Fairness without the Sensitive Attribute via Causal Variational Autoencoder

Vincent Grari[1,2*] , **Sylvain Lamprier**[1] , **Marcin Detyniecki**[2,3]
[1] Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
[2]AXA, Paris, France
[3] Polish Academy of Science, IBS PAN, Warsaw, Poland
{vincent.grari, sylvain.lamprier}@isir.upmc.fr, marcin.detyniecki@axa.com

Abstract

In recent years, most fairness strategies in machine learning have focused on mitigating unwanted biases by assuming that the sensitive information is available. However, in practice this is not always the case: due to privacy purposes and regulations such as RGPD in EU, many personal sensitive attributes are frequently not collected. Yet, only a few prior works address the issue of mitigating bias in this difficult setting, in particular to meet classical fairness objectives such as Demographic Parity and Equalized Odds. By leveraging recent developments for approximate inference, we propose in this paper an approach to fill this gap. To infer a sensitive information proxy, we introduce a new variational auto-encoding-based framework named SRCVAE that relies on knowledge of the underlying causal graph. The bias mitigation is then done in an adversarial fairness approach. Our proposed method empirically achieves significant improvement over existing works in the field. We observe that the generated proxy's latent space correctly recovers sensitive information and that our approach achieves a higher accuracy while obtaining the same level of fairness on two real datasets.

1 Introduction

Over the past few years, machine learning algorithms have emerged in many different fields of application. However, this development is accompanied with a growing concern about their potential threats, such as their ability to reproduce discrimination against a particular group of people based on sensitive characteristics (e.g., religion, race, gender, etc.). In particular, algorithms trained on biased data have been shown to be prone to learn, perpetuate or even reinforce these biases [Bolukbasi *et al.*, 2016], leading to numerous incidents being reported in recent studies [Angwin *et al.*, 2016; Lambrecht and E. Tucker, 2016]. To address this issue, there has been a growing interest for fair machine learning in the academic community, and a high variety of bias mitigation strategies have been proposed in the last decade [Zhang *et al.*, 2018; Adel *et al.*, 2019; Hardt *et al.*, 2016; Grari *et al.*, 2020b; Chen *et al.*, 2019; Zafar *et al.*, 2015; Celis *et al.*, 2019; Wadsworth *et al.*, 2018]. Currently, the vast majority of these state-of-the-art approaches rely on having access to the sensitive information to be mitigated during training (though sometimes encrypted as in [Veale and Binns, 2017; Kilbertus *et al.*, 2018]). However, in practice, it is often unrealistic to assume that this sensitive information is available or even collected. In Europe, for example, a car insurance company cannot ask a potential client about his/her origin or religion, as this is strictly regulated. Furthermore, in May 2018, the EU introduced the General Data Protection Regulation (GDPR), representing one of the most important changes in the regulation of data privacy in 20 years. It strictly regulates the collection and usage of sensitive personal data. Ignoring sensitive attributes as input of predictive models in order to achieve fairness is known as "fairness through unawareness" [Pedreshi *et al.*, 2008], but was shown to be insufficient since complex correlations in the data may provide unexpected links to sensitive information [Dwork *et al.*, 2012].

For this reason, some approaches have attempted to obtain a fair predictor model without the sensitive information. Most of them leverage the use of external data or prior knowledge on correlations [Zhao *et al.*, 2021; Madras *et al.*, 2018; Schumann *et al.*, 2019; Gupta *et al.*, 2018]. Others pursue fairness implicitly, by ensuring local smoothness in the decision function, rather than explicitly focusing on subgroups to be protected [Hashimoto *et al.*, 2018; Lahoti *et al.*, 2020].

To overcome limitations of these approaches, we propose a novel approach that leverages a causal graph to reconstruct sensitive information using Bayesian variational autoencoders (VaEs). The inferred information is then used as a proxy for mitigating biases in a adversarial fairness training setting. We empirically show experiments that this approach, based on sensitive reconstruction, is significantly more effective for achieving usual fairness objectives than its competitors, with a more direct control on mitigated biases.

2 Background and Related Work

In this paper, we consider training data which consists of n examples $(x_i, y_i)_{i=1}^n$, where $x_i \in \mathbb{R}^p$ is the feature vector of the i-th example and y_i its binary outcome. In our context the training sample x_i is decomposed into two feature vectors $x_{c_i} \in \mathbb{R}^{p_c}$ and $x_{d_i} \in \mathbb{R}^{p_d}$. In addition, we consider an -

[*]Contact Author

unobserved - binary sensitive attribute s_i for all i. We study fairness under the two following definitions.

