Identifying Labeling Mechanism in Positive–Unlabeled Learning under Unknown Class Prior

Siying He[*1] **Weijuan Liang**[*2] **Jiatong Liu**[†3]

[1]Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, Xiamen, China
[2]School of Mathematical Science, Xiamen University, Xiamen, China
[3]Paula and Gregory Chow Institute for Studies in Economics, Xiamen University, Xiamen, China

Abstract

Positive–Unlabeled (PU) learning critically depends on assumptions about how positive instances are labeled, yet the labeling mechanism is typically unknown in practice. Misspecification of the labeling mechanism can therefore lead to systematic bias and unstable learning. We formulate labeling mechanism identification as a statistical inference problem under an unknown class prior using only positive and unlabeled data. We propose a two-stage testing framework that first constructs a family of empirical positive sets through false discovery rate (FDR)-controlled multiple testing at multiple FDR levels, and then performs bootstrap-based selected completely at random (SCAR) consistency tests on each empirical positive set, whose evidence is aggregated via a Bonferroni correction to produce a global decision. We establish that each empirical positive set admits provably controlled contamination, and further show that the Bonferroni-aggregated bootstrap test provides valid inference. Experiments on synthetic and real-world datasets demonstrate that the proposed framework reliably distinguishes SCAR from selected at random (SAR) mechanisms and improves the robustness of downstream PU learning.

1 INTRODUCTION

In the Positive–Unlabeled (PU) learning setting, the training data consist of a set of labeled positive instances together with a large collection of unlabeled samples that contain both positive and negative cases. A central challenge in PU learning lies in identifying the unknown labeling mechanism that governs how positive instances are selected for labeling [Bekker and Davis, 2020]. Most existing PU learning methods implicitly assume a specific labeling mechanism when deriving their learning objectives [Song and Raskutti, 2020], yet this assumption is rarely verifiable in practice. When the assumed mechanism is misspecified, the resulting estimators may be biased and downstream training can be unstable [Teisseyre et al., 2024].

This naturally raises the problem of labeling mechanism identification: determining which labeling mechanism is compatible with the observed PU data. The challenge is further compounded by the fact that the class prior—the proportion of positive instances in the population—is typically unknown. As a result, reliable PU learning requires not only robust modeling under a given mechanism, but also principled tools for distinguishing between competing labeling assumptions under unknown class prior.

Two labeling mechanisms dominate the PU learning literature. Under the Selected Completely at Random (SCAR) assumption, each positive instance is labeled independently with a constant probability, implying that labeled positives form an unbiased sample of the positive class[Zhao et al., 2022, Chen et al., 2020, Li et al., 2022]. In contrast, the Selected at Random (SAR) assumption allows labeling probabilities to depend on instance-specific features, capturing more realistic and heterogeneous labeling behaviors[Gerych et al., 2022b]. In this work, we focus on a broad and practically relevant subclass of SAR mechanisms satisfying the Probabilistic Gap PU (PGPU) assumption [Gerych et al., 2022a]. The PGPU assumption posits that positive instances are labeled in a biased manner such that the labeling probability increases with the probabilistic gap $\Delta\Pr(x)$,

$$\Delta\Pr(x) = P(Y = 1 \mid X = x) - P(Y = 0 \mid X = x),$$

implying that positive instances that are less distinguishable from negative ones are less likely to be labeled. This is widely adopted to restore identifiability under SAR.

We formulate labeling mechanism identification as a statis-

[*]Equal contribution.
[†]Corresponding author: liujiatong220@stu.xmu.edu.cn.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:2119–2136.

tical inference problem that serves as a preprocessing step for PU learning. Rather than committing to a specific labeling mechanism a priori, we test whether the observed PU data are statistically consistent with the SCAR assumption, using only positive and unlabeled data and without requiring knowledge of the class prior. To this end, we propose a unified two-stage hypothesis testing framework. In the first stage, we consider a collection of FDR levels $\mathcal{Q} = \{q_1, q_2, \ldots, q_K\}$ to avoid truncation bias induced by a single FDR threshold, and construct a sequence of empirical positive sets $\{\widehat{P}^{(q_k)}\}_{k=1}^{K}$ via an FDR-controlled multiple testing procedure. In the second stage, we perform a bootstrap-based SCAR consistency test for each empirical positive set $\widehat{P}^{(q_k)}$, yielding a set of p-values $\{\hat{p}^{(q_k)}\}_{k=1}^{K}$. We then aggregate these p-values using a Bonferroni correction to obtain a global decision:

$$\hat{p}_{\text{global}} = \min_{k=1,\ldots,K} \hat{p}^{(q_k)} \cdot K.$$

We reject the null hypothesis if $\hat{p}_{\text{global}}$ is smaller than the pre-specified significance level.

We establish formal guarantees for the proposed framework under mild assumptions. In particular, we show that each empirical positive set constructed in the first stage admits provably controlled contamination, and characterize how this contamination propagates to the second-stage bootstrap test. We further show that Bonferroni aggregation yields a valid global testing procedure under the SCAR mechanism. Extensive experiments on synthetic and real-world datasets demonstrate that the proposed framework reliably distinguishes SCAR from SAR mechanisms and improves the robustness of downstream PU learning.

Contributions. The main contributions of this work are as follows:

- **Mechanism identification under unknown class prior.** We study labeling mechanism identification in PU learning without requiring knowledge of the class prior, enabling applicability in realistic settings.
- **FDR-controlled empirical positive construction with robustness guarantees.** We consider a family of FDR thresholds and construct corresponding empirical positive sets via a multiple testing procedure. We further establish contamination control for each empirical positive set and show how it propagates to downstream inference.
- **Bootstrap-based SCAR testing with family-wise aggregation.** We propose a bootstrap-based SCAR consistency test applied to each empirical positive set and aggregate the p-values via Bonferroni correction, yielding a valid global test.
- **Weaker identification assumption.** We replace the strong invariance-of-order assumption with a weaker condition requiring only a local deviation of positives from the low-score background, allowing overlap between positive and negative distributions.

2 RELATED WORK

2.1 PU LEARNING UNDER SCAR AND SAR

A large body of work on PU learning have focused on designing learning algorithms under a fixed labeling mechanism assumption. Under the SCAR assumption, the labeling propensity for positive instances is constant, which leads to particularly simple and computationally efficient PU learning formulations [Bekker and Davis, 2020]. Early approaches exploit this structure by estimating the labeling frequency and using it to scale posterior probabilities obtained from naive classifiers [Bekker and Davis, 2018, Jain et al., 2016, Ramaswamy et al., 2016, Lazecka et al., 2021] or to reweight empirical risk functions accordingly [Elkan and Noto, 2008]. This perspective has given rise to a broad class of risk reweighting methods [Yang et al., 2005, Teisseyre et al., 2020] and unbiased or non-negative risk estimators [Kiryo et al., 2017], [Wu et al., 2021].

To model more realistic feature-dependent labeling processes, subsequent work has studied PU learning under SAR assumption. Most existing SAR-based PU learning methods follow an alternating estimation strategy involving two coupled models: a classifier for the latent class posterior and a propensity model for the labeling mechanism. Representative examples include EM-based approaches that jointly update class posteriors and labeling probabilities through iterative steps [Bekker et al., 2019, Gong et al., 2021], as well as methods that iteratively approximate a surrogate positive set and exploit it to estimate the propensity score [Furmańczyk et al., 2023].

Importantly, most existing PU learning algorithms presuppose a known labeling mechanism—either SCAR or SAR—and therefore do not offer a principled procedure for inferring the mechanism from data. This implicit assumption is rarely verifiable in practice, yet it plays a critical role in the design and validity of downstream PU learning methods.

2.2 LABELING MECHANISM IDENTIFICATION

Despite its practical importance, explicit identification of labeling mechanisms in PU learning has received relatively limited attention [Bekker and Davis, 2020]. To our knowledge, the only systematic attempt is the recent work of [Teisseyre et al., 2024], which tests the SCAR assumption under a known class prior. Their method constructs an empirical positive set by selecting high-labeling-score unlabeled instances and subsequently evaluates the distributional consistency

implied by SCAR. Their framework further relies on the invariance-of-order assumption [Kato et al., 2019], which requires a global ordering between positive and negative instances in the score space.

In contrast, our work removes the requirement of a known class prior and simultaneously relaxes the strong invariance-of-order assumption. We only assume a mild separation between the corresponding score distributions, allowing distributional overlap. As a result, the proposed framework enables statistically valid labeling mechanism identification under more realistic PU learning settings.

3 BACKGROUND

3.1 POSITIVE–UNLABELED LEARNING

In PU learning, each observation is described by a triple (X, Y, S), where $X \in \mathbb{R}^d$ denotes the feature vector, $Y \in \{0, 1\}$ is the latent class label, and $S \in \{0, 1\}$ is a labeling indicator. We assume that negative instances cannot be labeled as positive, i.e. $\mathbb{P}(S = 1 \mid Y = 0) = 0$. Let $c = \mathbb{P}(S = 1 \mid Y = 1)$ denote the labeling frequency and $\pi = \mathbb{P}(Y = 1)$ the class prior. Then $\mathbb{P}(S = 1) = c\pi$, which is observable from data, while c and π are typically unknown.

We adopt the single-training-set setting and assume that $\{(X_i, Y_i, S_i)\}_{i=1}^n$ are i.i.d. samples from an unknown distribution $\mathbb{P}_{X,Y,S}$, where Y_i is unobserved. The available data thus consist of $\mathcal{D} = \{(X_i, S_i)\}_{i=1}^n$. The goal of PU learning is to predict the latent label Y for a new instance X.

