

Structure-Adaptive Conformal Inference for Large-Scale Out-of-Distribution Testing

Rongyi Sun*, Wenguang Sun[†] and Zinan Zhao[‡]

Abstract

This paper addresses the challenge of structured out-of-distribution (OOD) testing in high-stakes machine learning (ML) applications. Traditional conformal methods rely on the strict joint exchangeability assumption, rendering them unsuitable for data-rich scenarios where valuable auxiliary information – such as spatiotemporal or grouping structures – is available. To overcome this limitation, we present the structure-adaptive conformal q-value (SCQ), a novel significance index that integrates individual test evidence with structural patterns. Additionally, our work develops the pseudo-score-guided transductive automated model selection (P-TAMS) algorithm, which adapts conformalized model selection techniques to leverage structural insights and optimize OOD testing performance across a toolbox of candidate ML models. Together, SCQ and P-TAMS form a unified and flexible framework based on the weaker assumption of pairwise exchangeability, offering guaranteed error rate control, enhanced statistical power, and improved interpretability in detecting structured anomalies. We assess the performance of our approach through extensive experiments on both simulated and real data. The results demonstrate that our proposed approach effectively controls the false discovery rate and outperforms competing methods in power across a variety of settings.

Keywords: Automated Model Selection; Conformal q-values; False Discovery Rate; Pairwise Exchangeability; Side Information

1 Introduction

Out-of-distribution (OOD) testing, also known as outlier or novelty detection, is a crucial task in a range of important domains such as medical diagnostics, image detection, security monitoring, and autonomous driving (Tarassenko et al., 1995; Pimentel et al., 2014; Lee et al., 2018; Yang et al., 2024). The goal is to identify instances that deviate from a reference distribution of labeled inliers. State-of-the-art OOD testing methods often rely on complex machine learning (ML) algorithms that do not provide rigorous uncertainty quantification for their outputs. These limitations present significant challenges in high-stakes

*Center for Data Science and School of Mathematical Sciences, Zhejiang University.

[†]Center for Data Science and School of Management, Zhejiang University.

[‡]Center for Data Science and School of Mathematical Sciences, Zhejiang University.

settings, where precise risk assessment and strict error control are essential. Conformal inference (Vovk et al., 1999, 2005; Balasubramanian et al., 2014; Lei and Wasserman, 2014) has emerged as a powerful framework that grounds ML algorithms in rigorous theoretical foundations; particularly, methods based on conformal p-values (Bates et al., 2023; Marandon et al., 2024; Liang et al., 2024b) provide guaranteed finite-sample error rate control, enabling reliable out-of-distribution testing in risk-sensitive environments.

When conducting thousands of tests simultaneously, controlling the false discovery rate (FDR; Benjamini and Hochberg, 1995) provides a practical strategy to balance the benefits of detecting true outliers against the costs of reporting false discoveries. In data-intensive applications, such as genomics, neuroimaging, or IoT sensor monitoring, valuable auxiliary information (derived from spatiotemporal proximities, data-driven clusters, or domain knowledge) is frequently available. These structures provide critical and valuable insights into the analysis. For instance, in spatiotemporal anomaly detection, outliers often occur in clusters rather than in isolation. Exploiting these dependencies enables the aggregation of weak signals from nearby locations, thereby boosting detection power and producing more meaningful findings. Additionally, by considering hierarchical correlations and grouping structures in the data, localized comparisons can be performed to enhance both sensitivity and interpretability of the OOD testing procedure.

However, structured FDR methods (Li and Barber, 2019; Ignatiadis and Huber, 2021; Cai et al., 2022) are not directly applicable to OOD testing as they typically depend on strict distributional assumptions for deriving p-values – assumptions that are often violated or unverifiable in modern ML applications with complex data structures. Our goal is to harness the power of conformal inference to develop principled and effective procedures for structured OOD testing; this entails addressing two key challenges.

The first challenge arises from a fundamental conflict: incorporating structural information intentionally places test units on unequal footing, thereby violating the joint exchangeability assumption required by standard conformal methods. Enforcing this assumption runs counter to the goal of detecting structured anomalies, leading to inferences that are invalid, inefficient, and less interpretable. A second challenge stems from a significant limitation in model selection. Although the conformal framework is inherently model-agnostic and can, in theory, be deployed with any off-the-shelf ML model, its implementation typically relies on a fixed pre-trained model. This pre-training strategy overlooks the fact that model performance can vary significantly across diverse data structures and contextual settings. To overcome this limitation, we propose data-adaptive, provably valid strategies that directly integrate structural knowledge into the model selection process, thereby fully leveraging the capabilities of the conformal paradigm.

1.1 A preview of our proposal and contributions

To bypass the strict requirement of joint exchangeability, Section 2 introduces the structure-adaptive conformal q-value (SCQ), which operates under the weaker assumption of pairwise exchangeability. The SCQ combines unit-specific evidence with relevant structural information, resulting in a structure-aware risk measure with guaranteed statistical validity.

To address the challenge of model selection, Section 3 extends conformalized automated model selection (CAMS) methods (Liang et al., 2023; Magnani et al., 2023; Marandon et al., 2024; Liang et al., 2024b) to structured settings. The proposed algorithm, termed pseudo-

score-guided transductive automated model selection (P-TAMS), employs innovative min-max pairing and coin-flipping strategies to maintain pairwise exchangeability throughout the model selection process. P-TAMS cohesively integrates with the SCQ procedure, forming a unified framework for the efficient detection of structured anomalies.

The contributions of our work include:

- *A structure-aware significance index:* SCQ offers a principled, user-friendly, and interpretable significance index in structured settings, enabling a precise and informative assessment of the relative importance of test units in light of side information.
- *A novel CAMS strategy:* P-TAMS extends existing CAMS methods to structured settings, thereby enhancing SCQ by offering provable validity and improved efficiency for OOD testing tasks.
- *A unified framework under weaker assumptions:* Both SCQ and P-TAMS rely solely on the weaker assumption of pairwise exchangeability, resulting in a coherent framework specifically tailored for structured OOD testing problems. This approach delivers increased flexibility, enhanced power, and improved interpretability.
- *An in-depth theoretical analysis of the SCQ procedure:* We develop an asymptotic framework for analyzing the theoretical properties of SCQ. Our theory addresses two critical gaps in the literature. First, we characterize the conditions under which SCQ are anti-conservative in FDR control. Second, we establish when and why integrating structural knowledge yields increased power. These contributions are enabled by a set of novel theoretical tools designed to accommodate weak dependence, structural patterns, and data-driven thresholding.

1.2 Related work

Our work integrates several latest advances from related areas, such as conformal inference for OOD testing, structured multiple testing, model-free FDR control, and conformalized automatic model selection. We discuss these related works to clarify our contribution.

Conformal Inference for OOD Testing. Recent studies (e.g., [Mary and Roquain, 2022](#); [Bates et al., 2023](#); [Marandon et al., 2024](#); [Liang et al., 2024b](#)) have successfully employed conformal p-values to perform OOD testing with finite-sample guarantees. However, the theoretical validity of existing methods critically depends on the joint exchangeability assumption, which fundamentally limits their applicability to structured testing scenarios. Our research directly addresses this limitation by relaxing this stringent assumption.

Structured Multiple Testing. The design of SCQ is inspired by research on structured multiple testing, which demonstrates that incorporating side information enhances both power and interpretability ([Genovese et al., 2006](#); [Benjamini and Heller, 2007](#); [Sun and Cai, 2009](#); [Sun et al., 2015](#)). Our work closely aligns with weighted p-value methods, such as IHW ([Ignatiadis and Huber, 2021](#)), SABHA ([Li and Barber, 2019](#)), and LAWS ([Cai et al., 2022](#)). However, these methods become infeasible in modern ML applications; our SCQ framework addresses this limitation by offering a model-free, theoretically grounded alternative that can be deployed with black-box classifiers.

Automated Model Selection. The effectiveness of OOD testing methods critically depends on the choice of underlying classifier, which can be a one-class classifier (OCC)

(Moya et al., 1993; Khan and Madden, 2014; Sabokrou et al., 2018), a binary classifier (BIC) (Kumari and Srivastava, 2017; Haroush et al., 2021), or a classifier constructed via positive-unlabeled learning techniques (PUC; Du Plessis et al., 2014, Bekker and Davis, 2020). The problem of automatically selecting among these different types of classifiers was recently explored by Liang et al. (2023), Magnani et al. (2023), Liang et al. (2024b), Marandon et al. (2024), and Bai and Jin (2024). However, existing conformalized model selection methods cannot handle structured setups; we provide a detailed comparison with these methods in Appendix C.2. Moreover, the BONuS (Yang et al., 2021), MetaOD (Zhao et al., 2021) and AutoMS (Zhang et al., 2022) methods explored model selection ideas but not within the conformal context, while Yang and Kuchibhotla (2021), Stutz et al. (2021), Einbinder et al. (2022), and Liang et al. (2024a) studied classifier selection for conformal prediction, but the ideas are inapplicable to OOD testing.

Model-Free FDR Control. The construction of SCQ is conceptually inspired by FDR methodologies employing mirror processes. This line of work includes knockoffs filters (Barber and Candès, 2015), AdaPT (Lei and Fithian, 2018), ZAP (Leung and Sun, 2022), SDA (Du et al., 2023), adaptive knockoffs (Ren and Candès, 2023), and PLIS (Zhao and Sun, 2025b), among others. However, these works are not designed for OOD testing tasks and lack the machinery to integrate automated model selection strategies.

Post-Selection Inference. The literature on post-selection inference with false coverage-statement control (Benjamini and Yekutieli, 2005; Bao et al., 2024; Gazin et al., 2025; Jin and Ren, 2025; Gui et al., 2025) focuses on drawing inferences after the *selection of test units*. In contrast, our method operates along a different dimension – namely, *selecting among different classifiers*.

2 FDR Control for Structured OOD Testing

This section develops a generic methodology for structured OOD testing that integrates OCCs, BICs, and PUCs into a unified framework; this lays the groundwork for Section 3, where we present the P-TAMS algorithm for model selection across these classifier families. In Section 2.1, we begin by formulating the problem and introducing a unified approach to constructing conformal p-values that accommodates various classifiers. Section 2.2 provides detailed descriptions of the SCQ procedure and establishes its theoretical properties.

