A Burden Shared is a Burden Halved: A Fairness-Adjusted Approach to Classification

Bradley Rava[1], Wenguang Sun[1,2], Gareth M. James[1] and Xin Tong[1]

Abstract

We study fairness in classification, where one wishes to make automated decisions for people from different protected groups. When individuals are classified, the decision errors can be unfairly concentrated in certain protected groups. We develop a fairness-adjusted selective inference (FASI) framework and data-driven algorithms that achieve statistical parity in the sense that the false selection rate (FSR) is controlled and equalized among protected groups. The FASI algorithm operates by converting the outputs from black-box classifiers to R-values, which are intuitively appealing and easy to compute. Selection rules based on R-values are provably valid for FSR control, and avoid disparate impacts on protected groups. The effectiveness of FASI is demonstrated through both simulated and real data.

Keywords: Calibration by group; Fairness in classification; False selection rate; Selective Inference; Statistical parity

1 Introduction

In a broad range of applications, artificial intelligence (AI) systems are rapidly replacing human decision-making. Many of these scenarios are sensitive in nature, where the AI's decision, correct or not, can directly impact one's social or economic status. A few examples include a bank determining credit card limits, stores using facial recognition systems to detect shoplifters, and hospitals attempting to identify which of their patients has a specific disorder. Unfortunately, despite their supposedly unbiased approach to decision-making, there has been increasing evidence that AI algorithms often fail to treat equally people of different genders, races, religions, or other protected attributes. Whether this is due to the historical bias in one's training data, or otherwise, it is important, for both legal and policy reasons, that we make ethical use of data and ensure that decisions are made fairly for everyone regardless of their protected attributes.

Although a large literature has been devoted to developing supervised learning algorithms for improving the prediction accuracy, making reliable and fair decisions in the classification setting remains a critical and challenging problem. AI algorithms typically are forced to make classifications on all new observations without a careful assessment of associated uncertainty or ambiguity. The ambiguity can be inherent to the classification problem – even an oracle knowing exactly the probability distribution of the classes cannot control the classification errors at a desired level when the signal to noise ratio is too low. This article develops a "fairness-adjusted selective inference" (FASI) framework to address the critical issues of uncertainty assessment, error rate control and statistical parity in classification. We provide an *indecision* option for observations who cannot be selected into any classes with confidence. These observations can then be separately evaluated.

[1]Department of Data Sciences and Operations, University of Southern California.

[2]*Address for correspondence:* Wenguang Sun, Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA 90089. E-mail: wenguans@marshall.usc.edu

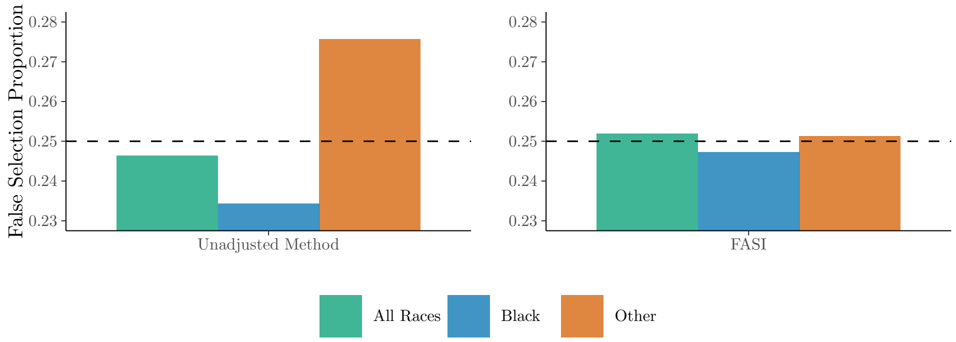

Figure 1: The selection of recidivists from a pool of criminal defendants (Broward County, Florida). The target FSR is 25%. Left: the unadjusted approach. Right: the proposed FASI approach.

This practice often aligns with the policy objectives in many real world scenarios. For example, incorrectly classifying a low-risk individual as a recidivist or rejecting a well-deserving candidate for the loan request is much more expensive than turning the case over for a more careful review. A mis-classification is an error, the probability of which must be controlled to be small as its consequence can be severe. By contrast, the cost of an indecision is usually much less. For example, the ambiguity can be mitigated by collecting additional contextual knowledge of the convicted individual or requesting more information from the loan applicant. Under the selective inference (Benjamini, 2010) framework, we only make definitive decisions on a *selected subset* of all individuals; the less consequential indecision option is considered as a wasted opportunity rather than an error. A natural error rate notion under this framework is the false selection rate (FSR), which is defined as the expected fraction of erroneous classifications among the selected subset of individuals. The goal is to develop decision rules to ensure that the FSR is effectively controlled and equalized across protected groups, while minimizing the total wasted opportunities.

However, a classification rule that controls the overall FSR may have disparate impacts on different protected groups. We illustrate the point using the COMPAS data (Angwin et al., 2016; Dieterich et al., 2016). The COMPAS algorithm has been widely used in the US to help inform courts about a defendant's recidivism likelihood, i.e., the likelihood of a convicted criminal recommitting a crime, so any prediction errors could have significant implications. The left hand plot of Figure 1 shows the *False Selection Proportions* (FSP), i.e. the fraction of individuals who did not recommit a crime among those who were classified as recidivists. The classification rule was generated via a Generalized Additive Model (GAM) to achieve the target FSR of 25%. We can see that the green bar, which provides the overall FSP for all races, is close to the target value. Moreover, the rule appears to be "fair" for all individuals, regardless of their protected attributes, in the sense that the *same* threshold has been applied to the base scores (i.e. estimated class probabilities) produced by the *same* GAM fit. However, the blue and orange bars show that the FSPs for different racial groups differ significantly from 25%, which is clearly not a desirable situation.

This article introduces a new notion of fairness that requires parity in FSR control across various protected groups. This aligns with the social and policy goals in various decision-making scenarios such as selecting recidivists or determining risky loan applicants, where the burden of erroneous classifications should be shared equally among different genders and/or races. However, the development of effective and fair FSR rules is challenging. First, controlling the error rate associated with a classifier, such as one built around the GAM procedure, critically depends on the accuracy of the scores. However, the assessment of the accuracy/uncertainty of these scores largely remains unknown. Second, we wish to provide practitioners with theoretical guarantees on the parity and validity for FSR control, regardless of the algorithm being used, including complex black-box classifiers. If we build an algorithm around black-box models then it often becomes intractable to compute thresholds for controlling the FSR over multiple protected groups.

To address these issues, we develop a data-driven FASI algorithm, which is specifically designed to control both the overall FSR, and the (protected) group-wise FSRs, below a user specified level of risk α. The right hand plot of Figure 1 illustrates the FSPs of FASI on the recidivism data. Now, not only is the overall FSP controlled at 25%, but also so are the individual race FSPs. FASI works by converting the base score from a black-box classifier to an R-value, which is intuitive, easy to compute, and comparable across different protected groups. We then show that selecting all observations with R-value no greater than α will result in an FSR of approximately α. Hence, we can directly use this R-value to assign new observations a class label or, for those with high R-values, to assign them to the *indecision* class.

This paper makes several contributions. First, we introduce a new notion of fairness involving controlling, not only the overall FSR, but also the FSR for designated sub-groups. Ours is not the only approach to fairness and we do not claim that it is universally superior relative to alternative approaches (Dwork et al., 2012; Hardt et al., 2016; Romano et al., 2020b). However, for the settings we consider, controlling all of the sub-group FSRs appears to be a reasonable fairness approach. Second, we develop a data-driven FASI algorithm based on the R-value. The algorithm, which can be implemented with a user-specified classifier, is intuitively appealing and easy to intepret. Third, we provide theoretical results both justifying the use of R-values and the effectiveness of the FASI algorithm for FSR control. The finite-sample theory is established with minimal assumptions: we allow the base scores to be generated from black-box classifiers, and make no assumptions on the accuracy of these scores. Finally, the strong empirical performance of FASI is demonstrated via both simulated and real data.

The rest of the paper is structured as follows. In Section 2 we define the FSR and describe the problem formulation. Section 3 introduces the R-value and FASI algorithm. We establish theoretical properties of FASI in Section 4. The numerical results for simulated and real data are presented in Sections 5 and 6, respectively. Section 7 concludes the article with a discussion of related works and possible extensions.

2 Problem Formulation

Suppose we observe a data set $\mathcal{D} = \{(X_i, A_i, Y_i) : i = 1, \ldots, n\}$, where $X_i \in R^p$ is a p-dimensional vector of features, $A_i \in \mathcal{A}$ is an additional feature representing the protected or sensitive attribute, and Y_i is a class label taking values in $\mathcal{C} = \{1, \ldots, C\}$. The goal is to predict the classes for multiple individuals with instances (X_{n+j}, A_{n+j}), $j = 1, \ldots, m$.

2.1 A selective inference framework for binary classification

To focus on key ideas, we mainly consider the binary classification problem in this article. The extension to the general multi-class setting is discussed in Section 7.

One important application scenario is the prediction of mortgage default, where $Y = 2$ indicates default and $Y = 1$ otherwise. The current practice is to use risk assessment software to produce a *confidence score*, based on which an individual is classified into "high", "medium" or "low" risk classes. Let $S(x, a)$ denote such a score that maps an instance (x, a) to a real value, with a higher value indicating a higher risk of default. Suppose we observe a new instance $(X^*, A^*) = (x, a)$. Consider a class of decision rules of the form: $\hat{Y} = \mathbb{I}\{S(x, a) < t_l\} + 2\mathbb{I}\{S(x, a) > t_u\}$, where t_l and t_u are thresholds chosen by the investigator to characterize the lower and upper limits of potential risks and $\mathbb{I}(\cdot)$ is the indicator function. $\hat{Y}$ takes three possible values in the action space $\Lambda = \{1, 2, 0\}$, respectively indicating that an individual has low, high and medium risks of default. The value "0", which is referred to as an *indecision* or *reject option* in classification (Herbei and Wegkamp, 2006; Sun and Wei, 2011; Lei, 2014), is used to express "doubt", reflecting that there is not sufficient confidence to make a definitive decision. For example, an individual with $\hat{Y} = 1$ will be approved for a mortgage, with $\hat{Y} = 2$ will be rejected, whereas with $\hat{Y} = 0$ will be asked to provide additional information and resubmit the application.

Now we turn to a classification task with multiple individuals whose risk scores are given by $\boldsymbol{S}^{test} = \{S_{n+j} : 1 \le j \le m\}$. Consider the following decision rule

$$\hat{\boldsymbol{Y}} = \{\hat{Y}_{n+j} : 1 \le j \le m\} = \{\mathbb{I}(S_{n+j} < t_l) + 2\mathbb{I}(S_{n+j} > t_u) : 1 \le j \le m\}. \tag{1}$$

We can view (1) as a *selective inference* procedure, which selects individuals with extreme scores into the high and low risk classes, while returning an indecision on the remainder. The selective inference view provides a flexible framework that allows other types of classification rules. For example, if it is only of interest to select high-risk individuals, then the action space is $\Lambda = \{0, 2\}$, and one can use the following rule

$$\hat{\boldsymbol{Y}} = \{\hat{Y}_{n+j} : 1 \le j \le m\} = \{2 \cdot \mathbb{I}(S_{n+j} > t_u) : 1 \le j \le m\}. \tag{2}$$

2.2 False selection rate and the fairness issue

In practice, it is desirable to avoid erroneous selections, which often have negative social or economic impacts. In the context of the mortgage example, approving an individual who will default (i.e., $\hat{Y} = 1$ but $Y = 2$) would increase the financial burden of the lender, while rejecting an individual who will not default (i.e., $\hat{Y} = 2$ but $Y = 1$) would lead to loss of profit. In situations where m is large, controlling the inflation of selection errors is a crucial task for policy makers. A practically useful notion is the false selection rate (FSR), which is defined as the expected fraction of erroneous decisions among all definitive decisions. We use the notation $\text{FSR}^{\mathcal{C}'}$, where $\mathcal{C}' \subset \mathcal{C} = \{1, 2\}$ is the set of class labels that we are interested in selecting. To illustrate the definition, we consider two scenarios. In the first, we select individuals from both classes using rule (1). Denote $\mathcal{S} = \{1 \le j \le m : \hat{Y}_{n+j} \ne 0\}$ the index set of the selected cases and $|\mathcal{S}|$ its cardinality. Then we have

$$\text{FSR}^{\{1,2\}} = \mathbb{E}\left[\frac{\sum_{j\in\mathcal{S}} \mathbb{I}(\hat{Y}_{n+j} \ne Y_{n+j})}{|\mathcal{S}| \vee 1}\right], \tag{3}$$

where $x \vee y = \max\{x, y\}$, and the exact meaning of $\mathbb{E}$ will become clear in Section 3 after we explicitly spell out our algorithm. In the second scenario, the goal is to select individuals in class $c = 2$ using rule (2). Then

$$\text{FSR}^{\{2\}} = \mathbb{E}\left[\frac{\sum_{j=1}^{m} \mathbb{I}(\hat{Y}_{n+j} = 2, Y_{n+j} \ne 2)}{\left\{\sum_{j=1}^{m} \mathbb{I}(\hat{Y}_{n+j} = 2)\right\} \vee 1}\right]. \tag{4}$$

$\text{FSR}^{\{1\}}$ can be defined similarly. By allowing for indecisions, one can find a decision rule that simultaneously controls both $\text{FSR}^{\{1\}}$ and $\text{FSR}^{\{2\}}$ at a small user-specified level[1].

[1]However, such a goal may not be achievable under the standard setup for classification, which does not allow indecisions. For example, it is impossible to make FSR^1 and FSR^2 to be simultaneously small when the minimum condition on the *classification boundary* (Meinshausen and Rice, 2006; Cai and Sun, 2017) fails to hold.