3.2 LABELING MECHANISMS: SCAR AND SAR

The labeling process in PU learning is characterized by the propensity score

$$e(x) := \mathbb{P}(S = 1 \mid X = x, Y = 1).$$

Selected Completely at Random (SCAR). Under the SCAR assumption, the propensity score is constant, as a result, labeled positives form an unbiased sample of the positive class and satisfy

$$P_{X|S=1} = P_{X|Y=1}.$$

Selected at Random (SAR). Under the SAR assumption, the labeling probability depends on the features, and the propensity score is non-constant. Consequently, the distribution of labeled positives differs from that of the full positive class,

$$P_{X|S=1} \neq P_{X|Y=1}.$$

Figure 1 provides an intuitive illustration of the key distinction between SCAR and SAR labeling mechanisms.

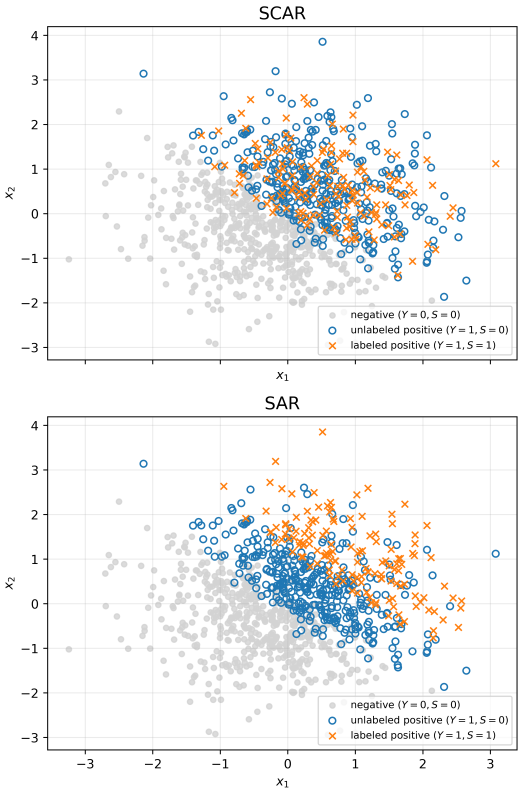


Figure 1: Illustration of SCAR and SAR labeling mechanisms. Under SCAR(top), labeled positives form an unbiased sample of the positive class, whereas under SAR (down), labeled positives exhibit a distributional shift relative to the true positive class.

Implications for mechanism identification. Distinguishing SCAR from SAR therefore amounts to testing whether the SCAR-induced distributional invariance holds in the observed PU data. The main notation used throughout the paper is summarized in Table 1.

4 TWO-STAGE HYPOTHESIS TESTING

4.1 PROBLEM FORMULATION

We consider the following hypothesis testing problem:

$$H_0 : P_{X|S=1} = P_{X|Y=1} \quad \text{(SCAR)},$$

$$H_1 : P_{X|S=1} \neq P_{X|Y=1} \quad \text{(SAR)}.$$

As established in Section 2, the SCAR assumption is characterized by the distributional equivalence, whereas SAR mechanisms violate this property. A key challenge is that both the positive instances and their total proportion are unobserved in PU data. Our approach addresses this challenge by decoupling labeling mechanism identification into two stages: (i) constructing a family of empirical positive

Table 1: Summary of notation.

Symbol	Description
x	Feature vector
Y	Latent class label (1 for positive)
S	Labeling indicator (1 for labeled)
π	Class prior $\mathbb{P}(Y=1)$
c	Labeling frequency $\mathbb{P}(S=1 \mid Y=1)$
$e(x)$	Propensity score $\mathbb{P}(S=1 \mid X=x, Y=1)$
L	Labeled positive set
U	Unlabeled set
T	Estimated labeling score $T=\hat{s}(X)$
$\widehat{\mathcal{P}}^{(q_k)}$	Empirical positive set at FDR level q_k
F_1	Score distribution of positive-but-unlabeled instances
F_0	Score distribution of negatives
q_k	Nominal FDR level q_k
ρ	Contamination rate of the empirical positive set
$K(\cdot,\cdot)$	Distributional discrepancy statistic

sets with controlled contamination at multiple FDR levels, and (ii) testing distributional consistency under the SCAR assumption via bootstrap tests.

4.2 STAGE I: FDR-CONTROLLED EMPIRICAL POSITIVE SET CONSTRUCTION

4.2.1 Scoring Unlabeled Instances

We train a probabilistic classifier to estimate the labeling score $s(x) = \mathbb{P}(S = 1 \mid X = x)$ as $\hat{s}(x)$ using the observed data $\{(X_i, S_i)\}_{i=1}^n$. The scores $\hat{s}(X_i)$ are evaluated on unlabeled instances $i \in \mathcal{U} = \{i : S_i = 0\}$ and used for subsequent inference.

4.2.2 Mixture Modeling of Unlabeled Scores

The distribution of $\{\hat{s}(X_i) : i \in \mathcal{U}\}$ is modeled as a two-component mixture corresponding to negative and positive-but-unlabeled instances. We identify the component with smaller mean as the score distribution for negatives, yielding an estimate cumulative distribution function $\widehat{F}_0$.

4.2.3 Instance-wise Hypothesis Testing and FDR Control

For each unlabeled instance $i \in \mathcal{U}$, we consider the hypothesis test,

$H_{0,i}:$ i is generated from the negative score distribution,

$H_{1,i}:$ i is not generated from the negative score distribution.

We define a plug-in p-value for each unlabeled instance $i \in \mathcal{U}$ as

$$\widehat{p}_i := 1 - \widehat{F}_0\big(\hat{s}(X_i)\big).$$

The p-value $\widehat{p}_i$ measures how extreme the estimated score is in the direction of positive instances, relative to the estimated negative score distribution. Under mild regularity conditions, these plug-in p-values are approximately super-uniform for negative instances, which enables valid false discovery rate control in the subsequent step.

We consider a collection of FDR levels $\mathcal{Q} = \{q_1, q_2, \ldots, q_K\}$, instead of relying on a single FDR threshold. As the FDR level increases, the empirical positive set progressively expands from high-confidence positives toward positive instances with lower labeling scores that are closer to the negative population. This multi-resolution construction alleviates the truncation bias associated with selecting a single FDR threshold.For each $q_k \in \mathcal{Q}$, we apply the Benjamini–Hochberg procedure at level q_k to the collection of p-values $\{\widehat{p}_i\}_{i\in\mathcal{U}}$ [Benjamini and Hochberg, 1995].Let $\widehat{\mathcal{U}}_+^{(q_k)}$ denote the set of unlabeled instances whose null hypotheses are rejected. By construction,instances in $\widehat{\mathcal{U}}_+^{(q_k)}$ exhibit labeling scores that are unlikely to be generated from the negative distribution, and are therefore treated as positive-but-unlabeled examples.

The corresponding empirical positive set is constructed as

$$\widehat{\mathcal{P}}^{(q_k)} = \mathcal{L} \cup \widehat{\mathcal{U}}_+^{(q_k)}, \qquad q_k \in \mathcal{Q}.$$

where $\mathcal{L}$ denotes the set of positively labeled instances. This procedure yields a family of empirical positive sets $\{\widehat{\mathcal{P}}^{(q_k)}\}_{q_k\in\mathcal{Q}}$, which are subsequently used in the second-stage bootstrap testing procedure.

4.3 STAGE II: BOOTSTRAP-BASED LABELING MECHANISM IDENTIFICATION

In the second stage, we test the SCAR hypothesis formulated in Section 4.1. A bootstrap-based SCAR consistency test is performed for each empirical positive set, yielding a collection of bootstrap p-values. These p-values are then aggregated using the Bonferroni correction to obtain a global decision on the labeling mechanism.

4.3.1 Bootstrap Procedure

For each empirical positive set $\widehat{\mathcal{P}}^{(q_k)}$, we perform the following bootstrap procedure. For notational simplicity, the dependence on q_k is omitted whenever there is no ambiguity.

Under the SCAR null hypothesis, labeled instances arise as an unbiased random subset of positive instances. We exploit this property to construct a bootstrap procedure that approximates the sampling distribution of a discrepancy statistic under H_0.

We first define artificial class labels

$$\tilde{Y}_i = \mathbb{I}\{i \in \widehat{\mathcal{P}}\},$$

treating all instances in the empirical positive set as positive. Let$\hat{c} = \frac{|\mathcal{L}|}{|\widehat{\mathcal{P}}|}$ denote the empirical labeling frequency. For each bootstrap replicate $b = 1, \ldots, B$, we generate artificial labeling indicators

$$\tilde{S}_i^{(b)} \sim \text{Bernoulli}(\hat{c}), \quad i \in \widehat{\mathcal{P}},$$

and set $\tilde{S}_i^{(b)} = 0$ for $i \notin \widehat{\mathcal{P}}$.Given $(\tilde{Y}, \tilde{S}^{(b)})$, we compute the bootstrap statistic

$$\widehat{K}^{(b)} = \mathcal{K}\Big(\widehat{P}_{X|\tilde{S}^{(b)}=1}, \widehat{P}_{X|\tilde{Y}=1}\Big),$$

which measures the distributional discrepancy induced by random labeling under SCAR.

4.3.2 Test Statistics

We consider several discrepancy statistics $\mathcal{K}(\cdot, \cdot)$ to capture different forms of distributional deviation [Teisseyre et al., 2024].

KL-type statistics. We approximate both $\widehat{P}_{X|\tilde{S}=1}$ and $\widehat{P}_{X|\tilde{Y}=1}$ by multivariate normal distributions. With diagonal covariance matrices, the resulting KL divergence is

$$\mathcal{K}_{\text{KL}} = \frac{1}{2}\Big[(\hat{\mu}_Y - \hat{\mu}_S)^\top \hat{\Sigma}_Y^{-1}(\hat{\mu}_Y - \hat{\mu}_S) + \text{tr}(\hat{\Sigma}_Y^{-1}\hat{\Sigma}_S) - \log\frac{\det(\hat{\Sigma}_S)}{\det(\hat{\Sigma}_Y)} - p\Big].$$

where p is the feature dimension. We also consider a fully multivariate variant $\mathcal{K}_{\text{KL}}^{\text{full}}$ based on unrestricted covariance estimates, which additionally accounts for feature correlations.