2.1 Problem formulation

Suppose we observe n labeled data points $\{(X_i, Y_i) : i \in [n]\}$, where $X_i \in \mathbb{R}^p$ is a p -dimensional feature and $Y_i \in \{0, 1\}$ denotes the label, with $Y_i = 0$ and $Y_i = 1$ corresponding to an inlier and an outlier, respectively. Define the index sets $\mathcal{D}_0 = \{i \in [n] : Y_i = 0\}$ and $\mathcal{D}_1 = \{i \in [n] : Y_i = 1\}$. In addition, we have m unlabeled features $(X_{n+1}, \dots, X_{n+m})$ with corresponding unknown labels $(Y_{n+1}, \dots, Y_{n+m})$. Let $\mathcal{D}^{\text{test}} = \{n+1, \dots, n+m\}$. Denote \mathcal{H}_0 and \mathcal{H}_1 the index sets of inliers and outliers in $\mathcal{D}^{\text{test}}$, respectively. Structural information associated with each test sample X_j – such as group membership or spatial location – is captured by an external covariate S_j . Denote $\mathbf{S} = \{S_j : j \in \mathcal{D}^{\text{test}}\}$.

The decision rule for determining which of the m instances are outliers can be represented by $(\hat{Y}_{n+1}, \dots, \hat{Y}_{n+m}) \in \{0, 1\}^m$, where $\hat{Y}_j = 1$ indicates that we classify the j -th instance as an outlier, and $\hat{Y}_j = 0$ otherwise. Define the false discovery rate

$$\text{FDR} := \mathbb{E} \left[\frac{\sum_{j \in \mathcal{D}^{\text{test}}} (1 - Y_j) \hat{Y}_j}{\left(\sum_{j \in \mathcal{D}^{\text{test}}} \hat{Y}_j \right) \vee 1} \right],$$

where the expectation is taken over the labeled and test data.

We briefly outline the basic workflow of structured OOD testing. First, a collection of null samples is partitioned into three distinct subsets, each serving a specific purpose: (i) a pre-training set ($\mathcal{D}_0^{\text{tr}}$) for model training; (ii) a calibration set ($\mathcal{D}_0^{\text{cal}}$) for quantifying the uncertainty of model outputs; and (iii) a mirror set ($\mathcal{D}_0^{\text{mir}}$) for recalibrating with structural knowledge (see Section 2.2). Our method requires that $|\mathcal{D}_0^{\text{mir}}| = |\mathcal{D}^{\text{test}}| = m$. Denote the test data as $\mathbf{X} := (X_j : j \in \mathcal{D}^{\text{test}})$ and rewrite the mirror data ($X_i : i \in \mathcal{D}_0^{\text{mir}}$) as $\tilde{\mathbf{X}} := (\tilde{X}_j : j \in \mathcal{D}^{\text{test}})$. This modified notation allows us to pair the observations from $\mathcal{D}_0^{\text{mir}}$ and $\mathcal{D}^{\text{test}}$ as $\{(X_j, \tilde{X}_j) : j \in \mathcal{D}^{\text{test}}\}$, thereby simplifying the description of the method. Define $\mathbf{X}_1 = \{X_i : i \in \mathcal{D}_1\}$, $\mathbf{X}_0^{\text{tr}} = \{X_i : i \in \mathcal{D}_0^{\text{tr}}\}$, and $\mathbf{X}_0^{\text{cal}} = \{X_i : i \in \mathcal{D}_0^{\text{cal}}\}$.

Next, we train a score function $s(\cdot)$, applicable across various classifier families, that satisfies the following permutation-invariance condition:

$$s(\cdot; (\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{X}_0^{\text{cal}})_{\Pi}, \mathbf{X}_0^{\text{tr}}, \mathbf{X}_1, \mathbf{S}) = s(\cdot; (\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{X}_0^{\text{cal}}), \mathbf{X}_0^{\text{tr}}, \mathbf{X}_1, \mathbf{S}), \quad (1)$$

where Π denotes an arbitrary permutation of the data points in \mathbf{X} , $\tilde{\mathbf{X}}$, and $\mathbf{X}_0^{\text{cal}}$. Define the observed training data as \mathbf{O}^{tr} , which may consist solely of null samples or may also include labeled outliers. The permutation-invariance principle in (1) permits the use of mirror and test data together with \mathbf{O}^{tr} and encompasses several important special cases:

- OCC: $s(\cdot) := s(\cdot; \mathbf{O}^{\text{tr}})$ with $\mathbf{O}^{\text{tr}} = \{X_i : i \in \mathcal{D}_0^{\text{tr}}\}$;
- BIC: $s(\cdot) := s(\cdot; \mathbf{O}^{\text{tr}})$ with $\mathbf{O}^{\text{tr}} = \{(X_i, Y_i) : i \in \mathcal{D}_0^{\text{tr}} \cup \mathcal{D}_1\}$;
- PUC: $s(\cdot) := s(\cdot; \mathbf{O}^{\text{tr}}, (\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{X}_0^{\text{cal}})_{\Pi}) = s(\cdot; \mathbf{O}^{\text{tr}}, (\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{X}_0^{\text{cal}}))$, $\mathbf{O}^{\text{tr}} = \{X_i : i \in \mathcal{D}_0^{\text{tr}}\}$.

In OCC and BIC cases (e.g., Bates et al., 2023; Liang et al., 2024b), the score function $s(\cdot)$ does not depend on the test or mirror data; hence, property (1) is automatically satisfied. In contrast, the PUC approach (e.g., Marandon et al., 2024; Zhao and Sun, 2025a) is carefully designed to fulfill property (1). As elaborated in Appendix C.2, each classifier family has distinct advantages and limitations; a key advantage of the generic formulation outlined in (1) is its ability to provide a unified framework for SCQ construction (Section 2.2), model selection (Section 3), and theoretical analyses (Section 4) across a broad class of classifier families.

2.2 The SCQ procedure

Consider the partition introduced in Section 2.1, where $\mathcal{D}_0 = \mathcal{D}_0^{\text{tr}} \cup \mathcal{D}_0^{\text{cal}} \cup \mathcal{D}_0^{\text{mir}}$ with $|\mathcal{D}_0^{\text{mir}}| = |\mathcal{D}^{\text{test}}| = m$. The SCQ procedure operates in three steps.

Step 1: Compute conformal p-values (first calibration). Let $s(\cdot) \in \mathcal{G}$ denote a score function that takes the generic form given in (1), with smaller values indicating stronger evidence against the null hypothesis. The conformal p-value function is defined as:

$$p(x) = \frac{1 + \left| \left\{ i \in \mathcal{D}_0^{\text{cal}} : s(X_i) \leq s(x) \right\} \right|}{1 + |\mathcal{D}_0^{\text{cal}}|}. \quad (2)$$

Through (1) and (2), we first build a predictive model to compute preliminary conformity scores, and then construct m pairs of conformal p-values: $\{(p(X_j), p(\tilde{X}_j)) : j \in \mathcal{D}^{\text{test}}\}$.

The conformal p-values are inadequate for structured OOD testing problems, as the heterogeneity among test units has been ignored. The next step details a weighting strategy to incorporate structural knowledge.

Step 2: Compute weighted p-values. Let $w(S_j)$ represent the weight associated with covariate S_j , where a larger value of $w(S_j)$ provides stronger indication that the j -th instance is an outlier. The key idea is to prioritize instances that are more likely to be outliers through the use of weighted p-values:

$$\{(V_j := p(X_j)/w(S_j), \tilde{V}_j := p(\tilde{X}_j)/w(S_j)) : j \in \mathcal{D}^{\text{test}}\}, \quad (3)$$

where (V_j, \tilde{V}_j) will be utilized as new conformity scores in the next step.

The weight function $w(S_j)$ can be determined based on prior knowledge or structural information. For example, in the spatial multiple testing setting where S_j denotes the spatial location, let $\pi(S_j)$ represent the local sparsity level, with higher values of $\pi(S_j)$ indicating a greater density of outliers near S_j . In this context, higher weights $w(S_j)$ should be assigned to locations with higher $\pi(S_j)$ (Li and Barber, 2019; Cai et al., 2022). We require that if data-driven weights are used, the weight function $w(S_j; (\mathbf{X}, \tilde{\mathbf{X}}), \mathbf{X}_0^{\text{tr}}, \mathbf{X}_0^{\text{cal}}, \mathbf{X}_1, \mathbf{S})$ must belong to a generic class of swap-invariance functions \mathcal{G} , where for each $g \in \mathcal{G}$, we have

$$g(\cdot; (\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(\mathcal{J})}, \mathbf{X}_0^{\text{tr}}, \mathbf{X}_0^{\text{cal}}, \mathbf{X}_1, \mathbf{S}) = g(\cdot; (\mathbf{X}, \tilde{\mathbf{X}}), \mathbf{X}_0^{\text{tr}}, \mathbf{X}_0^{\text{cal}}, \mathbf{X}_1, \mathbf{S}), \quad \forall \mathcal{J} \subset \mathcal{D}^{\text{test}}. \quad (4)$$

Here, $(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(\mathcal{J})}$ denotes the swapping of X_j and \tilde{X}_j for each $j \in \mathcal{J}$. We detail how to construct data-driven weights fulfilling (4) in Appendix B.1.

Step 3: Calculate SCQs (second calibration). Unlike standard conformal p-values, the weighted p-values constructed from (3) are not directly interpretable and do not preserve theoretical properties such as super-uniformity under the null. Hence, we apply a second calibration step employing the following mirror process $H(t)$:

$$H(t) = \frac{1 + \sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{I}(\tilde{V}_j \leq t, V_j < \tilde{V}_j)}{\left[\sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{I}(V_j \leq t, V_j < \tilde{V}_j) \right] \vee 1}, \quad t > 0. \quad (5)$$

Let $\mathcal{V} = \{V_j : j \in \mathcal{D}^{\text{test}}\}$ and $\tilde{\mathcal{V}} = \{\tilde{V}_j : j \in \mathcal{D}^{\text{test}}\}$. Define the conformal q-values:

$$q_j := \begin{cases} \min_{t \in \mathcal{V} \cup \tilde{\mathcal{V}}, t \geq V_j} H(t), & \text{if } V_j < \tilde{V}_j \\ 1, & \text{if } V_j \geq \tilde{V}_j \end{cases}, \quad j \in \mathcal{D}^{\text{test}}. \quad (6)$$

The rejection set of the SCQ procedure is then given by

$$\mathcal{R}_{\text{scq}} = \{j \in \mathcal{D}^{\text{test}} : q_j \leq \alpha\}. \quad (7)$$

Algorithm A.2 in Appendix B.2 summarizes the key steps of the proposed method.

