The Annals of Statistics
2024, Vol. 52, No. 1, 157–183
https://doi.org/10.1214/23-AOS2338

ADAPTIVE NOVELTY DETECTION WITH FALSE DISCOVERY RATE GUARANTEE

BY ARIANE MARANDON[1,a], LIHUA LEI[2,c], DAVID MARY[3,d] AND ETIENNE ROQUAIN[1,b]

[1]*LPSM, Sorbonne Université, Université de Paris & CNRS,* [a]*ariane.marandon-carlhian@sorbonne-universite.fr,* [b]*etienne.roquain@sorbonne-universite.fr*

[2]*Graduate School of Business, Stanford University,* [c]*lihualei@stanford.edu*

[3]*Laboratoire Lagrange, Observatoire de la Côte d'Azur, Université Côte d'Azur & CNRS,* [d]*david.mary@unice.fr*

This paper studies the semisupervised novelty detection problem where a set of "typical" measurements is available to the researcher. Motivated by recent advances in multiple testing and conformal inference, we propose AdaDetect, a flexible method that is able to wrap around any probabilistic classification algorithm and control the false discovery rate (FDR) on detected novelties in finite samples without any distributional assumption other than exchangeability. In contrast to classical FDR-controlling procedures that are often committed to a pre-specified p-value function, AdaDetect learns the transformation in a data-adaptive manner to focus the power on the directions that distinguish between inliers and outliers. Inspired by the multiple testing literature, we further propose variants of AdaDetect that are adaptive to the proportion of nulls while maintaining the finite-sample FDR control. The methods are illustrated on synthetic datasets and real-world datasets, including an application in astrophysics.

1. Introduction.

1.1. *Novelty detection.* In this paper, we consider a novelty detection problem (see, e.g., Blanchard, Lee and Scott (2010) and references therein) where we observe:

- a null training sample (NTS hereafter) $Y = (Y_1, \ldots, Y_n)$ of "typical" measurements where Y_is share a common marginal distribution P_0 which we refer to as the null distribution;
- and a test sample $X = (X_1, \ldots, X_m)$ of "unlabeled" measurements for which the marginal distribution of X_i is denoted by P_i, which might be different from P_0.

These measurements are assumed to take values in a general space $\mathcal{Z}$ endowed with a prescribed σ-field. For example, the space can be the set of real matrices ($\mathcal{Z} = \mathbb{R}^{d \times d'}$) or real vectors ($\mathcal{Z} = \mathbb{R}^d$), whose dimension is potentially large.

Putting two samples together, we observe $Z = (Z_1, \ldots, Z_{n+m}) = (Y_1, \ldots, Y_n, X_1, \ldots, X_m)$. The aim is to detect novelties, namely X_is with $P_i \neq P_0$. This task is illustrated in Figure 1 on a classical image dataset, where we want to detect hand-written digit '9's in the test sample based on an NTS of digits '4'. The procedure, that declares as novelties the images with red boxes, can make false discoveries (digit '4') and true discoveries (digit '9').

To avoid false positives that might be costly in practice, we seek to control the false discovery rate (FDR), defined as the average proportion of errors among the discoveries, while attempting to maximize the true discovery rate (TDR), defined as the average portion of detected novelties. FDR has been a very popular criterion in multiple testing and exploratory analysis since its introduction by Benjamini and Hochberg (1995); see Benjamini (2010) for

Received October 2022; revised October 2023.
MSC2020 subject classifications. Primary 62J15, 62G10; secondary 62H30.
Key words and phrases. Adaptive multiple testing, novelty detection, false discovery rate, conformal p-values, machine learning, classification, neural network, knockoff.

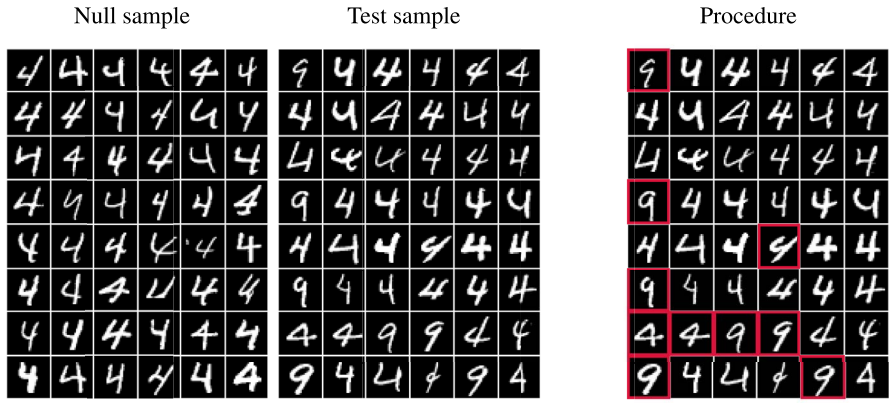

FIG. 1. *Illustration of the novelty detection task on the MNIST dataset (LeCun and Cortes (2010)); see Section 6 for more details on the setting.*

a detailed discussion and Barber and Candès (2015), Barber, Candès and Samworth (2020), Bogdan et al. (2015), Javanmard and Javadi (2019), Ma, Cai and Li (2021) for recent developments, among others.

1.2. *Existing strategies.* For a standard multiple testing problem where the null distribution P_0 is known, the celebrated Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg (1995)) controls the FDR in finite samples uniformly over all alternative distributions, when the test statistics are independent or satisfy the positive regression dependency on each one from a subset (PRDS) property; see Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001). Variants of the BH procedure have been proposed to relax the conservatism when the fraction of true nulls is not close to 1, such as the Storey-BH (Storey, Taylor and Siegmund (2004)) or Quantile-BH procedure (Benjamini, Krieger and Yekutieli (2006), Blanchard and Roquain (2009), Sarkar (2008)), and to robustify the FDR control under more general dependence structures (see Fithian and Lei (2022) and references therein).

Despite this generality, BH-like methods have two major limitations in novelty detection problems with multivariate measurements:

(i) it is based on p-values or, more generally, univariate scores with a *known distribution under the null*, which is typically out of reach for such problems;

(ii) the score function that transforms the multivariate measurements into univariate test statistics (e.g., the p-value transformation) is *pre-specified*, while it should be *learned from data* for the sake of power.

We now discuss several existing solutions that partially circumvent these limitations. Table 1 provides a summary of the properties of each method, along with the corresponding applicable settings.

A popular solution in the multiple testing literature is the empirical Bayes approach, which operates on the local FDR instead of the p-values. Assuming a two-group mixture model (Efron et al. (2001)), the local FDR is defined as the probability of being null conditional on the observed measurement values. The latter can be estimated by estimating the null and alternative densities together with the proportion of nulls; see Efron (2004), Efron (2007), Efron (2008), Efron (2009). Combining local FDRs appropriately controls FDR asymptotically, under the assumptions that allow the model to be consistently estimated, and achieves optimal power, as shown in a series of paper by Cai and Sun (2009), Cai, Sun and Wang (2019), Sun and Cai (2007), Sun and Cai (2009). We refer to this procedure as the SC procedure hereafter. Despite the appealing optimality guarantees, the model assumptions tend

TABLE 1
Properties of different methods and the specific settings in which they can be applied for novelty detection

Method	Finite sample FDR control	Adaptative score	Learning alternative	Unknown null
Benjamini and Hochberg (1995)	**yes**	*no*	*no*	*no*
Sun and Cai (2007)	*no*	**yes**	**yes**	**yes**
Weinstein, Barber and Candès (2017)				
Mary and Roquain (2022)	**yes**	*no*	*no*	**yes**
Bates et al. (2023)	**yes**	**yes**	*no*	**yes**
Yang et al. (2021)	**yes**	**yes**	**yes**	*no*
AdaDetect (our approach)	**yes**	**yes**	**yes**	**yes**

to be fragile when the dimension d of the test statistics is moderately high. In such cases, accurate model estimation is hard to come by and the FDR of the SC procedure can thus be inflated; see our numerical experiments in Section 6 for an illustration.

Another line of research stems from conformal inference. While this technique is designed for prediction inference (see Angelopoulos and Bates (2021) for a recent review), it can also be employed in the novelty detection problem. In particular, it can generate *conformal p-values* that are super-uniform under the null without any model assumption beyond that the data are exchangeable (e.g., Balasubramanian, Ho and Vovk (2014), Bates et al. (2023), Vovk, Gammerman and Shafer (2005)). This approach starts by transforming Z_j into a univariate score S_j, called the *nonconformity score*, that measures the conformity to the data and then computes an *empirical p-value*, also known as the conformal p-value, to evaluate the statistical evidence of being a novelty:

$$p_j = (n+1)^{-1}\left(1 + \sum_{i=1}^{n} \mathbb{1}\{S_i \geq S_{n+j}\}\right). \tag{1}$$

Each p-value is marginally super-uniform under the null due to exchangeability and hence yields a valid test. Nonetheless, since the conformal p-values all use the same null sample, the above operation induces dependence between the p-values, making it unclear whether common multiple testing procedures are guaranteed to control FDR. Bates et al. (2023) carefully study the dependence structure and show that the split (or inductive) conformal p-values are PRDS. As a consequence, BH procedure applied on these conformal p-values controls the FDR. However, the approach limits the construction of the scores to be based solely on null examples and hence cannot learn the patterns of novelties in the mixed samples, unless extra labelled novelties are available (Liang, Sesia and Sun (2022)), which are not always possible. Even when labelled novelties are present, they may behave differently than the ones in the mixed sample that we aim to detect. For this reason, Bates et al. (2023) apply the one class classification techniques (e.g., Schölkopf et al. (2001)) that are not adaptive to the novelties. In sum, while the method successfully solves the issue (i), it falls short of adequately addressing issue (ii). On the other hand, while other versions of conformal p-values, like full conformal p-values (Vovk, Gammerman and Shafer (2005)) and cross conformal p-values (Barber et al. (2021), Vovk (2015)), can use test samples and yield marginally valid p-values, they generally fail to satisfy the PRDS property, making it unclear whether the BH procedure would control FDR.

A subsequent work by Yang et al. (2021) proposes the Bag Of Null Statistics (BONuS) procedure for multiple testing problems with high dimensional test statistics, which largely motivates our method. The BONuS procedure learns a score function of the form $S_i = g(Z_i, (Z_1, \ldots, Z_n))$ and the method is valid as long as $g(Z_i, \cdot)$ is permutation invariant

thereby allowing the transformation to be adapted to novelties. While the framework is flexible, they focus on the parametric setting where the null distribution is known, like Gaussian or multinomial, and the measurements are independent. In these cases, they propose using the estimated local FDR as the score function for which the alternative distribution and null proportion are learned by an empirical Bayes approach. The BONuS procedure controls the FDR in finite samples regardless of the quality of the estimates, even if the working model is completely wrong. However, for novelty detection problems, the local FDR involves unknown null and alternative densities, which are difficult to fit in high dimensions. Hence, point (ii) mentioned earlier remains partially addressed.

Lastly, we briefly review other related work that study different settings. The "counting knockoffs" procedure introduced by Weinstein, Barber and Candès (2017) is designed for multiple testing for high-dimensional linear models with random design matrices. Mary and Roquain (2022) show that it is equivalent to applying the BH procedure to the scores $S_1, \dots, S_{n+m}$ and closely related to the BONuS procedure. More recently, Rava et al. (2021) develop a method that is equivalent to applying the BH procedure on the conformal p-values to obtain a finite sample control of the false selection rate (FSR) for the task of (supervised) classification.

1.3. *Contributions*. In this work we introduce AdaDetect, an extension[1] of the BONuS procedure for novelty detection problems. In particular, we show how to leverage flexible off-the-shelf classification algorithms in machine learning to address both issues (i) and (ii) without compromising the FDR-controlling guarantees. In a nutshell, AdaDetect operates by initially splitting the null sample in two parts, $(Y_1, \dots, Y_k)$ and $(Y_{k+1}, \dots, Y_n)$, generating a membership label $A_j = -1$ if $Z_j \in \{Y_1, \dots, Y_k\}$ and $A_j = 1$ otherwise, and subsequently calculating a score function using a binary classifier trained on $(Z_i, A_i)_{i=1}^{n+m}$ and applying the BH procedure on the empirical p-values. For the example illustrated in Figure 1, Adadetect would split the null samples (digits '4') into two subsets and train a probabilistic classifier using both the null and test samples to distinguish the first subset of the null sample and the mix of the second subset of the null sample and the test sample (digits '4' and '9'). The predicted probability to be in the mixed sample is taken as the score. When the classification algorithm performs well, the scores tend to be larger for novelties than for nulls, because novelties are only present in the mixed sample. A comprehensive description of the procedure can be found in Section 2.4.

We summarize our main results below.

- In Section 3, we revisit the theoretical guarantees in Bates et al. (2023), Mary and Roquain (2022), Weinstein, Barber and Candès (2017) and provide new FDR bounds based on an extension of the leave-one-out technique in the multiple testing literature. The bounds show that AdaDetect, as well as its π_0-adaptive variants Storey–AdaDetect and Quantile–AdaDetect, controls the FDR in finite samples with *arbitrary* classification algorithms even if the algorithm performs poorly. This is in sharp contrast to the SC procedure which heavily relies on correct model specification and consistent estimation.
- In addition, we extend the result in Bates et al. (2023) to show that the empirical p-values are PRDS under a more general exchangeability assumption, even if the score function depends on both null and test samples. For instance, our condition covers the Gaussian distributions with equi-correlation (Example 3.1). This PRDS property suggests that the resulting p-values can be applied in other contexts beyond the FDR control (e.g., Goeman and Solari (2011)).