The FSR is a general concept for selective inference that encompasses important special cases such as the standard misclassification rate, the false discovery rate (FDR; Benjamini and Hochberg, 1995) and beyond. If we set both the state space and action space to be $\{1, 2\}$, so there are no indecisions, then the FSR defined by (3) reduces to $m^{-1}\mathbb{E}\left\{\sum_{j=1}^{m}(\hat{Y}_{n+j} \neq Y_{n+j})\right\}$, i.e. the standard misclassification rate. To see the connection to the FDR, consider a multiple testing problem with

$$H_{j0}: \ Y_{n+j} = 2 \quad vs. \quad H_{j1}: \ Y_{n+j} = 1, \quad j = 1, \dots, m.$$

The state space is $\mathcal{C} = \{1, 2\}$. A multiple testing procedure $\hat{\boldsymbol{Y}} = \{\hat{Y}_{n+j} : 1 \leq j \leq m\} \in \{0, 1\}^m$ corresponds to a selection rule that aims to select cases in class 1 only. The action space $\Lambda = \{1, 0\}$ differs from the state space $\mathcal{C}$, with $\hat{Y}_{n+j} = 1$ indicating that H_{j0} is rejected, and $\hat{Y}_{n+j} = 0$ indicating that there is not enough evidence to reject H_{j0}. Then $\text{FSR}^{\{1\}}$ precisely yields the widely used FDR, the expected fraction of false rejections among all rejections.

We use the expected proportion of indecisions (EPI) to describe the power concept (the smaller the EPI the larger the power):

$$\text{EPI} = \frac{1}{m}\mathbb{E}\left\{\sum_{j=1}^{m}\mathbb{I}\left(\hat{Y}_{n+j} = 0\right)\right\} = 1 - E(|\mathcal{S}|)/m. \tag{5}$$

Compared to erroneous decisions, the losses incurred due to indecisions are less consequential. This leads to a constrained optimization problem where the goal is to develop a selective rule that satisfies $\text{FSR} \leq \alpha$ while making the EPI as small as possible.

Next we turn to the important fairness issue in selective inference. We argue that simply controlling FSR^c may not be satisfactory in many contexts. A major concern is that the rate of erroneous decisions might be unequally shared between the protected groups, as illustrated in the COMPAS example. To address this issue, it is desirable to control the FSR for each protected attribute A. In particular, we require that the group-wise FSRs must satisfy:

$$\text{FSR}_a^{\{c\}} = \mathbb{E}\left[\frac{\sum_{j=1}^{m}\mathbb{I}(\hat{Y}_{n+j} = c, Y_{n+j} \neq c, A_{n+j} = a)}{\left\{\sum_{j=1}^{m}\mathbb{I}(\hat{Y}_{n+j} = c, A_{n+j} = a)\right\} \vee 1}\right] \leq \alpha^c, \quad \text{for all } a \in \mathcal{A}. \tag{6}$$

We aim to develop a classification rule that fulfills the fairness criterion (6). This formulation, which adopts a fairness-adjusted error rate constraint, equally bounds the fraction of erroneous decisions among protected groups.

2.3 From fair scores to fair classifiers: issues and roadmap

We investigate the important issue of what makes a "fair" classifier. In most classification tasks, the standard operation is to first construct a "base score", and then turn the score into a decision by setting a threshold. Consider a thresholding rule of the form

$$\hat{\boldsymbol{Y}} = \left\{c \cdot \mathbb{I}(S_{n+j}^c > t) : 1 \leq j \leq m\right\}. \tag{7}$$

Now we present two possible approaches to calculating S^c, and hence, selecting individuals into class c. The first, referred to as the "full covariate classifier" (FCC), thresholds the scores calculated based on

$$S^c(x, a) = \Pr\left\{Y = c | X = x, A = a\right\}. \tag{8}$$

The score is used to assess the likelihood of being in class c given all observed characteristics of an individual. However, incorporating a sensitive attribute into a classifier without appropriate

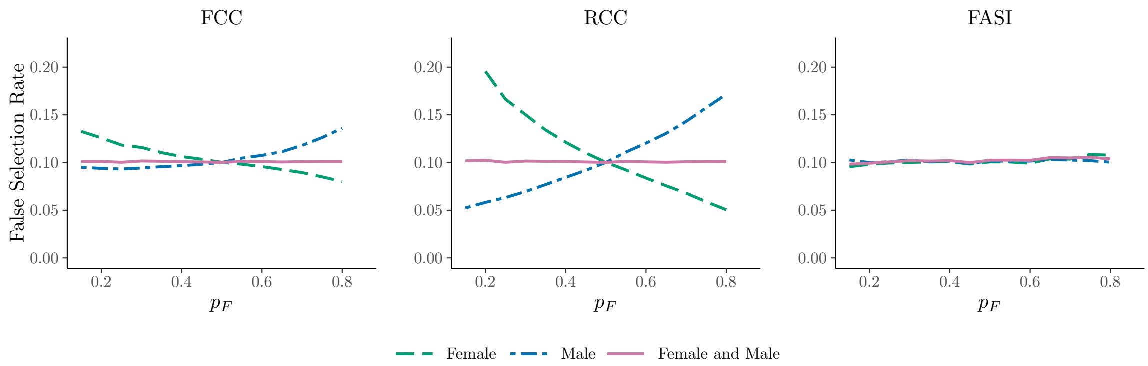

Figure 2: All plots have a fixed proportion of signals (50%) from the male group and varying proportions of signal (from 15% to 85%) for the female group. For FCC and RCC, the degree of unfairness increases as p_M and p_F become more disparate. FASI ensures that the group-wise FSRs are effectively controlled and equalized.

adjustments can lead to unfair decisions. To illustrate, consider the mortgage setup where we simulate a data set that contains a sensitive attribute "gender". The goal is to select individuals into the high risk class with FSR control at 10%; the simulation setup is detailed in Section 5. We highlight here that the proportions of individuals with label "2" (default) are different across the groups: for the male group, the proportion p_M is fixed at 50%, whereas for the female group the proportion p_F varies from 15% to 85%. We plot the overall FSR and group-wise FSRs as functions of p_F on the left panel of Figure 2. We can see that FCC controls the overall FSR but not the group-wise FSRs. Thresholding rules based on (8) are harmful in the sense that the burden of erroneous decisions is not shared equally among the two gender groups.

The second approach, referred to as the "reduced covariate classifier" (RCC), is to apply (7) by removing the sensitive attribute A:

$$S^c(x) = \Pr\{Y = c|X = x\}. \tag{9}$$

However, this approach, as illustrated in the middle panel of Figure 2, can be harmful as well. While the overall FSR is still controlled at 10%, the issue of unfairness is not mitigated but aggravates with widened gaps in the group-wise FSRs. We mention two additional drawbacks of the RCC approach. First, ignoring an informative sensitive attribute can lead to substantial power loss. Second, the feature X can be highly predictive of the sensitive attribute A; hence the classifier is likely to form a *surrogate encoding* of the sensitive attribute based on other features, leading to unfair decisions in a similar fashion as if (8) were used.

To equalize the FSRs among the protected groups, we can consider two strategies: (a) construct new scores by adjusting the current base scores so that the new scores are fair and comparable across groups, or (b) stick to the original base scores and set varied group-adjusted thresholds. Strategy (a), which applies a universal threshold to all individuals, is appealing because the decision making process would be straightforward once adjusted scores are given to practitioners, and the new scores are comparable across the groups. By contrast, strategy (b), although working equally effectively for addressing the fairness issue, can be less intuitive and nontrivial to implement. For practitioners without a full understanding of the underlying algorithm, strategy (b) can be confusing and even controversial as varied thresholds are being used for different protected groups, causing another level of concern about possible discrimination.

As a result, we adopt strategy (a). We call a score S^c (with respect to class c) *unfair* or *illegal* if thresholding rule (7), which applies a universal cutoff t to all scores, produces larger FSRs for one group over the others. We can see from Figure 2 that both conditional probabilities (8) and (9) are unfair scores. Two key issues in strategy (a) are: (i) how to construct fair scores and (ii) how to go from fair scores to fair classifiers. Next we develop a data-driven FASI algorithm based on the R-value to address the two issues integrally. The empirical performance of our algorithm is illustrated on the right panel of Figure 2. We can see that the FSRs are effectively controlled at the nominal level and properly equalized across the protected groups.

3 Methodology

This section develops a fairness-adjusted FSR controlling procedure for binary classification. We focus on the goal of controlling the FSR^c defined in (4). The methodologies for controlling the FSR of the form (3) and the case of multinomial classification will be briefly discussed in Section 7.

A major challenge in our methodological development is that most state-of-the-art classifiers are constructed based on complex black-box models, which do not offer theoretical guarantees about the outputs. This makes the uncertainty assessment, particularly the analysis of error rates, very challenging and even intractable. We address the challenge by developing a model-free framework that applies to any black-box classifier, and our methodology only leverages mild assumptions on exchangeability between the observed and future data. In contrast with existing theories that are asymptotic and require strong assumptions on models or classifiers, the new framework not only eliminates these assumptions, but also yields a powerful finite-sample theory on FSR control.

3.1 The R-value and FASI algorithm

We first introduce a significance index, called the R-value, for ranking individuals and then discuss how the R-values can be converted to a selection rule via thresholding.

The FASI algorithm consists of three steps: training, calibrating and thresholding. The observed data set $\mathcal{D} = \{(X_i, A_i, Y_i) : 1 \le i \le n\}$ is divided into a training set and a calibration set: $\mathcal{D} = \mathcal{D}^{train} \cup \mathcal{D}^{cal}$. The testing set is denoted $\mathcal{D}^{test}$. The first step trains a score function, denoted $\hat{S}^c(x, a)$, on $\mathcal{D}^{train}$. By convention a larger score indicates a higher probability of belonging to class c. The scores, often representing estimated class probabilities, can be generated from any user-specified classifier. We make no assumptions on the accuracy of these scores. The only consequence of inaccurate scores is that the number of indecisions will increase, meaning that more effort will be needed to perform separate evaluations.

In the second step, we first calculate $\hat{S}^c(x, a)$ for individuals in both $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$, then calibrate an *R-value* for every individual in $\mathcal{D}^{test}$. Specifically, denote $\hat{R}^c_{n+j} := \hat{R}^c_{n+j}(X_{n+j} = x, A_{n+j} = a)$ and $\hat{s} := \hat{S}^c(X_{n+j} = x, A_{n+j} = a)$, we define

$$\hat{R}^c_{n+j} = \frac{\frac{1}{n^{cal}_a+1}\left\{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}\left(A_i = a, \hat{S}^c_i \ge \hat{s}, Y_i \ne c\right) + 1\right\}}{\frac{1}{m_a}\sum_{i\in\mathcal{D}^{test}} \mathbb{I}\left(A_i = a, \hat{S}^c_i \ge \hat{s}\right)} \quad \text{if } A_{n+j} = a, \tag{10}$$

where $m_a = \sum_{i\in\mathcal{D}^{test}} \mathbb{I}(A_i = a)$ and $n^{cal}_a = \sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(A_i = a)$. In situations where the test set is small, we consider the following modified R-value that utilizes both the test and calibration set in the denominator:

$$\hat{R}^{c,+}_{n+j} = \frac{\frac{1}{n^{cal}_a+1}\left\{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}\left(A_i = a, \hat{S}^c_i \ge \hat{s}, Y_i \ne c\right) + 1\right\}}{\frac{1}{m_a+n^{cal}_a+1}\left\{\sum_{i\in\mathcal{D}^{test}\cup\mathcal{D}^{cal}} \mathbb{I}\left(A_i = a, \hat{S}^c_i \ge \hat{s}\right) + 1\right\}} \quad \text{if } A_{n+j} = a, \tag{11}$$

for $1 \leq j \leq m$. Section D in the supplement shows that the R^{+}-value (11) provides a more stable score than the R-value (10) when $|\mathcal{D}^{test}|$ is small.

Now we provide some intuition behind the R-value. Roughly speaking, the R-value corresponds to the smallest group-wise FSR such that the $(n+j)$th individual is *just selected*. In other words, if we make the cut at $\hat{R} = r$, e.g. selecting all instances with R-values less than or equal to r into class c, then we expect that, for every group $a \in \mathcal{A}$, approximately $100r\%$ of the selected cases do not belong to class c. This interpretation is similar to the q-value (Storey, 2003) in FDR analysis.

As the R-value can be interpreted as a fraction, we set the R-value to 1 if the quantity in (10) or (11) exceeds 1. The fairness notion has been naturally incorporated into the (group-adjusted) R-value, making it possible to calibrate a universal threshold equalizing the FSRs across the groups. More explanations of (10) and (11) are provided in the next section.

In the third thresholding step, we compare the R-value defined in (10) or (11) with a pre-specified FSR level α_c. For example, if we are interested in selecting individuals into class c, then the decision rule is

$$\hat{\boldsymbol{Y}} = \left\{c \cdot \mathbb{I}(R^{c}_{n+j} \leq \alpha_c) : 1 \leq j \leq m\right\},$$

where the threshold is simply the user-specified α_c. According to the interpretation of the R-value, this step amounts to choosing the largest possible threshold subject to the constraint on group-wise FSRs. Intuitively this can best reduce the wasted opportunities induced by indecisions. If we are interested in selecting both classes, then the decision rule is $\hat{\boldsymbol{Y}} = \sum_{c=1}^{2}\left\{c \cdot \mathbb{I}(R^{c}_{n+j} \leq \alpha_c) : 1 \leq j \leq m\right\}$. To avoid assigning an individual to multiple classes, we classify the individual to the class with the smaller R-value when there is overlapping selection.

The proposed FASI algorithm has several attractive properties. First, as we explain shortly, the R-value provides an estimate for a fraction, which is standardized between 0 and 1, comparable across protected groups, and easily interpretable. Second, the FSR analysis via R-values is straightforward: practitioners can make decisions directly with the R-values; the threshold is simply the user-specified FSR level. The fairness consideration, wrapped up in the R-value definition, is addressed in a proper and clean way. Finally, the FASI algorithm is model-free and offers a powerful theory on FSR control; this is discussed next.