Remark 1. As explained in Section 4.1, the mirror process $H(t)$ is carefully constructed to estimate the false discovery proportion; hence, the q-value provides an intuitive and

interpretable structure-aware risk measure. Moreover, the SCQ procedure is simple to use, enabling practitioners to make decisions by directly comparing q-values to a pre-specified level α . Finally, the validity of SCQ requires only the pairwise exchangeability of scores – a considerably weaker condition than joint exchangeability; this flexibility allows SCQ to effectively incorporate local heterogeneities while ensuring its validity for FDR control.

We now establish the theoretical properties of the SCQ procedure. Consider m pairs of conformity scores $\{(V_j, \tilde{V}_j) : j \in \mathcal{D}^{\text{test}}\}$. Let $\mathcal{V}_{-i} = \{V_j : j \in \mathcal{D}^{\text{test}} \setminus \{n+i\}\}$ and $\tilde{\mathcal{V}}_{-i} = \{\tilde{V}_j : j \in \mathcal{D}^{\text{test}} \setminus \{n+i\}\}$. The following proposition shows that the conformity scores constructed via Steps 1-2 are pairwise exchangeable:

$$(V_i, \tilde{V}_i, \mathcal{V}_{-i}, \tilde{\mathcal{V}}_{-i}) \stackrel{d}{=} (\tilde{V}_i, V_i, \mathcal{V}_{-i}, \tilde{\mathcal{V}}_{-i}), \quad \forall i \in \mathcal{H}_0. \quad (8)$$

Proposition 1. *Suppose the data points satisfy the pairwise exchangeability condition:*

$$(X_i : i \in \mathcal{D}_0; X_j : j \in \mathcal{H}_0) \text{ are exchangeable conditional on } (X_j : j \in \mathcal{H}_1, \mathbf{X}_1, \mathbf{S}). \quad (9)$$

Consider the weighted p-values $\{(V_j, \tilde{V}_j) : j \in \mathcal{D}^{\text{test}}\}$ constructed via (2) and (3). Assume that $s(\cdot)$ and $w(\cdot)$ satisfy the conditions in (1) and (4), respectively. Then $\{(V_j, \tilde{V}_j) : j \in \mathcal{D}^{\text{test}}\}$ satisfy the pairwise exchangeability condition (8).

The next theorem establishes the finite-sample validity of the SCQ procedure.

Theorem 1. *Under the conditions in Proposition 1, the SCQ procedure controls the FDR at level α .*

Remark 2. The conformal p-values, constructed using a continuous score function $s(\cdot)$ that satisfies (1), are super-uniform under the null when conditions (1) and (9) hold. To see this, observe that (9) satisfies the exchangeability requirement of Assumption 1 in [Marandon et al. \(2024\)](#), then the result follows from Theorem 3.3 in [Marandon et al. \(2024\)](#).

Remark 3. Similar assumptions to (9) are widely adopted in the structured multiple testing literature ([Ignatiadis and Huber, 2021](#); [Li and Barber, 2019](#); [Cai et al., 2022](#)), where null p-values are typically assumed to be independent and super-uniform conditional on the auxiliary covariates and the non-null p-values. For clarity and interpretability, Theorem 1 is presented under a slightly stronger set of conditions. However, the result remains valid under strictly weaker assumptions: (a) the score function $s(\cdot)$ need only satisfy swap-invariance (4) in place of the permutation-invariance required in (1), and (b) the joint exchangeability condition (9) can be relaxed to the following pairwise exchangeability:

$$((\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(\mathcal{J})} \mid \mathbf{X}_0^{\text{tr}}, \mathbf{X}_0^{\text{cal}}, \mathbf{X}_1, \mathbf{S}) \stackrel{d}{=} (\mathbf{X}, \tilde{\mathbf{X}} \mid \mathbf{X}_0^{\text{tr}}, \mathbf{X}_0^{\text{cal}}, \mathbf{X}_1, \mathbf{S}), \quad \forall \mathcal{J} \subset \mathcal{H}_0. \quad (10)$$

These relaxations do not compromise the FDR guarantee established in Theorem 1.

3 Conformalized Model Selection with P-TAMS

3.1 General considerations

The selection of an effective classifier – whether OCC, BIC, or PUC – is critical to enhancing the power of OOD testing. The performance of classifier families can vary considerably

across settings, as each family possesses distinct strengths and weaknesses that are influenced by factors such as data dimensionality and class imbalance (cf. Liang et al., 2024b and Appendix C.2). Moreover, within each family, specific implementations (such as random forests, support vector machines, or neural networks) can yield substantially different results across various datasets. This variability motivates the development of a principled, data-driven automated model selection (AMS) approach to adaptively identify the most suitable model from a toolbox of candidates.

However, if applied without proper adjustment, a naive model selection strategy that simply cherry-picks the best-performing model may lead to an inflation of the error rate. Within the conformal framework, Magnani et al. (2023) demonstrated that unadjusted model selection breaks the required exchangeability between test and calibration samples, leading to invalid inference – a phenomenon we demonstrate numerically in Appendix D.2. Similar concerns about double-dipping bias in other contexts of conformal inference have been discussed by Einbinder et al. (2022) and Liang et al. (2024a).

Conformalized Automated Model Selection (CAMS; Magnani et al., 2023; Liang et al., 2023, 2024b; Marandon et al., 2024, among others) employs carefully designed calibration techniques to preserve exchangeability between test and calibration samples throughout the model training and selection process, ensuring valid inference conditional on the selected model. The CAMS framework represents a paradigm shift from conventional post-selection inference, which relies on fixed models and strong parametric assumptions, such as linearity and sparsity. In contrast, CAMS is applicable to a randomly selected model, chosen in a data-driven manner from a flexible toolbox, fully leveraging the power of the conformal approach. By unifying model training and statistical validation in a single, coherent process, CAMS allows practitioners to harness the flexibility of ML toolboxes without compromising statistical guarantees.

Let $\mathbf{C} = \{\mathcal{C}_k : k \in [K]\}$ denote a toolbox of models. Existing CAMS strategies can be categorized into two types: (a) model-oriented selection and (b) task-oriented selection. Strategy (a) evaluates a classifier’s ability to distinguish between inliers (null model) and outliers (non-null model). For instance, as proposed by Liang et al. (2024b), the model is selected by maximizing the median difference between two sets of conformity scores computed for labeled inliers and labeled outliers, respectively. By contrast, Strategy (b) directly assesses a classifier’s utility in performing a specific task. For example, the methods in Magnani et al. (2023) and Marandon et al. (2024) select a model that maximizes (a proxy of) the number of rejections in the test samples. However, existing CAMS methods rely on joint exchangeability and are consequently ill-suited for structured inference. The next subsection develops new methodology to address the above issues.

3.2 The P-TAMS algorithm

We introduce the Pseudo-score-guided Transductive Automated Model Selection (P-TAMS) algorithm, which employs a transductive, task-oriented model selection strategy for two primary reasons. First, unlike model-oriented approaches that require labeled outliers (Liang et al., 2024b) – which are often unavailable in OOD settings – our model selection process operates without such supervision. Moreover, even when labeled outliers are available, they may exhibit distributional shifts relative to test outliers, potentially misleading model selection (cf. Section 5.3). Second, our design capitalizes on the structural knowledge inherent

in test samples by formulating the selection criterion directly on the test set (transductive) while aligning it with the utility for performing a specific task, such as maximizing the number of rejections in OOD testing (task-oriented).

A central challenge is ensuring pairwise exchangeability throughout the inferential process, where the SCQ procedure is deployed *conditional on the selected model*. One may naturally consider using the number of rejections computed from the true scores \mathcal{V} and $\tilde{\mathcal{V}}$ as the selection criterion. However, this approach is not permissible because the criterion is contingent on $H(t)$ [Eq. (5)], which treats \mathcal{V} and $\tilde{\mathcal{V}}$ asymmetrically. This asymmetry violates the pairwise exchangeability assumption essential for valid inference.

To address this challenge, we introduce carefully constructed pseudo-scores that serve as substitutes for the true scores to ensure statistical validity. Our selection criterion is defined as the *pseudo number of rejections*, computed from these pseudo-scores. The design of the pseudo-score functions is guided by two key considerations. First, swap invariance with respect to \mathbf{X} and $\tilde{\mathbf{X}}$ must be fulfilled throughout the entire training and selection process. Second, the resulting pseudo number of rejections should effectively capture the model's utility, mirroring the behavior observed with true scores. The P-TAMS algorithm proceeds in three main steps, each accompanied by detailed explanatory remarks.

Step 1: Initial partition of test samples. For each classifier $\mathcal{C}_k \in \mathbf{C}$, let $s_k(\cdot)$ denote the corresponding score function computed according to the invariance principle (4). We construct m pairs of preliminary conformal p-values via (2): $\{(p_i^k \equiv p^k(X_i), \tilde{p}_i^k \equiv p^k(\tilde{X}_i)) : i \in \mathcal{D}^{\text{test}}\}$, then form m pseudo conformal p-values by taking the pairwise minimum: $\bar{\mathcal{P}}_k = \left\{ \min(p_i^k, \tilde{p}_i^k) : i \in \mathcal{D}^{\text{test}} \right\}$. Finally, we apply the BH procedure to $\bar{\mathcal{P}}_k$ at level α_0 to obtain an initial rejection set $\bar{\mathcal{R}}_k$.

Remark 4. The primary goal of this step is to partition test samples into likely outliers and likely inliers while preserving swap invariance, which underpins the subsequent construction of pseudo-scores. After processing with the minimum operator, the likely outliers (and likely inliers) tend to exhibit smaller (and larger) p-values compared to other test units, making them more (or less) likely to be rejected. Although the minimum operator distorts the original p-values, it retains sufficient discriminatory power to distinguish between the two groups while preserving swap invariance. In contrast, alternative swap invariance operators, such as the maximum or sum, would result in significant information loss and fail to achieve effective separation. Moreover, the P-TAMS procedure is not sensitive to the choice of α_0 , which may be set slightly above the nominal FDR level α (with 2α as the default).

