Fair learning with bagging

Jean-David Fermanian, Dominique Guegan

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers

Fair learning with bagging

Jean-David FERMANIAN, Dominique GUEGAN

2021.34

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
https://centredeconomiesorbonne.cnrs.fr/
ISSN : 1955-611X

Fair learning with bagging.

Jean-David Fermanian*and Dominique Guégan†

November 22, 2021

Abstract

The central question of this paper is how to enhance supervised learning algorithms with fairness requirement ensuring that any sensitive input does not "'unfairly"' influence the outcome of the learning algorithm. To attain this objective we proceed by three steps. First after introducing several notions of fairness in a uniform approach, we introduce a more general notion through conditional fairness definition which englobes most of the well known fairness definitions. Second we use a ensemble of binary and continuous classifiers to get an optimal solution for a fair predictive outcome using a related-post-processing procedure without any transformation on the data, nor on the training algorithms. Finally we introduce several tests to verify the fairness of the predictions. Some empirics are provided to illustrate our approach.

Keywords: fairness; nonparametric regression; classification; accuracy.

JEL classification: C10, C38, C53.

1 Introduction

Machine learning is increasingly used to take decision that can severely affect people's life in education, living, working, lending or criminal risk assessment, for instance. This phenomena has been accompanied by an increase in concern about disparate treatment caused

*Ensae-Crest, 5 Avenue Le Chatelier, 91120 Palaiseau, France. Email: jean-david.fermanian@ensae.fr

†Université Paris 1 Panthéon Sorbonne, France and University Ca'Foscari, Venezia, Italy. Email: dguegan@univ-paris1.fr

by models errors and bias in the data. The potential for unfairness in algorithmic decisions being growing, an important literature around notions of fairness was recently largely developed inspired by the concept of discrimination in social sciences and law. Thus, to ensure that learned classifiers are non discriminatory or fair with respect to some sensitive feature is a crucial question. To study this subject we need first an agreement on the notion of fairness, second the necessity to study it with an holistic approach, considering at the same time the features (input data) and the algorithm used. The concept of bias in machine learning is not new, and has been investigated from different points of view, sometimes focusing on data introducing different concepts like statistical bias, demographic parity, equalized odds, unbalanced sets, disparate parity, opportunity parity, etc., Madras et al. (2018), sometimes focusing on algorithms considering their robustness, accuracy, interpretability, Lundberg and Lee (2017), sometimes on both with the adversarial, counterfactual or agnostic approaches, for a review Bogroff and Guégan (2019).

In this paper we investigate the notion of fairness associated to a specific machine learning modelling with respect to the data set used in order to provide fair predictions. We focus on an ensemble of classifiers, called random forest, Breiman (2001), and analyse the existence of an optimal solution depending on the notion of fairness we are interested in. Ensemble learning algorithms use the same base classifier to produce repeated multiple classifications of the same data, Kotsiantis and Pitelas (2001). In order to attain our objective, we provide some classification of the different definitions proposed in the literature introducing a more general concept with the notion of conditional fairness which can encompass most of the classical fairness notions. A novel testing procedure is associated to the learning procedure we consider in order to test the fairness of the predictions obtaiend through the classifiers. Our approach is related to a post processing procedure avoiding the modification of the data representation and the modification of the trained classifier to achieve desired outputs.

The fair machine learning literature has tended to focus on domains such as recidivism prediction, automated hiring, and face recognition, where fairness can be understood, at least partially, in terms of well-defined quantitative metrics. It has recently been shown that algo-

rithms trained with biased data have resulted in algorithmic discrimination. Indeed, recent analyzes have questioned the statistical methods used in the US judicial system, pointing to the bias against African-American accused, considering that African-American accused were more likely to be wrongly labeled as higher risk of recidivism. Thus, a fairly extensive literature has proposed different criteria to avoid these biases. All these works has been enriched by research on discrimination involving gender, then race, color, religion, disability, family status, children health, facial recognition. For instance, Buolamini and Gebru (2018) analyze pictures to evaluate how bias can be present in automated facial analysis algorithms and data sets with respect to phenotypic subgroups. Chouldechova et al. (2018) investigate the decision of children's placement in case of abuse based on administrative information and how the choice of the information can impact the decision.

Significant effort in the fair machine learning community has focused on the development of statistical definitions of fairness, Hardt et al. (2016) and Berk et aL. (2018), and algorithmic methods, Agarwal (2018) and Kusner et al. (2017), to assess and mitigate biases in relation to these definitions, Zafar et al. (2017). The interest to study fairness in classification is to be able to make prediction 'responsably'. A model should predict only if its predictions are reliable aligned with the system objectives which often include accuracy (predictions should mostly indicate ground truth) and fairness (*predictions should be unbiased with respect of different subgroups*).

The first notion of fairness which was introduced is *statistical parity*, called also *group fairness* or *demographic parity* which equalizes outcomes across protected and non protected groups. The marginal distributions of the predicted class are the same for both protected group categories. Demographic parity requires that a decision is independent of a protected attribute. In other words, membership in a protected class should have no correlation with the decision. From the point of view of an individual the outcome can be unfair. Indeed, this approach can create highly undesirable decision: for instance if the protected attribute is gender, one might incarcerate women who pose no public safety risk so that the same proportions of men and women are released on probation. This problem has been illustrated in Dwork et al. (2012).

In order to avoid the limitation of the previous definition, another approach based on the

use of the joint distribution of the output and the protected attribute provides a decision that does not discriminate with respect to the protected attribute. This approach is called *equalized odds* with respect to a protected attribute : the predictor and the protected attribute are independent conditionally on the output, Hardt et al. (2016). This notion is also called *conditional procedure accuracy equality*, Berk et al. (2018). A particular case of the *equalized odds* notion is the *equal opportunity*, when the value of the outcome is specified. This notion is used in particular when we focus on a decision concerning an 'advantaged' outcome, like for instance 'not defaulting to a loan' or 'admission to a college', etc. It is a weaker notion of discrimination, but still interesting. A close notion is also discussed in Kleinberg et al. (2016): the authors introduce the notion of *well calibration* or *calibration within each group* which corresponds to *equality of opportunity* for the output labelled positive.

An unfairness metric has been introduced by Zafaz et al. (2017) called *disparate mistreatment* which is defined in terms of mis-classification rates. Disparate mistreatment seems especially well-suited for scenarii where ground truth is available for historical decisions used during the training phase. The authors call a decision making process to be suffering from disparate mistreatment with respect to a given sensitive attribute (e.g., race) if the mis-classification rates differ for groups of people having different values of that sensitive attribute (e.g., black or white).

To avoid discrimination, other approaches have been used prohibiting the use of certain kinds of features, not based on the simple elimination of sensitive features which can be insufficient for avoiding inappropriate determination processes, due to the indirect influence of sensitive information. Feldman et al (2015) introduce the notion of *indirect impact*, called *statistical discrimination* in economics. The notion is also close to the *indirect prejudice* definition introduced in Kamishima et al. (2012) and to the concept of *group-based fairness* developed in Zemel et al. (2013). This approach develops a procedure that predicts the protected attribute from the other features, and is totally different of the approaches introduced previously.

Thus, having fixed the notion of fairness developed in the literature, in order to get fair

predictions several approaches have been considered. We can distinguish mainly two ways. The pre-process training ensures fairness of any learned model eliminating any sources of unfairness in the data before the algorithm is formulated. A major problem with this approach is that interaction effects (e.g., with race and gender) containing information leading to unfairness are not removed unless they are explicitly included in the residualizing regression even if all of the additive contaminants are removed. In short, all interactions effects, even higher order ones, would need to be anticipated. Rebalancing the information set is another possibility, Zemel et al. (2013), Berk et al. (2018). Lu et al. (2017) use actionable plans to transform the predictions of the input to a desired output with a minimum cost. Another procedure to ensure the predictions fairness is the post-processing training. After the algorithm is applied, its performance is adjusted to make it more fair. To date, perhaps the best example of this approach draws on the idea of random reassignment of the class label previously assigned by the algorithm, Feldman et al. (2015) and Hardt et al. (2016). An optimal action extraction for random forest and boosted tree is proposed for post processing by Cui et al. (2015), see also Yang et al. (2003). A direct approach ensuring fairness by optimisation is proposed by Zafar et al. (2017) and Agarwal et al. (2018).

On the basis of the previously recalled papers our objective is threefold. In order to provide a complete procedure to quantifying the notion of fairness for applications, first we introduce the conditional fairness definition which means that the probability of missclassification does not depend on a sensitive variable for any given value of a covariate. By this way we integrate most of the classical definitions of fairness developed in the literature that we recall in an uniform way. Second we propose a fair bagging class of classifiers to obtain fair predictions considering a related-post-processing approach relying on two main ideas: we do not use pre-processing considering that any changes in the data can be sources of bias, and we do not make any change inside the trained classifier we consider to avoid also approximations. Indeed we privilege the use of replication of the classifiers through bagging. Indeed, group of classifiers performs more accurately than any single classifier, and utilizes the strengths of the individual group of classifiers - which can be different - while at the same time the classifier weaknesses are circumvented. By definition this approach can decrease

the errors.

The rest of the paper is organized as follows. In Section two we introduce the formalism to define all the fairness measures we investigate, and introduce our new proposal. In Section three, we introduce the additive tree classifier we use to train the data that we call a related-post-processing procedure. In Section four, we provide optimal solutions with this ensemble of classifiers. Section five reports numerical experiments and illustrates our approach on some real data sets. Section six concludes.

2 Definitions of fairness: the role of input variables

Let X be a vector of individual characteristics (exogenous variables) in a finite dimensional space $\mathcal{X}$. We want to explain/predict the univariate random variable Y, that can be discrete or continuous. Beside, a variable S is considered as sensitive: gender, ethnical background, etc. In line with the literature, the latter sensitive feature is discrete. To simplify, it will take only two values zero and one. The sensitive feature S may be a component of Xor X may contain other features that are arbitrarily indicative of S. For example, if the classification task is to predict whether or not someone will default on a loan, each training example might correspond to a person, where X represents their demographics, income level, past payment history, and loan amount; S represents their race; and Y represents whether or not they defaulted on that loan. Note that X might contain their race as one of the features or, for example, contain their zipcode - a feature that is often correlated with race. Our goal is to build a predictor that is "fair" w.r.t S, not transforming the input variables.

Strictly speaking, a predictor is a mapping $g : \mathcal{X} \to \mathbb{R}$ that seeks to predict Y (which could be a vector containing discrete or continuous components) from individual characteristics X, possibly including S. Denote by $\hat{Y}$ the g-predictions, i.e. $\hat{Y} := g(X)$ for any X. Several definitions of perfect fairness have been proposed in the literature (see Williamson and Menon, 2019), and the references therein), we formalize them now in an uniform way.

(i) *Demographic parity or statistical parity*: $\hat{Y}$ and S are statistically independent. If we assume that the random variable $\hat{Y}$ is discrete and potentially takes p values

Documents de travail du Centre d'Economie de la Sorbonne 2021.34

$0, 1, \ldots, p-1$, then, *demographic parity* is equivalent to satisfying

$$\mathbb{P}(\hat{Y} = j|S = k) = \mathbb{P}(\hat{Y} = j),\ j = 0, \ldots, p-1; k = 0, 1. \tag{2.1}$$

When $p = 2$, it can be seen this is equivalent to

$$\mathbb{P}(\hat{Y} = 1|S = 0) = \mathbb{P}(\hat{Y} = 1|S = 1),$$

and the equation (2.1) reduces to $\mathbb{E}_X[\hat{Y}|S = 0] = \mathbb{E}_X[\hat{Y}|S = 1] = \mathbb{E}_X[\hat{Y}]$ because $\hat{Y} \in \{0, 1\}$.[1] In that latter case we say also that we avoid *disparate impact*, say "if the probability that a classifier assigns a user to the positive class, i.e. $\hat{Y} = 1$, is the same for both values of the sensitive feature S, then there is no *disparate impact*", Zafar et al. (2017).

In the case of a continuous predicted variable $\hat{Y}$, *demographic parity* may be rewritten as

$$\mathbb{P}(\hat{Y} \leqslant y|S = 0) = \mathbb{P}(\hat{Y} \leqslant y|S = 1) = \mathbb{P}(\hat{Y} \leqslant y), \tag{2.2}$$

for every real number y. This implies (but is not equivalent to)

$$\mathbb{E}(\hat{Y}|S = 0) = \mathbb{E}(\hat{Y}|S = 1) = \mathbb{E}(\hat{Y}). \tag{2.3}$$

Now Y can be a vector. If we assume that Y is a bivariate vector $Y = (Y_1, Y_2)$ where Y_1 represents, for instance, the annual fixed salary and Y_2 some bonus, thus we can be interested to check whether women and men are fairly equally paid, then the equation (2.2) becomes:

$$\mathbb{P}(\hat{Y}_1 \leqslant y_1, \hat{Y}_2 \leqslant y_2|S = 0) = \mathbb{P}(\hat{Y}_1 \leqslant y_1, \hat{Y}_2 \leqslant y_2|S = 1) = \mathbb{P}(\hat{Y}_1 \leqslant y_1, \hat{Y}_2 \leqslant y_2). \tag{2.4}$$

All the following equations can be written for a multivariate vector Y. For simplicity

[1] If S represents the gender and Y the recidivism variable, one wants that the probability of recidivism of a person (for instance) is the same conditionally (only) that this person is a male or a female.

we restrict the presentation to discrete random variables Y. While simple and intuitive, demographic parity has serious conceptual limitations as a fairness notion, many of which are pointed out in Dwork et al. (2012).

(ii) *Equalized odds*: $\hat{Y}$ and S are statistically independent, given Y. Assume that the random variables Y and $\hat{Y}$ are discrete and potentially takes p values $0, 1, \ldots, p-1$. Then, *equalized odds* reduces to

$$\mathbb{P}(\hat{Y} = j | S = k, Y = l) = \mathbb{P}(\hat{Y} = j | Y = l),\ j, l = 0, \ldots, p-1; k = 0, 1. \tag{2.5}$$

When $p = 2$ ($\hat{Y}$ takes only two values), this means

$$\mathbb{E}[\hat{Y} | S = 0, Y = l] = \mathbb{E}[\hat{Y} | S = 1, Y = l] = \mathbb{E}[\hat{Y} | Y = l],\ l = 0, 1. \tag{2.6}$$

Assuming S represents the gender and Y the recidivism variable, this means one wants that the probability of recidivism of a person (for instance) is the same conditionally that this person is a male or a female *and* he/she reoffends.

