as a function of the proportion of selected test points and of the total number of outliers in the test set. In these experiments, ACODE is applied using a one-class support vector classifier to compute the conformity scores.

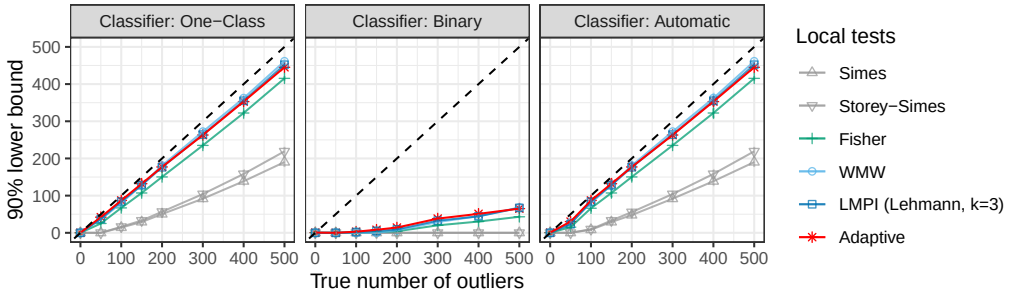

Figure 6.8: Empirical 90-th quantile for a 90% lower confidence bounds on the total number of outliers in a test set, computed by ACODE on synthetic data based on different classifiers and local testing procedures. Other details are as in Figure 6.1.

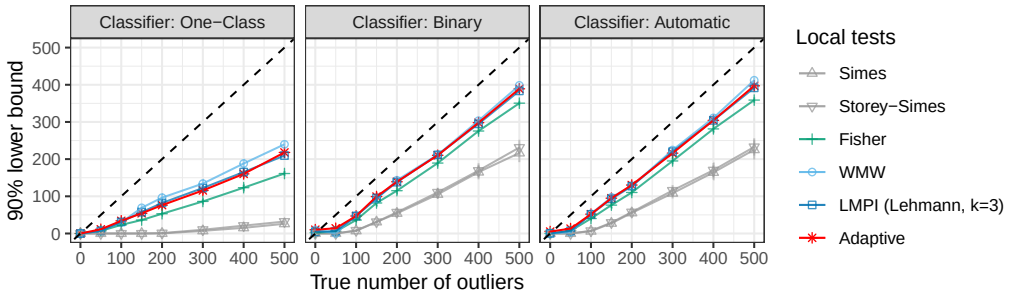

Figure 6.9: Empirical 90-th quantile for a 90% lower confidence bounds on the total number of outliers in a test set, computed by ACODE on synthetic data based on different classifiers and local testing procedures. Other details are as in Figure 6.6.

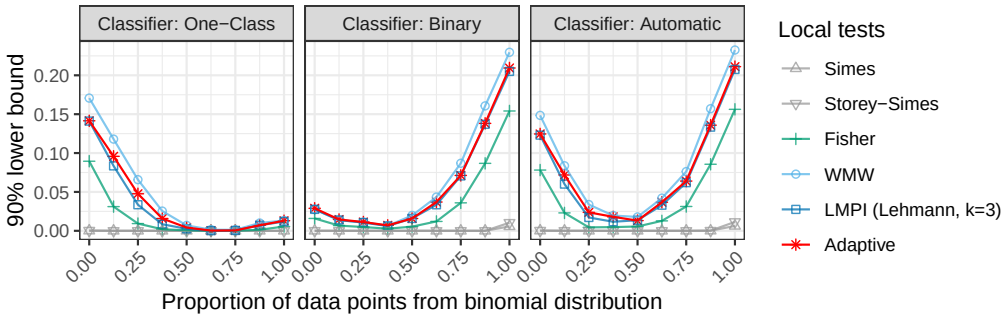

Figure 6.10: Lower confidence bounds for the total number of outliers in a test set, computed by ACODE on synthetic data based on different classifiers and local testing procedures. The results are shown as a function of the true number of outliers within a test set of size 1000. In these experiments, the synthetic data are generated from a binomial model borrowed from Liang et al. (2024). Other details are as in Figure 6.6.

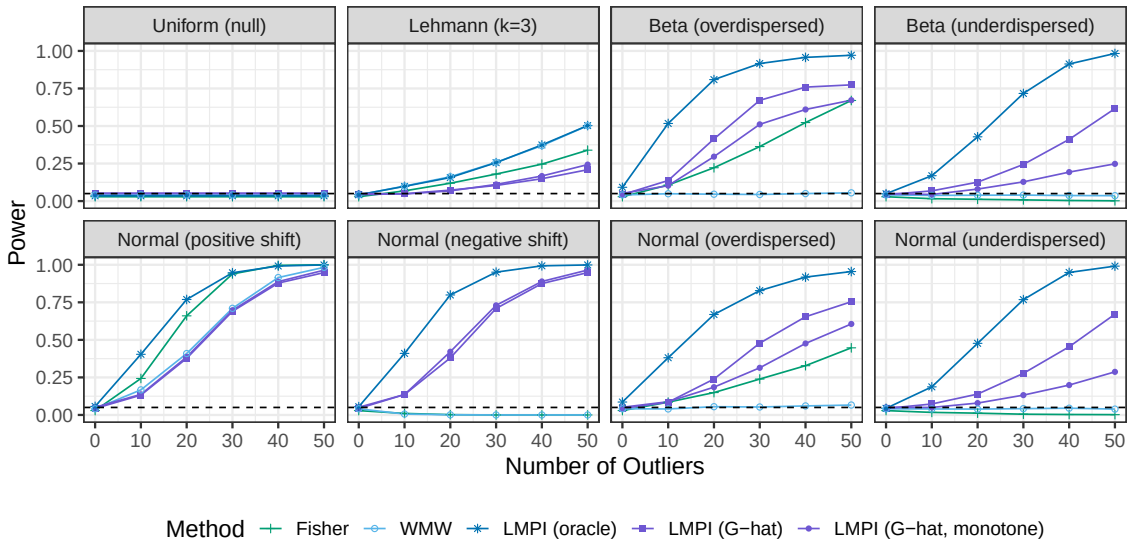

Figure 6.11: Power of different rank tests for the global null hypothesis of no outliers in a test set containing 200 independent observations of a simulated univariate score, using a calibration set containing 500 independent inlier scores. The nominal level is $\alpha = 0.05$ (horizontal dashed line). The results are shown as a function of the true number of outliers in the test set. Top: the inliers are randomly sampled from a uniform distribution on $[0, 1]$, while the outliers are sampled from the alternative distribution indicated in each panel. Bottom: the inliers are standard normal, while the outlier scores follow a non-standard normal distribution, as indicated in each panel. Three versions of the Shiraishi test from Section 5.3.2 are compared. The first version uses oracle knowledge of the true transformation G linking the outlier distribution to the inlier distribution. The second version of the Shiraishi test uses an empirical estimate of G obtained as described in Section 5.3.3. The third version relies on a monotone approximation of the derivative $g = G'$, which (in Figure 6.12) enables a computationally convenient shortcut in the context of closed testing, as also explained in Section 5.3.3.