3.2 Why FASI works?

Now we explain why the FASI algorithm works. The effectiveness of our algorithm only leverages a mild exchangeability assumption.

Assumption 1. *The calibration data and testing data are independent. The triples* $\{(X_i, A_i, Y_i) : i \in \mathcal{D}^{cal} \cup \mathcal{D}^{test}\}$ *are exchangeable.*

We start by explaining why the R-value provides a sensible estimate of the FSR. To simplify the discussion, we ignore the sensitive attribute A for the moment and consider a thresholding rule of the form $\hat{\boldsymbol{Y}} = \{\mathbb{I}(\hat{S}^{c}_{n+j} \geq t) : 1 \leq j \leq m\}$. Consider the false selection proportion (FSP) process for $\mathcal{D}^{test}$:

$$\text{FSP}(t) = \frac{\sum_{i \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}^{c}_{i} \geq t, Y_i \neq c\right)}{\sum_{i \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}^{c}_{i} \geq t\right)}, \tag{12}$$

with $\text{FSP}(t) = 0$ if no individual is selected. The FSP cannot be computed from data because we do not observe the true states $\{Y_i : i \in \mathcal{D}^{test}\}$. The good news is that under Assumption 1 on exchangeability, the unobserved process $\sum_{i \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}^{c}_{i} \geq t, Y_i \neq c\right)$ will strongly resemble its "mirror process" in the calibration data $\sum_{i \in \mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}_i^c \geq t, Y_i \neq c\right)$. Constructing a mirror process and exploiting the symmetry property to make inference is a powerful idea that has been explored in recent works (e.g. Barber and Candès, 2015; Du et al., 2021). Finally, adjusting for the possible unequal sample sizes between $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$, we obtain the R-value process

$$\widehat{R}^c(t) = \frac{\frac{1}{n^{cal}+1}\left\{\sum_{i \in \mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}_i^c \geq t, Y_i \neq c\right) + 1\right\}}{\frac{1}{m}\sum_{i \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}_i^c \geq t\right)}, \tag{13}$$

where m and n^{cal} are respectively the cardinalities of $\mathcal{D}^{test}$ and $\mathcal{D}^{cal}$. Finally, the fairness-adjusted R-value defined in (10) can be recovered by restricting the R-value process to a specific group $a \in \mathcal{A}$ and substituting $\hat{s}$ in place of t in (13). The R^+-value defined by (11) can be conceptualized in a similar fashion.

Remark 1. Comparing (12) and (13), we note that "+1" has been incorporated into the count of false selections on $\mathcal{D}^{cal}$. This technical adjustment has virtually no impact on the empirical performance of FASI. However, it ensures that (13) effectively leads to a martingale, which is essential for proving the theory. It is natural to apply the same "+1" adjustment to n^{cal}, which makes the algorithm slightly more powerful.

Next we state a theorem that establishes the finite-sample property of FASI. Our theory fundamentally departs from those in existing works: we do not make assumptions regarding the accuracy of $\hat{S}_i^c$. The accuracy of the classifier only affects the power, not the validity for FSR control. See Section 4.3 for practical guidelines on how to construct more informative classifiers/R-values.

Theorem 1. *Suppose Assumption 1 holds. Then we have:*

1. *The FASI algorithm with R-value (10) controls the group-wise FSRs at level α_c, e.g., $FSR_a^{\{c\}} \leq \alpha_c$ for all $a \in \mathcal{A}$.*
2. *The FASI algorithm with R-value (11) satisfies $FSR_a^{\{c\},*} \leq \alpha_c$, where*

$$FSR_a^{\{c\},*} = \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{I}(\hat{Y}_{n+j} = c, Y_{n+j} \neq c, A_{n+j} = a)}{\sum_{j=1}^m \mathbb{I}(\hat{Y}_{n+j} = c, A_{n+j} = a) + 1}\right]. \tag{14}$$

Remark 2. In the modified FSR definition (14), the "+1" adjustment is used to account for the extra uncertainty in the approximation of the number of rejections. A similar modification, in the context of FDR analysis but for different reasons, has been used in Theorem 1 of Barber and Candès (2015).

Finally, we mention that the R-value has a nice interpretation under the *conformal inference* framework. Section A.3 in the Supplementary Material shows that a variation of our R-value corresponds to the Benjamini-Hochberg (BH) adjusted q-value of the *conformal p-values* (Bates et al., 2021) under the *one-class classification* setting. The connection to conformal inference and the BH method, both of which are model-free, provides insights on why the FASI algorithm is assumption-lean and offers exact FSR control in finite samples as claimed in Theorem 1.

4 Theoretical Issues

In this section we discuss a few important theoretical issues, which provide an in-depth investigation of the R-value and place our contributions in context. Section 4.1 explains the challenges and novelty of our FSR theory and sketches the main ideas in the proof of Theorem 1. In Sections 4.2 and 4.3, we introduce the theoretical R-value and derive the optimal score function under a simplified setup. The theory, which is intuitive and sensible, provides insights for practitioners regarding how to train score functions for constructing informative R-values.

4.1 Theory on FSR control: main ideas and contributions

We explain the main idea behind the proof of Theorem 1. The discussion will focus on the R-value process (13), but can be easily extended to the group-adjusted R-value (10). Three major challenges in the theoretical analysis include (a) how to handle the unknown and complex dependence between the scores $\hat{S}_i^c$ [as the same training data have been used to compute the scores in (13)], (b) how to evaluate the FSR in classification without knowledge about the theoretical properties of the scores, and (c) how to develop non-asymptotic guarantees on the performance of the FASI algorithm in finite samples.

Inspired by the elegant ideas in the FDR literature (Storey et al., 2004; Barber and Candès, 2015), we have carefully constructed the R-values so that the corresponding FSP process (13) can be stochastically bounded above by a martingale. In the proof of Theorem 1, we first show that the threshold induced by the FSP process is a *stopping time*, and then apply Doob's optional stopping theorem to obtain an upper bound for the expectation of the martingale. Finally we leverage the exchangeability assumption to cancel out the cardinality adjustments and establish the upper bound for the FSR. We stress that our theory utilizes no assumptions on the underlying models or quality of scores, and the control is *exact for finite samples*.

A closely related idea is the *conformal p-value*, which was independently proposed in a recent work by Bates et al. (2021). In Section A.3 of the Supplementary Material, we show that the R-value, which was motivated from a very different perspective, can be derived as the BH q-value of conformal p-values under the one-class classification setting (Moya and Hush, 1996; Khan and Madden, 2009; Kemmler et al., 2013). The theory on FDR control in Bates et al. (2021) needs to deal with a similar complication as the conformal p-values are also dependent. Bates et al. (2021) adopted a novel approach by first showing that the conformal p-values satisfy the PRDS (i.e. *positive regression dependent on a subset*) condition, and second applying the theory in Benjamini and Yekutieli (2001) to prove the FDR control. We conjecture that the PRDS approach could be extended to establish our theory. However, the extension seems to be non-trivial because, as we point out at the end of Section A.3, under the binary classification setup our R-values do not utilize conformal p-values explicitly. Our direct application of the martingale theory seems to be simpler and equally effective.

4.2 Theoretical R-value

Our discussions in the next two subsections assume an oracle with access to all distributional information, make several simplifying assumptions and are purely theoretical. The major goal is to develop a theoretical version of the R-value and an optimality theory for FSR control.

Our discussions are based on the following mixture model

$$F(x) = \sum_{a \in \mathcal{A}} \{\pi_{1,a} F_{1,a}(x) + \pi_{2,a} F_{2,a}(x)\}, \tag{15}$$

where $F_{c,a}$ is the conditional CDF of X from class c with attribute a and $\pi_{c,a} = P(Y = c, A = a)$, $c = 1, 2$. Denote $\pi_a = P(A = a)$, $\pi_{c|a} = P(Y = c|A = a)$ and $f_{c,a}(x)$ the corresponding conditional density functions. Consider a selection rule of the form $\hat{Y}(t) = c \cdot I(S^c \geq t)$, where t denotes a threshold. Suppose an oracle knows the conditional probabilities and conditional CDFs defined above. In Appendix A.1, we discuss an algorithm that converts the base score $S^c = s$ (possibly unfair) to a fair score. Specifically, for an individual from group a, with $S^c = s$, the conversion algorithm yields the following *theoretical R-value*:

$$R^c(s) = \inf_{t \leq s} \left\{ Q_a^c(t) := \mathbb{P}(Y \neq c|\hat{Y}(t) = c, A = a) \right\}, \tag{16}$$

where $Q_a^c(t)$ is the conditional error probability when the threshold is t[2]. We shall see that the group-wise FSR (6) is closely connected to $Q_a^c(t)$ (Appendix A.2), and that the theoretical R-value can be viewed as the counterpart of the data-driven R-value defined in (10). It corresponds to the smallest conditional probability such that the individual with score $S^c = s$ is *just selected* into class c.

The next proposition, which follows directly from (16), shows that thresholding the theoretical R-value leads to a fair selective inference procedure.

Proposition 1. *Consider a classifier that claims $\hat{Y} = c$ if $R^c \leq \alpha$. Then*

$$\mathbb{P}(Y \neq c|\hat{Y} = c, A = a) \leq \alpha \text{ for all } a \in \mathcal{A}. \tag{17}$$

The theoretical R-value is a fundamental quantity that is closely connected to the sufficiency principle in the fairness literature (Section 7.1). It also plays a central role in developing the optimality theory (next section).

4.3 A sketch of the optimality theory

We state and prove an intuitive result that $S^c(x, a) = \mathbb{P}(Y = c|X = x, A = a)$ is the optimal choice of score function for calibrating the theoretical R-value. A few simplifying assumptions will be required.

Consider random mixture model (15). Suppose an oracle knows the score function $S^c(x, a) = \mathbb{P}(Y = c|X = x, A = a)$. The goal is to assign labels "0", "1" and "2" to new instances $\{(X_{n+j}, A_{n+j}) : 1 \leq j \leq m\}$. We assume that the instances (X_j, A_j) are independent draws from an underlying distribution $F(x, a)$. Our optimality theory will be developed based on a variation of the FSR referred to as the marginal FSR:

$$\text{mFSR}_a^c = \frac{\mathbb{E}\left\{\sum_{1 \leq j \leq m: A_{n+j}=a} \mathbb{I}(\hat{Y}_{n+j} = c, Y_{n+j} \neq c)\right\}}{\mathbb{E}\left\{\sum_{1 \leq j \leq m: A_{n+j}=a} \mathbb{I}(\hat{Y}_{n+j} = c)\right\}}.$$

The relationship between the mFSR and FSR is discussed in Section A.2. We aim to develop a selection rule under the binary classification setting that solves the following constrained optimization problem:

$$\text{minimize the EPI subject to } \text{mFSR}_a^c \leq \alpha_c,\ c = 1, 2 \text{ for all } a \in \mathcal{A}. \tag{18}$$

[2] If the base score satisfies the monotone likelihood ratio condition (MLRC, Cao et al., 2013) then the infimum is achieved at s exactly. See Section C.1 for related discussions.

Denote $S^c_{n+j} = \mathbb{P}(Y_{n+j} = c | X_{n+j} = x_{n+j}, A_{n+j} = a_{n+j})$. The scores can be converted to theoretical R-values, denoted R^1_{n+j} and R^2_{n+j}. The process of conversion is described in the proof of Theorem 2. Define the oracle procedure

$$\boldsymbol{\delta}_{OR} = \{\delta^j_{OR} : 1 \le j \le m\}, \text{ where } \delta^j_{OR} = \mathbb{I}(R^1_{n+j} \le \alpha_1) + 2\mathbb{I}(R^2_{n+j} \le \alpha_2). \tag{19}$$

The optimality of the oracle procedure is established in the next theorem.

Theorem 2. *Consider random mixture model (15). Assume that α_1 and α_2 have been properly chosen such that (19) does not have overlapping selections. Let $\mathcal{D}_{\alpha_1,\alpha_2}$ denote the collection of selection rules that satisfy $mFSR^c_a \le \alpha_c$ for $c = 1, 2$ and all $a \in \mathcal{A}$. Let $EPI_{\boldsymbol{\delta}}$ denote the EPI of an arbitrary decision rule $\boldsymbol{\delta}$. Then the oracle procedure (19) is optimal in the sense that $EPI_{\boldsymbol{\delta}_{OR}} \le EPI_{\boldsymbol{\delta}}$ for any $\boldsymbol{\delta} \in \mathcal{D}_{\alpha_1,\alpha_2}$.*

The optimality theory has several important implications for practitioners. The choice of optimal score indicates that in the training stage, we should construct the score functions using *all features*, including the sensitive attribute A, to best capture the individual level information. The scores trained without the sensitive attribute [e.g. (9)], are usually suboptimal. The adjustments for fairness should not be made in the training stage but in the calibration stage, where the fully informative scores can be converted to the R-values to adjust the disparity in error rates across the groups.

A by-product of our theory is the finding that the optimal selection rule also equalizes the group-wise error rates. The intuition is that in order to maximize the EPI, the pre-specified mFSRs must be *exhausted in all separate groups*; hence the group-wise mFSRs are all equal to the nominal level (hence automatically equalized). We conjecture that asymptotic equality holds for our FASI algorithm, which has been corroborated by our numerical studies. A full analysis is complicated due to the dependence between the scores; we leave this for future research.

5 Simulation Results

This section presents the results from two simulation scenarios comparing FASI to the Full Covariate Classifier (FCC). The Restricted Covariate Classifier (RCC) is not included since, in our simulations, it has systematically larger deviations from the target group-wise FSR levels compared to FCC. We demonstrate that both the oracle and data-driven versions of FASI are able to correctly control the group-wise FSRs, while RCC fails to do so. The oracle versions of FASI and FCC use the exact posterior probabilities for S_i, defined in Equation 8, while the data-driven versions estimate S_i via a GAM classifier (Hastie et al., 2009; James et al., 2013). In our experience, similar patterns are observed when other classifiers are used to construct the base scores.