In the case of continuous explained variables Y and $\hat{Y}$, equalized odds may be rewritten as

$$\mathbb{P}(\hat{Y} \leqslant y | S = 0, Y = y') = \mathbb{P}(\hat{Y} \leqslant y | S = 1, Y = y') = \mathbb{P}(\hat{Y} \leqslant y | Y = y'), \tag{2.7}$$

for every real numbers y and y'. This implies (but is not equivalent to)

$$\mathbb{E}[\hat{Y} | S = 0, Y = y'] = \mathbb{E}[\hat{Y} | S = 1, Y = y'] = \mathbb{E}[\hat{Y} | Y = y'], \tag{2.8}$$

for every $y' \in \mathbb{R}$.

A particular case of equalized odds is *equal opportunity*, Hardt et al. (2016). In that case, if the random variables Y and $\hat{Y}$ are discrete and potentially take 2 values $0, 1$, then, if we privilege that $Y = 1$ (the "advantaged" outcome, for instance "receiving a

promotion" or "not defaulting on a loan"), we have

$$\mathbb{E}[\hat{Y}|S=0,Y=1] = \mathbb{E}[\hat{Y}|S=1,Y=1] = \mathbb{E}[\hat{Y}|Y=1]. \tag{2.9}$$

(iii) In the case of discrete outcomes, the *lack of disparate mistreatment* (Zafar et al., 2017) is defined as

$$\mathbb{P}(\hat{Y} \neq Y|S=0) = \mathbb{P}(\hat{Y} \neq Y|S=1).$$

In order to insure that classifiers do not suffer from disparate mistreatment a practical approach for model calibration is based on the minimisation of a convex loss $L(\theta)$: $\min L(\theta)$ under the constraints $|\mathbb{P}(\hat{Y} \neq Y|S=0) - \mathbb{P}(\hat{Y} \neq Y|S=1)| \leqslant \varepsilon$ and the smaller ε is, the more fair the decision boundary would be.

With binary outcomes, the latter definition is equivalent to

$$\mathbb{E}[|Y-\hat{Y}|^{\alpha} \,|\, S=0] = \mathbb{E}[|Y-\hat{Y}|^{\alpha} \,|\, S=1], \tag{2.10}$$

for some constant $\alpha > 0$, or even

$$\mathbb{P}(Y-\hat{Y} \leqslant y \,|\, S=0) = \mathbb{P}(Y-\hat{Y} \leqslant y \,|\, S=1),\ y \in \mathbb{R}. \tag{2.11}$$

Then, it makes sense to extend the concept of *lack of disparate mistreatment* for continuous outcomes through the relationships (2.10) or even the stronger requirement (2.11). The latter definition is stronger than the former: (2.11) means an equality in law, when (2.10) is related to an equality in expectations. Hereafter, such concepts will be called strong and weak lack of disparate mistreatment respectively.

The discrepancy between the law of $\hat{Y} \neq Y$ given $(S=0)$ and this law given $(S=1)$ is a global measure of (un)fairness. To inspect more closely-related but more peculiar features, Zafar et al. (2017), proposed to check whether the following relationships apply:

- $\mathbb{P}(\hat{Y} \neq Y|S=0,Y=1) \neq \mathbb{P}(\hat{Y} \neq Y|S=1,Y=1)$ (fairness bias by false

negative signals);

- $\mathbb{P}(\hat{Y} \neq Y | S = 0, Y = 0) \neq \mathbb{P}(\hat{Y} \neq Y | S = 1, Y = 0)$ (fairness bias by false positive signals);
- $\mathbb{P}(\hat{Y} \neq Y | S = 0, \hat{Y} = 1) \neq \mathbb{P}(\hat{Y} \neq Y | S = 1, \hat{Y} = 1)$ (fairness bias by false discovery signals);
- $\mathbb{P}(\hat{Y} \neq Y | S = 0, \hat{Y} = 0) \neq \mathbb{P}(\hat{Y} \neq Y | S = 1, \hat{Y} = 0)$ (fairness bias by false omission signals).

(iv) The criteria above encourage us to introduce new notions of fairness based on conditional probabilities/expectations/laws. The idea is based on the concept of *conditional fairness*. Assume that X includes the sensitive feature S plus some other variables, among which some are of particular interest. To fix the ideas, $X = (Z_1, Z_2, S)$ and S does not belong to $Z := (Z_1, Z_2) \in \mathcal{Z} := \mathcal{Z}_1 \times \mathcal{Z}_2$. We focus on the sub-vector of covariates Z_1 that will be the conditioning variable "of interest". Now, a classifier will be considered as fair when adding the information about the sensitive feature S does not change the prediction of Y, given the value of the sub-covariate Z_1. In other words, a predictor will be said *conditionally fair given* Z_1 if the sensitive feature does not bring any additional information to predict Y, compared to the set of explanatory variables Z, given any fixed value of Z_1.

In the case of *demographic parity*, this means

$$\mathbb{P}(\hat{Y} = j | S = 0, Z_1 = z_1) = \mathbb{P}(\hat{Y} = j | S = 1, Z_1 = z_1),\ j = 0, \ldots, p-1,\ z_1 \in \mathcal{Z}_1. \tag{2.12}$$

In particular, the latter relationships imply

$$\mathbb{E}[\hat{Y} | S = 0, Z_1 = z_1] = \mathbb{E}[\hat{Y} | S = 1, Z_1 = z_1] = \mathbb{E}[\hat{Y} | Z_1],\ z_1 \in \mathcal{Z}_1. \tag{2.13}$$

For example, a judicial process can be considered as fair relatively to ethnical background given earnings, if the probability of being convicted is the same for a black

person and a white person with the same income. In terms of terminology, we call such stronger properties with the adjective "conditional". Here, we would say that the judicial process is fair in terms of *conditional demographic parity given earnings*. Endowed with this additional information about the "not sensitive" features Z (including the interesting features Z_1), it is possible to revisit our first notions of fairness, as introduced in (i)-(iii): simply, all such relationships have to be fulfilled given every conditioning events $(Z_1 = z_1)$ for any $z_1 \in \mathcal{Z}_1$. In particular, a binary classifier does not suffer from *conditional disparate treatment given* Z_1 if (using the same notations as above), for every $z_1 \in \mathcal{Z}_1$,

$$\mathbb{P}(\hat{Y} \neq Y | Z_1 = z_1, S = 0) = \mathbb{P}(\hat{Y} \neq Y | Z_1 = z_1, S = 1) = \mathbb{P}(\hat{Y} \neq Y | Z_1 = z_1).$$

This means that the probability of misclassification does not depend on the sensitive feature S, for any given value of Z_1.

All the concepts of conditional fairness - i.e. that involve conditioning events of the type $(Z_1 = z_1)$ - impose stronger constraints than the previous criteria. For instance, in the case of binary outcomes, if a classifier has the *conditional demographic parity* property, it satisfies the *demographic parity* property itself: $\mathbb{E}[\hat{Y}|S = k, Z_1 = z_1] = \mathbb{E}[\hat{Y}|Z_1 = z_1]$ for every (k, z_1) implies $\mathbb{E}_X[\hat{Y}|S = k] = \mathbb{E}_X[\hat{Y}]$.

3 A related-post processing approach

The idea to build fair classifiers in order to make fair predictions is not new and several ideas have been proposed in the literature from different strategies. Indeed, in order to address the ethic problem posed through the different definitions we have previously introduced, several fairness-aware machine learning algorithms have been investigated that can be categorized as (i) pre- processing techniques designed to modify the input data so that the outcome of any machine learning algorithm applied to that data will be fair, (ii) algorithm modification techniques that modify an existing algorithm or create a new one that will be fair under any inputs, and (iii) postprocessing techniques that take the output of any model

and modify that output to be fair.

The motivation behind pre-processing algorithms is the idea that training data is the cause of the discrimination that a machine learning algorithm might learn. This could be because the training data itself captures historical discrimination or because there are more subtle patterns in the data, such as for instance an under-representation of a minority group. In order to suppress the bias introduced by the training set, several approaches on the input data have been used. We can distinguish three strategies. One involves modifying the labels of the attributes. This means for instance that the proportion of positive labels are equal in the protected and unprotected groups. A classifier is then trained with these new labels assuming that equal opportunity of positive labeling will generalize to the test set. Calders and Kamiran (2009) and Kamiran and Calders (2012) use three approaches to attain this objective: suppression of input data (looking at correlations), change in the labellisation of the data, assign specific weights to the data considering frequency counts. A second approach which is a regularization strategy adds a regularizer to the classification training which quantifies the degree of bias or discrimination. For instance considering disparate impact fairness Feldman et al. (2015) modify each attribute (not the training label) so that the marginal distributions based on the subsets of that attribute with a given sensitive value are all equal. They impose to the binary classifier a constraint to minimize that corresponds to a balanced error rate . A previous work similar to this one is Kamishima et al. (2012). Calmon et al. (2017) have chosen to work by randomisation following the works of Pedreschi et al. (2009) and Zemel et al. (2013). The third approach was proposed by Dwork et al. (2012) for statistical parity. It consists in a mapping to an intermediate representation by optimizing the classification decision criteria while satisfying a Lipschitz condition on individuals which stipulates that nearby individuals should be mapped similarly. They build a randomized classification procedure minimizing an arbitrary loss function.

Algorithm modifications are present in the form of additional constraints, considering that removing sensitive features is insufficient. Kamishima et al. (2012) introduce a fairness focused regularization term and apply it to a logistic regression classifier while still allow-

ing for classification. Calders and Verwer (2010) build separate models for each value of a sensitive attribute and use the appropriate model for inputs with the corresponding value of the attribute. In their paper Zemel et al. (2013) combine pre-processing and algorithm modification. Their approach is to learn a modified representation of the data that is most effective at classification while still being free of signals pertaining to the sensitive attribute. Zafar et al. (2017) observing that standard fairness constraints are non convex and hard to satisfy directly introduce a convex relaxation for purpose of optimization. For a review on these strategies we refer to Friedler et al (2019).

A third approach to building fairness into algorithm design is by modifying the results of a previously trained classifier to achieve the desired results on different groups. This is categorized as post processing techniques. In spite of 'cleaning away' the discrimination from the dataset before a classifier is learned Kamiran et al. (2010) propose an approach in which the non-discriminatory constraint is pushed deeply into a decision tree learner by changing its splitting criterion and pruning strategy by using a novel leaf re-labeling approach after training in order to satisfy fairness constraints. The discrimination of the decision tree before the label of a leaf is changed according to the majority class of this leaf. The same approach is proposed in Zliobaite (2015) who splits the dataset, trains a logistic regression, outputs class probability scores for the test set and varies the classification threshold from 0 to 1, which changes the acceptance rate. Hardt et al. (2016) propose to enforce equalized odds binary choice situations based on univariate scores by conveniently choosing thresholds. They propose the introduction of a predictive score $R = f(X, S)$ with the enforcement that higher values of R correspond to greater likelihood of Y = 1 and thus a bias toward predicting $\hat{Y} = 1$. A binary classifier $\hat{Y}$ can be obtained by thresholding the score, i.e. setting $\hat{Y} = I[R > t]$ for some threshold t, when $S = a$, $a \in \{0, 1\}$. A score R satisfies equalized odds if R is independent of S given Y. The idea that the prediction error restricted to any protected group remains below some pre-determined level has also be investigated introducing fair regression under statistical parity by Agarwal et al. (2019), and also in a related work by Chzhen et al. (2020) in which the authors measure unfairness by the Total Variation distance in place of Kolmogorov-Smirnov distance.

Woodworth et al. (2017) explored the use of post-processing as a way to ensure fairness with respect to error profiles directly incorporating non-discrimination into the learning process. They define a randomized binary predictor $\hat{Y}$ as α-discriminatory with respect to a binary protected attribute S on a given population if $\Gamma(\hat{Y}) = max_{y\in\{0,1\}}|\gamma_{y0}(\hat{Y}) - \gamma_{y1}(\hat{Y})| \leqslant \alpha$ where $\gamma_{ya}(\hat{Y}) = P[\hat{Y} = 1|Y = y, S = a]$. Then they use the following test for detecting α-discrimination on a sample V : $T(\hat{Y}) = I[\Gamma^V_{ya}(\hat{Y}) > \alpha]$. A similar approach is developed in Kearns et al. (2018) who work with subgroups in place of individuals. Agarwal et al. (2018) used a weighted classification implementation of logistic regression and gradient-boosted trees in order to find the lowest-error distribution over a class of classifiers. Following the same ides Donini et al. (2018) enforces fairness directly during the training step of the classifier optimizing a fairness constraint related to the notion of equalized odds.

Our approach is different of the previous methods from several points. First it cannot be associated to a pre-processing technique : we do not modify the input data to make fair the outcomes of the classifiers applied to the data. Indeed, the motivation behind preprocessing algorithms is the idea that training data is the cause of the discrimination that a machine learning algorithm might learn, and so modifying it can keep a learning algorithm trained on it from discriminating. However, to avoid this pitfall, one must be sure that all correlations or other more subtle relationships between the data have been removed. This requires having anticipated all the hidden relationships between these variables. Second it cannot be assimilated to a post-processing approach: we do not introduce modifications in the classifier to create a new one which can be fair under any inputs, and we do not use techniques that take the output of any model and modify it to be fair. This fairness-aware machine learning methods is limited to batch-learning-based interventions.

Our approach which can be associated to a post -processing strategy does not introduce in our algorithm any specific constraints to assure fairness apart from the notion of conditional fairness definition which we retained. We argue that using an additive and repetitive classification can provide smallest errors on the predictors through an optimal solution. The fairness of the predictors is tested through a battery of specific tests with specific confidence level.