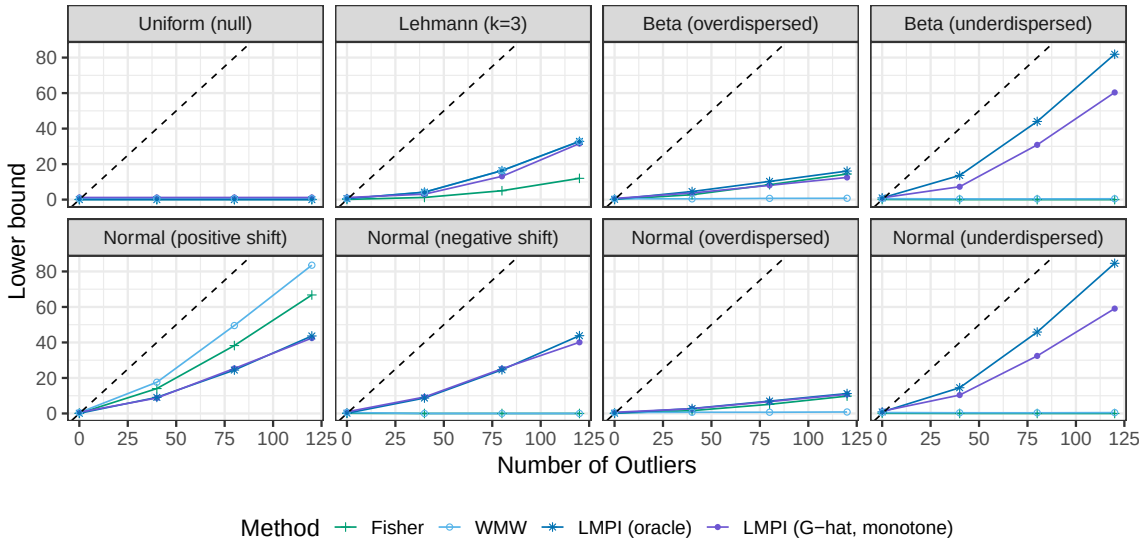

Figure 6.12: Median values for a 90% lower confidence bound on the total number of outliers in the test set, in the same experiments of Figure 6.11. The lower bound is calculated through closed testing, using different local testing procedures. For the Shiraishi approach, if the derivative g of the oracle function G is not monotone, it is in practice replaced by a monotone approximation that enables the application of a computationally efficient closed testing shortcut.

6.3.2 Numerical Experiments with Particle Collision Data

Figure 6.13 is a more detailed version of Figure 1.1, including a broader range for the choice of the local test. Figures 6.14 and 6.15 report on experiments similar to those of Figures 1.1 and 6.5, respectively, but leveraging a larger training set containing 100,000 inliers, which leads to higher power for all methods. Tables 6.1 and 6.2 provide a more detailed view of the results, respectively, in Figures 6.13 and 6.14. Finally, Figures 6.16 and 6.17 summarize the 90-th empirical quantiles of the lower confidence bounds for the numbers of outliers presented in Figures 6.5 and 6.15, respectively, confirming again the validity of ACODE.

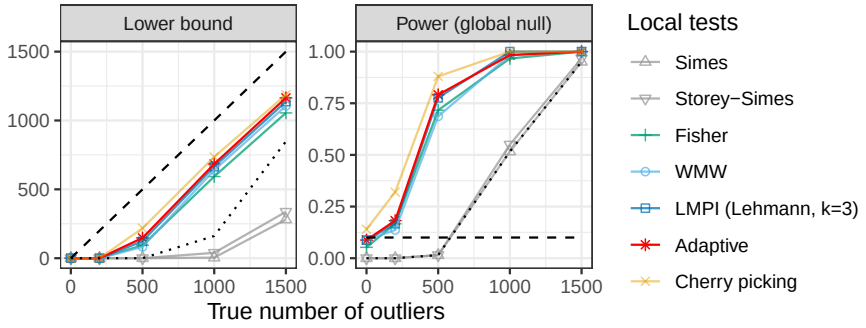

Figure 6.13: Performance of ACODE for collective outlier detection with the LHCO data, in the experiments of Figure 1.1. ACODE utilizes a local testing procedure that may be adaptively selected (red curve) or fixed (other solid curves). Compared to Figure 1.1, these results include a more detailed comparison of the performance of ACODE applied using different choices of local testing procedures.

	Local testing procedure						
Outliers	Simes	Storey-Simes	Fisher	WMW	LMPI (Lehmann, k=3)	Adaptive	Cherry picking
90% Lower bound (median)							
0	0 (0)	0 (0)	0 (2)	0 (5)	0 (4)	0 (5)	0 (4)
200	0 (0)	0 (0)	0 (1)	0 (5)	0 (4)	0 (5)	0 (5)
500	0 (0)	0 (0)	94 (11)	82 (17)	123 (16)	147 (17)	224 (14)
1000	4 (12)	39 (13)	593 (24)	640 (29)	665 (26)	684 (28)	736 (14)
1500	281 (20)	337 (19)	1055 (23)	1106 (27)	1137 (25)	1164 (24)	1190 (14)
90% Lower bound (90-th quantile)							
0	0 (0)	0 (0)	0 (2)	0 (5)	0 (4)	0 (5)	27 (4)
200	0 (0)	0 (0)	14 (1)	28 (5)	42 (4)	82 (5)	104 (5)
500	0 (0)	0 (0)	227 (11)	310 (17)	301 (16)	341 (17)	404 (14)
1000	188 (12)	198 (13)	700 (24)	871 (29)	840 (26)	864 (28)	910 (14)
1500	458 (20)	501 (19)	1188 (23)	1341 (27)	1347 (25)	1358 (24)	1376 (14)
Power (global null)							
0	0.00 (0.00)	0.00 (0.00)	0.05 (0.03)	0.09 (0.04)	0.09 (0.04)	0.09 (0.04)	0.15 (0.04)
200	0.00 (0.00)	0.00 (0.00)	0.17 (0.05)	0.14 (0.04)	0.17 (0.05)	0.18 (0.05)	0.33 (0.05)
500	0.01 (0.01)	0.01 (0.01)	0.72 (0.06)	0.69 (0.06)	0.78 (0.05)	0.79 (0.05)	0.88 (0.03)
1000	0.52 (0.07)	0.55 (0.06)	0.97 (0.02)	0.98 (0.02)	1.00 (0.00)	0.98 (0.02)	1.00 (0.00)
1500	0.95 (0.03)	0.97 (0.02)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)

Table 6.1: Performance of ACODE for collective outlier detection with the LHCO data, as a function of the true number of outliers. The performance is measured in terms of median and 90-th percentile values of a 90% lower confidence bound for the number of outliers (top and center) and the power to reject the global null hypothesis of no outliers (bottom). The numbers in parenthesis are standard errors. Other details are as in Figures 1.1 and 6.13.

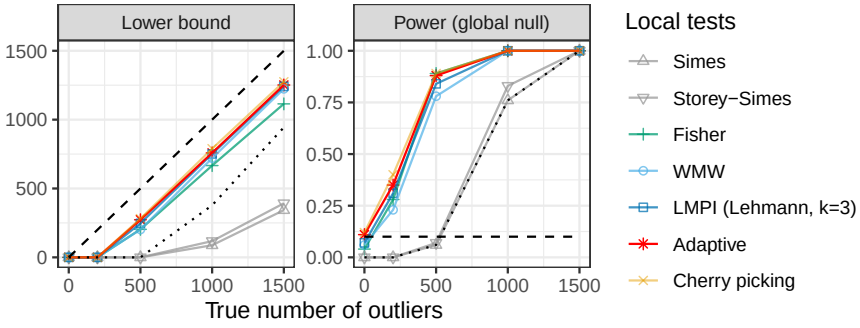

Figure 6.14: Performance of ACODE for collective outlier detection with the LHCO data, using a training set containing 100,000 inliers. In the interest of computational efficiency, in these experiments, the conformity scores leveraged by ACODE are computed by a fixed AdaBoost classifier. Other details are as in Figure 1.1.