Definition 1. ***Demographic Parity***: *A classifier is considered fair under the demographic parity criterion if the prediction $\widehat{Y}$ from features X is independent from the protected attribute S [Dwork* et al.*, 2012]. The underlying idea is that each demographic group has the same chance for a positive outcome. The p-rule assessment considers the likelihood ratio for the unprivileged group (the higher the more fair):*

$$\text{P-rule}(\widehat{Y}, S) = \min\left(\frac{P(\widehat{Y}=1|S=1)}{P(\widehat{Y}=1|S=0)}, \frac{P(\widehat{Y}=1|S=0)}{P(\widehat{Y}=1|S=1)}\right)$$

Definition 2. ***Equalized Odds***: *A classifier is considered fair according to this criterion if the outcome $\widehat{Y}$ has equal false positive rates and false negative rates for both demographics $S = 0$ and $S = 1$ [Hardt* et al.*, 2016]. A metric to assess this is the disparate mistreatment (DM) [Zafar* et al.*, 2015], which we report as the sum of the two following quantities:*

$$\Delta_{FPR} : |P(\widehat{Y}=1|Y=0, S=1) - P(\widehat{Y}=1|Y=0, S=0)|$$
$$\Delta_{FNR} : |P(\widehat{Y}=0|Y=1, S=1) - P(\widehat{Y}=0|Y=1, S=0)|$$

From the state-of-the-art literature, one possible way to achieve fairness despite the unavailability of sensitive attributes during training is to use transfer learning methods from external sources of data where the sensitive group labels are known. For example, [Madras *et al.*, 2018] proposed to learn fair representations via adversarial learning on a specific downstream task and transfer it to the targeted one. [Schumann *et al.*, 2019] and [Coston *et al.*, 2019] focus on domain adaptation. [Mohri *et al.*, 2019] considers an agnostic federated learning context by equalizing the performance of all participants through the lens of minimax optimization and fair resource allocation. However, this makes the actual desired bias mitigation highly dependent on the distribution of the external data. Other methods require prior knowledge on sensitive correlations. With prior assumptions, [Gupta *et al.*, 2018] and [Zhao *et al.*, 2021] mitigate the dependence of the predictions on the available features that are known to be likely correlated with the sensitive attribute. However, such strongly correlated features do not always exist in the data.

Finally, a few approaches address this objective without any prior knowledge on the sensitive information. Some of these works aim at improving the accuracy for the worst-case protected group (Rawlsian Max-Min objective) by leveraging techniques from distributionally robust optimization [Hashimoto *et al.*, 2018] or adversarial learning [Lahoti *et al.*, 2020]. Other works act on the input data using a cluster-based balancing strategy in order to minimize the biases locally [Yan *et al.*, 2020]. However, such methods are usually ineffective for traditional group fairness definitions such as *demographic parity* and *equalized odds*. Their blind way of mitigation affects non-sensitive information, likely implying a degradation of the predictor accuracy.

Our approach is inherently different from the aforementioned approaches. Based on minimal prior knowledge of causal relationships in the data, we perform Bayesian inference of latent sensitive proxies, whose dependencies with prediction outputs are mitigated in a second training step.

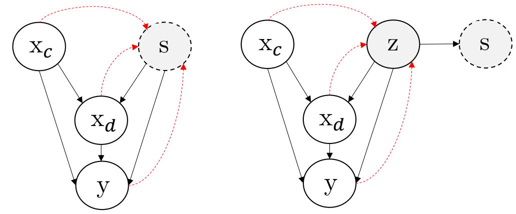


Figure 1: Causal graphs of SRCVAE: Left graph represents prior expert knowledge, where x is mapped into two components x_c and x_d. Right graph denotes the graph considered in our approach, with a multivariate confounder z inferred to be used as a proxy of the sensitive attribute s. Solid arrows denote causal links, red dashed arrows denote inference, grey circles denote missing attributes.

3 Methodology

In our approach, we first assume the existence and availability of a specific causal graph which underlies the training data, as discussed in subsection 3.1. The causal graph allows us to infer, through Bayesian inference, a latent representation containing as much information as possible about the sensitive feature. This process is described in subsection 3.2. Finally, we present in subsection 3.3 our methodology to mitigate fairness biases while preserving as much as possible prediction accuracy using this latent representation.

3.1 Causal Structure of SRCVAE

Our work relies on the assumption of having an underlying causal graph describing the data, where causal interactions are indicated as directed edges between subsets of features (nodes). We consider the training data

In particular, we suppose that the graph can be represented by the illustration shown in Figure 1. This structure is aimed to be generic enough to fit with most real world settings (slightly different graphs are studied in appendix). In the leftmost graph, parents of the output y are split into three components x_c, x_d and s. The subsets x_c and x_d regroup together all of the features that are given as input x to the model. The distinction between the two is made depending on the existence or absence of a causal relationship with the missing sensitive information s: no interaction is assumed with x_c, while some is with x_d. In addition, some causal relationship may exist between x_c and x_d.