[1]More precisely, we extend the version of BONuS where the score function is fit only in the initial stage; see the discussion in Section 8 for more details.

- In Section 4, we show that *any* score function that is monotone in the ratio between the average density of novelties and the null density yields the optimal power. In particular, the optimal classifier to distinguish between the null and mixed samples is efficient despite that the null training is split and that the mixed sample is contaminated by nulls. The optimal score function can be obtained by minimizing certain loss function such as the cross-entropy loss that is commonly used in neural networks (NN hereafter).
- We provide nonasymptotic power analyses for AdaDetect in Section 5. First, we investigate AdaDetect with the score function given by a constrained empirical risk minimizer (ERM) of the 0-1 loss and show it approaches the optimal likelihood ratio test in an appropriate sense. Next, we provide an oracle inequality for general score functions and conditions under which the procedure mimics its oracle version. We apply the results to analyze power for AdaDetect procedures based on NN and on nonparametric kernel density estimation.
- We demonstrate the efficiency, flexibility, and robustness of AdaDetect[2] in Sections 6 and 7 on synthetic, semisynthetic, and real datasets, including the MNIST image dataset and an astronomy dataset from the 'Sloan Digital Sky Survey'.

2. Preliminaries.

2.1. *Notation.* As in Section 1.1, we let $Y = (Y_1, \dots, Y_n)$ denote the null training sample (NTS) with a common marginal distribution P_0, $X = (X_1, \dots, X_m)$ the test sample with $X_i \sim P_i$ $(1 \le i \le m)$, $Z = (Z_1, \dots, Z_{n+m}) = (Y_1, \dots, Y_n, X_1, \dots, X_m)$ the full sample, $\mathcal{H}_0 = \{1 \le i \le m : P_i = P_0\}$ the set of nulls in the test sample with $m_0 = |\mathcal{H}_0|$, $\pi_0 = m_0/m$, and $\mathcal{H}_1 = \{1, \dots, m\} \setminus \mathcal{H}_0$ the set of novelties with $m_1 = |\mathcal{H}_1|$, $\pi_1 = m_1/m$. For notational convenience, we write $n + \mathcal{H}_0$ for the set $\{n + i, i \in \mathcal{H}_0\}$. Furthermore, we denote by P the joint distribution of Z, which belongs to a family of distributions $\mathcal{P}$ (model).

Throughout the paper, we consider the semisupervised setting (Mary and Roquain (2022)) where the null distribution P_0 is unknown and one can access it only through the measurements in the NTS. In practice, the NTS can be obtained from external data, past experiments or black-box samplers.

2.2. *Criteria.* A novelty detection procedure is a measurable function $R(\cdot)$ that takes Z as input and returns a subset of $\{1, \dots, m\}$ corresponding to the indices of detected novelties within $\{X_1, \dots, X_m\}$. Throughout the paper, we will slightly abuse the notation by using R to refer to both the procedure and the rejection set given by the procedure. Ideally, we want $R(Z)$ to capture novelties (i.e., alternative hypotheses in $\mathcal{H}_1$) and avoid inliers (i.e., null hypotheses in $\mathcal{H}_0$). Given a procedure R, the false discovery rate (FDR) is defined as the expectation of the false discovery proportion (FDP) with respect to the distribution $P \in \mathcal{P}$:

$$\text{FDR}(P, R) = \mathbb{E}_{Z\sim P}\big[\text{FDP}(P, R)\big], \qquad \text{FDP}(P, R) = \frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}\{i \in R\}}{1 \vee |R|}. \tag{2}$$

Similarly, the true discovery rate (TDR) is defined as the expectation of the true discovery proportion (TDP):

$$\text{TDR}(P, R) = \mathbb{E}_{Z\sim P}\big[\text{TDP}(P, R)\big], \qquad \text{TDP}(P, R) = \frac{\sum_{i\in\mathcal{H}_1} \mathbb{1}\{i \in R\}}{1 \vee m_1(P)}. \tag{3}$$

Note that $m_1(P) = 0$ implies $\text{TDP}(P, R) = 0$. Our goal is to build a procedure R that controls the FDR and maximizes the TDR to the fullest extent.

[2]The code is publicly available at https://github.com/arianemarandon/adadetect.

2.3. *BH algorithm and its π_0-adaptive variants.* Suppose a set of p-values $(p_i, 1 \le i \le m)$ is available, the BH algorithm (Benjamini and Hochberg (1995)) returns $R = \{i \in \{1, \dots, m\} : p_i \le \alpha \hat{k}/m\}$, where α is the target FDR level and

$$\hat{k} = \max\left\{k \in \{0, \dots, m\} : \sum_{i=1}^{m} \mathbb{1}\{p_i \le \alpha k/m\} \ge k\right\}. \tag{4}$$

When the null p-values $(p_i, i \in \mathcal{H}_0)$ are independent, super-uniform, and independent of alternative p-values $(p_i, i \in \mathcal{H}_1)$, the BH procedure is proved to control the FDR at level $\alpha\pi_0$ in finite samples (Benjamini and Hochberg (1995)). The independence assumption can be further relaxed to the PRDS condition (Benjamini and Yekutieli (2001)).

When π_0 is not close to 1, the BH procedure is conservative because $\alpha\pi_0 < \pi_0$. When π_0 is known, it can be applied at level α/π_0 to close the gap. In practice, π_0 is usually unknown though. Nonetheless, there exists estimators $\hat{\pi}_0$ of π_0 such that the BH procedure with level $\alpha/\hat{\pi}_0$ continues to control the FDR under independence. Two celebrated estimators are introduced by Storey, Taylor and Siegmund (2004) and Benjamini, Krieger and Yekutieli (2006):

$$\hat{\pi}_0^{\text{Storey}} = \frac{1 + \sum_{i=1}^{m} \mathbb{1}\{p_i \ge \lambda\}}{m(1-\lambda)}, \quad \lambda > 0; \quad \text{or} \tag{5}$$

$$\hat{\pi}_0^{\text{Quant}} = \frac{m - k_0 + 1}{m(1 - p_{(k_0)})}, \quad k_0 \in \{1, \dots, m\}, \tag{6}$$

where $p_{(k_0)}$ is the k_0th smallest[3] p-value. These procedures are often called the π_0-adaptive versions of the BH algorithm.

2.4. *Our method.* In this paper, we propose a method called AdaDetect. It is an adaptive novelty detection procedure that extends the existing strategies described in Weinstein, Barber and Candès (2017), Yang et al. (2021), Mary and Roquain (2022), and Bates et al. (2023). It starts by splitting the null sample $(Y_1, \dots, Y_n)$ in two samples $(Y_1, \dots, Y_k)$ and $(Y_{k+1}, \dots, Y_n)$ with $k \ge 0$. To avoid cluttering our notation, we define ℓ as the size of the second null sample, that is, $\ell = n - k$. It proceeds with the following steps:

1. Compute a data-driven score function of form

$$g(z) = g\big(z, (Z_1, \dots, Z_k), (Z_{k+1}, \dots, Z_{n+m})\big), \quad z \in \mathcal{Z}, \tag{7}$$

which satisfies the following invariance property: for any permutation π of $\{k+1, \dots, n+m\}$ and $z, z_1, \dots, z_{n+m} \in \mathcal{Z}$, we have

$$g\big(z, (z_1, \dots, z_k), (z_{\pi(k+1)}, \dots, z_{\pi(n+m)})\big) = g\big(z, (z_1, \dots, z_k), (z_{k+1}, \dots, z_{n+m})\big). \tag{8}$$

2. Transform the raw data into univariate scores

$$S_i = g\big(Z_i; (Z_1, \dots, Z_k), (Z_{k+1}, \dots, Z_{n+m})\big), \quad i \in \{k+1, \dots, n+m\}. \tag{9}$$

Here, we assume that novelties typically have large scores.

3. For each test point X_j, generate the empirical p-value by comparing S_i with the scores in the NTS:

$$p_j = \frac{1}{\ell + 1}\left(1 + \sum_{i=k+1}^{n} \mathbb{1}\{S_i > S_{n+j}\}\right), \quad j \in \{1, \dots, m\}. \tag{10}$$

[3]In this paper, a convention is to order the p-values from the smallest to the largest, while the test statistics are ordered from the largest to the smallest.

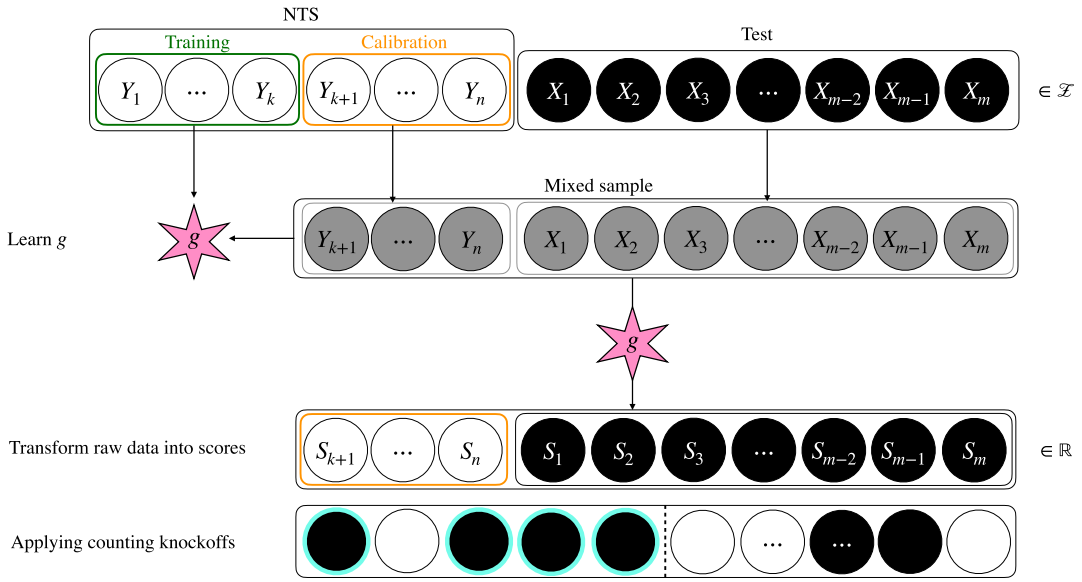

FIG. 2. *A schematic illustration of AdaDetect*: •/○ *stands for a test/null observation, respectively. The vertical dashed line corresponds to the largest threshold* t *for which* $\widehat{\mathrm{FDP}}(t) \le \alpha$ *and the* • *circled in blue correspond to the discoveries of AdaDetect procedure.*

4. Apply the BH algorithm to $(p_1, \dots, p_m)$ at the target level α.

We will call this procedure AdaDetect$_\alpha$ in the sequel to emphasize the target level. By simple algebra, the last two steps together are equivalent to the "counting knockoff" algorithm proposed by Weinstein, Barber and Candès (2017) applied to the scores $S_{k+1}, \dots, S_{n+m}$. Specifically, the method declares i as a novelty if $S_i \ge \hat{t}$ where $\hat{t}$ is the threshold defined by $\min\{t \in \{S_i : k+1 \le i \le n+m\} : \widehat{\mathrm{FDP}}(t) \le \alpha\}$ for $\widehat{\mathrm{FDP}}(t) = \frac{m}{\ell+1}(1 + \sum_{i=k+1}^{n} \mathbb{1}\{S_i \ge t\}) / \sum_{i=n+1}^{n+m} \mathbb{1}\{S_i \ge t\}$. Therefore, the counting knockoff procedure can be seen as a shortcut that avoids computing the empirical p-values explicitly. The pipeline for AdaDetect is illustrated in Figure 2.

AdaDetect offers greater flexibility than existing methods in the types of score functions that can be employed.

- Prespecified p-value transformations are score functions that do not depend on $(Z_1, \dots, Z_k)$ and $(Z_{k+1}, \dots, Z_{n+m})$. For example, when $\mathcal{Z} = \mathbb{R}^d$, the χ^2 test chooses the nonadaptive score $g(z) = \sum_{j=1}^{d} z_j^2$, $z \in \mathbb{R}^d$.
- The one-class classification approach considered by Bates et al. (2023) corresponds to score functions that only depend on $(Z_1, \dots, Z_k)$, but not $(Z_{k+1}, \dots, Z_{n+m})$.
- The BONuS procedure (Yang et al. (2021)) considers empirical Bayes-based score functions that depend on the pooled sample $\{Z_1, \dots, Z_{n+m}\}$ without distinguishing between the null and mixed samples.
- Our proposed method constructs the score function $g(\cdot, (Z_1, \dots, Z_k), (Z_{k+1}, \dots, Z_{n+m}))$ as the estimated probability by any probabilistic classifier that distinguishes between $(Z_1, \dots, Z_k)$ and $(Z_{k+1}, \dots, Z_{n+m})$; see Section 4 for details.

Lastly, we propose the Storey–AdaDetect and Quantile–AdaDetect as the π_0-adaptive versions of AdaDetect applied at level $\alpha/\widehat{\pi}_0^{\mathrm{Storey}}$ and $\alpha/\widehat{\pi}_0^{\mathrm{Quant}}$, respectively, in which the p-values have been replaced by the empirical ones.