Now, assume that any classifier g is based on a regressor (if Y is continuous or ordinal). In case of M regressors, they are denoted as $r_j : X \to \mathbb{R}$, $j = 2, \ldots, M$[2]. Then any quantity $r_j(X)$ is an estimator of $\mathbb{E}[Y|X]$. Note that, in the case of binary outcomes, every $r_j(X)$ is an estimator of $\mathbb{P}(Y = 1|X)$, that is $\mathbb{E}[Y|X]$ too, and any classify $g_j \in \{0, 1\}$ is typically chosen as $g_j(x) = \mathbf{1}\big(r_j(x) > 1/2\big)$.

To illustrate this idea, assume $r_1(X)$ is a Logistic classifier. In this case, this model yields an estimator $\hat{\mathbb{P}}(Y = 1|X = x)$ of $\mathbb{P}(Y = 1|X = x)$ for any $x \in \mathcal{X}$. Then, the predicted class given $X = x$ is typically $\hat{Y} = \mathbf{1}\big(\hat{\mathbb{P}}(Y = 1|X = x) > 1/2\big) = \mathbf{1}(r_1(x) > 1/2)$.
In the next Section, we use this approach to investigate the existence of a solution for the different notions of fairness we have introduced.

4 Fair "bagging" classifiers/regressors

An additive tree model (ATM) is an ensemble of M decision trees. Let $X \in \mathcal{X}$ be the vector of individual features. Each decision tree outputs a real value. Let $g_j(x)$ be the output obtained from tree j, $j = 1, \ldots, M$. For classification purpose, the output G of the additive tree model is a weighted sum of all the tree outputs as follows:

$$G(x) = \sum_{j=1}^{M} \omega_j g_j(x),$$

where $\omega_j \in \mathbb{R}$ is the weight associated to tree j. Then, in the case of binary outputs, fix a threshold a so that $g(x) = (G(x) \geqslant a)$. Typically, $g_j(x) \in \{0, 1\}$, $\omega_j = 1/M$ for every j and $a = 1/2$.

In the case of regressions, the idea of bagging, Breiman (1996), is similar: combine M regression-type predictors r_j to build another predictor

$$r_\omega(x) = \sum_{j=1}^{M} \omega_j r_j(x) =: \omega' \vec{r}(x),$$

[2] g_j could be an indicator or any other continuous function

for every $x \in \mathcal{X}$. For the properties on bagging estimates, we refer to Breiman (2001) and Friedman and Hall (2007), among others. We will see that we do not need many competing classifiers/predictors to build (approximately) fair new ones. Even with $M = 2$, this can be made (as we illustrate below).

The vector of weights $\omega = (\omega_1, \ldots, \omega_M)$ has to be chosen carefully. Even if the weights could depend on x, we keep them constant for the sake of simplicity. In the usual case of random forests, they are all equal even if a few attempts have been made to add another degree of flexibility with different weights: see Maudes et al. (2012), Kim et al. (2011), e.g. Here, we only impose $\sum_{j=1}^{M} \omega_j = \omega' e = 1$, but we do not restrict ourselves to nonnegative weights. Therefore, ω may be any arbitrary vector in $\mathbb{R}^M$ so that $\omega' e = 1$.

The previous formulation is very general and includes some popular models as special cases, like random forests. This additive tree model is widely used in real-world applications and appears as the most popular and powerful off-the-shelf classifer, Breiman (2001). It can cope with regression and multi-class classification on both categorical and numerical datasets with superior accuracy. In essence, random forest is a bagging model, Breiman, (1996), of trees where each tree is trained independently on a group of randomly sampled instances with randomly selected features.

Following Biau (2012, theorem 1), it can be shown that the estimate $r_\omega(x)$, once calibrated using the data set $\mathcal{X}$, is consistent in the sense that $\mathbb{E}[(r_\omega(X) - r(X)^2] = 0$ as $n \to \infty$, with $r(x) := \mathbb{E}[\hat{Y}|X = x]$. This result assumes that, at each node j, d variables X are chosen with a probability $p_{nj}, \sum_{j=1}^{d} p_{nj} = 1$. If each tree has approximatively k_n terminal nodes, then one needs that $p_{nj} k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Other conditions are provided in this paper on the variance of the estimates under more general conditions. Asymptotic normality is obtained in Wager and Athey (2018, Theorem 3.1) under more complex conditions on the choice of d, M, moments and Lipschitz conditions on $\mathbb{E}[Y|X = x]$, we refer also to Biau and Scornet (2015).

We would like to calibrate ω to obtain a predictor r_ω that fulfills "at most" some of the latter concepts of fairness. In other words, how can we combine several possibly "unfair" classifiers so that the bagging predictor becomes "fair"? For instance, in the case of demographic parity

and equalised odds, this predictor should satisfy some moment conditions as (2.3) or (2.8) respectively, at least approximately. Note that such conditions are necessary but in general not sufficient to insure fairness. For obtaining approximate moment condition, we rely on a sample $(X_i, Y_i)_{i=1,\dots,n}$, $X_i = (S_i, Z_i)$, that can reasonably be assumed i.i.d.

1. Let us deal first with the constraint (2.3) to satisfy *demographic parity*. The empirical version of the latter constraint is obtained by replacing the theoretical expectation by the empirical measure, as given by the calibration subset. A convenient choice of ω should minimize a distance between $\sum_{i=1}^n \mathbf{1}(S_i = 0)\omega'\vec{r}(X_i)/\sum_{i=1}^n \mathbf{1}(S_i = 0)$ and $\sum_{i=1}^n \omega'\vec{r}(X_i)/n$. Actually, in general, there exist an infinity of vectors ω s.t.

$$\sum_{i=1}^n \mathbf{1}(S_i = 0)\omega'\vec{r}(X_i)/\sum_{i=1}^n \mathbf{1}(S_i = 0) = \sum_{i=1}^n \omega'\vec{r}(X_i)/n, \tag{4.1}$$

strictly speaking given by the intersection of two hyperplans in $\mathbb{R}^M$.

Remark 4.1. *If, in addition, we impose that the weights have to belong into* $[0, 1]$*, the existence of a solution is no longer guaranteed. In this case, it is necessary to rely on approximated solutions* ω*. For instance, it makes sense to propose a vector of weights as a solution of the program*

$$\arg\min_{\omega} \Big(\sum_{i=1}^n \mathbf{1}(S_i = 0)\omega'\vec{r}(X_i)/\sum_{i=1}^n \mathbf{1}(S_i = 0) - \sum_{i=1}^n \omega'\vec{r}(X_i)/n \Big)^2, \tag{4.2}$$

under the constraints $\omega'e = 1$ *and* $\omega_k \in [0, 1]$ *for every* $k = 1, \dots, M$*. The latter optimization program can be easily solved by quadratic programming. Sometimes, it yields degenerate solutions for which all the weights are zero, except for a single component. Nonethless, in general, there are an infinite number of solutions of (4.2). To obtain unicity, it will be necessary to introduce the prediction accuracy of the new bagging classifer/regressor: see Section 5.*

2. More interesting is the case of *equalized odds*, and the constraint (2.8), that has to be empirically approximated. It is necessary to introduce some nonparametric estimators of $\mathbb{E}[\hat{Y}|S = k, Y = y]$, $k = 0, 1$, and $\mathbb{E}[\hat{Y}|Y = y]$. We choose the usual Nadaraya-

Watson estimator of regression functions, w.l.o.g.:

$$\mathbb{E}[\hat{Y}|S=k,Y=y] \simeq \hat{r}_k(y) := \frac{\sum_{i=1}^n \omega'\vec{r}(X_i)\mathbf{1}(S_i=k)K_h(Y_i-y)}{\sum_{i=1}^n \mathbf{1}(S_i=k)K_h(Y_i-y)} =: \omega' v_{k,h}(y),$$

for some univariate kernel K, a bandwidth sequence $h = h(n) \to 0$ and $K_h(t) := K(t/h)/h$ for every y. The kernel K_h can be seen as a proximity measure between two forecasted points[3].

Similarly,

$$\mathbb{E}[\hat{Y}|Y=y] \simeq \hat{r}(y) := \frac{\sum_{i=1}^n \omega'\vec{r}(X_i)K_h(Y_i-y)}{\sum_{i=1}^n K_h(Y_i-y)} =: \omega' v_h(y).$$

Note that $\hat{r}_k(y)$ and $\hat{r}(y)$ are linear function of ω, for any y. This yields an estimator of ω as

$$\hat{\omega}_{e,0} := \arg\min_{\omega} \sum_{i=1}^n \left|\hat{r}_0(Y_i) - \hat{r}(Y_i)\right|^\alpha$$

for some $\alpha > 0$, under the constraint $\omega' e = 1$.

Let us impose $\alpha = 2$. Then, this yields

$$\hat{\omega}_{e,0} := \arg\min_{\omega} \sum_{i=1}^n \omega'\big(v_{0,h}(Y_i) - v_h(Y_i)\big)\big(v_{0,h}(Y_i) - v_h(Y_i)\big)'\omega, \tag{4.3}$$

under the constraint $\omega' e = 1$. Set the $M \times M$ real matrix

$$\Sigma_h = n^{-1}\sum_{i=1}^n \big(v_{0,h}(Y_i) - v_h(Y_i)\big)\big(v_{0,h}(Y_i) - v_h(Y_i)\big)'.$$

Then, by the usual optimization of a quadratic form under a linear constraint and the Lagrangian method, we obtain a unique solution $\hat{\omega}_{e,0} = \Sigma_h^{-1}e/e'\Sigma_h^{-1}e$.

Note that Σ_h is a consistent estimator of $\Sigma = \mathbb{E}[\big(v_{0,h}(Y_i) - v_h(Y_i)\big)\big(v_{0,h}(Y_i) - v_h(Y_i)\big)']$. Unfortunately, when $M >> 1$ as in the case of random forests, the covariance matrix Σ may be high dimension, and the sample covariance matrix Σ_h could suffer from a

[3]Hence, any forest has its own metric K_h, but unfortunately the one associated with the CART-splitting strategy is strongly data-dependent and therefore complicated to work with. For discussion on the introduction of kernel in random forest, see Biau et al. (2015).

significant amount of sampling error. Its inverse may be a poor and non-invertible estimator for Σ. Moreover, when Y is discrete and may take q values ($q = 2$ in the binary case, typically), it is easy to check that the rank of Σ_h is at most $\min(q, M)$. In other terms, for a binary Y, the latter matrix is no longer invertible when we consider a linear combination of three or more individual classifiers ! To avoid these problems, we could replace Σ_h using the shrinkage method [4].

3. As for equalized odds, we could evaluate an optimal ω under the point of view of *lack of disparate mistreatment*. Let us choose a L^2 distance, i.e. $\alpha = 2$. An estimator of $\mathbb{E}[(\hat{Y} - Y)^2 \,|\, S = k]$, $k = 0, 1$, is given by

$$\begin{aligned}\mathbb{E}[(\hat{Y} - Y)^2 | S = k] &\simeq \frac{\sum_{i=1}^n (\omega' \vec{r}(X_i) - Y_i \omega' e)^2 \mathbf{1}(S_i = k)}{\sum_{i=1}^n \mathbf{1}(S_i = k)} \\ &= \omega' \frac{\sum_{i=1}^n (\vec{r}(X_i) - Y_i e)(\vec{r}(X_i) - Y_i e)' \mathbf{1}(S_i = k)}{\sum_{i=1}^n \mathbf{1}(S_i = k)} \omega =: \omega' \bar{\Sigma}_k \omega,\end{aligned}$$

under the constraint $\omega' e = 1$ and with obvious notations. Similarly,

$$\mathbb{E}[(\hat{Y} - Y)^2] \simeq \frac{\sum_{i=1}^n (\omega' \vec{r}(X_i) - Y_i \omega' e)^2}{n} =: \omega' \bar{\Sigma} \omega,$$

denoting $\bar{\Sigma} := \sum_{i=1}^n \big(\vec{r}(X_i) - Y_i e\big)\big(\vec{r}(X_i) - Y_i e\big)'/n$. Note that $\bar{\Sigma}$ is a convex combination of $\bar{\Sigma}_0$ and $\bar{\Sigma}_1$. The goal would be to find a vector ω so that, for some $k = 0, 1$, we have

$$\omega' \big(\bar{\Sigma} - \bar{\Sigma}_k\big) \omega = 0, \text{ and } \omega' e = 1. \tag{4.4}$$

The latter program has no solution if the symmetric matrix $\bar{\Sigma} - \bar{\Sigma}_k$ is strictly positive or strictly negative. Otherwise, there are an infinite number of solutions. Indeed, in this case, all $\omega \in \mathbb{R}^M$ s.t. $\omega' \big(\bar{\Sigma} - \bar{\Sigma}_k\big) \omega = 0$ is a linear subspace of $\mathbb{R}^M$. The kernel $N_k = \{\omega \,|\, \omega' \big(\bar{\Sigma} - \bar{\Sigma}_k\big) \omega = 0\}$ of this quadratic form can be specified after a Gram Schmidt process that yields an orthonormal basis, typically. We promote to restrict ourselves to a compact set of ω, and to solve the program $\arg\min_\omega \omega' \Omega \omega$, under the linear constraints $\omega' e = 1$, $\omega \in N_k$ and $\|\omega\|_\infty \leqslant 1$, for any positive definite

[4] The shrinkage estimator is a linear combination of the sample estimator and a fixed matrix or another estimator, Hastie et al. (2009), e.g.

matrix Ω. When the kernel of $\bar{\Sigma} - \bar{\Sigma}_k$ is not reduced to 0, the latter program has (at least) a solution, because one has to minimize a continuous function of ω on a non-empty compact subset. Alternatively, a simpler solution would be to calculate $\arg\min_\omega \left\{\omega'\big(\bar{\Sigma} - \bar{\Sigma}_k\big)\omega\right\}^2$, under $\omega' e = 1$. The solution is not unique in general and the optimization is not longer quadratic, but this avoids the choice of an arbitrary matrix Ω.