All simulations are run with sample sizes of $|\mathcal{D}| = 2{,}500$ and $|\mathcal{D}^{test}| = 1{,}000$. We generate $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$ using a random split of $\mathcal{D}$, with $|\mathcal{D}^{train}| = 1{,}500$ and $|\mathcal{D}^{cal}| = 1{,}000$. We use gender as our protected attribute taking two values $A = F$ (females) and $A = M$ (males). The feature vectors $\boldsymbol{X} \in R^3$ are simulated according to Model 15 with four components:

$$F(\cdot) = \pi_M\{\pi_{1|M}F_{1,M}(\cdot) + \pi_{2|M}F_{2,M}(\cdot)\} + \pi_F\{\pi_{1|F}F_{1,F}(\cdot) + \pi_{2|F}F_{2,F}(\cdot)\},$$

where $\pi_a = \mathbb{P}(A = a)$, $\pi_{c|a} = \mathbb{P}(Y = c|A = a)$ and $F_{c,a}$ is the conditional distribution of $\boldsymbol{X}$ given $Y = c$ and $A = a$. Let $\pi_M = \pi_F = 0.5$, i.e. the numbers of females and males in the data set are equal. We will consider two scenarios in our simulation study that follow this setup.

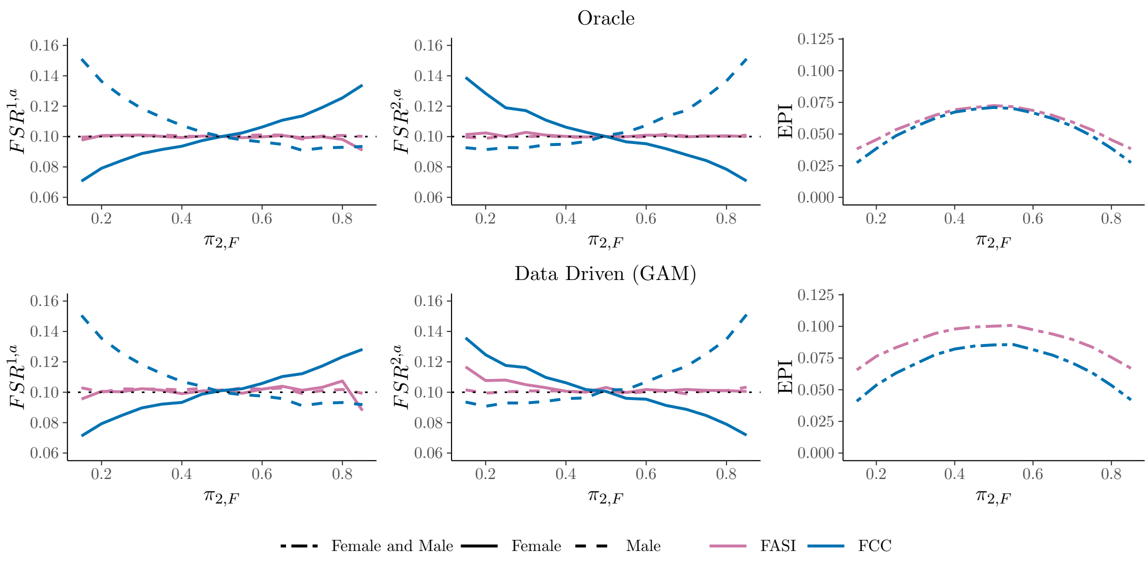

Figure 3: Simulation 1. Top row: The oracle procedure. Bottom row: A data-driven GAM fitting procedure. Left and middle column: $FSR^{1,a}$ and $FSR^{2,a}$ for both females and males. Right column: The expected proportion of indecision's (EPI). All plots vary the proportion of true class 2 observations from the Female protected group, $\pi_{2,F}$.

In the first scenario, the conditional distributions of $\boldsymbol{X}$ given class Y are assumed to be multivariate normal and are identical for males and females:

$$F_{1,M} = F_{1,F} = \mathcal{N}(\boldsymbol{\mu}_1, 2 \cdot \mathbf{I}_3), \quad F_{2,M} = F_{2,F} = \mathcal{N}(\boldsymbol{\mu}_2, 2 \cdot \mathbf{I}_3),$$

where $\mathbf{I}_3$ is a 3×3 identity matrix, $\boldsymbol{\mu}_1 = (0, 1, 6)^\top$ and $\boldsymbol{\mu}_2 = (2, 3, 7)^\top$. The only difference between males and females is in the conditional proportions: we fix $\pi_{2|M} = \mathbb{P}(Y = 2|A = M) = 0.5$, while varying $\pi_{2|F} = \mathbb{P}(Y = 2|A = F)$ from 0.15 to 0.85.

We simulate 1,000 data sets and apply both the FCC and FASI [with R-values defined in (11)] at FSR level 0.1 to the simulated data sets. The FCC method ignores the protected attributes when computing the R-values, i.e.

$$\hat{R}^{c,\mathrm{FCC}}_{n+j} = \frac{\frac{1}{n^{cal}_a+1} \left\{ \sum_{i \in \mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}^c_i \geq \hat{s}, Y_i \neq c \right) + 1 \right\}}{\frac{1}{m_a + n^{cal}_a + 1} \left\{ \sum_{i \in \mathcal{D}^{test} \cup \mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}^c_i \geq \hat{s} \right) + 1 \right\}}.$$

The corresponding selection rule is $\hat{\boldsymbol{Y}}^{\mathrm{FCC}} = \left\{ c \cdot \mathbb{I}(R^{c,\mathrm{FCC}}_{n+j} \leq \alpha_c) : 1 \leq j \leq m \right\}$.

The FSR levels are computed by averaging the respective FSPs from 1,000 replications. The simulation results are summarized in Figure 3. The first and second rows respectively correspond to the oracle and data-driven versions of each method. The first two columns respectively plot the group-wise FSRs for class 1 and class 2 as functions of $\pi_{2|F}$. The final column plots the expected proportion of indecisions (EPI) obtained by averaging the results from 1,000 replications. The following patterns can be observed.

- Both the FASI method and FCC control the global FSR. For simplicity, we do not include these results in the figures below.

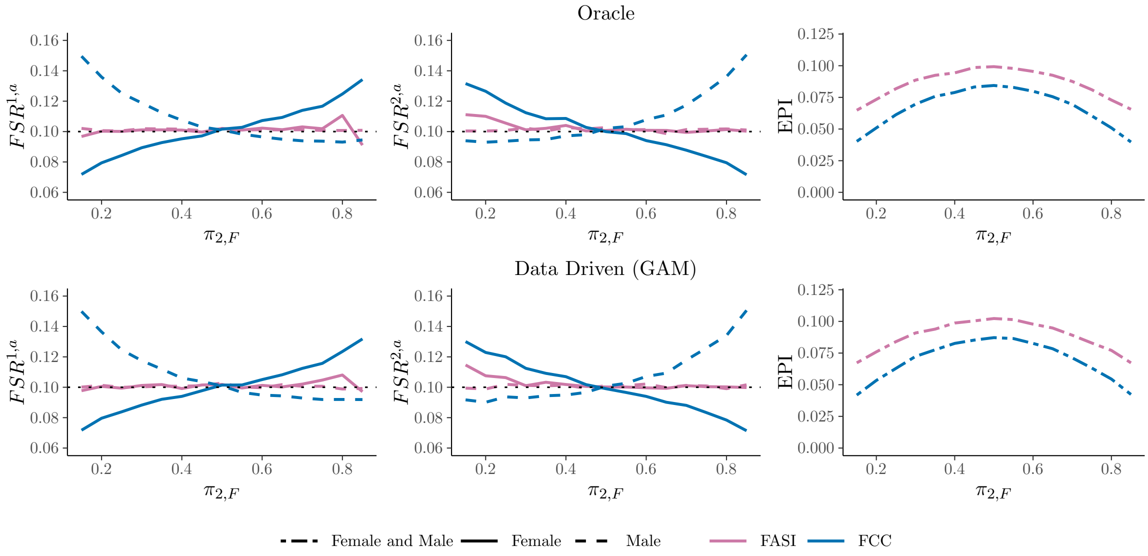

Figure 4: Simulation 2. Comparable setup to Simulation 1 except that the female and male distributions now differ from each other. However, the results are similar with FASI correctly controlling the FSRs.

- Shifting our focus to the group-wise FSRs, FCC fails to control the error rate. When $\pi_{2|F} = 0.5$, by construction we have $\pi_{2|F} = \pi_{2|M}$, making the Female and Male attributes indistinguishable. However, as $\pi_{2|F}$ moves away from $\pi_{2|M} = 0.5$, the gap between the FSR control for Females and Males dramatically widens due to the asymmetry in the proportions of the signals (true class 2 observations) in the male and female groups.
- In comparison, both oracle and data-driven FASI algorithms are able to roughly equalize the group-wise FSRs between the Female and Male groups. The data-driven version of FASI is able to closely mirror the behavior of the oracle method. The FSR control is in general effective except that the FSR levels are slightly elevated in the tails.
- The parity in FSR control is achieved at the price of slightly higher EPI levels.

Our second simulation scenario considers the setting where $F_{c,M} \neq F_{c,F}$, for both $c = 1, 2$. Denoting the mean of each distribution for class c and protected attribute a as $\boldsymbol{\mu}_{c,a}$, the data is generated from $F_{c,a} = \mathcal{N}(\boldsymbol{\mu}_{c,a}, 2 \cdot \mathbf{I}_3)$, with components $\boldsymbol{\mu}_{1,M} = (0, 1, 6)^\top$, $\boldsymbol{\mu}_{2,M} = (2, 3, 7)^\top$, $\boldsymbol{\mu}_{1,F} = (1, 2, 7)^\top$ and $\boldsymbol{\mu}_{2,F} = (3, 4, 8)^\top$. In all other respects Simulations 1 and 2 are identical. The results for the second simulation scenario are provided in in Figure 4.

We notice very similar patterns to our first simulation setup. Both FASI and FCC are able to control the global FSR (omitted from the figure). FASI controls the group-wise FSRs for all values of $\pi_{2|F}$ while the FCC fails to do so. The data-driven FASI closely emulates the oracle procedure, for both the FSR and EPI levels.

6 Real Data Examples

In this section we demonstrate FASI's effectiveness on two real world case studies. In Section 6.1 we examine the Compas recidivism algorithm made popular by ProPublica in 2016 (Angwin et al., 2016), while in Section 6.2 we use US census data from 1994 to predict an individual's salary (Dua and Graff, 2017). In both cases we compare FASI to the FCC approach described in Section 5 by

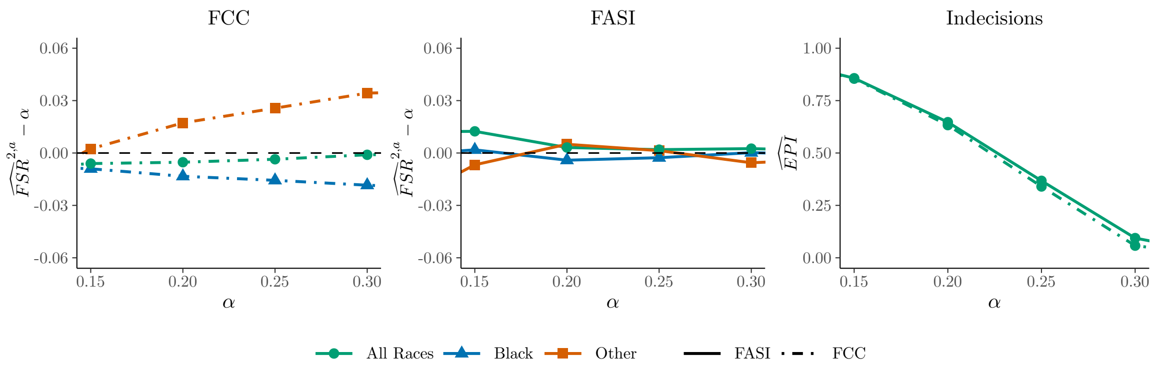

Figure 5: COMPAS data analysis for predicting recidivists. Left and Middle: False Selection Rate minus the desired control level for varying levels of α for the FCC and FASI method respectively. Right: The EPI for both the FCC and FASI method.

randomly assigning 70% of our data set to $\mathcal{D}$ and the remaining 30% to $\mathcal{D}^{test}$. We further evenly split $\mathcal{D}$ into $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$.

Since this is a real data setting, the true posterior probabilities for S_i, defined in Equation 8, are unavailable to us. To estimate them, we used a GAM fitting procedure for the compas case study and an Adaboost fitting procedure for the census income prediction case study (Hastie et al., 2009; James et al., 2013).

This method can be applied to other applications through the *"fasi"* package available in the R language on CRAN.

6.1 COMPAS Data Analysis

In 2016, ProPublica's investigative journalists curated a data set of 6,172 individuals, 3,175 of whom were Black and the remaining 2,997 other races, that were arrested in Broward County, Florida. In this study, Black and Other are our protected attributes.

The Black and Other groups respectively had 1,773 and 1,217 individuals who actually recidivated in the 2-year time frame that the study considered. We used this two year window as a proxy for the true label of identifying recdivists.

All individuals were assigned a risk score by the compas algorithm (a whole number between 1 and 10) developed by NorthPointe Inc. This score was used to inform the judge of each person's risk of recidivating during their bail hearing. The data set contains demographic information about each person including their race, age, number of previous offenses, sex, number of prior offenses, and their assigned compas risk score.

We randomly split the data set 1,000 times into $\mathcal{D}$ and $\mathcal{D}^{test}$, and averaged the difference between the actual and target recidivism FSR's for a range of α between 0.15 and 0.30. The first two columns of Figure 5 provide the difference between true and target FSR_2 for FCC and FASI respectively. The last column plots the overall EPI for each method.

