Journal of Machine Learning Research 24 (2023) 1-41 Submitted 10/22; Revised 2/23; Published 4/23

Selection by Prediction with Conformal p-values

Ying Jin YING531@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Emmanuel J. Candès CANDES@STANFORD.EDU
Department of Statistics and Department of Mathematics
Stanford University
Stanford, CA 94305, USA

Editor: Genevera Allen

Abstract

Decision making or scientific discovery pipelines such as job hiring and drug discovery often involve multiple stages: before any resource-intensive step, there is often an initial screening that uses predictions from a machine learning model to shortlist a few candidates from a large pool. We study screening procedures that aim to select candidates whose unobserved outcomes exceed user-specified values. We develop a method that wraps around any prediction model to produce a subset of candidates while controlling the proportion of falsely selected units. Building upon the conformal inference framework, our method first constructs p-values that quantify the statistical evidence for large outcomes; it then determines the shortlist by comparing the p-values to a threshold introduced in the multiple testing literature. In many cases, the procedure selects candidates whose predictions are above a data-dependent threshold. Our theoretical guarantee holds under mild exchangeability conditions on the samples, generalizing existing results on multiple conformal p-values. We demonstrate the empirical performance of our method via simulations, and apply it to job hiring and drug discovery datasets.

Keywords: Conformal inference, selective inference, multiple testing, p-values, false discovery rate

1. Introduction

Decision making and scientific discovery are resource intensive tasks: human evaluation is needed before high-stakes decisions such as job hiring (Shen et al., 2019) and disease diagnosis (Etzioni et al., 2003); several rounds of expensive clinical trials are required before a drug can receive FDA approval (FDA, 2018). Early on, we often hope to identify viable candidates from a very large pool—consider hundreds of applicants to a position or hundreds of thousands of potential compounds that may bind to the target. In such problems, machine learning prediction is useful for an initial screening step to shortlist a few candidates; in later, more costly stages, only these shortlisted candidates are carefully investigated to confirm the interesting cases.

This paper concerns scenarios where outcomes taking on higher values are of interest. Formally, suppose we have access to a set of training data $\{(X_i, Y_i)\}_{i=1}^n$ and a set of test

samples $\{X_{n+j}\}_{j=1}^m$ whose outcomes $\{Y_{n+j}\}_{j=1}^m$ are unobserved, all $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ pairs being i.i.d. from some arbitrary and unknown distribution.[1] Given some thresholds $\{c_j\}_{j=1}^m$, our goal is to find as many test units with $Y_{n+j} > c_j$ as possible, while ensuring the false discovery rate (FDR), the expected proportion of errors ($Y_{n+j} \le c_j$) among all shortlisted candidates, is controlled. To be specific, letting $\mathcal{R} \subseteq \{1, \dots, m\}$ be the selection set, we define FDR as the expectation of the false discovery proportion (FDP), so that

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}], \quad \mathrm{FDP} = \frac{\sum_{j=1}^m \mathbb{1}\{j \in \mathcal{R}, Y_{n+j} \le c_j\}}{1 \vee |\mathcal{R}|}, \tag{1}$$

where we denote $a \vee b = \max\{a, b\}$ for any $a, b \in \mathbb{R}$, and the expectation is over the randomness of all training data and all test samples. The FDR is a natural measure of type-I error for binary classification (Hastie et al., 2009). For regression problems with a continuous response, counting the error is reasonable if each selected candidate incurs a similar cost. We discuss below potential applications with binary or quantitative outcomes.

Candidate screening. Companies are turning to machine learning to support recruitment decisions (Faliagka et al., 2012; Shehu and Saeed, 2016). Predictions using automatic resume screening (Amdouni and abdessalem Karaa, 2010; Faliagka et al., 2014), semantic matching (Mochol et al., 2007) or AI-assisted interviews are used to screen and select candidates from a large pool. Related tasks include talent sourcing, e.g., finding people who are likely to search for new opportunities, and candidate screening, i.e., selecting qualified applicants before further human evaluation (Heaslip, 2022). One may be interested in controlling the FDR (1) for resource efficiency: each shortlisted candidate incurs similar costs such as communication for talent sourcing and interviews before the hiring decisions. In candidate screening, controlling FDR ensures most of the costs are devoted to evaluating and ranking qualified candidates. Job recruitment also has fairness concerns; an alternative goal is to ensure that qualified candidates do not get screened out before human evaluation. To this end, one can flip the sign of the outcomes, and (1) represents the proportion of qualified candidates in the filtered out ones.

Drug discovery. Machine learning is playing a similar role in accelerating the drug discovery pipeline. Early stages of drug discovery aim at finding molecules or compounds—from a diverse library (Szymański et al., 2011) developed by institutions across the world (Kim et al., 2021)—with strong effects such as high binding affinity to a specific target. The activity of drug candidates can be evaluated by high-throughput screening (HTS) (Macarron et al., 2011). However, the capacity of this approach is quite limited in practice, and it is generally infeasible to screen the whole library of readily synthesized compounds. Instead, virtual screening (Huang, 2007) by machine learning models has enabled the automatic search of promising drugs. Often, a representative (ideally diverse) subset of the whole library is evaluated by HTS; machine learning models are then trained on these data to predict other candidates' activity based on their physical, geometric and chemical features (Carracedo-Reboredo et al., 2021; Koutsoukas et al., 2017; Vamathevan et al., 2019; Dara et al., 2021) and select promising ones for further HTS and/or clinical trials. Given the cost of subsequent investigation, false positives in this process are a major concern (Sink

1. Later on, we will relax the i.i.d. assumption to exchangeablity conditions.

et al., 2010). Ensuring that a sufficiently large proportion of resources is devoted to promising drugs is thus important for the efficiency of the whole pipeline.

In these two examples, the FDR quantifies a trade-off between the resources devoted to shortlisted candidates (the selection set) and the benefits from finding interesting candidates (the true positives). This interpretation is similar to the justification of FDR in multiple testing (Benjamini and Hochberg, 1995, 1997; Benjamini and Yekutieli, 2001): when evaluating a large number of hypotheses, the FDR measures the proportion of "false leads" for follow-up confirmatory studies. However, in our *prediction problem*, the affinity of a new drug is inferred not from the observations, but from other similar compounds, i.e., other drugs in the training data. This perspective also blurs the distinction between statistical inference and prediction; we will draw more connections between these sub-fields later.

The FDR may not necessarily be interpreted as a resource-efficiency measure. In the next two examples, controlling the FDR, which limits the error in inferring the direction of outcomes, is relevant to monitoring risk in healthcare and counterfactual inference.

Healthcare. With increasingly available patient data, machine learning is widely adopted to assist human decisions in healthcare. For example, many works use machine learning prediction for large-scale early disease diagnosis (Shen et al., 2019; Richens et al., 2020) and patient risk prediction (Rahimi et al., 2014; Corey et al., 2018; Jalali et al., 2020). Calibrating black-box models is important in such high-stakes problems. When it is more desired to limit false negatives than false positives, machine learning prediction might be used to filter out low-risk cases, leaving other cases for careful human evaluation. It is sensible to control the proportion of high-risk cases among all filtered out samples.

Counterfactual inference. In randomized experiments that run over a period of time, inferring whether the patients have benefited from the treatment option compared to an alternative might inform decisions such as early stopping of the trial for some patients. More generally, inferring the benefit of certain patients also provides evidence on treatment effect heterogeneity. This is a counterfactual inference problem (Lei and Candès, 2021; Jin et al., 2023) in which one could predict the counterfactuals, i.e., what would happen *should* one takes an alternative option, by learning from the outcomes of patients under that option, and then compare the prediction to the realized outcomes. In this case, the set of those declared as having benefited from the treatment is informative if the FDR is controlled.

The generic task underlying these applications is to find a subset of candidates whose not-yet-observed outcomes are of interest (e.g., qualification or high binding affinity to the target) from a potentially enormous pool of test samples. This is often achieved by thresholding their test scores—the model prediction on the test samples—from models built on a set of training data that are assumed to be from the same distribution. However, controlling the error in the selected set is a nontrivial task.

1.1 Why calibrated predictive inference is insufficient

We consider a binary example to fix ideas, so that $\mathcal{Y} = \{0, 1\}$. Our goal is to find test samples with $Y_{n+j} = 1$. A natural starting point is to train a machine learning model that predicts (classifies) Y given X, with the hope that test samples with higher predicted values

are more promising. To achieve valid prediction, one could calibrate the model (Vovk et al., 2005) to output a prediction set $\widehat{C}_{1-\alpha}(X)$ taking the form $\varnothing$, $\{0\}$, $\{1\}$ or $\{0,1\}$ with the prescription that $\widehat{C}_{1-\alpha}(X)$ must cover the outcome Y with probability at least $1-\alpha$ for some user-specified $\alpha \in (0,1)$. The probability is averaged over the randomness in the test sample and the training process.

However, a prediction set with marginal coverage guarantees is insufficient for selection. For instance, one might consider selecting all test samples j with $\widehat{C}_{1-\alpha}(X_{n+j}) = \{1\}$. The FDR of the selected set would then be below α if $\widehat{C}_{1-\alpha}(X_{n+j})$ covers 1 with probability at least $1-\alpha$ *for the selected units*, that is, conditional on selection. However, this is clearly a false statement because predictive inference only ensures $(1-\alpha)$ coverage averaged over *all* test samples. In fact, no matter how large we set the coverage $(1-\alpha)$, such a naive approach might still return a selection set that contains too many uninteresting candidates.

It might be best to preview our results on a real-world drug discovery dataset properly introduced and studied later in the paper (Section 4.2.1). In short, the goal is to find promising drug candidates, among thousands of molecules, that are active ($Y = 1$) for the HIV target. This dataset is highly imbalanced in the sense that only 3% of the drugs are active, as is often the case in studies on drug discovery. Our main purpose is here to rapidly demonstrate that a straightforward application of conformal prediction methods, selecting those leads with $\widehat{C}_{1-\alpha}(X_{n+j}) = \{1\}$, results in over-confident predictions in a sense described below.

We use a deep learning model (this is introduced in Section 4.2.1) to construct conformity scores, and ultimately, conformal prediction sets that are one-sided in the sense that they only take on three possible values: $\varnothing$, $\{1\}$ and $\{0,1\}$; (see Appendix C.1 for details). The left panel of Figure 1 shows the FDR of the naive approach as a function of the confidence level $1-\alpha \in \{0.99, 0.98, \ldots, 0.70\}$, along with the marginal miscoverage of conformal prediction sets and the proportion of cases in the test set for which $\widehat{C}_{1-\alpha}(X) = \{1\}$.

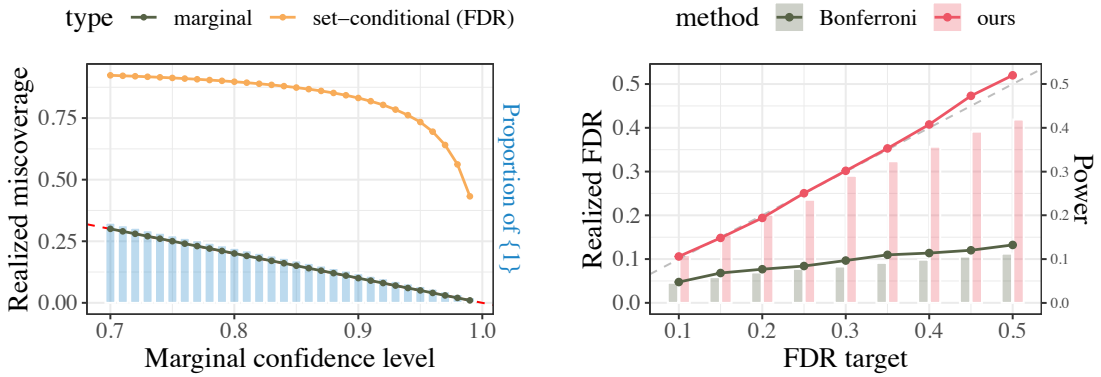

Figure 1: Left: FDR (set-conditional miscoverage) of the naive approach and marginal miscoverage as a function of the parameter α; the light blue bars are the proportion of cases among all test samples for which $\widehat{C}_{1-\alpha}(X) = \{1\}$. Right: FDR (curve) and power (bar) of our selective inference approach and of Bonferroni's method as a function of the nominal FDR target q. The FDR (resp. power) is computed by averaging the FDP (resp. proportion of true positives) in $N = 100$ independent splits of training, calibration, and test data.

While conformal prediction always achieves nearly exact marginal validity (brown), it is overconfident for seemingly promising candidates, as the error rate among the selected (FDR)— those with $\widehat{C}_{1-\alpha}(X) = \{1\}$—is very high (orange). When $1 - \alpha = 0.90$, we witness an error rate of about 80%, meaning that 4 out of 5 'discoveries' are false. Even in the extremely conservative case where $1 - \alpha = 0.99$, the FDR exceeds 35%. Note that this phenomenon is independent of the target FDR level. We can thus see that the selection issue would be especially pressing if, say, we aim for a small FDR level. In fact, conformal prediction outputs a large proportion of uninformative sets: as seen from the light bars, about $1-\alpha$ of the prediction sets are $\widehat{C}_{1-\alpha}(X) = \{0, 1\}$ (we observe no empty prediction sets for this data). Thus, it ensures valid marginal coverage even though those $\widehat{C}_{1-\alpha}(X) = \{1\}$ seldom cover the true label.

To make sure the FDR falls below a user specified tolerance $q \in \{0, 1\}$, one might want to employ a Bonferroni correction. To do this we would pick test cases for which $\widehat{C}_{1-q/m}(X_{n+j}) = \{1\}$, where m is the number of test samples. That is, we apply a Bonferroni correction to the marginal coverage, and this ensures that the probability of making a single false selection—which upper bounds the FDR—is below q. In the right panel of Figure 1, we compare the FDR and power of our approach and Bonferroni's method applied to a range of nominal FDR levels $q \in \{0.11, 0.15, \ldots, 0.3\}$.[2] Our approach yields almost exact FDR control and much higher power than Bonferroni's.

To ensure calibration on the selected, we will bridge conformal inference and selective inference and devise `cfBH`, an algorithm that turns any prediction model into a screening mechanism. In a nutshell, instead of calibrating to a fixed confidence level α, we will use tools from conformal inference to quantify the model confidence in outcomes with larger values, and then employ multiple testing ideas to construct a shortlist of candidates with statistical guarantees.

Returning to the drug discovery application, we acknowledge a substantial literature using conformal inference for uncertainty quantification in compound activity prediction, see Lampa et al. (2018); Eklund et al. (2015); Svensson et al. (2018, 2017); Lindh et al. (2017), and Cortés-Ciriano and Bender (2019) for a recent review. Whether explicitly stated or not, the goal is eventually to select or prioritize compounds that progress to later stages of drug discovery (Ahlberg et al., 2017a,b) after constructing valid prediction intervals. That said, current tools for selection are all heuristic, e.g., picking cases with a high predicted value and a relatively short prediction interval. As already mentioned, a marginally valid prediction set does not necessarily imply reliable selection. The method from this paper fills this gap, and can wrap around the predictions from the literature to produce reliable selection rules for drug discovery.

1.2 Hypothesis testing and conformal p-values

One may view our problem as testing the *random* hypotheses

$$H_j \colon Y_{n+j} \leq c_j, \quad j = 1, \ldots, m. \tag{2}$$

From now on, we denote $\mathcal{H}_0 = \{j \colon Y_{n+j} \leq c_j\}$ as the set of null hypotheses. That is, we define a hypothesis H_j for each test sample j, and we say H_j is non-null if Y_{n+j} exceeds the

2. Here we take a subset of $m = 1000$, as otherwise q/m exceeds the resolution of conformal prediction.

threshold c_j. This is perhaps non-classical since the hypothesis H_j is random: it concerns a random variable rather than a model parameter. However, we show that we can still use p-values and rely on multiple hypothesis testing ideas to construct the "rejection" set $\mathcal{R}$.

We start by introducing the tool we use to quantify model confidence: *conformal p-values*; as its name suggests, these p-values build upon the conformal inference framework (Vovk et al., 2005, 1999). Suppose we are given any prediction model from a training process that is independent of the calibration and test samples. We condition on the training process and view the prediction model as given. We first define a nonconformity score $V\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ based on the prediction model. Intuitively, $V(x, y)$ measures how well a value y *conforms* to the prediction of the model at x. For example, given a prediction $\widehat{\mu}\colon \mathcal{X} \to \mathbb{R}$, one could use $V(x, y) = |y - \widehat{\mu}(x)|$; other popular choices in the literature include ideas based on quantile regression (Romano et al., 2019) and conditional density estimation (Chernozhukov et al., 2021). Should Y_{n+j} be observed, one could compute the nonconformity scores $V_i = V(X_i, Y_i)$ for $i = 1, \ldots, n$ and $V_{n+j} = V(X_{n+j}, Y_{n+j})$. The corresponding conformal p-value (Vovk et al., 2005, 1999; Bates et al., 2021) is defined as

$$p_j^* = \frac{\sum_{i=1}^n \mathbb{1}\{V_i < V_{n+j}\} + U_j \cdot (1 + \sum_{i=1}^n \mathbb{1}\{V_i = V_{n+j}\})}{n+1}, \tag{3}$$

where $U_j \sim \text{Unif}[0,1]$ are i.i.d. random variables to break ties. If the test sample (X_{n+j}, Y_{n+j}) follows the same distribution as the training data, then $p_j^* \sim \text{Unif}\,[0, 1]$. However, the mutual dependence among $\{p_j^*\}$ is complicated as they all depend on the same calibration data. A recent paper (Bates et al., 2021) used conformal p-values for outlier detection; in their setting, observations $\{(X_{n+j}, Y_{n+j})\}_{j=1}^m$ are available, and the null set $\{j\colon H_j \text{ is true}\}$ is deterministic since the null hypothesis H_j posits that (X_{n+j}, Y_{n+j}) follows the same distribution as the training samples. In our setting, the response Y_{n+j} is not observed. This leads us to introduce a different set of conformal p-values. Our analysis also generalizes Bates et al. (2021) to exchangeable data.