	Local testing procedure						
Outliers	Simes	Storey-Simes	Fisher	WMW	LMPI (Lehmann, k=3)	Adaptive	Cherry picking
90% Lower bound (median)							
0	0 (0)	0 (0)	0 (1)	0 (3)	0 (2)	0 (2)	0 (3)
200	0 (0)	0 (0)	0 (4)	0 (7)	0 (7)	0 (7)	0 (8)
500	0 (0)	0 (0)	204 (11)	202 (16)	247 (15)	273 (15)	284 (15)
1000	88 (10)	118 (10)	666 (10)	720 (18)	751 (14)	758 (13)	796 (13)
1500	343 (16)	392 (15)	1114 (10)	1224 (20)	1242 (16)	1252 (15)	1276 (15)
90% Lower bound (90-th quantile)							
0	0 (0)	0 (0)	0 (1)	0 (3)	0 (2)	2 (2)	8 (3)
200	0 (0)	0 (0)	46 (4)	107 (7)	131 (7)	133 (7)	156 (8)
500	0 (0)	0 (0)	328 (11)	421 (16)	436 (15)	456 (15)	474 (15)
1000	238 (10)	251 (10)	782 (10)	952 (18)	919 (14)	932 (13)	957 (13)
1500	541 (16)	563 (15)	1255 (10)	1467 (20)	1430 (16)	1452 (15)	1478 (15)
Power (global null)							
0	0.00 (0.00)	0.00 (0.00)	0.04 (0.02)	0.06 (0.02)	0.07 (0.03)	0.11 (0.03)	0.14 (0.03)
200	0.00 (0.00)	0.00 (0.00)	0.28 (0.05)	0.23 (0.04)	0.31 (0.05)	0.35 (0.05)	0.40 (0.05)
500	0.06 (0.02)	0.07 (0.03)	0.89 (0.03)	0.78 (0.04)	0.84 (0.04)	0.88 (0.03)	0.89 (0.03)
1000	0.76 (0.04)	0.83 (0.04)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
1500	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)

Table 6.2: Performance of ACODE for collective outlier detection with the LHCO data, using a training set containing 100,000 inliers. The performance is measured in terms of median and 90-th percentile values of a 90% lower confidence bound for the number of outliers (top and center) and the power to reject the global null hypothesis (bottom). The numbers in parenthesis are standard errors. Other details are as in Figure 6.14.

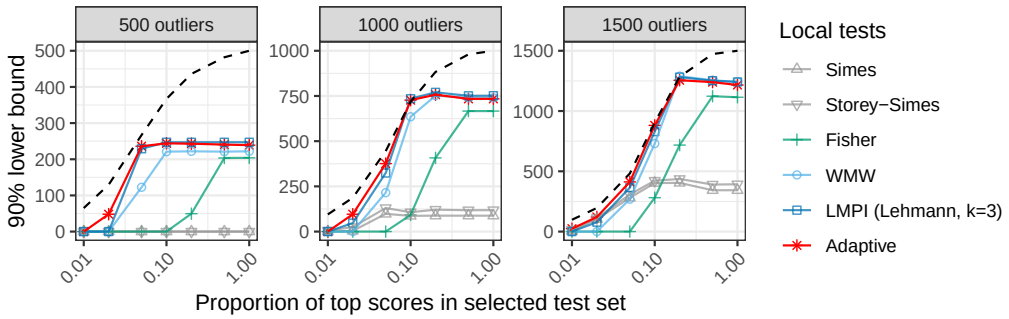

Figure 6.15: Lower confidence bounds for the number of outliers within an adaptively selected subset of 1000 test points, in numerical experiments otherwise similar to those of Figure 6.1. The results are shown as a function of the proportion of selected test points and of the total number of outliers in the test set. The dashed curve corresponds to the true number of outliers in this selected set. In these experiments, ACODE is applied using a one-class support vector classifier to compute the conformity scores.

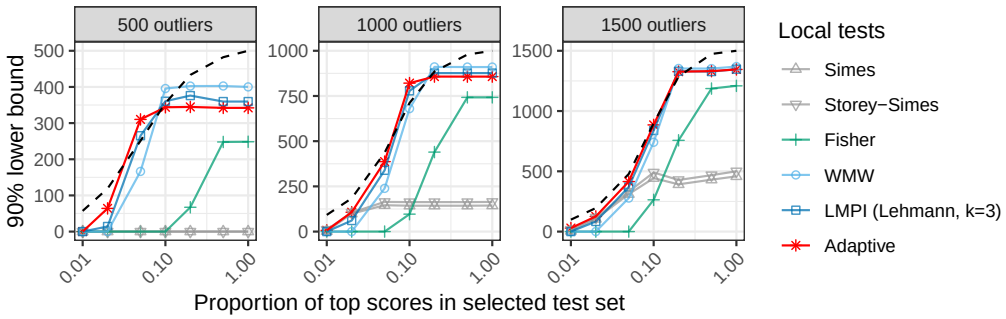

Figure 6.16: Empirical 90-th quantile for a 90% lower confidence bound on the number of outliers within an adaptively selected subset of 1000 test points, in the same numerical experiments of Figure 6.5.

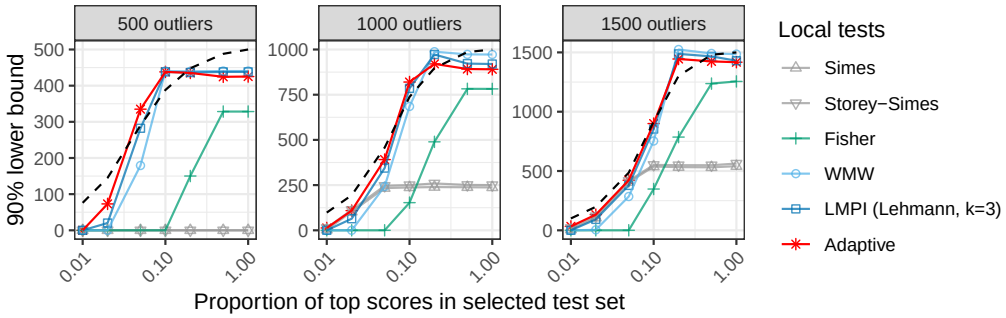

Figure 6.17: Empirical 90-th quantile for a 90% lower confidence bound on the number of outliers within an adaptively selected subset of 1000 test points, in the same numerical experiments of Figure 6.15.

6.3.3 Additional Experiments with Real Data

This section, further investigates the performance of ACODE by applying it to 6 additional data sets previously utilized in the related literature. These data sets are:

- `CreditCard` (Dataset, c), containing 30 variables, 492 outliers, 284315 inliers;
- `Pendigits` (Dataset, e), containing 16 variables, 156 outliers, 6714 inliers;
- `Covertype` (Dataset, b), containing 10 variables, 2747 outliers, 286048 inliers;
- `Shuttle` (Dataset, f), containing 9 variables, 3511 outliers, 45586 inliers;
- `Mammography` (Dataset, d), containing 6 variables, 260 outliers, 10923 inliers;
- `ALOI` (Dataset, a), containing 27 variables, 1508 outliers, 48026 inliers.

These experiments are conducted following an approach similar to that described in Section 6.1, applying ACODE based on the same suite of 6 classification algorithms and 6 local testing procedures. Each experiment is independently repeated 100 times, each time randomly sampling disjoint training, calibration, tuning, and test sets of sizes 1000, 100, 100, and 100, respectively. The training, calibration, and tuning sets include only inliers, while the proportion of outliers in the test set is varied as a control parameter between 0 and 1.

Figures 6.18–6.23 summarize, separately for each data set, the performances of the lower confidence bounds and global tests obtained with ACODE. The results are reported as a function of the true number of outliers in the test set and are also stratified based on the type of classifier utilized by ACODE and on the underlying local testing procedure. The findings overall indicate that binary classification models generally yield higher power in these experiments, which can be attributed to the relatively low dimensionality of these data sets. Moreover, among the local testing methods considered, both Fisher's method and the WMW method demonstrate similarly high power, typically outperforming Simes' method. Crucially, the fully automatic implementation of ACODE, which selects both the classification model and the local testing procedure in a data-adaptive way, leads to near-oracle performance in all cases.

Finally, Figures 6.24 and 6.25 describe experiments aimed at constructing 90% lower confidence bounds for the number of outliers within a data-dependent subset of test points, specifically those selected for their high conformity scores. For easier understanding of these experiments, ACODE is now applied with a fixed classification algorithm, the one-class isolation forest, rather than leveraging the full suite of 6 different algorithms. The results show that the WMW test consistently outperforms the Fisher combination method in terms of power, aligning with findings previously documented in Sections 6.1 and 6.2. Furthermore, the fully adaptive implementation of ACODE effectively maximizes power by automatically identifying the most powerful local testing procedure for each case, again achieving oracle-like performance.