Marginal KS statistic. As a nonparametric alternative, we compute for each feature j the univariate Kolmogorov–Smirnov distance

$$D_j = \sup_t \big|\widehat{F}_{S,j}(t) - \widehat{F}_{Y,j}(t)\big|,$$

and aggregate marginal discrepancies via $\mathcal{K}_{\text{KS}} = \sum_{j=1}^p D_j$.

Classifier-based statistic. We consider a classifier-based discrepancy measure. A binary classifier is trained to distinguish samples from $\widehat{P}_{X|\tilde{S}=1}$ and $\widehat{P}_{X|\tilde{Y}=1}$, and the statistic is defined as

$$\mathcal{K}_{\text{NB}} = \text{AUC} - \tfrac{1}{2}.$$

These statistics capture complementary forms of distributional discrepancy and can be seamlessly integrated into the bootstrap procedure described in Section 4.3.1.

4.3.3 Global Decision via Bonferroni Aggregation

For each empirical positive set $\widehat{\mathcal{P}}^{(q_k)}$, let

$$\widehat{K}_{\text{obs}}^{(q_k)} = \mathcal{K}\Big(\widehat{P}_{X|S=1}, \widehat{P}_{X|\tilde{Y}=1}\Big)$$

denote the observed discrepancy statistic. The corresponding bootstrap p-value is computed as

$$\hat{p}_{\text{boot}}^{(q_k)} = \frac{1}{B}\sum_{b=1}^B \mathbb{I}\Big(\widehat{K}^{(q_k,b)} \geq \widehat{K}_{\text{obs}}^{(q_k)}\Big).$$

Applying the bootstrap procedure to every empirical positive set produces a collection of p-values

$$\{\hat{p}_{\text{boot}}^{(q_1)}, \ldots, \hat{p}_{\text{boot}}^{(q_K)}\}.$$

These p-values are aggregated using the Bonferroni correction to obtain a single global p-value,

$$\hat{p}_{\text{global}} = K \min_{1 \leq k \leq K} \hat{p}_{\text{boot}}^{(q_k)}.$$

The final decision on the labeling mechanism is made by comparing $\hat{p}_{\text{global}}$ with the pre-specified significance level.

The overall procedure of the proposed framework is summarized in Algorithm 1.

5 THEORETICAL JUSTIFICATIONS

5.1 VALIDITY OF FIRST-STAGE P-VALUES

We establish the validity of the instance-wise plug-in p-values used in Stage I. Our analysis relies on asymptotic super-uniformity under the negative class, which is sufficient for false discovery rate control via the Benjamini–Hochberg procedure. Let $T = \hat{s}(X)$ denote the labeling score. By the law of total probability, the score distribution among unlabeled instances admits the following mixture representation:

$$T \mid S = 0 \sim \beta F_0 + (1 - \beta)F_1,$$

where F_0 and F_1 denote the conditional score distributions of negative instances ($Y = 0$) and positive-but-unlabeled instances ($Y = 1, S = 0$), respectively, and

$$\beta := \mathbb{P}(Y = 0 \mid S = 0) = \frac{1 - \pi}{1 - c\pi}.$$

Assumption 5.1 (Identifiability of the negative component)**.** The mixture components F_0 and F_1 are identifiable from the unlabeled score distribution. Specifically,

$$\mu_0 := \mathbb{E}[T \mid Y = 0, S = 0] < \mu_1 := \mathbb{E}[T \mid Y = 1, S = 0],$$

with separation $\mu_1 - \mu_0 \geq \Delta$ for some $\Delta > 0$.

Algorithm 1 Two-stage Labeling Mechanism Identification

Require: PU data $\{(X_i, S_i)\}_{i=1}^n$, collection of FDR levels $\mathcal{Q} = \{q_1, \ldots, q_K\}$, significance level α
Ensure: Global decision on the labeling mechanism

Stage I: Empirical Positive Set Construction
Train a scoring model to estimate $s(x) = P(S = 1 \mid X)$
Fit a two-component mixture model to $\{s(X_i) : i \in \mathcal{U}\}$
Identify the score distribution of the negatives $\widehat{F}_0$
Compute plug-in p-values $\hat{p}_i = 1 - \widehat{F}_0(\hat{s}(X_i))$ for $i \in \mathcal{U}$
for each $q_k \in \mathcal{Q}$ **do**
    Apply the BH procedure at level q_k
    Construct empirical positive set $\widehat{\mathcal{P}}^{(q_k)}$
    Estimate labeling frequency $\hat{c}^{(q_k)}$
end for

Stage II: Bootstrap-based Mechanism Test
for each $\widehat{\mathcal{P}}^{(q_k)}$ **do**
    Define empirical labels $\tilde{Y}_i = \mathbf{1}\{i \in \widehat{\mathcal{P}}^{(q_k)}\}$
    Compute the observed statistic $\widehat{K}_{\text{obs}}^{(q_k)}$
    for $b = 1$ **to** B **do**
        Generate $\tilde{S}_i^{(b)} \sim \text{Bernoulli}(\hat{c}^{(q_k)})$ for $i \in \widehat{\mathcal{P}}^{(q_k)}$
        Set $\tilde{S}_i^{(b)} = 0$ for $i \notin \widehat{\mathcal{P}}^{(q_k)}$
        Compute $\widehat{K}^{(q_k,b)}$
    end for
    Compute bootstrap p-value $\hat{p}_{\text{boot}}^{(q_k)}$
end for
Compute the Bonferroni-adjusted global p-value

$$\hat{p}_{\text{global}} = K \min_{q_k \in \mathcal{Q}} \hat{p}_{\text{boot}}^{(q_k)}.$$

if $\hat{p}_{\text{global}}$ is below the pre-specified significance level **then**
    Reject SCAR
else
    Fail to reject SCAR
end if

Remark 5.2. The above assumption is implied by the PGPU assumption [Gerych et al., 2022a], a standard identifiability condition in PU learning. Without such an assumption, the decomposition of the unlabeled distribution is information-theoretically non-identifiable, which is an intrinsic limitation of PU learning.

Assumption 5.3 (Approximation of the negative score distribution)**.** Let $\widehat{F}_0$ denote an estimator of score distribution of negatives obtained from the unlabeled scores. With probability at least $1 - \delta_n$,

$$\sup_{t \in \mathbb{R}} \left| \widehat{F}_0(t) - F_0(t) \right| \leq \varepsilon_n,$$

where $\varepsilon_n \to 0$ and $\delta_n \to 0$ as $n \to \infty$.

Theorem 5.4 (Validity of plug-in p-values)**.** *Under Assumptions 5.1,5.3, define the plug-in p-value for each unlabeled instance $i \in U$ by*

$$\widehat{p}_i := 1 - \widehat{F}_0(T_i),$$

Then for any unlabeled instance i with $Y_i = 0$ and for any $u \in [0, 1]$,

$$P(\widehat{p}_i \leq u) \leq u + \varepsilon_n,$$

where ε_n is the same approximation error appearing in Assumption 5.3. In particular, the plug-in p-values are asymptotically super-uniform under the negative class.

Corollary 5.5 (FDR control via BH procedure)**.** *Let $\widehat{\mathcal{U}}_+ \subseteq U$ be the set of positive but unlabeled instances selected by applying the Benjamini–Hochberg procedure at nominal level q to the plug-in p-values $\{\widehat{p}_i\}_{i \in U}$. Under the conditions of Theorem 5.4, if the p-values corresponding to true null hypotheses are independent or satisfy the PRDS condition, then the resulting false discovery rate satisfies*

$$\text{FDR} := \mathbb{E}\left[\frac{|\widehat{\mathcal{U}}_+ \cap \{i : Y_i = 0\}|}{|\widehat{\mathcal{U}}_+| \vee 1} \right] \leq q + O(\varepsilon_n).$$

In particular, the false discovery rate is asymptotically controlled at level q.

5.2 FROM CONTAMINATED EMPIRICAL POSITIVES TO VALID BOOTSTRAP INFERENCE

We analyze how contamination in the empirical positive set constructed in Stage I affects the validity of the second-stage bootstrap test under the null hypothesis. Throughout this section, we assume that the SCAR labeling mechanism holds.

Let ρ_q denote the contamination rate of the empirical positive set $\widehat{\mathcal{P}}^{(q)}$ constructed at FDR level q. Therefore we have $\rho_q \leq q + O(\varepsilon_n)$, where ε_n captures estimation error in the score model.

Assumption 5.6 (Stability of the discrepancy statistic)**.** The discrepancy statistic $\mathcal{K}(\cdot, \cdot)$ is stable under one-sided contamination in the sense that

$$\mathcal{K}\big((1 - \rho_q)P_+ + \rho_q P_-,\, P_+\big) \leq \rho_q, \qquad \forall \rho_q \in [0, 1].$$

Theorem 5.7 (Validity of the bootstrap test under contamination)**.** *Let c_α be the $(1 - \alpha)$-quantile of the bootstrap distribution. Then under the SCAR hypothesis,*

$$\mathbb{P}_{H_0}\big(\textit{reject at FDR level } q\big) \leq \alpha + O(q + \varepsilon_n).$$

Theorem 5.8 (Validity of the Bonferroni-aggregated test)**.** *Let $\hat{p}_{\text{global}}$ denote the Bonferroni-adjusted p-value obtained*

by aggregating the bootstrap p-values from Stage II. Then, under the SCAR hypothesis,

$$\mathbb{P}_{H_0}(\hat{p}_{\text{global}} \leq \alpha) \leq \alpha + \sum_{k=1}^{K} O(q_k + \varepsilon_n).$$

Proof. The proof and additional technical discussions are provided in Appendix.