Step 2: Constructing pseudo conformity scores. This step represents the core innovation of P-TAMS. Denote $\mathcal{V}_k = \{V_i^k : i \in \mathcal{D}^{\text{test}}\}$ and $\tilde{\mathcal{V}}_k = \{\tilde{V}_i^k : i \in \mathcal{D}^{\text{test}}\}$ the true conformity scores for classifier \mathcal{C}_k . The pseudo scores $\mathcal{U}^k = \{U_i^k : i \in \mathcal{D}^{\text{test}}\}$ and $\tilde{\mathcal{U}}^k = \{\tilde{U}_i^k : i \in \mathcal{D}^{\text{test}}\}$ are then computed as follows.

- For likely outliers, we apply the min-max operator:

$$U_i^k = \min(V_i^k, \tilde{V}_i^k), \quad \tilde{U}_i^k = \max(V_i^k, \tilde{V}_i^k), \quad i \in \bar{\mathcal{R}}_k. \quad (11)$$

- For the remaining units (likely inliers), let $\{b_i\}$ denote i.i.d. Bernoulli(1/2) variables, we apply a random coin-flipping operator:

$$U_i^k = (1 - b_i)V_i^k + b_i\tilde{V}_i^k, \quad \tilde{U}_i^k = b_iV_i^k + (1 - b_i)\tilde{V}_i^k, \quad i \notin \bar{\mathcal{R}}_k. \quad (12)$$

Remark 5. The utilization of pseudo-scores effectively avoids the pitfalls associated with using true scores for model selection. Specifically, the mirror process $H(t)$ computed with true scores breaks the pairwise exchangeability (conditional on the selected model). This issue is mitigated by employing pseudo-scores, which faithfully preserve the swap invariance with respect to \mathbf{X} and $\tilde{\mathbf{X}}$. Additionally, our choice of symmetrization operators reduces the distortion of the true scores: (a) for likely outliers, operator (11) preserves the tendency for V_i^k to be smaller than \tilde{V}_i^k under the alternative; (b) for likely inliers, operator (12) reflects the random ordering between V_i^k and \tilde{V}_i^k under the null.

Step 3: Model selection and final decisions. For each $k \in [K]$, we implement the SCQ procedure using pseudo scores \mathcal{U}^k and $\tilde{\mathcal{U}}^k$, and record the number of pseudo rejections r_k . We then select the classifier with the largest r_k , denoted by \mathcal{C}_* . Finally, we apply the SCQ procedure again with the selected model \mathcal{C}_* to output the final decision \mathcal{R}_* .

Remark 6. By leveraging the inherent structure of the mirror-based algorithms, P-TAMS integrates seamlessly with SCQ through an innovative design. Together, SCQ and P-TAMS form a unified framework that relies solely on the mild assumption of pairwise exchangeability, thereby offering enhanced flexibility and broader applicability – making it particularly well-suited for detecting structured patterns. Moreover, rather than resorting to additional sample splitting (cf. Magnani et al., 2023; Marandon et al., 2024), P-TAMS employs carefully constructed symmetrization operators that intrinsically satisfy the swap invariance requirement. This novel design not only preserves the validity of the SCQ but also delivers efficiency gains at no extra cost of valuable samples. A schematic illustration of P-TAMS is provided in Figure 1, with its key steps outlined in Algorithm A.3 in Appendix B.2. An extension of P-TAMS is provided in Algorithm A.4 in Appendix B.3.

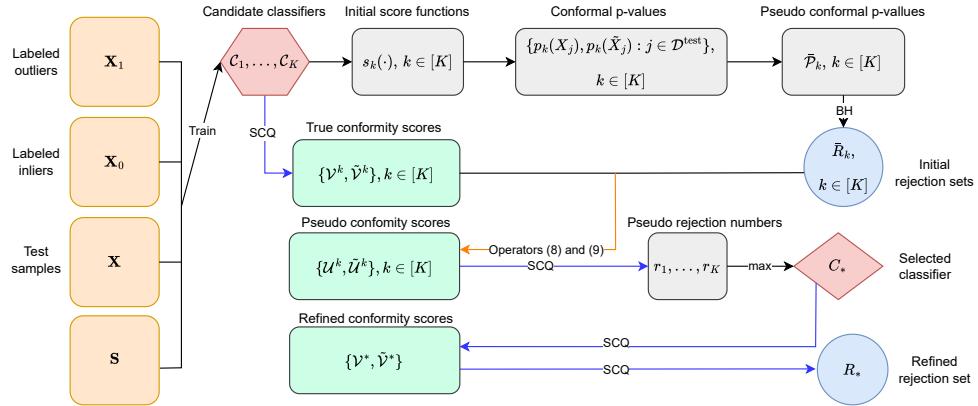


Figure 1: Schematic representation of the P-TAMS algorithm.

The following theorem establishes the finite-sample validity of P-TAMS for FDR control.

Theorem 2. Under the conditions in Proposition 1, the SCQ procedure refined with P-TAMS controls the FDR at level α .

4 Asymptotic Theories

4.1 Preliminaries

Existing conformalized OOD testing methods involve constructing conformal p-values followed by applying well-established multiple testing procedures. For example, Bates et al. (2023) showed that the conformal p-values satisfy the PRDS (positive regression dependence on subsets) condition and utilize the theory of Benjamini and Yekutieli (2001) to prove the validity of FDR control. Meanwhile, recognizing that recalibrated conformal p-values exhibit more complex dependence, Liang et al. (2024b) employed the conditional calibration framework of Fithian and Lei (2022) to establish the FDR theory. However, in structured OOD testing scenarios, the frameworks in Benjamini and Yekutieli (2001) and Fithian and Lei (2022) are inapplicable due to the violation of joint exchangeability. Instead, we establish the finite-sample FDR theory in Section 2.2 using the martingale theory in Zhao and Sun (2025b) and the e-BH theory in Wang and Ramdas (2022).

This section introduces novel tools to analyze the asymptotic properties of SCQ. The asymptotic theory provides insights into two key questions. First, while SCQ controls the FDR in finite samples, does it tend to be overly conservative? Second, SCQ employs a weighting strategy to leverage side information – when and why is this approach effective? Sections 4.2 and 4.3 develop a theoretical framework to characterize the conditions under which SCQ is asymptotically anti-conservative and demonstrates gains in power.

We begin with some preliminaries. Suppose we have constructed m pairs of conformity scores, $\{(V_j, \tilde{V}_j) : j \in \mathcal{D}^{\text{test}}\}$, that satisfy condition (8). Two thresholding rules can be considered. The first thresholds conformal q-values defined in (6); its rejection set is denoted by $\mathcal{R}_{scq} = \{j \in \mathcal{D}^{\text{test}} : q_j \leq \alpha\}$. The second thresholds conformity scores directly. Define

$$\tau = \max \left\{ t \in \mathcal{V} \cup \tilde{\mathcal{V}} : H(t) := \frac{1 + \sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{I}(\tilde{V}_j \leq t, \tilde{V}_j < V_j)}{\left[\sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{I}(V_j \leq t, V_j < \tilde{V}_j) \right] \vee 1} \leq \alpha \right\}. \quad (13)$$

The subsequent rejection set is given by

$$\mathcal{R}_{bc} = \{j \in \mathcal{D}^{\text{test}} : V_j \leq \tau, V_j < \tilde{V}_j\}. \quad (14)$$

Remark 7. The subscript ‘‘bc’’ in (14) denotes a Barber–Candès-type (BC) procedure (Barber and Candès, 2015). The generic BC algorithm – also known as the Selective SeqStep+ algorithm – was originally developed for knockoff filters in regression and recently adapted to conformalized multiple testing (Zhao and Sun, 2025b).

The q-value thresholding rule is both intuitive and interpretable, but its direct theoretical analysis can be quite challenging. In contrast, we can leverage exchangeability properties and asymptotic arguments to give a rigorous theoretical treatment of the score-thresholding rule (14). The following proposition establishes the equivalence between the two thresholding rules; as a result, our subsequent analysis will focus on the BC-type algorithm (14), which applies equally to the SCQ procedure.

Proposition 2. *Given m pairs of conformity scores $\{(V_j, \tilde{V}_j) : j \in \mathcal{D}^{\text{test}}\}$. Consider the rejection sets \mathcal{R}_{bc} and \mathcal{R}_{scq} defined above. Then $\mathcal{R}_{bc} = \mathcal{R}_{scq}$.*

4.2 Anti-conservativeness of FDR control

The BC-type procedure \mathcal{R}_{bc} given in (14) can be conservative. However, as signal strength increases, BC can adaptively attain the nominal FDR level – a phenomenon initially noted by Barber and Candès (2015); see also Appendix D.4 for an illustration. This section formally establishes this asymptotic attainment theory, tailored for conformal setups.

Consider the paired conformal p -values $\{(p_j \equiv p(X_j), \tilde{p}_j \equiv p(\tilde{X}_j)) : j \in \mathcal{D}^{\text{test}}\}$ constructed via (2), where the score function $s(\cdot)$ is assumed continuous. Let $\mathbf{w} = (w_j : j \in \mathcal{D}^{\text{test}})$ denote a generic sequence of p -value weights, and define the weighted p -values $(V_j \equiv p_j/w_j, \tilde{V}_j \equiv \tilde{p}_j/w_j)$ for each $j \in \mathcal{D}^{\text{test}}$. Consider a class of thresholding rules $\delta^{\mathbf{w}}(t) = \{\delta^{w_j}(t) : j \in \mathcal{D}^{\text{test}}\} \in \{0, 1\}^m$, where $t > 0$ is a threshold and $\delta^{w_j}(t) = \mathbb{I}(p_j/w_j \leq t, p_j/w_j < \tilde{p}_j/w_j)$. In addition, define the marginal FDR of $\delta^{\mathbf{w}}(t)$ and the corresponding oracle threshold:

$$Q(t, \mathbf{w}) = \frac{\mathbb{E} \left[\sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{I}(Y_j = 0, \delta^{w_j}(t) = 1) \right]}{\mathbb{E} \left[\sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{I}(\delta^{w_j}(t) = 1) \right]}, \quad t_{\text{OR}}^{\alpha}(\mathbf{w}) = \sup \{t \in (0, 1) : Q(t, \mathbf{w}) \leq \alpha\}. \quad (15)$$

Asymptotic FDR theory generally requires some form of weak dependence. Let $R_j(t) = \mathbb{I}(V_j \leq t, V_j < \tilde{V}_j)$ and $\tilde{R}_j(t) = \mathbb{I}(\tilde{V}_j \leq t, \tilde{V}_j < V_j)$. To lay the groundwork for our analysis, we adopt the following condition to characterize weak dependence:

$$\frac{1}{m^2} \sum_{\substack{i < j, \\ i, j \in \mathcal{D}^{\text{test}}}} \text{Cov}[R_i(t), R_j(t)] = o(1), \quad \frac{1}{m^2} \sum_{\substack{i < j, \\ i, j \in \mathcal{D}^{\text{test}}}} \text{Cov}[\tilde{R}_i(t), \tilde{R}_j(t)] = o(1). \quad (16)$$

Denote the numbers of test inliers, test outliers, and calibration data as m_0 , m_1 , and N , respectively. The lemma below shows that (V_j, \tilde{V}_j) are weakly dependent.