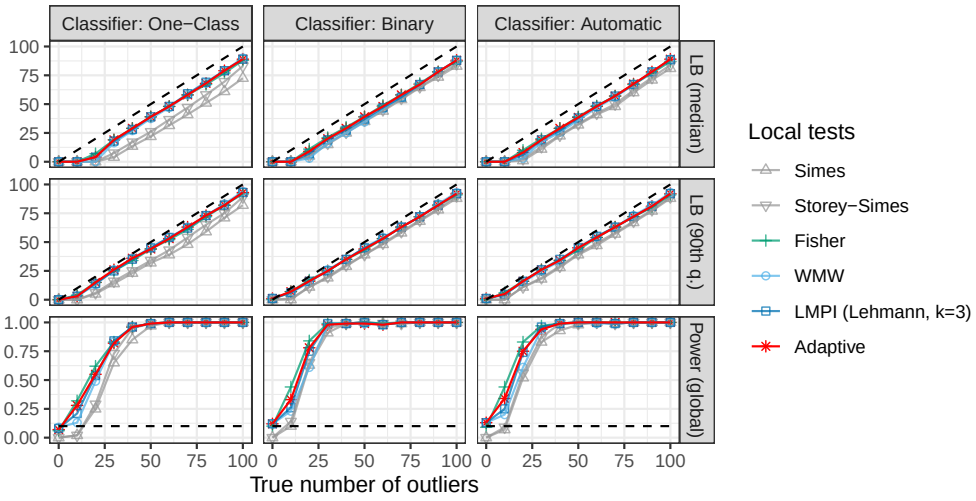

Figure 6.18: Performance of ACODE for collective outlier detection with the `creditcard` data set, as a function of the true number of outliers in a test set of size 100. The performance is measured in terms of median and 90-th percentile values of a 90% lower confidence bound for the number of outliers (top and center) and the power to reject the global null hypothesis of no outliers (bottom). Other details are as in Figure 6.1.

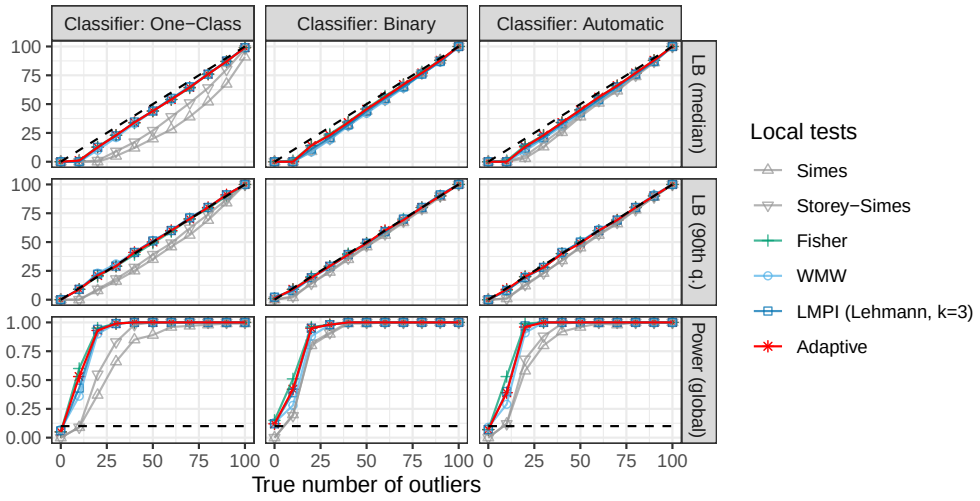

Figure 6.19: Performance of ACODE for collective outlier detection with the `pendigits` data set, as a function of the true number of outliers in a test set of size 100. The performance is measured in terms of median and 90-th percentile values of a 90% lower confidence bound for the number of outliers (top and center) and the power to reject the global null hypothesis of no outliers (bottom). Other details are as in Figure 6.18.

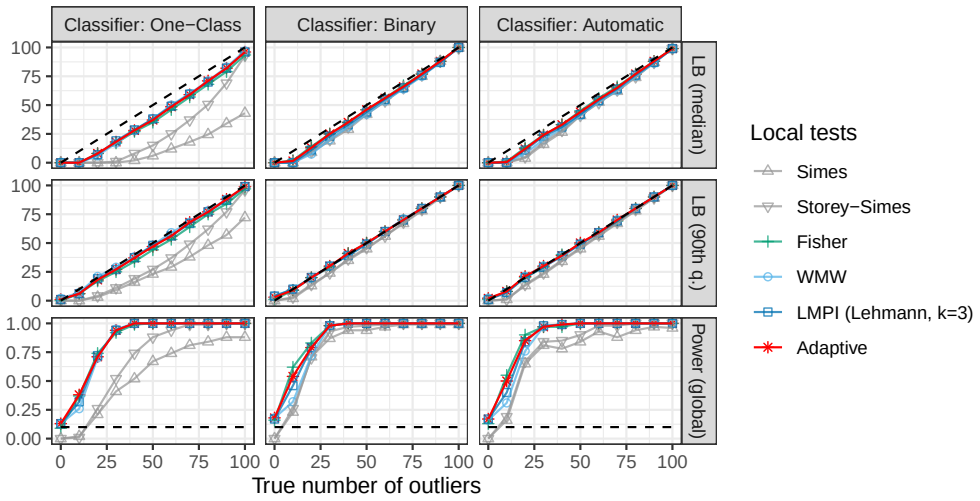

Figure 6.20: Performance of ACODE for collective outlier detection with the `covertype` data set, as a function of the true number of outliers in a test set of size 100. The performance is measured in terms of median and 90-th percentile values of a 90% lower confidence bound for the number of outliers (top and center) and the power to reject the global null hypothesis of no outliers (bottom). Other details are as in Figure 6.18.

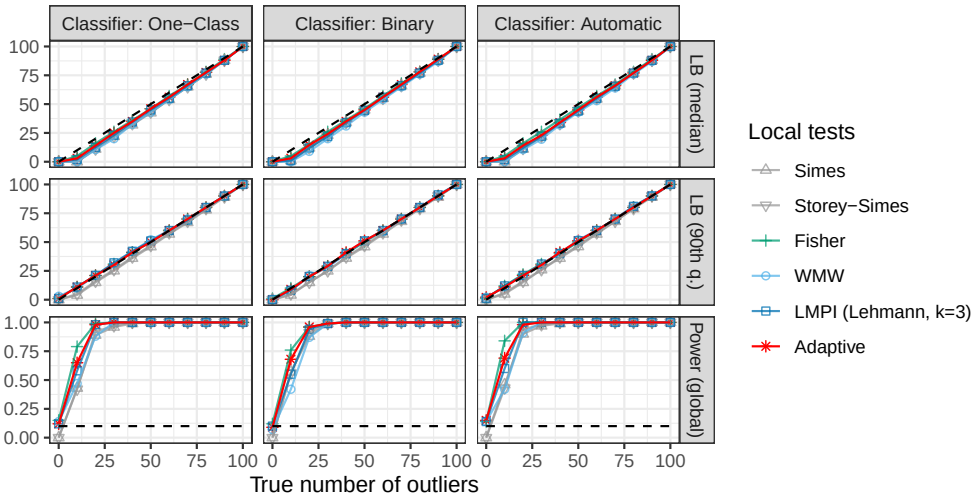

Figure 6.21: Performance of ACODE for collective outlier detection with the `shuttle` data set, as a function of the true number of outliers in a test set of size 100. The performance is measured in terms of median and 90-th percentile values of a 90% lower confidence bound for the number of outliers (top and center) and the power to reject the global null hypothesis of no outliers (bottom). Other details are as in Figure 6.18.