As we noted with the simulated data, while the FCC does a good job at controlling the overall FSR (green / circle) of recidivists, it is unable to do this at the race level. In the left hand plot of Figure 5, the breakdown of the race-wise FSR control for the FCC is shown. The Black attribute (blue / triangle) systematically has FSR control lower than the desired target level. While the Other attribute (orange / square) systematically has higher FSR control than the target level. This observation holds for all values of α considered.

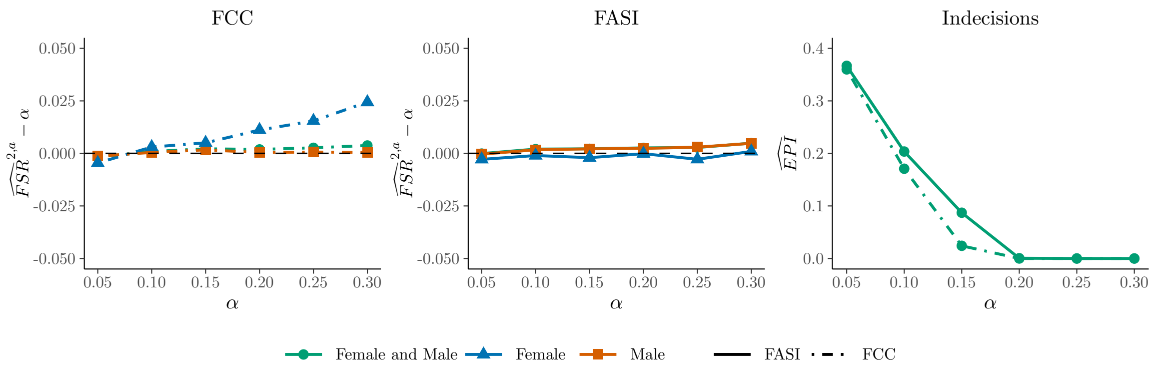

Figure 6: Census income prediction for individuals that earn more than 50K a year. Left and Middle: False Selection Rate minus the desired control level for varying levels of α for the FCC and FASI method respectively. Right: The EPI for both the FCC and FASI method.

In comparison, the middle plot in Figure 5 shows the FASI method. For all values of α, the FSR is controlled at the desired level for both the protected attributes and for all observations. The right plot in Figure 5 also demonstrates that, in this study, FASI is able to obtain a nearly identical EPI to the FCC. This demonstrates that the price of our fairness constraint, in terms of the size of the indecision group, is nearly zero.

6.2 1994 Census Income Data Analysis

The US census is the leading body of information for producing information about the American people. Naturally, the data that they collect can directly inform future policy decisions, such as funding programs that provide economic assistance for populations in need. In particular, resources such as food, health care, job training, housing, and other economic assistance rely upon good estimates of a population's income levels. The cost of making unfair decisions when predicting ones income can be severe since the prediction helps determine how hundreds of billions of dollars in federal funding are spent for the next decade. In this case study, we use the 1994 US Census Data set from the UCI Machine Learning Repository to predict if an individual earns more than 50,000 dollars a year.

The data consist of 32,561 observations on 14, largely demographic, variables including education level, age, hours worked per week, and others. The protected attributes in this study are Female and Male. The Female group has 10,771 total observations, of which 1,179 make over \$50K a year. Similarly, the remaining 21,790 observations are from the Male attribute, of which 6,662 make over \$50K a year.

As in Section 6.1, we compare the FCC approach to the FASI method for many values of FSR control, α, ranging from 0.05 to 0.3. The left most plot of Figure 6 shows results from the FCC method, where both the overall (green / circle) and Male (orange / square) has the desired FSR control. However, the FCC is unable to maintain the desired FSR control for the Female attribute (blue / triangle) across all values of α. In comparison, the FASI method shown in the middle plot is able to maintain both the overall FSR control as well as the FSR control for Females and Males across all values of α.

The right most plot provides the estimated EPI of each approach. Unlike for the Compas data, for some values of α the FASI method returns a slightly larger indecision group on average in comparison to the FCC method.

7 Discussions

We conclude this article with a discussion of other fairness notions and related error rate concepts.

7.1 Connection to the sufficiency principle

Consider the theoretical R-value (16). Theorem 1 shows that thresholding the theoretical R-value at level α ensures that $\mathbb{P}(Y \neq c|\hat{Y} = c, A = a) \leq \alpha$ for all $a \in \mathcal{A}$, which is referred to as the *sufficiency principle* or *predictive parity* in the fairness literature (Crisp, 2003; Barocas et al., 2017; Chouldechova, 2017). The conversion algorithm described in Appendix A.1 can be viewed as a method of *calibration by group* (Barocas et al., 2017), which leads to the fulfillment of the sufficiency principle automatically.

The connection between the theoretical R-value (16) and our (empirical) R-value (10) becomes clear by noting that (a) the empirical R-value can be conceptualized as the smallest group-wise FSR such that the individual is just selected, whereas the theoretical R-value corresponds to the smallest conditional probability that the individual is just selected; (b) the group-adjusted FSRs (6) can be viewed as empirical counterparts of the conditional probabilities (17) (Appendix A.2). This interesting connection indicates that our FASI algorithm leads to classification rules that approximately satisfy the sufficiency principle.

Our work departs from existing machine learning algorithms tailored to fulfill the sufficiency principle in several ways. The FASI algorithm in Section 3.1 is intuitive and easily implementable. However, existing calibration by group methods are often complicated and computationally intensive; the score conversion strategy discussed above, which assumes an oracle and involves many complicated unknown distributions, can be intractable in practice. Moreover, existing machine learning algorithms have no statistical guarantees on the error rate control, and there is no discussion about the inflation of decision errors when multiple individuals must be classified. By contrast, the FASI algorithm offers exact error rate control in finite samples and addresses the multiplicity issue by controlling the FSR, which is motivated by the powerful and popular FDR idea in multiple hypothesis testing. In contrast with existing methods, our theory works for sophisticated classifiers trained from black-box models without requiring conditions on the accuracy of the outputs.

7.2 Other fairness notions

In addition to the sufficiency principle, a widely used fairness notion is the *separation principle* (Barocas et al., 2017), which requires that

$$P(Y \neq \hat{Y}|Y = c, A = a) \text{ are the same for all } a \in \mathcal{A}. \tag{20}$$

A third notion on fairness, in the context of prediction intervals, has been considered in Romano et al. (2020a). Rather than conditioning on either Y or $\hat{Y}$, these works are concerned with the joint probabilities of $(\hat{Y}, Y)$. This fairness criterion requires that the misclassification rates are equalized across all protected groups:

$$P(Y \neq \hat{Y}|A = a) \text{ are the same for all } a \in \mathcal{A}. \tag{21}$$

Other popular fairness notions include *equalized odds* (Hardt et al., 2016; Romano et al., 2020b) and *equalized risks* (Corbett-Davies and Goel, 2018). A highly controversial issue is that different fairness criteria often lead to different algorithms and decisions in practice. For example, the sufficiency and separation principles can be incompatible with each other (Kleinberg et al., 2016; Friedler et al., 2021), and the classification parity and calibration can harm the very groups that the algorithms are designed to protect (Corbett-Davies and Goel, 2018). We do not claim that FASI is universally superior than competitive approaches but adjusting group-wise FSRs appears to be a reasonable fairness criterion for the applications under our consideration. Much research is still needed to fully understand the trade-offs and caveats between different approaches to fairness-adjusted inference.

7.3 FSR concepts in multinomial classification

The selective inference framework and FSR concepts can be extended from the binary classification setting to more general settings. Denote the collection of all class labels by $\mathcal{C} = \{1, \ldots, C\}$. The case with $C = 1$ corresponds to the one-class classification problem discussed in Appendix A.3; particularly the outlier detection problem recently considered in Guan and Tibshirani (2021) and Bates et al. (2021) can be encompassed by our general framework.

For situations with $C \geq 2$, denote the set of classes to be selected by $\mathcal{C}'$, and assume $\mathcal{C}' \subset \mathcal{C}$. With indecisions being allowed, the action space is given by $\Lambda = \{0, \mathcal{C}'\}$. Denote the selection rule $\hat{\boldsymbol{Y}} = \{\hat{Y}_{n+j} : 1 \leq j \leq m\} \in \Lambda^m$. Then the FSR with respect to subset $\mathcal{C}'$ is defined as the expected fraction of erroneous selections among all selections:

$$\mathrm{FSR}^{\mathcal{C}'} = \mathbb{E}\left[\frac{\sum_{j=1}^{m} \mathbb{I}(\hat{Y}_{n+j} \in \mathcal{C}', \hat{Y}_{n+j} \neq Y_{n+j})}{\left\{\sum_{j=1}^{m} \mathbb{I}(\hat{Y}_{n+j} \in \mathcal{C}')\right\} \vee 1}\right]. \tag{22}$$

The group-wise FSRs taking into account the protected attribute A can be defined analogously to (6) by restricting the selections to specific groups. The EPI (5), which characterizes the power of the selection procedure, remains the same. The development of the R-values and corresponding fairness algorithms is more complicated and will be left for future research.

Acknowledgments

Bradley Rava was supported in part by the NSF GRFP fellowship. The research of Wenguang Sun was supported in part by NSF grant DMS-2015339. We thank Prof. Matteo Sesia for the reference of Bates et al. (2021) and his insightful comments on both theory and methodology.

References

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, **23**, 77–91.

Barber, R. F. and Candès, E. J. (2015) Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, **43**, 2055–2085.

Barocas, S., Hardt, M. and Narayanan, A. (2017) Fairness in machine learning. *Nips tutorial*, **1**, 2.

Bates, S., Candès, E., Lei, L., Romano, Y. and Sesia, M. (2021) Testing for outliers with conformal p-values. *arXiv preprint arXiv:2104.08279*.

Benjamini, Y. (2010) Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal*, **52**, 708–721.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.

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

Cai, T. T. and Sun, W. (2009) Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.*, **104**, 1467–1481.

Cai, T. T., Sun, W. and Wang, W. (2019) CARS: Covariate assisted ranking and screening for large-scale two-sample inference (with discussion). *J. Roy. Statist. Soc. B*, **81**, 187–234.

Cao, H., Sun, W. and Kosorok, M. R. (2013) The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing. *Biometrika*, **100**, 495–502.

Chouldechova, A. (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, **5**, 153–163.

Corbett-Davies, S. and Goel, S. (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Crisp, R. (2003) Equality, priority, and compassion. *Ethics*, **113**, 745–763.

Dieterich, W., Mendoza, C. and Brennan, T. (2016) Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*.

Du, L., Guo, X., Sun, W. and Zou, C. (2021) False discovery rate control under general dependence by symmetrized data aggregation. *Journal of the American Statistical Association,*, **to appear**, arXiv:2002.11992.

Dua, D. and Graff, C. (2017) UCI machine learning repository. URL: http://archive.ics.uci.edu/ml.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2012) Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Friedler, S. A., Scheidegger, C. and Venkatasubramanian, S. (2021) The (im) possibility of fairness: different value systems require different mechanisms for fair decision making. *Communications of the ACM*, **64**, 136–143.

Guan, L. and Tibshirani, R. (2021) Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **to appear**, arXiv:1905.04396.

Hardt, M., Price, E. and Srebro, N. (2016) Equality of opportunity in supervised learning. *Advances in neural information processing systems*, **29**, 3315–3323.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer series in statistics. Springer.

Herbei, R. and Wegkamp, M. H. (2006) Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 709–721.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An Introduction to Statistical Learning: with Applications in R.* Springer.

Kemmler, M., Rodner, E., Wacker, E.-S. and Denzler, J. (2013) One-class classification with gaussian processes. *Pattern recognition*, **46**, 3507–3518.

Khan, S. S. and Madden, M. G. (2009) A survey of recent trends in one class classification. In *Irish conference on artificial intelligence and cognitive science*, 188–197. Springer.

Kleinberg, J., Mullainathan, S. and Raghavan, M. (2016) Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Lei, J. (2014) Classification with confidence. *Biometrika*, **101**, 755–769.

Meinshausen, N. and Rice, J. (2006) Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, **34**, 373–393.

Moya, M. M. and Hush, D. R. (1996) Network constraints and multi-objective optimization for one-class classification. *Neural networks*, **9**, 463–474.

Romano, Y., Barber, R. F., Sabatti, C. and Candès, E. (2020a) With malice toward none: Assessing uncertainty via equalized coverage. Https://hdsr.mitpress.mit.edu/pub/qedrwcz3.

Romano, Y., Bates, S. and Candès, E. J. (2020b) Achieving equalized odds by resampling sensitive attributes. In *Advances in Neural Information Processing Systems 33 (NIPS 2020)*. Curran Associates, Inc. To appear.

Storey, J. D. (2002) A direct approach to false discovery rates. *J. Roy. Statist. Soc. B*, **64**, 479–498.

— (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.*, **31**, 2013–2035.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Statist. Soc. B*, **66**, 187–205.

Sun, W. and Wei, Z. (2011) Large-scale multiple testing for pattern identification, with applications to time-course microarray experiments. *J. Amer. Statist. Assoc.*, **106**, 73–88.

Supplementary Material for "A Burden Shared is a Burden Halved: A Fairness-Adjusted Approach to Binary Classification"

This supplement provides additional technical results (Section A), proofs of all theorems (Sections B-C) and additional numerical results (Section D).

A Supplementary Technical Results

A.1 The conversion algorithm and connections to Storey's q-value

This section presents a general strategy for converting an arbitrary score $S^c(x,a)$ to a fair score $R^c(S^c)$, which is referred to as the theoretical R-value. Although the discussion is purely theoretical, it shows that for every base score there always exists a corresponding fair score. The algorithm can be viewed as a method of calibration by group.