To illustrate the generic aspect of this framework, we apply it to the Adult UCI dataset. The assumed causal graph of this dataset, with *Gender* as the sensitive attribute s and *Income* as the expected output y, is shown in Figure 2. In this context, x_c is the set of variables *Race*, *Age* and *Native_Country* which do not depend on the sensitive attribute, while x_d corresponds to all remaining variables that are generated from x_c and s (i.e., $x_d = \{Education, Work\_Class, ...\}$).

Assuming all of the variables except s are available, our purpose is to recover all the hidden information not caused by the set x_c but responsible of x_d and y. In a real world scenario, it is noteworthy that the accuracy with which one can recover the real sensitive s depends on the right representation of the complementary set x_c. Yet, it is possible that the set x_c is under-represented. In such a case, there is a risk that

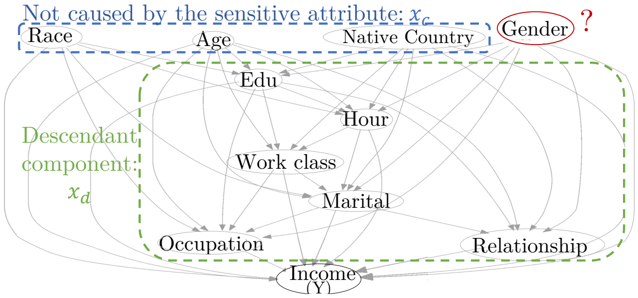


Figure 2: Causal Graph - Adult UCI

the reconstruction of s may contain some of this missing additional information. For instance, assuming that the graph from Figure 2 is the exact causal graph that underlies the Adult UCI, let us consider a setting where the variable *Race* is hidden. Hence, this variable would be likely to leak in the sensitive variable reconstruction. In such a leakage setting, we argue that working with a binary sensitive proxy would strongly degrade the inferred sensitive information, by introducing noise in the reconstruction. This is what motivated us to rather consider the rightmost graph from Figure 1. It considers a multivariate continuous intermediate confounder z that both causes the sensitive s and the observed variables in x_d and y. As long as the confounder z contains the real sensitive information, removing the corresponding dependence with the output prediction is guaranteed to ensure fairness for the model (we prove this in 1). As we observe in the experiments section, such a multivariate proxy also allows for better generalization abilities for mitigated prediction.

3.2 Reconstructing the Sensitive Attributes

We describe in this section the first step of our SRCVAE (Sensitive Retrieval Causal Variational Autoencoder) framework, which aims to generate a latent representation z that contains as much information as possible about the real sensitive feature s. As discussed above, our strategy is to use Bayesian inference approximation, using the pre-defined causal graph represented in Figure 1.

VAE Leveraging recent developments for approximate inference with deep learning, many different works proposed to use Variational Autoencoding methods (VAE) [Kingma and Welling, 2013] to model exogenous variables in causal graphs. It has been shown to achieve successful results, in particular in the sub-field of counterfactual fairness [Louizos *et al.*, 2017; Grari *et al.*, 2020a]. We propose to apply VAE for our setting of fairness with hidden sensitive attribute.

Following the rightmost causal graph from Figure 1, the decoder distribution $p_\theta(x_c, x_d, y|z)$ can be factorized as:

$$p_\theta(x_c, x_d, y|z) = p(x_c)p_\theta(x_d|x_c, z)p_\theta(y|x_c, x_d, z)$$

Given an approximate posterior $q_\phi(z|x_c, x_d, y)$, we obtain the following variational lower bound:

$$\begin{aligned} log(p_\theta(x_c,x_d,y)) \geq \mathbb{E}_{\substack{(x_c,x_d,y)\sim\mathcal{D},\\ z\sim q_\phi(z|x_c,x_d,y)}} &[\log p_\theta(x_d, y|x_c, z) \\ &+ log(p(x_c)) - D_{KL}(q_\phi(z|x_c, x_d, y)||p(z))] \end{aligned} \tag{1}$$

where D_{KL} denotes the Kullback-Leibler divergence of the posterior $q_\phi(z|x_c, x_d, y)$ from a prior $p(z)$, typically a standard Gaussian distribution $\mathcal{N}(0, I)$. The posterior $q_\phi(z|x_c, x_d, y)$ is estimated using a deep neural network with parameters ϕ, which typically outputs the mean μ_ϕ and the variance σ_ϕ of a diagonal Gaussian distribution $\mathcal{N}(\mu_\phi, \sigma_\phi I)$.

