Fairness-Aware Grouping for Continuous Sensitive Variables: Application for Debiasing Face Analysis with respect to Skin Tone

Veronika Shilova[a,b], Emmanuel Malherbe[a], Giovanni Palma[c], Laurent Risser[b] and Jean-Michel Loubes[b,d]

[a]Artefact Research Center, Paris, France, [b]Institut de Mathématiques de Toulouse (UMR 5219), CNRS, Université de Toulouse, F-31062 Toulouse, France, [c]L'Oréal Research and Innovation, Paris, France, [d]INRIA, ANITI, France

Table of content

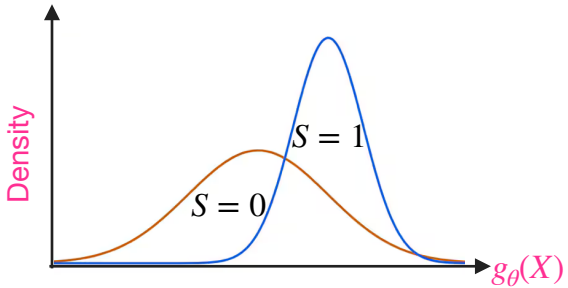
1. Motivation
2. Problem setting
3. FairGroups: Fairness aware grouping algorithm
4. Experiments
    4.1. Synthetic data
    4.2. CelebA and FFHQ
5. Post-processing using FairGroups
6. Conclusion

Motivation

- (X, Y) – data: features and target (ground truth, $Y \in \{0,1\}$)
- g_θ – model ($g_\theta(X) \in [0,1]$)
- $\hat{Y}$ – model's prediction ($\hat{Y} = \mathbf{1}_{\{g_\theta(X)>0.5\}} \in \{0,1\}$)
- S – sensitive attribute (e.g. gender, age, skin color, etc.)

Fairness Definition: Independence: $\hat{Y} \perp\!\!\!\perp S$

Predefined groups



Disparate Impact:

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

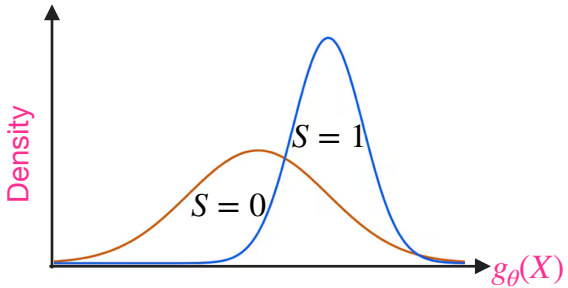
ARTEFACT RESEARCH CENTER

Motivation

- (X, Y) – data: features and target (ground truth, $Y \in \{0,1\}$)
- g_θ – model ($g_\theta(X) \in [0,1]$)
- $\hat{Y}$ – model's prediction ($\hat{Y} = \mathbf{1}_{\{g_\theta(X)>0.5\}} \in \{0,1\}$)
- S – sensitive attribute (e.g. gender, age, skin tone, etc.)

Fairness Definition: Independence: $\hat{Y} \perp\!\!\!\perp S$

Predefined groups



Disparate Impact:

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

- Prone to annotator bias
- Overlook both the underlying data distribution and the target variable
- Assume samples inside a group to be homogeneous in terms of fairness, thus, underestimating the level of discrimination

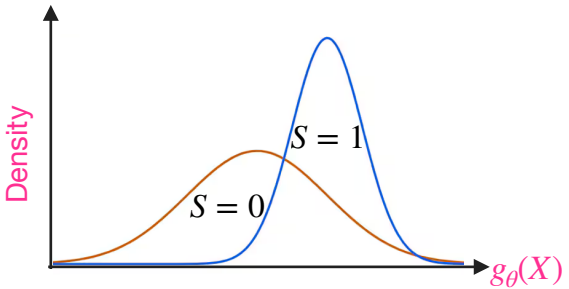
ARTEFACT RESEARCH CENTER

Motivation

- (X, Y) – data: features and target (ground truth, $Y \in \{0,1\}$)
- g_θ – model ($g_\theta(X) \in [0,1]$)
- $\hat{Y}$ – model's prediction ($\hat{Y} = \mathbf{1}_{\{g_\theta(X)>0.5\}} \in \{0,1\}$)
- S – sensitive attribute (e.g. gender, age, skin color, etc.)