1.3 Related work

This work concerns calibrating prediction models to obtain correct directional conclusions on the outcomes. In situations where one cares more about the mistakes on the selected subset, our error notion, the FDR, might be more relevant than average prediction errors. That said, several works have studied FDR control in prediction problems, especially in binary classification. Among them, Dickhaus (2014) connects classification to multiple testing, showing that controlling type-I error (FDR) at certain levels by thresholding an oracle classifier asymptotically achieves the optimal (Bayes) classification risk; Scott et al. (2009) provides high-probability bounds for estimating the FDR achieved by classification rules, rather than adaptively controlling it at a specific level.

Our problem setup is close to several recent works on calibrated screening or thresholding (Wang et al., 2022; Sahoo et al., 2021) in classification or regression problems. These works however focus on different targets; Wang et al. (2022) focuses on selecting a subset with a prescribed expected number of qualified candidates; Sahoo et al. (2021) focuses on the calibration of the predicted score itself to achieve a similar notion of error control as ours, but at varying levels for all thresholds. The difference is that our method rigorously

controls FDR in finite samples, while it might be difficult to obtain such guarantees for the targets in Wang et al. (2022); Sahoo et al. (2021).

Our methods build upon the conformal inference framework (Vovk et al., 2005, 1999). Although conformal-inference-based methods have been developed for reliable uncertainty quantification in various problems (Lei and Candès, 2021; Candès et al., 2021; Jin et al., 2023; Tibshirani et al., 2019), the theoretical guarantee usually concerns a single test point. However, in many applications, one might be interested in a batch of individuals and desire uncertainty quantification for multiple test samples simultaneously; in such situations, these methods are insufficient due to the complex dependence structure of test scores and p-values as well as multiplicity issues.

This work is closely related to Bates et al. (2021), in which the authors use conformal p-values (3) to test for multiple outliers. Our conformal p-values differ from theirs as the outcomes are not observed. A few works (Mary and Roquain, 2021; Roquain and Verzelen, 2022) are parallel to Bates et al. (2021), studying multiple testing in a setting where all null hypotheses specify an identical null distribution; they are further generalized by Rava et al. (2021) to achieve subgroup FDR control in classification. Our method is similar to this line of work in constructing a threshold for certain "scores" and selecting candidates with scores above that threshold. However, we work with random hypotheses and propose distinct procedures, whereas in their works, the hypotheses are deterministic (or conditioned on). We will discuss these distinctions in more detail as we present our results.

Our perspective on the problem is also generally related to the multiple hypothesis testing literature where the FDR is a popular notion of type-I error. Since we pay more attention to one particular direction (e.g., we are interested in finding those $Y_{n+j} > c_j$), our work is related to testing the signs of statistical parameters (Bohrer, 1979; Bohrer and Schervish, 1980; Hochberg, 1986; Guo et al., 2010; Weinstein and Ramdas, 2020). Our framework differs from the existing directional testing literature in important ways. Firstly, we test for the direction of a random outcome instead of a model parameter. This leads to random null sets, whose dependence structure is complicated. Secondly, our inference relies on exchangeability of the data while imposing no assumption on their distribution; this differs from standard practice, in which a null hypothesis specifies the distribution of test statistics and leads to a uniform p-value.

2. Methodology

2.1 Selection by prediction with conformal p-values

We construct new conformal p-values to test the random hypotheses (2) regarding Y_{n+j}, building on an arbitrary nonconformity score V obeying the *monotone* property.

Definition 1 *A nonconformity score $V(\cdot,\cdot)\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is monotone if $V(x,y)\le V(x,y')$ holds for any $x\in\mathcal{X}$ and any $y,y'\in\mathcal{Y}$ obeying $y\le y'$.*

In general, $V(x,y)$ should measure how extreme the value y is compared to the normal behavior of the outcome for a value x of the covariate. For example, one could let $V(x,y) = y-\widehat{\mu}(x)$ if the prediction machine outputs some estimate $\widehat{\mu}(x)$ of the conditional mean function or conditional quantile function. One could also set $V(x,y)$ as an estimator for $F(x,y) := \mathbb{P}(Y\le y\,|\,X=x)$ that obeys monotonicity (Chernozhukov et al., 2021).

The choice of V may also account for the form of $\mathcal{R}$: given a *monotone* nonconformity score V, our method includes case j in $\mathcal{R}$ if $V(X_{n+j}, c_j)$ is sufficiently small. As warm-up, consider binary classification. To find samples with $Y = 1$, one could set $c_j \equiv 0.5$. Suppose $\widehat{\mu}(\cdot)$ is some prediction from a machine learning algorithm; for instance, one could think of $\widehat{\mu}(x)$ as an estimate for $\mathbb{P}(Y = 1 \,|\, X = x)$ obtained by means of a neural network. If one would like to select individuals with larger fitted probability $\widehat{\mu}(X_{n+j})$, then $V(x, y)$ can be chosen in a way such that $V(x, c_j)$ is decreasing in $\widehat{\mu}(x)$. One such choice is $V(x, y) = y - \widehat{\mu}(x)$. This reasoning also applies to continuous responses with regression modeling.

We first compute $V_i = V(X_i, Y_i)$ for $i \in \mathcal{D}_{\text{calib}} = \{1, \ldots, n\}$ and $\widehat{V}_{n+j} = V(X_{n+j}, c_j)$ for $j = 1, \ldots, m$; here we use the notation $\widehat{V}_{n+j}$ to distinguish from the unobserved scores $V_{n+j} := V(X_{n+j}, Y_{n+j})$. Then for each $j = 1, \ldots, m$, we construct the conformal p-values

$$p_j = \frac{\sum_{i=1}^{n} \mathbb{1}\{V_i < \widehat{V}_{n+j}\} + (1 + \sum_{i=1}^{n} \mathbb{1}\{V_i = \widehat{V}_{n+j}\}) \cdot U_j}{n+1}, \tag{4}$$

where $U_j \sim \text{Unif}(0, 1)$ are i.i.d. random variables to break ties.

Remark 2 Our conformal p-values have an intuitive interpretation: p_j is the smallest significance level such that a one-sided conformal prediction interval for Y_{n+j} excludes c_j. Indeed, the split conformal inference procedure (Lei et al., 2018) (using $-V$ as the nonconformity score) yields the one-sided prediction interval $\widehat{C}(X_{n+j}, 1-\alpha) = [\eta(X_{n+j}, 1-\alpha), +\infty)$ for Y_{n+j}, where

$$-\eta(X_{n+j}, 1-\alpha) = \text{Quantile}\big(1-\alpha; \tfrac{1}{n+1} \textstyle\sum_{i=1}^{n} \delta_{-V_i} + \tfrac{1}{n+1}\delta_{-\infty}\big).$$

It is guaranteed that $\mathbb{P}(Y_{n+j} \in \widehat{C}(X_{n+j}, 1-\alpha)) \geq 1-\alpha$ where the expectation is taken over the randomness in $\mathcal{D}_{\text{calib}}$ and $\{X_{n+j}, Y_{n+j}\}$. Thus, ignoring the tie-breaking U_j, we see that the conformal p-value p_j is the smallest α such that $c_j < \eta(X_{n+j}, 1-\alpha)$. Put it another way, we have confidence of at least $1 - p_j$ that $Y_{n+j} > c_j$.

The difference between $\{p_j\}$ in (4) and $\{p_j^*\}$ in (3) is whether we use $\widehat{V}_{n+j}$ or V_{n+j} to contruct the p-values. To distinguish, we call $\{p_j^*\}$ the *oracle* conformal p-values hereafter, which are not observable in our setting. In Bates et al. (2021), p_j^* quantifies how extreme a score is and is used to test whether the test sample j is an outlier. In our context, p_j quantifies how extreme the threshold is compared to the usual behavior of the outcomes.

We then run the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) with the conformal p-values $\{p_j\}$. The whole procedure for `cfBH` is summarized in Algorithm 1.

2.2 Finite sample FDR control

Throughout, we define the random vector $Z_i = (X_i, Y_i)$ for $1 \leq j \leq n+m$, and $\widetilde{Z}_{n+j} = (X_{n+j}, c_j)$ for $1 \leq j \leq m$. The following theorem establishes generic conditions under which `cfBH` controls the FDR (1) using $\{p_j\}$ in (4) with i.i.d. calibration and test samples.

Theorem 3 *Suppose V is monotone, the calibration data $\{Z_i\}_{i=1}^n$ and test data $\{Z_{n+j}\}_{j=1}^m$ are i.i.d., and data in $\{Z_i\}_{i=1}^n \cup \{\widetilde{Z}_{n+\ell}\}_{\ell \neq j} \cup \{Z_{n+j}\}$ are mutually independent for any j. Then, for any $q \in (0, 1)$, the output $\mathcal{R}$ of Algorithm 1 satisfies* $\text{FDR} \leq q$.

Algorithm 1 `cfBH`: Selection by prediction with conformal p-values

Input: Calibration data $\{(X_i, Y_i)\}_{i=1}^n$, test data covariates $\{X_{n+j}\}_{j=1}^m$, thresholds $\{c_j\}_{j=1}^m$, FDR target $q \in (0,1)$, monotone nonconformity score $V\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

1: Compute $V_i = V(X_i, Y_i)$ for $i = 1, \ldots, n$, and $\widehat{V}_{n+j} = V(X_{n+j}, c_j)$ for $j = 1, \ldots, m$.
2: Construct conformal p-values $\{p_j\}_{j=1}^m$ as in (4).
3: (BH procedure) Compute $k^* = \max\big\{k\colon \sum_{j=1}^m \mathbb{1}\{p_j \le qk/m\} \ge k\big\}$.

Output: Selection set $\mathcal{R} = \{j\colon p_j \le qk^*/m\}$.

Our framework applies to scenarios where c_j are random variables (see Examples 1, 2 and 3 in the next subsection). In this case, all data being i.i.d. does not necessarily imply the mutual independence of data in $\{Z_i\}_{i=1}^n \cup \{\widetilde{Z}_{n+\ell}\}_{\ell \neq j} \cup \{Z_{n+j}\}$.

We note some conceptual novelty regarding our p-values in `cfBH`. In conventional statistical inference, a null hypothesis (approximately) specifies the distribution of a test statistic, e.g., a p-value dominating Unif$[0,1]$. In sharp contrast, p-values defined in (4) do not satisfy such a property, i.e., it does not necessarily hold that $\mathbb{P}(p_j \le \alpha \,|\, j \in \mathcal{H}_0) \le \alpha$ for $\alpha \in (0,1)$. This is because p_j and the random hypothesis H_j are dependent in an unknown fashion.[3] Instead, they obey a generalized notion which states that

$$\mathbb{P}(p_j \le \alpha \text{ and } j \in \mathcal{H}_0) \le \alpha, \quad \text{for all } \alpha \in [0,1]. \tag{5}$$

That is, for testing one single hypothesis H_j at level α, rejecting $p_j \le \alpha$ yields the control of the generalized error (5) that accounts for the randomness in H_j as well. This might connect to the Bayesian perspective where parameters are themselves random variables.

We outline some important properties of our p-values to develop intuitions regarding the FDR control of `cfBH`. It is proved in Bates et al. (2021) that the oracle p-values $\{p_j^*\}$ satisfy a specific dependence structure called *positive regression dependent on a subset* (PRDS) (Benjamini and Yekutieli, 2001).

Definition 4 (PRDS) *A random vector $X = (X_1, \ldots, X_m)$ is PRDS on a subset $\mathcal{I}$ if for any $i \in \mathcal{I}$ and any increasing set D, the probability $\mathbb{P}(X \in D \,|\, X_i = x)$ is increasing in x.*

Here a set $D \subseteq \mathbb{R}^m$ is *increasing* if $a \in D$ and $b \succeq a$ implies $b \in D$, where $\succeq$ denote coordinate-wise inequality. To be specific, Bates et al. (2021) proved that the random vector of oracle conformal p-values $(p_1^*, \ldots, p_m^*)$ is PRDS on the index set $\mathcal{I}$ consisting of all cases (X_{n+j}, Y_{n+j}) that follow the same distribution as the training data. Relying on this result, we show that our p-values are PRDS after modifying one coordinate. The following lemma shows the key properties for proving Theorem 3.

Lemma 5 *(i) If V is monotone, then $p_j \ge p_j^*$ on the event $\{j \in \mathcal{H}_0\}$ for each j. (ii) If the calibration and test samples are i.i.d., then $p_j^* \sim$ Unif$([0,1])$. (iii) If data in $\{Z_i\}_{i=1}^n \cup \{\widetilde{Z}_{n+\ell}\}_{\ell \neq j} \cup \{Z_{n+j}\}$ are independent, then $(p_1, \ldots, p_{j-1}, p_j^*, p_{j+1}, \ldots, p_m)$ is PRDS on p_j^*.*

3. To see this, consider a special case where $V(x,y) = y - x$ for $Y = -X + \epsilon$ with $(X, \epsilon) \overset{\text{i.i.d.}}{\sim} N(0,1)$, and $c = 0$. With sufficiently many calibration data ($n \to \infty$), one can show that $p_j = \Phi(-X_{n+j})$ where Φ is the c.d.f. of standard normal distribution. One can check that in this case, $\mathbb{P}(p_j \le 0.05 \,|\, Y \le 0) > 0.09$.

The first two properties in Lemma 5 mean that p_j is more conservative than $p_j^* \sim \mathrm{Unif}\,[0,1]$ on the null event, hence leading to (5). Generalizing to multiple hypotheses testing, one could expect that the false discoveries from $\{p_j\}$ can be controlled with those using $\{p_j^*\}$; the latter is studied in Bates et al. (2021) and works well with the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) because of the PRDS property (c.f. Definition 4). These heuristics are made concrete by a careful leave-one-out analysis for FDR control, as well as the general PRDS property of conformal p-values in (iii) of Lemma 5. We defer the detailed proofs to Appendix B.2.

2.3 Extension to exchangeable data

Our previous result extends to the situation where the calibration and test samples obey a natural exchangeability condition.

Theorem 6 *Suppose V is monotone, and that for any $j = 1, \ldots, m$, the random variables $\{V_1, \ldots, V_n, V_{n+j}\}$ are exchangeable conditional on $\{\widehat{V}_{n+\ell}\}_{\ell \neq j}$. Also, the random variables $\{V_i\}_{i=1}^n \cup \{\widehat{V}_{n+\ell}\}_{\ell=1}^m$ have no ties almost surely. Then* `cfBH` *applied to p-values defined as*

$$p_j^{\mathrm{dtm}} = \frac{1 + \sum_{i=1}^n \mathbb{1}\{V_i < \widehat{V}_{n+j}\}}{n+1},$$

satisfies $\mathrm{FDR} \leq q$.

This result may be of interest in the case where one is sampling from a finite set without replacement. For instance, in drug discovery, the calibration data may be molecules that have already been evaluated; to curate such dataset, it is common to randomly sample and evluate a fixed number of molecules from a fixed drug library. Conditional on all the data in the library, the randomness from sampling still ensures the exchangeability conditions in Theorem 6. On the contrary, the i.i.d. assumptions in Theorem 3 may not hold under sampling without replacement. The proof of Theorem 6 is in Appendix B.3. Its analysis no longer relies on the PRDS property developed in Bates et al. (2021) for i.i.d. data.[4]

2.4 Setting the testing thresholds

Our framework allows us to test the random hypothesis (2) for general thresholds $\{c_j\}$. The simplest case is to set $c_j = \tau$, where τ is some constant which could possibly be obtained from an independent training process. It covers binary or one-versus-all classification problems as well as many scenarios in regression modeling where a threshold on the outcomes can be decided beforehand. Consider an application of large-scale screening in early-stage diagnosis (Shen et al., 2019), where the practitioner would like to find individuals with high unobserved health risk. The threshold of the risk measure as being hazardous could be decided by domain knowledge, or by looking at the experience of former patients. In this case, as τ is independent of all subsequent steps, it can be viewed as fixed and the conditions in Theorem 3 are thus satisfied. However, we do note that τ should not depend on the calibration data, because this would potentially break the mutual independence condition on $\{Z_i\}_{i=1}^n$ and $\{\widetilde{Z}_{n+\ell}\}_{\ell \neq j}$ in Theorem 3 and invalidate FDR control.

4. A consequence is that our new technique can be used to show finite sample FDR control for the outlier detection problem in Bates et al. (2021) under a similar exchangeability condition.

More generally, c_j can also be a random variable for test sample j as discussed below.

Example 1 (Random variable associated with test sample) In early disease diagnosis, one might want to identify individuals whose future cholestrol level Y_{n+j} in a month would be higher than $c_j := W_{n+j}$, their current measurements upon a first visit. In this case, we simply modify the construction of p_j in Algorithm 1 to $\widehat{V}_{n+j} = V(X_{n+j}, W_{n+j})$. All the conditions in Theorem 3 remain true as long as $\{(X_{n+j}, Y_{n+j})\}_{j=1}^m$ are i.i.d. pairs, and $\{(X_{n+j}, Y_{n+j}, W_{n+j})\}_{j=1}^m$ are i.i.d. triples, hence our method still yields valid FDR control.