5.3 COMPUTATIONAL COMPLEXITY

The computational cost is dominated by two components: (i) training the scoring model in Stage I, and (ii) the bootstrap procedure in Stage II. Specifically, the overall complexity is

$$O(\mathcal{C}_{\text{score}} + KBm)\,,$$

where $\mathcal{C}_{\text{score}}$ is the complexity of the scoring model, B is the number of bootstrap resamples, K is the number of FDR levels, and m denotes the average size of the empirical positive sets. Moreover, the bootstrap replications are independent and can be executed in parallel.

6 EXPERIMENT

6.1 EXPERIMENTAL SETUP

We evaluate the proposed framework along the following aspects:

(i) **Type-I error control under SCAR**: we evaluate whether the proposed procedure avoids falsely rejecting the SCAR assumption when it holds.

(ii) **Detection power under SAR**: we assess whether the test can detect violations of SCAR under different SAR mechanisms and strengths of dependence.

(iii) **Sensitivity to FDR levels**: we study how different FDR levels affect the construction of empirical positive sets, in terms of selection accuracy and coverage, and their impact on Stage II testing.

(iv) **Robustness analysis**: we investigate stability under varying labeling probabilities and class imbalance, as well as different choices of score distribution modeling for mixture estimation.

(v) **Scalability analysis**: we evaluate computational efficiency and statistical stability under increasing sample sizes and feature dimensions, to assess the practicality of the proposed method in high-dimensional and large-scale settings.

(vi) **Comparison with alternative methods**: we compare the proposed approach with existing SCAR mechanism identification methods, highlighting the benefits of relaxing strong invariance assumptions.

Table 2: Real-world datasets

Dataset	#Samples	#Features	Pos. Ratio (%)
Breast-w	699	9	34.5
Wdbc	569	30	37.3
Banknote Auth.	1372	4	44.5
Vote	435	16	61.4
Segment	2310	19	57.1
Cleveland Heart Dis.	303	13	46.1

(vii) **Downstream impact of mechanism identification**: we evaluate how accurate SCAR/SAR identification improves downstream PU learning performance.

Datasets. We evaluate the proposed framework on both synthetic and real-world PU datasets. In particular, we consider two synthetic datasets, ARTIF1 and ARTIF2, each consisting of n samples with p features. Feature vectors are independently generated as

$$X \sim \mathcal{N}(0, I_p),$$

where I_p denotes the p-dimensional identity matrix. Binary class labels $Y \in \{0, 1\}$ are sampled according to a probabilistic model with a controllable class prior. To assess robustness under different noise characteristics, we consider two data-generating models for the class posterior. In ARTIF1, the conditional class probability follows a logistic model. In ARTIF2, we consider a heavy-tailed Cauchy link for the conditional class probability. In addition, we evaluate our method on multiple real-world benchmark datasets commonly used in PU learning. Dataset characteristics are summarized in Table 2.

Labeling mechanisms. To simulate the PU setting, we generate labels according to several labeling mechanisms as follows:

- **S0 (SCAR):** Positive instances are labeled independently with a constant labelling probability c, i.e., $e(x) = c$.
- **S1 (SAR-1D):** The labeling probability depends on the first feature, $e(x) = \sigma(gx_1)$, where $\sigma(\cdot)$ is the sigmoid function and g controls the strength of feature dependence.
- **S2 (SAR-linear):** The labeling probability depends on all features through a linear model, $e(x) = \sigma(g\, a^\top x)$, where a is chosen to achieve a target labeling frequency.
- **S3 (SAR-nonlinear):** A nonlinear extension of S2, where the linear predictor is transformed by a higher-order function to induce stronger distributional shifts.

By varying g, we smoothly interpolate between the SCAR regime ($g = 0$) and increasingly strong SAR violations.

Table 3: Type-I error under the SCAR mechanism for different labeling probabilities c.Values exceeding the significance level α are highlighted in bold.

Dataset	c	KL	KLCOV	KS	NB
Artif1	0.1	0.006	0.000	0.000	0.008
	0.2	0.014	0.000	0.000	0.008
	0.3	0.004	0.002	0.000	0.004
Breast-w	0.1	0.000	0.000	0.000	0.000
	0.2	0.008	0.000	0.002	0.002
	0.3	0.022	0.038	0.026	0.008
Wdbc	0.1	0.002	0.000	0.000	0.000
	0.2	0.004	0.000	0.006	0.008
	0.3	0.038	0.002	**0.070**	**0.058**
Banknote Auth.	0.1	0.000	0.010	0.000	0.000
	0.2	0.000	0.048	0.000	0.002
	0.3	0.004	**0.304**	0.006	0.006

Implementation details. Throughout all experiments, we use a random forest classifier as the base model to estimate the labeling score. The second-stage test is implemented using a bootstrap procedure with $B = 300$ resamples, significance level $\alpha = 0.05$, and results averaged over 500 independent repetitions. The collection of FDR levels is fixed as $Q = \{0.001, 0.005, 0.007, 0.01\}$.

6.2 TYPE-I ERROR CONTROL UNDER SCAR

We evaluate whether the proposed two-stage testing framework controls the type-I error under the SCAR labeling mechanism. To investigate the effect of the labeling frequency, we vary the positive labeling probability over $c \in \{0.1, 0.2, 0.3\}$.

Under SCAR, varying the labeling probability c changes the composition of the unlabeled set. As c increases, more positive instances are labeled, yielding fewer hidden positives and a higher proportion of negatives among unlabeled samples. This evaluates the stability of Type-I error control under different unlabeled-set compositions.

Tables 3 report Type-I error using different discrepancy statistics on the ARTIF1 and real datasets. Overall, the results demonstrate stable Type-I error control across different labeling frequencies and unlabeled-set compositions.

6.3 POWER ANALYSIS UNDER SAR

We evaluate the power of the test for detecting violations of the SCAR assumption under SAR labeling mechanisms.

We consider the three SAR mechanisms introduced in Section 6.1: S1, S2, and S3 and vary the SAR strength parameter g. Larger values of g make the labeling propensity more strongly feature-dependent, thereby inducing stronger violations of the SCAR assumption.

Figure 2 illustrates the rejection rates on synthetic datasets as the SAR strength parameter g increases. The rejection rate increases consistently with g, indicating that the proposed test becomes increasingly sensitive as the labeling mechanism deviates further from SCAR. Although different discrepancy statistics exhibit different sensitivities, they all demonstrate the same increasing trend, confirming the effectiveness of the proposed framework for identifying violations of the SCAR assumption. Additional results on the real-world datasets are presented in Appendix B.1

6.4 POWER COMPARISON WITH ALTERNATIVE MECHANISM IDENTIFICATION METHODS UNDER SAR

We compare the proposed method with the prior method as introduced in Section 2.2 under SAR labeling mechanisms with varying class priors π and labeling probabilities c. The class prior π controls the degree of class imbalance, while c determines the proportion of labeled positive instances.

To ensure a fair comparison, we provide the prior method with access to the true class prior π. In contrast, our method does not assume knowledge of π. Table 4 reports the test power across different settings of $\pi \in \{0.2, 0.3, 0.5, 0.7\}$ and $c \in \{0.2, 0.3, 0.5\}$.

The results show that the proposed method achieves comparable or higher power than the prior method across most settings, despite not using the true class prior.In contrast, the prior method relies on a strong invariance-of-order assumption, which can be overly restrictive in practice. By relaxing this assumption and constructing accurate empirical positives, the proposed method remains more robust under distributional overlap.

In addition, we observe that the rejection rates of the KL-based statistic remain consistently high and stable across different values of π, indicating strong robustness to varying class imbalance under SAR mechanisms. As π increases, the proportion of negatives decreases, resulting in a few-negatives scenario that poses additional challenges. Nevertheless, the KL-based statistic continues to achieve stable rejection rates, whereas the other discrepancy measures exhibit noticeable variability and a mild degradation in performance. This demonstrates that the KL-based statistic is less sensitive to severe class imbalance and remains reliable even in few-negatives settings.

6.5 SENSITIVITY ANALYSIS OF FDR LEVELS

We study the effect of the nominal FDR level q on the construction of the empirical positive set in Stage I and its impact on the Stage II test. Table 5,6in appendix shows that

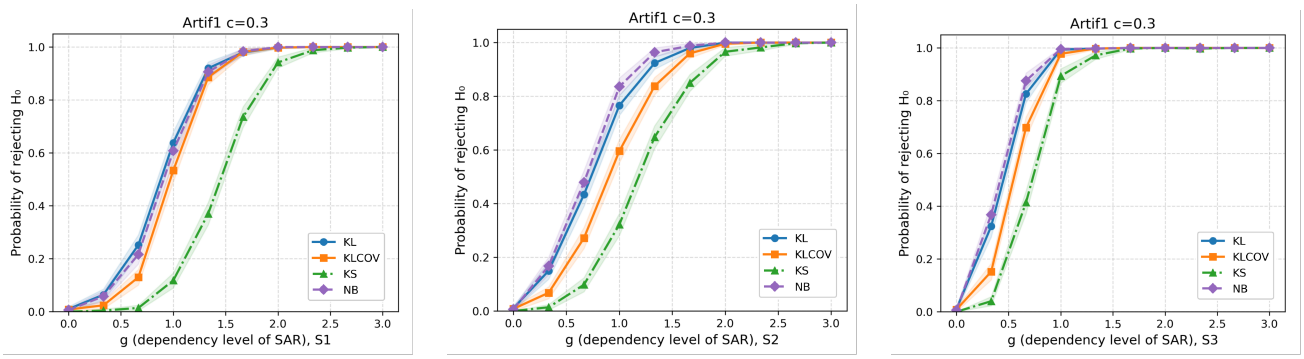


Figure 2: Power under SAR mechanisms. This shows the synthetic ARTIF1 dataset under the S1, S2, and S3 labeling mechanisms, respectively.