The likelihood term, which factorizes as $p_\theta(x_d, y|x_c, z) = p_\theta(x_d|x_c, z)p_\theta(y|x_c, x_d, z)$, is defined as the output of a neural network with parameters θ. Since attracted by a standard prior, the posterior is supposed to remove the probability mass for any information of z that is not involved in the reconstruction of x_d and y. Since x_c is given together with z as input of the likelihoods, all the information from x_c should be removed from the posterior distribution of z. In this paper, we employ a variant of the ELBO optimization as done in [Pfohl *et al.*, 2019], where the term $D_{KL}(q_\phi(z|x_c, x_d, y)||p(z))$ is replaced by a Maximum Mean Discrepancy (MMD) term $\mathcal{L}_{MMD}(q_\phi(z)||p(z))$ between the aggregated posterior $q_\phi(z)$ and the prior. This has been shown to be more powerful than the classical D_{KL} for ELBO optimization in [Zhao *et al.*, 2017], as the latter may be too restrictive [Chen *et al.*, 2016; Sønderby *et al.*, 2016], and also tends to overfit the data.

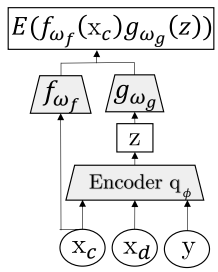
HGR Minimization To be accurate, inference must ensure that no dependence is created between x_c and z (no arrow is linking x_c to z in the rightmost graph in Figure 1). This ensures the generation of a proper sensitive proxy that is not linked to the complementary x_c. However, by optimizing the ELBO Equation 1, some dependence may still be observed empirically between x_c and z, as we show in Section 4. This is due to some information from x_c leaking to the inferred z. In order to ensure some minimum independence level, we add a penalisation term in the proposed loss function. Leveraging recent research for mitigating the dependence between continuous variables, we extend the main idea of [Grari *et al.*, 2021; Grari *et al.*, 2020b] by adapting this penalization to the case of variational autoencoders. Following this idea, we consider the Hirschfeld-Gebelein-Renyi (HGR) coefficient [Rényi, 1959] to measure the (possibly non linear) dependence between two (possibly multidimensional) variables. In the following, we denote as $\widehat{HGR}^{w_f,w_g}_{U\sim\mathcal{D}_\mathcal{U},V\sim\mathcal{D}_\mathcal{V}}(U, V)$ the neural estimation of HGR between two variables U and V, computed via two inter-connected neural networks f and g with parameters w_f and w_g [Grari *et al.*, 2020b; Grari *et al.*, 2021]:

$$\widehat{HGR}^{w_f,w_g}_{U\sim\mathcal{D}_\mathcal{U},V\sim\mathcal{D}_\mathcal{V}}(U, V) = \max_{w_f,w_g} \mathbb{E}_{U\sim\mathcal{D}_\mathcal{U},V\sim\mathcal{D}_\mathcal{V}}(\widehat{f}_{w_f}(U)\widehat{g}_{w_g}(V))$$

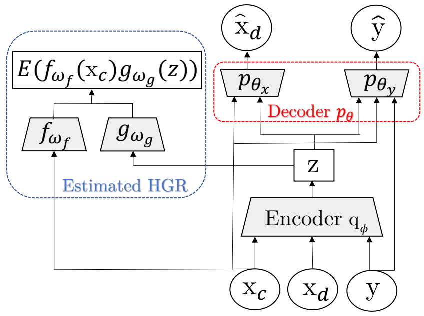
where $\mathcal{D}_\mathcal{U}$ (resp. $\mathcal{D}_\mathcal{V}$) is the distribution of U (resp. V), and $\widehat{f}$ (resp. $\widehat{g}$) refer to standardized outputs of network f (resp. g).

Reconstruction Objective Altogether, the final objective of our SRCVAE approach is given as:

$$\begin{aligned} \arg\min_{\theta,\phi}\max_{w_f,w_g} -\mathbb{E}_{\substack{(x_c,x_d,y)\sim\mathcal{D},\\ z\sim q_\phi(z|x_c,x_d,y)}} &[\log p_\theta(x_d, y|x_c, z) \\ &+ \lambda_{mmd}\mathcal{L}_{MMD}(q_\phi(z)||p(z))] \\ &+ \lambda_{inf} \widehat{HGR}^{w_f,w_g}_{\substack{(x_c,x_d,y)\sim\mathcal{D},\\ z\sim q_\phi(z|x_c,x_d,y)}}(x_c, z) \end{aligned}$$



(a) Max Phase: HGR



(b) Min Phase: Reconstruction

Figure 3: Neural architecture of SRCVAE in max phase for the HGR estimation between x_c and z via gradient ascent (a) and Variational autoencoder structure of SRCVAE in min phase (b).

where $\lambda_{mmd}, \lambda_{inf}$ are scalar hyperparameters. The additional MMD objective can be interpreted as minimizing the distance between all moments of each aggregated latent code distribution and the prior distribution. Note that giving y as input of the inference scheme $q(z|x_c, x_d, y)$ is allowed since z is only used during training (see next section).