REMARK 2.1. An appealing property of Adadetect and its adaptive versions is that the rejection is invariant to strictly increasing transformations of score function. This feature proves useful in the power analysis of AdaDetect; see Section 4.

REMARK 2.2. By construction, the empirical p-values are multiples of $1/(\ell+1)$. As Mary and Roquain (2022) point out, the number of null samples ℓ needs to be larger than $m/(\alpha(1 \vee M))$ in order to guarantee sufficient resolution of the p-values for the BH procedure, where $M \geq 0$ is some high-probability lower bound on the number of rejections. Typically, if M is of the order of m, a constant ℓ would suffice, while if $M = 0$ (i.e., without any prior knowledge on the number of rejections), ℓ should be larger than m/α. In general practical situations where $n \gtrsim m$, we recommend setting $\ell = m$ and this choice works reasonably well in our numerical experiments. When $m > n$, it might be more appropriate to impose further assumptions on the distribution (e.g., the knowledge of M).

3. FDR control. In this section, we prove that AdaDetect and its π_0-adaptive variants control the FDR. In Section 3.1, we state the key assumption of exchangeability and show it translates to the scores as long as g satisfies the condition (8). Based on this observation, we prove in Section 3.2 that the empirical p-values are PRDS, which is a highly nontrivial extension of the results by Bates et al. (2023). Though the PRDS property implies the FDR control of AdaDetect as a result of Benjamini and Yekutieli (2001), we present in Section 3.3 an alternative proof based on a new FDR expression that unify and extend the previous FDR bounds. Lastly, in Section 3.4, we prove the FDR control for Storey–AdaDetect and Quantile–AdaDetect based on an FDR bound for general π_0-adaptive versions of AdaDetect.

3.1. *Exchangeability.* We make the following assumption on the raw measurements throughout the paper.

ASSUMPTION 1. $(Y_1, \ldots, Y_n, X_i, i \in \mathcal{H}_0)$ are exchangeable conditional on $(X_i, i \in \mathcal{H}_1)$.[4]

Clearly, Assumption 1 holds when the measurements are independent, as assumed by Yang et al. (2021) and Bates et al. (2023). In general, Assumption 1 allows for dependencies among the measurements.

EXAMPLE 3.1. Consider the observation where $Z_i = \mu_i + \rho^{1/2}\xi + (1-\rho)^{1/2}\varepsilon_i$, $1 \leq i \leq n+m$, with the variables $\xi, \varepsilon_1, \ldots, \varepsilon_{n+m}$ being i.i.d. $\sim \mathcal{N}(0, I_d)$, ρ being a nonnegative correlation coefficient, and $\mu_i = 0$ for $i \in \{1, \ldots, n\} \cup (n + \mathcal{H}_0)$ (hence $\mathcal{Z} = \mathbb{R}^d$). Then Assumption 1 holds. The case $d = 1$ corresponds to the Gaussian equi-correlated case, which is widely studied in the multiple testing literature (e.g., Korn et al. (2004)).

For our results, a necessary assumption is exchangeability of the scores under the null.

ASSUMPTION 2. $(S_{k+1}, \ldots, S_n, S_{n+i}, i \in \mathcal{H}_0)$ is exchangeable conditionally on $(S_{n+i}, i \in \mathcal{H}_1)$.

It turns out the exchangeability of the raw measurements translates to the scores.

[4]Note that such an assumption implicitly assumes that such a conditional distribution exists, which is always the case for instance when $\mathcal{Z} = \mathbb{R}^d$ or $\mathcal{Z}$ is discrete.

LEMMA 3.2. *Under Assumption 1, the adaptive scores defined by (9) satisfy Assumption 2 for any score function that satisfies the condition (8).*

This result substantially simplifies the FDR analysis presented in the next section. To avoid unnecessary mathematical complications, we make the following mild assumption.

ASSUMPTION 3. $(S_{k+1}, \ldots, S_{n+m})$ have no ties almost surely.

3.2. *The p-values are PRDS.* Following Benjamini and Yekutieli (2001), we say a family of p-values $(p_i, 1 \le i \le m)$ is PRDS on $\mathcal{H}_0$ if, for any $i \in \mathcal{H}_0$ and nondecreasing[5] measurable set $D \subset [0,1]^m$, the function $u \in [0,1] \mapsto \mathbb{P}((p_j, 1 \le j \le m) \in D | p_i = u)$ is nondecreasing.

THEOREM 3.3. *For any family of scores $(S_{k+1}, \ldots, S_{n+m})$ satisfying Assumptions 2 and 3, the empirical p-values defined in (10) are PRDS on $\mathcal{H}_0$ and the null p-values are super-uniform. In particular, under Assumptions 1 and 3, this result holds for the p-values generated by AdaDetect with a score function satisfying (8).*

We present a proof of Theorem 3.3 in Section A.3 of the Supplementary Material (Marandon et al. (2024)). It extends Theorem 2 in Bates et al. (2023) to dependent scores. Notably, the AdaDetect scores are dependent in general even if the measurements Z_i's are independent because the data-adaptive score function depends on the entire dataset.

Theorem 3.3 has interesting consequences. First, the celebrated result for the BH procedure (Benjamini and Yekutieli (2001), Romano and Wolf (2005)) implies that AdaDetect strongly controls the FDR at level $\alpha\pi_0$. Second, the PRDS property is also useful for other purposes, such as post hoc inference (Goeman and Solari (2011)), FDR control with structural constraints (Loper et al. (2022), Ramdas et al. (2019a)), online FDR control (Fisher (2021), Zrnic, Ramdas and Jordan (2021)), hierarchical FDR control (Barber and Ramdas (2017)) and weighted FDR control with prior knowledge (Ramdas et al. (2019b)). Hence, our result paves the way for developing similar AdaDetect-style procedures in these contexts.

3.3. *A new FDR expression.* While the PRDS property implies the FDR control for AdaDetect, we pursue an alternative way based on a new expression for the FDR of the BH procedure in our setting, which would also yield a lower bound for FDR that is not implied by the PRDS property.

THEOREM 3.4. *Consider any family of scores $(S_{k+1}, \ldots, S_{n+m})$ satisfying Assumptions 2 and 3. Let R_α denote the rejection set of BH procedure applied to p-values defined in (10) at level α. Then, for any distribution $P \in \mathcal{P}$,*

$$\mathrm{FDR}(P, R_\alpha) = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left(\frac{\lfloor \alpha(\ell+1)K_i/m \rfloor}{(\ell+1)K_i} \right), \tag{11}$$

where K_i is a random variable that takes values in $\{1, \ldots, m\}$ for any $i \in \mathcal{H}_0$. In particular, under Assumptions 1 and 3, (11) holds with $R_\alpha = \mathit{AdaDetect}_\alpha$, the AdaDetect procedure at level α.

[5]A set $D \subset [0,1]^m$ is said to be nondecreasing if for any $x \in D$ and $y \in [0,1]^m$, we have $y \in D$ provided that $y_i \ge x_i$ for all i.

The proof of Theorem 3.4 is presented in Section A.4 of the Supplementary Material. It is similar to the classical leave-one-out technique to prove the FDR control for step-up procedure (e.g., Ferreira and Zwinderman (2006), Giraud (2022), Ramdas et al. (2019b), Roquain and Villers (2011)), though it is nontrivial to handle empirical p-values. Since for any $x > 0$ and integer k, we have $\lfloor x \rfloor k \leq \lfloor xk \rfloor \leq xk$, expression (11) immediately implies the following bounds.

COROLLARY 3.5. *Under Assumptions 1 and 3, the following holds, for any values of* $k, \ell, m \geq 1$ *and any parameter* $P \in \mathcal{P}$:

$$m_0 \lfloor \alpha(\ell+1)/m \rfloor / (\ell+1) \leq \mathrm{FDR}(P, \mathit{AdaDetect}_\alpha) \leq \alpha m_0/m. \tag{12}$$

In particular, $\mathrm{FDR}(P, \mathit{AdaDetect}_\alpha) = \alpha\pi_0$ *when* $\alpha(\ell+1)/m$ *is an integer.*

Corollary 3.5 recovers Theorem 3.1 in Mary and Roquain (2022) which imposes a slightly more restrictive condition than Assumption 2. Their proofs are based on martingale techniques and the proof for the lower bound is particularly involved. Here, we rely instead on the exact expression (11), which is arguably simpler and more comprehensible.

3.4. *New FDR bounds for π_0-adaptive procedures.* For each $i \in \mathcal{H}_0$, let $\mathcal{D}_i$ be the distribution of $(p'_j, 1 \leq j \leq m)$, where

$$\begin{cases} p'_j = 0,\ j \in \mathcal{H}_1,\ p'_i = 1/(\ell+1); \\ p'_j,\ j \in \mathcal{H}_0 \backslash \{i\} \text{ are i.i.d. conditionally on } U \text{ with a common c.d.f. } F^U; \\ U = (U_1, \ldots, U_{\ell+1}) \text{ has i.i.d. } U(0,1) \text{ components,} \end{cases} \tag{13}$$

where F^U denotes the discrete c.d.f. $F^U(x) = (1 - U_{(\lfloor x(\ell+1) \rfloor + 1)}) \mathbb{1}\{1/(\ell+1) \leq x < 1\} + \mathbb{1}\{x \geq 1\}$, $x \in \mathbb{R}$, and $U_{(1)} > \cdots > U_{(\ell+1)}$ denote the order statistics of the vector U. Note that the distribution $\mathcal{D}_i$ only depends on i, m, ℓ and $\mathcal{H}_0$. The following general result holds.

THEOREM 3.6. *In the setting of Theorem 3.4, denote* $p = (p_i, 1 \leq i \leq m)$ *the family of empirical p-values defined in (10) and consider any function* $G : [0,1]^m \to (0, \infty)$ *that is coordinatewise nondecreasing. Then the procedure, denoted by* $R_{\alpha m / G(p)}$, *combining the BH algorithm at level* $\alpha m / G(p)$ *with these empirical p-values is such that, for any parameter* $P \in \mathcal{P}$,

$$\mathrm{FDR}(P, R_{\alpha m/G(p)}) \leq \alpha \sum_{i \in \mathcal{H}_0} \mathbb{E}_{p' \sim \mathcal{D}_i} \left(\frac{1}{G(p')} \right), \tag{14}$$

where $\mathcal{D}_i$ is defined by (13). In particular, this FDR expression holds for $R_{\alpha m/G(p)} = \mathit{AdaDetect}_{\alpha m/G(p)}$ *under Assumptions 1 and 3.*

Theorem 3.6 is proved in Section A.5 of the Supplementary Material. In a nutshell, the distribution $\mathcal{D}_i$ is a least favorable distribution for the FDR of the adaptive BH procedure applied to empirical p-values defined in (10). It can be seen as an adaptation of the classical leave-one-out technique for adaptive BH procedures; see Benjamini, Krieger and Yekutieli (2006) and Theorem 11 of Blanchard and Roquain (2009).

This result generalizes Theorem 6 of Bates et al. (2023) which only works for the Storey-BH procedure. Our proof technique is fundamentally different and works for a broad class of estimators of π_0. Applying Theorem 3.6 to the estimators defined in (5) and (6), we obtain the following result.

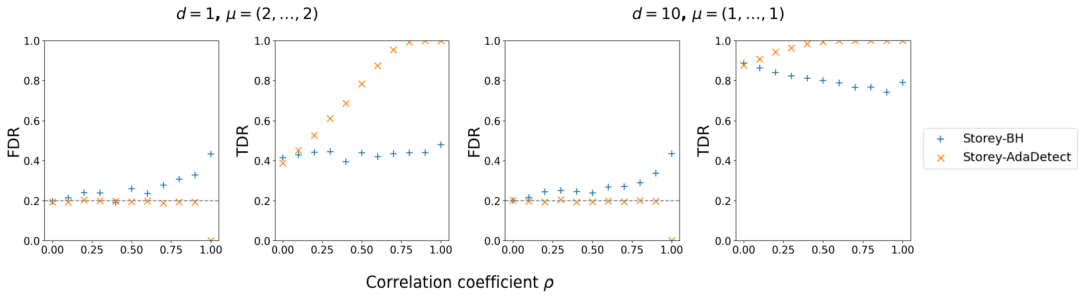

FIG. 3. *FDR and TDR for Storey-BH and Storey–AdaDetect (both with oracle test statistics/scores) in Example 3.1 with varying correlation* $\rho \in [0, 1]$. *The dimension* $d = 1$ *in the two left panels and* $d = 10$ *in the two right panels. In all settings,* $m = 100$, $n = \ell = 1000$, $\alpha = 0.2$, $\pi_0 = 0.9$, *and* $\lambda = 500/1001$.

COROLLARY 3.7. *Under Assumptions 1 and 3, the following hold*:

- *Storey–AdaDetect controls the FDR at level* α *for any* $\lambda = K/(\ell + 1)$ *and* $K \in \{2, \dots, \ell\}$.
- *Quantile–AdaDetect controls the FDR at level* α *for any* $k_0 \in \{1, \dots, m\}$.

The proof of Corollary 3.7 is presented in Section A.6 of the Supplementary Material. It bounds the RHS of (14) via combinatoric arguments. The result for Quantile–AdaDetect is novel. The result for Storey–AdaDetect was proved in Yang et al. (2021) for BONuS, with a different proof technique, in the special case where the scores are independent. Hence, we extend it to the exchangeable case.