Table 4: Comparison of test power under SAR mechanisms between the prior baseline (known class prior) and the proposed method (unknown class prior). Results are reported as mean ± std over 50 runs.

π	c	Prior baseline (known class prior)				Proposed method			
		KL	KLCOV	KS	NB	KL	KLCOV	KS	NB
0.2	0.2	0.82±0.38	0.42±0.49	0.48±0.49	0.46±0.49	0.97±0.18	0.70±0.45	0.80±0.40	0.70±0.46
	0.3	0.92±0.27	0.72±0.45	0.72±0.45	0.76±0.43	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
	0.5	0.96±0.19	0.94±0.24	0.68±0.46	0.80±0.40	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
0.3	0.2	0.92±0.27	0.72±0.45	0.70±0.46	0.68±0.47	1.00±0.00	0.93±0.25	0.87±0.34	0.80±0.40
	0.3	0.90±0.30	0.80±0.40	0.70±0.46	0.78±0.41	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
	0.5	0.90±0.30	0.92±0.27	0.88±0.33	0.86±0.35	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
0.5	0.2	0.96±0.19	0.90±0.30	0.78±0.41	0.78±0.41	1.00±0.00	0.97±0.18	0.83±0.37	0.87±0.34
	0.3	0.92±0.27	0.84±0.37	0.76±0.43	0.82±0.38	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
	0.5	0.96±0.19	0.92±0.27	0.88±0.33	0.96±0.19	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
0.7	0.2	0.90±0.30	0.86±0.35	0.70±0.46	0.80±0.40	0.60±0.49	0.27±0.44	0.17±0.37	0.30±0.46
	0.3	0.96±0.19	0.94±0.24	0.86±0.35	0.90±0.30	0.83±0.37	0.83±0.37	0.63±0.48	0.70±0.46
	0.5	0.96±0.19	0.96±0.19	0.92±0.27	0.96±0.19	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00

increasing q enlarges the empirical positive set by including additional low-score positive instances that are excluded under more stringent thresholds. As a result, using a collection of FDR levels mitigates the truncation bias induced by any single FDR threshold, by providing empirical positive sets at multiple selection granularities.

Consistent with this observation, Figure 4 in appendix shows that the power of the test increases as the FDR level becomes larger. By incorporating lower-score positive instances, the empirical positive set provides a more representative approximation to the complete positive distribution, leading to stronger sensitivity for detecting SAR mechanisms.

7 CONCLUSION

In this work, we study the problem of labeling mechanism identification in positive–unlabeled learning under unknown class prior. We propose a two-stage statistical framework that first constructs a family of empirical positive sets via FDR-controlled multiple testing, and then performs bootstrap-based tests of the SCAR assumption with Bonferroni aggregation across multiple resolutions.

Our method is developed under the PGPU assumption, which serves as a key condition ensuring identifiability in PU learning. However, when the PGPU assumption is violated, particularly in extreme SAR regimes where unlabeled-positive and negative instances become nearly indistinguishable in the score space, the proposed method is no longer applicable and should not be expected to yield meaningful inference.

Extensive experiments on synthetic and real-world datasets demonstrate that the proposed framework consistently controls Type-I error under SCAR, achieves strong detection power under SAR, and remains robust across different class priors, labeling probabilities, and distributional overlaps. These results confirm the effectiveness and practical reliability of the proposed labeling mechanism identification framework.

Acknowledgements

This work was partially supported by the Fujian Provincial Natural Science Foundation of China under Grant 2026J008018 and the Natural Science Foundation of Xiamen, China under Grant 3502Z202671023.

References

Jessa Bekker and Jesse Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

Jessa Bekker, Pieter Robberechts, and Jesse Davis. Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 71–85, 2019.

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.

Honglei Chen, Feng Liu, Yao Wang, Lei Zhao, and Hao Wu. A variational approach for learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, pages 14844–14854, 2020.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 213–220, New York, NY, USA, 2008. Association for Computing Machinery.

Krzysztof Furmańczyk, Jan Mielniczuk, Wojciech Rejchel, and Paweł Teisseyre. Double logistic regression approach to biased positive–unlabeled data. In *Proceedings of the European Conference on Artificial Intelligence*, pages 764–771, 2023.

W. Gerych, T. Hartvigsen, L. Buquicchio, E. Agu, and E. Rundensteiner. Recovering the propensity score from biased positive unlabeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6694–6702, 2022a.

William Gerych, Thomas Hartvigsen, Luca Buquicchio, Emmanuel Agu, and Elke Rundensteiner. Recovering the propensity score from biased positive and unlabeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6694–6702, 2022b.

Chen Gong, Qian Wang, Tongliang Liu, Bo Han, Jianfeng You, Jian Yang, and Dacheng Tao. Instance-dependent positive and unlabeled learning with labeling bias estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Saurabh Jain, Martha White, and Predrag Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Advances in Neural Information Processing Systems*, pages 2693–2701, 2016.

Masahiro Kato, Tatsuya Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Magdalena Lazecka, Jan Mielniczuk, and Paweł Teisseyre. Estimating the class prior for positive and unlabelled data via logistic regression. *Advances in Data Analysis and Classification*, 15:1039–1068, 2021.

Changchun Li, Ximing Li, Lei Feng, and Jihong Ouyang. Who is your right mixup partner in positive and unlabeled learning. In *Proceedings of the International Conference on Learning Representations*, 2022.

Harish G. Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 2052–2060, 2016.

Huan Song and Garvesh Raskutti. Pulasso: High-dimensional variable selection with presence-only data. *Journal of the American Statistical Association*, 115(529): 334–347, 2020.

Paweł Teisseyre, Jan Mielniczuk, and Małgorzata Łażecka. Different strategies of fitting logistic regression for positive and unlabelled data. In *Proceedings of the International Conference on Computational Science*, 2020.

Piotr Teisseyre, Krzysztof Furmańczyk, and Jan Mielniczuk. Verifying the selected completely at random assumption in positive-unlabeled learning. *arXiv preprint arXiv:2404.00145*, 2024.

Man Wu, Shirui Pan, Lan Du, and Xingquan Zhu. Learning graph neural networks with positive and unlabeled nodes. *ACM Transactions on Knowledge Discovery from Data*, 15(6):1–25, 2021.

Xulei Yang, Qing Song, and A. Cao. Weighted support vector machine for data classification. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, pages 859–864, 2005.

Yifan Zhao, Qian Xu, Yifan Jiang, Peng Wen, and Qingming Huang. Dist-pu: Positive-unlabeled learning from a label distribution perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14461–14470, 2022.

Identifying Labeling Mechanism in Positive–Unlabeled Learning under Unknown Class Prior (Supplementary Material)

Siying He[*1] **Weijuan Liang**[*2] **Jiatong Liu**[†3]

[1]Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, Xiamen, China
[2]School of Mathematical Science, Xiamen University, Xiamen, China
[3]Paula and Gregory Chow Institute for Studies in Economics, Xiamen University, Xiamen, China

A ADDITIONAL THEORETICAL BACKGROUND

A.1 VALIDITY OF FIRST-STAGE P-VALUES

Remark A.1 (On the use of Gaussian mixture models). We use Gaussian mixture models (GMMs) as a convenient estimator for the unlabeled score distribution. Under mild regularity conditions, finite Gaussian mixtures are dense in the space of continuous distributions, so sufficiently smooth component distributions can be approximated arbitrarily well. Our analysis only requires uniform consistency of the estimated negative component $\widehat{F}_0$ in terms of the cumulative distribution function, and does not rely on the true score distributions being Gaussian.

Remark A.2 (On dependence assumptions). Approximate independence of null p-values can be enforced via sample splitting: one subset of the data is used to train the score function $\hat{s}(\cdot)$ and to estimate the negative score distribution $\widehat{F}_0$, while plug-in p-values $\widehat{p}_i = 1 - \widehat{F}_0(T_i)$ are computed on a disjoint subset with $\hat{s}$ and $\widehat{F}_0$ held fixed. Under this construction, the resulting p-values depend only on individual samples and therefore satisfy weak dependence or asymptotic independence.

Proof. Let

$$\mathcal{E}_n := \Big\{ \sup_{t\in\mathbb{R}} \big|\widehat{F}_0(t) - F_0(t)\big| \le \varepsilon_n \Big\}$$

denote the high-probability event specified in Assumption 5.3. By definition, $\mathbb{P}(\mathcal{E}_n) \ge 1 - \delta_n$ for some $\delta_n \to 0$.

We first work on the event $\mathcal{E}_n$. For any $t \in \mathbb{R}$, the definition of $\mathcal{E}_n$ implies

$$\widehat{F}_0(t) \le F_0(t) + \varepsilon_n.$$

Hence, on $\mathcal{E}_n$, we have

$$\begin{aligned} \widehat{p}_i \le u &\iff 1 - \widehat{F}_0(T_i) \le u \\ &\iff \widehat{F}_0(T_i) \ge 1 - u \\ &\Rightarrow F_0(T_i) \ge 1 - u - \varepsilon_n \\ &\iff p_i^\star := 1 - F_0(T_i) \le u + \varepsilon_n. \end{aligned}$$

where $p_i^\star := 1 - F_0(T_i)$ denotes the oracle p-value.

Since $Y_i = 0$ implies $T_i \sim F_0$, the probability integral transform yields that $p_i^\star$ is super-uniform on $[0,1]$, i.e., $\mathbb{P}(p_i^\star \le v) \le v$ for all $v \in [0,1]$. Therefore, conditioning on $\mathcal{E}_n$,

$$\mathbb{P}(\widehat{p}_i \le u \mid \mathcal{E}_n) \le \mathbb{P}(p_i^\star \le u + \varepsilon_n) \le u + \varepsilon_n.$$

Finally, removing the conditioning, we obtain

$$\mathbb{P}(\widehat{p}_i \le u) = \mathbb{P}(\widehat{p}_i \le u,\, \mathcal{E}_n) + \mathbb{P}(\widehat{p}_i \le u,\, \mathcal{E}_n^c) \le (u + \varepsilon_n) + \delta_n,$$

where $\delta_n := \mathbb{P}(\mathcal{E}_n^c)$, so that the bound can be written as $\mathbb{P}(\widehat{p}_i \le u) \le u + O(\varepsilon_n)$. □

A.2 FROM CONTAMINATED EMPIRICAL POSITIVES TO VALID BOOTSTRAP INFERENCE

We analyze the effect of contamination in the empirical positive set on the validity of the Stage II bootstrap procedure.