Fairness Definition: Independence: $\hat{Y} \perp\!\!\!\perp S$

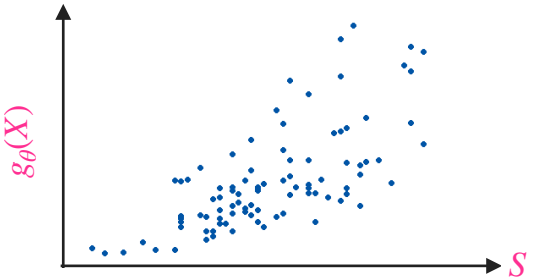
Predefined groups



Disparate Impact:

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

Continuous sensitive variable



Pearson Correlation:

$$\rho_{\hat{Y},S} = 0$$

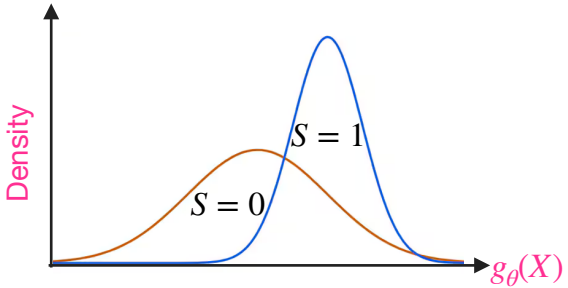
ARTEFACT RESEARCH CENTER

Motivation

- (X, Y) – data: features and target (ground truth, $Y \in \{0,1\}$)
- g_θ – model ($g_\theta(X) \in [0,1]$)
- $\hat{Y}$ – model's prediction ($\hat{Y} = \mathbf{1}_{\{g_\theta(X)>0.5\}} \in \{0,1\}$)
- S – sensitive attribute (e.g. gender, age, skin color, etc.)

Fairness Definition: Independence: $\hat{Y} \perp\!\!\!\perp S$

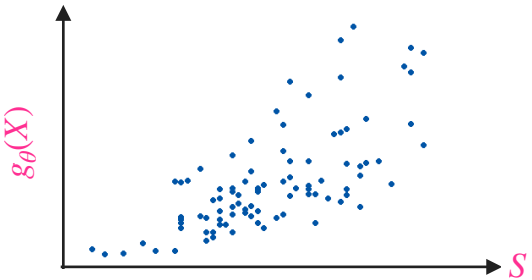
Predefined groups



Disparate Impact:

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

Continuous sensitive variable



Pearson Correlation:

$$\rho_{\hat{Y},S} = 0$$

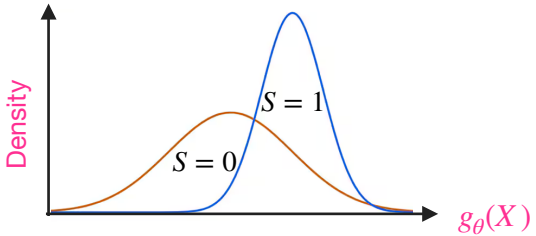
- Non-linear dependence is computationally intractable
- Evaluate the overall dependence between a continuous sensitive variable and a target variable

ARTEFACT RESEARCH CENTER

Motivation

Fairness Definition: Independence: $\hat{Y} \perp\!\!\!\perp S$

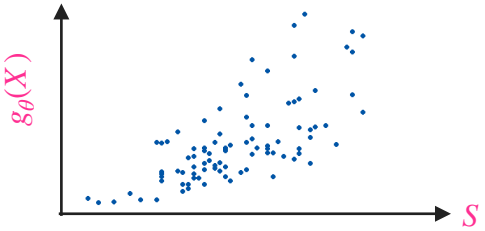
Predefined groups



Disparate Impact:

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

Continuous sensitive variable



Pearson Correlation:

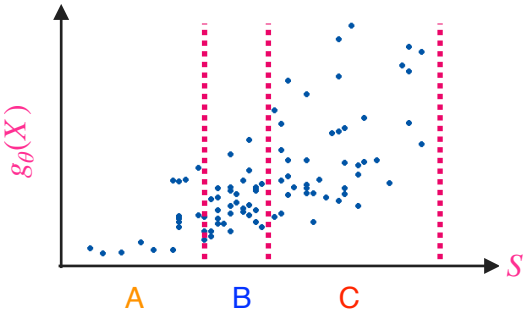
$$\rho_{\hat{Y},S} = 0$$

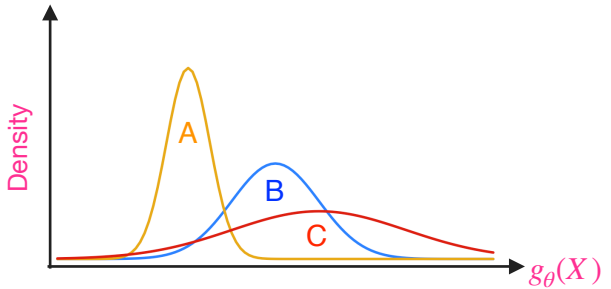
Our paper



Worst case discrimination: Groups with the most different levels of discrimination