To illustrate the robustness of π_0-adaptive AdaDetect under dependence, consider Example 3.1 with common alternative means $\mu_i \equiv \mu \in \mathbb{R}^d$ and a fixed score function $S_i = \mu^T Z_i$, $1 \le i \le n + m$. One alternative approach to Storey–AdaDetect is to apply the Storey-BH procedure on the marginal p-values $p_i = \bar{\Phi}(S_{n+i}/\|\mu\|)$, $1 \le i \le m$. Interestingly, Figure 3 shows that the Storey-BH procedure inflates the FDR substantially in the presence of high correlation while Storey–AdaDetect with $k = 0$ controls the FDR for any correlation ρ (as implied by Corollary 3.7). Hence, while Storey–AdaDetect is only based on an NTS without the knowledge of the true null distribution, it is more robust to dependence than Storey-BH that requires more information. Furthermore, Storey–AdaDetect is more powerful than Storey-BH because the effect of the common variable ξ is cancelled out in the calculation of empirical p-values.

REMARK 3.8. Assumptions 2 and 3 hold true in other contexts. For example, this is the case for LASSO-based scores in the Gaussian linear model where the design matrix has i.i.d. entries with a known distribution (Weinstein, Barber and Candès (2017)). Hence, the FDR bounds we developed also hold in those cases.

4. Constructing score functions. While any score function satisfying (8) can be used in AdaDetect, we discuss principles and various techniques to construct score functions that yield high power. Section 4.1 introduces the assumptions and notation. In Section 4.2, we show that the optimal score function is given by any monotone function of the ratio between the average density of novelties and the average density of all points. We proceed by discussing two methods to approach the optimal score based on direct density estimation (Section 4.3) and classification (Section 4.4). The latter is more scalable and flexible in the sense that it is able to wrap around any probabilistic classification algorithms. In Section 4.5, we discuss a cross-validation approach for hyper-parameter tuning and model selection without compromising the finite-sample FDR control.

4.1. *Assumptions and notation.* In this section, we make the following two assumptions.

ASSUMPTION 4. $Y_1, \dots, Y_n, X_1, \dots, X_m$ are mutually independent.

Given the setting of Section 1.1, we thus have under Assumption 4 that $(Y_1, \dots, Y_n, X_i, i \in \mathcal{H}_0)$ are i.i.d. $\sim P_0$ and independent of $(X_i, i \in \mathcal{H}_1)$ which are mutually independent.

ASSUMPTION 5. For each $i \in \{0\} \cup \mathcal{H}_1$, P_i has a positive density f_i w.r.t. a measure ν.

Let

$$f = \pi_0 f_0 + \pi_1 \bar{f}_1, \tag{15}$$

$$\bar{f}_1 = m_1^{-1} \sum_{i \in \mathcal{H}_1} f_i. \tag{16}$$

Under Assumptions 4 and 5, f_0 is the average density of $(Z_1, \dots, Z_k)$, $\bar{f}_1$ is the average alternative density, f is the average density of the test sample $(X_1, \dots, X_m)$. Similarly the average density of $(Z_{k+1}, \dots, Z_{n+m})$ is f_γ where

$$\gamma = \frac{m_1}{\ell + m}; \qquad f_\gamma = (1-\gamma) f_0 + \gamma \bar{f}_1 = \frac{\ell}{\ell + m} f_0 + \frac{m}{\ell + m} f. \tag{17}$$

Compared to f, the mixture f_γ is contaminated by more nulls, that is, $\pi_0 \le 1 - \gamma = \frac{\ell + m_0}{\ell + m}$. Lastly, we define the density ratio

$$r(x) = \frac{\pi_1 \bar{f}_1(x)}{f(x)}, \quad x \in \mathcal{Z}. \tag{18}$$

Note that $r(x) \in (0, 1)$ for ν-almost every $x \in \mathcal{Z}$ by Assumption 5.

4.2. *Optimal score function.* Recall that AdaDetect is equivalent to applying the counting knockoff on the scores which relies on an estimator $\widehat{\mathrm{FDP}}$ (Section 2.4). For each given t, when ℓ and m is large, $\widehat{\mathrm{FDP}}(t) \approx m \mathbb{P}_{S_i \sim P_0}(S_i \ge t) / \mathbb{E}[|R(t)|] \approx \mathbb{E}[|R(t) \cap \{k+1, \dots, n\}|] / \mathbb{E}[|R(t)|]$, where $R(t)$ is set of rejections at threshold t. The RHS is called the marginal FDR (mFDR), an error metric that is close to FDR when $|R(t)|$ is large and often used for asymptotic power analysis of FDR-controlling procedures (e.g., Lei and Fithian (2018), Sun and Cai (2007)). The following theorem derives the optimal score function among all procedures that reject hypotheses with S_i above some thresholds subject to mFDR control (see Rosset et al. (2022), Weinstein (2021) for results for FDR instead of mFDR).

THEOREM 4.1. *Assume Assumptions 4 and 5 hold. The likelihood ratio function $r(\cdot)$ defined in (18) is an optimal score function in the sense that the rejection set $R = \{i \in \{1, \dots, m\} : r(X_i) \ge c(\alpha)\}$, where $c(\alpha) \in (0, 1)$ is chosen such that mFDR$(R) = \alpha$ (assuming it exists), has a higher TPR than any rejection set $R' = \{i \in \{1, \dots, m\} : r'(X_i) \ge c'\}$ where $c' \in \mathbb{R}$ and $r' : \mathcal{Z} \mapsto \mathbb{R}$ is measurable with mFDR at most α.*

The proof can be found in Section B.1 of the Supplementary Material. Theorem 4.1 suggests the following oracle procedure.

DEFINITION 4.2. The oracle AdaDetect procedure, denoted by AdaDetect*, is defined as the AdaDetect procedure with the score function $r(\cdot)$ defined in (18).

Since AdaDetect is invariant under any strictly monotone transformation of the score function (see Remark 2.1), AdaDetect* can be realized as any AdaDetect procedure with a score function of the form

$$g^* = \Psi \circ r, \text{ for some increasing continuous } \Psi : (0,1) \to \mathbb{R}, \tag{19}$$

where Ψ could depend on unknown parameters. This is a crucial property of AdaDetect that enables flexible classification methods to construct score functions without concerning about the composition of nulls and novelties that may change the oracle score r.

Since r (or g^*) is unknown, the oracle procedure AdaDetect* is not directly accessible in practice. Our goal is to learn a g^* in the form of (7) that satisfies the constraint (8).

4.3. *Density estimation.* A first example of score function is built from density estimation. From (15) and (17), the following score:

$$g^*(x) = f_\gamma(x)/f_0(x) = 1 - \gamma/\pi_1 + (\pi_0\gamma/\pi_1)\big(1 - r(x)\big)^{-1} \tag{20}$$

is indeed of the form (19). A straightforward approach is to directly estimate the densities as follows:

- Estimate f_0 by a density estimator $\widehat{f_0}$ based on the sample $(Z_1, \dots, Z_k)$
- Estimate f_γ by a density estimator $\widehat{f_\gamma}$ based on the mixed sample $(Z_{k+1}, \dots, Z_{n+m})$ via a mixture estimation approach.
- Estimate $g^*(x)$ by $\widehat{g}(x) = \widehat{f_\gamma}(x)/\widehat{f_0}(x)$ assuming that $\widehat{f_0}(Z_i) > 0$.

Above, the density estimators can be either parametric or nonparametric. Both versions will be considered in the sequel (see Section 4.3 and the numerical experiments in Section 6). Note that Yang et al. (2021) applies this approach when f_0 is known.

4.4. *PU classification.* While density estimation is straightforward, it is not scalable when the dimension d is large; see the numerical experiments in Section 6 for an illustration. In this section, we consider a different strategy that estimate density ratios through probabilistic classification (e.g., Friedman (2003), Lei et al. (2021), Sugiyama, Suzuki and Kanamori (2012), Wang, Kaji and Rockova (2022)).

Define $(Z_1, \dots, Z_k)$ as the "positive sample" and the sample $(Z_{k+1}, \dots, Z_{n+m})$ as the "unlabeled sample", and let $(A_1, \dots, A_k) = (-1, \dots, -1)$ and $(A_{k+1}, \dots, A_{n+m}) = (1, \dots, 1)$ the corresponding labels. In this context, the classification task is typically referred to as the PU (positive unlabeled) classification, which is an active research area; see Calvo, Larranaga and Lozano (2007), Du Plessis, Niu and Sugiyama (2014), Guo et al. (2020), Ivanov (2020) among others and Bekker and Davis (2020) for a recent review. Here, we are considering a slightly different setting where the unlabeled samples are independent but not identically distributed.

Usually, the classifier is learned by empirical risk minimization (ERM) where the objective function is in the form of $\widehat{J}_\lambda(g) = \sum_{i=1}^{n+m} \lambda_{A_i} \ell(A_i, g(Z_i)) = \sum_{i=1}^{k} \ell(-1, g(Z_i)) + \lambda \sum_{i=k+1}^{n+m} \ell(1, g(Z_i))$, where $\ell : \{-1, +1\} \times \mathbb{R} \to \mathbb{R}_+$ is a loss function and $\lambda_a = \lambda \mathbb{1}\{a \geq 0\} + \mathbb{1}\{a \leq 0\}$ with $\lambda > 0$ measuring the relative cost misclassifying a positive sample to misclassifying an unlabeled sample. Here, g is a function that belongs to $\mathcal{G}$, a class of measurable functions from $\mathcal{Z}$ to $\mathbb{R}$ and the classifier corresponds to the sign of g. Typical choices of the loss function include the hinge loss $\ell(a, u) = 0.5(1 - au)_+$ and the cross entropy loss $\ell(a, u) = -\log(1 - u)\mathbb{1}\{a = -1\} - \log(u)\mathbb{1}\{a = +1\}$. The population objective function is given by

$$J_\lambda(g) = \mathbb{E}\widehat{J}_\lambda(g) = k\mathbb{E}_{Z \sim f_0} \ell\big(-1, g(Z)\big) + \lambda(\ell + m)\mathbb{E}_{Z \sim f_\gamma} \ell\big(1, g(Z)\big), \tag{21}$$

where f_γ is defined in (17). The following result shows that the minimizer of (21) over all measurable functions yields an optimal score in the form of (19) when the loss function ℓ is appropriately chosen.

LEMMA 4.3. *Let $g^\sharp$ denote the minimizer of (21) over all measurable functions.*

(i) *When $\ell(\cdot,\cdot)$ is the hinge loss, assuming that the set $\{x \in \mathcal{Z} : f_\gamma(x) = cf_0(x)\}$ is of ν-measure zero for any $c > 0$, where ν is defined in Assumption 5,*

$$g^\sharp(x) = \operatorname{sign}\left(\frac{\lambda(\ell+m)}{k}\frac{f_\gamma(x)}{f_0(x)} - 1\right) = \operatorname{sign}\left(\frac{\lambda\ell}{k} + \frac{\lambda m_0}{k}\big(1 - r(x)\big)^{-1} - 1\right),$$

and the minimum is unique ν-almost everywhere.

(ii) *When $\ell(\cdot,\cdot)$ is the cross entropy,*

$$g^\sharp(x) = \frac{\lambda(\ell+m)f_\gamma(x)}{\lambda(\ell+m)f_\gamma(x) + kf_0(x)} = \left(1 + \left\{\frac{\lambda\ell}{k} + \frac{\lambda m_0}{k}\big(1 - r(x)\big)^{-1}\right\}^{-1}\right)^{-1},$$

and the minimum is unique ν-almost everywhere.

The proof is presented in Section B.2 of the Supplementary Material. Clearly, $g^\sharp$ is an optimal score function in the form of (19) with the cross-entropy loss but not so with the hinge loss because the sign function is not strictly monotone. For cross-entropy loss, when $\lambda = 1$,

$$g^\sharp(x) = \frac{\frac{\ell+m}{n+m}f_\gamma(x)}{\frac{\ell+m}{n+m}f_\gamma(x) + \frac{k}{n+m}f_0(x)}, \tag{22}$$

which can be roughly interpreted as the posterior probability to be in class 1.

In practice, it is computationally infeasible and statistically inefficient to optimize over all measurable functions. Instead, we often choose a function class $\mathcal{G}$ and estimate the score function by

$$\widehat{g} \in \operatorname*{arg\,min}_{g\in\mathcal{G}} \widehat{J}_\lambda(g). \tag{23}$$

By construction, $\widehat{J}_\lambda(g)$ is invariant to permutations of $(Z_{k+1}, \ldots, Z_{n+m})$, $\widehat{g}$ always satisfies the condition (8). When $\mathcal{G}$ has low complexity, we should expect $\widehat{g} \approx g^\sharp_{\mathcal{G}}$ where

$$g^\sharp_{\mathcal{G}} \in \operatorname*{arg\,min}_{g\in\mathcal{G}} J_\lambda(g). \tag{24}$$

On the other hand, when $\mathcal{G}$ is sufficiently rich, we can expect $g^\sharp_{\mathcal{G}} \approx g^\sharp$. In summary, when the function class $\mathcal{G}$ and the loss function $\ell(\cdot,\cdot)$ are chosen appropriately, $\hat{g} \approx g^\sharp_{\mathcal{G}} \approx g^\sharp$, which is an optimal score function.