Consider random mixture model (15). The conversion algorithm consists of three steps. In Step 1, we find the distributional information with respect to the score S^c. Let $G^c(s) = \sum_{a\in\mathcal{A}} \pi_a G_a^c(s)$ be the CDF of S^c, where $G_a^c(s) = \pi_{1|a} G_{1,a}^c(s) + \pi_{2|a} G_{2,a}^c(s)$. $G_{1,a}^c(s)$ and $G_{2,a}^c(s)$ are the conditional CDFs of S^c given $A = a$ and Y. Suppose an oracle knows the conditional probabilities and conditional CDFs defined above. In Step 2, we compute the conditional error probability when the threshold is t:

$$Q_a^c(t) := \mathbb{P}(Y \neq c | S^c \geq t, A = a) = \begin{cases} \frac{\pi_{2|a}\left\{1 - G_{2,a}^1(t)\right\}}{1 - G_a^1(t)}, & c = 1 \\ \frac{\pi_{1|a}\left\{1 - G_{1,a}^2(t)\right\}}{1 - G_a^2(t)}, & c = 2. \end{cases}$$

Finally, in Step 3 we compute a fair score, referred to the theoretical R-value, for an individual from group a with $S^c = s^c$: $R^c(s^c) = \inf_{t \leq s^c} \left\{Q_a^c(t)\right\}$.

The theoretical R-value is closely related to the q-value Storey (2003), which provides a useful tool in large-scale testing problems due to its intuitive interpretation and easy operation. Suppose we are interested in testing hypotheses $\{H_j : 1 \leq j \leq m\}$ with associated p-values $\{p_j : 1 \leq j \leq m\}$. Let π be the proportion of non-nulls and $G(t)$ the alternative distribution of p-values. Define the q-value for hypothesis H_j

$$q_j = \inf_{t \geq p_j} \left\{ \text{pFDR}(t) := \frac{(1-\pi)t}{(1-\pi)t + \pi G(t)} \right\}. \tag{A.1}$$

Roughly speaking, q_j measures the fraction of false discoveries when H_j is just rejected. The operations of the q-value and R-value algorithms are similar. Specifically, an FDR/FSR analysis at level α can be carried out in two steps: first obtain the q-value/R-value for hypothesis/individual j, and then reject/select hypothesis/individual j if its q-value/r-value is less than α.

A.2 FSR and mFSR

It can be shown that, under the random mixture model (15), $\text{mFSR}^{c,a} = \mathbb{P}(Y \neq c | \hat{Y} = c, A = a)$, the conditional probability needed in the sufficiency principle (17); see Storey (2003) for a similar result in the FDR literature. Following Cai et al. (2019), we can similarly show that under mild conditions, $\text{FSR}_a^c = \text{mFSR}_a^c + o(1)$ when $m_a := |\{1 \leq j \leq m : A_{n+j} = a\}| \to \infty$. This connection shows that, under the simplifying assumptions we have made, the two criteria, namely group-wise FSR control (6) and sufficiency principle (17), are asymptotically equivalent.

A.3 R-value is the (BH) q-value of conformal p-values

In this section, we take a multiple testing point of view to provide further insights on the R-value. We show that the R-value, which was motivated from a very different perspective, can be derived as the q-value of the *conformal p-values* (Bates et al., 2021) under the one-class classification setting. To make different ideas comparable, we drop the sensitive attribute A and focus on the unadjusted R-value defined in (13).

The conformal p-value (Bates et al., 2021) is developed under a different context for nonparametric outlier detection, which is also referred to as *one-class classification* in machine learning (Moya and Hush, 1996; Khan and Madden, 2009; Kemmler et al., 2013). The basic setup is that one collects a training data set containing only the subjects of one class. Without loss of generality we label the class as "1" and denote the data set as $\mathcal{D}^1$. The goal is to detect "unusual" subjects in $\mathcal{D}^{test}$, which can be a mixture of observations from class 1 and some other classes. The unusual subjects, referred to as *outliers*, are defined as subjects which exhibit differential characteristics from the *inliers*, i.e. subjects of class "1" in the "pure" training data set.

The one-class classification problem can be formulated under the selective inference framework discussed in Section 2.1. Suppose we observe data from two classes and divide the observed data into two subsets

$$\mathcal{D} = \{(X_i, Y_i) : 1 \le i \le n\} = \mathcal{D}^1 \cup \mathcal{D}^2,$$

with $\mathcal{D}^c = \{X_i : \text{subject } i \text{ is observed with label } Y = c\}$, $c = 1, 2$. Imagine that we simply discard $\mathcal{D}^2$ and focus on a one-class classification problem, where the goal is to detect outliers in $\mathcal{D}^{test} = \{X_{n+j} : 1 \le j \le m\}$, which can be a mixture of observations from both classes. Viewing individuals in class "1" as the null cases, we can formulate an equivalent multiple testing problem:

$$H_{j0} : \; Y_{n+j} = 1 \quad vs. \quad H_{j1} : \; Y_{n+j} \neq 1 \; (i.e. \; Y_{n+j} = 2), \quad j = 1, \cdots, m.$$

It is important to note that the practice of discarding $\mathcal{D}^2$ leads to information loss. We revisit this point at the end of the section.

The construction of conformal p-values involves a sample splitting step, which divides the observed data $\mathcal{D}^1$ into two parts: $\mathcal{D}^{train}$ for training a score function $\hat{S}$ and $\mathcal{D}^{cal}$ for calibrating a significance index. We view $\hat{S}(X)$ as a *conformity score*, which measures how different X is from the inliers. The conformal p-value for testing H_{j0}, under our notational system[3], corresponds to

$$\hat{p}(t) = \frac{\sum_{i \in D^{cal}} \mathbb{I}\{\hat{S}_i^c \ge t)\} + 1}{n^{cal} + 1}, \tag{A.2}$$

where c is taken to be 2 and t is the observed score $\hat{S}_{n+j}^c = t$.

To see the connection of our R-value to the conformal p-value (A.2), recall Storey's q-value (Storey, 2002):

$$\hat{q}^{ST}\{\hat{p}(t)\} = (1 - \pi)\hat{p}(t)/G\{\hat{p}(t)\},$$

where π is the proportion of non-null cases in $\mathcal{D}^{test}$ and $G(\cdot)$ is the cumulative distribution function (CDF) of the p-values. Now let $\hat{G}(\cdot)$ denote the empirical process of the scores $\{\hat{S}_i : i \in \mathcal{D}^{test}\}$:

$$\hat{G}(t) = \frac{1}{n_{test}} \sum_{i \in \mathcal{D}^{test}} \mathbb{I}\{\hat{p}(S_i) \le \hat{p}(t)\} = \frac{1}{n_{test}} \sum_{i \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}_i^c \ge t\right), \tag{A.3}$$

[3]We mention a minor point to avoid confusions. Under our setup, a larger score indicates a greater likelihood of being an outlier. This interpretation makes sense in our problem but is in the opposite direction compared to that in Bates et al. (2021). To make the two definitions equivalent, the expression "$S \le t$" in the conformal p-value definition in Bates et al. (2021) has been swapped to "$S \ge t$" in our equation (A.2).

where the last equality holds because, by (A.2), a larger score corresponds to a smaller conformal p-value. Next we consider a modification of Storey's q-value, referred to as the BH q-value, which ignores the $(1-\pi)$ term and substitutes $\hat{G}$ in place of G in Storey's q-value:

$$\hat{q}^{BH}(t) = \frac{\hat{p}_j(t)}{\hat{G}(t)}. \tag{A.4}$$

The superscript "BH" is used because the thresholding rule

$$\hat{\boldsymbol{Y}} = \left[\mathbb{I}\left\{\hat{q}^{BH}(\hat{S}_{n+j}) \leq \alpha\right\} : 1 \leq j \leq m\right]$$

is equivalent to applying the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to the conformal p-values $\{\hat{p}(\hat{S}^c_{n+j}) : 1 \leq j \leq m\}$. Combining (A.2), (A.3) and (A.4), we can precisely recover the R-value defined in (13). Concretely, we have

$$\begin{aligned} \hat{q}^{BH}(t) &= \frac{m}{n^{cal}+1} \cdot \frac{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}^c_i \geq t\right) + 1}{\sum_{i\in\mathcal{D}^{test}} \mathbb{I}\left(\hat{S}^c_i \geq t\right)} & \text{(A.5)} \\ &= \frac{m}{n^{cal}+1} \cdot \frac{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}^c_i \geq t, Y_i \neq c\right) + 1}{\sum_{i\in\mathcal{D}^{test}} \mathbb{I}\left(\hat{S}^c_i \geq t\right)}. & \text{(A.6)} \end{aligned}$$

The last equality (A.6) holds because under the one-class classification setup, $\mathcal{D}^{cal}$ is a "pure" training set in which all observations satisfy $Y_i \neq c$ trivially. We conclude that our unadjusted R-values (13), which can be called the *conformal q-value* under the one-class classification setup, is the BH q-value of conformal p-values.

Finally we point out that this fundamental connection only holds under the one-class classification setup. The BH q-value (A.5) will be different from the R-value (13) under the binary classification setup. Specifically, the cardinalities of the calibration sets will be different under the two setups, and the equality (A.6) does not hold. The conformal p-value approach by Bates et al. (2021) can still be applied for selective inference under the binary classification setup if the goal is only to detect cases from class 2. However, the conformal p-value method utilizes fewer data points than our R-value approach since the data set $\mathcal{D}^2$ has been discarded. This may lead to loss of information and decreased power. Related issues have gone beyond the scope of this work and will be pursued in future research.

B Proof of Theorem 1

B.1 Proof of Part (a)

B.1.1 An empirical process description of the FASI algorithm

Suppose we select subjects into class c if the base score S^c is great than t. The estimated false discovery proportion (FSP), as a function of t, in group a is given by:

$$\hat{Q}_c(t) = \frac{\frac{1}{n^{cal}_a+1}\left\{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(\hat{S}^c_i \geq t, Y_i \neq c, A_i = a) + 1\right\}}{\frac{1}{m_a}\left\{\sum_{(n+j)\in\mathcal{D}^{test}} \mathbb{I}(\hat{S}^c_{n+j} \geq t, A_j = a)\right\} \vee 1}. \tag{B.1}$$

We choose the smallest t such that the estimated FSP is less than α. Define

$$\tau = \hat{Q}_c^{-1}(\alpha) = \inf\left\{t : \hat{Q}_c(t) \leq \alpha\right\}. \tag{B.2}$$

Consider the R-value defined in (10). For $(n+j)^{th}$ observation in $\mathcal{D}^{test}$, it is easy to see that $\hat{R}^c_{n+j} = \inf_{t\leq\hat{s}}\left\{\hat{Q}_c(t)\right\}$, where $\hat{s} := \hat{S}^c(X_{n+j} = x, A_{n+j} = a)$. The FASI algorithm can be represented in two equivalent ways:

$$\mathbb{I}(\hat{R}^c_{n+j} \leq \alpha) \quad \Longleftrightarrow \quad \mathbb{I}(\hat{S}^c_{n+j} \leq \tau). \tag{B.3}$$

Next we turn to the description of the true FSP process of the FASI algorithm (B.3) via the representation via $\hat{S}^c$. Let

$$\begin{aligned} V^{test}(t) &= \sum_{j\in\mathcal{D}^{test}} \mathbb{I}(\hat{S}^c_j \geq t, Y_j \neq c, A_j = a), \quad \text{and} \\ R^{test}(t) &= \sum_{j\in\mathcal{D}^{test}} \mathbb{I}(\hat{S}^c_j \geq t, A_j = a) \end{aligned}$$

be the number of false selections and total selections in $\mathcal{D}^{test}$ when the threshold is t. Furthermore, denote $V^{cal}(t) = \sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(\hat{S}^c_i \geq t, Y_i \neq c, A_i = a)$ and $R^{cal}(t) = \sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(\hat{S}^c_i \geq t, A_i = a)$ the corresponding quantities in $\mathcal{D}^{cal}$. The FSP of the proposed FASI algorithm is given by

$$\text{FSP}^{\{c\}}_a(\tau) = \frac{V^{test}(\tau)}{R^{test}(\tau) \vee 1}.$$

The operation of the FASI algorithm implies that

$$\begin{aligned} \text{FSP}^{\{c\}}_a(\tau) &= \frac{V^{test}(\tau)}{V^{cal}(\tau)+1} \cdot \frac{V^{cal}(\tau)+1}{R^{test}(\tau) \vee 1} \\ &= \hat{Q}^c(\tau) \cdot \frac{n^{cal}_a + 1}{m_a} \cdot \frac{V^{test}(\tau)}{V^{cal}(\tau)+1} \\ &\leq \alpha \cdot \frac{n^{cal}_a + 1}{m_a} \cdot \frac{V^{test}(\tau)}{V^{cal}(\tau)+1}, \end{aligned}$$

where the last two steps utilize definitions (B.1) and (B.2), respectively.

B.1.2 Martingale arguments

A key step to establish the FSR control, i.e. $\mathbb{E}\left\{\text{FSP}^{\{c\}}_a(\tau)\right\} \leq \alpha$, is to show that the ratio

$$\frac{V_{test}(t)}{V^{cal}(t)+1} \tag{B.4}$$

is a martingale. Suppose that both the calibration and test data (without labels) have been given. It is natural to consider the following filtration that involves two parallel processes: $\mathcal{F}_t = \left\{\sigma\left(V^{test}(s), V^{cal}(s)\right)\right\}_{t_l\leq s\leq t}$, where t_l is lower limit of the threshold. If t_l is used, then all subjects are classified to class c.