Use both the underlying data distribution and the target variable





ARTEFACT RESEARCH CENTER

Motivation

AI Act emphasizes the need for reporting algorithmic performance across demographic groups.



Our research question: How to partition a *continuous S* into groups ?

Problem setting

- (X, Y) – data: features and target (ground truth, $Y \in \{0,1\}$)
- g_θ – model ($g_\theta(X) \in [0,1]$)
- $\hat{Y}$ – model's prediction ($\hat{Y} = \mathbf{1}_{\{g_\theta(X)>0.5\}} \in \{0,1\}$)
- $L \in \mathscr{L} \subset \mathbb{R}^p$, $(p = 1$ *or* $2)$ – sensitive attribute – skin color, e.g. $L^*, ITA,$ *or* (L^*, h^*), etc.

We denote a partition $\mathscr{P} = \{\mathscr{P}_k\}_{k=1}^K$ as a set of connected subsets $\mathscr{P}_k \subset \mathscr{L}$ with the two following properties:

1. there is no overlap between elements of the partition: $\forall k_1, k_2 \in \{1,\dots,K\}, \mathscr{P}_{k_1} \cap \mathscr{P}_{k_2} = \varnothing$;
2. $\mathscr{P}$ totally covers $\mathscr{L}$: $\cup_{k=1}^K \mathscr{P}_k = \mathscr{L}$.

We consider that the number of groups K is given, as a user-defined variable.

Problem setting

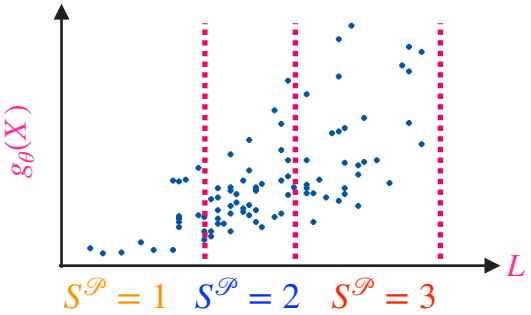
For a partition $\mathscr{P}$ we define $S^{\mathscr{P}}$ to be a multinomial random variable associating the observations of L to their group $\mathscr{P}_k$:

$$S^{\mathscr{P}} = k \iff L \in \mathscr{P}_k, \text{ where } k = 1,\dots,K\,.$$

We introduce and formalize a new task: find the partition $\mathscr{P}$ of the continuous sensitive attribute L, such that its components $\mathscr{P}_k$ have the most different levels of discrimination according to the fairness measured by Φ:

$$\arg\max_{\mathscr{P}} \mathbb{V}ar\Big(\Phi\big(S^{\mathscr{P}}\big)\Big),$$

where, e.g. $\Phi(S^{\mathscr{P}}) = \mathbb{P}(\hat{Y} = 1 \,|\, S^{\mathscr{P}}) - \mathbb{P}(\hat{Y} = 1)$.



ARTEFACT RESEARCH CENTER

Problem setting

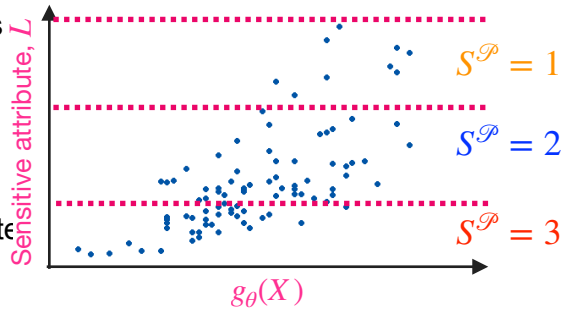
For a partition $\mathscr{P}$ we define $S^{\mathscr{P}}$ to be a multinomial random variable associating the observations of L to their group $\mathscr{P}_k$:

$$S^{\mathscr{P}} = k \iff L \in \mathscr{P}_k, \text{ where } k = 1, \ldots, K.$$