We illustrate the roles of function classes and loss functions in a simple setting where the positive class and the unlabeled class are generated from two gaussian distributions with dimension 1 or 2. The results are presented in Figure 4, with each row corresponding to a data-generating process. For all settings, the first panel displays the null and alternative distributions and the second panel displays the distributions of the positive and unlabeled classes. In all settings, we plot $\hat{g}$ and $g^\sharp$ for hinged loss (SVM) and cross-entropy losses with two function classes, an inaccurate one (Logistic Regression) and an accurate one (Neural Networks). In the two-dimensional settings, we display the functions by contour plots. For instance, for the cross entropy loss, we can observe the NN function class outperforms the logistic function class for approximating $g^\sharp$.

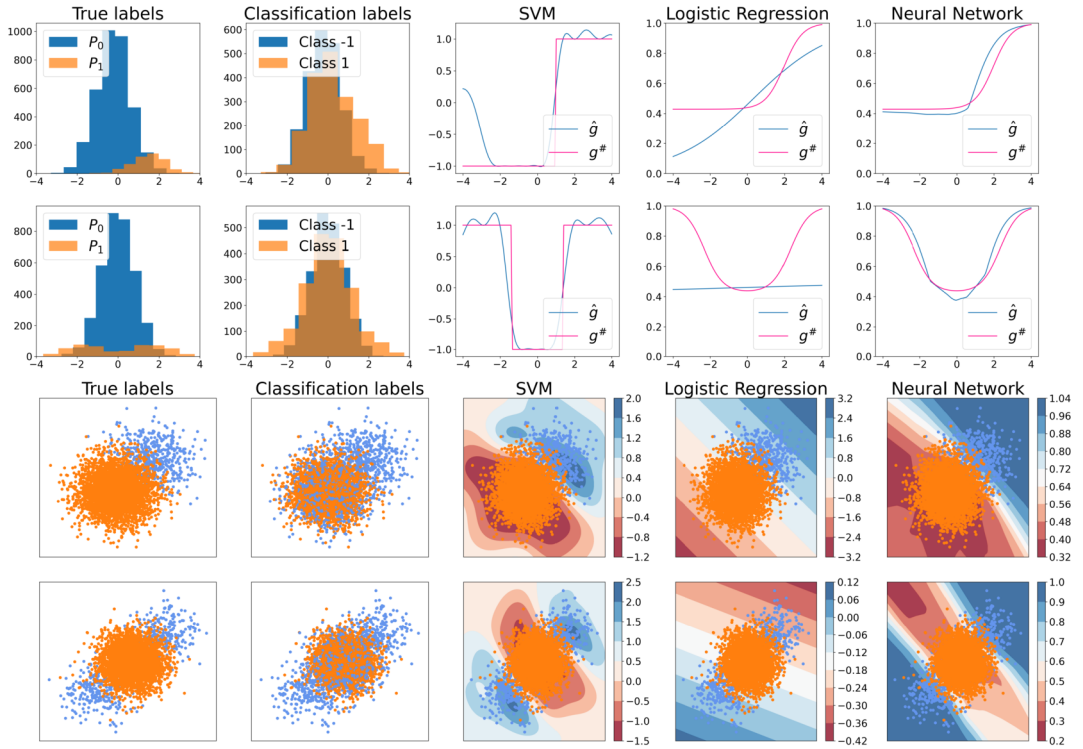

FIG. 4. *Plot of g^* and $g^\sharp$ in different settings (rows) with different loss functions $\ell(\cdot,\cdot)$ and function classes $\mathcal{G}$ (with default parameters in* `scikit-learn`*). In all settings, $m = 1000$, $m_0 = 500$, $m_1 = 500$, $n = 3000$, and $k = 2000$. The top two rows correspond to $d = 1$ and the bottom two rows correspond to $d = 2$. In all cases, $P_0 = \mathcal{N}(0, I_d)$. For the first and third rows, $P_1 = \mathcal{N}((2,\dots,2), I_d)$ (one-sided alternatives); for the second and fourth rows, $P_1 = 0.5\mathcal{N}((2,\dots,2), I_d) + 0.5\mathcal{N}((-2,\dots,-2), I_d)$ (two-sided alternatives).*

In conclusion, both the loss function $\ell(\cdot,\cdot)$ and the function class $\mathcal{G}$ are pivotal. Among the two loss functions we discuss, the cross entropy loss with a sufficiently rich function class (e.g., fully-connected neural networks) is particularly suitable for AdaDetect in that it is computationally feasible and approximately optimal. In contrast to the classification literature, hinged loss is undesirable for our purpose since the estimator does not converge to an optimal score.

4.5. *AdaDetect with cross-validation.* In previous sections, we focus on a single score function. Nevertheless, most density estimation and classification algorithms involve hyper-parameters that require data-driven tuning to maximize the power. Examples include the bandwidth for kernel density estimation, the maximum depth for random forests, the width and number of hidden layers for neural networks, and the numerical algorithm to optimize the loss.

Formally, we assume the researcher has a class of candidate score functions $\{g_\upsilon, \upsilon \in \mathcal{U}\}$ indexed by the hyper-parameter υ. The goal is to choose $\hat{\upsilon}$ based on data and use $g_{\hat{\upsilon}}$ as the score function without breaking the FDR guarantee. By Theorem 3.4, the FDR is controlled so long as $g_{\hat{\upsilon}}$ satisfies the condition (8). Motivated by the "double BONuS" procedure proposed in Yang et al. (2021), we propose the following version of AdaDetect with cross-validation, which we abbreviate as the AdaDetect cv procedure.

1. Split $(Y_1,\dots,Y_k)$ further into two parts $(Y_1,\dots,Y_s)$ and $(Y_{s+1},\dots,Y_k)$ for some $s < k$.
2. Generate a class of score functions g_υ that satisfy a stronger condition than (8):

$$g_\upsilon\big(z,(z_1,\dots,z_s),(z_{\pi(s+1)},\dots,z_{\pi(n+m)})\big) = g\big(z,(z_1,\dots,z_s),(z_{s+1},\dots,z_{n+m})\big).$$

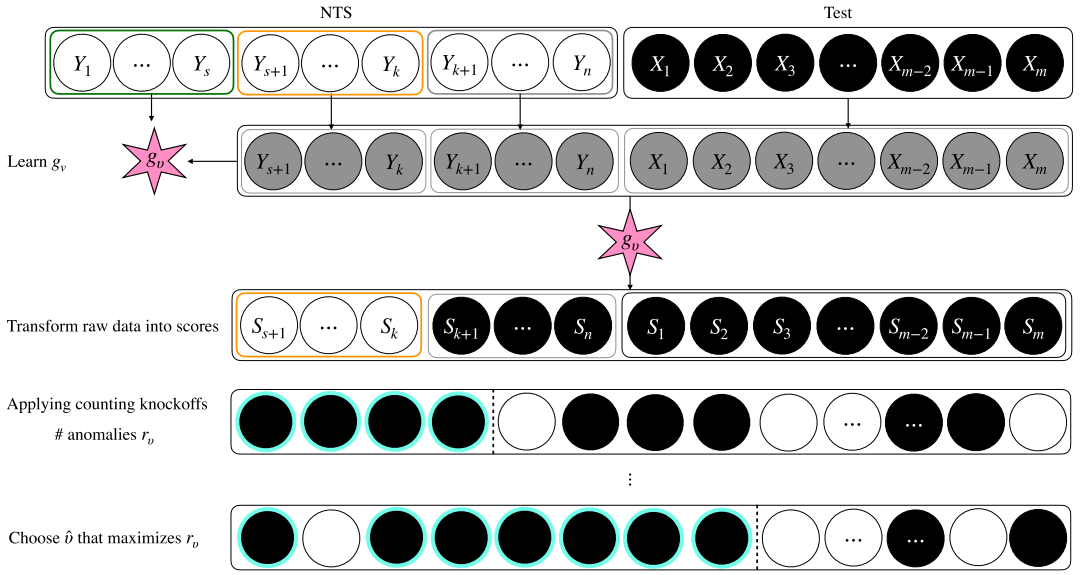

FIG. 5. *The pipeline to compute score function $g_{\hat{\upsilon}}$ for AdaDetect cv. Same pictural conventions as in Figure 2.*

3. For each g_υ, apply AdaDetect with $(Y_{k+1}, \dots, Y_n, X_1, \dots, X_m)$ being the test sample and $(Y_1, \dots, Y_k) = (Y_1, \dots, Y_s; Y_{s+1}, \dots, Y_k)$ being the NTS. Denote by r_υ the number of rejections.
4. Choose $\hat{\upsilon} \in \arg\max_{\upsilon \in \mathcal{U}} r_\upsilon$.
5. Apply AdaDetect with score function $g_{\hat{\upsilon}}$ to the original problem (with $(X_1, \dots, X_m)$ being the test sample and $(Y_1, \dots, Y_n) = (Y_1, \dots, Y_k; Y_{k+1}, \dots, Y_n)$ being the NTS).

The pipeline to compute $g_{\hat{\upsilon}}$ is illustrated in Figure 5. By definition, each g_υ is invariant to permutation of $(Y_{s+1}, \dots, Y_n, X_1, \dots, X_m)$ and hence invariant to permutation of the mixed sample $(Y_{k+1}, \dots, Y_n, X_1, \dots, X_m)$. Thus, r_υ is also invariant to $(Y_{k+1}, \dots, Y_n, X_1, \dots, X_m)$, implying that $\hat{\upsilon}$ is so as well. As a result, $g_{\hat{\upsilon}}$ satisfies the condition (8). Therefore, the results in Section 3 all carry over to the AdaDetect cv procedure.

In principle, we can use any other objective function that is invariant to the mixed sample than the number of rejections r_υ. Nonetheless, r_υ tends to be a good proxy for the number of rejections in the last step and hence a better objective to optimize than the indirect ones like classification accuracy.

REMARK 4.4. When fitting the hyper-parameter υ, the sample sizes $s, k-s, \ell+m, \ell$ do not maintain the same proportions as the original sizes k, ℓ, m. Our recommendation, following the guidelines in Remark 2.2, is to choose s such that $k-s$ is of the same order as $\ell+m$ and s is of the same order as m (e.g., $\ell = m$, $s = 3m$, $k = 4m$).

REMARK 4.5. The cross-validation can rule out overfitted models that performs well in training data but does poorly out of sample. By including nonsophisticated baseline models that likely generalize, the power of AdaDetect becomes less sensitive to overfitting of other complicated models or other failure modes that we have yet discovered. For example, the researcher can always add a nonadaptive score that cannot incur overfitting and might be underpowered.

5. Power results. In this section, we analyze the power of AdaDetect with appropriately chosen score functions. Throughout this section, we assume that the measurements take values in $\mathcal{Z} = \mathbb{R}^d$. We start in Section 5.1 with a specific score function given by a constrained

empirical risk minimizer (ERM) with the 0-1 loss and show it is as powerful as the classification approach based on the optimal score functions defined in (19) when the function class is sufficiently flexible and up to asymptotically vanishing remainder terms. In Section 5.2, we turn to a general estimated score function that is close to an oracle (deterministic) score function on all measurements in the mixed sample. When the latter is sufficiently smooth, we show that AdaDetect with the estimated score function is as efficient as AdaDetect with the oracle score function, up to explicit remainder terms that are asymptotically vanishing.

5.1. *A constrained ERM score function.* For the convenience of theoretical analysis, we study a constrained empirical risk minimizer (ERM) score function with 0-1 loss motivated by the Neyman–Pearson (NP) formulation of classification problems given in Blanchard, Lee and Scott (2010); see also Cannon et al. (2002) and Scott and Nowak (2005). Define

$$\hat{R}_0(g) = k^{-1}\sum_{i=1}^{k} \mathbb{1}\{g(Z_i) \geq 0\}, \qquad R_0(g) = \mathbb{E}\hat{R}_0(g) = \mathbb{P}_{Z\sim f_0}\big(g(Z) \geq 0\big),$$

$$\hat{R}_\gamma(g) = (m+\ell)^{-1}\sum_{i=k+1}^{n+m} \mathbb{1}\{g(Z_i) < 0\}, \tag{25}$$

$$R_\gamma(g) = \mathbb{E}\hat{R}_\gamma(g) = (1-\gamma)\big(1 - R_0(g)\big) + \gamma R_1(g),$$

$$R_1(g) = \mathbb{P}_{Z\sim \bar{f}_1}\big(g(Z) < 0\big),$$

where γ, f_0, and $\bar{f}_1$ are defined in (17). We consider a function class $\mathcal{G}$ with a finite Vapnik–Chervonenkis (VC) dimension $V(\mathcal{G})$ (Vapnik (1998)) and the following constrained ERM score function:

$$\hat{g} \in \arg\min_{g\in\mathcal{G}}\big\{\hat{R}_\gamma(g) : \hat{R}_0(g) \leq \beta + \epsilon_0\big\}, \tag{26}$$

for some $\epsilon_0 > 0$, as well as its population version

$$g_{\mathcal{G}}^\sharp \in \arg\min_{g\in\mathcal{G}}\big\{R_\gamma(g) : R_0(g) \leq \beta\big\}. \tag{27}$$

THEOREM 5.1. *Consider the setting of Theorem 4.1. Assume $\alpha, \beta \in (0,1)$, $k, m_1 \geq 1$, and $g_{\mathcal{G}}^\sharp$, defined in (27), satisfies $R_0(g_{\mathcal{G}}^\sharp) = \beta$. Fix any $\delta \in (0, 1/2)$. Then there exist constants $C, C' > 0$ that only depend on δ such that, if*

$$\epsilon_0 = C\sqrt{\frac{V(\mathcal{G}) + \log(1/\delta)}{k}}, \qquad \Delta = C'\gamma^{-1}\sqrt{\frac{V(\mathcal{G}) + \log(1/\delta)}{k \wedge \ell}}, \tag{28}$$

where γ is defined in (17), the following results hold:

(i) *With probability at least $1-\delta$, $R_0(\hat{g}) \leq \beta + \Delta$ and $R_1(\hat{g}) \leq R_1(g_{\mathcal{G}}^\sharp) + \Delta$.*

(ii) *Let $M = \lceil (1 - R_1(g_{\mathcal{G}}^\sharp) - \Delta)m_1 \rceil$. Assume that*

$$1 - R_1\big(g_{\mathcal{G}}^\sharp\big) \geq \big(1 + \alpha^{-1}\big)\Delta, \qquad \ell \geq \frac{2m}{\alpha M}, \qquad \beta \leq \frac{0.4\alpha M}{m}. \tag{29}$$

Then, with probability at least $1-\delta$,

$$\mathit{AdaDetect}_\alpha \supset \big\{i \in \{1,\dots,m\} : \hat{g}(X_i) \geq 0\big\}, \tag{30}$$

$$|\mathit{AdaDetect}_\alpha \cap \mathcal{H}_1|/m_1 \geq 1 - R_1\big(g_{\mathcal{G}}^\sharp\big) - \Delta, \tag{31}$$

where $\mathit{AdaDetect}_\alpha$ denotes the rejection set of AdaDetect with score function $\hat{g}$.