4. Concerning *conditional fairness given* Z, it is unrealistic to impose the identity between $r_\omega(z,0)$ and $r_\omega(z,1)$ for every z of interest. We rather promote a "best solution" in average, in the same spirit of *equalized odds*. Depending on any definition of fairness, denote by W the relevant conditioning vector. With the definitions of Section 2, to obtain *conditional demographic parity given* Z_1, consider $W = Z_1$; in the case of *conditional equalized odds given* Z_1, set $W = (Z_1, Y)$. To build a fair classifier/regressor by linear combination of , we want that a relationship $\mathbb{E}[\omega'\vec{r}(W,0)|S=0, W=w] = \mathbb{E}[\omega'\vec{r}(W,1)|S=1, W=w]$ is satisfied "at best" for all values of w. Set

$$\hat{\omega}_{sf} := \arg\min_\omega \sum_{i=1}^n \left|\omega'\vec{r}(W_i,0) - \omega'\vec{r}(W_i,1)\right|^\alpha,$$

under the constraint $\omega' e = 1$. As in Section 2, we have defined

$$\vec{r}(w,k) := \frac{\sum_{i=1}^n \vec{r}(X_i)\mathbf{1}(S_i=k)K_h(W_i-w)}{\sum_{i=1}^n \mathbf{1}(S_i=k)K_h(W_i-w)},\ k \in \{0,1\}.$$

As above and when $\alpha = 2$, the solution is $\hat{\omega}_{sf} = \Sigma_{sf}^{-1}e/e'\Sigma_{sf}^{-1}e$, where

$$\Sigma_{sf} := n^{-1}\sum_{i=1}^n \big(\vec{r}(W_i,0) - \vec{r}(W_i,1)\big)\big(\vec{r}(W_i,0) - \vec{r}(W_i,1)\big)'.$$

Concerning the lack of disparate mistreatment, the goal is to satisfy

$$\mathbb{E}\Big[\big(\omega'\vec{r}(W_i,0) - Y\big)^2 | S=0, Z_1=z\Big] = \mathbb{E}\Big[\big(\omega'\vec{r}(W_i,1) - Y)^2\big) | S=1, Z_1=z\Big],$$

for every z. Note that the latter identity involves a quadratic form in ω, but indexed by z. For $k \in \{0,1\}$, define

$$\bar{\Sigma}_k(z) := \frac{\sum_{i=1}^n (\vec{r}(X_i) - Y_i e)(\vec{r}(X_i) - Y_i e)' \mathbf{1}(S_i = k) K_h(Z_i - z)}{\sum_{i=1}^n \mathbf{1}(S_i = k) K_h(Z_i - z)}.$$

Thus, we propose to solve

$$\hat{\omega}_{ldm} := \arg\min_{\omega} \sum_{i=1}^n \Big\{ \omega' \big(\bar{\Sigma}_0(Z_{1,i}) - \bar{\Sigma}_1(Z_{1,i}) \big) \omega \Big\}^2,$$

under the constraint $\omega' e = 1$ and, possibly, $\omega_j \in [0,1]$ for every j.

Thus, under realistic constraints we have calibrated ω to get fair bagging predictors for demographic parity, equalized odds, lack of disparate mistreatment and conditional fairness given some covariate.

5 Fairness and predictive power

In the previous section, the proposed (possibly approximated) fair procedures based on "bagging" do not consider the predictive power of our final classifier or regressor. In other words, does our weighting scheme improve or deteriorate the performances of prediction, compared to the initial classifiers/regressors ? Intuitively, when we use fair predictors based on weighted averages of some initial "reasonable" predictors, it is likely that the new performances should be "average", in a rough sense, i.e. in the range of the performances obtained with the initial predictors. We could blindly stay with such an intuition, but it is possible to go one step forward by directly adding a constraint in terms of performance during the calibration stage of our weights.

To this goal, let us reconsider the fair classifiers built in Section 4. Generally spealing, the quality of a predictor is measured by a comparison between predicted values - here $\omega' \vec{r}(X_i)$ - and the true realizations Y_i. Let d be a distance between two real numbers, typically $d_1(x,y) := |x-y|$ or $d_2(x,y) := (x-y)^2$. The idea will be to penalize our criteria to impose more or less prediction accuracy, on the calibration dataset and through a tuning parameter

$\lambda \geqslant 0$.

In the case of *demographic parity*, an extended criterion for finding optimal weights would be

$$\arg\min_{\omega} \Big(\frac{\sum_{i=1}^n \mathbf{1}(S_i = 0)\omega' \vec{r}(X_i)}{\sum_{i=1}^n \mathbf{1}(S_i = 0)} - \frac{1}{n}\sum_{i=1}^n \omega' \vec{r}(X_i) \Big)^2 + \frac{\lambda}{n}\sum_{i=1}^n d\big(\omega' \vec{r}(X_i), Y_i\big), \tag{5.1}$$

possibly under the constraints $\omega' e = 1$ and/or $\omega_k \in [0,1]$, $k = 1, \ldots, M$. Note that, if $d = d_2$ and without imposing $\omega_k \in [0,1]$, then the latter program has a unique solution $\omega^*_{DP} := \Sigma_{DP}^{-1} e / e' \Sigma_{DP}^{-1} e$, by setting

$$\Sigma_{DP} = \Delta_{DP}\Delta'_{DP} + \frac{\lambda}{n}\sum_{i=1}^n \big(\vec{r}(X_i - Y_i e)\big)\big(\vec{r}(X_i - Y_i e)\big)' = \Delta_{DP}\Delta'_{DP} + \lambda\bar{\Sigma},$$

$$\Delta_{DP} := \frac{\sum_{i=1}^n \mathbf{1}(S_i = 0)\vec{r}(X_i)}{\sum_{i=1}^n \mathbf{1}(S_i = 0)} - \frac{1}{n}\sum_{i=1}^n \vec{r}(X_i).$$

By choosing the parameter λ, we manage a trade-off between targeting perfect fairness ($\lambda = 0$) and picking-up the "best" initial classifier ($\lambda >> 1$).

Alternatively, assume there is an infinity of solutions ω of (4.1), that we rewrite $\omega' v_n = 0$. When one does not impose $\omega \in [0,1]^M$, this is the case a.e. A natural idea would be to select the best "fair bagging predictor" among the latter ones. This can be done by solving the program

$$\arg\min_{\omega} \sum_{i=1}^n d\big(\omega' \vec{r}(X_i), Y_i\big), \text{ with } \omega' e = 1 \text{ and } \omega' v_n = 0. \tag{5.2}$$

The latter program has always a solution $\omega \in \mathbb{R}^M$, for "reasonable" distances d, as $d\big(\omega' \vec{r}(X_i), Y_i\big) = \big(\omega' \vec{r}(X_i) - Y_i\big)^2$. If we impose $\omega \in [0,1]^M$ in addition, (5.4) may have no solution, but only when all the components of v_n have the same sign. Under such circumstances, a practical solution would be to try different individual classifiers until the vector v_n has some componnents of different signs, or to (slightly) disturb one of the existing basis predictors.

In the case of *equalized odds*, the new program would be

$$\arg\min_{\omega} \frac{1}{n}\sum_{i=1}^n \Big|\hat{r}_0(Y_i) - \hat{r}(Y_i)\Big|^{\alpha} + \frac{\lambda}{n}\sum_{i=1}^n d\big(\omega' \vec{r}(X_i), Y_i\big),$$

for some $\alpha > 0$, under the constraint $\omega' e = 1$. When $d = d_2$ and with the same notations as in Section 4, this means solving

$$\arg\min_{\omega} \omega' \Sigma_h \omega + \frac{\lambda}{n} \sum_{i=1}^{n} \big(\omega' \vec{r}(X_i) - Y_i\big)^2,$$

whose solution is $\hat{\omega}^*_{EO} := \Sigma^{-1}_{EO} e / e' \Sigma^{-1}_{EO} e$, with $\Sigma_{EO} := \Sigma_h + +\lambda\bar{\Sigma}$. Obviously, it is possible to add the constraints $\omega_k \in [0,1]$ for every k. This would mean losing analytic solutions and relying on numerical optimization (quadratic programming).

By a similar reasoning, *lack of disparate mistreatment* implies solving

$$\arg\min_{\omega} \omega' \big(\bar{\Sigma} - \bar{\Sigma}_k\big)\omega + \frac{\lambda}{n} \sum_{i=1}^{n} d\big(\omega' \vec{r}(X_i), Y_i\big), \tag{5.3}$$

under the constraint $\omega' e = 1$, and possibly under $\omega_k \in [0,1]$ for every $k \in \{1, \ldots, M\}$. Again, when $d = d_2$ and without the latter constraints, there is an unique analytic solution $\hat{\omega}^*_{LDM} := \Sigma^{-1}_{LDM} e / e' \Sigma^{-1}_{LDM} e$, where $\Sigma_{LDM,k} := \bar{\Sigma} - \bar{\Sigma}_k + \lambda\bar{\Sigma}$.

Similar reasoning apply with *conditional fairness*. Details are left to the reader.

Actually, in the case of binary explained variables $Y \in \{0,1\}$, it may appear strange to calibrate our weights ω, or, even more, to evaluate the quality of the new classifier by comparing continuous outcomes $\omega' \vec{r}(X_i)$ to binary observations Y_i. Alternatively, it is tempting to replace the predictor $p_{i,0}(\omega) := \omega' \vec{r}(X_i)$ by $p_{i,1}(\omega) := \mathbf{1}\big(\omega' \vec{r}(X_i) > 0.5\big)$, for every i. Therefore, this opens to many slightly different calibration criteria. When numerical solutions are invoked instead of closed form ones, all of them yield reasonable alternatives.

(a) In the case of *demographic parity*, our penalized criterion for finding optimal weights could be

$$\arg\min_{\omega} L^{DP}_{n,k,l} := \Big(\frac{\sum_{i=1}^n \mathbf{1}(S_i = 0) p_{i,k}(\omega)}{\sum_{i=1}^n \mathbf{1}(S_i = 0)} - \frac{\sum_{i=1}^n \mathbf{1}(S_i = 1) p_{i,k}(\omega)}{\sum_{i=1}^n \mathbf{1}(S_i = 1)}\Big)^2 + \frac{\lambda}{n} \sum_{i=1}^{n} d\big(p_{i,l}(\omega), Y_i\big), \tag{5.4}$$

for every choice of $(k,l) \in \{0,1\}^2$, possibly under the constraints $\omega' e = 1$ (and possibly $\omega_k \in [0,1]$, $k = 1, \ldots, M$).

(b) In the case of *equalized odds*, it is necessary to redefine some quantities of interest: for every $(k,s) \in \{0,1\}^2$,

$$\mathbb{E}[\hat{Y}|S=s,Y=y] \simeq q_k(s,y,\omega) := \frac{\sum_{i=1}^n p_{i,k}(\omega)\mathbf{1}\big(S_i=s\big)K_h(Y_i-y)}{\sum_{i=1}^n \mathbf{1}(S_i=s)K_h(Y_i-y)}.$$

When Y is binary and $h<0.5$, the latter quantity is equal to

$$q_k(s,y,\omega) := \frac{\sum_{i=1}^n p_{i,k}(\omega)\mathbf{1}\big(S_i=s,Y_i=y\big)}{\sum_{i=1}^n \mathbf{1}(S_i=s,Y_i=y)}.$$

When $k=1$, note that the latter function of ω is not differentiable.

The new penalized-EO program could be

$$\arg\min_{\omega} L_{n,k,l}^{EO} := \frac{1}{n}\sum_{i=1}^n \Big|q_k(0,Y_i,\omega) - q_k(1,Y_i,\omega)\Big|^{\alpha} + \frac{\lambda}{n}\sum_{i=1}^n d\big(p_{i,l}(\omega),Y_i\big), \tag{5.5}$$

for some $\alpha>0$ and every couple $(k,l)\in\{0,1\}^2$, under the constraint $\omega'e=1$ (and possibly $\omega_k\in[0,1]$, $k=1,\ldots,M$).

(c) For the lack of disparate mistreatment, for every s and k in $\{0,1\}$,

$$\mathbb{E}[(\hat{Y}-Y)^2|S=s] \simeq \frac{\sum_{i=1}^n (p_{i,k}(\omega)-Y_i\omega'e)^2\mathbf{1}(S_i=s)}{\sum_{i=1}^n \mathbf{1}(S_i=s)} =: \ell_{k,s}(\omega)$$

This leads to new generalized LDM criteria

$$\arg\min_{\omega} L_{n,k,l}^{LDM} := \big(\ell_{k,0}(\omega)-\ell_{k,1}(\omega)\big)^2 + \frac{\lambda}{n}\sum_{i=1}^n d\big(p_{i,l}(\omega),Y_i\big), \tag{5.6}$$

for every choice of $(k,l)\in\{0,1\}^2$, possibly under the constraints $\omega'e=1$ (and possibly $\omega_k\in[0,1]$, $k=1,\ldots,M$).

When $(k,l)=(1,1)$, the latter programs involve non-differentiable functions of ω. Thus, there is no unique solution. To enforce a single solution, we use a numerical trick, by smoothing the step functions $\omega\mapsto\mathbf{1}(\omega'\vec{r}(X_i)>0.5)$ in the "accuracy terms" (second terms of the penalized criteria). The latter functions are replaced by $\omega\mapsto\Phi\Big(\big(\omega'\vec{r}(X_i)-0.5\big)/\sigma\Big)$, for some (small) constant $\sigma>0$ (Φ denotes the cdf of a $\mathcal{N}(0,1)$). JDF: In the R-code,

the parameters discr-fair and discr-prec are related to the indices k and l in the formulas (5.4), (5.5) and (5.6).

To evaluate the performances of all the new classifiers, a measure of fairness and a measure of accuracy (prediction power) are calculated (out-of-sample on some test sub-samples). For this task, we are free of choosing between continuous and/or discrete predictors, as above during the calibration stage. JDF: This additional degree of freedom is managed by the variables discr-fair-out and discr-prec-out in the code.

6 Tests of fairness based on expectations and conditional expectations

6.1 The problem

The previous definitions of fairness can be statistically tested, a highly desirable feature. Strictly speaking, any of our zero assumptions will be related to a particular definition of "fairness", as written in terms of equality between some (conditional) probabilities, expectations or laws, possibly given some vectors of covariates Z. In practice, two independent samples are available: the first sample is corresponding to $S = 0$, and the second to $S = 1$. Therefore, we are facing a two sample problem. When conditioning variables apart S appear in the considered definition of fairness, our problem is reduced to checking a particular case of conditional independence.