Continuing the above example, we allow for situations where the random variable W is unobserved (missing) in the calibration data.

Example 2 (Missing variable in the calibration data) While the goal is to compare the future cholesterol level of a new patient to that upon admission, the cholesterol level upon a first visit, W_i, may not be measured for former patients $i = 1, \ldots, n$ in the calibration set. This setting cannot be turned into a classification problem because the (W, Y) pair is never simultaneously observed for calibration samples. However, our method is still applicable with $\widehat{V}_{n+j} = V(X_{n+j}, W_{n+j})$ as the conditions in Theorem 3 still hold.

Another example that is closely related to the missing variable setting is to infer whether the counterfactual is larger or smaller than the realized outcome in causal inference.

Example 3 (Counterfactual inference) Predicting counterfactuals is another application whose setup is similar to Example 2 (Lei and Candès, 2021; Jin et al., 2023). Under the potential outcomes framework (Imbens and Rubin, 2015), we let $\{X_i, T_i, O_i(1), O_i(0)\}_{i=1}^N$ be i.i.d. tuples from an unknown super-population $\mathbb{P}$, where $X_i \in \mathcal{X}$ is the vector of covariates, $T_i \in \{0, 1\}$ is the treatment indicator, and $(O_i(1), O_i(0)) \in \mathbb{R}^2$ are potential outcomes under treatment and control, respectively. We observe $\{(X_i, O_i, T_i)\}_{i=1}^N$ where $O_i = O_i(T_i)$. We consider completely randomized experiments where for some $p \in (0, 1)$, $T_i \sim \text{Bern}(p)$ are independent of all other quantities. Counterfactual inference (e.g. predicting $O_i(0)$ when $T_i = 1$) asks what would happen should unit i receive another treatment status. A potential application is to find treated units with $O_i(1) \leq O_i(0)$ so that they might drop out early from the experiment to avoid adverse effects. In this case, we take all the control units $\{(X_i, O_i(0))\}_{i=1}^n$ as the calibration data, each i.i.d. from $\mathbb{P}_{X,O(0)\,|\,T=0}$. The test data are $\{X_{n+j}\}_{j=1}^m$ for which $\{O_{n+j}(0)\}_{j=1}^m$ are not observed. That is, we set the response as $Y = O(0)$ for all samples and $c_j = O_{n+j}(1)$ (the observed outcome) in (2) which is a random variable. This task cannot be turned into a classification problem because $(O_i(1), O_i(0))$ is never simultaneously observed for any unit. However, letting $\widehat{V}_{n+j} = V(X_{n+j}, c_j)$, the conditions in Theorem 3 still hold, because randomized treatments imply $\mathbb{P}_{X,Y\,|\,T=1} = \mathbb{P}_{X,O(0)} = \mathbb{P}_{X,Y\,|\,T=0}$. In this way, our method identifies multiple treated units among whom a prescribed proportion have negative individual treatment effects.

2.5 Asymptotic analysis and choice of nonconformity score

While the only requirement for the validity of `cfBH` is the monotonicity of the nonconformity score function, a carefully constructed score might enhance the power. This concerns two aspects: (i) what should the prediction machine pursue (as a function of x), and (ii) given

any output of the prediction machine, which nonconformity score (as a function of both x and y) should the practitioner use. We offer some heuristics through asymptotic analysis.

To simplify the discussion, we take the thresholds as $c_j = 0$, and assume the samples $\{(X_i, Y_i)\}_{i=1}^{n+m}$ are i.i.d. from an unknown distribution. In this case, the outcomes can be encoded into a binary variable $\mathbb{1}\{Y_i > 0\}$ for $i \in \mathcal{D}_{\text{calib}} \cup \mathcal{D}_{\text{test}}$. We thus consider the case $Y \in \{0, 1\}$, and define

$$\text{Power} = \mathbb{E}\left[\frac{\sum_{j=1}^{m} \mathbb{1}\{j \in \mathcal{R}, Y_{n+j} = 1\}}{1 \vee \sum_{j=1}^{m} \mathbb{1}\{Y_{n+j} = 1\}}\right].$$

We consider the regime where both n, the size of the calibration set, and m, the size of the test set, tend to infinity. Throughout, we hold the nonconformity score V as fixed, and the randomness is only in the calibration and test samples. The proof of the next proposition is in Appendix B.4.

Proposition 7 *Let V be any fixed monotone nonconformity score, and suppose $\{(X_i, Y_i)\}_{i=1}^{n+m}$ are i.i.d. Define $F(v, u) = \mathbb{P}(V(X, Y) < v) + u \cdot \mathbb{P}(V(X, Y) = v)$ for any $v \in \mathbb{R}$ and $u \in [0, 1]$. Define $t^* = \sup\left\{t \in [0, 1] \colon t/\mathbb{P}(F(V(X, 0), U) \le t) \le q\right\}$. Suppose that for any sufficiently small $\epsilon > 0$, there exists some $t \in (t^* - \epsilon, t^*)$ such that $t/\mathbb{P}(F(V(X, 0), U) \le t) < q$. Then the output $\mathcal{R}$ of `cfBH` satisfies*

$$\lim_{n,m\to\infty} \text{FDR} = \frac{\mathbb{P}\{F(V(X, 0), U) \le t^*, Y \le 0\}}{\mathbb{P}\{F(V(X, 0), U) \le t^*\}}, \quad \textit{and}$$
$$\lim_{n,m\to\infty} \text{Power} = \frac{\mathbb{P}\{F(V(X, 0), U) \le t^*, Y > 0\}}{\mathbb{P}\{Y > 0\}}.$$

In words, Proposition 7 states that the rejection threshold for conformal p-values in `cfBH` converges to t^*, leading to the convergence of FDR. By definition, t^* is the largest value of t such that the mass of $F(V(X, Y), U)$ below t is less than q times that of $F(V(X, 0), U)$. Since $F(v, u)$ is increasing in the first argument, t^* is large if $V(X, 0)$ has a far more heavier left tail than $V(X, Y)$. One such case is when the probability of $Y = 1$ is large for small $V(X, 0)$. For instance, if we use $V(x, y) := y - \widehat{\mu}(x)$ for a point prediction $\widehat{\mu}(\cdot)$, then t^* is large if $Y = 1$ is more probable for $X = x$ with large $\widehat{\mu}(x)$. The convergence is not necessarily true if the asymptotic FDR lies on the critical line near t^*, i.e., if there exists some $t_1 < t^*$ such that $t/\mathbb{P}(F(V(X, 0), U) \le t) = q$ for all $t \in [t_1, t]$. The technical condition in Proposition 7 rules out this situation, and actually guarantees that

$$q = \frac{t^*}{\mathbb{P}\{F(V(X, 0), U) \le t^*\}} = \frac{\mathbb{P}\{F(V(X, Y), U) \le t^*\}}{\mathbb{P}\{F(V(X, 0), U) \le t^*\}}$$

since the distribution of $F(V(X, 0), U)$ has no point mass. In particular, the asymptotic FDR of `cfBH` satisfies

$$\begin{aligned}\frac{\mathbb{P}\{F(V(X, 0), U) \le t^*, Y \le 0\}}{\mathbb{P}\{F(V(X, 0), U) \le t^*\}} &\le \frac{\mathbb{P}\{F(V(X, Y), U) \le t^*, Y \le 0\}}{\mathbb{P}\{F(V(X, 0), U) \le t^*\}} \\ &\le \frac{\mathbb{P}\{F(V(X, Y), U) \le t^*\}}{\mathbb{P}\{F(V(X, 0), U) \le t^*\}} \le q. \end{aligned} \tag{6}$$

To further simplify, we suppose the distribution of $V(X,Y)$ does not have a point mass, hence $F(u,v) = \mathbb{P}(V(X,Y) \le v)$. We also let $v^* = \sup\{v \colon \mathbb{P}(V(X,Y) \le v) \le t^*)$, such that $F(u,v) \le t^*$ if and only if $v \le v^*$. Thus, (6) becomes

$$\frac{\mathbb{P}\{V(X,0) \le v^*, Y = 0\}}{\mathbb{P}\{V(X,0) \le s^*\}} = \frac{\mathbb{P}\{V(X,Y) \le v^*, Y = 0\}}{\mathbb{P}\{V(X,0) \le s^*\}} \le \frac{\mathbb{P}\{V(X,Y) \le v^*\}}{\mathbb{P}\{V(X,0) \le v^*\}} \le q. \tag{7}$$

Choice of V. We first investigate (7) to provide some heuristics on the choice of V. If we could design some nonconformity score V such that

$$V(X,Y) \le v^* \quad \Rightarrow \quad Y = 0, \tag{8}$$

then the first inequality in (7) becomes an equality. Given a prediction $\widehat{\mu}(x)$ from any machine learning algorithm, if one would like to select individuals with larger values of $\widehat{\mu}(X_{n+j})$, one might design a nonconformity score V such that

$$V(x,0) = -\widehat{\mu}(x), \quad V(x,1) = +\infty.$$

In this way, selecting cases where $V(X_{n+j},0)$ is small is equivalent to selecting large $\widehat{\mu}(X_{n+j})$, and this choice guarantees (8) as long as $v^* < \infty$. We recommend a relaxation given by

$$V(x,y) = M \cdot y - \widehat{\mu}(x) \tag{9}$$

for some sufficiently large constant M. This "clipped" score obeys $\inf_x V(x,1) = M - \sup_x \widehat{\mu}(x) \ge \sup_x V(x,0)$ if $M \ge 2\sup_x |\widehat{\mu}(x)|$. That is, the nonconformity score for $Y = 1$ is always larger than that for $Y = 0$, regardless of the value of x. Recalling the definitions, we know $t^* \le q \cdot \mathbb{P}(F(V(X,0),U) \le t^*) \le q$. Thus, when $q < \mathbb{P}(Y = 0)$, by definition, v^* is smaller than the q-th quantile of $V(X,Y)$. As a result, (8) holds exactly and the first inquality in (7) is an equality—that is, using (9) could potentially yield a value of FDR close to the nominal level q, using up all the FDR budget; we thus anticipate a higer power. We indeed verify these heuristics in our simulations.

Choice of $\widehat{\mu}$. We then discuss the choice of $\widehat{\mu}$ when $V(x,y) = My - \widehat{\mu}(x)$ and (8) holds. Recall that given the conditions in Proposition 7, the last inequalities in (6) and (7) are exact equalities. Hence

$$\lim_{n,m\to\infty} \mathrm{FDR} = \frac{\mathbb{P}\{-\widehat{\mu}(X) \le v^*, Y = 0\}}{\mathbb{P}\{-\widehat{\mu}(X) \le v^*\}}, \quad \lim_{n,m\to\infty} \mathrm{Power} = \frac{\mathbb{P}\{-\widehat{\mu}(X) \le v^*, Y = 1\}}{\mathbb{P}\{Y = 1\}}.$$

Since our procedure always ensures that the asymptotic FDR is below q, letting $f(x) = v^* + \widehat{\mu}(x)$, we could view asymptotic power maximization as solving an optimization problem

$$\begin{aligned} \text{maximize} \quad & \mathbb{P}\{f(X) \ge 0, Y = 1\} \\ \text{subject to} \quad & \frac{\mathbb{P}\{f(X) \ge 0, Y = 0\}}{\mathbb{P}\{f(X) \ge 0\}} \le q. \end{aligned}$$

Equivalently, this is

$$\text{maximize} \quad \mathbb{E}\big[\mathbb{1}\{f(X) \ge 0\}\mathbb{P}(Y = 1 \,|\, X)\big]$$

$$\text{subject to} \quad \mathbb{E}\big[\mathbb{1}\{f(X) \geq 0\}\big(\mathbb{P}(Y = 0 \,|\, X) - q\big)\big] \leq 0.$$

By Neyman-Pearson lemma, the optimal choice of f should be a monotone function of $\mathbb{P}(Y = 1 \,|\, X)$. That is, we should aim for some $\widehat{\mu}(x)$ that is monotone in $\mathbb{P}(Y = 1 \,|\, X = x)$. The most convenient option is to fit $\mathbb{P}(Y = 1 \,|\, X = x)$. This heuristic derivation leads to a quite intuitive recommendation: the predicted score should indeed aim to reflect how likely $Y = 1$ is given X.

3. Numerical experiments

We evaluate our method on simulated datasets, leading to some practical suggestions. We generate i.i.d. covariates $X_i \sim$Unif $[-1, 1]^{20}$ and responses $Y_i = \mu(X_i) + \epsilon_i$, where $\mu(x) = \mathbb{E}[Y \,|\, X = x]$ is nonlinear in x, and ϵ_i is the independent random noise. We design 8 simulation settings to demonstrate the performance of our methods under various data generating processes, with different configurations of $\mu(\cdot)$ and distributions for the ϵ_i's. In particular, we vary (i) whether the image $\{\mu(x) \colon x \in [-1, 1]^{20}\}$ is a continuous set, and (ii) whether the noise is heterogeneous, such that the hardness of correctly identifying those outcomes exceeding zero varies. The details of all settings are summarized in Appendix C.2. The reproduction codes for this part can be found at `https://github.com/ying531/selcf_paper`.

The task is to select individuals with $Y_{n+j} > 0$ among all test samples. We fix the sizes of training and calibration data at $n = |\mathcal{D}_{\text{train}}| = |\mathcal{D}_{\text{calib}}| = 1000$ and vary the test sample size $|\mathcal{D}_{\text{test}}| \in \{10, 100, 500, 1000\}$. We use gradient boosting, SVM with `rbf` kernel, and random forest to fit a regression model $\widehat{\mu}(\cdot)$ for $\mathbb{E}[Y \,|\, X]$, all from the `scikit-learn` Python library without fine tuning. We then apply `cfBH` and Algorithm 2 at the FDR target $q = 0.1$, which, together with the Bonferroni baseline, leads to four algorithm configurations:

1. `BH_sub`: `cfBH0` (Algorithm 2) with $\widehat{V}_{n+j} = -\widehat{\mu}(X_{n+j})$;
2. `BH_res`: `cfBH` with $V(x, y) = y - \widehat{\mu}(x)$;
3. `BH_clip`: `cfBH` with $V(x, y) = M \cdot \mathbb{1}\{y > 0\} - \widehat{\mu}(x)$ and a large constant $M = 100$; this value is chosen to ensure it is larger than $2 \sup_x |\widehat{\mu}(x)|$;
4. `Bonferroni`: Select all $p_j \leq q/m$ with $V(x, y)$ the same as in `BH_clip`.

Algorithm 2 used in `BH_sub` is formally introduced in Appendix A; when applied to classification problems, it is equivalent to the score-based methods of Mary and Roquain (2021) and Rava et al. (2021). The only difference from `cfBH` is that Algorithm 2 (`cfBH0`) uses $\{(X_i, Y_i) \colon i \in \mathcal{D}_{\text{calib}},\ Y_i = 0\}$ as the calibration data when constructing conformal p-values (4), which leads to a slightly stronger theoretical guarantee although it comes at a price: loss of power (see Appendix A.2). Other than this, we are not aware of alternative methods for exact control of FDR in classification.

3.1 Valid FDR control

We empirically evaluate the FDR by averaging the FDP $\frac{\sum_{j \in \mathcal{D}_{\text{test}}} \mathbb{1}\{j \in \mathcal{R}, Y_{n+j} > 0\}}{1 \vee |\mathcal{R}|}$ over $N = 1000$ independent runs ($\mathcal{R}$ is the rejection set). We observe similar power and FDR for different values of n_{test}, hence we only plot the results for $q = 0.1$ and $n_{\text{test}} = 100$ in

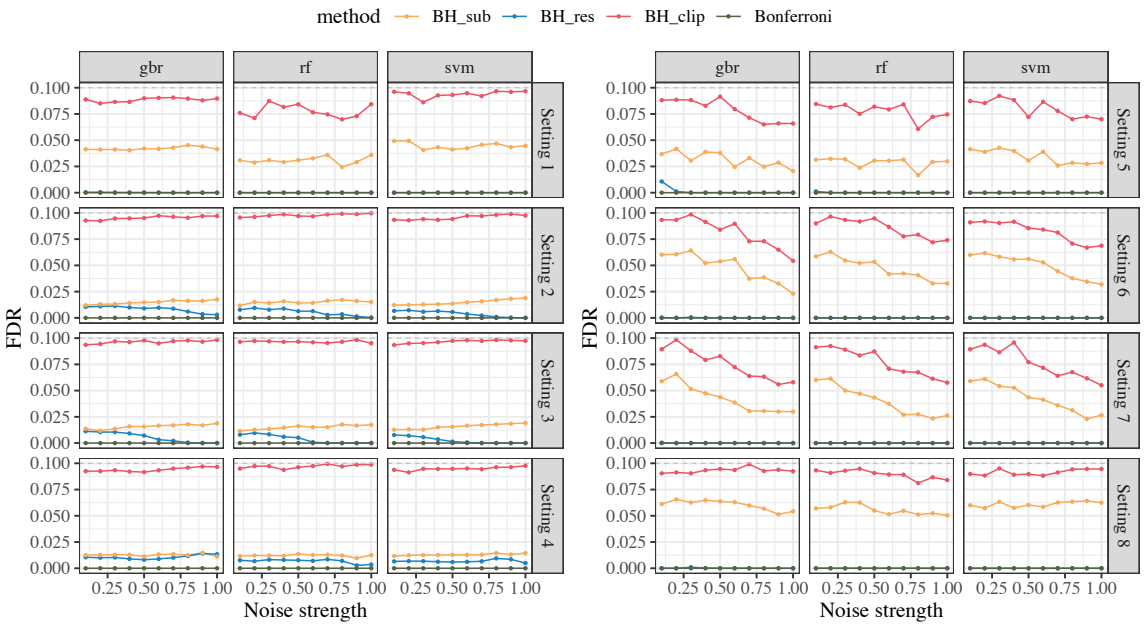

Figure 2: Realized FDR for four procedures at FDR target $q = 0.1$. Each row corresponds to one data generating process, and each column corresponds to one regressor (`gbr` for gradient boosting, `rf` for random forest, `svm` for support vector machine). The x-axis is the parameter σ for the noise level of ϵ_i, whose precise definition is in Appendix C.2.