Lemma 1. Suppose condition (9) holds and there exist $\epsilon_1, \epsilon_N \in (0, 1)$, such that $m_1 \asymp m^{1-\epsilon_1}$ and $N \asymp m^{1+\epsilon_N}$ ¹. Then $\{(V_j, \tilde{V}_j) : j \in \mathcal{D}^{\text{test}}\}$ satisfy (16).

Next, our asymptotic theory requires the following two regularity conditions.

Assumption 1. $\frac{1}{m} \sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{P}(\tilde{V}_j < V_j | Y_j = 1) = o(1)$.

Remark 8. This assumption is mild, essentially requiring only that a subset of signals can be reliably separated from noise – a prerequisite for any meaningful FDR analysis with vanishingly small α (cf. Cai and Sun, 2017). Under the sparse, high-dimensional settings commonly studied (e.g., Donoho and Jin, 2004; Meinshausen and Rice, 2006; Arias-Castro and Wang, 2017), Assumption 1 is satisfied by the weighted conformal p -values (V_j, \tilde{V}_j) ; detailed justifications are provided in Appendix A.11.

The second assumption, implied by the condition in Equation (7) of Storey et al. (2004), is also needed in our theoretical analysis.

Assumption 2. $\mathbb{E} \left[\sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{I}(V_j \leq t_{\text{OR}}^{\alpha}(\mathbf{w}), V_j < \tilde{V}_j) \right] \asymp m$.

¹For positive sequences $a(m)$ and $b(m)$, we write $a(m) \asymp b(m)$ if there exist $C_1, C_2 > 0$ such that $a(m) \leq C_1 \cdot b(m)$ and $b(m) \leq C_2 \cdot a(m)$ for all $m \geq 1$.

Finally, combining the results in Proposition 2 and Lemma 1, we show in the next theorem that the FDR of SCQ asymptotically attains the nominal level α .

Theorem 3. *Suppose we have weighted p-values $\{(V_j, \tilde{V}_j) : j \in \mathcal{D}^{\text{test}}\}$ obtained from the SCQ procedure in Section 2.2. Consider the rejection set \mathcal{R}_{scq} defined in (7). Then, under Assumptions 1–2 and the conditions in Lemma 1, we have $\lim_{m \rightarrow \infty} \mathbb{E} \left[\frac{|\mathcal{R}_{\text{scq}} \cap \mathcal{H}_0|}{|\mathcal{R}_{\text{scq}}| \vee 1} \right] = \alpha$.*

4.3 Asymptotic power analysis

We investigate the effectiveness of p-value weighting through a novel power analysis. Existing theory on p-value weighting (e.g., Genovese et al., 2006; Cai et al., 2022; Liang et al., 2024b), developed under the BH framework, is not applicable to BC-type algorithms. Furthermore, prior work based on conformal p-values (Liang et al., 2024b) has considered only oracle thresholds (defined in (15)), while theories for data-driven thresholds (14) remain undeveloped. We address these gaps by examining weight informativeness within the BC framework and developing theory for both oracle and data-driven setups.

Suppose we have constructed m pairs of weighted p-values $\{(V_j \equiv p_j/w_j, \tilde{V}_j \equiv \tilde{p}_j/w_j) : j \in \mathcal{D}^{\text{test}}\}$ via (2) and (3). Let $\eta_j = \mathbb{I}(p_j < \tilde{p}_j)$ for each $j \in \mathcal{D}^{\text{test}}$. The following assumption indicates that an effective weighting strategy should, in general, prioritize the rejection of units that are more likely to be outliers in light of side information.

Assumption 3. *The weights $(w_j : j \in \mathcal{D}^{\text{test}})$ satisfy:*

$$\frac{\sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{P}(\eta_j = 1, Y_j = 0 \mid S_j)}{\sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{P}(Y_j = 0 \mid S_j) w_j} \cdot \frac{\sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{P}(\eta_j = 1, Y_j = 1 \mid S_j)}{\sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{P}(\eta_j = 1, Y_j = 1 \mid S_j) w_j^{-1}} \geq 1. \quad (17)$$

Remark 9. This assumption, inspired by Genovese et al. (2006), provides an insightful principle for constructing data-driven weights, as implemented in Li and Barber (2019) and Cai et al. (2022). Following this principle, for instance, higher weights are assigned to spatial regions with greater outlier abundance in spatial multiple testing.

The next assumption is a regularity condition on the conditional alternative distribution $F_{1,1,j}(t)$ of p_j , which reduces to convexity of $F_{1,1,j}(t/x)$ under homogeneity across $j \in \mathcal{D}^{\text{test}}$. Similar conditions appear in Hu et al. (2010), Cai et al. (2022), and Liang et al. (2025).

Assumption 4. $\{F_{1,1,j}(t) := \mathbb{P}(p_j \leq t \mid \eta_j = 1, Y_j = 1, S_j) : j \in \mathcal{D}^{\text{test}}\}$ satisfy:

$$\sum_{j \in \mathcal{D}^{\text{test}}} a_j F_{1,1,j}(t/x_j) \geq \sum_{j \in \mathcal{D}^{\text{test}}} a_j F_{1,1,j} \left(\frac{t \sum_{i=1}^m a_i}{\sum_{i=1}^m a_i x_i} \right), \quad (18)$$

for any $0 \leq a_j \leq 1$, $\min_{1 \leq j \leq m} w_j^{*-1} \leq x_j \leq \max_{1 \leq j \leq m} w_j^{*-1}$, and $t > 0$, where

$$w_j^* = w_j \cdot \frac{\mathbb{P}(\eta_j = 1 \mid Y_j = 0, S_j) \sum_{i \in \mathcal{D}^{\text{test}}} \mathbb{P}(\eta_i = 1, Y_i = 0 \mid S_i)}{\sum_{i \in \mathcal{D}^{\text{test}}} w_i \mathbb{P}(\eta_i = 1, Y_i = 0 \mid S_i)}. \quad (19)$$

Our power analysis proceeds in two steps. First, under the oracle setup using the threshold in (15), we compare the power of weighted and unweighted schemes at the same

FDR level, revealing the benefits of weighting for producing superior rankings. Second, we examine the more intricate data-driven setup, analyzing power gains of weighted BC-type methods using thresholds in (13), a setting not previously addressed in the literature.

1. Power Analysis for the Oracle Setup. Let $\delta_{\text{OR}}^{\mathbf{w}} = (\delta_{\text{OR}}^{w_j} : j \in \mathcal{D}^{\text{test}}) \in \{0, 1\}^m$ denote the oracle rule with weights \mathbf{w} and threshold $t_{\text{OR}} \equiv t_{\text{OR}}^{\alpha}(\mathbf{w})$ from (15), where $\delta_{\text{OR}}^{w_j} = \mathbb{I}(p_j/w_j \leq t_{\text{OR}}, p_j/w_j < \tilde{p}_j/w_j)$. Define $\Psi_{\text{OR}}(\mathbf{w}) = \mathbb{E}[\sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{I}(Y_j = 1, \delta_{\text{OR}}^{w_j} = 1)]$ as the expected number of true positives. Denote the structure-adaptive weights utilized in the SCQ procedure as \mathbf{w}_s . The next theorem establishes the efficiency gain by weighting.

Theorem 4. Consider the oracle rules $\delta_{\text{OR}}^{\mathbf{w}_s}$ and $\delta_{\text{OR}}^{\mathbf{1}}$. Suppose that the structure-adaptive weights \mathbf{w}_s are swap invariant (cf. (4)). Under Assumptions 3–4 and the exchangeability condition (9), if the score function $s(\cdot)$ utilized in (2) is continuous and satisfies (1), then we have $\Psi_{\text{OR}}(\mathbf{w}_s) \geq \Psi_{\text{OR}}(\mathbf{1})$.

2. Power Analysis for the Data-Driven Setup. Let $\delta_{\text{DD}}^{\mathbf{w}} = (\delta_{\text{DD}}^{w_j} : j \in \mathcal{D}^{\text{test}}) \in \{0, 1\}^m$ denote the data-driven rule applying the BC algorithm with weights \mathbf{w} and threshold $\tau \equiv \tau_{\text{DD}}^{\alpha}(\mathbf{w})$ from (13), where $\delta_{\text{DD}}^{w_j} = \mathbb{I}(p_j/w_j \leq \tau, p_j/w_j < \tilde{p}_j/w_j)$. Define $\Psi_{\text{DD}}(\mathbf{w}) = \mathbb{E}[\sum_{j \in \mathcal{D}^{\text{test}}} \mathbb{I}(Y_j = 1, \delta_{\text{DD}}^{w_j} = 1)]$ as the expected number of true positives of the data-driven BC method with weights \mathbf{w} . Consider $t_{\text{OR}}^{\alpha}(\mathbf{w})$ in (15). The following theorem shows that for generic weights \mathbf{w} , the data-driven BC algorithm δ_{DD} asymptotically achieves the power performance of δ_{OR} .

Theorem 5. Suppose the generic weighted p -values $\{(p_j/w_j, \tilde{p}_j/w_j) : j \in \mathcal{D}^{\text{test}}\}$ satisfy pairwise exchangeability (8). Under Assumptions 1–2, and the conditions in Lemma 1, we have $\lim_{m \rightarrow \infty} \frac{\Psi_{\text{DD}}(\mathbf{w})}{\Psi_{\text{OR}}(\mathbf{w})} = 1$.

The next theorem shows the benefits of informative weighting in data-driven setups.