The proof of Theorem 5.1 is presented in Section C.1 of the Supplementary Material. The idea is to show that there are many alternatives with a nonnegative score, while there are only a few true nulls with nonnegative scores. This yields small empirical p-values for hypotheses with nonnegative scores, which implies that the procedure AdaDetect$_\alpha$ detects these nonnegative scores; see Lemma D.5 of the Supplementary Material.

Theorem 5.1(i) shows that $\hat{g}$ has a similar classification accuracy to $g^\sharp_{\mathcal{G}}$ on both the NTS and mixed sample. It is analogous to Theorem 2 in Blanchard, Lee and Scott (2010), though Blanchard, Lee and Scott (2010) considers a different setting where the proportion of nulls is random. Theorem 5.1(ii) entails that, with high probability, all hypotheses with nonnegative scores will be rejected and the power of AdaDetect with $\hat{g}$ is nearly as large as the power of the classification procedure given by $g^\sharp_{\mathcal{G}}$.

Note that the Lagrangian form of the above problem is in the form of the weighted loss defined in (21). Thus, by Lemma 4.3(i), there exists $\lambda_\beta > 0$ such that $g^\sharp_{\mathcal{G}}(x) = g^*(x) = \frac{\lambda_\beta(\ell+m)}{k}\frac{f_\gamma(x)}{f_0(x)} - 1$, if the constraint is feasible and $\mathcal{G}$ is sufficiently rich to include the above function. Above, $g^*(x)$ satisfies (19) and hence yields the optimal power. If we define $b = R_1(g^\sharp_{\mathcal{G}}) - R_1(g^*)$ as the bias due to the constraint, Theorem 5.1(ii) implies that, with probability $1-\delta$, the power of AdaDetect with $\hat{g}$ is at most $\Delta + b$ below the optimal power. Thus, the function class $\mathcal{G}$ incurs a tradeoff that a richer class yields a smaller b but a larger Δ and vice versa.

Here, we aim at making $\mathcal{G}$ as flexible as possible while ensuring $\Delta = o(1)$. When k, ℓ, and m are of the same order and δ is a constant, $\Delta \asymp \frac{m}{m_1}\sqrt{\frac{V(\mathcal{G})}{m}}$. Hence, $\Delta = o(1)$ if

$$\frac{m_1}{m} \gg \sqrt{\frac{V(\mathcal{G})}{m}}. \tag{32}$$

For illustration, consider the class $\mathcal{G}_{N,L,s}$ of ReLU feed forward neural networks with fixed topology, maximum width $N \asymp m^c$ ($c \in (0,1)$), depth $L \asymp \log m$ and sparsity $s \asymp N \log m$. Bartlett et al. (2019) show that $V(\mathcal{G}_{N,L,s}) \le 2sL\log(4eN) \lesssim m^c(\log m)^3$. Hence, condition (32) reads in this case $m_1/m \gg m^{\frac{c-1}{2}}(\log m)^{3/2}$. This implies that $\Delta = o(1)$ unless the novelties are too sparse. On the other hand, given the approximation ability of class of neural networks, we should expect $1 - R_1(g^\sharp_{\mathcal{G}_{N,L,s}}) \approx 1 - R_1(g^*)$. Thus, Theorem 5.1(ii) implies the resulting score function is nearly optimal.

REMARK 5.2 (Choice of β). Since AdaDetect$_\alpha$ controls FDR at level α, it is necessary to impose an upper bound on β in Theorem 5.1(ii). Roughly speaking, our condition on β guarantees that the classifier $g^\sharp_{\mathcal{G}}$ controls the FDR at level α, up to remainder terms.

REMARK 5.3. The condition on ℓ in (29) is needed to ensure that the minimum value $1/(1+\ell)$ that p-values can take is sufficiently small so that the BH procedure can reject. A similar condition was introduced in Mary and Roquain (2022); see also Remark 2.2.

5.2. *General score functions.* Now we move to general score functions. Let g^* be any measurable function $\mathbb{R}^d \to \mathbb{R}$ in the form of (19) and

$$\overline{G}_0(s) = \mathbb{P}_{X\sim P_0}\big(g^*(X) \ge s\big), \quad s \in \mathbb{R}; \tag{33}$$

$$\zeta_r(\eta) = \max_{u\in[\alpha(r\vee 1)/m,\alpha]}\left\{\frac{\overline{G}_0(\overline{G}_0^{-1}(u) - 2\eta) - u}{u}\right\}, \quad \eta > 0, r \in \{0,\dots,m\}. \tag{34}$$

Here, $\zeta_r(\cdot)$ measures the local fluctuation of $\overline{G}_0$. We suppress the dependence on α and m to ease notation. Furthermore, consider any data-driven score function $\hat{g}$ satisfying the condition (8) and let

$$\hat{\eta} = \max_{k+1\le i\le n+m} \big|\hat{g}\big(Z_i; (Z_1,\dots,Z_k), (Z_{k+1},\dots,Z_{n+m})\big) - g^*(Z_i)\big|, \tag{35}$$

which measures the maximal discrepancy of scores in the mixed sample. In the following, $\mathrm{AdaDetect}_\alpha$ denotes the procedure with score function $\hat{g}$ and $\mathrm{AdaDetect}^*_\alpha$ denotes the procedure with score function g^*.

THEOREM 5.4. *Fix any $r \in \{0,\dots,m\}$ and let $\mathcal{R} = \{|\mathit{AdaDetect}^*_\alpha| \ge r\}$. Assume $m \ge 1$, $\ell, k \ge 0$, $n = k + \ell \ge 1$, and $\overline{G}_0$ (33) is continuous and strictly decreasing. Under Assumptions 4 and 5, for any $\delta, \eta \in (0,1)$ such that $(\ell+1)\delta\alpha(r \vee 1)/m \ge 2$,*

$$\mathbb{P}\big(\mathcal{R} \cap \big\{\mathit{AdaDetect}^*_\alpha \subset \mathit{AdaDetect}_{\alpha'}\big\}^c\big) \le \mathbb{P}(\hat{\eta} > \eta) + 2me^{-(3/28)(\ell+1)\delta^2\alpha(r\vee 1)/m}, \tag{36}$$

where $\alpha' = \alpha(1+3\delta)(1+\zeta_r(\eta))$ and $\zeta_r(\cdot)$ and $\hat{\eta}$ are defined in (34) and (35), respectively. Furthermore, (36) is also true with $\mathit{AdaDetect}^*_\alpha$ replaced by BH^*_α, the BH algorithm applied to the oracle p-values $p^*_i = \overline{G}_0(g^*(X_i))$, $1 \le i \le m$.

The proof is presented in Section C.2 of the Supplementary Material. The condition $(\ell+1)\delta\alpha(r\vee 1)/m \ge 2$ is analogous to the one studied (Mary and Roquain (2022)) for fixed score functions. When we choose $r = 0$, we have $\mathbb{P}(\mathcal{R}) = 1$ and thus (36) implies

$$\mathbb{P}\big(\mathrm{AdaDetect}^*_\alpha \subset \mathrm{AdaDetect}_{\alpha'}\big) \ge 1 - \mathbb{P}(\hat{\eta} > \eta) - 2me^{-(3/28)(\ell+1)\delta^2\alpha/m},$$

for any δ with $(\ell+1)\delta\alpha/m \ge 2$. If $\ell/m >> \log m$, choosing $\delta = o(1)$ such that $(\ell+1)\delta^2\alpha/m >> \log m$ and η such that $\mathbb{P}(\hat{\eta} \ge \eta) = o(1)$, we have $\mathbb{P}(\mathrm{AdaDetect}^*_\alpha \subset \mathrm{AdaDetect}_{\alpha'}) = 1 - o(1)$, where $\alpha' = \alpha(1+\zeta_0(\eta))(1+o(1))$. Thus, when $\zeta_0(\eta)$ is small, we show that AdaDetect with the estimated score function and slight inflation of the target level is strictly more powerful than its oracle version.

In general, when $|\mathrm{AdaDetect}^*_\alpha|$ is larger with high probability, we can choose a larger r to relax the condition on δ, reduces $\zeta_r(\eta)$ (and hence α'), and improve the RHS of (36). In particular, we can set r appropriately to obtain the following result on the asymptotic TDR.

COROLLARY 5.5. *Consider the setting of Theorem 5.4. Fix any $\epsilon > 0$. Assume $m_1 \ge 1$ and $(\ell+1)\delta\alpha\lceil m_1\epsilon\rceil/m \ge 2$. Then*

$$\begin{aligned}\mathrm{TDR}(\mathit{AdaDetect}_{\alpha'}) \ge{}& \mathrm{TDR}\big(\mathit{AdaDetect}^*_\alpha\big)\\ &- \mathbb{P}(\hat{\eta} > \eta) - 2me^{-(3/28)(\ell+1)\delta^2\alpha\lceil m_1\epsilon\rceil/m} - \epsilon,\end{aligned} \tag{37}$$

where $\alpha' = \alpha(1+3\delta)(1+\zeta_{\lceil m_1\epsilon\rceil}(\eta))$. In particular, if there exist sequences $\delta = \delta(k,\ell,m,m_1)$, $\epsilon = \epsilon(k,\ell,m,m_1)$, and $\eta = \eta(k,\ell,m,m_1)$ such that, as ℓ, m, m_1 tend to infinity,

$$\delta, \epsilon \to 0,\ \ell\delta^2\epsilon m_1/m \to \infty,\ \mathbb{P}(\hat{\eta} > \eta) \to 0 \text{ and } \zeta_{\lceil m_1\epsilon\rceil}(\eta) \to 0, \tag{38}$$

then

$$\liminf_{\ell,m,m_1} \big\{\mathrm{TDR}(\mathit{AdaDetect}_{\tilde{\alpha}}) - \mathrm{TDR}\big(\mathit{AdaDetect}^*_\alpha\big)\big\} \ge 0, \quad \text{for any fixed } \tilde{\alpha} > \alpha. \tag{39}$$

Furthermore, these results hold with $\mathit{AdaDetect}^*_\alpha$ replaced by BH^*_α defined in Theorem 5.4.

TABLE 2
Summary of datasets

	Shuttle	Credit card	KDDCup99	Mammography	Musk	MNIST
Dimension d	9	30	40	6	166	28×28
Feature type	Real	Real	Real, categorical	Real	Real	Real
Inliers	45586	284315	47913	10923	5581	5842
Novelties	3511	492	200	260	1017	5949

The proof is presented in Section C.3 of the Supplementary Material. Corollary 5.5 shows that AdaDetect is nearly as powerful as the oracle version, as well as the BH procedure with the optimal score.

Now we discuss the choice of η. Note that η is a parameter that only shows up in the bound but not in the algorithm. It incurs a tradeoff that a larger η would improve the tail bound by decreasing $\mathbb{P}(\hat{\eta} > \eta)$ but inflate α' through increasing $\zeta_r(\eta)$. Ideally, we would want η so that $\mathbb{P}(\hat{\eta} > \eta)$ and $\zeta_r(\eta)$ are both negligible. For illustration, assume

$$\mathbb{P}\big(\hat{\eta} > (n+m)^{-\kappa}\big) = o(1), \tag{40}$$

for some $\kappa \in (0, 1/2)$ and $\zeta_r(\eta) \lesssim \eta/\gamma$, where γ is defined in (17). In this case, $\mathbb{P}(\hat{\eta} > \eta)$ and $\zeta_r(\eta)$ are both $o(1)$ if $(n+m)^{-\kappa} = o(\gamma) = o(m_1/(m+\ell))$. Again, this would hold unless the novelties are too sparse.