Figure 2. The FDR is controlled below $q = 0.1$ in all configurations, showing the validity of our procedure. In particular, the FDR of `Bonferroni` is always close to zero.

Among the three nonconformity scores, the realized FDR of `BH_clip` is the highest in all settings, and is often very close to the nominal level. FDR also varies with the regression algorithms that are adopted, but the variation is not very large.

In settings 1-4 and 8, the FDR levels of different methods are relatively stable across various noise strengths. In settings 5-7, the FDR decreases as the noise level (the x-axis) increases. It might seem counterintuitive because at first sight, one might think that a harder problem (larger noise) would lead to a higher error rate. However, we observe that it is accompanied by lower power (Figure 3 as we will present shortly) and a smaller rejection set (Figure 10 in Appendix C.3). This might be contributed by two factors. The first is the increased difficulty of prediction; with larger noise, machine learning is less capable of capturing the heterogeneity in the true conditional mean function $\mu(X_{n+j})$. Since a test sample needs to have a sufficiently small value of $V(X_{n+j}, 0)$ to be selected, such lack of heterogeneity leads to small selection sets. The second is the decreased confidence even with ground truth available: even when $\mu(X_{n+j})$ is known, when the noise is too large, there is hardly any value of the covariate for which one has a high confidence that $Y > 0$. To be more specific, when $c_j = \tau$ is a constant, the selection set is fully decided by $\mathcal{D}_{\text{calib}} \cup \{X_{n+j}\}_{j \in \mathcal{D}_{\text{test}}}$; in addition, $\{Y_{n+j}\}_{j \in \mathcal{D}_{\text{test}}}$ are independent of $\mathcal{D}_{\text{calib}}$ conditional on

$\{X_{n+j}\}_{j\in\mathcal{D}_{\text{test}}}$ by the i.i.d. assumption. The tower property then implies

$$\text{FDR} = \mathbb{E}\left[\frac{\sum_{j=1}^{m}\mathbb{1}\{j\in\mathcal{R}\}\mathbb{1}\{Y_{n+j}\le 0\}}{1\vee|\mathcal{R}|}\right] = \mathbb{E}\left[\frac{\sum_{j=1}^{m}\mathbb{1}\{j\in\mathcal{R}\}\mathbb{P}\big(Y_{n+j}\le 0\,|\,X_{n+j}\big)}{1\vee|\mathcal{R}|}\right] \quad (10)$$

which is roughly the average of $\mathbb{P}(Y_{n+j}\le 1\,|\,X_{n+j})$ among selected individuals. Thus, when $\mathbb{P}(Y\le 0\,|\,X=x) > q$ for almost all $x\in\mathcal{X}$, sometimes one needs to output $\mathcal{R}=\varnothing$ in order to keep the FDR below q, leading to smaller selection sets. In general, as the selection becomes difficult, selected units should have extremely small nonconformity scores and extremely strong confidence in a positive response, resulting in a lower FDR.

3.2 Power

We evaluate power by averaging $\frac{\sum_{j\in\mathcal{D}_{\text{test}}}\mathbb{1}\{j\in\mathcal{R},Y_{n+j}>0\}}{\sum_{j\in\mathcal{D}_{\text{test}}}\mathbb{1}\{Y_{n+j}>0\}}$, the proportion of correct selections among all positive test samples, over all replicates. We again observe stable power across different values of n_{test}, hence we only plot the average power for $q=0.1$ and $n_{\text{test}}=100$ in Figure 3. `BH_clip` always has the highest power while `BH_res` always has the lowest power (excluding `Bonferroni`); `BH_sub` is sometimes closer to `BH_clip` and sometimes closer to `BH_res`. We note that the general applicability of `BH_res` comes with its low power in such binary classification problems, while the other two (which are only applicable for fixed thresholds) are more powerful. Finally, the Bonferroni correction nearly has no power (even if we set $V(x,y)$ to be the same as the most powerful `BH_clip`).

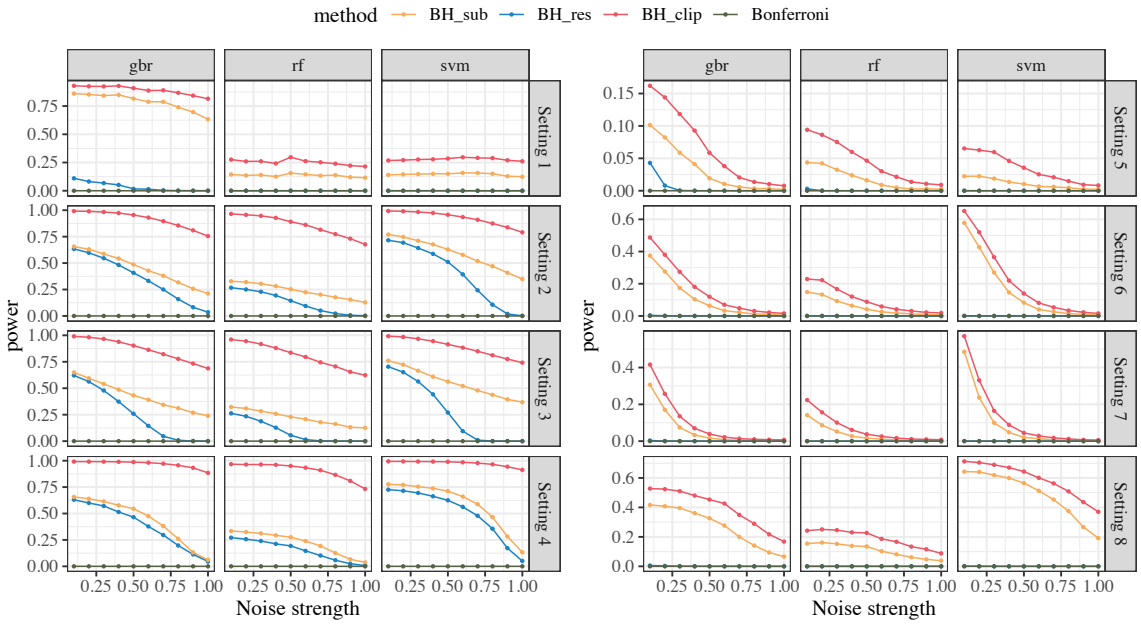

Figure 3: Realized power of four procedures at FDR target $q=0.1$ for various data generating processes. Details are otherwise the same as Figure 2.

The power of our procedure also varies with the prediction algorithms (the columns). In setting 1 where the FDR is similar across the three algorithms, the power is actually

drastically different: `gbr` performs the best in most settings; however, `svm` performs the best in settings 4 and 8 where there is strong heterogeneity in the distribution of $\epsilon_i \,|\, X_i$, in which case the other two prediction algorithms might fail to capture the true dependence between Y and X.

The power decreases as the noise strength increases in all settings. This is because larger noise makes it more difficult to fit the prediction model, and the fundamental detection hardness increases as sketched in (10). (10) also implies that in practice, the FDR target q should be properly chosen: in situations where $\mathbb{P}(Y_{n+j} > 0)$ is small, it might be too demanding to choose a very small q since X_{n+j} for which $\mathbb{P}(Y_{n+j} > 0 \,|\, X_{n+j})$ is very large may not exist. A practical choice might be some q that is moderately larger than the marginal proportion of positive Y's in the training data (our method is still valid). In this way, when there exists some region in $\mathcal{X}$ where $\mathbb{P}(Y > 0 \,|\, X) \geq 1 - q$, our method finds critical subsets while achieving some power.

4. Real data application

4.1 Candidate screening in recruitment

We apply `cfBH` as an automatic screening tool in recruiting, where a human resources staff uses machine learning prediction to screen all applicants and shortlist some for subsequent test and interviews. In this application, machine learning is used to predict whether a new candidate is qualified for the job (i.e., whether the recruitment is successful); a higher predicted value might indicate a better fit to the position, but no guarantee can be provided for a black-box prediction machine. We will use `cfBH` to calibrate the prediction and generate a shortlist of candidates with rigorous FDR control, i.e., limiting the proportion of unqualified individuals among the selected candidates.

We assume new applicants to the position and previous applicants on record (such as those who applied last year) are i.i.d. from the same distribution. This is reasonable if the pool of applicants for the position is stable over the years. The recruiters may train any prediction model on previous applicants and use any monotone nonconformity score as their choice. We use a small-scale recruitment dataset from Kaggle (Roshan, 2020), as recruitment datasets from companies are often confidential. There are $n_{\text{tot}} = 215$ samples in total. Each sample is from an applicant for the position; the data includes covariates about their education, work experience, gender, specialization, etc., and the response is a binary variable indicating whether the applicant is finally offered the job. Here, we use this binary outcome as a perfect proxy for the qualification of a candidate. We randomly split the data into a training set of size $|\mathcal{D}_{\text{train}}| = 86$ and a test set of size $|\mathcal{D}_{\text{test}}| = 43$. We first train a gradient boosting model to predict the job offering, using the `sckit-learn` Python library without fine tuning, and apply the three procedures (except Bonferroni since it is less powerful) in Section 3 for $q =\in \{0.1, 0.2, 0.5\}$. We plot the false discovery proportion (FDP) and power over $N = 100$ independent runs in Figures 4 and 5, respectively.

All three scores achieve valid FDR control (averaging the FDPs). `BH_res` and `BH_clip` have similar FDP (hence FDR). The FDP of `BH_sub` is lower, potentially because of its low power, and also less stable than the other two. As `BH_sub` only uses the subset of training samples with $Y = 0$ to calibrate the selection set as Bates et al. (2021); Rava et al. (2021) did (see discussion in Appendix B.5), when such subset is small, the calibration, hence the

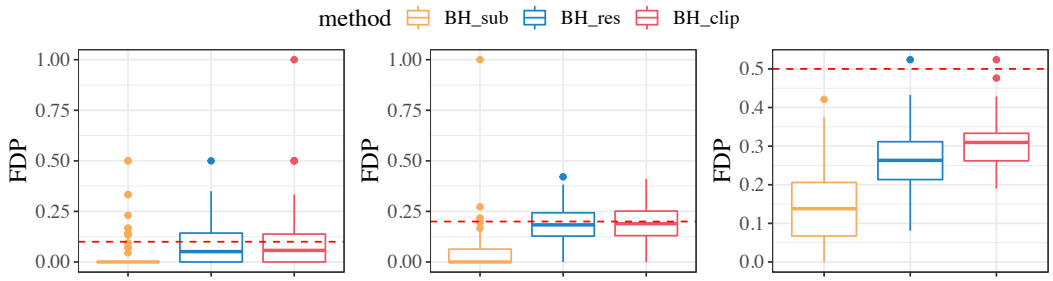

Figure 4: Boxplot for false discovery proportions over $N = 100$ independent runs for the recruitment dataset, with $q = 0.1$ (left), $q = 0.2$ (middle), and $q = 0.5$ (right). Red dashed lines are the nominal levels.

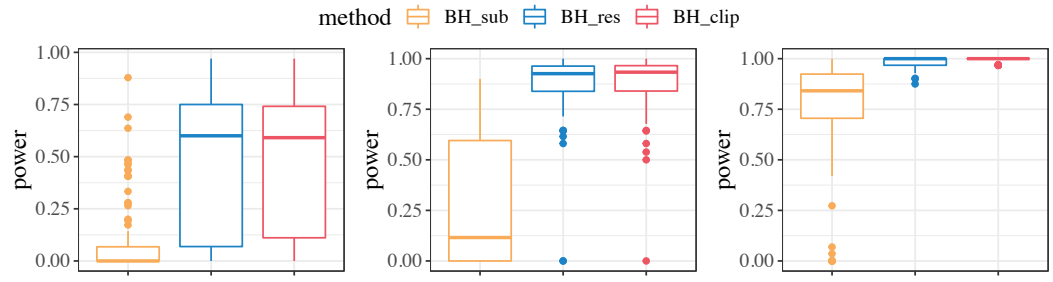

Figure 5: Boxplot for power over $N = 100$ independent runs for the recruitment dataset, with $q = 0.1$ (left), $q = 0.2$ (middle), and $q = 0.5$ (right).

FDP, can be unstable. Intuitively, `BH_res` and `BH_clip` achieve a more stable behavior by using all the calibration data.

The power of these methods differ more significantly. In general, predicting qualified candidates from our data is a relatively easy task: once we allow the FDR level to be 0.2 or 0.5, `BH_res` and `BH_clip` could almost identify all qualified candidates. Both these methods achieve similar power while `BH_clip` is again a little better. However, `BH_sub`, which only uses training samples with $Y = 0$ to calibrate the selection set as Bates et al. (2021) and Rava et al. (2021) did, has much lower power. We discuss this issue in Appendix A.1: in finding positives for a binary response, our method can be more powerful than `BH_sub` when there are many positive samples in the population.

4.2 Drug discovery

We apply `cfBH` to therapeutic datasets for drug discovery, focusing on two tasks: (i) selecting molecules that bind to a target protein for a certain disease, and (ii) selecting drug-target (molecule-protein) pairs with a high affinity score. Our main focus is to calibrate any given prediction model to limit false positives. Therefore, we use the pre-trained models and the prediction pipelines established in the DeepPurpose library (Huang et al., 2020).

4.2.1 Drug property prediction for HIV

We first consider the task of predicting drug properties for a certain protein target for HIV. As we mentioned in the introduction, given a specific target, machine learning models are often trained on a representative subset of the whole drug library screened by HTS, and then used to predict the activity of the remaining proteins to find promising candidates. It is important to control for false positives in the shortlisted candidates.

We use the HIV screening dataset with a total size of $n_{\text{tot}} = 41127$. We randomly split the data into three folds with ratio $6 : 2 : 2$ in size. The first two folds contain binary outcomes indicating whether the drugs interact with the disease. We use the first fold to train a machine learning model to predict the outcome, where the drugs are encoded into numerical features using Extended-Connectivity FingerPrints (ECFP) that characterize topological properties of molecules and compounds. We train a small neural network in only 3 epochs so that the whole procedure works well with CPUs; using more complicated or pre-trained networks might improve power but this is not the main focus here. The second fold serves as the calibration data. Our goal is to find active proteins in the last test fold while controlling for the false discovery rate.

In the training fold, about 3% of the drugs are active for the HIV disease. We choose the FDR levels among $q \in \{0.1, 0.2, 0.5\}$. We compute the empirical FDR, power, and average size of the selection set over $N = 100$ independent runs of the procedure in Table 1.

	FDR			Power			$\lvert\mathcal{R}\rvert$		
Level q	0.1	0.2	0.5	0.1	0.2	0.5	0.1	0.2	0.5
`BH_clip`	0.0957	0.196	0.495	0.0788	0.174	0.410	26.5	64.2	240
`BH_res`	0.0989	0.196	0.494	0.0766	0.174	0.410	25.8	64.4	239
`BH_sub`	0.0862	0.192	0.474	0.0739	0.169	0.397	24.8	61.8	222

Table 1: FDR and power of the three methods averaged over $N = 100$ random splits.

All three choices of nonconformity scores control FDR below the nominal levels. Their performance is also similar, while `BH_clip` has the highest power and `BH_sub` is the least powerful, both with a small margin. This is because the positive samples in the population is extremely small, so that using $Y = 0$ samples or the whole calibration set does not have a huge impact on the selection set.

Using all three methods, the selection set consists of all test samples whose predicted binding affinity is above some value. This value is specific to the training model we use. Figure 6 shows the selection threshold of the predicted value for all configurations. If we control FDR at $q = 0.1$, the predicted scores needs to be as large as 0.8 to be considered promising; this leads to around 25 candidates among about 8000 test samples. However, if we set $q = 0.5$, then the thresholds are in the range $[0.2, 0.4]$ most of the time: a moderately large score is sufficient to stand out.

4.2.2 Drug-target interaction (DTI) prediction

Last, we consider the task of predicting drug-target interactions (DTI) among a huge pool of drug-target pairs. This might be of use to a therapeutic company to prioritize its resources in

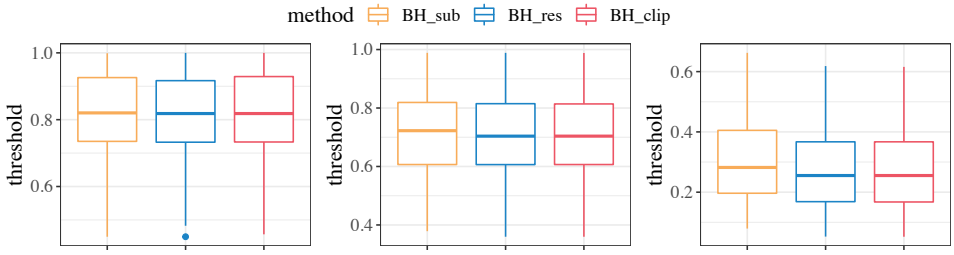

Figure 6: Selection thresholds over $N = 100$ runs, with $q = 0.1$ (left), $q = 0.2$ (middle), and $q = 0.5$ (right).

developing drugs that might be effective for any of the targets they happen to be interested in. In this application, one may need to be cautious about the i.i.d. assumption: this can be reasonable if the drugs and targets are drawn from a diverse library, and a representative subset of all pairs have been screened to form the training data.