Let $\widehat{\mathcal{P}}^{(q)}$ denote the empirical positive set constructed at FDR level q, and define its contamination rate as

$$\rho_q := \frac{|\widehat{\mathcal{P}}^{(q)} \cap \{i : Y_i = 0\}|}{|\widehat{\mathcal{P}}^{(q)}| \vee 1}.$$

From Corollary5.5, we have

$$\rho_q \leq q + O(\varepsilon_n).$$

(i) Ideal uncontaminated case. The bootstrap procedure resamples labels within $\widehat{\mathcal{P}}_q$. Conditional on $\widehat{\mathcal{P}}_q$, both samples involved in the discrepancy statistic are drawn from the same empirical distribution. Therefore, the bootstrap statistic is centered at zero (up to sampling variability).

Let c_α denote the $(1-\alpha)$ quantile of this bootstrap distribution. In the ideal uncontaminated case, the observed statistic T_{ideal} follows this bootstrap distribution, yielding

$$\mathbb{P}_{H_0}(T_{\text{ideal}} > c_\alpha) = \alpha,$$

(ii) Contaminated case and stability of $\mathcal{K}$. In the presence of contamination, we assume the discrepancy functional $\mathcal{K}(\cdot,\cdot)$ satisfies Assumption 5.6, then the observed statistic T_{obs} and the statistic in the uncontaminated case T_{ideal} satisfies

$$|T_{\text{obs}} - T_{\text{ideal}}| \leq C\rho_q.$$

hence

$$T_{\text{obs}} > c_\alpha \quad \text{implies} \quad T_{\text{ideal}} > c_\alpha - C\rho_q.$$

(iii) Effect on rejection probability. The above perturbation implies

$$\mathbb{P}_{H_0}(\text{reject}) \leq \alpha + O(\rho_q).$$

Using the Stage I result $\rho_q \leq q + \varepsilon_n$, we obtain

$$\mathbb{P}_{H_0}(\text{reject}) \leq \alpha + O(q + \varepsilon_n).$$

(iv) Multiple FDR levels. To reduce sensitivity to a single threshold, we consider multiple FDR levels $\mathcal{Q} = \{q_1, \ldots, q_K\}$. We reject if

$$\min_k p_{q_k} \leq \frac{\alpha}{K}.$$

By Bonferroni's inequality, we obtain

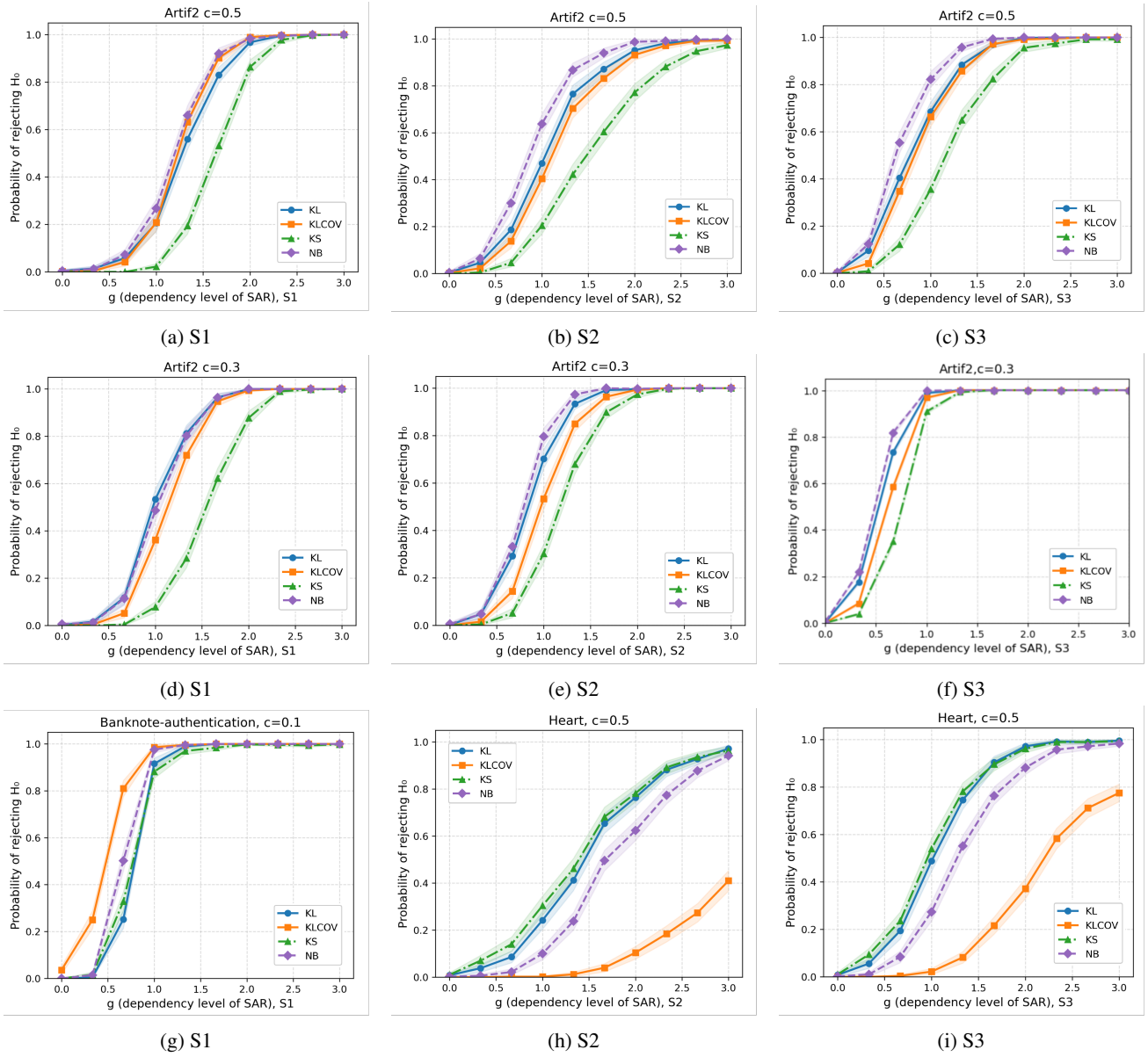
$$\mathbb{P}_{H_0}(\text{reject}) \leq \alpha + \sum_{j=1}^{K} O(q_j + \varepsilon_n).$$

We may derive that the use of multiple q values does not inflate the type-I error, since the Bonferroni correction controls the family-wise error. Thus, the procedure achieves valid type-I error control while improving robustness to threshold selection.

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 POWER ANALYSIS UNDER SAR

Figure3 provides supplementary experiments to 6.3, demonstrating the test power on Artif2 and synthetic datasets under SAR labeling mechanisms.



(a) S1 (b) S2 (c) S3

(d) S1 (e) S2 (f) S3

(g) S1 (h) S2 (i) S3

Figure 3: Test Power under SAR Labeling Mechanisms as a Function of Dependency Strength g on Synthetic and Real-World Datasets

B.2 SENSITIVITY ANALYSIS WITH RESPECT TO THE FDR LEVEL

Figure 4 and Tables 5 and 6 summarize how the nominal FDR level q affects both the power of the second-stage test and the empirical positive sets constructed in Stage I. Across datasets, increasing q generally enlarges the empirical positive set by including additional unlabeled instances with less extreme labeling scores. This expansion improves coverage of the latent positive population and often leads to higher detection power under SAR mechanisms, as shown in Figure 4.

The tables further illustrate the corresponding precision–coverage trade-off. For small values of q, the selected empirical positives are highly conservative and tend to have high precision, but may cover only a restricted region of the positive distribution. As q increases, the empirical positive set becomes larger and more representative, including more lower-score positive instances. Importantly, the resulting accuracy remains stable in most settings, indicating that the FDR-controlled construction does not introduce severe contamination even when the selection threshold is relaxed. These results support the use of multiple FDR levels in the proposed procedure, since different values of q capture complementary parts of the positive distribution and reduce sensitivity to any single threshold choice.