Theorem 6. Consider two sets of conformity scores employing structure-adaptive weights \mathbf{w}_s and constant weights $\mathbf{1}$, respectively. If both sets of scores satisfy condition (8) and $s(\cdot)$ utilized in (2) is continuous and satisfies (1), then under Condition (9), Assumptions 1–4, and conditions in Lemma 1, we have $\lim_{m \rightarrow \infty} \frac{\Psi_{\text{DD}}(\mathbf{w}_s)}{\Psi_{\text{DD}}(\mathbf{1})} \geq 1$.

5 Simulation Studies

This section investigates the numerical performance of our proposed methods for structured OOD testing. Section 5.1 outlines the basic setup, while Sections 5.2 through 5.4 assess the effects of feature dimensionality, distribution shifts, and data imbalance across different methods. Additional numerical experiments are provided in Appendices D.2–D.4.

5.1 Simulation setup

We generate test samples $X_j \in \mathbb{R}^p$ from the following hierarchical model:

$$\left(Y_j = 1 \mid S_j \right) \stackrel{\text{ind.}}{\sim} \text{Ber}\left(\pi(S_j)\right), \quad \left(X_j \mid Y_j, S_j \right) \stackrel{\text{ind.}}{\sim} (1 - Y_j) \cdot \mathcal{N}(\mathbf{0}, I_p) + Y_j \cdot F_{1S_j}. \quad (20)$$

Here, $\pi(S_j)$ denotes the local sparsity level, $\mathbf{c} \in \mathbb{R}^p$ is a constant vector with all components equal to c , and I_p denotes a $p \times p$ diagonal matrix. We fix the test sample size at $m = 3000$ and generate 5000 null samples from $\mathcal{N}(\mathbf{0}, I_p)$. These null samples are randomly split into training, calibration, and mirror subsets, with $|\mathcal{D}_0^{\text{tr}}| = |\mathcal{D}_0^{\text{cal}}| = 1000$ and $|\mathcal{D}_0^{\text{mir}}| = 3000$. The auxiliary variable is set as $S_j = j$, which may represent, for example, the time points or spatial locations at which the samples are collected. Accordingly, samples with close or adjacent values in the indices (e.g., samples collected at nearby time points or locations) tend to exhibit similar patterns.

The structural heterogeneity is characterized by varying sparsity levels, with elevated signal frequencies in the following intervals:

$$\pi(S_j) = 0.6, \quad S_j \in [201, 300] \cup [601, 700]; \quad \pi(S_j) = 0.9, \quad S_j \in [1000, 1100] \cup [1400, 1500],$$

and $\pi(S_j) = 0.01$ for all other locations. The alternative distribution F_{1S_j} is defined as

$$F_{1S_j} = \mathcal{N}(\mu, I_p), \quad S_j \in [1, 1500]; \quad F_{1S_j} = \mathcal{N}(-2, 0.5^2 \cdot I_p), \quad S_j \in [1501, 3000].$$

The weights $w(S_j)$ employed in the SCQ procedure are constructed based on the estimated $\pi(S_j)$ and are specifically designed to satisfy Condition (4); we provide detailed description on weight construction in Appendix B.1. In all experiments, the FDR level is set to 0.05. The efficiency is evaluated using the average power $\text{AP} := \mathbb{E} \left[\frac{|\mathcal{R} \cap \mathcal{H}_1|}{\max\{1, |\mathcal{H}_1|\}} \right]$. Both the FDR and AP are reported by empirically averaging the results over 500 replications.

We compare the following methods in our experiments; implementation details of each method are provided in Appendix D.1:

1. **cfBH-OCC** and **ICP-OCC** apply the Storey-BH procedure to the p-values (constructed via the split conformal method, Bates et al., 2023) and the integrative conformal p-values (ICP; Liang et al., 2024b), employing score functions pre-trained using one-class support vector machines (SVM).
2. **AdaDetect-KDE** and **CLAW-KDE** apply the Storey-AdaDetect (Marandon et al., 2024) and semi-supervised CLAW (Zhao and Sun, 2025a) procedures, respectively, employing score functions trained via kernel density estimation (KDE) methods.
3. **AdaDetect-PU** and **CLAW-PU** apply the Storey-AdaDetect and semi-supervised CLAW procedures, respectively, with score functions trained via random forests (RF) based PU-learning methods.
4. **SCQ-OCC** and **SCQ-BIC** utilize the SCQ procedure (Algorithm A.2) with score functions pre-trained using one-class SVM and RF, respectively.
5. **SCQ-KDE** and **SCQ-PU** apply the SCQ procedure, employing score functions trained via KDE and PU-learning methods, respectively.
6. **SCQ+P-TAMS** implements SCQ with a model selected by P-TAMS from a toolbox of ML models, which will be specified later.

5.2 Impacts of feature dimensionality

This section investigates how feature dimensionality affects the performance of the following methods: AdaDetect-KDE, AdaDetect-PU, CLAW-KDE, CLAW-PU, and four SCQ variants: SCQ-OCC, SCQ-KDE, SCQ-PU, and SCQ+P-TAMS. We report the AP and FDR as functions of μ for three settings ($p = 5$, $p = 10$, and $p = 200$), which are summarized in the left, middle, and right columns of Figure 2. The following observations can be made. First, all methods successfully control the FDR below 0.05. Second, both SCQ-KDE and SCQ-PU demonstrate noticeable power improvements over their AdaDetect counterparts by effectively leveraging structural information. Third, in the low-dimensional setting ($p = 5$), KDE-based methods achieve higher AP than PU-learning-based approaches, with SCQ-KDE attaining the best overall performance. As the dimensionality grows, however, AdaDetect-PU and CLAW-PU begin to outperform their KDE-based counterparts. This is because kernel density estimation methods become increasingly unstable in high-dimensional settings. In this regime, SCQ-PU achieves the highest power. Fourth, in the high-dimensional case ($p = 200$), KDE-based methods are at a clear disadvantage, whereas SCQ-OCC consistently delivers the strongest performance among all SCQ variants. Finally, across all settings, P-TAMS successfully identifies the best-performing SCQ variant, resulting in the uniformly superior performance of SCQ+P-TAMS.

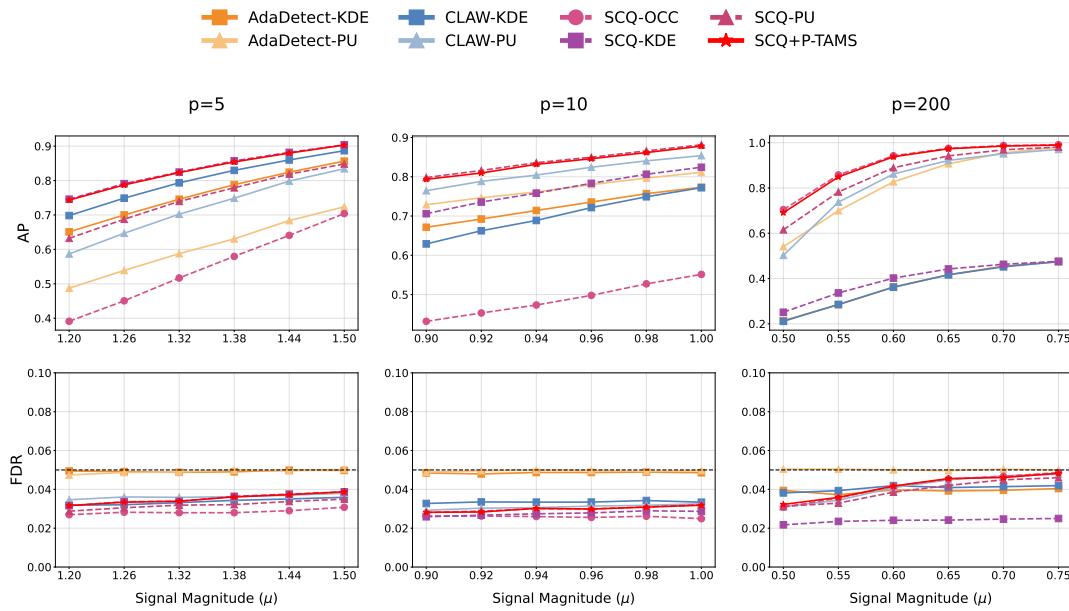


Figure 2: AP and FDR comparison of the SCQ variants, together with P-TAMS that selects among them, against density-ratio-based methods at $\alpha = 0.05$.

5.3 Performance under distribution shifts

To examine the effect of distribution shifts between labeled and test outliers, we evaluate the performance of SCQ-OCC, SCQ-BIC, cfBH-OCC, and ICP-OCC. Note that ICP (Liang et al., 2024b), which also uses a p-value weighting strategy, differs fundamentally from SCQ and may be adversely affected by distributional shifts. To ensure a fair comparison,

both SCQ and ICP are implemented using a fixed classifier (OneClassSVM). Additional comparisons regarding AMS strategies are provided in Appendix D.3.

Our experiments consider three scenarios, with the number of labeled outliers n_1 set to 60, 600, and 6000. To induce distribution shifts, half of the labeled outliers follow $\mathcal{N}(-\mathbf{2}, 0.5^2 \cdot I_p)$, matching the test outlier distribution, while the other half follow $\mathcal{N}(\mathbf{d}, I_p)$, differing from the test distribution. With μ fixed at 0.5 and p set to 200, we vary d from 0.5 (indicating no shift) to 2.5 to represent increasing degrees of distribution shift.