We use the DAVIS dataset published in Davis et al. (2011), which records real-valued binding affinities for $n_{\text{tot}} = 30060$ drug-target pairs. In this application, we mimic a scenario where a small proportion of the whole library has been screened, and one would like to find promising ones among a huge amount of pairs whose binding affinities are unknown. In particular, we randomly split the dataset into three folds of size $2 : 2 : 6$; we use the first fold for training the model, the scond for calibration, and the last largest set as test samples. We use ECFP and Conjoint triad feature (CTF) (Shen et al., 2007; Shao et al., 2009) to encode the drugs and the targets into numeric features, respectively. We train a small neural network over 10 epochs. These choices are suitable for experiments on CPUs (one might of course use other more computationally intensive alternatives).

Because the affinity is continuously valued, and to account for the heterogeneity in targets, we set c_j as the q_{pop}-th quantile of the outcomes of the training samples with the same binding target as sample j, where $q_{\text{pop}} \in \{0.7, 0.8, 0.9\}$. Given a predicted score, there is no natural way to use `BH_sub` in this setting, so we test the following two methods:

- `BH_res` with nonconformity score $V(x, y) = y - \widehat{\mu}(x)$,
- `BH_clip` with nonconformity score $V(x, y; c) = M\mathbb{1}\{y \geq c\} + c\mathbb{1}\{y < c\}$ for $M = 100$,

where x is the vector of features, y is the binding affinity ranging in $[5, 10]$, and c is the threshold that is computable for both calibration and test samples. We set the FDR level at $q \in \{0.1, 0.2, 0.5\}$. There are 18 configurations in total, with 2 nonconformity scores and 3×3 combinations of (q_{pop}, q). In this case, since the threshold c_j varies among the samples, the selection is not monotone in the predicted score.

The empirical FDP, computed as $\frac{\sum_{j=1}^{m} \mathbb{1}\{Y_{n+j} \leq c_j\}}{1 \vee |\mathcal{R}|}$ ($\mathcal{R}$ is the selection set), over $N = 100$ independent runs is plotted in Figure 7. For all configurations of q_{pop}, both methods control the FDR (average of FDPs) at the nominal level. However, there can be some variation in the FDP for $q = 0.1$; `BH_res` is less stable than `BH_clip`. Also, the FDR from `BH_clip` is very close to the nominal level while that from `BH_res` is much lower. This is due to the low power of `BH_res` as we show in the power plot (Figure 8).

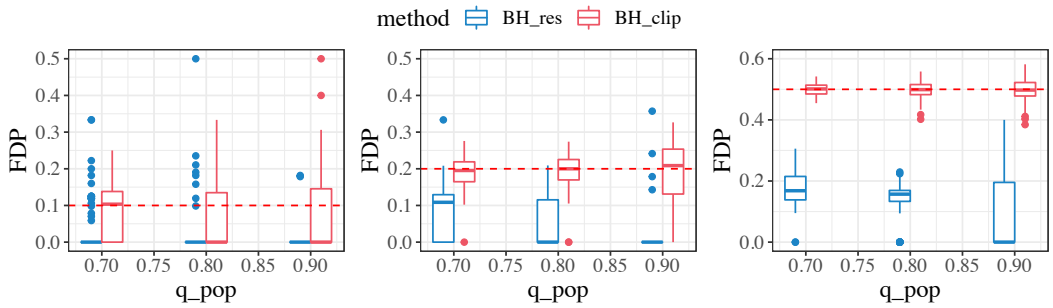

Figure 7: FDP over $N = 100$ runs, with $q = 0.1$ (left), $q = 0.2$ (middle), $q = 0.5$ (right).

The power of both methods decrease as q_{pop} increases, which is natural because this leads to higher thresholds for binding affinity. In all settings, `BH_clip` is more powerful and also yields larger selection sets. Thus, when the thresholds are computable for both the calibration and test samples (as Example 1), we recommend `BH_clip` for higher power.

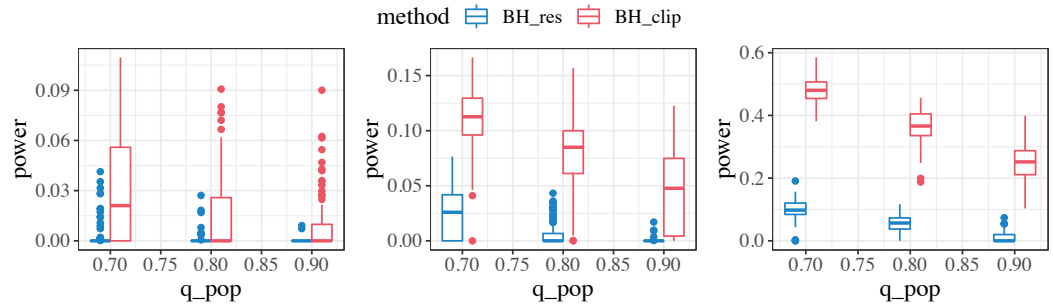

Figure 8: Power over $N = 100$ runs, with $q = 0.1$ (left), $q = 0.2$ (middle), $q = 0.5$ (right).

5. Discussion

In this paper, we introduce `cfBH`, which is a generic tool to turn any prediction model into a selection threshold for interesting outcomes. By constructing conformal p-values based on i.i.d. calibration data and leveraging multiple testing ideas, we guarantee that a prescribed proportion of the selected set is indeed of interest. Controlling the false discovery rate ensures efficient use of resources for follow-up investigations.

A crucial condition that `cfBH` relies on is that the calibration and test samples are i.i.d. or exchangeable. However, in practice, the two datasets might differ because of selection or distribution shift. For example, to infer the performance of this year's job candidates, last years' candidates that are documented might in general be more competent than average; to infer new drugs, the drugs that have been screened by HTS might be selected with varying preference based on the features; drug discovery also needs to deal with domain shift (repurposing) for completely unseen targets. Reliable selection under distribution shift, if not infeasible, may require more involved techniques.

FDR, as a measure of Type-I error, may be limited in applications such as healthcare, where both type-I and type-II errors are of concern. Therefore, it might also be interesting to

see whether our methodology can be extended to controlling a mixture of both error types. Meanwhile, counting the number of errors may be less sensible if the cost of making an error varies with individuals or depends on the outcomes. Developing calibration methods to control general risks in screening procedures is also an interesting direction to pursue.

Acknowledgement

We are grateful to two anonymous referees and the action editor for valuable comments and suggestions. We thank John Cherian, Issac Gibbs, Jayoon Jang, Lihua Lei, Shuangning Li, Zhimei Ren, Hui Xu, and Qian Zhao for helpful discussions. Y.J. would like to specially thank John Cherian for inspiring discussion on the applications. E.C. and Y.J. were supported by the Office of Naval Research grant N00014-20-1-2157, the National Science Foundation grant DMS-2032014, the Simons Foundation under award 814641, and the ARO grant 2003514594.

References

Ernst Ahlberg, Oscar Hammar, Claus Bendtsen, and Lars Carlsson. Current application of conformal prediction in drug discovery. *Annals of Mathematics and Artificial Intelligence*, 81(1):145–154, 2017a.

Ernst Ahlberg, Susanne Winiwarter, Henrik Boström, Henrik Linusson, Tuve Löfström, Ulf Norinder Ulf Johansson, Ola Engkvist, Oscar Hammar, Claus Bendtsen, and Lars Carlsson. Using conformal prediction to prioritize compound synthesis in drug discovery. In *Conformal and Probabilistic Prediction and Applications*, pages 174–184. PMLR, 2017b.

Soumaya Amdouni and Wahiba Ben abdessalem Karaa. Web-based recruiting. In *ACS/IEEE International Conference on Computer Systems and Applications-AICCSA 2010*, pages 1–7. IEEE, 2010.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *arXiv preprint arXiv:2104.08279*, 2021.

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

Yoav Benjamini and Yosef Hochberg. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418, 1997.

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

Robert Bohrer. Multiple three-decision rules for parametric signs. *Journal of the American Statistical Association*, 74(366a):432–437, 1979.

Robert Bohrer and Mark J Schervish. An optimal multiple decision rule for signs of parameters. *Proceedings of the National Academy of Sciences*, 77(1):52–56, 1980.

Emmanuel J Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*, 2021.

Paula Carracedo-Reboredo, Jose Liñares-Blanco, Nereida Rodríguez-Fernández, Francisco Cedrón, Francisco J Novoa, Adrian Carballal, Victor Maojo, Alejandro Pazos, and Carlos Fernandez-Lozano. A review on machine learning approaches and trends in drug discovery. *Computational and Structural Biotechnology Journal*, 19:4538–4558, 2021.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48), 2021.

Kristin M Corey, Sehj Kashyap, Elizabeth Lorenzi, Sandhya A Lagoo-Deenadayalan, Katherine Heller, Krista Whalen, Suresh Balu, Mitchell T Heflin, Shelley R McDonald, Madhav Swaminathan, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (pythia): A retrospective, single-site study. *PLoS medicine*, 15(11):e1002701, 2018.

Isidro Cortés-Ciriano and Andreas Bender. Concepts and applications of conformal prediction in computational drug discovery. *arXiv preprint arXiv:1908.03569*, 2019.

Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Babu, and Mohamed Jawed Ahsan. Machine learning in drug discovery: A review. *Artificial Intelligence Review*, pages 1–53, 2021.

Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

Thorsten Dickhaus. Multiple testing and binary classification. In *Simultaneous Statistical Inference*, pages 91–101. Springer, 2014.

Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. The application of conformal prediction to the drug discovery process. *Annals of Mathematics and Artificial Intelligence*, 74(1):117–132, 2015.

Ruth Etzioni, Nicole Urban, Scott Ramsey, Martin McIntosh, Stephen Schwartz, Brian Reid, Jerald Radich, Garnet Anderson, and Leland Hartwell. The case for early detection. *Nature reviews cancer*, 3(4):243–252, 2003.

Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, and Giannis Tzimas. Application of machine learning algorithms to an online recruitment system. In *Proc. International Conference on Internet and Web Applications and Services*, pages 215–220. Citeseer, 2012.

Evanthia Faliagka, Lazaros Iliadis, Ioannis Karydis, Maria Rigou, Spyros Sioutas, Athanasios Tsakalidis, and Giannis Tzimas. On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed cv. *Artificial Intelligence Review*, 42(3):515–528, 2014.

U.S. FDA. The drug development process, 2018. URL `https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process`.

Wenge Guo, Sanat K Sarkar, and Shyamal D Peddada. Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics*, 66(2):485–492, 2010.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Emily Heaslip. Ai tools for talent acquisition to help you hire, 2022. URL `https://vervoe.com/ai-tools-for-talent-acquisition/`.

Yosef Hochberg. Multiple classification rules for signs of parameters. *Journal of statistical planning and inference*, 15:177–188, 1986.

Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. Deeppurpose: A deep learning library for drug-target interaction prediction. *Bioinformatics*, 2020.

Ziwei Huang. *Drug discovery research: new frontiers in the post-genomic era*. John Wiley & Sons, 2007.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Ali Jalali, Hannah Lonsdale, Nhue Do, Jacquelin Peck, Monesha Gupta, Shelby Kutty, Sharon R Ghazarian, Jeffrey P Jacobs, Mohamed Rehman, and Luis M Ahumada. Deep learning for improved risk prediction in surgical outcomes. *Scientific reports*, 10(1):1–13, 2020.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120, 2023.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1):D1388–D1395, 2021.

Alexios Koutsoukas, Keith J Monaghan, Xiaoli Li, and Jun Huan. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of cheminformatics*, 9(1):1–13, 2017.

Samuel Lampa, Jonathan Alvarsson, Staffan Arvidsson Mc Shane, Arvid Berg, Ernst Ahlberg, and Ola Spjuth. Predicting off-target binding profiles with confidence using conformal prediction. *Frontiers in pharmacology*, 9:1256, 2018.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021.

Martin Lindh, Anders Karlén, and Ulf Norinder. Predicting the rate of skin penetration using an aggregated conformal prediction framework. *Molecular Pharmaceutics*, 14(5): 1571–1576, 2017.

Ricardo Macarron, Martyn N Banks, Dejan Bojanic, David J Burns, Dragan A Cirovic, Tina Garyantes, Darren VS Green, Robert P Hertzberg, William P Janzen, Jeff W Paslay, et al. Impact of high-throughput screening in biomedical research. *Nature reviews Drug discovery*, 10(3):188–195, 2011.

David Mary and Etienne Roquain. Semi-supervised multiple testing. *arXiv preprint arXiv:2106.13501*, 2021.

Malgorzata Mochol, Holger Wache, and Lyndon Nixon. Improving the accuracy of job search with semantic techniques. In *International Conference on Business Information Systems*, pages 301–313. Springer, 2007.

Kazem Rahimi, Derrick Bennett, Nathalie Conrad, Timothy M Williams, Joyee Basu, Jeremy Dwight, Mark Woodward, Anushka Patel, John McMurray, and Stephen MacMahon. Risk prediction in patients with heart failure: a systematic review and analysis. *JACC: Heart Failure*, 2(5):440–446, 2014.

Bradley Rava, Wenguang Sun, Gareth M James, and Xin Tong. A burden shared is a burden halved: A fairness-adjusted approach to classification. *arXiv preprint arXiv:2110.05720*, 2021.

Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):1–9, 2020.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Etienne Roquain and Nicolas Verzelen. False discovery rate control with unknown null distribution: Is it possible to mimic the oracle? *The Annals of Statistics*, 50(2):1095–1123, 2022.

Ben Roshan. Campus recruitment, 2020. Academic and Employability Factors influencing placement, `https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement`.

Roshni Sahoo, Shengjia Zhao, Alyssa Chen, and Stefano Ermon. Reliable decisions with threshold calibration. *Advances in Neural Information Processing Systems*, 34, 2021.

Clayton Scott, Gowtham Bellala, and Rebecca Willett. The false discovery rate for statistical pattern recognition. *Electronic Journal of Statistics*, 3:651–677, 2009.

Xiaojian Shao, Yingjie Tian, Lingyun Wu, Yong Wang, Ling Jing, and Naiyang Deng. Predicting dna-and rna-binding proteins from sequences with kernel methods. *Journal of Theoretical Biology*, 258(2):289–293, 2009.

Muhammad Ahmad Shehu and Faisal Saeed. An adaptive personnel selection model for recruitment using domain-driven data mining. *Journal of Theoretical and Applied Information Technology*, 91(1):117, 2016.

Juwen Shen, Jian Zhang, Xiaomin Luo, Weiliang Zhu, Kunqian Yu, Kaixian Chen, Yixue Li, and Hualiang Jiang. Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 104(11):4337–4341, 2007.

Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):1–12, 2019.

Roman Sink, Stanislav Gobec, S Pecar, and Anamarija Zega. False positives in the early stages of drug discovery. *Current medicinal chemistry*, 17(34):4231–4255, 2010.

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

Fredrik Svensson, Ulf Norinder, and Andreas Bender. Improving screening efficiency through iterative screening using docking and conformal prediction. *Journal of chemical information and modeling*, 57(3):439–444, 2017.

Fredrik Svensson, Natalia Aniceto, Ulf Norinder, Isidro Cortes-Ciriano, Ola Spjuth, Lars Carlsson, and Andreas Bender. Conformal regression for quantitative structure–activity relationship modeling—quantifying prediction uncertainty. *Journal of Chemical Information and Modeling*, 58(5):1132–1140, 2018.

Paweł Szymański, Magdalena Markowicz, and Elżbieta Mikiciuk-Olasik. Adaptation of high-throughput screening in drug discovery—toxicological screening tests. *International journal of molecular sciences*, 13(1):427–452, 2011.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. 1999.

Lequn Wang, Thorsten Joachims, and Manuel Gomez Rodriguez. Improving screening processes via calibrated subset selection. *arXiv preprint arXiv:2202.01147*, 2022.

Asaf Weinstein and Aaditya Ramdas. Online control of the false coverage rate and false sign rate. In *International Conference on Machine Learning*, pages 10193–10202. PMLR, 2020.

Appendix A. Connections to the literature

This section provides a detailed comparison of our method to related ones in the literature. In Section A.1, we present a variant of our method that applies to classification problems and controls FDR conditional on the test responses. We then compare it to the score-based methods of Mary and Roquain (2021) and Rava et al. (2021) for classification problems in Section A.2. In Section A.3, we show how the outlier detection problem in Bates et al. (2021) with conformal p-values can be turned into the classification setting and discuss its connection to our methods.

A.1 A variant for hypotheses-conditional FDR control

A variant of Algorithm 1 provides a slightly stronger guarantee, FDR control for test data conditional on $\mathcal{H}_0$. This variant applies to binary or one-versus-all classification; it also applies to our directional selection with constant thresholds, because it can be turned into a classification problem. The classification setting is close to that of Rava et al. (2021), while our analysis through conformal p-values offers a complementary perspective.