To be specific, our null assumption to be tested will be

$$\mathcal{H}_0^* : \text{"the classifier predictor is fair"}.$$

Assume we observe two iid samples $(Y_i, \hat{Y}_i, Z_i)_{i=1,\dots,n}$ and $(Y_j, \hat{Y}_j, Z_j)_{j=n+1,\dots,n+m}$. The former (resp. latter) sample is corresponding to observations for which $S = 0$ (resp. $S = 1$). In this paper, we take as given a particular classifier/predictor. The goal is to decide whether it is fair "ex post", i.e. after its learning/inference/calibration stage. In particular, this means the predictors have been estimated with another dataset (the so-called "learning" dataset).

Now, we have an access to additional data, called the "testing" dataset. In practice, this implies keeping aside some part of a given global and unique dataset for testing purpose, a usual procedure for calibration and validation in machine learning.

Remark 6.1. *A different perspective would be to build a fair classifier/predictor by using only covariates that are "informative", i.e. that add an useful amount of information to predict Y. This is related to the well-known problem of "variable importance", or "variable selection" in statistics and machine learning. See Watson and Wright (2019), and the references therein, for instance. It should be possible to impose some fairness constraints during such inference procedures but this will not be our approach in this paper.*

There are many ways of testing fairness, depending on its chosen definition. Now, let us review the main "omnibus" approaches, i.e. the methodologies that can be applied for any statistical model (parametric, semi- or non-parametric) and without additional assumptions, beside regularity.

6.2 Tests of fairness based on expectations

The simplest and oldest testing procedures do not try to test the identity between two (possibly conditional) laws strictly speaking but rather between their (possibly conditional) moments. This idea would a priori induce a loss of power [5]. Nonetheless, such test statistics are most often significantly simpler than those based on some distances between laws. In particular, their limiting distibutions under $\mathcal{H}_0^*$ and their asymptotic variances may often be easily evaluated, sometimes analytically.

Let us illustrate this method with *demographic/statistical parity*. The relationship (2.3) can be tested by comparing the two empirical means of $\hat{Y}$ given $S = 0$ and/or $S = 1$. Obviously, if $\hat{Y}$ is discrete, this is equivalent to testing the identity between $\mathbb{P}(\hat{Y} = 1|S = 0)$ and $\mathbb{P}(\hat{Y} = 1|S = 1)$. If we denote $m_{\hat{Y},0} := E[\hat{Y}|S = 0]$ and $m_{\hat{Y},1} := E[\hat{Y}|S = 1]$, one wants to test

$$\mathcal{H}_0^m : m_{\hat{Y},0} = m_{\hat{Y},1}, \text{ against } \mathcal{H}_1^{\mathrm{m}} : \mathrm{m}_{\hat{\mathrm{Y}},0} \neq \mathrm{m}_{\hat{\mathrm{Y}},1}.$$

[5] Indeed, there obviously exist some couples of laws that are different even if their means are equal !

This is corresponding to the classical Behrens-Fisher problem, for which many answers have been proposed in the literature for a long time: Student's t test, Whelch's t test (Y univariate), Hotelling's test (Y multivariate), etc. The same ideas can be applied for testing the *lack of disparate mistreatment* through the identity (2.10).

Let us illustrate. Classical solutions to the Behrens -Fisher problem are different when the variances of the explained variable $\hat{Y}$ in the two samples are unknown or not, unequal or not, and also when one works with univariate or multivariate variables. In the univariate case and when the two latter variances are supposed to be unknown and unequal, a classical test statistic is $T_1 := \big(\hat{m}_{\hat{Y},0} - \hat{m}_{\hat{Y},1}\big)/\hat{\sigma}_{T_1}$ where $\hat{m}_{\hat{Y},k}$ denotes the empirical mean of $\hat{Y}$ given $S = k$, $k \in \{0, 1\}$, and $\hat{\sigma}_{T_1}$ is an estimator for the standard deviation of $\hat{m}_{\hat{Y},0} - \hat{m}_{\hat{Y},1}$: $\hat{\sigma}_{T_1} := \sqrt{\frac{2}{\bar{n}}}\sqrt{\frac{(n-1)s_1^2+(m-1)s_2^2}{(n-1)+(m-1)}}$, where $\bar{n}$ is the harmonic mean of n and m, s_1^2 (resp. s_2^2) is the sample variance of $\hat{Y}$ the first (resp. second) sample. Under $\mathcal{H}_0^m$, the law of the statistic T_1 is approximated by a t-distribution with $\nu_1 := n + m - 2$ degrees of freedom. An alternative test statistic is $T_2 := \big(\hat{m}_{\hat{Y},0} - \hat{m}_{\hat{Y},1}\big)/\hat{\sigma}_{T_2}$ where $\hat{\sigma}_{T_2}^2 := s_1^2/n + s_2^2/m$, with the same notations as above. Under $\mathcal{H}_0^m$, the statistic T_2 is close to a t-distribution with ν_2 degrees of freedom, setting $\nu_2 := \frac{\left(s_1^2/n+s_2^2/m\right)^2}{(s_1^2/n)^2/(n-1)+((s_2^2/m)^2/(m-1).}$ These statistics have been established by Behrens (1929), Fisher (1935) and Welch (1938). Some more recent presentation and discussions are available in Scheffe (1970) or Derrick et al. (2016) for instance.

Dealing with d-dimensional vectors Y, the previous statistics have been extended through the so-called Hotelling test (Willems et al. 2002). Keeping the same notations as before, the null assumption is formally the same as $\mathcal{H}_0^m$, and a convenient test statistic T_3 is defined by $T_3^2 := \frac{nm}{n+m}(\hat{m}_{\hat{Y},0} - \hat{m}_{\hat{Y},1})'\hat{\Sigma}^{-1}(\hat{m}_{\hat{Y},0} - \hat{m}_{\hat{Y},1})$, where $\hat{\Sigma}$ is a "global" covariance matrix: denoting by $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ the sample covariances of $\hat{Y}$ in the first and second sample respectively, set $\hat{\Sigma} := \frac{(n-1)\hat{\Sigma}_1+(m-1)\hat{\Sigma}_2}{n+m-2}$. With Gaussian variables $\hat{Y}$ in each sample and under the null, the statistic T_3^2 follows the Hotelling distribution $\mathcal{T}^2(p, n + m - 2)$. Note that the latter law is linked to a non-central Fisher distribution F by the relationship $(n + m - 2)pF(p, n + m - 1 - p) \overset{\text{law}}{=} (n + m - p + 1)T_3^2$.

In the case of *Equalized odds*, the conditioning events are no longer $(S = k)$ for some $k \in \{0, 1\}$ but rather $(S = k, Y = y)$ for any y in the support of Y-law: see (2.5) and (2.8).

For every particular value of Y, it is tempting to lead a test of

$$\mathcal{H}_0^{m,y} : \mathbb{E}(\hat{Y} = 1|S = 0, Y = y) = \mathbb{E}(\hat{Y} = 1|S = 1, Y = y)$$

that would be built through a test statistic $T_{n,m}(y)$. Then, a general testing procedure of $\mathcal{H}_0^*$ may be obtained through the statistic $\mathcal{T}_{n,m} := \int T_{n,m}(y)\,\mu(dy)$, for some (discrete or continuous) measure μ on the support of Y.

6.3 Tests of fairness based on conditional expectations

The latter general idea has been applied by several authors and is a way of obtaining many test statistics of $\mathcal{H}_0^*$ by comparing conditional means for several notions of fairness. The case of continuous conditional variables Y (or (Y, Z) for conditional fairness) necessitates localization techniques (kernel smoothing, typically), at the price of numerical difficulties when the dimension of the conditioning variable is larger than three, typically (curse of dimensionality).

To keep things general and to potentially encompass all cases of interest (i.e. all concepts of fairness), let us introduce new notations. The explained/predicted variable will be denoted as $V \in \mathbb{R}$, and the potential covariates will be staked in a random vector $W \in \mathbb{R}^q$. For instance, in the case of demographic parity (resp. equalized odds), we have $V = \hat{Y}$ and $W = \varnothing$ (resp. $W = Y$). Adding a vector of covariates Z, conditional demographic parity (resp. equalized odds) corresponds to $W = Z$ (resp. $W = (Y, Z)$). Concerning the lack of disparate mistreatment, $V = Y - \hat{Y}$ and $W = \varnothing$ or $W = Z$ depending on we discuss usual or conditional fairness.

Then, the current statistical problem is reduced to testing the zero assumption

$$\mathcal{H}_0^m : \mathbb{E}[V|S = 0, W = w] = \mathbb{E}[V|S = 1, W = w], \quad \text{for every } w \in \mathbb{R}^d.$$

Again, keep in mind that $\mathcal{H}_0^m$ is a consequence of $\mathcal{H}_0^*$, but is not equivalent to $\mathcal{H}_0^*$.

In a nonparametric setting (i.e. no parametric assumption on (V, S, W)), Delgado and González Manteiga (2001) have proposed a direct comparison of kernel-based estimators of

conditional expectations. In contrast with Delgado and González Manteiga (2001) which employs indicator testing functions, Huang et al. (2016) use a distance between a family of conditional moments weighted by a set of so-called "Generically Comprehensively Revealing" functions (possibly parameterized). Such functions are more fexible than indicators and hence may better present the information.

When our conditioning variables W are discrete, not too numerous and take only a few values, things are significantly simpler. Indeed, conveniently splitting the testing dataset, it is possible to independently lead several Behrens-Fisher type tests. For instance, it is possible to test the relationships (2.5) and (2.6) by applying p times some usual tests proposed for checking demographic parity. We are facing p test statistics and would like to test whether the assumption of equalized odds is realistic. This problem of multiple testing is classical. Finding critical regions can be made after Bonferroni-type corrections, for instance. But, to avoid the induced loss of statistical power, it is probably better to directly test all the p relationships simultaneously, by building a relevant global test statistics $T_{n,m}(y)$ as for Hotelling's test.

Instead of working with (possibly conditional) moments, an alternative would be to compare (possibly conditional) medians, or even $\alpha-$ quantiles for any $\alpha \in (0,1)$. Indeed, under $\mathcal{H}_0^*$, the quantiles of order α of V given $(S=0, W=w)$ and of V given $(S=1, W=w)$ are the same. Such an approach may be considered as more robust than the previous moment-based one, as quantiles are not influenced by outliers. For instance, the Hodges-Lehmann two sample test is based on checking whether the median of the set $\{V_i - V_j; i = 1, \ldots, n \; j = n+1, \ldots, n+m\}$ can be considered as zero. We refer to Dehling and Fried (2012) for extensions towards arbitrary quantile levels. This idea is close to the rank-based Wilcoxon-Mann-Whitney (WMW) test procedure, when there are no covariates W. Formally, its zero assumption is $\mathcal{H}_0^w : \mathbb{P}(V_0 > V_1) = \mathbb{P}(V_1 > V_0)$, where V_k is the law of V given $S = k$, $k \in \{0,1\}$. In the case of a continuous r.v. V and using empirical quantiles, the WMW test statistics can be easily calculated as $WMW_{n,m} := \sum_{i=1}^{n} R_i$, where R_i is the rank of V_i inside the combined sample $(V_i)_{i=1,\ldots,n+m}$. See Lee (2019) for technical details and the asymptotic law of $WMW_{n,m}$. If the samples sizes n and m are large, the latter statistic

converges to a Gaussian distribution, and for small samples it is tabulated. Concerning a comparison between t-tests and rank tests, keeping in mind the zero assumption $\mathcal{H}_0^*$, see the survey of Fay and Proscham (2010) and the references therein.

Remark 6.2. *The topic of this section may be linked to the evaluation of treatment effects, that has induced a flourishing literature in econometrics. In particular, our zero assumption $\mathcal{H}_0^m$ can be interpreted as the property of "no selection bias". See Imbens and Wooldridge (2009) for a survey, and Lee and Wang (2009) for a general approach of testing. In particular, Heckman et al. (1998) proposed several kernel smoothing techniques.*

7 Tests of fairness based on comparisons between distributions

7.1 Introduction

Working with means or conditional means induces some lack of information, if the goal is to test $\mathcal{H}_0^*$. The disadvantage of the tests of Section 6.2 is that they are only consistent against a rather restricted set of alternatives and they could be quite ineffective in certain situations. In other words, such testing procedures may induce a significant loss of power, compared to alternative procedures that would focus on laws directly, i.e. that would compare (conditional) cdfs', densities, quantiles, characteristic functions, etc.

When there is no covariate, a general answer is to consider a distance D between two distributions. Once empirically estimated, the approximated distance becomes a natural candidate for defining a test statistic. To illustrate, in the case of our previous r.v. of interest V, denote by F_k (resp. f_k) the cdf (resp. density w.r.t. the Lebesgue measure, when it exists) of V given $(S = k)$, $k \in \{0, 1\}$. Usual distances are

- $D_\gamma(F_0, F_1) := \int \big(F_0 - F_1\big)^\gamma(y) w(y)\, dy$ or $D_\gamma(f_0, f_1) = \int \big(f_0 - f_1\big)^\gamma(y) w(y)\, dy$, for some weight function $w : \mathbb{R} \mapsto \mathbb{R}^+$ and some $\gamma > 0$;
- $D_\infty(F_0, F_1) := \sup_y \big|(F_0 - F_1)(y)\big|$ (Kolmogorov-Smirnov);

- $D_{TV}(F_0, F_1) := \sup_{A \in \mathcal{B}} \left|F_0(A) - F_1(A)\right| = \int |f_1 - f_0|/2(y)dy$ (Total variation), where $\mathcal{B}$ denotes the Borel subsets on the real line;
- $D_{Hel}(f_0, f_1) = \int \left(\sqrt{f_0} - \sqrt{f_1}\right)^2(y)dy$ (Hellinger); etc.