In Figure 3, we represent the min-max structure of SRCVAE. The left structure represents the max phase where the HGR between z and x_c is estimated by gradient ascent with multiple iterations. The right graph represents the min phase where the reconstruction of x_d and y is performed by the decoder p_θ (red frame) via the generated latent space z from the encoder q_ϕ. The adversarial HGR component (blue frame) ensures independence between the generated latent space z and x_c. The network f takes the set x_c as input, while g takes the continuous representation space z. This way, for each gradient iteration of SRCVAE we capture the estimated HGR between the set x_c and the generated proxy latent space z. At the end of each iteration, the algorithm updates the parameters of the decoder parameters θ as well as the encoder parameters ϕ by one step of gradient descent. Concerning the HGR adversary, the backpropagation of the parameters ω_f and ω_g is performed by multiple steps of gradient ascent. This allows for a more accurate estimation of the HGR at each step, leading to a far more stable learning process. λ_{inf} controls the importance of the dependence loss in the optimization.

3.3 Mitigating the Unwanted Biases

The sensitive reconstruction model can now be used for training a fair predictive function h_θ. Since z contains some continuous multidimensional information, we adopt an HGR-based approach inspired from [Grari *et al.*, 2020b; Grari *et al.*, 2021] which have shown superior performance in this context. In our setting, we also verify this claim empirically as shown in appendix. We propose to mitigate the unwanted bias via an adversarial penalization during the training phase that depends on the targeted fairness objective.

Demographic Parity We propose to find a mapping $h_\theta(x)$ that both minimizes the deviation with the expected target y and does not imply much dependency with the representation z, inferred from $q_\phi(z|x_c, x_d, y)$ as described in the previous section. We propose the following optimization, which considers a neural estimation of HGR as well, but this time applied to variables $h_\theta(x)$ (the output of the classifier) and z (the inferred latent representation):

$$\arg\min_{\theta} \max_{\psi_f,\psi_g} \mathcal{L}(h_\theta(x), y) + \lambda_{DP} \widehat{HGR}^{\psi_f,\psi_g}_{\substack{(x_c,x_d,y)\sim\mathcal{D},\\ z\sim q_\phi(z|x_c,x_d,y)}} (h_\theta(x), z)$$

where $\mathcal{L}$ is the predictor loss function (the log-loss function in our experiments) of the output $h_\theta(x) \in \mathbb{R}$ w.r.t. the target label y. The hyperparameter λ_{DP} controls the impact of dependence between the output prediction $h_\theta(x) \approx p(y = 1|x_d, x_c)$ and the sensitive proxy z. To assess this correlation, K different representations are sampled for each observation (x_{c_i}, x_{d_i}, y_i) from the causal model (200 in our experiments). As in the inference phase, the backpropagation of the HGR adversary with parameters ψ_f and ψ_g is performed by multiple steps of gradient ascent. This allows to optimize a more accurate estimation of the HGR at each step, leading to a greatly more stable predictive learning process.

Practice in real-world As mentioned in the first subsection, the assumed causal graph 1 requires the right representation of the complementary set x_c. If the set x_c is under-represented, some specific hidden attributes can be integrated with the sensitive information in the inferred sensitive latent space z. The following Theorem 1 allows us to ensure that mitigating the HGR between z and $\widehat{y}$ implies some upper-bound for the targeted objective (proof in appendix).

Theorem 1. *For two nonempty index sets S and Z such that $S \subset Z$ and $\hat{Y}$ the output prediction of the model, we have:*

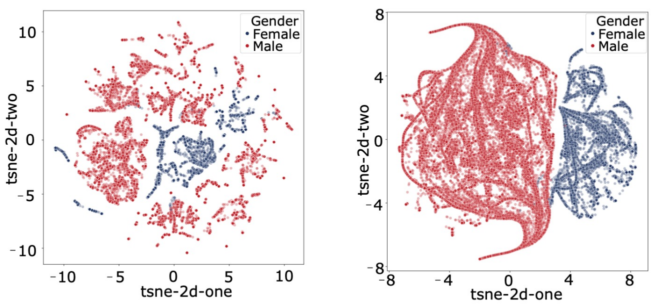
$$HGR(\hat{Y}, Z) \geq HGR(\hat{Y}, S) \tag{2}$$

Proof. in appendix

Therefore, minimizing $HGR(\hat{Y}, Z)$ tends to reduce the real bias objective $HGR(\hat{Y}, S)$. Results on benchmark and real-world datasets demonstrate below in part 1 that such an assumed graph demonstrates good robustness properties. This property is also held for equalized-odds we consider below, with $HGR(\hat{Y}, Z|Y) \geq HGR(\hat{Y}, S|Y)$.