We assume access to i.i.d. training data $\{(X_i, L_i)\}_{i=1}^n$ where L_i is the label of unit i and $X_i \in \mathcal{X}$ is the features. We also assume the covariates of the testing sample $\{X_{n+j}\}_{j=1}^m$ are observed, but not the label. Suppose one is interested in finding a subset $\mathcal{R}$ of test samples whose labels are in some user-specified class $\mathcal{C}$, while controlling the FDR, defined as

$$\mathrm{FDR} := \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{1}\{L_j \notin \mathcal{C}, j \in \mathcal{R}\}}{1 \vee |\mathcal{R}|}\right],$$

where the expectation is with respect to the randomness in all the training and testing data. One could encode a binary label $Y = \mathbb{1}\{L \in \mathcal{C}\}$, which turns it into our setting with $c_j \equiv 0.5$ in (2).

Since the label is observable in the calibration fold, we choose a subset $\mathcal{D}^0_{\mathrm{calib}} = \{i \in \mathcal{D}_{\mathrm{calib}} \colon Y_i = 0\}$ and let $n^0 = |\mathcal{D}^0_{\mathrm{calib}}|$. We construct conformal p-values via

$$p_j^0 = \frac{\sum_{i \in \mathcal{D}^0_{\mathrm{calib}}} \mathbb{1}\{\widehat{V}_i < \widehat{V}_{n+j}\} + (1 + \sum_{i \in \mathcal{D}^0_{\mathrm{calib}}} \mathbb{1}\{\widehat{V}_i = \widehat{V}_{n+j}\}) \cdot U_j}{n^0 + 1}, \tag{11}$$

where $U_j \sim \mathrm{Unif}\,[0,1]$ are i.i.d. random variables. The construction of p-values in (11) differs from (4) in that we use a smaller calibration set $\mathcal{D}^0_{\mathrm{calib}}$, and compare the test nonconformity scores $\widehat{V}_{n+j}$ to $\widehat{V}_i = V(X_i, 0)$, instead of $V_i = V(X_i, Y_i)$ for $i \in \mathcal{D}^0_{\mathrm{calib}}$. We then run BH with $\{p_j^0\}$. We name this procedure as `cfBH0`, which is summarized in Algorithm 2.

Algorithm 2 `cfBH0`: Selection by prediction with same-class calibration

Input: Calibration data $\{(X_i, Y_i)\}_{i \in \mathcal{D}^0_{\mathrm{calib}}}$, test data covariates $\{X_{n+j}\}_{j=1}^m$, FDR target $q \in (0,1)$, monotone nonconformity score $V \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

1: Compute $\widehat{V}_i = V(X_i, c_j)$ for $i \in \mathcal{D}_{\mathrm{calib}}$ and $\widehat{V}_{n+j} = V(X_{n+j}, 0)$ for $j = 1, \ldots, m$
2: Construct conformal p-values $\{p_j^0\}_{j=1}^m$ as in (11)
3: (BH procedure) Compute $k^* = \max\left\{k \colon \sum_{j=1}^m \mathbb{1}\{p_j^0 \le qk/m\} \ge k\right\}$

Output: Selection set $\mathcal{R} = \{j \colon p_j^0 \le qk^*/m\}$.

Using a slightly different argument, the following proposition shows the FDR control of `cfBH0` conditional on the labels of test samples. The proof is in Appendix B.5.

Proposition 8 *Suppose V is monotone, $\mathcal{D}_{\text{calib}}$ and test data are i.i.d., and $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{D}^0_{\text{calib}} = \{i \in \mathcal{D}_{\text{calib}} \colon Y_i = 0\}$. Given any $q \in [0, 1]$, the output of Algorithm 2 satisfies*

$$\mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{1}\{j \in \mathcal{R}, Y_{n+j} = 0\}}{1 \vee |\mathcal{R}|} \,\middle|\, \{Y_{n+j}\}_{j=1}^m\right] \le q.$$

A.2 Connection to classification methods

A recent paper (Rava et al., 2021) considers controlling the mis-classification rate (called FSR, the false selection rate), a similar notion as our FDR, for certain pre-specified subgroups in classification problems. Given any model $\widehat{s}(x)$ that predicts how likely the label of a sample with feature value x is in a specified class $\mathcal{C}$, they use an independent set of data from class $\mathcal{C}$ to calibrate a threshold $\widehat{s} \in \mathbb{R}$, and classify all the test samples with $\widehat{s}(X_{n+j}) \ge \widehat{s}$ into $\mathcal{C}$. Using a martingale argument, they show that the proportion of mis-classified units among the detected ones is controlled below a pre-specified level. Also related are Mary and Roquain (2021) and Roquain and Verzelen (2022) that consider testing whether the test data follow a "null" distribution and provide similar analysis.

Without the subgroup fairness aspect, the procedure in Rava et al. (2021) is equivalent to our `cfBH0` if we set $\widehat{V}(x, 0) = \widehat{s}(x)$. Our analysis for Proposition 8 is an alternative to the martingale argument of Rava et al. (2021); Mary and Roquain (2021); Roquain and Verzelen (2022). We note that both `cfBH` and `cfBH0` can be extended to account for subgroup fairness by applying the selection procedure separately to different subgroups.

While providing a stronger guarantee (i.e., conditional on the hypotheses), `cfBH0` is less general than `cfBH`. `cfBH` applies to problems that cannot be easily translated to a classfication problem (see, e.g., Examples 2 and 3). Moreover, in classification problems where `cfBH0` is applicable, `cfBH` can be more powerful with a suitable choice of V. We discuss this issue in the following.

Remark 9 (Power comparison in classification problems) Suppose $Y \in \{0, 1\}$ and the goal is to find $Y = 1$. We set $V(x, y) = My - \widehat{s}(x)$ as discussed in Section 2.5, where $\widehat{s}(X_i)$ is any score to predict how likely $Y_i = 1$ happens conditional on X_i (such as those used in Rava et al. (2021)), and M is a sufficiently large constant. We suppose $M > 2 \sup_x |\widehat{s}(x)|$, for which often $M = 2$ suffices. Since $M > 2 \sup_x |\widehat{s}(x)|$, for those $i \notin \mathcal{D}^0_{\text{calib}}$, i.e., $Y_i = 1$, we have $V(X_i, Y_i) = M - \widehat{s}(X_i) > -\widehat{s}(X_{n+j}) = \widehat{V}_{n+j}$. Thus our conformal p-values used in `cfBH` reduce to

$$p_j = \frac{\sum_{i \in \mathcal{D}^0_{\text{calib}}} \mathbb{1}\{V_i < \widehat{V}_{n+j}\} + U_j \cdot (1 + \sum_{i \in \mathcal{D}^0_{\text{calib}}} \mathbb{1}\{V_i = \widehat{V}_{n+j}\})}{n + 1},$$

while the p-values in `cfBH0` are

$$p^0_j = \frac{\sum_{i \in \mathcal{D}^0_{\text{calib}}} \mathbb{1}\{V_i < \widehat{V}_{n+j}\} + U_j \cdot (1 + \sum_{i \in \mathcal{D}^0_{\text{calib}}} \mathbb{1}\{V_i = \widehat{V}_{n+j}\})}{|\mathcal{D}^0_{\text{calib}}| + 1}.$$

That is, in classification problems, such a choice of nonconformity score leads to

$$p_j = \frac{|\mathcal{D}^0_{\text{calib}}| + 1}{n+1} p^0_j < p^0_j.$$

With strictly smaller p-values, the rejection set of `cfBH` is a superset of that of `cfBH0`, hence achieving strictly higher power. Also, the coefficient $\frac{|\mathcal{D}^0_{\text{calib}}|+1}{n+1}$ implies that the smaller the proportion of $Y_i = 0$ samples is among the calibration data, the greater the power gain of `cfBH` over `cfBH0`.

A.3 Connection to outlier detection

The outlier detection problem in Bates et al. (2021) can also be turned into the above classification setting. As stated in Lemma 10, Bates et al. (2021) studies a problem where given i.i.d. calibration data $\{Z_i\}_{i=1}^n$ from an unknown distribution $\mathbb{P}$, one would like to test for outliers in independent test samples $\{Z_{n+j}\}_{j=1}^m$, and the hypotheses are $H_j \colon Z_{n+j} \sim \mathbb{P}$. There, Z_{n+j} is called an *inlier* if H_j is true, and an *outlier* otherwise; the calibration data are all inliers.

To turn the outlier detection problem into our language, one could view Z as the covariate, and encode a label $Y = 0$ for inliers and $Y = 1$ for outliers. This setup is the same as the preceding subsection, where one would like to control the average proportion of inliers in those classified as $Y = 1$. Here, the inlier distribution is $\mathbb{P}_{Z \mid Y=0}$; it is then equivalent to the hypothesis testing problem in Mary and Roquain (2021) and Roquain and Verzelen (2022) as well, where data associated with null hypotheses are i.i.d. from a *null distribution*. The method in Bates et al. (2021) for marginal FDR control is the same as `cfBH0` if we set $V(x, 0) = -\widehat{s}(x)$ for the one-class-classifier $\widehat{s}(\cdot)$ trained on inliers for outlier detection in Bates et al. (2021). Similar to `cfBH0`, only inliers are used for calibration in Bates et al. (2021), and the FDR control is conditional on the test labels.

Besides the differences in the methodology and the FDR control guarantee, another distinction between `cfBH` and that of Bates et al. (2021) is the assumption on the data distributions. In Bates et al. (2021) (and also Mary and Roquain (2021); Roquain and Verzelen (2022)), it is only assumed that the inliers in the test samples are i.i.d. from $\mathbb{P}_{Z \mid Y=0}$, but the outliers can be arbitrary distributed and are not necessarily from the same distribution. In contrast, `cfBH` assumes all (Z_i, Y_i) pairs in the calibration and test data are from an i.i.d. super-population. Leveraging this additional structure, `cfBH` can deal with more general thresholds for continuous responses, see Examples 2 and 3, and achieve higher power in classification problem by including all observations in the calibration fold as we discussed in Section A.2. This super-population assumption can be reasonable in many cases, such as healthcare diagnosis, job hiring, and drug discovery.

At a high level, responses of higher values in our setting can be roughly viewed as "outliers". However, many applications such as job hiring and drug discovery may not be easily turned into an outlier detection problem. The one-class-classification in Bates et al. (2021) may also be insufficient in such applications when information from both positive and negative training samples is available. In contrast, `cfBH` is able to use any prediction model from an independent training process.

Appendix B. Technical proofs

B.1 Proof of Lemma 5

Recall that $V_{n+j} = V(X_{n+j}, Y_{n+j})$ is the unobserved score for the j-th test sample, and the oracle p-value is

$$p_j^* = \frac{\sum_{i=1}^n \mathbb{1}\{V_i < V_{n+j}\} + (1 + \sum_{i=1}^n \mathbb{1}\{V_i = V_{n+j}\}) \cdot U_j}{n+1}, \tag{12}$$

where U_j is the same as in p_j.

Proof [Proof of Lemma 5] The first property follows from the monotonicity of V. To be specific, on the event $\{Y_{n+j} \le c_j\}$, we have $V_{n+j} = V(X_{n+j}, Y_{n+j}) \le V(X_{n+j}, c_j) = \widehat{V}_{n+j}$ hence $p_j^* \le p_j$.

The second property follows from an application of the PRDS property proved in Bates et al. (2021). In the following lemma, we cite the results from Bates et al. (2021).

Lemma 10 (PRDS property in Bates et al. (2021)) *Let $Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ be the calibration data, and $Z_{n+1}, \dots, Z_{n+m}$ be independent test samples. For each $j = 1, \dots, m$, the null hypothesis is $H_j^*\colon Z_{n+j} \sim \mathbb{P}$. For any fixed nonconformity score $V\colon \mathcal{Z} \to \mathbb{R}$, we compute $V_i = V(Z_i)$ for $1 \le i \le m+n$, and construct p-values $\{p_j^*\}$ as in (12). If H_j^* is null, i.e., if $Z_{n+j} \sim \mathbb{P}$, then the vector $(p_1^*, \dots, p_m^*)$ is PRDS on p_j^*.*

We now construct another set of 'oracle p-values' in the setup of Bates et al. (2021) that coincide with our $\{p_j\}$. Let j be any fixed index. We keep p_j^* as it is, and let $Z_{n+\ell}^* = (X_{n+\ell}, c_\ell)$ for all $\ell \ne j$. Without loss of generality we define $Z_{n+j}^* = (X_{n+j}, Y_{n+j})$ and $Z_i^* = (X_i, Y_i)$ for $i = 1, \dots, n$. This can be viewed as the $(X_{n+\ell}, Y_{n+\ell})$ pair by setting $Y_{n+\ell} = c_\ell$ a.s. to be an "outlier" for all $\ell \ne j$. Then we let $\widetilde{V}_i = V(Z_i^*)$ for $1 \le i \le n+m$, and construct p-values using the same $\{U_\ell\}$ as in (12):

$$\widetilde{p}_\ell^* = \frac{\sum_{i=1}^n \mathbb{1}\{\widetilde{V}_i < \widetilde{V}_{n+\ell}\} + (1 + \sum_{i=1}^n \mathbb{1}\{\widetilde{V}_i = \widetilde{V}_{n+\ell}\}) \cdot U_\ell}{n+1}, \qquad \text{for all } \ell = 1, \dots, m.$$

Note that $\widetilde{p}_\ell^* = p_\ell$ for $\ell \ne j$ and $\widetilde{p}_j^* = p_j^*$, where p_ℓ is our new conformal p-values in (4). In the view of Lemma 10, $Z_1^*, \dots, Z_n^*$ are the i.i.d. calibration data, $Z_{n+1}^*, \dots, Z_{n+m}^*$ are independent test samples, and Z_{n+j}^* follows the same distribution as $Z_1^*, \dots, Z_n^*$. Hence by Lemma 10, we know the vector $(\widetilde{p}_1^*, \dots, \widetilde{p}_m^*)$ is PRDS on p_j^*, which is equivalent to the second property of Lemma 5. This concludes the proof of Lemma 5. ∎

B.2 Proof of Theorem 3

Proof [Proof of Theorem 3] Let $\mathcal{R}$ be the rejection set, and let $R_j = \mathbb{1}\{j \in \mathcal{R}\}$ for $j = 1, \dots, m$. In the BH procedure, $j \in \mathcal{R}$ if and only if $p_j \le q|\mathcal{R}|/m$. The FDR can thus be decomposed as

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{1}\{Y_{n+j} \le c_j\} R_j}{1 \vee \sum_{j=1}^m R_j}\right] = \sum_{j=1}^m \sum_{k=1}^m \frac{1}{k} \mathbb{E}\big[\mathbb{1}\{|\mathcal{R}| = k\} \mathbb{1}\{Y_{n+j} \le c_j\} \mathbb{1}\{p_j \le qk/m\}\big].$$

Let $\mathcal{R}_{j\to *}$ be the rejection set obtained by setting p_j to p_j^* while keeping others fixed. By Lemma 5, we have $p_j \le p_j^*$ on the event $\{Y_{n+j} \le c_j\}$. In addition, by the property of the BH procedure, if $j \in \mathcal{R}$, i.e., if $p_j \ge q|\mathcal{R}|/m$, then sending p_j to a smaller value does not change the rejection set. Thus,

$$\begin{aligned}\mathbb{1}\{|\mathcal{R}| = k\}\mathbb{1}\{Y_{n+j} \le c_j\}\mathbb{1}\{p_j \le qk/m\} &= \mathbb{1}\{|\mathcal{R}_{j\to *}| = k\}\mathbb{1}\{Y_{n+j} \le c_j\}\mathbb{1}\{p_j \le qk/m\} \\ &\le \mathbb{1}\{|\mathcal{R}_{j\to *}| = k\}\mathbb{1}\{Y_{n+j} \le c_j\}\mathbb{1}\{p_j^* \le qk/m\}.\end{aligned}$$

Therefore, the FDR is bounded as

$$\begin{aligned}\mathrm{FDR} &\le \sum_{j=1}^{m}\sum_{k=1}^{m}\frac{1}{k}\mathbb{E}\big[\mathbb{1}\{|\mathcal{R}_{j\to *}| = k\}\mathbb{1}\{Y_{n+j} \le c_j\}\mathbb{1}\{p_j^* \le qk/m\}\big] \\ &\le \sum_{j=1}^{m}\sum_{k=1}^{m}\frac{1}{k}\mathbb{E}\big[\mathbb{1}\{|\mathcal{R}_{j\to *}| = k\}\mathbb{1}\{p_j^* \le qk/m\}\big].\end{aligned}$$

By Lemma 5, the vector of p-values $(p_1, \dots, p_{j-1}, p_j^*, p_{j+1}, \dots, p_m)$ is PRDS (Bates et al., 2021) on p_j^*. Thus, following standard proofs for the BH(q) procedure under PRDS condition Benjamini and Yekutieli (2001), each term can be controlled as

$$\sum_{k=1}^{m}\frac{1}{k}\mathbb{E}\big[\mathbb{1}\{|\mathcal{R}_{j\to *}| = k\}\mathbb{1}\{p_j^* \le qk/m\}\big] \le \frac{q}{m},$$

which completes the proof of Theorem 3. ■

B.3 Proof of Theorem 6

Proof [Proof of Theorem 6| For notational simplicity, set $p_j = p_j^{\mathrm{dtm}}$ (3) in this proof only. Also define the corresponding deterministic oracle p-values

$$p_j^* = \frac{1 + \sum_{i=1}^{n}\mathbb{1}\{V_i < V_{n+j}\}}{n+1},$$

and let this notation override (3) in this proof only.