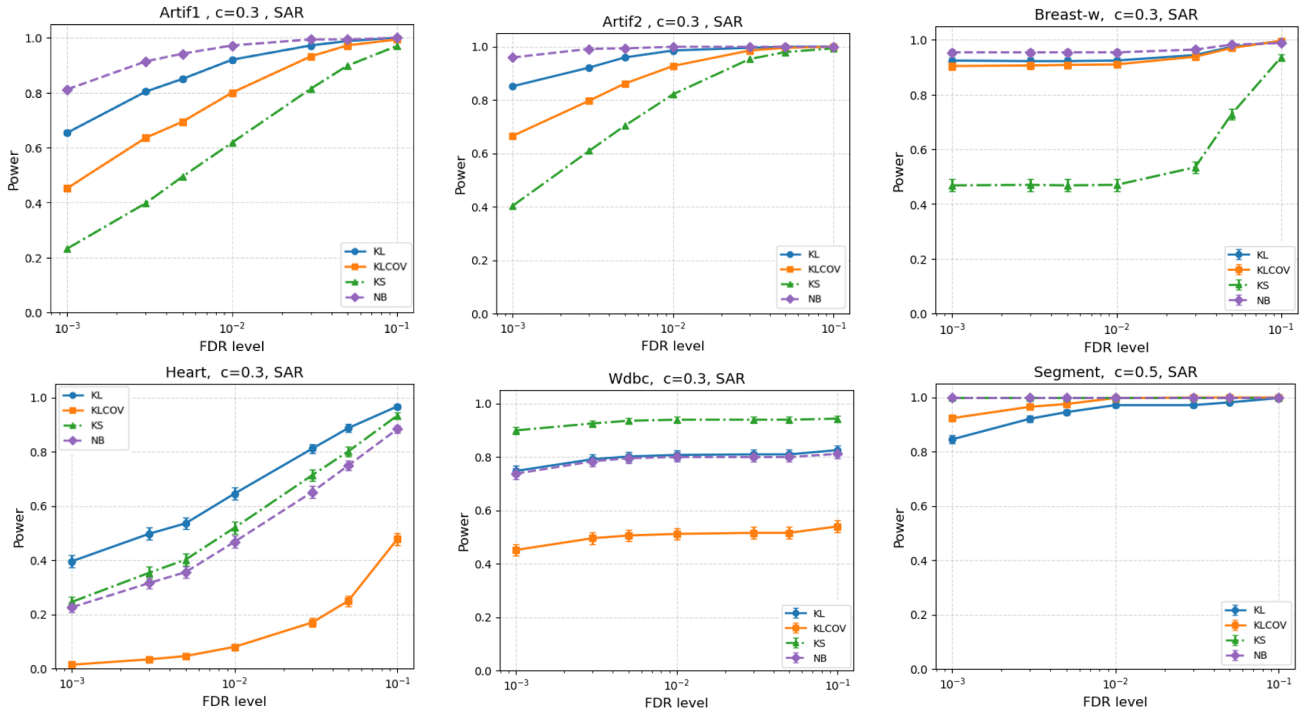


Figure 4: Detection power of the second-stage test as a function of the nominal FDR level on the synthetic and real datasets under SAR labeling mechanisms.

Table 5: Effect of the FDR level on empirical positive set construction on BREAST-W, comparing **SCAR** and **SAR** labeling mechanisms across different labeling probability c.

c	q	**SCAR**			**SAR**		
		$\lvert\widehat{\mathcal{P}}\rvert$	Precision	Accuracy	$\lvert\widehat{\mathcal{P}}\rvert$	Precision	Accuracy
0.2	0.001	203.7 ± 0.5	0.872 ± 0.001	0.936 ± 0.000	154.0 ± 0.6	0.948 ± 0.001	0.915 ± 0.001
	0.005	207.3 ± 0.6	0.866 ± 0.001	0.936 ± 0.000	156.6 ± 0.7	0.945 ± 0.001	0.917 ± 0.001
	0.010	210.9 ± 0.6	0.859 ± 0.001	0.935 ± 0.000	159.8 ± 0.7	0.942 ± 0.001	0.919 ± 0.001
	0.050	224.8 ± 0.4	0.829 ± 0.001	0.930 ± 0.000	178.6 ± 0.5	0.919 ± 0.001	0.933 ± 0.001
	0.100	225.0 ± 0.4	0.828 ± 0.001	0.930 ± 0.000	179.0 ± 0.5	0.918 ± 0.001	0.933 ± 0.001
0.3	0.001	195.2 ± 0.3	0.839 ± 0.001	0.942 ± 0.000	153.4 ± 0.4	0.928 ± 0.001	0.936 ± 0.000
	0.005	195.3 ± 0.3	0.839 ± 0.001	0.942 ± 0.000	153.5 ± 0.4	0.928 ± 0.001	0.936 ± 0.000
	0.010	195.3 ± 0.3	0.839 ± 0.001	0.942 ± 0.000	153.6 ± 0.4	0.928 ± 0.001	0.936 ± 0.000
	0.050	202.4 ± 0.6	0.818 ± 0.001	0.936 ± 0.000	168.6 ± 0.6	0.902 ± 0.001	0.942 ± 0.000
	0.100	211.7 ± 0.4	0.788 ± 0.001	0.925 ± 0.000	174.2 ± 0.4	0.890 ± 0.001	0.943 ± 0.000
0.5	0.001	157.8 ± 0.4	0.760 ± 0.001	0.933 ± 0.000	128.8 ± 0.3	0.856 ± 0.001	0.949 ± 0.000
	0.005	159.0 ± 0.4	0.754 ± 0.001	0.931 ± 0.000	128.8 ± 0.3	0.855 ± 0.001	0.949 ± 0.000
	0.010	159.0 ± 0.4	0.754 ± 0.001	0.931 ± 0.000	128.9 ± 0.3	0.855 ± 0.001	0.949 ± 0.000
	0.050	159.1 ± 0.4	0.753 ± 0.001	0.931 ± 0.000	129.0 ± 0.3	0.855 ± 0.001	0.949 ± 0.000
	0.100	159.1 ± 0.4	0.753 ± 0.001	0.931 ± 0.000	133.3 ± 0.4	0.840 ± 0.001	0.947 ± 0.000
0.7	0.001	109.4 ± 0.3	0.657 ± 0.001	0.929 ± 0.000	94.5 ± 0.3	0.720 ± 0.002	0.942 ± 0.000
	0.005	114.5 ± 0.4	0.628 ± 0.002	0.919 ± 0.000	94.5 ± 0.3	0.719 ± 0.002	0.942 ± 0.000
	0.010	116.7 ± 0.3	0.616 ± 0.001	0.915 ± 0.000	94.6 ± 0.3	0.719 ± 0.002	0.942 ± 0.000
	0.050	117.1 ± 0.3	0.614 ± 0.001	0.914 ± 0.000	94.7 ± 0.3	0.718 ± 0.002	0.942 ± 0.000
	0.100	117.1 ± 0.3	0.614 ± 0.001	0.914 ± 0.000	95.0 ± 0.3	0.716 ± 0.002	0.941 ± 0.000

Table 6: Effect of the FDR level on empirical positive set construction on BANKNOTE-AUTHENTICATION, comparing **SCAR** and **SAR** labeling mechanisms across different labelling probability c.

c	q	**SCAR**			**SAR**		
		$\lvert\widehat{\mathcal{P}}\rvert$	Precision	Accuracy	$\lvert\widehat{\mathcal{P}}\rvert$	Precision	Accuracy
0.2	0.001	444.9 ± 0.8	0.906 ± 0.001	0.899 ± 0.001	380.4 ± 1.2	0.924 ± 0.001	0.865 ± 0.001
	0.005	444.9 ± 0.8	0.906 ± 0.001	0.899 ± 0.001	380.4 ± 1.2	0.924 ± 0.001	0.865 ± 0.001
	0.010	444.9 ± 0.8	0.906 ± 0.001	0.899 ± 0.001	380.8 ± 1.2	0.923 ± 0.001	0.865 ± 0.001
	0.050	484.9 ± 2.3	0.871 ± 0.002	0.894 ± 0.001	455.1 ± 1.8	0.859 ± 0.002	0.867 ± 0.001
	0.100	552.5 ± 0.9	0.809 ± 0.001	0.883 ± 0.001	473.5 ± 0.9	0.842 ± 0.001	0.866 ± 0.001
0.3	0.001	446.5 ± 0.6	0.893 ± 0.001	0.936 ± 0.000	373.4 ± 0.7	0.915 ± 0.001	0.898 ± 0.001
	0.005	446.5 ± 0.6	0.893 ± 0.001	0.936 ± 0.000	373.4 ± 0.7	0.915 ± 0.001	0.898 ± 0.001
	0.010	446.5 ± 0.6	0.893 ± 0.001	0.936 ± 0.000	373.4 ± 0.7	0.915 ± 0.001	0.898 ± 0.001
	0.050	447.3 ± 0.7	0.892 ± 0.001	0.935 ± 0.000	382.1 ± 1.3	0.906 ± 0.001	0.897 ± 0.001
	0.100	506.1 ± 1.9	0.816 ± 0.002	0.907 ± 0.001	451.7 ± 1.1	0.825 ± 0.001	0.883 ± 0.001
0.5	0.001	343.2 ± 0.9	0.880 ± 0.001	0.957 ± 0.000	290.5 ± 0.6	0.873 ± 0.001	0.927 ± 0.000
	0.005	354.0 ± 0.7	0.855 ± 0.001	0.949 ± 0.000	290.8 ± 0.6	0.872 ± 0.001	0.927 ± 0.000
	0.010	355.3 ± 0.6	0.852 ± 0.001	0.948 ± 0.000	290.8 ± 0.6	0.872 ± 0.001	0.927 ± 0.000
	0.050	355.5 ± 0.6	0.852 ± 0.001	0.948 ± 0.000	290.8 ± 0.6	0.872 ± 0.001	0.927 ± 0.000
	0.100	355.5 ± 0.6	0.852 ± 0.001	0.948 ± 0.000	290.8 ± 0.6	0.872 ± 0.001	0.927 ± 0.000
0.7	0.001	215.9 ± 0.5	0.846 ± 0.001	0.964 ± 0.000	203.0 ± 0.5	0.792 ± 0.001	0.935 ± 0.000
	0.005	219.0 ± 0.6	0.835 ± 0.001	0.961 ± 0.000	204.0 ± 0.5	0.789 ± 0.001	0.934 ± 0.000
	0.010	231.1 ± 0.8	0.792 ± 0.002	0.949 ± 0.001	204.1 ± 0.5	0.788 ± 0.001	0.934 ± 0.000
	0.050	239.1 ± 0.6	0.765 ± 0.001	0.940 ± 0.000	204.1 ± 0.5	0.788 ± 0.001	0.934 ± 0.000
	0.100	239.1 ± 0.6	0.765 ± 0.001	0.940 ± 0.000	204.1 ± 0.5	0.788 ± 0.001	0.934 ± 0.000

B.3 COMPARISON OF ALTERNATIVE STRATEGIES FOR CONSTRUCTING THE EMPIRICAL POSITIVE SET

In the first-stage test, our main approach constructs the empirical positive set by fitting a one-dimensional Gaussian mixture model to the labeling scores. To assess the robustness of this design choice, we additionally compared it with two alternative strategies that differ in the underlying modeling space.