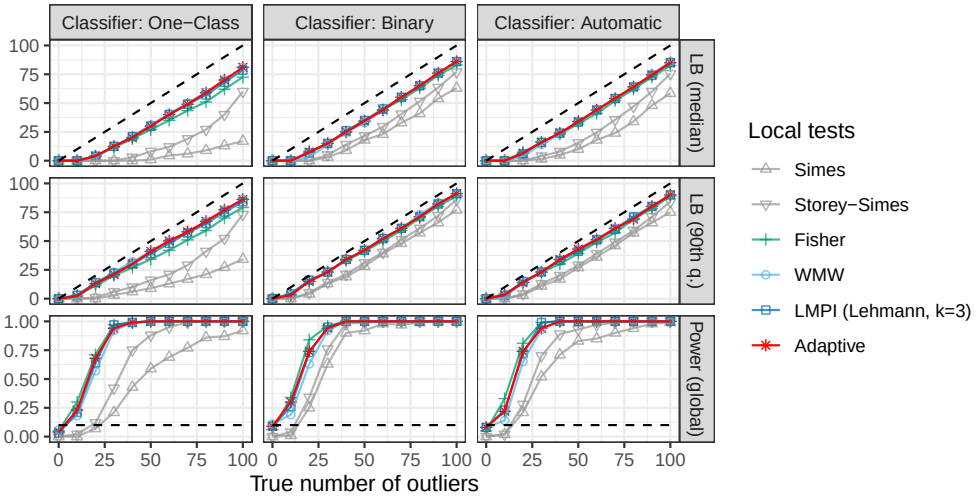

Figure 6.22: Performance of ACODE for collective outlier detection with the `mammography` data set, as a function of the true number of outliers in a test set of size 100. The performance is measured in terms of median and 90-th percentile values of a 90% lower confidence bound for the number of outliers (top and center) and the power to reject the global null hypothesis of no outliers (bottom). Other details are as in Figure 6.18.

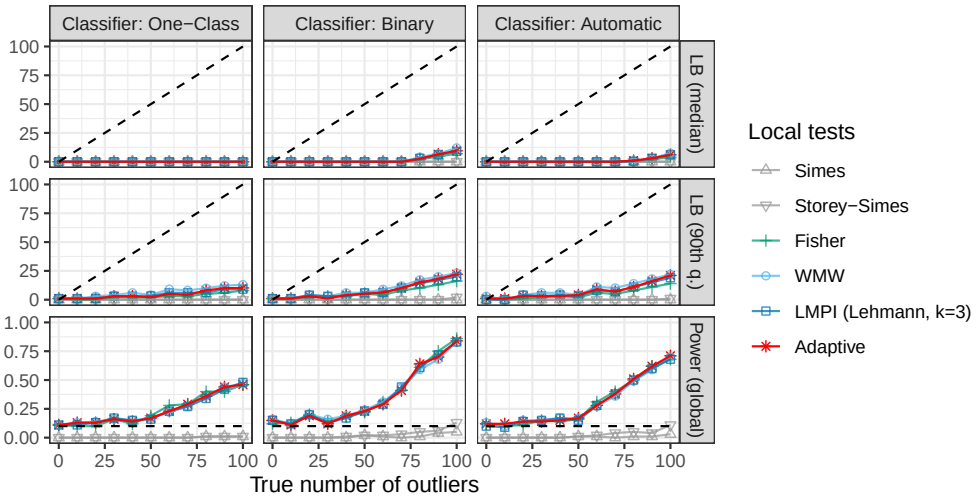

Figure 6.23: Performance of ACODE for collective outlier detection with the `aloi` data set, as a function of the true number of outliers in a test set of size 100. The performance is measured in terms of median and 90-th percentile values of a 90% lower confidence bound for the number of outliers (top and center) and the power to reject the global null hypothesis of no outliers (bottom). Other details are as in Figure 6.18.

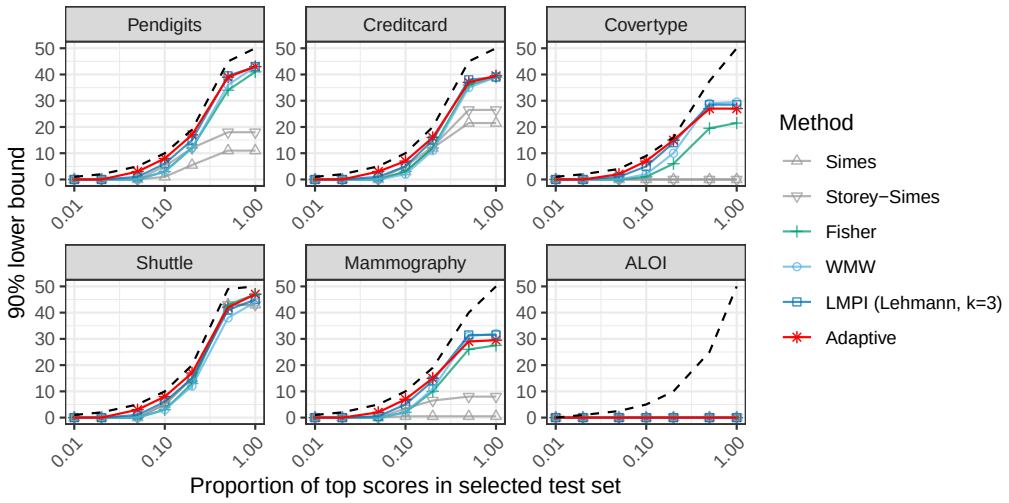

Figure 6.24: Median values for a lower confidence bound on the number of outliers within an adaptively selected subset of 1000 test points, in numerical experiments with several different data sets. The results are shown as a function of the proportion of selected test points. In these experiments, ACODE is applied using a one-class isolation forest model to compute the conformity scores. Other details are as in Figure 6.3.

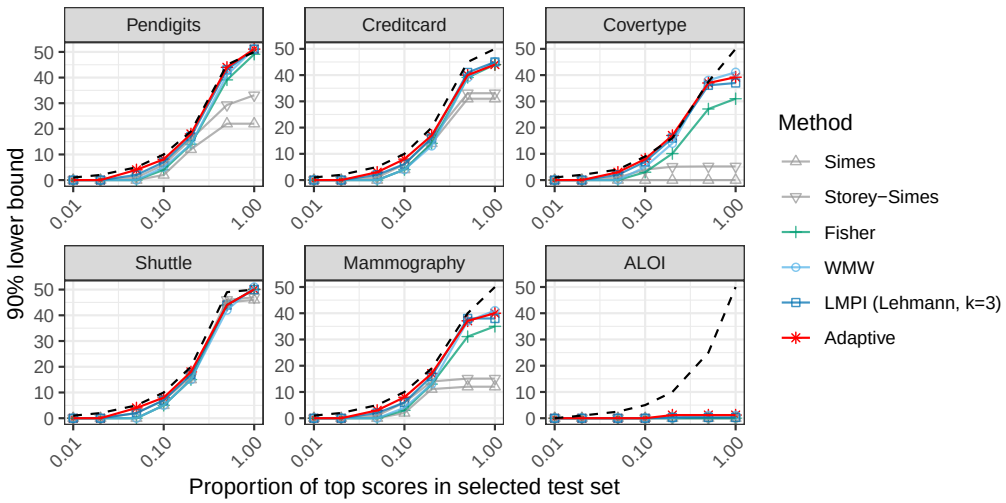

Figure 6.25: Empirical 90-th quantile for a lower confidence bound on the number of outliers within an adaptively selected subset of 1000 test points, in numerical experiments with several different data sets. Other details are as in Figure 6.24.

Chapter 7

Conclusions

7.1 Discussion

This thesis presents a novel distribution-free methodology for collective outlier detection that extends traditional approaches, by providing post-hoc confidence bounds for the total number of outliers in a dataset. Unlike conventional methods that focus primarily on controlling the false discovery rate, this new method offers explicit guarantees about the false discovery proportion.