For any $j = 1, \dots, m$, define a set of slightly modified p-values

$$p_\ell^{(j)} = \frac{\sum_{i=1}^{n}\mathbb{1}\{V_i < \widehat{V}_{n+\ell}\} + \mathbb{1}\{V_{n+j} < \widehat{V}_{n+\ell}\}}{n+1}, \quad \forall\ \ell \ne j.$$

These p-values are only used in our analysis (our method cannot use them since they cannot be computed from the observations). Also define $\mathcal{R}(a_1, \dots, a_m) \subseteq \{1, \dots, m\}$ as the rejection (indices) set obtained by the BH procedure, from p-values taking on the values $a_1, \dots, a_m$.

Recall that the output of Algorithm 1 is $\mathcal{R} = \mathcal{R}(p_1, \dots, p_m)$. In the sequel, we will compare $\mathcal{R}$ to

$$\mathcal{R}(p_1^{(j)}, \dots, p_{j-1}^{(j)}, p_j^*, p_{j+1}^{(j)}, \dots, p_m^{(j)})$$

on the event $\{Y_{n+j} \le c_j, j \in \mathcal{R}\}$. First, on this event, since V is monotone, we have $V_{n+j} = V(X_{n+j}, Y_{n+j}) \le V(X_{n+j}, c_j)$, whence $p_j^* \le p_j$. For the remaining p-values, since the scores have no ties, we consider two cases:

(i) If $\widehat{V}_{n+\ell} > \widehat{V}_{n+j}$, then $\widehat{V}_{n+\ell} > V_{n+j}$ since $\widehat{V}_{n+j} > V_{n+j}$. This means

$$p_\ell^{(j)} = \frac{1 + \sum_{i=1}^n \mathbb{1}\{V_i < \widehat{V}_{n+\ell}\}}{n+1} = p_\ell.$$

(ii) If $\widehat{V}_{n+\ell} < \widehat{V}_{n+j}$, then $p_\ell \le p_j$. Since $j \in \mathcal{R}$, the BH procedure implies $\ell \in \mathcal{R}$. By definition, we have

$$p_\ell^{(j)} \le \frac{1 + \sum_{j=1}^n \mathbb{1}\{V_i < \widehat{V}_{n+\ell}\}}{n+1} \le \frac{1 + \sum_{j=1}^n \mathbb{1}\{V_i < \widehat{V}_{n+j}\}}{n+1} = p_j.$$

To summarize, suppose we are to replace p_j by p_j^* and p_ℓ by $p_\ell^{(j)}$ for all $\ell \neq j$. Then on the event $\{Y_{n+j} \le c_j, j \in \mathcal{R}\}$, such a replacement does not change any of those $p_\ell \ge p_j$; also, all those $p_\ell \le p_j$ including p_j itself (they are rejected in $\mathcal{R}$) are still no greater than p_j after the replacement. Thus, by the step-up nature of the BH procedure, such a replacement does not change the rejection set, meaning that

$$\begin{aligned}\mathcal{R} &= \mathcal{R}(p_1^{(j)}, \ldots, p_{j-1}^{(j)}, p_j, p_{j+1}^{(j)}, \ldots, p_m^{(j)}) \\ &= \mathcal{R}(p_1^{(j)}, \ldots, p_{j-1}^{(j)}, p_j^*, p_{j+1}^{(j)}, \ldots, p_m^{(j)}) =: \mathcal{R}_j^*\end{aligned}$$

on the event $\{Y_{n+j} \le c_j, j \in \mathcal{R}\}$. As in the proof of Theorem 3, a leave-one-out analysis of the FDR then implies

$$\begin{aligned}\mathrm{FDR} &= \mathbb{E}\Big[\frac{\sum_{j=1}^m \mathbb{1}\{Y_{n+j} \le c_j\} R_j}{1 \vee \sum_{j=1}^m R_j}\Big] \\ &= \sum_{j=1}^m \sum_{k=1}^m \frac{1}{k} \mathbb{E}\big[\mathbb{1}\{|\mathcal{R}| = k\} \mathbb{1}\{Y_{n+j} \le c_j\} \mathbb{1}\{p_j \le qk/m, j \in \mathcal{R}\}\big] \\ &\le \sum_{j=1}^m \sum_{k=1}^m \frac{1}{k} \mathbb{E}\big[\mathbb{1}\{|\mathcal{R}_j^*| = k\} \mathbb{1}\{Y_{n+j} \le c_j\} \mathbb{1}\{p_j^* \le qk/m\}\big] \\ &\le \sum_{j=1}^m \sum_{k=1}^m \frac{1}{k} \mathbb{E}\big[\mathbb{1}\{|\mathcal{R}_j^*| = k\} \mathbb{1}\{p_j^* \le qk/m\}\big] \\ &= \sum_{j=1}^m \sum_{k=1}^m \frac{1}{k} \mathbb{E}\big[\mathbb{1}\{|\mathcal{R}_j^*| = k\} \mathbb{1}\{p_j^* \in \mathcal{R}_j^*\}\big];\end{aligned}$$

the second and the last lines use the property of the BH procedure, whereas the third uses the facts stated just above. By the step-up nature of the BH procedure, we know that on the event $\{p_j^* \in \mathcal{R}_j^*\}$, sending p_j^* to zero does not change the rejection set, i.e., we have

$$\mathcal{R}_j^* = \mathcal{R}(p_1^{(j)}, \ldots, p_{j-1}^{(j)}, 0, p_{j+1}^{(j)}, \ldots, p_m^{(j)}) =: \mathcal{R}_{j\to 0}^*.$$

Thus

$$\mathrm{FDR} \le \sum_{j=1}^{m}\sum_{k=1}^{m}\frac{1}{k}\mathbb{E}\big[\mathbb{1}\{|\mathcal{R}_j^*|=k\}\mathbb{1}\{p_j^*\in\mathcal{R}_{j\to 0}^*\}\big] = \sum_{j=1}^{m}\mathbb{E}\bigg[\frac{\mathbb{1}\{p_j^*\le q|\mathcal{R}_{j\to 0}^*|/m\}}{1\vee|\mathcal{R}_{j\to 0}^*|}\bigg].$$

Note that by definition, $\{p_\ell^{(j)}\}_{\ell\neq j}$ is invariant after permuting $\{V_i\}_{i=1}^n\cup\{V_{n+j}\}$. Since $\{V_i\}_{i=1}^n\cup\{V_{n+j}\}$ are exchangeable conditional on $\{\widehat{V}_{n+\ell}\}_{\ell\neq j}$, we know that the distribution of $\{p_\ell^{(j)}\}_{\ell\neq j}$ is independent of $\{V_i\}_{i=1}^n\cup\{V_{n+j}\}$ conditional on the unordered set $[V_1,\dots,V_n,V_{n+j}]$. Also note that $\mathcal{R}_{j\to 0}^*$ only depends on $\{p_j^{(\ell)}\}_{\ell\neq j}$, and p_j^* only depends on $\{V_i\}_{i=1}^n\cup\{V_{n+j}\}$. This implies that $\mathcal{R}_{j\to 0}^*$ is independent of p_j^* conditional on the unordered set $[V_1,\dots,V_n,V_{n+j}]$. The tower property yields

$$\mathbb{E}\bigg[\frac{\mathbb{1}\{p_j^*\le q|\mathcal{R}_{j\to 0}^*|/m\}}{1\vee|\mathcal{R}_{j\to 0}^*|}\bigg] = \mathbb{E}\bigg[\mathbb{E}\bigg[\frac{\mathbb{1}\{p_j^*\le q|\mathcal{R}_{j\to 0}^*|/m\}}{1\vee|\mathcal{R}_{j\to 0}^*|}\,\bigg|\,[V_1,\dots,V_n,V_{n+j}]\bigg]\bigg].$$

In addition, since the variables $\{V_i\}_{i=1}^n\cup\{V_{n+j}\}$ are conditionally exchangeable, they are also marginally exchangeable. Thus, for any random variable $t\in\mathbb{R}$ that is measurable with respect to the unordered set $[V_1,\dots,V_n,V_{n+j}]$, we have

$$\mathbb{P}\big(p_j^*\le t\,\big|\,[V_1,\dots,V_n,V_{n+j}]\big)\le t.$$

This gives

$$\mathbb{E}\bigg[\frac{\mathbb{1}\{p_j^*\le q|\mathcal{R}_{j\to 0}^*|/m\}}{1\vee|\mathcal{R}_{j\to 0}^*|}\,\bigg|\,[V_1,\dots,V_n,V_{n+j}]\bigg]\le\frac{q}{m}.$$

Summing over $j\in\{1,\dots,m\}$ concludes the proof. ■

B.4 Proof of Proposition 7

Proof [Proof of Proposition 7] We utilize an equivalent representation of the BH(q) procedure, communicated in Storey et al. (2004): the rejection set is $\mathcal{R}=\{j\colon p_j\le\widehat{\tau}\}$, where

$$\widehat{\tau}=\sup\bigg\{t\in[0,1]\colon\frac{mt}{\sum_{j=1}^m\mathbb{1}\{p_j\le t\}}\le q\bigg\}. \tag{13}$$

To clarify the dependence on the calibration data, we denote the p-values as $p_j=\widehat{F}_n(V_j^0,U_j)$, where for simplicity we denote $V_j^0=\widehat{V}_{n+j}=V(X_{n+j},0)$, and define

$$\widehat{F}_n(v,u)=\frac{\sum_{i=1}^n\mathbb{1}\{V_i<v\}+(1+\sum_{i=1}^n\mathbb{1}\{V_i=v\})\cdot u}{n+1}$$

for any $(v,u)\in\mathbb{R}\times[0,1]$. We know that $\{(V_j^0,U_j)\colon 1\le j\le m\}$ are i.i.d., and independent of $\mathcal{D}_{\mathrm{calib}}$. We first define

$$F(v,u)=\mathbb{P}\big(V(X,Y)<v\big)+\mathbb{P}\big(V(X,Y)=v\big)\cdot u.$$

Then by the uniform law of large numbers, we have

$$\sup_{v\in\mathbb{R},u\in[0,1]} \left|\widehat{F}_n(v,u) - F(v,u)\right| \overset{\text{a.s.}}{\to} 0, \quad \text{as } n\to\infty. \tag{14}$$

We then repeatly employ the (uniform) strong law of large numbers to show the asymptotic behavior of the testing procedure. Based on (14), we show the uniform convergence of the criterion in (13).

Lemma 11 *With the same setup as in the proof of Proposition 7, suppose* $\sup_{x\in\mathcal{X}} w(x) \le M$ *for some constant* $M > 0$. *Then*

$$\sup_{t\in[0,1]} \left| \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\widehat{F}_n(V_j^0, U_j) \le t\} - \mathbb{P}\big(F(V_j^0, U_j) \le t\big) \right| \overset{a.s.}{\to} 0,$$

as $m, n \to \infty$, *where* $\mathbb{P}$ *is taken with respect to* $V_j^0 = V(X_{n+j}, 0)$ *and an independent* $U_j \sim$ *Unif*$[0,1]$.

Proof [Proof of Lemma 11] Let $0 = t_0 < t_1 < \cdots < t_K = 1$ be a partition of $[0,1]$. Then for each $t \in [0,1]$, there exists some k such that $t_k \le t < t_{k+1}$, whence

$$\begin{aligned}
&\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\widehat{F}_n(V_j^0, U_j) \le t_k\} - \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{F(V_j^0, U_j) \le t_{k+1}\} \qquad (15)\\
&\le \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\widehat{F}_n(V_j^0, U_j) \le t\} - \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{F(V_j^0, U_j) \le t\}\\
&\le \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\widehat{F}_n(V_j^0, U_j) \le t_{k+1}\} - \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{F(V_j^0, U_j) \le t_k\}.\\
&\le \left|\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\widehat{F}_n(V_j^0, U_j) \le t_{k+1}\} - \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{F(V_j^0, U_j) \le t_k\}\right|\\
&\le \left|\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\widehat{F}_n(V_j^0, U_j) \le t_{k+1}\} - \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{F(V_j^0, U_j) \le t_{k+1}\}\right|\\
&\quad + \left|\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{F(V_j^0, U_j) \le t_{k+1}\} - \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{F(V_j^0, U_j) \le t_k\}\right|. \qquad (16)
\end{aligned}$$

For any fixed $\delta > 0$, we have

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\widehat{F}_n(V_j^0, U_j) \le t_{k+1}\} - \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{F(V_j^0, U_j) \le t_{k+1}\}\right|\\
&\le \frac{1}{m}\sum_{j=1}^{m} \Big(\mathbb{1}\{\widehat{F}_n(V_j^0, U_j) \le t_{k+1}, F(V_j^0, U_j) > t_{k+1}\}\\
&\qquad + \mathbb{1}\{\widehat{F}_n(V_j^0, U_j) > t_{k+1}, F(V_j^0, U_j) \le t_{k+1}\}\Big)
\end{aligned}$$

$$\le \mathbb{1}\big\{\sup_j \big|\widehat{F}_n(V_j^0,U_j) - F(V_j^0,U_j)\big| \ge \delta\big\}$$
$$+ \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{\widehat{F}_n(V_j^0,U_j) \le t_{k+1}, t_{k+1}+\delta \ge F(V_j^0,U_j) > t_{k+1}\}$$
$$+ \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{\widehat{F}_n(V_j^0,U_j) > t_{k+1}, t_{k+1}-\delta < F(V_j^0,U_j) \le t_{k+1}\}$$
$$\le \mathbb{1}\big\{\sup_j \big|\widehat{F}_n(V_j^0,U_j) - F(V_j^0,U_j)\big| \ge \delta\big\} + \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{t_{k+1}-\delta \le F(V_j^0,U_j) \le t_{k+1}+\delta\}.$$

Then (14) implies $\limsup_{n\to\infty} \mathbb{1}\big\{\sup_j \big|\widehat{F}_n(V_j^0,U_j) - F(V_j^0,U_j)\big| \ge \delta\big\} = 0$ almost surely. Combinining with the decomposition (16), we have

$$\limsup_{n\to\infty} \left|\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{\widehat{F}_n(V_j^0,U_j) \le t_{k+1}\} - \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{F(V_j^0,U_j) \le t_k\}\right|$$
$$\le \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{t_{k+1}-\delta \le F(V_j^0,U_j) \le t_{k+1}+\delta\} + \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{t_k < F(V_j^0,U_j) \le t_{k+1}\}$$

almost surely. Again invoking the (uniform) law of large numbers for i.i.d. random variables $F(V_j^0,U_j)$,

$$\limsup_{m\to\infty}\sup_k \Big|\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{t_{k+1}-\delta \le F(V_j^0,U_j) \le t_{k+1}+\delta\} - \mathbb{P}\big(t_{k+1}-\delta \le F(V_j^0,U_j) \le t_{k+1}+\delta\big)$$
$$+ \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{t_k < F(V_j^0,U_j) \le t_{k+1}\} - \mathbb{P}\big(t_k < F(V_j^0,U_j) \le t_{k+1}\big)\Big| = 0$$

with probability one. We thus have

$$\limsup_{m,n\to\infty}\sup_k \left|\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{\widehat{F}_n(V_j^0,U_j) \le t_{k+1}\} - \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{F(V_j^0,U_j) \le t_k\}\right|$$
$$\le \sup_k \Big|\mathbb{P}\big(t_{k+1}-\delta \le F(V_j^0,U_j) \le t_{k+1}+\delta\big) + \mathbb{P}\big(t_k < F(V_j^0,U_j) \le t_{k+1}\big)\Big| \qquad (17)$$

for any partition $\{t_k\}$ and $\delta > 0$. On the other hand, we note that since $U_j \sim$ Unif[0,1] is independent of the observations, $F(V_j^0,U_j)$ are i.i.d. with continuous distributions. Letting $\delta \to 0$ and $\{t_k\}$ be fine enough sends the supremum in (17) to zero. With similar arguments, we can show a lower bound for (15) that leads to

$$\limsup_{m,n\to\infty}\sup_k \left|\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{\widehat{F}_n(V_j^0,U_j) \le t_k\} - \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{F(V_j^0,U_j) \le t_{k+1}\}\right| = 0$$

almost surely. Combining the above two results, we then have

$$\sup_{t\in[0,1]} \left|\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{\widehat{F}_n(V_j^0,U_j) \le t\} - \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{F(V_j^0,U_j) \le t\}\right| \overset{\text{a.s.}}{\to} 0. \qquad (18)$$

Invoking the uniform strong law of large numbers, we have

$$\sup_{t\in[0,1]}\left|\frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\{F(V_j^0,U_j)\le t\}-\mathbb{P}\big(F(V_j^0,U_j)\le t\big)\right|\overset{\text{a.s.}}{\to}0,$$

hence by the triangular inequality we complete the proof of Lemma 11. ∎

With similar arguments as in the proof of Lemma 11, we can also show that as $n\to\infty$,

$$\sup_{t\in[0,1]}\left|\frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\{\widehat{F}_n(V_j^0,U_j)\le t, j\in\mathcal{H}_0\}-\mathbb{P}\big(F(V_j^0,U_j)\le t, Y(1)\le Y(0)\big)\right|\overset{\text{a.s.}}{\to}0. \quad (19)$$

Suppose there exists some $t'\in(0,1]$ such that $\frac{\mathbb{P}(F(V_j^0,U_j)\le t')}{t'}>\frac{1}{q}$. We then define

$$t^*=\sup\left\{t\in[0,1]\colon\frac{\mathbb{P}\big(F(V_j^0,U_j)\le t\big)}{t}\ge\frac{1}{q}\right\}=\sup\big\{t\in[0,1]\colon G_\infty(t)\le q\big\},$$

where $G_\infty(t)=t/\mathbb{P}(F(V_j^0,U_j)\le t)$. It is well-defined and $t^*\ge t'$. Fix any $\delta\in(0,t')$. By Lemma 11,

$$\sup_{t\in[\delta,1]}\left|\frac{\sum_{j=1}^m\mathbb{1}\{F(V_j^0,U_j)\le t\}}{mt}-\frac{\mathbb{P}(F(V_j^0,U_j)\le t)}{t}\right|\overset{\text{a.s.}}{\to}0. \quad (20)$$

In particular, $\frac{\sum_{j=1}^m\mathbb{1}\{F(V_j^0,U_j)\le t'\}}{mt'}\overset{\text{a.s.}}{\to}\frac{\mathbb{P}(F(V_j^0,U_j)\le t')}{t'}>\frac{1}{q}$, hence $\widehat{\tau}\ge t'\ge\delta$ eventually. Furthermore, since $F(V_j^0,U_j)$ admits a continuous distribution, the function $t\mapsto\mathbb{P}\big(F(V_j^0,U_j)\le t\big)$ is continuous in $t\in[0,1]$. Under the assumption that for any $\epsilon>0$, there exists some $|t-t^*|\le\epsilon$ such that $\frac{\mathbb{P}(F(V_j^0,U_j)\le t)}{t}>1/q$, the uniform convergence in (20) implies $\widehat{\tau}\overset{\text{a.s.}}{\to}t^*$.