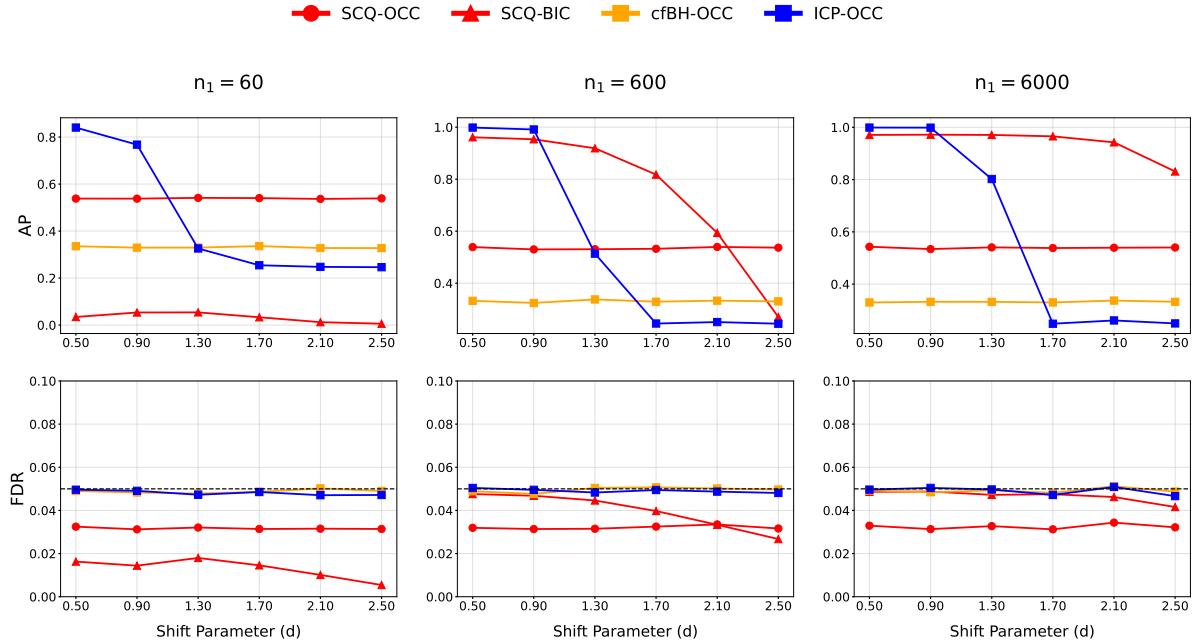


Figure 3: Comparison of AP and FDR between SCQ, cfBH, and ICP at $\alpha = 0.05$.

The results, shown in Figure 3, reveal several noteworthy patterns. First, when the distribution shift is absent or minimal, ICP-OCC performs well, effectively leveraging the labeled outliers to construct informative weights. However, its performance deteriorates progressively as the degree of shift increases, eventually falling below that of the unweighted cfBH-OCC. Second, unlike ICP – which may suffer from negative learning – SCQ-OCC consistently outperforms cfBH-OCC across all settings. Third, the flexibility of the SCQ procedure allows it to incorporate labeled outliers by employing a pre-trained BIC. In contrast to ICP, SCQ-BIC demonstrates robust performance and, in most cases, achieves the highest AP, even when the degree of distribution shift is substantial. Finally, the performance of SCQ-BIC relative to SCQ-OCC varies, with the former sometimes outperforming or underperforming the latter. This observation motivates the use of P-TAMS to uniformly enhance both variants of SCQ, as illustrated in the next subsection.

5.4 Performance under data imbalance

We generate data from a multivariate Gaussian mixture model (see Appendix D.1 for details) following the approach of Bates et al. (2023) and Liang et al. (2024b) to illustrate the selection dilemma under data imbalance and to demonstrate the effectiveness of P-TAMS. In our experiments, we set $\mu = 0.5$ and $p = 150$, generate 9000 labeled inliers, and

vary the number of labeled outliers from 0 to 3000 to examine the impact of data imbalance. We employ two OCC-based score functions (one-class SVM with sigmoid and polynomial kernels, denoted by OCC-SVM.S and OCC-SVM.P, respectively) and two BIC-based score functions (k -nearest neighbors and multi-layer perceptron, denoted by BIC-KNN and BIC-MLP, respectively) to implement the SCQ procedure.

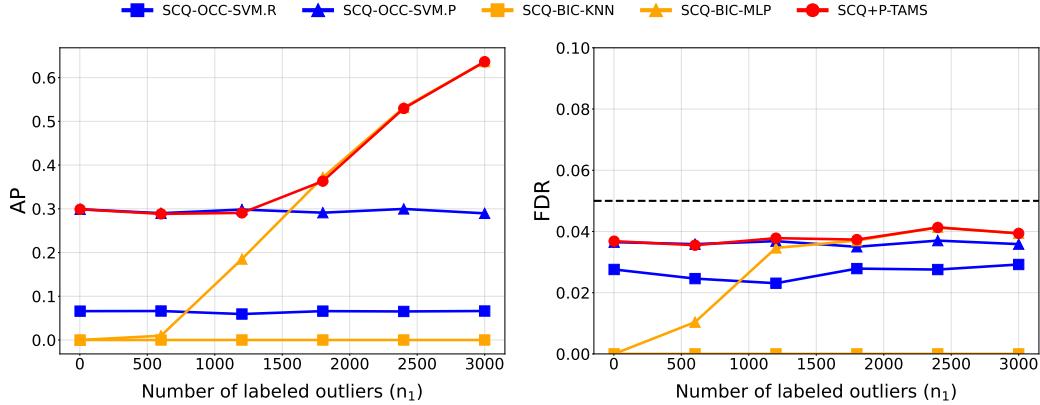


Figure 4: Comparison of AP and FDR for SCQ variants (implemented with two OCC-based and two BIC-based score functions) and P-TAMS, which uniformly enhances SCQ.

Several trends can be observed in the results presented in Figure 4. First, the SCQ-OCC variants, which rely solely on inlier samples, offer enhanced stability but fail to leverage the information provided by labeled outliers. Second, the SCQ-BIC variants incorporate labeled outliers, but may become unstable and underperform the SCQ-OCC variants when the data are highly imbalanced. Finally, SCQ+P-TAMS adaptively switches between the OCC and BIC approaches to maximize detection power, thereby providing an effective solution to the model selection dilemma. Across all settings, P-TAMS consistently identifies the best-performing SCQ variant, leading to uniformly improved performance compared to either the OCC or BIC variants of SCQ.

6 Experiments with real data

6.1 Analysis of the cybersecurity data

This section evaluates the performance of the proposed SCQ procedure on the CICIDS2017 dataset (Sharafaldin et al., 2018). The dataset, provided by the Canadian Institute for Cybersecurity (CIC), consists of normalized and deduplicated network traffic data with 80 features, encompassing both benign traffic and multiple attack categories. From this data, we construct a test set $\mathcal{D}^{\text{test}}$ of size m , consisting of m_0 benign (inlier) samples and $m_1 = 1000$ attack (outlier) samples. The latter includes three distinct attack types: distributed denial-of-service, port scan, and botnet. We consider settings where only labeled inliers are available. A null set \mathcal{D}_0 of 30,000 samples is constructed and partitioned into three parts: a mirror dataset of size m , while the remaining samples are divided equally into a training set and a calibration set.

Our objective is to efficiently identify network intrusions (outliers) by leveraging available side information to improve detection power and resource allocation. Specifically, we

use the timestamp of each sample as the side information variable S_j . This variable naturally defines 9 groups (corresponding to 9 working hours from 9 AM to 5 PM) and informs the construction of structure-aware weights (cf. Appendix B.1 of the Supplement). As illustrated in the right column of Figure 5, attack events are highly non-uniform over time, with pronounced peaks during specific hours. This temporal pattern provides valuable contextual information about the likelihood of anomalous activity. We demonstrate that this informative pattern can be exploited by SCQ to prioritize high-risk periods for anomaly detection, thereby enhancing the overall effectiveness of OOD testing.

We compare the performance of different methods, including AdaDetect-KDE, AdaDetect-PU-RF, CLAW-KDE, CLAW-PU-RF, SCQ-PU-RF, and SCQ+P-TAMS, for OOD testing at a target FDR level of 0.05. SCQ+P-TAMS selects the optimal classifier from a toolbox comprising SCQ-PU-RF, SCQ-KDE, and SCQ-OCC-SVM. The number of test outliers m_1 is fixed at 1000, while the number of test inliers m_0 is varied to create different sparsity settings. Each method is applied to 200 independent datasets with $\mathcal{D}^{\text{test}}$ and \mathcal{D}_0 being randomly sampled from the original dataset. The FDR and the expected number of true positives (ETP) are computed by averaging results over these 200 replications. Results are presented for both dense setting (with m_0 between 8000 and 10000) and sparse setting (with m_0 between 24000 and 26000), shown in the left and right panels of Figure 5, respectively.

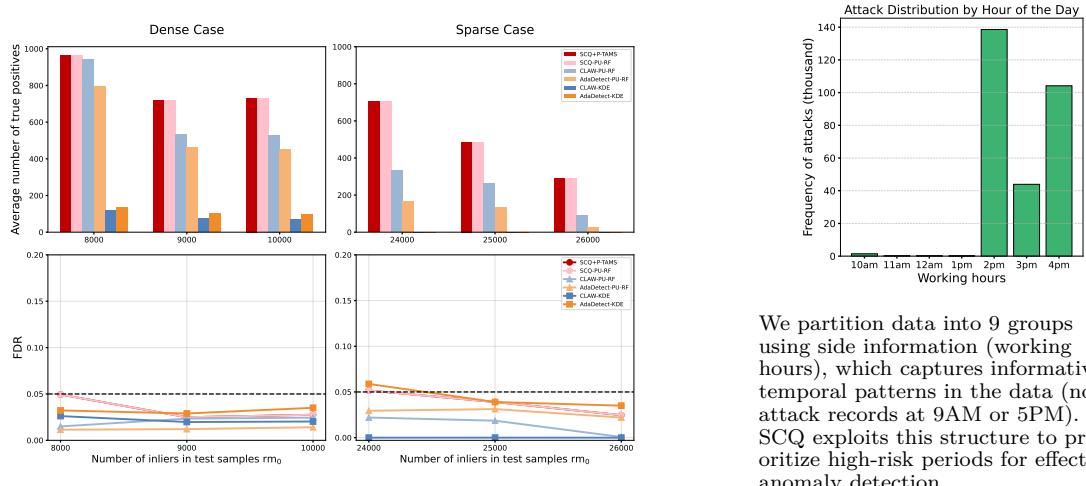


Figure 5: Left two columns: FDR and ETP of different methods; right column: hourly distribution of attack samples in the test set.

The comparisons reveal several important patterns: (1) leveraging side information (temporal patterns) significantly improves performance of OOD testing ($\text{SCQ-PU-RF} > \text{AdaDetect-PU-RF}$), with the efficiency gain being especially pronounced in sparse settings; (2) PU methods outperform KDE methods in high-dimensional settings; and (3) P-TAMS consistently selects the SCQ-PU-RF variant, and the resulting SCQ+P-TAMS method achieves the best overall performance.