The proposed approach involves a trade-off. On one hand, it exhibits somewhat reduced power for the precise localization of individual outliers compared to standard FDR-controlling procedures. On the other hand, it provides significantly stronger protection against the inherent variability of the FDP around the FDR. Furthermore, the method enables researchers to make simultaneous, valid inferences about the number of outliers across multiple subsets of the test data.

The novel method builds upon conformal inference techniques, specifically leveraging conformity scores and conformal p-values to enable valid data-driven inference through a closed testing procedure with local rank tests.

A key insight underlying this approach is that the optimal choice of conformity score function and local testing procedure can vary dramatically across different data scenarios, demonstrated through extensive empirical investigations. To address this variability, a fully data-driven approach has been adopted for selecting both the classification method and the local test.

This adaptive strategy is made rigorous through the use of split conformal techniques, which enable the incorporation of test set observations into the selection process while maintaining the essential property of inlier score exchangeability. By preserving this exchangeability through sample splitting, the method ensures that the resulting inference remains statistically valid despite the data-driven nature of the procedural choices, thereby avoiding selection bias.

7.2 Future Directions of Research

Our work inspires several avenues for future research. For example, while our method employs sample splitting, one could explore extensions to more sophisticated conformal inference frameworks that use cross-validation, which tends to be more computationally expensive but also more powerful for small sample sizes (Barber et al., 2021).

Additionally, this paper assumes an ideal reference data set of inliers; future work could examine relaxations of this assumption, accounting for possible contamination of the calibration data (Barber et al., 2023), or other distribution shifts (Tibshirani et al., 2019) following a weighting approach similar to that of Hu and Lei (2024). Indeed, distribution shifts and contamination in calibration data for outlier detection might lead to a loss of power (Bashari et al., 2025). Recent literature explores *double-robustness* (Lei and Candès, 2021; Candès et al., 2023; Yang et al., 2024; Sesia and Svetnik, 2025) and *weighted methods* (Tibshirani et al., 2019; Hore and Barber, 2024; Bhattacharyya and Barber, 2025) as promising approaches.

Another intriguing question arises about the implications of employing an optimal local test within a closed testing procedure, as it may not directly lead to an optimal closed testing procedure. Nevertheless, it is worth noting that closed testing procedures hold the potential for optimality. As highlighted by Goeman et al. (2021), admissible methods are required to be closed testing procedures, and optimal methods must adhere to admissibility.

Chapter A

Appendix

A1 Proofs

A1.1 Theorem 13

Proof of Theorem 13. The proof of Theorem 13 begins by establishing that the global testing problem on π

$$\tilde{H} : \pi = 0 \qquad \text{vs} \qquad \tilde{K} : \pi = \bar{\pi} > 0$$

is equivalent to the testing problem on θ

$$H : \theta = 0 \qquad \text{vs} \qquad K : \theta = \bar{\theta} > 0.$$

The equivalence follows by the fact that the optimal score vector $(Y_1^*, \ldots, Y_n^*)$ is a sufficient statistic for π. The likelihood function of the multivariate data $L(Z_{m+1}, \ldots, Z_{m+n} \; ; \; \pi)$ can be factorized into the product of two functions: one that depends on the parameter π and the data only through the transformation $(Y_1^*, \ldots, Y_n^*)$, and another that is independent of

π:

$$\begin{aligned}
L(Z_{m+1},\ldots,Z_{m+n}\,;\,\pi) &= \prod_{j\in[n]}\left[(1-\pi)p_0(Z_{m+j})+\pi p_1(Z_{m+j})\right] = \\
&= \frac{\prod_{j\in[n]}\left[(1-\pi)p_0(Z_{m+j})+\pi p_1(Z_{m+j})\right]}{\prod_{j\in[n]}p_0(Z_{m+j})}\cdot\prod_{j\in[n]}p_0(Z_{m+j}) = \\
&= \prod_{j\in[n]}\left[(1-\pi)+\pi\frac{p_1(Z_{m+j})}{p_0(Z_{m+j})}\right]\cdot\prod_{j\in[n]}p_0(Z_{m+j}) = \\
&= \prod_{j\in[n]}\left[(1-\pi)+\pi r(Z_{m+j})\right]\cdot\prod_{j\in[n]}p_0(Z_{m+j}) = \\
&= \prod_{j\in[n]}\left[(1-\pi)+\pi Y_j^*\right]\cdot\prod_{j\in[n]}p_0(Z_{m+j}).
\end{aligned}$$

The sufficiency of $(Y_1^*,\ldots,Y_n^*)$ for π follows from the Factorization Theorem (Casella and Berger, 2002, Theorem 6.2.6).

By the sufficiency of $(Y_1^*,\ldots,Y_n^*)$ for π and Theorem 8.2.4 of Casella and Berger (2002), the likelihood ratio test for $\tilde{H}$ versus $\tilde{K}$ applied to $Z_{m+1},\ldots,Z_{m+n}$:

$$\rho(Z_{m+1},\ldots,Z_{m+n}) := \frac{\prod_{j\in[n]}\left[(1-\bar{\pi})p_0(Z_{m+j})+\bar{\pi}\bar{p}_1(Z_{m+j})\right]}{\prod_{j\in[n]}p_0(Z_{m+j})}. \tag{A1}$$

is equivalent to the likelihood ratio test ρ^{scores} applied to the optimal scores $(Y_1^*,\ldots,Y_n^*)$ since, when using the score function r, the hypothesis H holds if and only if $\tilde{H}$ holds.

Consequently, for any significance level $\alpha\in(0,1)$ the rejection region of the Neyman-Pearson optimal test ρ defined as

$$R_\alpha(\rho) = \{(Z_{m+1},\ldots,Z_{m+n}) : \rho(Z_{m+1},\ldots,Z_{m+n})\geq\rho_\alpha\}, \tag{A2}$$

with $\rho_\alpha>0$ such that $\mathbb{P}(\rho(Z_{m+1},\ldots,Z_{m+n})\geq\rho_\alpha\,;\,\tilde{H})\leq\alpha$ and the rejection region of the two-step oracle procedure

$$R_\alpha(r,\rho^{\text{scores}}) = \{(Z_{m+1},\ldots,Z_{m+n}) : \rho^{\text{scores}}(r(Z_{m+1}),\ldots,r(Z_{m+n}))\geq t_\alpha\} \tag{A3}$$

where $t_\alpha>0$ is such that $\mathbb{P}\left(\rho^{\text{scores}}(r(Z_{m+1}),\ldots,r(Z_{m+n}))\geq t_\alpha\,;\,H\right)\leq\alpha$ coincide.

□

Bibliography

Ahmed, M. and A. N. Mahmood (2014). Network traffic analysis based on collective anomaly detection. In *9th IEEE Conference on Industrial Electronics and Applications*, pp. 1141–1146. IEEE.

Andreella, A., J. Hemerik, L. Finos, W. Weeda, and J. Goeman (2023). Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Stat. Med. 42*(14), 2311–2340.

Angelopoulos, A. N., R. F. Barber, and S. Bates (2025). Theoretical foundations of conformal prediction.

Barber, R. F., E. Candès, A. Ramdas, and R. J. Tibshirani (2021). Predictive inference with the jackknife+. *Ann. Stat. 49*(1), 486–507.

Barber, R. F., E. J. Candès, A. Ramdas, and R. J. Tibshirani (2023). Conformal prediction beyond exchangeability. *Ann. Stat. 51*(2), 816–845.

Bashari, M., M. Sesia, and Y. Romano (2025). Robust conformal outlier detection under contaminated reference data.

Bates, S., E. Candès, L. Lei, Y. Romano, and M. Sesia (2023). Testing for outliers with conformal p-values. *Ann. Stat. 51*(1), 149–178.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B 57*(1), 289–300.

Benjamini, Y. and Y. Hochberg (2000). The adaptive control of the false discovery rate in multiple comparison problems. *J. Educ. Behav. Stat. 25*(1), 60–83.