In our proof, we focus on the following discrete-time filtration that describes the misclassification process:

$$\mathcal{F}_k = \left\{\sigma\left(V^{test}(s_k), V^{cal}(s_k)\right)\right\}_{t_l\leq s_k\leq t},$$

where s_k corresponds to the threshold (time) when exactly k subjects, combining the subjects in both $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$, are mistakenly classified as $Y = c$. Note that at time s_k, only one of the two following events are possible

$$
\begin{aligned}
A_1 &= \mathbb{I}\{V^{test}(s_{k-1}) = V^{test}(s_k), \text{ and } V^{cal}(s_{k-1}) = V^{cal}(s_k) - 1\}, \\
A_2 &= \mathbb{I}\{V^{test}(s_{k-1}) = V^{test}(s_k) - 1, \text{ and } V^{cal}(s_{k-1}) = V^{cal}(s_k)\}.
\end{aligned}
$$

According to Assumption 1 which claims that $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$ are exchangeable, and the fact that FASI uses same fitted model to compute the scores, we have

$$
\mathbb{P}(A_1|\mathcal{F}_k) = \frac{V^{cal}(s_k)}{V^{test}(s_k) + V^{cal}(s_k)}; \quad \mathbb{P}(A_2|\mathcal{F}_k) = \frac{V^{test}(s_k)}{V^{test}(s_k) + V^{cal}(s_k)}.
$$

To see why the ratio defined in (B.4) is a discrete-time martingale with respect to the filtration $\mathcal{F}_k$, note that

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{V^{test}(s_{k-1})}{V^{cal}(s_{k-1}) + 1}|\mathcal{F}_k\right\} \\
&= \frac{V^{test}(s_k)}{V^{cal}(s_k)} \cdot \frac{V^{cal}(s_k)}{V^{test}(s_k) + V^{cal}(s_k)} + \frac{V^{test}(s_k) - 1}{V^{cal}(s_k) + 1} \cdot \frac{V^{test}(s_k)}{V^{test}(s_k) + V^{cal}(s_k)} \\
&= \frac{V^{test}(s_k)}{V^{cal}(s_k) + 1},
\end{aligned}
$$

establishing the desired result.

B.1.3 FSR Control

The threshold τ defined by (B.2) is a stopping time with respect to the filtration $\mathcal{F}_k$ since $\{\tau \leq s_k\} \in \mathcal{F}_k$. In other words, the event whether the kth misclassification occurs completely depends on the information prior to time s_k (including s_k).

Let $\mathcal{D}^{test,0}/\mathcal{D}^{cal,0}$ be the index sets for subjects in the testing/calibration data that do not belong to class c. The following facts/strategies will be used in the proof.

(a). When t_l is used then all subjects are classified to class c.

(b). The sizes of the testing and calibration sets are random. The expectation is taken in steps, first conditional on fixed sample sizes and then taken over all possible sample sizes.

(c). The testing data and calibration data are independent (Assumption 1).

In the final step of our proof, we shall apply the optional stopping theorem to the filtration

$\{\mathcal{F}_k\}$. The group-wise FSR of the FASI algorithm is

$$\begin{aligned}
\text{FSR}_a^{\{c\}} &= \mathbb{E}\{\text{FSP}_a^{\{c\}}(\tau)\} \\
&\leq \alpha \cdot \mathbb{E}\left[\frac{|\mathcal{D}^{cal}|+1}{|\mathcal{D}^{test}|} \cdot \mathbb{E}\left\{\frac{V^{test}(\tau)}{V^{cal}(\tau)+1}|\mathcal{D}^{cal}, \mathcal{D}^{test}\right\}\right] \\
&= \alpha \cdot \mathbb{E}\left[\frac{|\mathcal{D}^{cal}|+1}{|\mathcal{D}^{test}|} \cdot \mathbb{E}\left\{\frac{V^{test}(t_l)}{V^{cal}(t_l)+1}|\mathcal{D}^{cal}, \mathcal{D}^{test}\right\}\right] \\
&= \alpha \cdot \mathbb{E}\left\{\frac{|\mathcal{D}^{cal}|+1}{|\mathcal{D}^{test}|} \cdot \frac{|\mathcal{D}^{test,0}|}{|\mathcal{D}^{cal,0}|+1}\right\} \qquad \text{(B.5)} \\
&= \alpha \cdot \mathbb{E}\left\{\frac{|\mathcal{D}^{cal}|+1}{|\mathcal{D}^{cal,0}|+1}\right\} \mathbb{E}\left\{\frac{|\mathcal{D}^{test,0}|}{|\mathcal{D}^{test}|}\right\} \qquad \text{(B.6)} \\
&\leq \alpha \cdot \mathbb{E}\left\{\frac{|\mathcal{D}^{cal}|}{|\mathcal{D}^{cal,0}|}\right\} \mathbb{E}\left\{\frac{|\mathcal{D}^{test,0}|}{|\mathcal{D}^{test}|}\right\} = \alpha.
\end{aligned}$$

To get Equation (B.5) we used fact(a) and to get Equation (B.6) we used fact (c). The two expectations in the last line must be reciprocal of each other since we assume that the distributions of testing and calibration data are identical (Assumption 1). This completes the proof.

B.2 Proof of Part (b)

The proof is more complicated as the arguments involve constructing two martingales. We follow the same organization of the proof for Part (a). Details are provided for new arguments and omitted for repeated arguments similar to those in Part (a).

B.2.1 The empirical process description

The estimated FSP in group a for a given threshold t is:

$$\hat{Q}_c(t) = \frac{\frac{1}{n_a^{cal}+1}\left\{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(\hat{S}_i^c \geq t, Y_i \neq c, A_i = a) + 1\right\}}{\frac{1}{m_a+n_a^{cal}+1}\left\{\sum_{j\in\mathcal{D}^{test}\cup\mathcal{D}^{cal}} \mathbb{I}(\hat{S}_j^c \geq t, A_j = a) + 1\right\}}.$$

Similar to (a) define $\tau = \hat{Q}_c^{-1}(\alpha) = \inf\left\{t : \hat{Q}_c(t) \leq \alpha\right\}$. Then our data-driven algorithm is given by $\mathbb{I}(\hat{S}_{n+j}^c \leq \tau)$. Define $V^{test}(t)$, $R^{test}(t)$, $V^{cal}(t)$ and $R^{cal}(t)$ as before. Following similar arguments as in (a), we have

$$\begin{aligned}
\text{FSP}_a^{\{c,*\}}(\tau) &= \frac{V^{test}(\tau)}{V^{cal}(\tau)+1} \cdot \frac{V^{cal}(\tau)+1}{R^{cal}(\tau)+R^{test}(\tau)+1} \cdot \frac{R^{cal}(\tau)+R^{test}(\tau)+1}{R^{test}(\tau)+1} \\
&\leq \alpha \cdot \frac{n_a^{cal}+1}{n_a^{cal}+m_a+1} \cdot \frac{V^{test}(\tau)}{V^{cal}(\tau)+1} \cdot \frac{R^{cal}(\tau)+R^{test}(\tau)+1}{R^{test}(\tau)+1}.
\end{aligned}$$

Next we shall show that the last two terms in the above product are both martingales.

B.2.2 Martingale arguments

In Part (a), we have shown that $V^{test}(t)/\{V^{cal}(t)+1\}$ is a discrete-time martingale with respect to the filtration $\mathcal{F}_k = \left\{\sigma\left(V^{test}(s_k), V^{cal}(s_k)\right)\right\}_{t_l \leq s_k \leq t}$, which is defined on the misclassification process. Next we show that $\{R^{cal}(t)+R^{test}(t)+1\}/\{R^{test}(t)+1\}$ is also a discrete-time martingale. Consider the filtration that describes the *selection process*:

$$\mathcal{F}_k^* = \{\sigma\left(R^{cal}(s_k^*), R^{test}(s_k^*)\right)\}_{t_l \le s_k^* \le t},$$

where s_k^* corresponds to the time when exactly k subjects are selected. At time s_k^*, only one of the following two events are possible:

$$\begin{aligned} A_1^* &= \left\{R^{cal}(s_{k-1}^*) = R^{cal}(s_k^*), R^{test}(s_{k-1}^*) = R^{test}(s_k^*) - 1\right\}; \\ A_2^* &= \left\{R^{cal}(s_{k-1}^*) = R^{cal}(s_k^*) - 1, R^{test}(s_{k-1}^*) = R^{test}(s_k^*)\right\}. \end{aligned}$$

On this backward running filtration, we have

$$P(A_1^*) = \frac{R^{test}(s_k^*)}{R^{cal}(s_k^*) + R^{test}(s_k^*)}, \quad P(A_2^*) = \frac{R^{cal}(s_k^*)}{R^{cal}(s_k^*) + R^{test}(s_k^*)}.$$

It follows that $\{R^{cal}(t) + R^{test}(t) + 1\}/\{R^{test}(t) + 1\}$ is a martingale since

$$\begin{aligned} &\mathbb{E}\left\{\frac{R^{cal}(s_{k-1}^*) + R^{test}(s_{k-1}^*) + 1}{R^{test}(s_{k-1}^*) + 1} | \mathcal{F}_k^*\right\} \\ = \; & \frac{R^{cal}(s_k^*) + R^{test}(s_k^*)}{R^{test}(s_k^*)} \cdot \frac{R^{test}(s_k^*)}{R^{cal}(s_k^*) + R^{test}(s_k^*)} + \frac{R^{cal}(s_k^*) + R^{test}(s_k^*)}{R^{test}(s_k^*) + 1} \cdot \frac{R^{cal}(s_k^*)}{R^{cal}(s_k^*) + R^{test}(s_k^*)} \\ = \; & \frac{R^{cal}(s_k^*) + R^{test}(s_k^*) + 1}{R^{test}(s_k^*) + 1}. \end{aligned}$$

B.2.3 FSR Control

Note that the threshold τ is a stopping time with respect to the filtration $\mathcal{F}_k^*$. Let $\mathcal{D}^{test,0}/\mathcal{D}^{cal,0}$ be the index sets for subjects in the testing/calibration data that do not belong to class c.

$$\begin{aligned} \text{FSR}_a^{c,*} &= \mathbb{E}\{\text{FSP}_a^{c,*}(\tau)\} \\ &\le \alpha \mathbb{E}\left[\frac{|D^{cal}| + 1}{|D^{cal}| + |D^{test}| + 1} \cdot \frac{R^{cal}(\tau) + R^{test}(\tau) + 1}{R^{test}(\tau) + 1} \cdot \mathbb{E}\left\{\frac{V^{test}(\tau)}{V^{cal}(\tau) + 1} | \mathcal{D}^{cal}, \mathcal{D}^{test}\right\}\right]. \end{aligned}$$

The term $\{R^{cal}(\tau)+R^{test}(\tau)+1\}/\{R^{test}(\tau)+1\}$ can be factored out because $\{R^{cal}(\tau)$ and $R^{test}(\tau)$ are constant when $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$ are given. According to Part (a), $\{V^{test}(t)\}/\{V^{cal}(t) + 1\}$ is a backward martingale on $\mathcal{F}_k$. When t_l is used then all subjects are classified to class c. According to the optional stopping theorem we have

$$\mathbb{E}\left\{\frac{V^{test}(\tau)}{V^{cal}(\tau) + 1} | \mathcal{D}^{cal}, \mathcal{D}^{test}\right\} = \frac{V^{test}(t_l)}{V^{cal}(t) + 1} = \frac{|\mathcal{D}^{test,0}|}{|\mathcal{D}^{cal,0}| + 1}.$$

Next, conditional on the filtration defined on the selection process, we have

$$\mathbb{E}\left\{\frac{R^{cal}(\tau) + R^{test}(\tau) + 1}{R^{test}(\tau) + 1}\right\} = \mathbb{E}\left\{\frac{R^{cal}(t_l) + R^{test}(t_l) + 1}{R^{test}(t_l) + 1}\right\} = \mathbb{E}\left\{\frac{|D^{cal}| + |D^{test}| + 1}{|D^{test}| + 1}\right\}.$$

Combining the above results, we have

$$\begin{aligned}\text{FSR}_a^{c,*} &\leq \alpha\mathbb{E}\left[\frac{|D^{cal}|+1}{|D^{cal}|+|D^{test}|+1}\cdot\frac{|D^{cal}|+|D^{test}|+1}{|D^{test}|+1}\cdot\frac{|\mathcal{D}^{test,0}|}{|\mathcal{D}^{cal,0}|+1}\right]\\ &= \alpha\cdot\mathbb{E}\left\{\frac{|\mathcal{D}^{cal}|+1}{|\mathcal{D}^{cal,0}|+1}\right\}\mathbb{E}\left\{\frac{|\mathcal{D}^{test,0}|}{|\mathcal{D}^{test}|}\right\}\\ &\leq \alpha\cdot\mathbb{E}\left\{\frac{|\mathcal{D}^{cal}|}{|\mathcal{D}^{cal,0}|}\right\}\mathbb{E}\left\{\frac{|\mathcal{D}^{test,0}|}{|\mathcal{D}^{test}|}\right\} = \alpha.\end{aligned}$$

The last two lines have used similar arguments as those in Part (a) hence we skipped the explanations. The proof is compete.

C Proof of Theorem 2

The theorem implies that the optimal base score for constructing R-values should be $S^c(x,a) = \mathbb{P}(Y = c|X = x, A = a)$. A similar optimality theory has been developed in the context of multiple testing with groups (Cai and Sun, 2009). However, the proof for the binary classification setup with the indecision option is much more complicated; we provide the proof here for completeness. We first establish an essential monotonicity property in Section C.1, then prove the optimality theory in Section C.2.

C.1 A monotonicity property

Suppose we use $S^c_{n+j}(x,a) = \mathbb{P}(Y_{n+j} = c|X_{n+j} = x, A_{n+j} = a)$ as the base score. The corresponding theoretical R-values can be obtained via the conversion algorithm in Appendix A.1. Under Model 15, the mFSR level with threshold t is $\text{mFSR}_a^{\{c\}}(t) = \mathbb{P}(Y \neq c|S^c \geq t, A = a)$. The theoretical R-values is defined as $R^c(s^c) = \inf_{t\leq s^c}\left\{\text{mFSR}_a^{\{c\}}(t)\right\}$. Let $Q_a^c(t)$ be the mFSR level when the threshold is t. The next proposition characterizes the monotonic relationship between $Q_a^c(t)$ and t.