The first alternative fits a multivariate Gaussian mixture model directly to the feature vectors of unlabeled samples and identifies the negative component based on the average Mahalanobis distance to labeled positives. Candidate positives are then selected via χ^2-based p-values. The second alternative models the labeling scores using a Beta mixture model and similarly selects samples according to tail probabilities under the estimated negative component.

These two approaches are conceptually reasonable, as they respectively rely on feature-space density modeling and probability-space mixture modeling. However, empirical results show that both alternatives produce empirical positive sets with noticeably lower purity and stability compared to the proposed score-based GMM–FDR method. In particular, direct feature-space modeling is sensitive to high-dimensional noise and local density variations, while Beta mixture modeling, despite its flexibility on $[0, 1]$, exhibits weaker separation under distributional overlap.

Overall, these comparisons support our design choice of constructing the empirical positive set from a one-dimensional labeling score representation with GMM and explicit FDR control, which yields more stable and statistically reliable candidate positives across both SCAR and SAR settings. To evaluate the quality of the constructed empirical positive sets, Table 7 reports their accuracy and precision under both SCAR and SAR mechanisms.

Table 7: Comparison of empirical positive set construction methods under the **SCAR** and **SAR** mechanisms. Only Accuracy (Acc) and Precision (Prec) are reported. Best results for each dataset are highlighted in red.

Dataset	SCAR					
	GMM (labeling score)		Beta		GMM (multivariate)	
	Acc	Prec	Acc	Prec	Acc	Prec
Wdbc	0.881±0.001	0.664±0.002	0.742±0.002	0.473±0.002	0.879±0.000	0.667±0.001
Breast-w	0.942±0.000	0.839±0.001	0.924±0.000	0.785±0.001	0.825±0.000	0.606±0.000
Banknote	0.936±0.000	0.893±0.001	0.892±0.000	0.781±0.001	0.572±0.001	0.388±0.001
Segment	0.862±0.004	0.866±0.002	0.813±0.000	0.725±0.001	0.686±0.003	0.707±0.003
Vote	0.896±0.001	0.840±0.002	0.824±0.001	0.651±0.002	0.826±0.006	0.657±0.006
Heart	0.741±0.001	0.708±0.003	0.525±0.002	0.441±0.001	0.560±0.005	0.428±0.006

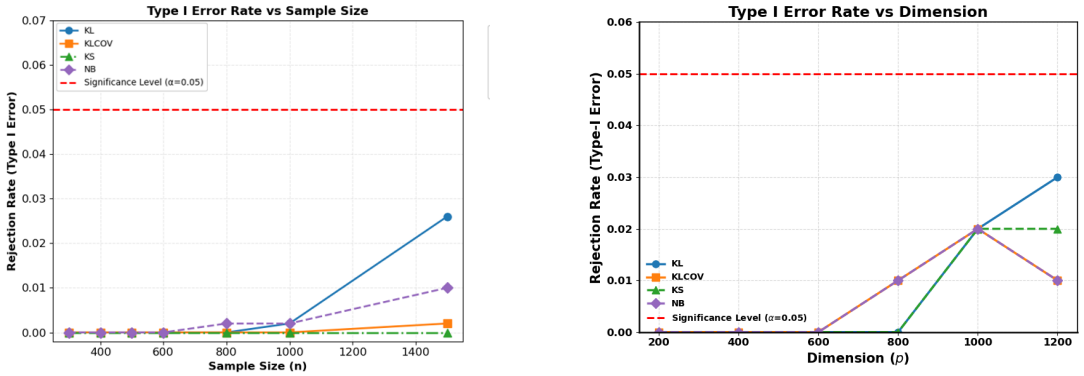
Dataset	SAR					
	GMM (labeling score)		Beta		GMM (multivariate)	
	Acc	Prec	Acc	Prec	Acc	Prec
Wdbc	0.918±0.000	0.850±0.001	0.883±0.001	0.741±0.002	0.864±0.001	0.712±0.001
Breast-w	0.923±0.001	0.943±0.001	0.938±0.000	0.911±0.001	0.825±0.000	0.606±0.000
Banknote	0.901±0.000	0.905±0.001	0.879±0.000	0.798±0.001	0.562±0.001	0.353±0.001
Segment	0.817±0.001	0.910±0.001	0.808±0.001	0.763±0.001	0.663±0.002	0.669±0.002
Vote	0.902±0.001	0.914±0.001	0.878±0.001	0.747±0.002	0.851±0.001	0.679±0.001
Heart	0.768±0.001	0.747±0.002	0.659±0.001	0.527±0.001	0.496±0.005	0.404±0.003

B.4 SCALABILITY ANALYSIS

We further evaluate the scalability of the proposed framework in terms of sample size and feature dimensionality. Specifically, we investigate whether Type-I error control under the SCAR mechanism remains stable as data scale increases.

Effect of Sample Size. We vary the sample size $n \in \{400, 600, 800, 1000, 1200\}$. For each setting, we generate data under the SCAR labeling mechanism and compute the rejection rate of the proposed test. As shown in Figure 5a, the empirical Type-I error remains consistently close to the nominal significance level $\alpha = 0.05$ across all sample sizes. This indicates that the bootstrap-based inference procedure is stable with respect to increasing sample size and does not exhibit systematic inflation of Type-I error.

Effect of Feature Dimension. We next vary the feature dimension $p \in \{200, 400, 600, 800, 1000, 1200\}$. This setting evaluates whether high-dimensional feature spaces affect the validity of the proposed test. As shown in Figure 5b, the empirical Type-I error remains well controlled across all dimensional settings. Although minor fluctuations are observed , no systematic deviation from the nominal level is detected. This suggests that the proposed test is robust to high-dimensional settings.



(a) Effect of sample size.

(b) Effect of feature dimension.

Figure 5: Type-I error under SCAR on ARTIF2 with varying sample size and feature dimension.

B.5 ROBUSTNESS UNDER DISTRIBUTIONAL OVERLAP

We study the robustness of the proposed procedure under varying degrees of overlap between positive and negative distributions in the unlabeled set. We control the mean separation Δ, where smaller values indicate stronger overlap.

Table 8 reports the rejection rates under different values of Δ. Overall, the proposed procedure demonstrates strong robustness to distributional overlap. Even under the most challenging setting ($\Delta = 1.0$), the KL, KLCOV, and NB statistics maintain high rejection rates, indicating stable detection performance despite substantial overlap . In contrast, the KS statistic is considerably more sensitive to severe overlap, exhibiting a noticeable reduction in rejection rate when Δ is small. As the class separation increases, the performance of all statistics improves, with the KS statistic benefiting the most.

Table 8: Rejection Rate under SAR with Different Class Separation Levels Δ.

Δ	KL	KLCOV	KS	NB
1.0	0.93 ± 0.25	0.86 ± 0.34	0.32 ± 0.46	0.90 ± 0.30
1.5	1.00 ± 0.00	1.00 ± 0.00	0.95 ± 0.22	1.00 ± 0.00
2.0	1.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.17	1.00 ± 0.00

B.6 IMPACT OF MECHANISM IDENTIFICATION ON DOWNSTREAM PU LEARNING

We finally investigate the practical impact of labeling mechanism identification on downstream PU learning performance. Table 9 reports the downstream performance when the PU data are generated under the SCAR mechanism and SAR mechanism respectively. In each setting, we apply both SCAR-based and SAR-based PU learning algorithms, and directly compare their prediction accuracy and runtime to assess the impact of mechanism mismatch.

Overall, the proposed two-stage mechanism identification framework serves not only as a diagnostic tool, but also as a principled decision mechanism that directly translates into improved downstream accuracy and learning efficiency in practical PU learning scenarios.

Table 9: Impact of mechanism identification on downstream PU learning performance.

SCAR-generated PU data

Dataset	Accuracy (SCAR)	Time (SCAR)	Accuracy (SAR)	Time (SAR)
Breast-W	$\mathbf{0.8462 \pm 0.0043}$	0.2139 ± 0.0039	0.8319 ± 0.0442	5.1223 ± 1.1893
Banknote	$\mathbf{0.9878 \pm 0.0007}$	0.7544 ± 0.0568	0.7857 ± 0.0023	19.0641 ± 4.5178
Wdbc	$\mathbf{0.9486 \pm 0.0049}$	0.7078 ± 0.0741	0.8184 ± 0.0030	54.1093 ± 11.2344
Heart	$\mathbf{0.7696 \pm 0.0027}$	0.1721 ± 0.0028	0.7688 ± 0.0081	4.4695 ± 0.5613
Vote	0.8372 ± 0.0018	0.4034 ± 0.0373	$\mathbf{0.9222 \pm 0.0038}$	4.6685 ± 3.0104

SAR-generated PU data

Dataset	Accuracy (SCAR)	Time (SCAR)	Accuracy (SAR)	Time (SAR)
Breast-W	0.6690 ± 0.0014	0.4500 ± 0.0259	$\mathbf{0.7818 \pm 0.0023}$	2.5495 ± 1.9238
Banknote	0.7250 ± 0.0009	0.7233 ± 0.0232	$\mathbf{0.8152 \pm 0.0015}$	5.3066 ± 1.1559
Wdbc	0.8063 ± 0.0039	0.4902 ± 0.0415	$\mathbf{0.8479 \pm 0.0028}$	82.1031 ± 12.5006
Heart	$\mathbf{0.7457 \pm 0.0031}$	0.4059 ± 0.0195	0.7283 ± 0.0068	52.1725 ± 123.0163
Vote	0.7871 ± 0.0026	0.4242 ± 0.0302	$\mathbf{0.8335 \pm 0.0055}$	62.0909 ± 21.2133