We show in Lemma E.4 of the Supplementary Material that the score function given by density estimation satisfies (40) under regularity conditions. Another example is given by Theorem 3.2 in Audibert and Tsybakov (2007) in the case where g^* is the posterior probability under a different setting; see Section E.3 of the Supplementary Material). For $\zeta_r(\eta)$, we provide bounds in Section E.1 of the Supplementary Material for two examples. In the Gaussian example, we show that $\zeta_r(\eta) \lesssim \eta/\gamma$, where γ is defined in (17).

REMARK 5.6. Theorem 3 in Yang et al. (2021) provides another asymptotic power analysis showing that the symmetric difference between the rejection set for the data-driven score function and its oracle version has a size $o_P(m)$. Unlike Theorem 5.4 and Corollary 5.5, it does not have implications when the oracle procedure can only reject $o_P(m)$ hypotheses, as in the case where $m_1/m = o(1)$.

6. Experiments. In this section, we examine the performance of AdaDetect on real data (experiments on simulated data are postponed to Section F.4 of the Supplementary Material). Each dataset contains measurements that are labeled as either typical or novelty. We summarize the datasets in Table 2 and provide further details in Section F.1 of the Supplementary Material.

We apply AdaDetect with various score functions, including the density estimation-based score (`AdaDetect parametric` and `AdaDetect KDE`), the PU classification-based score (`AdaDetect SVM`, `AdaDetect RF`, `AdaDetect NN`, and `AdaDetect NN cv`). We also include the conformal novelty detection procedures proposed by Bates et al. (2023) (`CAD SVM` and `CAD IForest`). Note that both `CAD SVM` and `CAD IForest` are instances of AdaDetect with one-class classification-based scores. For all our experiments, we use the Python package `scikit-learn` for Expectation Maximization (EM) algorithm, kernel density estimation, random forests, and neural networks, with the default hyper-parameters from the packages unless otherwise specified (a full description of these methods is provided in Section F.2 of the Supplementary Material). For the MNIST dataset,

we consider two more methods based on a convolutional neural network (CNN), with two convolution layers and one fully connected layer. The first method `CAD SVDD CNN` is the conformal novelty detection procedure of Bates et al. (2023) with a special one-class classifier, given by the Support Vector Data Description (SVDD) method introduced in Ruff et al. (2018) used with a family of functions given by the CNN. The second method is AdaDetect with the two-class classifier based on the CNN, denoted by `AdaDetect CNN`.

We construct test samples and null training samples by subsampling the dataset with $n = 5000$, $m = 1000$, and a fixed null proportion $\pi_0 = m_0/m = 0.9$. For AdaDetect, we choose $k = 4m$, $\ell = m$, $s = 3m$ for cross-validation, and the target level $\alpha = 0.1$. The FDR and TDR for the methods are evaluated by using 100 runs and the results are reported in Table 3. As expected, all methods control the FDR. Compared to Bates et al. (2023), AdaDetect with classification-based scores substantially boosts the power because it incorporates the novelties in learning the score function. Overall, the best performing method is `AdaDetect RF`, with `AdaDetect NN` (possibly cross-validated) coming in second. `AdaDetect CNN` is particularly efficient on the classical MNIST dataset, which is unsurprising because CNN-type classifiers are appropriate for such an image dataset (Goodfellow, Bengio and Courville (2016)). We however note that the one-class classifier based upon CNN behaves poorly, which shows that two-class classification is the key for the power boost instead of the better representation given by CNN. In addition, further comparisons are made in Section F.3 of the Supplementary Material for other values of n, m, m_1 in more challenging regimes and the conclusions are qualitatively similar.

To conclude, if a classification method is expected to distinguish between typical and anomalous measurements, combining it with AdaDetect is expected to achieve high power without threatening FDR control.

7. An astronomy application. In this section, we apply AdaDetect to detect variable stars using the Sloan Digital Sky Survey (Ivezić et al. (2005)), a large labeled dataset with $92,658$ nonvariable (null) and 483 variable (novelties) stars. Each star is encoded as a 4-dimensional vector containing the star's flux in specific bands (colors) of the visible light. This dataset is particularly appealing for demonstrating our method. First, the two classes occupy similar regions in the considered color space, with slight overlap leading to complex decision boundaries. Second, the large number of nonvariable stars allows us to vary the size of the NTS in a large range in the Monte Carlo simulations. Third, this dataset has been extensively studied by astronomers and has become a standard for benchmarking classification methods (see Chapter 9 of Ivezić et al. (2019)). Lastly, we can compute the achieved FDR and TDR for any novelty detection method based on the labeled data.

For each experiment, we sample n nonvariable stars as the NTS along with m_1 variable stars and $m_0 = m - m_1$ additional nonvariable stars as the test sample. We set $m = 100$ and vary n and m_1 across experiments. We apply AdaDetect with Kernel Density Estimation (KDE), Random Forest (RF), and Neural Networks (NN). For comparison, we also include two Empirical BH procedures (Mary and Roquain (2022)), which are special cases of AdaDetect with nonadaptive scores as the squared ℓ_2 norm of the demeaned vectors, where the mean is calculated on all nulls outside of the NTS ("Emp BH full") and on the NTS ("Emp BH current"), respectively. The "Emp BH current" method is closer to the current practice, though it is not granted to control the FDR since the score function does not satisfy (8). In addition, we apply the Empirical BH procedure without demeaning the data as well as the SC procedure with estimated local FDR. Neither detects any novelties so we will not report them.

Figure 6 presents the results for $m_1 = 50$ and varying n with target FDR level $\alpha = 0.05$. The FDR and TDR are calculated based on 100 Monte Carlo simulations. To aid visualization, we represents the uncertainty by a shaded area whose width is equal to the standard

TABLE 3

FDR (top) and TDR (bottom) of AdaDetect with different score functions on real datasets. The target FDR level is $\alpha = 0.1$. We report the mean value and the standard deviation (in brackets) over 100 runs. The two best-performing methods are highlighted in bold

	Shuttle	Credit card	KDDCup99	Mammography	Musk	MNIST
			FDR			
CAD SVM	0.04 (0.08)	0.00 (0.00)	0.00 (0.000)	0.05(0.10)	0.00 (0.00)	0.00 (0.00)
CAD IForest	0.10 (0.07)	0.09 (0.06)	0.08 (0.07)	0.05 (0.09)	0.00 (0.00)	0.00 (0.00)
AdaDetect parametric	0.01 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AdaDetect KDE	0.07 (0.07)	0.05 (0.08)	0.02 (0.06)	0.08 (0.07)	0.02 (0.08)	0.00 (0.00)
AdaDetect SVM	0.08 (0.04)	0.07 (0.05)	0.07 (0.05)	0.07 (0.06)	0.08 (0.06)	0.02 (0.03)
AdaDetect RF	0.08 (0.04)	0.09 (0.04)	0.08 (0.04)	0.04 (0.10)	0.03 (0.06)	0.03 (0.07)
AdaDetect NN	0.07 (0.05)	0.09 (0.04)	0.06 (0.07)	0.09 (0.06)	0.06 (0.09)	0.06 (0.08)
AdaDetect cv NN	0.08 (0.04)	0.09 (0.05)	0.08 (0.11)	0.08 (0.05)	0.06 (0.08)	0.01 (0.03)
CAD SVDD + CNN	–	–	–	–	–	0.03 (0.14)
AdaDetect CNN	–	–	–	–	–	0.09 (0.05)
			TDR			
CAD SVM	0.10 (0.18)	0.00 (0.00)	0.00 (0.00)	0.03 (0.06)	0.00 (0.00)	0.00 (0.00)
CAD IForest	0.45 (0.09)	0.39 (0.22)	0.56 (0.35)	0.05 (0.09)	0.00 (0.00)	0.00 (0.00)
AdaDetect parametric	0.02 (0.07)	0.00 (0.00)	0.00 (0.00)	0.07 (0.09)	0.00 (0.00)	0.00 (0.00)
AdaDetect KDE	0.44 (0.33)	0.12 (0.20)	0.11 (0.24)	0.22 (0.17)	0.02 (0.06)	0.00 (0.00)
AdaDetect SVM	**0.85 (0.17)**	0.68 (0.28)	0.66 (0.32)	0.43 (0.13)	**0.40 (0.17)**	**0.52 (0.21)**
AdaDetect RF	**0.99 (0.01)**	**0.85 (0.03)**	**0.99 (0.01)**	**0.48 (0.10)**	0.04 (0.09)	0.03 (0.08)
AdaDetect NN	0.76 (0.15)	**0.80 (0.07)**	0.52 (0.41)	**0.47 (0.14)**	0.11 (0.13)	0.01 (0.03)
AdaDetect cv NN	0.84 (0.12)	0.76 (0.13)	**0.74 (0.41)**	0.42 (0.16)	**0.13 (0.12)**	0.01 (0.03)
CAD SVDD + CNN	–	–	–	–	–	0.03 (0.15)
AdaDetect CNN	–	–	–	–	–	**0.93 (0.06)**

error of estimated FDR/TDR divided by 10. This can be viewed as an approximation of the standard error with $10,000$ Monte Carlo simulations. In this setting, $\pi_0 = 0.5$ and thus all methods provably control the FDR at level $\pi_0\alpha = 0.025$ (except "Emp BH current"). This is confirmed in the left panel of Figure 6. From the right panel, we observe that AdaDetect with RF achieves the highest power, substantially improving upon AdaDetect with nonadaptive scores (Emp BH). This demonstrates the advantages of utilizing classification-based score functions. In Section G of the Supplementary Material, we present results in additional experimental settings that exhibit qualitative similarities to Figure 6.

8. Conclusion and discussion. In this work, we propose AdaDetect as a generic framework that can wrap around any classification methods and provably control the FDR in finite samples when the null measurements are exchangeable conditional on the novelties. It generalizes and often substantially outperforms previous methods that only work with one-class classification methods which are not adaptive to the novelty distribution. We also develop the π_0-adaptive AdaDetect that further improves the power in the presence of many novelties as well as the cross-validated AdaDetect that allows model selection. The theoretical analysis is based on a novel FDR expression that unifies and generalizes the existing results. In addition, we provide power analysis showing that (1) the optimal score function is given by any monotonic transformation of the ratio between the average density of novelties and the null density and (2) the estimated score function can be asymptotically optimal in terms of power. We demonstrate the versatility of AdaDetect on a variety of tasks.

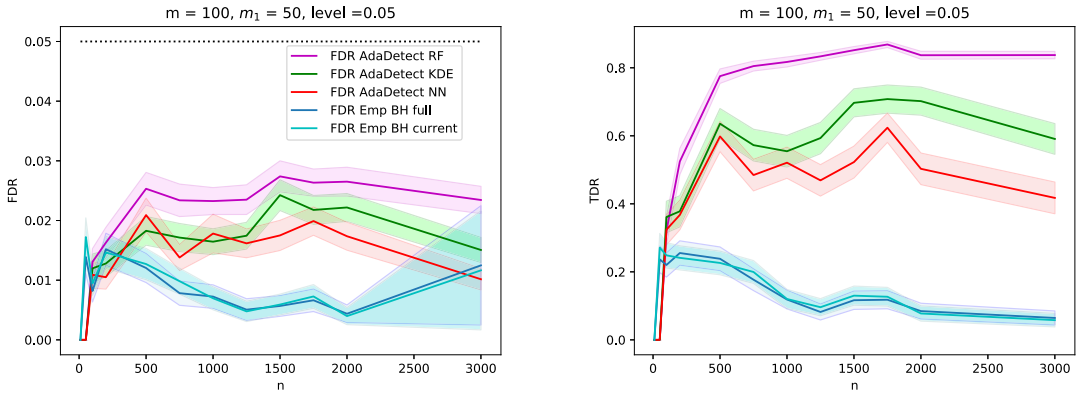

FIG. 6. *Estimated FDR (left) and TDR (right) as a function of n, the size of the NTS, with* $m = 100, m_1 = 50$ *and* $\alpha = 0.05$. *All methods shown in the plot provably control the FDR at level* $\pi_0\alpha = 0.025$ (*except "Emp BH current"*).

8.1. *Limitations of AdaDetect.* Here we discuss several limitations of our method and potential solutions.

- Heterogeneous null distributions. A key assumption for the FDR control is that the null distribution P_0 is the same across the NTS and the test sample. This excludes the case where the null can be generated from a bag of distributions $\{P_{0,k}, 1 \le k \le K\}$. Under heterogeneity, the empirical p-values can be invalid even marginally since the nulls are no longer exchangeable. One possible way to reconcile this issue is to assume the nulls are generated from a mixture distribution $\sum_{k=1}^{K} \pi_k P_{0,k}$, thereby retaining the exchangeability. We leave this for future research.
- Directional null hypotheses. Throughout the paper, we focus on testing whether a new observation has the same distribution as the typical measurements. In some applications, it may be more appropriate to test directional nulls, which are often characterized by the sign of a parameter for parametric models. However, it is unclear how this can be done in nonparametric cases. One possibility is to consider the nulls $P_i \preceq P_0$ where $\preceq$ denotes the stochastic dominance, meaning that there exists a random vector (A, B) such that $A \sim P_0$, $B \sim P_i$ and $A \ge B$ in an entrywise fashion. By restricting the score function to be entrywise increasing, we may still apply AdaDetect and retain the FDR control.
- Randomness of data splitting. AdaDetect is intrinsically randomized due to the data splitting step. Without carefully documenting random seeds, the researcher can "hack" the results by reporting the best results across different splits. A subsequent work by Bashari et al. (2023) proposes an elegant solution to derandomize AdaDetect by treating the test statistics as e-values and aggregating over all data splits. They show that the e-AdaDetect successfully stabilizes the output of AdaDetect.
- Semisupervised data. In some applications, labeled novelties are available in the training sample. For example, the researcher may have historical data on fraud transactions recorded in the system and can train a two-class classifier to distinguish between the nulls and labeled novelties. When future novelties are similar to labeled novelties, it should yield an efficient score function. This has been studied by Liang, Sesia and Sun (2022). Combining their approach with ours in a nonstationary setting where future novelties behave differently from the past ones is a promising avenue for future research.
- Sparse novelties: when the novelties are too sparse, two-class classifiers may not be the best at discriminating between nominals and novelties and can be out-performed by simpler one-class classifiers; see Liang, Sesia and Sun (2022). One possible solution is to apply AdaDetect cv by including both one-class and two-class classification methods and let

data decide which score function is more efficient. We leave the full examination of this approach for future research.