Note that diverse notions of contrasts/divergences/dissimilarities (Bregman, Kullback-Leibler, etc) can be invoked too, instead of D, even if they are not always true distances. The natural counterparts for F_k are the empirical distributions $\hat{F}_0(t) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(V_i \leqslant t)$, $\hat{F}_1(t) = m^{-1} \sum_{i=n+1}^{n+m} \mathbf{1}(V_i \leqslant t)$, for every t. In the case of densities, there are more competitors, but the simplest ones are the Nadaraya-Watson kernel estimators $\hat{f}_0(t) = \dfrac{1}{nh_n} \displaystyle\sum_{i=1}^{n} K\big(\frac{V_i - t}{h_n}\big)$, $\hat{f}_1(t) = \dfrac{1}{mh_m} \displaystyle\sum_{i=n+1}^{n+m} K\big(\frac{V_i - t}{h_m}\big)$, for some bandwidth sequence (h_n), $h_n > 0$, $h_n \to 0$ when $n \to \infty$, and K denotes a univariate kernel function i.e. a real map s.t. $\int K = 1$. For instance, an estimator of $D_\gamma(F_0, F_1)$ would be $\hat{D}_\gamma(F_0, F_1) := \int \left(\hat{F}_0 - \hat{F}_1\right)^\gamma(y) w(y)\, dy$. The consistency of the estimated distances is insured when the estimators $\hat{F}_k$ and/or $\hat{f}_k$ are uniformly convergent on the w support. Therefore, under $\mathcal{H}_0^*$, $\hat{D}_\gamma(F_0, F_1)$ tends to zero a.s. due to Glivenko-Cantelli theorem. Moreover, $(n+m)^{\gamma/2} \hat{D}_\gamma(F_0, F_1)$ is weakly convergent when n and m both tend to the infinity "at the same rate" because of the weak convergence of empirical processes (Donsker theorem). The theory of the two-sample Kolmogorov–Smirnov test is detailed in Van der Vaart and Wellner(1996), Section 3.7. Other traditional two-sample goodness-of-fit tests based on empirical distribution functions include the Cramer–von Mises and Anderson–Darling ones (Anderson and Darling, 1954; Pettitt, 1976). In the case of kernel densities, a similar result applies under some conditions of regularity for the underlying law, the kernel and for some range of the bandwidth sequences (see Einmahl and Mason, 2005 e.g.). Anderson et al. (1994) and Li (1996) studied different empirical counterparts of $D_2(f_0, f_1)$, e.g. The former authors directly considered $\int (\hat{f}_1 - \hat{f}_0)^2$, when the latter one studied $\int \hat{f}_1(t)\, \hat{F}_1(dt) + \int \hat{f}_0(t)\, \hat{F}_0(dt) - \int \hat{f}_1(t)\, \hat{F}_0(dt) - \int \hat{f}_0(t)\, \hat{F}_1(dt)$. Interestingly, this latter statistic is virtually the same as those recently proposed in Gretton et al. (2012) in a Reproducing Hilbert Kernel Space (RKHS) framework. For instance, Cao and Van Keilegom (2006) considered also tests for the two-sample problem with empirical likelihood techniques based on kernel density estimates. Beside cdfs' and densities, note that it is straightfor-

ward to build a similar non-parametric approach by comparing the empirical characteristic functions associated to the two samples, Epps and Singleton (1986). Nonetheless, most of the limiting laws of the previous test statistics are complex. The calculation of critical values has to be led by resampling techniques most of the time, particularly by bootstrap (Mammen 2012, Section 3).

Remark 7.1. *To discriminate between different testing procedures, an analysis of local power is most often useful, i.e. of the ability of a test to detect local departures from the null hypothesis. For the sake of illustration, the test based on* $D_2(\hat{f}_0, \hat{f}_1)$ *is still consistent against (a sequence of) alternatives as* $\mathcal{H}_1 : f_1 = f_0 + \delta_{n,m} g$ *where* $(\delta_{n,m})$ *is an array of real numbers s.t.* $\delta \asymp (n+m)^{-1/2} h_{n+m}^{-d/2}$*, and* g *denotes any integrable function chosen to ensure that* f_1 *is a valid density (*$\int g = 0$ *simply). Details are in Gretton et al. (2012). Other reasonings in terms of local power can be found in (Pagan and Ullah 1999, Section 2.9).*

After this presentation of several tests for fairness based on distributions, we are interested to see what can be done in the case of covariates and the test of the so-called "conditional fairness" property ? Indeed, under such circumstances, testing all the previous fairness concepts is similar to testing a conditional independence property, when one of the variables is discrete. Therefore, $\mathcal{H}_0^*$ is strictly equivalent to

$$\mathcal{H}_0 : V \text{ is independent of } S \text{ given } W,$$

sometimes simply rewritten $\mathcal{H}_0 : V \perp\!\!\!\perp S \,|\, W$, with the same notations as in Section 6.3.

Since many tests for conditional independence have been proposed in the literature, it is tempting to build a bridge towards such techniques/proposals. We will focus on general "omnibus" tests of conditional independence, that previously defined for arbitrary random vectors (V, W, S), which may have continuous and/or discrete components. For instance, Linton and Gozalo (2014) proposed a general procedure to discretize the spaces of (V, W, S) realizations to circumvent the curse of dimensionality. The latter general test has been recently revisited by Mittag (2018) when dealing with a discrete variable S. There are many competitors in this stream of the literature, where authors invoked diverse distances between

distributions, many techniques and testing procédures of $\mathcal{H}_0$. For example, Huang (2010) proposes tests for conditional independence using maximal nonlinear conditional correlation. Huang et al. (2013) introduce a nonparametric test for conditional independence based on an estimator of the topological "distance" between restricted and unrestricted probability measures corresponding to conditional independence or its absence, respectively. Su and White (2008) propose a nonparametric test based on density functions and the weighted Hellinger distance. Su and White (2012) propose a nonparametric test for conditional independence using local polynomial quantile regression, and Su and White (2014) construct a class of smoothed empirical likelihood-based tests for $\mathcal{H}_0$. Liu et al. (2018) and Su and Spindler (2019) compare different smoothed conditional cdfs' through a Cramer von-Mises distance. Su and White (2007), Wang et al. (2015), Wang and Hong (2018) prefer to work with some distances between smoothed conditional empirical characteristic functions.[6]

Testing for conditional independence is generally a challenging problem due to the curse of dimensionality (Bergsma 2004). All the previously cited papers rely on kernel smoothing and potentially suffer from this curse of dimensionality when $d > 2$. Nonetheless, in some particular cases, the latter testing procedures do not suffer from this problem (reduced rates of convergence, requirement of huge datasets). This is obviously the case when the number of continuous components of W is reasonable (i.e. not larger than three), and when, simultaneously, the discrete components of W are not too numerous, with relatively few items. Another possitibility is to assume that the classifier/regressor V is a map from a few "indices" $\gamma_k' W$, $k = 1, \ldots, r$, where r is less than three typically. In other words $V = R(\gamma_1' W, \ldots, \gamma_r' W)$. When $r = 1$, this is the classical single-index situation. Then, instead of smoothing w.r.t. W, an adapted testing procedures would rely on smoothing w.r.t. the vector of indices only, whose dimension is a lot smaller in general. In this vein, to illustrate, Song (2009) proposed tests of conditional independence of two random variables given a single-index involving an unknown finite dimensional parameter through Rosenblatt transforms. Therefore, it is here sufficient to invoke univariate kernel estimates, even if the

[6] Other references are Bergsma (2004), Huang et al. (2007), de Matos and Fernandes (2007), Geenens and Simar (2010), Bouezmarni et al. (2012), Bouezmarni and Taamouti (2014), Liu et al. (2018), Zhou et al. (2020) too, among others.

dimension of W is large.

More recently and with the develoment of Reproducing Kernel Hilbert Spaces (RKHS) methods, some new distances between distributions and conditional distributions have been successfully introduced: see Muandet et al. (2017) and the references therein. In particular, Fukumizu et al. (2008) propose a measure of conditional dependence of random variables, based on normalized cross-covariance operators on RKHS spaces. Such a measure is zero if and only if $\mathcal{H}_0$ is satisfied (not only "if"!) and it can easily be estimated. This allows the building of a consistent nonparametric tests of conditional independence. Alternatively Doran et al. (2014) calibrate an "optimal" permutation of the values of S (or V, if discrete) so that their laws remain approximately unchanged given the covariate W. Afterwards, a two sample test can be invoked to check whether such a permutation has disturbed the joint law of (V, S, W). If this is the case, $\mathcal{H}_0$ would be rejected. At the second stage, the kernel-based two sample test of Gretton et al. (2012) is invoked. Such techniques are less prone to the curse of dimensionality and admits fast convergence in terms of sample size, Grünewälder et al. (2012). On the negative side, some finite distance bounds and the asymptotic behavior of such RKHS-based test statistics have not been stated yet. Even consistent, it is not clear whether such statistics are weakly convergent under $\mathcal{H}_0$. Finding their critical values may be uncertain.

Remark 7.2. *Actually, in our case, a key feature comes from the fact the variables S is always discrete. Moreover, V is discrete too in a lot of interesting situations. Therefore, our problem of conditional independence is equivalent to a two sample problem for conditional distributions: denote by $\mu_{V|w,0}$ (resp. $\mu_{V|w,1}$) the laws of V given $(W = w)$ and $S = 0$ (resp. $S = 1$). Then, still in the case of conditional fairness, $\mathcal{H}_0^*$ (or $\mathcal{H}_0$, equivalently) may be rewritten as*

$$\mathcal{H}_0^* : \mu_{V|W=w,S=0} = \mu_{V|W=w,S=1} \text{ for every } w \in \mathbb{R}^q.$$

Surprisingly, we are not aware of papers that directly tackle such a testing problem, instead of relying on "omnibus" conditional indepdence procedures.

7.2 A simple testing procedure in the case of many covariates

All the testing statistics that have been quickly reviewed above can be applied in the case of *demographic parity*, equalized odds and lack of disparate mistreatment. Nonetheless, in the case of *conditional fairness*, predictors involve many covariates in general. As a consequence, we are expsoed to the curse of dimensionality and usual smoothing techniques cannot be applied in practice. Thus, as in Linton and Gozalo's (2014), we promote the idea of discretizing the space of W (that includes our so-called covariates Z in the case of conditional fairness). Consider a family of subsets $\mathcal{A} := \{A_1, \ldots, A_m\}$ in $\mathbb{R}^q$, typically a partition of the latter space. It is tempting to build a test of

$$\bar{\mathcal{H}}_0 : f_{V|W\in A_k,S=0} = f_{V|W\in A_k,S=1} \text{ for every } k = 1, \ldots, m, \text{ or equivalently}$$

$$\bar{\mathcal{H}}_0 : f_{V|W\in A_k,S=0} = f_{V|W\in A_k} \text{ for every } k = 1, \ldots, m.$$

Indeed, if we consider many "boxes" A_k that shrink towards singletons $\{z\}$ in $\mathbb{R}^q$, and if it can be done for many values of z, $\bar{\mathcal{H}}_0$ would be reduced to $\mathcal{H}_0^*$. Unfortunately, this "shrinking procedure" is unfeasible for obvious reasons. Moreover, $\mathcal{H}_0$ may be true when $\bar{\mathcal{H}}_0$ is not. This can be checked easily, because $\bar{\mathcal{H}}_0$ is equivalent to

$$\int_{A_k} f_{V|W=w,S=0}(v)\frac{\mathbb{P}(S=0|W=w)}{\mathbb{P}(S=0|W\in A_k)}f_W(w)\,dw = \int_{A_k} f_{V|W=w}(w)f_W(w)\,dw, \tag{7.1}$$

for every $k = 1, \ldots, m$. Under $\mathcal{H}_0$, the relationship (7.1) (and then $\bar{\mathcal{H}}_0$) is satisfied too if we have

$$\mathbb{P}(S=0|W=w) = \mathbb{P}(S=0|W\in A_k) \text{ for every } w \in A_k, k = 1, \ldots, m, \tag{7.2}$$

a relatively strong requirement. The latter condition could be approximately enforced by conveniently building the family of subsets $\mathcal{A}$. To this goal, we propose to build a tree, that allocates any observation W to a class $S \in \{0, 1\}$. The goal of such a classifier is to make disjoint boxes A_k (the leafs of a tree, possibly after grouping/prunig). Such boxes have to be as much homogenous as possible, relative to S. In other words, in any subset A_k, we expect that the vast majority of points W_i in A_k will be allocated to the same feature S.

This means that the variability of the map $w \mapsto \mathbb{P}(S = 0|W = w)$ will be weak, when $w \in A_k$. We hope that such maps will take relatively large (resp. small) values for every $w \in A_k$, depending on the considered subset A_k.

Therefore, let us propose a simple nonparametric test of $\bar{\mathcal{H}}_0$ that will be relevant to test $\mathcal{H}_0$, once the family $\mathcal{A}$ hase been conveniently chosen. Many tests are possible, replacing the theoretical conditional probabilities by empirical counterparts. For instance, a Kolmogorov-Smirnov-type test may be defined as

$$KS_n := \sup_v \sup_k \Big|\mathbb{P}_n(W \in A_k)\mathbb{P}_n(V \leqslant v, W \in A_k|S = 0) - \mathbb{P}_n(W \in A_k|S = 0)\mathbb{P}_n(V \leqslant v, W \in A_k)\Big|,$$

$$\mathbb{P}_n(V \leqslant v, W \in A_k|S = 0) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(V_i \leqslant v, W_i \in A_k),$$

$$\mathbb{P}_n(V \leqslant v, W \in A_k) := \frac{1}{n+m}\sum_{i=1}^{n+m} \mathbf{1}(V_i \leqslant v, W_i \in A_k),$$

$$\mathbb{P}_n(W \in A_k|S = 0) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(W_i \in A_k),\ \mathbb{P}_n(W \in A_k) := \frac{1}{n+m}\sum_{i=1}^{n+m} \mathbf{1}(W_i \in A_k).$$

Obviously, it is possible to build a similar statistic replacing the event $(S_i = 0)$ by $(S_i = 1)$ everywhere. Moreover, we could replace the supremum over $k = 1, \ldots, m$ by a sum over all boxes.