6.2 Analysis of the PageBlocks data

This section analyzes the normalized and deduplicated PageBlocks dataset (Campos et al., 2016). The dataset comprises 5,139 document blocks, each described by 10 layout-based

We partition data into 9 groups using side information (working hours), which captures informative temporal patterns in the data (no attack records at 9AM or 5PM). SCQ exploits this structure to prioritize high-risk periods for effective anomaly detection.

features, including 4,883 text blocks (inliers) and 256 non-text blocks (outliers) such as images, graphics, or separator lines. Accurately identifying these non-text elements is practically significant in several document analysis and digitization tasks. For instance, in automated document parsing and information extraction, distinguishing text from non-text regions enables optical character recognition systems to focus processing resources appropriately, improving both efficiency and accuracy.

In Section 6.1, the performance of SCQ+P-TAMS coincides with the same model – SCQ-PU-RF – across all settings; however, there are situations where P-TAMS may select different models according to different situations, as we shall see shortly in analysis of the PageBlocks data. We apply SCQ with four candidate classifiers – two OCCs (LOF and GMM) and two BICs (KNN and MLP), and utilizes P-TAMS to select the best classifier among these SCQ variants (denoted SCQ+P-TAMS).

We randomly select 600 inliers and 150 outliers to form the test data with three distinct subsets: $\mathcal{D}_1^{\text{test}}$ (120 outliers and 200 inliers), $\mathcal{D}_2^{\text{test}}$ (25 outliers and 200 inliers), and $\mathcal{D}_3^{\text{test}}$ (5 outliers and 200 inliers). The side information $S_j \in \{1, 2, 3\}$ is set according to the group membership. This reflects that in some books or sources we have more non-text elements and in other tasks the proportion is lower. The remaining data are used as the labeled data with the number of labeled inliers fixed. We report the empirical ETP and FDR over 500 independent repetitions with randomly sampled data as done in the previous subsection. We vary the number of outliers and summarize the results in Figure 6.

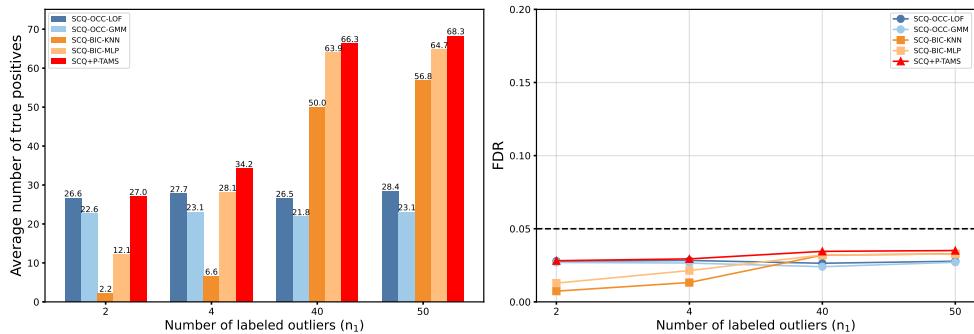


Figure 6: The FDR and ETP levels of four SCQ variants and SCQ+P-TAMS.

The following patterns can be observed. (1) OCC methods outperform BIC methods when the number of outliers is small. (2) Conversely, BIC methods are superior to OCC methods when the number of outliers is large. (3) The P-TAMS procedure consistently achieves the best overall performance. Notably, the optimal classifier can vary across replications for a fixed n_1 . P-TAMS adaptively switches among candidate models, identifying the optimal one for each dataset, which yields superior average performance compared to any pre-specified classifier.

References

- Ery Arias-Castro and Meng Wang. Distribution-free tests for sparse heterogeneous mixtures. *Test*, 26: 71–94, 2017.
- Tian Bai and Ying Jin. Optimized conformal selection: Powerful selective inference after conformity score optimization. *arXiv preprint arXiv:2411.17983*, 2024.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Yajie Bao, Yuyang Huo, Haojie Ren, and Changliang Zou. Selective conformal inference with false coverage-statement rate control. *Biometrika*, 111(3):727–742, 2024.

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055 – 2085, 2015.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine learning*, 109(4):719–760, 2020.

Yoav Benjamini and Ruth Heller. False discovery rates for spatial signals. *Journal of the American Statistical Association*, 102(480):1272–1281, 2007.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

Yoav Benjamini and Daniel Yekutieli. False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

T Cai, Wenguang Sun, and Weinan Wang. Covariate-assisted ranking and screening for large-scale two-sample inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):187–234, 2019.

T. Tony Cai and Wenguang Sun. Optimal screening and discovery of sparse signals with applications to multistage high throughput studies. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(1):197–223, 2017.

T. Tony Cai, Jiashun Jin, and Mark G. Low. Estimation and confidence sets for sparse normal mixtures. *The Annals of Statistics*, 35(6):2421–2449, 2007.

T Tony Cai, Wenguang Sun, and Yin Xia. Laws: A locally adaptive weighting and screening approach to spatial multiple testing. *Journal of the American Statistical Association*, 117(539):1370–1383, 2022.

Guilherme O Campos, Arthur Zimek, Jörg Sander, Ricardo JGB Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data mining and knowledge discovery*, 30:891–927, 2016.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

Lilun Du, Xu Guo, Wenguang Sun, and Changliang Zou. False discovery rate control under general dependence by symmetrized data aggregation. *Journal of the American Statistical Association*, 118 (541):607–621, 2023.

Marthinus C Du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems*, 27, 2014.

Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in neural information processing systems*, 35:22380–22395, 2022.

William Fithian and Lihua Lei. Conditional calibration for false discovery rate control under dependence. *The Annals of Statistics*, 50(6):3091–3118, 2022.

Ulysse Gazin, Ruth Heller, Ariane Marandon, and Etienne Roquain. Selecting informative conformal prediction sets with false coverage rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae120, 2025.

Christopher R Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

Yu Gui, Ying Jin, Yash Nair, and Zhimei Ren. Acs: An interactive framework for conformal selection. *arXiv preprint arXiv:2507.15825*, 2025.

Matan Haroush, Tzviel Frostig, Ruth Heller, and Daniel Soudry. A statistical framework for efficient out-of-distribution detection in deep neural networks. *arXiv preprint arXiv:2102.12967*, 2021.

James X Hu, Hongyu Zhao, and Harrison H Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491):1215–1227, 2010.

Nikolaos Ignatiadis and Wolfgang Huber. Covariate powered cross-weighted multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):720–751, 2021.

Ying Jin and Zhimei Ren. Confidence on the focal: Conformal prediction with selection-conditional coverage. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf016, 2025.

Shehroz S Khan and Michael G Madden. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374, 2014.

Roshan Kumari and Saurabh Kr Srivastava. Machine learning: A review on binary classification. *International Journal of Computer Applications*, 160(7), 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

Lihua Lei and William Fithian. Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):649–679, 2018.

Dennis Leung and Wenguang Sun. Zap: Z z-value adaptive procedures for false discovery rate control with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(5):1886–1946, 2022.

Ang Li and Rina Foygel Barber. Multiple testing with the structure-adaptive benjamini–hochberg algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(1):45–74, 2019.

Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after efficiency-oriented model selection. *arXiv e-prints*, pages arXiv-2408, 2024a.

Ziyi Liang, Yanfei Zhou, and Matteo Sesia. Conformal inference is (almost) free for neural networks trained with early stopping. In *International Conference on Machine Learning*, pages 20810–20851. PMLR, 2023.

Ziyi Liang, Matteo Sesia, and Wenguang Sun. Integrative conformal p-values for out-of-distribution testing with labelled outliers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad138, 2024b.

Ziyi Liang, T. Tony Cai, Wenguang Sun, and Yin Xia. A locally adaptive algorithm for multiple testing with network structure. *Statistica Sinica*, to appear, 2025. ISSN 1017-0405. doi: 10.5705/ss.202024.0002.

Chiara G Magnani, Matteo Sesia, and Aldo Solari. Collective outlier detection and enumeration with conformalized closed testing. *arXiv preprint arXiv:2308.05534*, 2023.

Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157–183, 2024.

David Mary and Etienne Roquain. Semi-supervised multiple testing. *Electronic Journal of Statistics*, 16(2):4926–4981, 2022.

Nicolai Meinshausen and John Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics*, 34(1):373–393, 2006.

Mary M Moya, Mark W Koch, and Larry D Hostetler. One-class classifier networks for target recognition applications. Technical report, Sandia National Labs., Albuquerque, NM (United States), 1993.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014. ISSN 0165-1684. doi: <https://doi.org/10.1016/j.sigpro.2013.12.026>.

Zhimei Ren and Emmanuel Candès. Knockoffs with side information. *The Annals of Applied Statistics*, 17(2):1152–1174, 2023.

Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3379–3388, 2018.

Iman Sharafaldin, Arash Habibi Lashkari, and Ali A Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*, pages 108–116. SciTePress, 2018.

John D Storey, Jonathan E Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):187–205, 2004.

David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

Wenguang Sun and T Cai. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):393–424, 2009.

Wenguang Sun, Brian J Reich, T Tony Cai, Michele Guindani, and Armin Schwartzman. False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1):59–83, 2015.

L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. *Novelty detection for the identification of masses in mammograms*, pages 442–447. Fourth International Conference on ‘Artificial Neural Networks’ (Conf. Publ. No.409), 1995. doi: 10.1049/cp:19950597.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML ’99, page 444–453, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022.

Chiao-Yu Yang, Lihua Lei, Nhat Ho, and Will Fithian. Bonus: Multiple multivariate testing with a data-adaptivetest statistic. *arXiv preprint arXiv:2106.15743*, 2021.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *Int. J. Comput. Vision*, 132(12):5635–5662, June 2024. ISSN 0920-5691. doi: 10.1007/s11263-024-02117-4.

Yachong Yang and Arun Kumar Kuchibhotla. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 5, 2021.

Yifan Zhang, Haiyan Jiang, Haojie Ren, Changliang Zou, and Dejing Dou. Automs: automatic model selection for novelty detection with error rate control. *Advances in Neural Information Processing Systems*, 35:19917–19929, 2022.

Yue Zhao, Ryan Rossi, and Leman Akoglu. Automatic unsupervised outlier model selection. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4489–4502. Curran Associates, Inc., 2021.

Zinan Zhao and Wenguang Sun. A conformalized empirical bayes method for multiple testing with side information. *arXiv preprint arXiv:2502.19667*, 2025a.

Zinan Zhao and Wenguang Sun. False discovery rate control for structured multiple testing: Asymmetric rules and conformal q-values. *Journal of the American Statistical Association*, 120(550):805–817, 2025b.