8.2. *Other future works*. First, we could provide a more detailed power analysis by quantifying the bias term $R_1(g^\sharp_{\mathcal{G}}) - R_1(g^\sharp)$ for a broader class of algorithms. For example, we can consider $\mathcal{G} = \mathcal{G}_{N,L,s}$, the set of realizations of NN with width N, depth L and sparsity s (Bos and Schmidt-Hieber (2022)). Such a quantitative analysis could provide guidelines for choosing hyper-parameters or at least a default range in the cross-validated AdaDetect procedure.

Next, a core assumption of the FDR controlling theory is exchangeability of the null scores conditional on the novelties. This can be satisfied beyond our setting, for example, the knockoff setting discussed in Remark 3.8. This suggests a possible path to further improve the knockoffs method.

Lastly, the BONuS algorithm in Yang et al. (2021) can iteratively remove null observations and update the score function correspondingly using a masking technique introduced by Lei and Fithian (2018). While this increases the computation cost, it gradually reduces the attenuation caused by the null sample in the mixed sample and hence improves the accuracy of the estimated score function. It would be interesting to apply their idea in AdaDetect.

Acknowledgments. We would like to thank Gilles Blanchard, Will Fithian, Aaditya Ramdas, Fanny Villers and Asaf Weinstein for constructive discussions and feedback.

Funding. The authors acknowledge the grants ANR-16-CE40-0019 (project SansSouci), ANR-17-CE40-0001 (BASICS), ANR-21-CE23-0035 (ASCAI) and ANR-23-CE40-0018-01 (BACKUP) of the French National Research Agency ANR, the program Emergence of Sorbonne University (project MARS) and the GDR ISIS through the "projets exploratoires" program (project TASTY).

SUPPLEMENTARY MATERIAL

Supplementary material for "Adaptive novelty detection with FDR guarantee" (DOI: 10.1214/23-AOS2338SUPP; .pdf). This supplement contains the proofs for the theoretical results in the main paper, additional simulations, and details for the numerical experiments.

REFERENCES

ANGELOPOULOS, A. N. and BATES, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. ArXiv preprint. Available at arXiv:2107.07511.

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

BALASUBRAMANIAN, V., HO, S.-S. and VOVK, V. (2014). *Conformal Prediction for Reliable Machine Learning*: *Theory, Adaptations and Applications*. Newnes.

BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

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

BARBER, R. F., CANDÈS, E. J. and SAMWORTH, R. J. (2020). Robust inference with knockoffs. *Ann. Statist.* **48** 1409–1431. MR4124328 https://doi.org/10.1214/19-AOS1852

BARBER, R. F. and RAMDAS, A. (2017). The p-filter: Multilayer false discovery rate control for grouped hypotheses. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1247–1268. MR3689317 https://doi.org/10.1111/rssb.12218

BARTLETT, P. L., HARVEY, N., LIAW, C. and MEHRABIAN, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.* **20** Paper No. 63. MR3960917

BASHARI, M., EPSTEIN, A., ROMANO, Y. and SESIA, M. (2023). Derandomized novelty detection with fdr control via conformal e-values.

BATES, S., CANDÈS, E., LEI, L., ROMANO, Y. and SESIA, M. (2023). Testing for outliers with conformal p-values. *Ann. Statist.* **51** 149–178. MR4564852 https://doi.org/10.1214/22-aos2244

BEKKER, J. and DAVIS, J. (2020). Learning from positive and unlabeled data: A survey. *Mach. Learn.* **109** 719–760. MR4094532 https://doi.org/10.1007/s10994-020-05877-5

BENJAMINI, Y. (2010). Discovering the false discovery rate. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 405–416. MR2758522 https://doi.org/10.1111/j.1467-9868.2010.00746.x

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

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

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

BLANCHARD, G., LEE, G. and SCOTT, C. (2010). Semi-supervised novelty detection. *J. Mach. Learn. Res.* **11** 2973–3009. MR2746544

BLANCHARD, G. and ROQUAIN, É. (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.* **10** 2837–2871. MR2579914

BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717 https://doi.org/10.1214/15-AOAS842

BOS, T. and SCHMIDT-HIEBER, J. (2022). Convergence rates of deep ReLU networks for multiclass classification. *Electron. J. Stat.* **16** 2724–2773. MR4406243 https://doi.org/10.1214/22-ejs2011

CAI, T. T. and SUN, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* **104** 1467–1481. MR2597000 https://doi.org/10.1198/jasa.2009.tm08415

CAI, T. T., SUN, W. and WANG, W. (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 187–234. MR3928141

CALVO, B., LARRANAGA, P. and LOZANO, J. A. (2007). Learning Bayesian classifiers from positive and unlabeled examples. *Pattern Recogn. Lett.* **28** 2375–2384.

CANNON, A., HOWSE, J., HUSH, D. and SCOVEL, C. (2002). Learning with the neyman-pearson and min-max criteria. Los Alamos National Laboratory, Tech. Rep. LA-UR, pages 02–2951.

DU PLESSIS, M. C., NIU, G. and SUGIYAMA, M. (2014). Analysis of learning from positive and unlabeled data. *Adv. Neural Inf. Process. Syst.* **27** 703–711.

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289 https://doi.org/10.1198/016214504000000089

EFRON, B. (2007). Doing thousands of hypothesis tests at the same time. *Metron* **LXV** 3–21.

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866 https://doi.org/10.1214/07-STS236

EFRON, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.* **104** 1015–1028. MR2562003 https://doi.org/10.1198/jasa.2009.tm08523

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571 https://doi.org/10.1198/016214501753382129

FERREIRA, J. A. and ZWINDERMAN, A. H. (2006). On the Benjamini–Hochberg method. *Ann. Statist.* **34** 1827–1849. MR2283719 https://doi.org/10.1214/009053606000000425

FISHER, A. (2021). Saffron and lord ensure online control of the false discovery rate under positive dependence. ArXiv preprint. Available at arXiv:2110.08161.

FITHIAN, W. and LEI, L. (2022). Conditional calibration for false discovery rate control under dependence. *Ann. Statist.* **50** 3091–3118. MR4524490 https://doi.org/10.1214/21-aos2137

FRIEDMAN, J. H. (2003). On multivariate goodness–of–fit and two–sample testing. *Stat. Probl. Part. Phys. Astrophys. Cosmol.* **1** 311.

GIRAUD, C. (2022). *Introduction to High-Dimensional Statistics. Monographs on Statistics and Applied Probability* **168**. CRC Press, Boca Raton, FL. MR4436193 https://doi.org/10.1201/9781003158745

GOEMAN, J. J. and SOLARI, A. (2011). Multiple testing for exploratory research. *Statist. Sci.* **26** 584–597. MR2951390 https://doi.org/10.1214/11-STS356

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3617773

GUO, T., XU, C., HUANG, J., WANG, Y., SHI, B., XU, C. and TAO, D. (2020). On positive-unlabeled classification in gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*).

IVANOV, D. (2020). Dedpul: Difference-of-estimated-densities-based positive-unlabeled learning. In 2020 19*th IEEE International Conference on Machine Learning and Applications* (*ICMLA*) 782–790. IEEE, New York.

IVEZIĆ, Ž., CONNOLLY, A. J., VANDERPLAS, J. T. and GRAY, A. (2019). *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton Univ. Press, Princeton.

IVEZIĆ, Ž., VIVAS, A. K., LUPTON, R. H. and ZINN, R. (2005). The selection of RR lyrae stars using single-epoch data. *Astron. J.* **129** 1096.

JAVANMARD, A. and JAVADI, H. (2019). False discovery rate control via debiased lasso. *Electron. J. Stat.* **13** 1212–1253. MR3935848 https://doi.org/10.1214/19-ejs1554

KORN, E. L., TROENDLE, J. F., MCSHANE, L. M. and SIMON, R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *J. Statist. Plann. Inference* **124** 379–398. MR2080371 https://doi.org/10.1016/S0378-3758(03)00211-8

LECUN, Y. and CORTES, C. (2010). MNIST handwritten digit database.

LEI, L., D'AMOUR, A., DING, P., FELLER, A. and SEKHON, J. (2021). Distribution-free assessment of population overlap in observational studies Technical report.

LEI, L. and FITHIAN, W. (2018). AdaPT: An interactive procedure for multiple testing with side information. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 649–679. MR3849338 https://doi.org/10.1111/rssb.12253

LIANG, Z., SESIA, M. and SUN, W. (2022). Integrative conformal p-values for powerful out-of-distribution testing with labeled outliers.

LOPER, J. H., LEI, L., FITHIAN, W. and TANSEY, W. (2022). Smoothed nested testing on directed acyclic graphs. *Biometrika* **109** 457–471. MR4430968 https://doi.org/10.1093/biomet/asab041

MA, R., CAI, T. T. and LI, H. (2021). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *J. Amer. Statist. Assoc.* **116** 984–998. MR4270038 https://doi.org/10.1080/01621459.2019.1699421

MARANDON, A., LEI, L., MARY, D. and ROQUAIN, E. (2024). Supplement to "Adaptive Novelty Detection with false discovery rate guarantee." https://doi.org/10.1214/23-AOS2338SUPP

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

RAMDAS, A., CHEN, J., WAINWRIGHT, M. J. and JORDAN, M. I. (2019a). A sequential algorithm for false discovery rate control on directed acyclic graphs. *Biometrika* **106** 69–86. MR3912384 https://doi.org/10.1093/biomet/asy066

RAMDAS, A. K., BARBER, R. F., WAINWRIGHT, M. J. and JORDAN, M. I. (2019b). A unified treatment of multiple testing with prior knowledge using the p-filter. *Ann. Statist.* **47** 2790–2821. MR3988773 https://doi.org/10.1214/18-AOS1765

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

ROMANO, J. P. and WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100** 94–108. MR2156821 https://doi.org/10.1198/016214504000000539

ROQUAIN, E. and VILLERS, F. (2011). Exact calculations for false discovery proportion with application to least favorable configurations. *Ann. Statist.* **39** 584–612. MR2797857 https://doi.org/10.1214/10-AOS847

ROSSET, S., HELLER, R., PAINSKY, A. and AHARONI, E. (2022). Optimal and maximin procedures for multiple testing problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1105–1128. MR4494154 https://doi.org/10.1111/rssb.12507

RUFF, L., VANDERMEULEN, R. A., GÖRNITZ, N., DEECKE, L., SIDDIQUI, S. A., BINDER, A., MÜLLER, E. and KLOFT, M. (2018). Deep one-class classification. In *Proceedings of the* 35*th International Conference on Machine Learning* **80** 4393–4402.

SARKAR, S. K. (2008). On methods controlling the false discovery rate. *Sankhyā* **70** 135–168. MR2551809

SCHÖLKOPF, B., PLATT, J. C., SHAWE-TAYLOR, J., SMOLA, A. J. and WILLIAMSON, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Comput.* **13** 1443–1471.

SCOTT, C. and NOWAK, R. (2005). A Neyman–Pearson approach to statistical learning. *IEEE Trans. Inf. Theory* **51** 3806–3819. MR2239000 https://doi.org/10.1109/TIT.2005.856955

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

SUGIYAMA, M., SUZUKI, T. and KANAMORI, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge Univ. Press, Cambridge. MR2895762 https://doi.org/10.1017/CBO9781139035613

SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657 https://doi.org/10.1198/016214507000000545

SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 393–424. MR2649603 https://doi.org/10.1111/j.1467-9868.2008.00694.x

VAPNIK, V. N. (1998). *Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control*. Wiley, New York. MR1641250

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

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

WANG, Y., KAJI, T. and ROCKOVA, V. (2022). Approximate Bayesian computation via classification. *J. Mach. Learn. Res.* **23** Paper No. 350. MR4577789

WEINSTEIN, A. (2021). On permutation invariant problems in large-scale inference. ArXiv preprint. Available at arXiv:2110.06250.

WEINSTEIN, A., BARBER, R. and CANDÈS, E. (2017). A power and prediction analysis for knockoffs with lasso statistics. Available at arXiv:1712.06465.

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

ZRNIC, T., RAMDAS, A. and JORDAN, M. I. (2021). Asynchronous online testing of multiple hypotheses. *J. Mach. Learn. Res.* **22** Paper No. 33. MR4253726 https://doi.org/10.1515/ijnsns-2019-0210