In the case of a discrete variable $V \in \{0, 1, \ldots, p\}$, a well-suited test statistic is

$$\widetilde{KS}_n := \sup_{j=1,\ldots,p} \sup_k \Big|\mathbb{P}_n(W \in A_k)\mathbb{P}_n(V = j, W \in A_k|S = 0) - \mathbb{P}_n(W \in A_k|S = 0)\mathbb{P}_n(V = j, W \in A_k)\Big|,$$

$$\mathbb{P}_n(V = j, W \in A_k|S = 0) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(V_i = j, W_i \in A_k),$$

$$\mathbb{P}_n(V = j, W \in A_k) := \frac{1}{n+m}\sum_{i=1}^{n+m} \mathbf{1}(V_i = j, W_i \in A_k).$$

7.3 The case of pure discrete covariates

In the particular case of discrete covariates, our so-called vector $W \in \mathbb{R}^d$ takes only a finite number of values and smoothing is non longer necessary in theory. To build a test of $\mathcal{H}_0$, a

simple approach is to consider as many bins as the number N of modalities of W. To fix the ideas, it is possible to lead a test of conditional independence

$$\mathcal{H}_{0,k} : V \perp\!\!\!\perp S \,|\, W = k, k = 1, \ldots, N \text{ in every bin.}$$

For instance by a simple chi-square test of independence. Under $\mathcal{H}_{0,k}$, assume a test statistic $\mathcal{T}_k$ weakly tends to a distribution μ_k. If we lead the same testing procedure in every bin and if our test statistics $\mathcal{T}_k$ are distribution-free, then μ_k does not depend on k and is simply denoted by μ. Now, to test $\mathcal{H}_0$, there remains a problem of aggregation because there are potentially many bins N and many test statistics. In the case of purely discrete binary variables V and W, Chiappori and Salanié (2000) were facing the same problem. They proposed three ways to aggregate these test statistics to obtain three overall test statistics for conditional independence. Here, we recall and adapt their simple proposals in our setting:

(i) build a Kolmogorov-Smirnoff test statistic that compares the empirical distribution function of the test statistics $\mathcal{T}_k$, $k = 1, \ldots, N$ with the cdf of the distribution μ;

(ii) alternatively, count the number of rejections for the independence test for each cell and a given (identical) significance level α. It is asymptotically distributed as binomial $Bin(N, \alpha)$ under the null;

(iii) add up all the test statistics for each individual cell. Under the null, the resulting r.v. $\mathcal{T}_{tot}$ is asymptotically distributed as the sum of N mutually independent variables that follows the same law μ. In the particular case of chi-square tests, $\mathcal{T}_k \sim \chi^2(1)$ and $\mathcal{T}_{tot} \sim \chi^2(N)$. This is inline with the classical conditional independence tests in the discrete setting through linear combination of chi-squared test distributions (Agresti 1992).

In case (iii), when the test statistics $\mathcal{T}_k$ are based on the squared L^2 distance between the joint distribution of (V, S) and the product of its margins, a particular weighted average of these $\mathcal{T}_k$ is optimal in terms of sample complexity (Canonne et al. 2018). In the case of general divergence functions and contingency tables, Taniechi et al. (2019) recently pro-

posed to approximate the distributions of chi-squared type test statistics through asymptotic expansions.

Moreover, it has to be stressed that all "omnibus" tests of conditional independence can be rewritten in the particular case of discrete variables, particularly discrete W. Indeed, even if it is no longer necessary to smooth the distribution of W or the distibution of V given W, these test statistics can be formally written and calculated in this particular case. And the choice of smoothing parameters (bandwidths for kernel estimators, number of neighbors for k-Nearest Neighbors, etc) does no longer matter. Nonetheless, the definition of such omnibus test statistics, the associated estimators and their asymptotic theory, have not been "fine tuned" for the particular case of discrete variables most of the time. There is an (most often implicit and hidden) overlap between this stream of the literature and some measures of association that have been historical proposed for discrete variables (Read and Cressie (2012), e.g.). For example, the "normalized cross-covariance" between V and S (no covariate), as introduced in Fukumizu et al. (2008), is reduced to the so-called mean square contingency (also called ϕ-coefficient), that is a commonly used as a dependence measure between discrete distribtions. For the sake of illustration, the particular version of W of our Kolmogorov-Smirnov type statistic KS_n of Section 7.2 would become

$$KS_n := \sup_v \sup_k \Big| \mathbb{P}_n(W = w_k)\mathbb{P}_n(V \leqslant v, W = w_k | S = 0) - \mathbb{P}_n(W = a_k | S = 0)\mathbb{P}_n(V \leqslant v, W = a_k) \Big|,$$

when $W \in \{w_1, \ldots, w_N\}$

8 Empirics

Let us evaluate the numerical performances of our "fair bagging" classifiers and regressors on a simulated experiment. We will consider the family of binary and continuous models $Y = g_\theta(X)$ for some finite dimensional parameter θ, and $X = (Z_1, Z_2, S)$ with the same notations as before: $S \in \{0, 1\}$ is the sensitive feature, $Z_1 \in \mathbb{R}$ is a (potential) continuous or discrete conditional feature of interest for testing conditional fairness, and $Z_2 \in \mathbb{R}^{p-1}$ is stacking the other underlying explanatory variables.

To fix the ideas, our Data Generating Process will be a mixture of Gaussian random variables:

$$h_\theta(X) = (1-S)\mathcal{N}\big(m_0(X), \sigma_0^2(X)\big) + S\mathcal{N}\big(m_1(X), \sigma_1^2(X)\big),$$

$$m_k(X) = a_k'[1, Z_1, Z_2],\ \sigma_k^2(X) = \exp(b_k'[1, Z_1, Z_2]),\ k \in \{0, 1\},$$

for some vector of coefficients a_k and b_k in $\mathbb{R}^{p+1}$. In other words, $\theta = (a_0, a_1, b_0, b_1)$. The sensitive feature S will simply be drawn through a Bernoulli r.v. with parameter $1/2$, independently of X and the gaussian random variables.

When Y is binary, set $Y = g_\theta(X) = \mathbf{1}(h_\theta(X) > 0)$. When Y is continuous, simply set $Y = g_\theta(X) = h_\theta(X)$. Our competitors will be

- usual classifiers as logistic, classification trees (CART) and SVM;
- usual regressors as k-Nearest Neighbors, kernel regression, SVM;
- some so-called "fair" classifiers/regressors as proposed in the literature: Hardt et al. (2016), Zafar et al. (2017), Donini et al. (2018), and possibly others.

All the latter competitors will try to predict Y from the available information, i.e. from X, including S possibly. Note that, when $X = S$ only, such classifiers will be unfair by construction. When $X = (Z_1, Z_2)$ but does not include S, this lack of fairness will be less obvious. It will depend on the way the vector Z will be drawn, mainly whether the DGP of Z is independent or not of S (to be measured by some indicators as detailed below). Our "corrected" classifier will be those proposed in Section 4, i.e. given by (4.2), (4.3) and (4.4).

See Donini et al. (2018) for a comparable simulation scheme. See Friedler et al. (2019) too. Interestingly, instead of starting from X and then generating Y, Donini et al. (2018) do the opposite: given the value of Y (a binary variable) and S, they propose to generate Z from different bivariate Gaussian vectors. For instance, given $(Y, S) = (1, 1)$, the law of Z is drawing along a $\mathcal{N}\big((-1, -1), 0.8I_2\big)$. Given $(Y, S) = (1, 0)$, the law of Z is drawing along a $\mathcal{N}\big((1, 1), 0.8I_2\big)$. Given $(Y, S) = (0, 1)$ (resp. $(Y, S) = (0, 0)$), $Z \sim \mathcal{N}\big((0.5, -0.5), 0.5I_2\big)$ (resp. $Z \sim \mathcal{N}\big((0.5, 0.5), 0.5I_2\big)$). In such a case, realizations of the bivariate vector Z are

clearly dependent on S and there is some hope to build performing (unfair) classifiers that only use the information Z (but not S) as inputs.

For every experiment, we draw a n sample $\mathcal{S}_n := (X_i, Y_i)_{i=1,\dots,n}$ of i.i.d. data; every classifier is estimated on $\mathcal{S}_n$, including ours (through the estimation of ω as in Section 4). Their predictive powers will be evaluated on another sample $\mathcal{S}_{n'} := (X_i, Y_i)_{i=n+1,\dots,n+n'}$ of n' data. Note that we do not try to estimate θ because the functional link between X and S is unknown for modellers. Predicted values $\hat{Y}_i$ are then available for measuring in-sample (resp. out-of-sample) accuracy with $\mathcal{S}_n$ (resp. $\mathcal{S}_{n'}$). In the case of binary outcomes, classical measures of accuracy are the True Positive Rate (resp. True Negative Rate) $\mathbb{P}(\hat{Y} = 1|Y = 1)$ (resp. $\mathbb{P}(\hat{Y} = 0|Y = 0)$). In this study, we consider their simple average, the so-called Balanced Classification Rate

$$BCR := \frac{1}{2}\big(\mathbb{P}(\hat{Y} = 1|Y = 1) + \mathbb{P}(\hat{Y} = 0|Y = 0)\big) \in [0, 1],$$

that can easily be estimated empirically. When dealing with conditinuous outcomes, we propose to generalize such a measure as

$$BCR_c := \frac{1}{q}\sum_{k=1}^{q} \mathbb{P}(\hat{Y} \in A_k|Y \in A_k) \in [0, 1],$$

for a given partition $\mathcal{A} := \cup_{k=1}^{q} A_k$ of the real line. For example, set $q = 10$ and $A_k = (q^Y_{k-1}, q^Y_k)$, denoting by q^Y_k the k-th decile of Y (to be empirically estimated) for $k \in \{1, \dots, 9\}$, $q_0 = -\infty$ and $q_{10} = +\infty$.

Many measures of fairness may be chosen, depending on the concepts we consider. Let us work with the following ones:

- in the case of demographic parity, the so-called "disparate impact" (Feldman et al. 2015) $DI_{dp} := \mathbb{E}(\hat{Y}|S = 0)/\mathbb{E}(\hat{Y}|S = 1)$ for binary and continuous outcomes;
- in the case of equalized odds (equal opportunity, to be specific), set

$$DI_{eo} := \frac{1}{2}\Big\{\frac{\mathbb{P}(\hat{Y} = 1|S = 0, Y = 1)}{\mathbb{P}(\hat{Y} = 1|Y = 1)} + \frac{\mathbb{P}(\hat{Y} = 1|S = 0, Y = 0)}{\mathbb{P}(\hat{Y} = 1|Y = 0)}\Big\},$$

for binary outcomes. Dealing with continuous Y, we propose the measure of "mean L^1 fairness"

$$MLF := 1 - \mathbb{E}_Y\big[|\mathbb{E}(\hat{Y}|S=0,Y) - \mathbb{E}(\hat{Y}|Y)|\big]/\mathbb{E}(Y).$$

Note that the latter measure may not belong to $[0,1]$ in general, even if $MLF = 1$ in case of perfect fairness.

- in the case of lack of disparate mistreatment, set $DI_{ldm} := \mathbb{E}\big[|\hat{Y} - Y||S=0\big]/\mathbb{E}\big[|\hat{Y} - Y||S=1\big]$ for binary and continuous outcomes;

Note that calibration procedures of our "fair" classifiers/predictors will be led with $\mathcal{S}_n$, as detailed in Section 4. All the latter measures of fairness will be empirically evaluated on the testing dataset $\mathcal{S}_{n'}$. For instance, we will calculate estimated disparate impact of binary outcomes as $\widehat{DI} := \mathbb{P}_{n'}(\hat{Y}=1|S=0)/\mathbb{P}_{n'}(\hat{Y}=1|S=1)$, denoting by $\mathbb{P}_{n'}$ the empirical measure associated to $\mathcal{S}_{n'}$. With continuous variables, it is necessary to rely on estimators of regression functions. For instance, an empirical counterpart of MLF may be

$$\widehat{MLF} := 1 - \sum_{j=n+1}^{n+n'} |\hat{r}_0(Y_j) - \hat{r}(Y_j)| / \sum_{j=n+1}^{n+n'} \hat{Y}_j,$$

$$\hat{r}_0(y) = \frac{\sum_{i=n+1}^{n+n'} \hat{Y}_i \mathbf{1}(S_i=0) K_h(y-Y_i)}{\sum_{i=n+1}^{n+n'} \mathbf{1}(S_i=0) K_h(y-Y_i)},\ \hat{r}(y) = \frac{\sum_{i=n+1}^{n+n'} \hat{Y}_i K_h(y-Y_i)}{\sum_{i=n+1}^{n+n'} K_h(y-Y_i)}.$$

To deal with conditional fairness, we can rely on the previous measures. Indeed, all of them can be calcuted for every fixed value of $Z_1 = z$, after localization (by kernel smoothing, for instance). To illustrate, the conditional disparate impact for demographic parity is defined as $DI^c_{dp}(z) := \mathbb{E}(\hat{Y}|S=0,Z_1=z)/\mathbb{E}(\hat{Y}|S=1,Z_1=z)$ for binary and continuous outcomes; it will be estimated as $\widehat{DI}^c_{dp}(z) := \hat{r}^Z_0(z)/\hat{r}^Z_1(z)$, where

$$\hat{r}^Z_k(z) = \frac{\sum_{i=n+1}^{n+n'} \hat{Y}_i \mathbf{1}(S_i=k) K_h(z-Z_i)}{\sum_{i=n+1}^{n+n'} \mathbf{1}(S_i=k) K_h(z-Z_i)},\ k \in \{0,1\}.$$

If the conditioning variable Z_1 is discrete and have integer values, the latter formulas still apply by replacing $K_h(z-Z_i)$ by $\mathbf{1}(Z_i = z)$ (this can be numerically obtained with $h << 1$). To obtain a global picture, we will calculate averages of the latter measuress w.r.t. the

distribution of the covariate Z_1. This yields

$$DI_{cdp}^{c,tot} := \mathbb{E}_{Z_1}\big[DI_{cdp}(Z)\big] \simeq \widehat{DI}_{dp}^{c,tot} := \frac{1}{n'}\sum_{i=n+1}^{n+n'} \widehat{DI}_{cdp}(Z_i).$$

Once some classifiers are combined to impose fairness and to build a new classifier, what would be the prediction performances of the latter one, compared to the former ones ? It is likely such performances will be "average", compared to the initial set of predictors. This will be empirically checked: for every model, calculate the prediction error on the testing subset. Check that this error for our fair classifier belongs to the convex hull of the prediction errors of the initial classifiers. If this not the case, this necessitates investigation.