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika 93*(3), 491–507.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat. 29*(4), 1165–1188.

Bhattacharyya, A. and R. F. Barber (2025). Group-weighted conformal prediction.

Birnbaum, A. (1954). Combining independent tests of significance. *J. Am. Stat. Assoc. 49*(267), 559–574.

Blain, A., B. Thirion, and P. Neuvial (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage 260*, 119492.

Blanchard, G., P. Neuvial, and E. Roquain (2020). Post hoc confidence bounds on false positives using reference families. *Ann. Stat. 48*(3), 1281–1303.

Blanchard, G. and E. Roquain (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res. 10*, 2837–2871.

Bogomolov, M. (2023). Testing partial conjunction hypotheses under dependency, with applications to meta-analysis. *Electron. J. Stat. 17*(1), 102–155.

Bretz, F., T. Hothorn, and P. Westfall (2016). *Multiple comparisons using R*. CRC press.

Brus, T. (1988). A recurrence formula for the distribution of the wilcoxon rank sum statistic. *Stat. Probab. Lett. 7*(2), 161–165.

Buckle, N., H. Kraft, Charles, and C. van Eeden (1969). An approximation to the wilcoxon-mann-whitney distribution. *J. Am. Stat. Assoc. 64*(326), 225–251.

Cai, T. T. and W. Sun (2017). Large-scale global and simultaneous inference: Estimation and testing in very high dimensions. *Annu. Rev. Econ. 9*, 411–439.

Candès, E., L. Lei, and Z. Ren (2023). Conformalized survival analysis. *J. R. Stat. Soc. B 85*(1), 24–45.

Casella, G. and R. Berger (2002). *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning.

Chen, Y., P. Liu, K. S. Tan, and R. Wang (2023). Trade-off between validity and efficiency of merging p-values under arbitrary dependence. *Stat. Sin. 33*(2), 851–872.

Chernoff, H. and I. R. Savage (1958). Asymptotic normality and efficiency of certain non-parametric test statistics. *Ann. Math. Stat. 29*, 972–994.

Choi, W. and I. Kim (2023). Averaging p-values under exchangeability. *Statist. Probab. Lett. 194*, 109748.

Cox, D. R. and D. V. Hinkley (1979). *Theoretical statistics*. CRC Press.

Dataset. 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.

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

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

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

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

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

Dobriban, E. (2020). Fast closed testing for exchangeable local tests. *Biometrika 107*(3), 761–768.

Dobriban, E., K. Fortney, S. K. Kim, and A. B. Owen (2015). Optimal multiple testing under a gaussian prior on the effect sizes. *Biometrika 102*(4), 753–766.

Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Stat. 32*(3), 962–994.

Donoho, D. and J. Jin (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci. 30*(1), 1–25.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc. 50*(272), 1096–1121.

Ebrahimpoor, M., P. Spitali, K. Hettne, R. Tsonaka, and J. Goeman (2020). Simultaneous enrichment analysis of all possible gene-sets: unifying self-contained and competitive methods. *Brief. Bioinform. 21*(4), 1302–1312.

Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *J. Clin. Psychol. 80*(2), 351–363.

Feroze, A., A. Daud, T. Amjad, and M. K. Hayat (2021). Group anomaly detection: past notions, present insights, and future prospects. *SN Computer Science 2*, 1–27.

Fisher, R. and F. Yates (1938). *Statistical Tables for Biological, Agricultural, and Medical Research (Table XX)*. Oliver and Boyd, London.

Fisher, R. A. (1925). Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pp. 66–70. Springer.

Fix, E. and L. J. Hodges, Joseph (1955). Significance probabilities of the wilcoxon test. *Ann. Math. Stat. 26*(2), 301–312.

Gammerman, A., K. S. Azoury, and V. Vapnik (1998). Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 148–156. Morgan Kaufmann.

Gazin, U., G. Blanchard, and E. Roquain (2024). Transductive conformal inference with adaptive scores. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, Volume 238 of *Proceedings of Machine Learning Research*, pp. 1504–1512.

Genovese, C. R. and L. Wasserman (2006). Exceedance control of the false discovery proportion. *J. Am. Stat. Assoc. 101*(476), 1408–1417.

Goeman, J. J., P. Górecki, R. Monajemi, X. Chen, T. E. Nichols, and W. Weeda (2023). Cluster extent inference revisited: quantification and localisation of brain activity. *J. R. Stat. Soc. B 85*(4), 1128–1153.

Goeman, J. J., J. Hemerik, and A. Solari (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *Ann. Stat. 49*(2), 1218–1238.

Goeman, J. J., R. J. Meijer, T. J. Krebs, and A. Solari (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika 106*(4), 841–856.

Goeman, J. J. and A. Solari (2011). Multiple testing for exploratory research. *Stat. Sci. 26*(4), 584–597.

Guan, L. and R. Tibshirani (2022). Prediction and outlier detection in classification problems. *J. R. Stat. Soc. B 84*(2), 524–546.

Heard, N. A. and P. Rubin-Delanchy (2018). Choosing between methods of combining-values. *Biometrika 105*(1), 239–246.

Heller, R., A. Krieger, and S. Rosset (2023). Optimal multiple testing and design in clinical trials. *Biometrics 79*(3), 1908–1919.

Heller, R. and S. Rosset (2020). Optimal control of false discovery criteria in the two-group model. *J. R. Stat. Soc. B 83*(1), 133–155.

Heller, R. and A. Solari (2023). Simultaneous directional inference. *J. R. Stat. Soc. B 86*(3), 650–670.

Hemerik, J. and J. Goeman (2018). Exact testing with random permutations. *Test 27*(4), 811–825.

Hemerik, J. and J. J. Goeman (2021). Another look at the lady tasting tea and differences between permutation tests and randomisation tests. *Int. Stat. Rev. 89*(2), 367–381.

Hemerik, J., A. Solari, and J. J. Goeman (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika 106*(3), 635–649.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika 75*(4), 800–802.

Hodges, J. L. and E. L. Lehmann (1961). Comparison of the normal scores and wilcoxon tests. In J. Neyman (Ed.), *Berkeley Symposium on Mathematical Statistics and Probability*, pp. 307–317. Berkeley, CA: University of California Press.

Hodges, Jr., J. L. and E. L. Lehmann (1956). The efficiency of some nonparametric competitors of the t-test. *Ann. Math. Stat. 27*(2), 324–335.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Stat. 19*(3), 293–325.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Ann. Math. Stat. 23*(3), 169–192.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. Stat. 6*(2), 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika 75*(2), 383–386.

Hore, R. and R. F. Barber (2024). Conformal prediction with local weights: randomization enables robust guarantees. *J. R. Stat. Soc. B 87*(2), 549–578.

Hu, X. and J. Lei (2024). A two-sample conditional distribution test using conformal prediction and weighted rank sum. *J. Am. Stat. Assoc. 119*(546), 1136–1154.

Hájek, J., Z. Šidák, and P. K. Sen (1999). *Theory of Rank Tests* (Second ed.). Academic Press.

Kasieczka, G., B. Nachman, D. Shih, O. Amram, A. Andreassen, K. Benkendorfer, B. Bortolato, G. Brooijmans, F. Canelli, J. H. Collins, et al. (2021). The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics. *Reports on progress in physics 84*(12), 124201.

Katsevich, E. and A. Ramdas (2020). Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *Ann. Stat. 48*(6), 3465–3487.

Koning, N. W. and J. Hemerik (2024). More efficient exact group invariance testing: using a representative subgroup. *Biometrika 111*(2), 441–458.

Kuchibhotla, A. K. (2021). Exchangeability, conformal prediction, and rank tests.

Laxhammar, R. and G. Falkman (2015). Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Ann. Math. Artif. Intell. 74*, 67–94.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *Ann. Math. Stat. 22*(2), 165–179.

Lehmann, E. L. (1953). The power of rank tests. *Ann. Math. Stat. 24*, 23–42.