Equalized odds We extend the demographic parity optimization to the equalized odds task. The objective is to find a mapping $h_\theta(x)$ which both minimizes the deviation with the expected target y and does not imply too much dependency with the representation z conditioned on the actual outcome y. For the decomposition of disparate mistreatment, we propose to divide the mitigation based on the two different values of y. Identification and mitigation of the specific non linear dependence for these two subgroups leads to the same false positive and the same false negative rates for each demographic. We propose the following optimization:

$$\arg\min_{\theta} \max_{\psi_{f_0},\psi_{g_0},\psi_{f_1},\psi_{g_1}} \mathcal{L}(h_\theta(x), y) +$$
$$\lambda_0 \widehat{HGR}^{\psi_{f_0},\psi_{g_0}}_{\substack{(x,y)\sim\mathcal{D}_0,\\ z\sim q_\phi(z|x,y)}} (h_\theta(x), z) + \lambda_1 \widehat{HGR}^{\psi_{f_1},\psi_{g_1}}_{\substack{(x,y)\sim\mathcal{D}_1,\\ z\sim q_\phi(z|x,y)}} (h_\theta(x), z)$$



(a) λ_{inf}=0.0; $HGR(x_c, z)$=81.7% (b) λ_{inf}=0.2; $HGR(x_c, z)$=22.6%

Figure 4: Inference phase for Adult UCI: t-SNE of the sensitive latent reconstruction Z. Blue points are males ($S = 1$), red ones are females ($S = 0$). Increasing λ_{inf} improves the independence of z from x_c. This leads to a better separation between male and female data points, which indicates a proper sensitive proxy.

with $\mathcal{D}_0$ (resp. $\mathcal{D}_1$) corresponding to the observations set (x, y) verifying $y = 0$ (resp. $y = 1$). The hyperparameters λ_0 and λ_1 control the impact of the dependence loss for the false positive and the false negative objective respectively. The first penalisation (controlled by λ_0) enforces the independence between the output prediction $h_\theta(x) \approx p_\theta(y = 1|x)$ and the sensitive proxy z only for the cases where $y = 0$. It enforces the mitigation of the difference of false positive rates between demographics, since at optimum for θ^* with no trade-off (i.e., with infinite λ_0) and $(x, y) \sim \mathcal{D}_0$, $HGR(h_{\theta^*}(x), z) = 0$ and implies theoretically: $h_{\theta^*}(x) \perp z|y = 0$. The second one enforces the mitigation of the difference between the true positive rates, since the dependence loss is performed between the output prediction $h_\theta(x)$ and the sensitive proxy only for cases where $y = 1$ (i.e., mitigation of Δ_{FNR}).

4 Experimental Results

For our experiments, we empirically evaluate the performance of our contribution on real-world data sets where the sensitive s is available. This allows to assess the fairness of the output prediction, obtained without the use of the sensitive attribute, w.r.t. this ground truth. For this purpose, we use the popular Adult UCI and Default datasets (descriptions in Appendix), often used in fair classification.

Sensitive Reconstruction In order to understand the interest of mitigating the dependence between the latent space z and the complementary set x_c during the inference phase, we plot the t-SNE of z with two different inference models for the Adult UCI dataset in Figure 4. We consider a version of our model trained without the penalization term ($\lambda_{inf} = 0.00$) as a baseline. It is then compared to a version trained with a penalization term equal to 0.20. As expected, training the inference model without the penalization term results in a poor reconstruction of the z proxy, where the dependence on x_c is observed. We can observe that the separation between the men (blue points) and women (red points) data is not significant. We also observe that increasing this hyper-parameter (λ_{inf}) allows to decrease the HGR estimation from 81.7% to 22.6% and to greatly increase the separation between male and female data points.

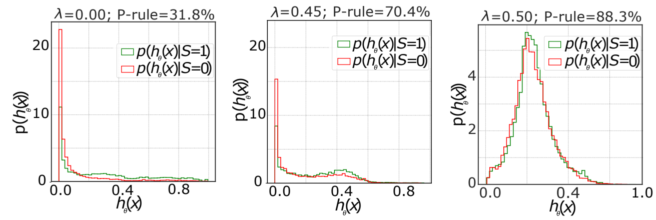
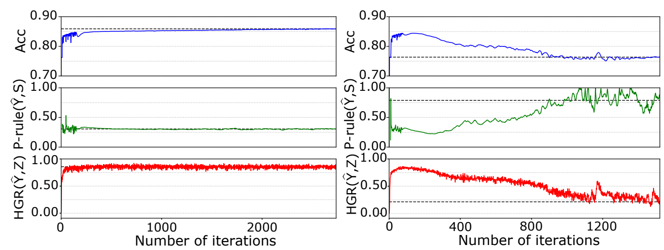


Figure 5: Distributions of the predicted probabilities given the real sensitive s (Adult UCI data set) for the Demographic Parity task.