Let $\delta\in(0,t^*)$ be any fixed value such that $\mathbb{P}(F(V_j^0,U_j)\le\delta)>2\epsilon$ for some constant $\epsilon>0$. Then we know $\inf_{t\in[\delta,1]}\mathbb{P}\{F(V(X,Y(0)),U)\le t\}\ge\epsilon$, and by Lemma 11,

$$\liminf_{m\to\infty}\inf_{t\in[\delta,1]}\left\{\frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\{\widehat{F}_n(V_j^0,U_j)\le t\}\right\}\ge\epsilon$$

almost surely. Combining this lower boundedness property with the uniform convergence results in Lemma 11, equation (18), and (19), we know that

$$\sup_{t\in[\delta,1]}\left|\frac{\sum_{j=1}^m\mathbb{1}\{\widehat{F}_n(V_j^0,U_j)\le t, j\in\mathcal{H}_0\}}{1\vee\sum_{j=1}^m\mathbb{1}\{\widehat{F}_n(V_j^0,U_j)\le t\}}-\frac{\mathbb{P}\{F(V(X,Y(0)),U)\le t, Y(1)\le Y(0)\}}{\mathbb{P}\{F(V(X,Y(0)),U)\le t\}}\right|\overset{\text{a.s.}}{\to}0.$$

Since $\widehat{\tau}\overset{\text{a.s.}}{\to}t^*$ and the distribution functions are continuous, the asymptotic FDR is

$$\lim_{m,n\to\infty}\text{FDR}=\lim_{m,n\to\infty}\mathbb{E}\left[\frac{\sum_{j=1}^m\mathbb{1}\{\widehat{F}_n(V_j^0,U_j)\le\widehat{\tau},Y_{n+j}\le0\}}{1\vee\sum_{j=1}^m\mathbb{1}\{\widehat{F}_n(V_j^0,U_j)\le\widehat{\tau}\}}\right]$$

$$= \mathbb{E}\left[\lim_{m,n\to\infty} \frac{\sum_{j=1}^m \mathbb{1}\{\widehat{F}_n(V_j^0, U_j) \le \widehat{\tau}, Y_{n+j} \le 0\}}{1 \vee \sum_{j=1}^m \mathbb{1}\{\widehat{F}_n(V_j^0, U_j) \le \widehat{\tau}\}}\right]$$
$$= \frac{\mathbb{P}\{F(V(X,0),U) \le t^*, Y \le 0\}}{\mathbb{P}\{F(V(X,0),U) \le t^*\}},$$

where the second line follows from the Dominated Convergence Theorem. With similar arguments, we can show that the asymptotic power of the procedure is

$$\lim_{m,n\to\infty} \text{Power} = \lim_{m,n\to\infty} \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{1}\{\widehat{F}_n(V_j^0, U_j) \le \widehat{\tau}, Y_{n+j} > 0\}}{1 \vee \sum_{j=1}^m \mathbb{1}\{Y_{n+j} > 0\}}\right]$$
$$= \frac{\mathbb{P}\{F(V(X,0),U) \le t^*, Y > 0\}}{\mathbb{P}\{Y > 0\}}.$$

Therefore, we complete the proof of Proposition 7. ■

B.5 Proof of Proposition 8

Proof [Proof of Proposition 8] We show that the FDR conditional on all signs of the data is controlled, i.e.,

$$\mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{1}\{Y_{n+j} \le 0, j \in \mathcal{R}\}}{1 \vee \sum_{j=1}^m R_j} \,\middle|\, \mathbb{1}\{Y_i \le 0\} \colon i \in \mathcal{D}_{\text{calib}}^- \cup \mathcal{D}_{\text{test}}\right] \le q.$$

Following the same arguments as in the proof of Theorem 3, it suffices to show

$$\sum_{k=1}^m \frac{1}{k}\mathbb{E}\big[\mathbb{1}\{|\mathcal{R}_{j\to *}| = k\}\mathbb{1}\{p_j^* \le qk/m\}\mathbb{1}\{Y_{n+j} \le 0\} \,\big|\, \mathbb{1}\{Y_i \le 0\} \colon i \in \mathcal{D}_{\text{calib}}^- \cup \mathcal{D}_{\text{test}}\big] \le \frac{q}{m},$$

where $\mathcal{R}_{j\to *}$ is the rejection set obtained by changing p_j to its oracle counterpart

$$p_j^* = \frac{\sum_{i\in\mathcal{D}_{\text{calib}}^-} \mathbb{1}\{V_i < V_{n+j}\} + (1 + \sum_{i\in\mathcal{D}_{\text{calib}}^-} \mathbb{1}\{V_i = V_{n+j}\}) \cdot U_j}{n+1}.$$

Recall that we define $Z_i = (X_i, Y_i)$ for $1 \le i \le n+m$ and $\widetilde{Z}_{n+j} = (X_{n+j}, c_j) = (X_{n+j}, 0)$ for $j = 1, \ldots, m$. Conditional on all signs, for any fixed j with $Y_{n+j} \le 0$, $\{Z_i\}_{i=1}^n \cup \{\widetilde{Z}_{n+\ell}\}_{\ell\neq j} \cup \{Z_{n+j}\}$ are mutually independent, and $\{Z_i \colon i = 1, \ldots, n, n+j\}$ are i.i.d. The desired result thus follows from the PRDS of conformal p-values, or equivalently the conditions in Theorem 3. ■

Appendix C. Additional details and results

C.1 Detailed setup for the data illustration in Section 1.1

We use the same dataset, same data splitting scheme, and the same machine learning model $\widehat{\mu}(\cdot)$ as in Section 4.2.1. That is, we repeat the procedure independently for $N = 100$ times;

in each time, we randomly split the whole dataset into training $\mathcal{D}_{\text{train}}$, calibration $\mathcal{D}_{\text{calib}}$, and test data $\mathcal{D}_{\text{test}}$ with ratio 6 : 2 : 2 in size. The model $\widehat{\mu}(\cdot)$ is trained on $\mathcal{D}_{\text{train}}$; $\mathcal{D}_{\text{calib}}$ is used to construct conformal prediction sets; the prediction sets on $\mathcal{D}_{\text{test}}$ are used to evaluate FDP, miscoverage rate, and proportion of $\widehat{C}_{1-\alpha}(X) = \{1\}$. These quantities are then averaged over $N = 100$ runs to estimate the FDR, marginal miscoverage, and average proportion of $\widehat{C}_{1-\alpha}(X) = \{1\}$. In constructing conformal prediction sets, we set $V(x, y) = y - \widehat{\mu}(x)$ for all three methods. The split conformal prediction set with $(1 - \alpha)$ marginal coverage is

$$\widehat{C}_{1-\alpha}(x) = \{y \in \{0, 1\} : V(x, y) \geq \widehat{\eta}\},$$

where $\widehat{\eta}$ is the $1 - (1 - \alpha)(1 + 1/n_{\text{calib}})$-th empirical quantile of $\{V(X_i, Y_i) : i \in \mathcal{D}_{\text{calib}}\}$. The naive approach takes $\mathcal{R} = \{j \in \mathcal{D}_{\text{test}} : \widehat{C}_{1-\alpha}(X_j) = \{1\}\}$. Our approach is Algorithm 1. Bonferroni correction sets $\mathcal{R} = \{j \in \mathcal{D}_{\text{test}} : p_j \leq q/m\}$, where $\{p_j\}$ are the conformal p-values constructed in our approach. When evaluating our approach and Bonferroni correction (i.e., producing the plot on the right panel), we randomly take a subset of $m = 1000$ test samples, such that conformal prediction at resolution q/m is still feasible.

C.2 Data generating processes

To better illustrate the data distributions, Figure 9 shows the scatterplots of data from our eight settings.

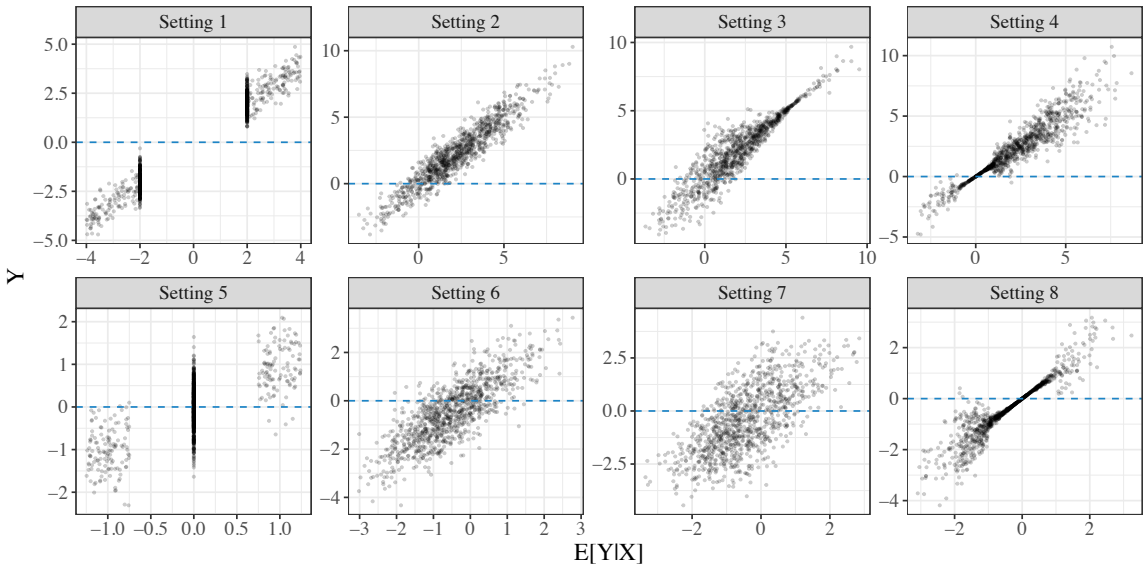

Figure 9: Scatter plots of i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^{1000}$ from the data generating processes in our simulations. The x-axis is the conditional mean function $\mu(X_i)$; the y-axis is the actual outcome Y_i.

In particular, we vary the following two aspects such that the hardness of correctly identifying those $Y > 0$ is not the same.

1. Continuity of the range of μ: settings 1 and 5 have disjoint ranges of $\mu(\mathcal{X})$ for negative, zero, and positive mean outcomes. Among the remainings with continuous ranges of

μ, settings 2, 3, 4 have more samples with positive $\mu(x)$, and settings 6, 7, 8 have more samples with negative $\mu(x)$. While directional selection seems to be easy in setting 1 and 5 (where a perfect selection would be to choose those $\mu(X_i) > 0$), with large noise level it may still be diffcult to learn the true model. The hardness of prediction in continuous-range settings also depends on the absolute scale of the mean functions.

2. Noise heterogeneity: in settings 1, 2, and 5, 6, we set $\epsilon_i \sim N(0, \sigma^2)$ with homogenous variance. Other continuous-range settings all have heterogenous noise $\epsilon_i \,|\, X_i \sim N(0, \sigma(X_i)^2)$ for some function $\sigma(\cdot)$. To be specific, in settings 3, 4, 8, the variance of noise increases with $|\mu(x)|$, showing more difficulty in the direction of interest. In setting 7, the noise is smaller for larger $|\mu(x)|$. In general, the task would be easier if the mean value is large and the noise variance is small for the direction of interest.

The specific configurations of $\mu(x)$ and $\sigma(x)$ in all of our simulation settings are detailed in Table 2 to reproduce the results in Section 3.

Setting	$\mu(\cdot)$	$\sigma(\cdot)$ for $\epsilon_i \,\|\, X_i = x \sim N(0, \sigma(x)^2)$
1	$4x_1\mathbb{1}\{x_2 > 0\} \cdot \max\{0.5, x_3\}$ $+4x_1\mathbb{1}\{x_2 \le 0\} \cdot \min\{x_3, -0.5\}$	σ^2
2	$5(x_1x_2 + e^{x_4-1})$	$2.25\sigma^2$
3	$5(x_1x_2 + e^{x_4-1})$	$\sigma \cdot (5.5 - \lvert\mu(x)\rvert)/2$
4	$5(x_1x_2 + e^{x_4-1})$	$\sigma \cdot 0.25\mu(x)^2\mathbb{1}\{\lvert\mu(x)\rvert < 2\}$ $+\sigma \cdot 0.5\lvert\mu(x)\rvert\mathbb{1}\{\lvert\mu(x)\rvert \ge 1\}$
5	$x_1\mathbb{1}\{x_2 > 0, x_4 > 0.5\} \cdot (0.25 + x_4)$ $+x_1\mathbb{1}\{x_2 \le 0, x_4 < -0.5\} \cdot (x_4 - 0.25)$	σ
6	$2(x_1x_2 + x_3^2 + e^{x_4-1} - 1)$	1.5σ
7	$2(x_1x_2 + x_3^2 + e^{x_4-1} - 1)$	$\sigma \cdot (5.5 - \lvert\mu(x)\rvert)/2$
8	$2(x_1x_2 + x_3^2 + e^{x_4-1} - 1)$	$\sigma \cdot (0.25\mu(x)^2\mathbb{1}\{\lvert\mu(x)\rvert < 2\}$ $+0.5\lvert\mu(x)\rvert\mathbb{1}\{\lvert\mu(x)\rvert \ge 1\})$

Table 2: Details of the eight data generating processes in the simulations of Section 3. The parameter σ corresponds to the noise strength on the x-axis in Figures 2 and 3.

C.3 Additional plots

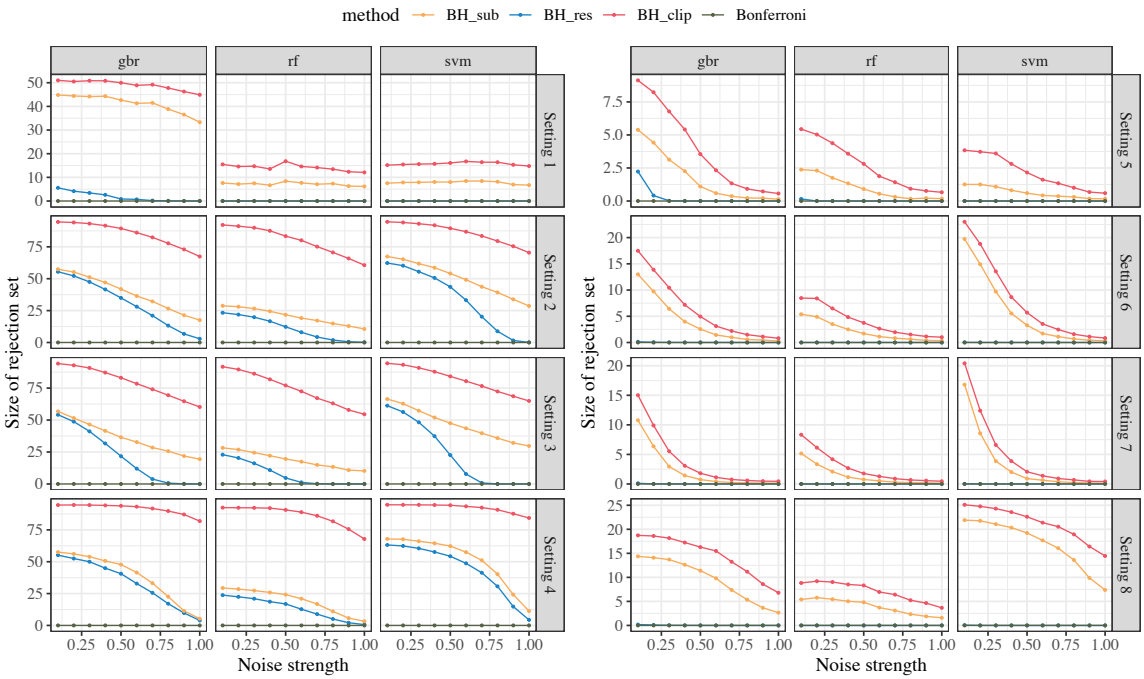

Figure 10: Average size of rejection set for four procedures at FDR target $q = 0.1$ for various data generating processes. Details of the plots are otherwise the same as Figure 2.