Proposition 2. *$Q_a^c(t)$ is monotonically decreasing in t.*

The proposition is essential for expressing the oracle procedure as a thresholding rule based on S^c. Specifically, denote $Q_a^{c,-1}(\cdot)$ the inverse of $Q_a^c(\cdot)$. The monotonicity of $Q_a^c(t)$ and the definition of the theoretical R-value together imply that $S_j^c(x,a) = Q_a^{c,-1}(R_j^c)$ for $a \in \mathcal{A}$. For notational convenience, let $T_{n+j}(x,a) = \mathbb{P}(Y_{n+j} = 2|X_{n+j} = x, A_{n+j} = a)$. Then $S^1_{n+j} = 1 - T_{n+j}$ and $S^2_{n+j} = T_{n+j}$. Therefore the oracle rule

$$\delta_{OR}^{n+j} = \mathbb{I}(R^1_{n+j} \leq \alpha_1) + 2\mathbb{I}(R^2_{n+j} \leq \alpha_2).$$

can be equivalently written as

$$\begin{aligned}\delta_{OR}^{n+j} &= \mathbb{I}\left\{S^1_{n+j} \geq Q^{-1}_{1,a}(\alpha_1)\right\} + 2\mathbb{I}\left\{S^2_{n+j} \geq Q^{-1}_{2,a}(\alpha_2)\right\}\\ &= \mathbb{I}\left\{T_{n+j} \leq 1 - Q^{-1}_{1,a}(\alpha_1)\right\} + 2\mathbb{I}\left\{T_{n+j} \geq Q^{-1}_{2,a}(\alpha_2)\right\}.\end{aligned} \quad (C.1)$$

for $1 \leq j \leq m$. This provides a key technical tool in Section C.2.

Proof of Proposition 2.

Define $\tilde{Q}_a^c(t) = 1 - Q_a^c(t)$. We only need to show that $Q_a^c(t)$ is monotonically increasing in t. Let $\mathcal{M}_a = \{n+1 \le j \le n+m : A_j = a\}$. According to the definition of the mFSR and the definition of S_j^c, we have

$$\mathbb{E}\left\{\sum_{j\in\mathcal{M}_a}\left\{S_j^c - \tilde{Q}_a^c(t)\right\}\mathbb{I}(S_j^c > t)\right\} = 0, \tag{C.2}$$

where the expectation is taken over both $\mathcal{D}^{test}$. It is important to note that the oracle procedure, which assumes that all distributional information is known, does not utilize $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$. It is easy to see from Equation (C.2) that $\tilde{Q}_a^c(t) > t$ otherwise the summation on the LHS must be positive, leading to a contradiction.

Next we show that $t_1 < t_2$ implies $\tilde{Q}_a^c(t_1) \le \tilde{Q}_a^c(t_2)$. Assume instead that $\tilde{Q}_a^c(t_1) > \tilde{Q}_a^c(t_2)$. We focus on group a, then

$$\begin{aligned}
& \sum_{j\in\mathcal{M}_a}\{S_j^c - \tilde{Q}_a^c(t_1)\}\mathbb{I}(S_j^c > t_1) \\
= & \sum_{j\in\mathcal{M}_a}\{S_j^c - \tilde{Q}_a^c(t_2) + \tilde{Q}_a^c(t_2) - \tilde{Q}_a^c(t_1)\}\mathbb{I}(S_j^c > t_1) \\
= & \sum_{j\in\mathcal{M}_a}\{S_j^c - \tilde{Q}_a^c(t_2)\}\mathbb{I}(S_j^c > t_2) + \sum_{j\in\mathcal{M}_a}\{S_j^c - \tilde{Q}_a^c(t_2)\}\mathbb{I}(t_1 \le S_j^c \le t_2) \\
& + \sum_{j\in\mathcal{M}_a}\left\{\tilde{Q}_a^c(t_2) - \tilde{Q}_a^c(t_1)\right\}\mathbb{I}(S_j^c > t_1) \\
= & I + II + III.
\end{aligned}$$

Taking expectations on both sides, it is easy to see that the LHS is zero. However, the RHS is strictly greater than zero. For term I, we have $\mathbb{E}(I) = 0$ according to the definition of mFSR. For term II, we have $\mathbb{E}(II) < 0$ as we always have $\tilde{Q}_a^c(t) > t$. For term III, we have $\mathbb{E}(III) < 0$ since we assume $\tilde{Q}_a^c(t_1) > \tilde{Q}_a^c(t_2)$. It follows that the assumption $\tilde{Q}_a^c(t_1) > \tilde{Q}_a^c(t_2)$ cannot be true, and the proposition is proved.

C.2 Proof of the theorem

Define the expected number of true selections ETS $= \sum_{j=1}^m \mathbb{I}(Y_{n+j} = c, \hat{Y}_{n+j} = c)$. Then it can be shown that minimizing the EPI subject to the FSR constraint is equivalent to maximizing the ETS subject to the same constraint.

According to Proposition 2, the oracle rule can be written as

$$\delta_{OR}^{n+j} = \mathbb{I}\left\{T_{n+j} \le 1 - Q_{1,a}^{-1}(\alpha_1)\right\} + 2\mathbb{I}\left\{T_{n+j} \ge Q_{2,a}^{-1}(\alpha_2)\right\}.$$

The mFSR constraints for the oracle rule imply that

$$\mathbb{E}\left\{\sum_{j\in\mathcal{M}_a}(T_j - \alpha_1)\mathbb{I}(\delta_{OR}^j = 1)\right\} = 0,\ \mathbb{E}\left\{\sum_{j\in\mathcal{M}_a}(1 - T_j - \alpha_2)\mathbb{I}(\delta_{OR}^j = 2)\right\} = 0. \tag{C.3}$$

Let $\boldsymbol{\delta} \in \{0,1,2\}^m$ be a general selection rule in $\mathcal{D}_{\alpha_1,\alpha_2}$. Then the mFSR constraints for $\boldsymbol{\delta}$ implies that

$$\mathbb{E}\left\{\sum_{j\in\mathcal{M}_a}(T_j - \alpha_1)\mathbb{I}(\delta_j = 1)\right\} \le 0,\ \mathbb{E}\left\{\sum_{j\in\mathcal{M}_a}(1 - T_j - \alpha_2)\mathbb{I}(\delta_j = 2)\right\} \le 0. \tag{C.4}$$

The ETS of $\boldsymbol{\delta} = (\delta_j : n+1 \le j \le n+m)$ is given by

$$\begin{aligned} \text{ETS}_{\boldsymbol{\delta}} &= \mathbb{E}\left[\sum_{a\in\mathcal{A}}\sum_{j\in\mathcal{M}_a}\{\mathbb{I}(\delta_j=1)(1-T_j)+\mathbb{I}(\delta_j=2)T_j\}\right] \\ &= \sum_{a\in\mathcal{A}}\text{ETS}_{\boldsymbol{\delta}}^{1,a}+\text{ETS}_{\boldsymbol{\delta}}^{2,a}. \end{aligned}$$

The goal is to show that $\text{ETS}(\boldsymbol{\delta}_{OR}) \ge \text{ETS}(\boldsymbol{\delta})$. We only need to show $\text{ETS}_{\boldsymbol{\delta}_{OR}}^{c,a} \ge \text{ETS}_{\boldsymbol{\delta}}^{c,a}$ for all c and a. We will show $\text{ETS}_{\boldsymbol{\delta}_{OR}}^{1,a} \ge \text{ETS}_{\boldsymbol{\delta}}^{1,a}$ for a given a. The remaining inequalities follow similar arguments.

According to (C.3) and (C.4), we have

$$\mathbb{E}\left[\sum_{j\in\mathcal{M}_a}(T_j-\alpha_1)\left\{\mathbb{I}(\delta_{OR}^j=1)-\mathbb{I}(\delta_j=1)\right\}\right]\ge 0. \tag{C.5}$$

Let $\lambda_{1,a} = (1-Q_{1,a}^{-1}(\alpha_1)-\alpha_1)/Q_{1,a}^{-1}(\alpha_1)$. It can be shown that $\lambda_{1,a} > 0$. For $j \in \mathcal{M}_a$, we claim that the oracle rule can be equivalently written as

$$\delta_{OR}^j = \mathbb{I}\left\{\frac{T_j-\alpha_1}{1-T_j}<\lambda_{1,a}\right\}.$$

Using the previous expression and techniques similar to the Neyman-Pearson lemma, we claim that the following result holds for all $j \in \mathcal{M}_a$:

$$\left\{\mathbb{I}(\delta_{OR}^j=1)-\mathbb{I}(\delta_j=1)\right\}\{T_j-\alpha_1-\lambda_{1,a}(1-T_j)\}\le 0.$$

It follows that

$$\mathbb{E}\left[\sum_{j\in\mathcal{M}_a}\left\{\mathbb{I}(\delta_{OR}^j=1)-\mathbb{I}(\delta_j=1)\right\}\{T_j-\alpha_1-\lambda_{1,a}(1-T_j)\}\right]\le 0. \tag{C.6}$$

According to (C.5) and (C.6), we have

$$\lambda_{OR}\mathbb{E}\sum_{j\in\mathcal{M}_a}(1-T_j)\left\{\mathbb{I}(\delta_{OR}^j=1)-\mathbb{I}(\delta_j=1)\right\}=\lambda_{OR}\left(\text{ETS}_{\boldsymbol{\delta}_{OR}}^{1,a}-\text{ETS}_{\boldsymbol{\delta}}^{1,a}\right)\ge 0.$$

Note that $\lambda_{OR} > 0$, the desired result follows.

D Additional Numerical Results

In this section, we demonstrate through simulation that the R^+-value (11) is more stable than the R-value (10) when $|\mathcal{D}^{test}|$ is small. To do this, we will look at the distributions of R-value and R^+-value for a fixed base score of $s(x,a) = 0.9$.

We consider the setting described in Section 5 with $F_{1,M} = F_{1,F} = \mathcal{N}(\boldsymbol{\mu}_1, 2\cdot\mathbf{I}_3)$ and $F_{2,M} = F_{2,F} = \mathcal{N}(\boldsymbol{\mu}_2, 2\cdot\mathbf{I}_3)$. We set $\pi_{2|F} = \pi_{2|M} = 0.8$, $\boldsymbol{\mu}_1 = (1,1,1)^\top$ and $\boldsymbol{\mu}_2 = (2,2,2)^\top$. The base scores are constructed as the oracle class probabilities $P(Y=c|X,A)$.

In Figure 7, we compute 1,000 R-values and R^+-values for a fixed score of $s = 0.9$ based on randomly generated $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$. The size of the calibration set is fixed at $|\mathcal{D}^{cal}| = 1,000$ and the

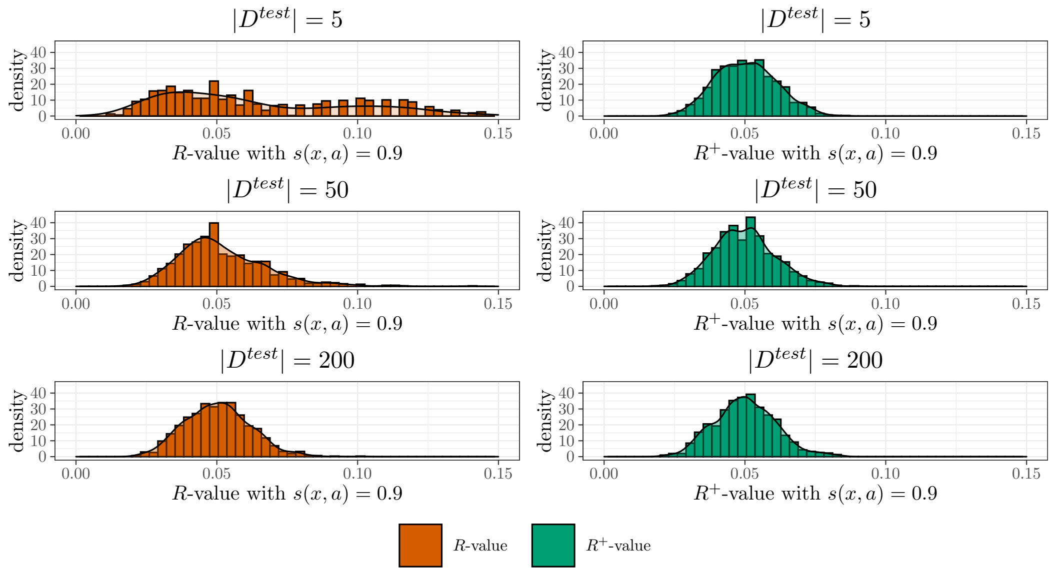

Figure 7: The comparison between the R-value and R^+-value for varying sizes of the test data set. The left column shows the histograms of the R-value (orange) and the right column shows the histograms of the R^+-value (green). The R-values and R^+-values are computed for a fixed base score of $s(x,a) = 0.9$ based on $1,000$ randomly generated data sets.

test set has sizes $|\mathcal{D}^{test}| \in \{5, 50, 200\}$. The columns of Figure 7 show the histograms the R-values (left) and R^+-values (right) with $\mathcal{D}^{test}$ increasing from 5 (first row) to 200 (last row).

When $|\mathcal{D}^{test}| = 5$, we notice that the R-value has much more variability than the R^+-value. This is because the denominator of the R-value only utilizes 5 observations when computing the total number of selections. By contrast, the R^+-value uses $1,005$ observations since it has access to data from both $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$. Moving further down the rows of Figure 7, the advantage of the R^+-value slowly disappears as $|\mathcal{D}^{test}|$ increases. This causes the variability of both R-value and R^+-value to become almost identical. We conclude from this simulation that the R^+-value is more desirable in settings where $|\mathcal{D}^{test}|$ is small since it can use more data to decrease its variability. However, while the R-value has more variability for small $|\mathcal{D}^{test}|$, this disadvantage can be quickly overcome through the introduction of a reasonably sized test set.