(a) No mitigation ($\lambda_{DP} = 0$) (b) Mitigation with $\lambda_{DP} = 0.5$

Figure 6: Dynamics of adversarial training

Bias Mitigation The dynamics of adversarial training for demographic parity is performed for Adult UCI with unfair ($\lambda_{DP} = 0$) and fair ($\lambda_{DP} = 0.5$) models as illustrated in Figure 6. Other values are presented in appendix. We represent the accuracy of the model (top), the P-rule metric between the prediction and the real sensitive s (middle), and the HGR between the prediction and the latent space z (bottom). For the unfair model (leftmost graph) we observe that the convergence is stable and achieves a P-rule of 29.5%. As expected, the penalization loss decreases (measured with the HGR) when the hyperparameter λ_{DP} is increased. It allows to increase the fairness metric P-rule to 83.1% with a slight drop of accuracy.

In Figure 5 we plot the distribution of the predicted probabilities for each sensitive attribute s for three different models: an unfair model with $\lambda_{DP} = 0$, and two fair models with $\lambda_{DP} = 0.45$ and 0.50, respectively. For the leftmost graph (i.e. $\lambda_{DP} = 0$) the model appears to be very unfair, since the distribution between the sensitive groups differs importantly. As expected, we observe that the distributions are more aligned as λ_{DP} values increase.

For the two datasets, we test different models where, for each, we repeat five runs by randomly sampling two subsets, 80% for the training set and 20% for the test set. As different optimization objectives result in different algorithms, we run separate experiments for the two fairness objectives of our interest. As an optimal baseline to be reached, we consider the approach from [Adel *et al.*, 2019] using observations of the sensitive s during training, which we denote as

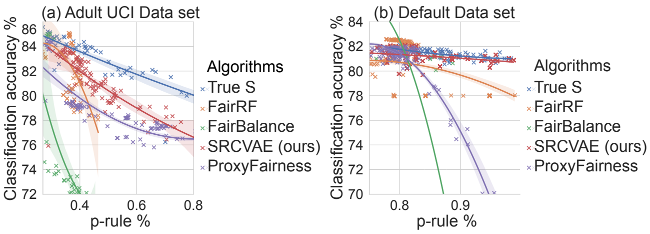


Figure 7: Demographic Parity task

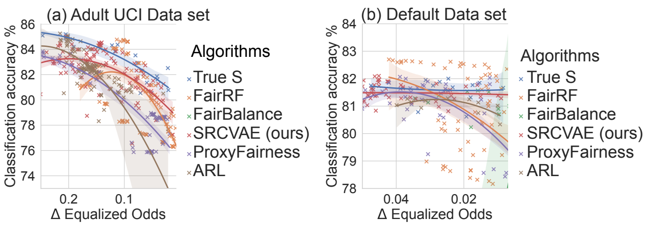


Figure 8: Equalized odds task

True S. We also compare various approaches specifically designed to be trained in the absence of the sensitive information during training: *FairRF* [Zhao *et al.*, 2021], *FairBalance* [Yan *et al.*, 2020], *ProxyFairness* [Gupta *et al.*, 2018] and *ARL* [Lahoti *et al.*, 2020]. The latter is only compared for the equalized odds task (i.e. discussion in [Zhao *et al.*, 2021]). We plot the performance of these approaches by displaying the Accuracy against the P-rule for Demographic Parity (Figure 7) and the Disparate Mistreatment (DM) for Equalized Odds (Figure 8). For all algorithms, we clearly observe that the Accuracy, or predictive performance, decreases when fairness increases. As expected, the baseline *True S* achieves the best performance for all the scenarios with the highest accuracy and fairness. We note that, for all levels of fairness (controlled by the mitigation weight in every approach), our method outperforms state-of-the-art algorithms for both fairness tasks (except some points for very low levels of fairness, on the left of the curves). We attribute this to the ability of SRCVAE to extract a useful sensitive proxy, while the approaches *FairRF* and *ProxyFairness* seem to greatly suffer from merely considering correlations present in the data for mitigating fairness. The approach *FairBalance*, which preprocessed the data with clustering, seems inefficient and degrades the predictive performance too significantly. The advantages of our approach are more pronounced on the Default dataset, where a less obvious correlation exists between observed variables and the sensitive attribute. In that setting, leveraging the knowledge of a causal graph appears to be crucial.

Proxy dimensions In figure 9(a), we perform an additional experiment on the sensitive proxy. For the two datasets we observe that increasing z dimensions results in increased accuracy. Increasing the dimensions to 5 for Adult UCI (same experiment for Default in appendix) allows to obtain better results in terms of accuracy and this for all levels of P-rule. We claim that mitigating biases in larger spaces allows better generalisation abilities at test time, as already observed in another context in [Grari *et al.*, 2021]. It supports the choice of considering a multivariate sensitive proxy z, rather than directly acting on a reconstruction of s as a univariate variable.