Lehmann, E. L. (2009). Parametric versus nonparametrics: two alternative methodologies. *J. Nonparametr. Stat. 21*(4), 397–405.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses* (3 ed.). Springer Texts in Statistics. Springer New York, NY.

Lei, L. and E. J. Candès (2021). Conformal inference of counterfactuals and individual treatment effects. *J. R. Stat. Soc. B 83*(5), 911–938.

Li, J., M. H. Maathuis, and J. J. Goeman (2024). Simultaneous false discovery proportion bounds via knockoffs and closed testing. *J. R. Stat. Soc. B 86*(4), 966–986.

Liang, Z., M. Sesia, and W. Sun (2024). Integrative conformal p-values for out-of-distribution testing with labelled outliers. *J. R. Stat. Soc. B 86*(3), 671–693.

Magnani, C. G., M. Sesia, and A. Solari (2024). Collective outlier detection and enumeration with conformalized closed testing. https://arxiv.org/abs/2308.05534.

Magnani, C. G. and A. Solari (2021). Hommel bh: an adaptive benjamini-hochberg procedure using hommel's estimator for the number of true hypotheses. In C. Perna, N. Salvati, and F. Schirripa Spagnolo (Eds.), *Book of Short Papers of the Italian Statistical Society 2021*. Pearson.

Mann, H. B. and D. R. Whitney (1947a). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat. 18*(1), 50–60.

Mann, H. B. and D. R. Whitney (1947b). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat. 18*(1), 50 – 60.

Marandon, A., L. Lei, D. Mary, and E. Roquain (2024). Adaptive novelty detection with false discovery rate guarantee. *Ann. Stat. 52*(1), 157–183.

Marcus, R., E. Peritz, and K. R. Gabriel (1976). Closed testing procedures with special reference to ordered analysis of variance. *Biometrika 1*(63), 655–660.

Meijer, R. J., T. J. P. Krebs, and J. J. Goeman (2019). Hommel's procedure in linear time. *Biometrical Journal 61*(1), 73–82.

Meinshausen, N. and J. Rice (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Stat. 34*(1), 373–393.

Meng, X.-L. (1994). Posterior predictive p-values. *Ann. Stat. 22*(3), 1142–1160.

Milton, R. C. (1964). An extended table of critical values for the mann-whitney (wilcoxon) two-sample statistic. *J. Am. Stat. Assoc. 59*(307), 925–934.

Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *Ann. Stat. 37*(6B), 3867 – 3892.

Patra, R. K. and B. Sen (2016). Estimation of a two-component mixture model with applications to multiple testing. *J. R. Statist. Soc. B 78*(4), 869—893.

Pesarin, F. (2001). *Multivariate Permutation Tests: With Applications in Biostatistics*. Wiley.

Pesarin, F. (2015). Some elementary theory of permutation tests. In *Nonparametric Statistics for the Behavioral Sciences*, pp. 107–137. Wiley.

Pesarin, F. and L. Salmaso (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

Pesarin, F. and L. Salmaso (2025). *Permutation Tests for Complex Data: Theory, Applications and Software, 2nd Edition*. Wiley.

Pitman, E. J. G. (1949). Lecture notes on nonparametric statistical inference. Given at Columbia University.

Rosenblatt, J. D., L. Finos, W. D. Weeda, A. Solari, and J. J. Goeman (2018). All-resolutions inference for brain imaging. *Neuroimage 181*, 786–796.

Rosset, S., R. Heller, A. Painsky, and E. Aharoni (2022). Optimal and maximin procedures for multiple testing problems. *J. R. Stat. Soc. B 84*, 1105–1128.

Rüschendorf, L. (1982). Random variables with maximum sums. *Adv. in Appl. Probab. 14*(3), 623–632.

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture. *Ann. Stat. 26*(2), 494–504.

Sarkar, S. K. (2008). On the Simes inequality and its generalization. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, Volume 1, pp. 231–243. Institute of Mathematical Statistics.

Sarkar, S. K. and C.-K. Chang (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Am. Stat. Assoc. 92*(440), 1601–1608.

Saunders, C., A. Gammerman, and V. Vovk (1999). Transduction with confidence and credibility. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pp. 722–726. Morgan Kaufmann.

Schweder, T. and E. Spjøtvoll (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika 69*(3), 493–502.

Sesia, M. and V. Svetnik (2025). Doubly robust conformalized survival analysis with right-censored data. In *Forty-second International Conference on Machine Learning*.

Shiraishi, T. (1985). Local powers of two-sample and multi-sample rank tests for lehmann's contaminated alternative. *Ann. Inst. Stat. Math. 37*, 519–527.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*(3), 751–754.

Solari, A. and J. J. Goeman (2017). Minimally adaptive bh: A tiny but uniform improvement of the procedure of benjamini and hochberg. *Biom. J. 59*(4), 776–780.

Sonnemann, E. (1982). Allgemeine lösungen multipler testprobleme. *EDV in Med. und Biol. 13*, 120–128.

Spjotvoll, E. (1972). On the optimality of some multiple comparison procedures. *Ann. Math. Stat. 43*(2), 398–411.

Steel, R. G. (1959). A multiple comparison rank sum test: treatments versus control. *Biometrics 15*(4), 560–572.

Stoepker, I. V., R. M. Castro, E. Arias-Castro, and E. van den Heuvel (2024). Anomaly detection for a large number of streams: A permutation-based higher criticism approach. *J. Am. Stat. Assoc. 119*(545), 461–474.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. B 64*(3), 479–498.

Storey, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Stat. Soc. B 69*(3), 347–368.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. B 66*(1), 187–205.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A. 100*(16), 9440–9445.

Student (1908). The probable error of a mean. *Biometrika 6*(1), 1–25.

Su, W. J. (2018). The fdr-linking theorem.

Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Stat. Assoc. 102*(479), 901–912.

Tian, J., X. Chen, E. Katsevich, J. J. Goeman, and A. Ramdas (2023). Large-scale simultaneous inference under dependence. *Scand. J. Stat. 50*(2), 750–796.

Tibshirani, R. J., R. Foygel Barber, E. Candès, and A. Ramdas (2019). Conformal prediction under covariate shift. In *Adv. Neural Inf. Process. Syst.*, Volume 32.

Tukey, J. W. (1976). T13 n: The higher criticism. Course Notes, Statistics 411, Princeton Univ.

Tukey, J. W. (1989). Higher criticism for individual significances in several tables or parts of tables. Working Paper, Princeton Univ.

van der Vaart, A. (2000). *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press.

Vatanen, T., M. Kuusela, E. Malmi, T. Raiko, T. Aaltonen, and Y. Nagai (2012). Semi-supervised detection of collective anomalies with an application in high energy particle physics. In *International Joint Conference on Neural Networks*, pp. 1–8. IEEE.

Vesely, A., L. Finos, and J. J. Goeman (2023). Permutation-based true discovery guarantee by sum tests. *J. R. Stat. Soc. B 85*(3), 664–683.

Vovk, V., A. Gammerman, and C. Saunders (1999). Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning*, pp. 444–453.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic learning in a random world*, Volume 29. Springer.

Vovk, V., B. Wang, and R. Wang (2022). Admissible ways of merging p-values under arbitrary dependence. *Ann. Stat. 50*(1), 351–375.

Vovk, V. and R. Wang (2020). Combining p-values via averaging. *Biometrika 107*(4), 791–808.

Westfall, P., A. Krishen, and S. Young (1998). Using prior information to allocate significance levels for multiple endpoints. *Stat. Med. 17*(18), 2107–2119.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin 1*(6), 80–83.

Xu, Z., A. Solari, L. Fischer, R. de Heide, A. Ramdas, and J. Goeman (2025). Bringing closure to false discovery rate control: A general principle for multiple testing.

Yang, Y., A. K. Kuchibhotla, and E. Tchetgen Tchetgen (2024). Doubly robust calibration of prediction sets under covariate shift. *J. R. Stat. Soc. B 86*(4), 943–965.