We will work with $n = n' = 1000$ and resample 100 times our dataset of size $n + n'$. We will consider

(a) models without any covariate Z (i.e. $p = 0$ and $X = S$). We will illustrate the results for a reference value θ_a^* for which $a_0 = 1$, $a_1 = (-1)$, $b_0 = b_1 = 0$;

(b) models for which there is a single variable $Z = Z_1$ (i.e. $p = 1$ and $X = (Z_1, S)$). Moreover, S and Z_1 will be dependent: $Z_1 \sim \mathcal{N}(S, 1)$. Here, the reference value will be θ_b^* for which $a_0 = (1, 1)$, $a_1 = (-1, 1)$, $b_0 = b_1 = (0, 1)$.

Afterwards, these experiments will be made for a grid value of parameters θ and the measures of accuracy/fairness will be averaged over all these values, to obtain so-called "aggregated" measures.

It will be interesting to apply the latter methdology on real datasets, in particular to deal with the case of numerous variables Z. In the latter case, we will be able to illustrate the test proposed in Section 7.2.

Acknowledgements

Jean-David Fermanian has been supported by the labex Ecodec (reference project ANR-11-LABEX-0047).

References

[1] Agarwal A, Beygelzimer A, Dudik M, Langford J, Wallach H (2018) A reductions approach to fair classification. arXiv preprint arXiv:1803.02453.

[2] Agarwal A, Dudík M, Wu ZS (2019) Fair Regression: Quantitative Definitions and Reduction-based Algorithms. arXiv preprint arXiv:1905.12843.

[3] Agresti A (1992) A survey of exact inference for contingency tables. *Statistical science*. 7(1): 131-153.

[4] Anderson TW, Darling DA (1952) Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes. *Annals of Mathematical Statistics*. 23: 193-212.

[5] Anderson NH, Hall P, Titterington DM (1994) Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*. 50(1): 41-54.

[6] Behrens WU (1929) Ein Betrag zur fehlerberechnung bei wenigen beobachtungen, *Landwirtschaftliche Jahrbucher*. 68: 807 - 837.

[7] Bergsma WP (2004) Testing conditional independence for continuous random variables. Preprint 2004-048, Eurandom, Eindhoven.

[8] Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2018) Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods and Research*. On-line publication, https://doi.org/10.1177/0049124118782533

[9] Biau G, Scornet E (2015) A Random Forest Guided Tour. arXiv preprint arXiv:1511.05741v1.

[10] Bogroff A, Guégan D (2019) Artificial intelligence, Data, Ethics: an holistic approach for risks and regulation. WP HaL, http://ideas.repec.org/p/hal/cesptp/halshs-02181597.html

[11] Bouezmarni T, Rombouts JV, Taamouti A (2012) Nonparametric copula-based test for conditional independence with applications to Granger causality. *Journal of Business & Economic Statistics*. 30(2): 275-287.

[12] Bouezmarni T, Taamouti A (2014) Nonparametric tests for conditional independence using conditional distributions. *Journal of Nonparametric Statistics*. 26(4): 697-719.

[13] Breiman L (1996) Bagging predictors. *Machine Learning*. 140: 26-123.

[14] Breiman L (2001) Random Forest. *Machine Learning*. 45: 5-32.

[15] Buolamwini J, Gebru T (2018) Gender Shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency* (FAT 2018, New York, USA), 81, 77-91.

[16] Calders T, Kamiran F, Pechenizkiy M (2009) Building classifiers with independency constraints. *2009 IEEE International Conference on Data Mining Workshops*, 13-18.

[17] Calders T, Verwer S (2010) Three naive Bayes approaches for discrimination-free classification. *datA Mining and Knowledge Discovery*. 21 (2): 277-292.

[18] Calmon F, Wei D, Vinzamuri B, Ramamurthy KN, Varshney KR (2017) Optimized preprocessing for discrimination prevention. *Advances in Neural Information Processing Systems*, 3992-4001.

[19] Canonne CL, Diakonikolas I, Kane DM, Stewart A (2018, February). Testing conditional independence of discrete distributions. *2018 Information Theory and Applications Workshop (ITA)*, 1-57.

[20] Cao R, Van Keilegom I (2006) Empirical likelihood tests for two?sample problems via nonparametric density estimation. *Canadian Journal of Statistics*. 34(1): 61-77.

[21] Chiappori PA, Salanie B (2000) Testing for asymmetric information in insurance markets. *Journal of political Economy*. 108(1): 56-78.

[22] Chouldechova A, Benavides-Prado D, Fialko O, Vaithianathan R (2018) A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *ACM Conference on Fairness, Accountability, and Transparency* (FAT 2018), 134-148.

[23] Chzhen E, Denis C, Hebiri M, Oneto L, Pontil M (2020) Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees. Hal-02501190, https://hal.archives-ouvertes.fr/hal-02501190.

[24] Cui Z, Chen W, He Y, Chen Y (2015) Optimal action extraction for random forests and boosted trees. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (ACM 2015), 179-188

[25] De Matos JA, Fernandes M (2007) Testing the Markov property with high frequency data. *Journal of Econometrics*. 141(1): 44-64.

[26] Dehling H, Fried R (2012) Asymptotic distribution of two-sample empirical U-quantiles with applications to robust tests for shifts in location. *Journal of Multivariate Analysis*. 105(1): 124-140.

[27] Delgado MA, González Manteiga W (2001) Significance testing in nonparametric regression based on the bootstrap. *The Annals of Statistics*. 29(5): 1469-1507.

[28] Derrick B, Toher D, White P (2016) Why Welch's test is type I error robust, *The quantitative methods for Psychology*. 12(1): 30-38.

[29] Donini M, Oneto L, Ben-David S, Shawe-Taylor JS,. Pontil M (2018) Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 2791-2801.

[30] Doran G, Muandet K, Zhang K, Schölkopf B (2014) A Permutation-Based Kernel Conditional Independence Test. *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence* (UAI 2014, Quebec City, Quebec, Canada), 132-141.

[31] Dwork C, Hardt Y, Pitassi T, Reingold O, Zemel R (2012) Fairness Through Awareness. *Proceedings of the 3rd Innovations of Theoretical Computer Science*, 214 -226.

[32] Einmahl U, Mason DM (2005) Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*. 33(3): 1380-1403.

[33] Epps TW, Singleton KJ (1986) An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*. 26(3-4): 177-203.

[34] Fay MP, Proschan MA (2010) Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*. 4: 1-39.

[35] Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259-268.

[36] Fisher RA (1935) The fiducial argument in statistical inference. *Annals of Eugenics*. 6(4): 391-398.

[37] Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 329-338.

[38] Friedman JH, Hall P (2007) On bagging and nonlinear estimation. *Journal of statistical planning and inference*. 1373(3): 669-683.

[39] Fukumizu K, Gretton A, Sun X, Schölkopf B (2008) Kernel measures of conditional dependence. *Advances in neural information processing systems*. 20: 489-496.

[40] Geenens G, Simar L (2010) Nonparametric tests for conditional independence in two-way contingency tables. *Journal of Multivariate Analysis*. 101(4): 765-788.

[41] Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A (2012) A kernel two-sample test. *Journal of Machine Learning Research* 13(Mar): 723-773.

[42] Grünewälder S, Lever G, Gretton A, Baldassarre L, Patterson S, Pontil M (2012) Conditional mean embeddings as regressors. arXiv preprint arXiv:1205.4656.

[43] Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 3315-3323.

[44] Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media, LLC).

[45] Heckman J, Ichimura H, Smith J, Todd P (1998) Characterizing Selection Bias Using Experimental Data. *Econometrica*. 66(5): 1017-1098.

[46] Huang TM (2010) Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*. 38(4): 2047-2091.

[47] Huang M, Sun Y, White H (2016) A flexible nonparametric test for conditional independence. *Econometric Theory*. 32(6): 1434-1482.

[48] Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. *Journal of economic literature*. 47(1): 5-86.

[49] Kamiran F, Calders T, Pechenizkiy M (2010) Discrimination aware decision tree learning. *2010 IEEE International Conference on Data Mining*, 869-874.

[50] Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*. 33(1): 1-33.

[51] Kamishima T, Akaho S, Asoh H, Sakuma J (2012). Fairness-aware classifier with prejudice remover regularizer. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Springer, Berlin, Heidelberg), 35-50.

[52] Kearns M, Neel S, Roth A, Wu ZS (2018) Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. arXiv preprint arXiv:1711.05144.

[53] Kim H, Kim H, Moon H, Ahn H (2011) A weight-adjusted voting algorithm for ensembles of classifiers. *Journal of the Korean Statistical Society*. 40(4): 437-449.

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

[55] Kotsiantis S, Pintelas P (2004) Combining bagging and boosting. *International Journal of Computational Intelligence*. 1(4): 324-333.

[56] Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. *Advances in Neural Information Processing Systems* (NeurIPS 2017), 4066-4076.

[57] Lee AJ (2019) *U-statistics: Theory and Practice* (Routledge).

[58] Lee S, Wang YJ (2009) Nonparametric tests of conditional treatment effects. WP Cemmap CWP36/09, UCL.

[59] Li Q (1996) Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews*. 15(3): 261-274.

[60] Linton O, Gozalo P (2014) Testing Conditional Independence Restrictions. *Econometric Reviews*. 33(5-6), 523-552.

[61] Liu Y, Wang Q, Liu X (2018) Testing conditional independence via integrating-up transform. *Statistics*. 52(4): 734-749.

[62] Lu Q, Cui Z, Chen Y, Chen X (2017) Extracting optimal actionable plans from additive tree models. *Frontiers of Compututational Science*. 11(1): 1-15.

[63] Lundberg, SM, Lee SI (2017) A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 4765–4774.

[64] Mammen E (2012) *When does bootstrap work?: asymptotic results and simulations* (Vol. 77, Springer Science & Business Media).

[65] Madras D, Creager E, Pitassi T, Zemel R (2018) Learning adversarially fair and transferable representations. arXiv preprint arXiv:1802.06309.

[66] Maudes J, Rodriguez JJ, Garcia-Osorio C, Garcia-Pedrajas N (2012). Random feature weights for decision tree ensemble construction. *Information Fusion*. 13(1): 20-30.

[67] Mittag N (2018) A Nonparametric k-Sample Test of Conditional Independence. Working paper of CERGE-EI, Prague, Czech Republic .

[68] Muandet K, Fukumizu K, Sriperumbudur B, Schölkopf B (2017) Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*. 10(1-2): 1-141.

[69] Pagan A, Ullah A (1999) *Nonparametric econometrics* (Cambridge University Press).

[70] Pedreschi D, Ruggieri S, Turini F (2009) Integrating induction and deduction for finding evidence of discrimination. *Proc. of Intl. Conf. on Artificial Intelligence and Law* (ICAIL '09), 157-166.

[71] Pettitt AN (1976) A two-sample Anderson-Darling rank statistic. *Biometrika*. 63(1): 161-168.

[72] Read TR, Cressie NA (2012) *Goodness-of-fit statistics for discrete multivariate data* (Springer Science & Business Media).

[73] Scheffé H (1970) Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association*. 65(322): 1501-1508

[74] Song K (2009) Testing conditional independence via Rosenblatt transforms. *The Annals of Statistics*. 37(6B): 4011-4045.

[75] Su L, Spindler M (2013) Nonparametric testing for asymmetric information. *Journal of Business & Economic Statistics*. 31(2): 208-225.

[76] Su L, White H (2007) A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*. 141(2): 807-834.

[77] Su L, White H (2008) A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*. 24(4): 829-864.

[78] Su L, White H (2012) Conditional independence specification testing for dependent processes with local polynomial quantile regression. *Advances in Econometrics*. 29: 355-434.

[79] Su L, White H (2014) Testing conditional independence via empirical likelihood. *Journal of Econometrics*. 182(1): 27-44.

[80] Taneichi N, Sekiya Y, Toyama J (2019) Transformed statistics for tests of conditional independence in $J \times K \times L$ contingency tables. *Journal of Multivariate Analysis*. 171: 193-208.

[81] Van der Vaart AW, Wellner JA (1996) *Weak convergence and empirical processes* (Springer, New York, NY).

[82] Wager S, Athey S.(2018) Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *JASA*. 113(523): 1228-1242.

[83] Wang X, Pan W, Hu W, Tian Y, Zhang H (2015) Conditional distance correlation. *Journal of the American Statistical Association*. 110(512): 1726-1734.

[84] Wang X, Hong Y (2018) Characteristic function based testing for conditional independence: a nonparametric regression approach. *Econometric Theory*: 34(4), 815-849.

[85] Watson DS, Wright MN (2019) Testing Conditional Predictive Independence in Supervised Learning Algorithms. arXiv preprint arXiv:1901.09917.

[86] Welch BL (1938) The significance or the difference between two means when the population variances are unequal, *Biometrika*. 29, 350-362.

[87] Willems G, Pison G, Rousseeuw PJ, Van Aelst S (2002) A robust Hotelling test. *Metrika*. 55: 125-138.

[88] Williamson RC, Menon AK (2019) Fairness risk measures. arXiv preprint arXiv:1901.08665.

[89] Yang O, Yin J, Ling CX, Chen T (2003) Postprocessing decision trees to extract actionable knowledge. *Third IEEE International Conference on Data Mining. IEEE*, 685-688.

[90] Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017) Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee*, 1171-1180

[91] Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. *International Conference on Machine Learning*, 325-333.

[92] Zhou Y, Liu J, Zhu L (2020) Test for conditional independence with application to conditional screening. *Journal of Multivariate Analysis*. 175(104557), 1-18.

[93] Zliobaite I (2015) On the relation between accuracy and fairness in binary classification. arXiv preprint arXiv:1505.05723.

[94] Woodworth B, Gunasekar S, Ohannessian MI, Srebro, N (2017) Learning non-discriminatory predictors, arXiv preprint arXiv:1702.06081.