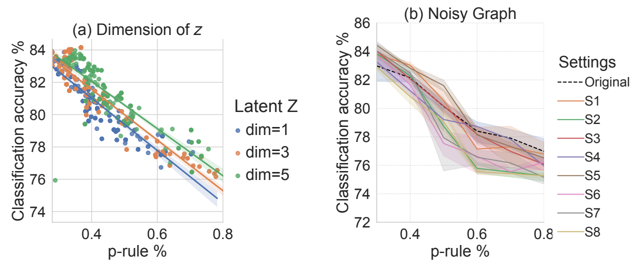


Figure 9: Additional Experiments

Noisy graph In figure 9(b), we analyse the impact of noise in the causal graph. To do this, we focus on cases where the decomposition of x in sets x_c and x_d is noisy, or sets of variables are under-represented. For this purpose, we experimented 8 scenarios on the Adult UCI data set. First, we removed features from x_c: the *race* (S1), the *age* (S2). Then, we removed features from x_d: the *education* (S3) and the *hour* (S4). Finally, we moved features from x_c to x_d and reversely: membership inversion between *race* and *education* (S5), membership inversion between *age* and *hour* (S6), inclusion of *age* in x_d (S7) and inclusion of *hour* in x_c (S8). From the results, our approach appears greatly robust to noise, with results in every scenario at least comparable to the best considered competitors (which all present settings where performances catastrophically drop as observed in Fig. 7 and 8). This robustness is partly achieved thanks to the use of a multivariate continuous proxy z, which limits the possible lack of sensitive information that would occur with a scalar proxy of s, if non-sensitive information leaks in the reconstruction. While the inclusion of variables from x_d to x_c may induce the removal of some useful sensitive information from the proxy, the inclusion of variables from x_c to x_d may lead to optimize the independence of some non sensitive information with model outputs. If fairness needs to be guaranteed, the expert must thus tend to favor false x_d variables rather than false x_c, the former only inducing a slight accuracy loss in most cases (as demonstrated in Theorem 1).

5 Conclusion and Future Work

This paper proposed a new way to mitigate undesired bias without the availability of the sensitive demographic information in training. To generate a latent representation which is expected to contain the most sensitive information as possible, the approach relies on a new variational auto-encoding based framework named SRCVAE. In a second phase, inferred proxies serve to mitigate biases in an adversarial fairness training of a prediction model. Compared with other state-of-the-art algorithms, our method proves to be more efficient in terms of accuracy for similar levels of fairness. For further investigation, we are interested in extending this work to settings where the actual sensitive can be continuous (e.g. age or weight attribute) and/or multivariate.

References

[Adel *et al.*, 2019] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *AAAI'19*, volume 33, pages 2412–2420, 2019.

[Angwin *et al.*, 2016] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. ProPublica, May 23, 2016, 2016.

[Bolukbasi *et al.*, 2016] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*. Curran Associates, Inc., 2016.

[Celis *et al.*, 2019] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328, 2019.

[Chen *et al.*, 2016] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

[Chen *et al.*, 2019] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 339–348, 2019.

[Coston *et al.*, 2019] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *AAAI*, 2019.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Grari *et al.*, 2020a] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Adversarial learning for counterfactual fairness. *arXiv preprint arXiv:2008.13122*, 2020.

[Grari *et al.*, 2020b] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Fairness-aware neural rényi minimization for continuous features. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2262–2268. ijcai.org, 2020.

[Grari *et al.*, 2021] Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. Learning unbiased representations via rényi minimization. *ECML PKDD*, 2021.

[Gupta *et al.*, 2018] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[Hashimoto *et al.*, 2018] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *ICML*, pages 1929–1938. PMLR, 2018.

[Kilbertus *et al.*, 2018] Niki Kilbertus, Adrià Gascón, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. In *ICML*, 2018.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Lahoti *et al.*, 2020] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H Chi. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

[Lambrecht and E. Tucker, 2016] Anja Lambrecht and Catherine E. Tucker. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *SSRN Electronic Journal*, 2016.

[Louizos *et al.*, 2017] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[Madras *et al.*, 2018] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[Mohri *et al.*, 2019] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, 2019.

[Pedreshi *et al.*, 2008] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *KDD'08*, page 560, 2008.

[Pfohl *et al.*, 2019] Stephen Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H Shah. Counterfactual reasoning for fair clinical risk prediction. *arXiv preprint arXiv:1907.06260*, 2019.

[Rényi, 1959] Alfréd Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959.

[Schumann *et al.*, 2019] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688*, 2019.

[Sønderby *et al.*, 2016] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *NIPS'16*, pages 3738–3746, 2016.

[Veale and Binns, 2017] Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

[Wadsworth *et al.*, 2018] Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv:1807.00199*, 2018.

[Yan *et al.*, 2020] Shen Yan, Hsien-te Kao, and Emilio Ferrara. Fair class balancing: enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020.

[Zafar *et al.*, 2015] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

[Zhang *et al.*, 2018] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI'18*, pages 335–340, 2018.

[Zhao *et al.*, 2017] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

[Zhao *et al.*, 2021] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. You can still achieve fairness without sensitive attributes: Exploring biases in non-sensitive features. *arXiv preprint arXiv:2104.14537*, 2021.