We introduce and formalize a new task: find the partition $\mathscr{P}$ of the continuous sensitive attribute L, such that its components $\mathscr{P}_k$ have the most different levels of discrimination according to the fairness measured by Φ:

$$\arg\max_{\mathscr{P}} \mathbb{V}ar\Big(\Phi\big(S^{\mathscr{P}}\big)\Big),$$

where, $S^{\mathscr{P}} = k \iff L \in \mathscr{P}_k,\ k = 1, \ldots, K$, and e.g. $\Phi(S^{\mathscr{P}}) = \mathbb{P}(\hat{Y} = 1 \mid S^{\mathscr{P}}) - \mathbb{P}(\hat{Y} = 1)$.

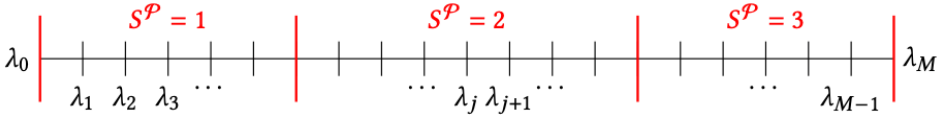


ARTEFACT RESEARCH CENTER

FairGroups: Fairness aware grouping algorithm

We consider partitions $\mathscr{P}$ as follows:

- In the one-dimensional case, we consider $\mathscr{P}_k = [l_{k-1}, l_k]$ to be segments, with $l_{k-1} < l_k$.
- In the two-dimensional case, we consider $\mathscr{P}_k$ to be rectangles, i.e. of the shape $[l, l'] \times [h, h']$.

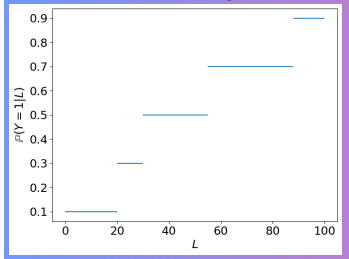


We solve the task of fairness-aware grouping: our method, ***FairGroups***, performs an exhaustive search with optimal pre-computation of the fairness metric Φ using dynamic programming.

ARTEFACT RESEARCH CENTER

Experiments: Synthetic data

Experimental Setting



$\mathbb{P}(Y = 1 \mid L)$

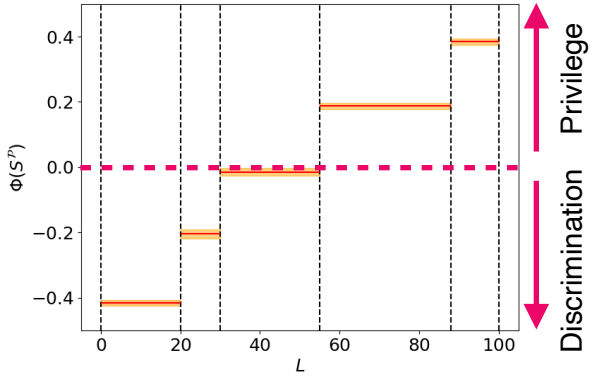
Generated observations

1. $L \sim \mathcal{U}(0,100)$
2. $L \sim \mathcal{N}(50,20)$ truncated on $[0,100]$

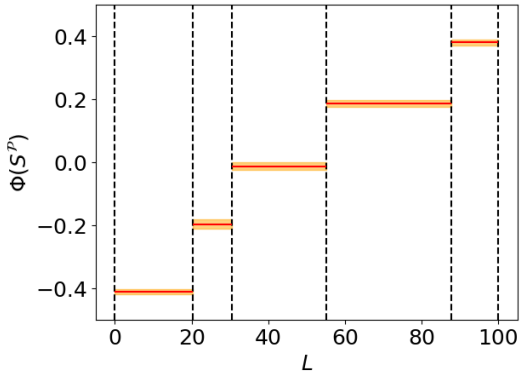
Takeaways

- Validated FairGroups on synthetic data with known ground truth.
- FairGroups: found the ground truth partition of L regardless of its distribution.

Comparison of partitions of L based on $\Phi(S^{\mathcal{P}}) = \mathbb{P}(Y = 1 \mid S^{\mathcal{P}}) - \mathbb{P}(Y = 1)$, $K = 5$



Ground truth partition of L



FairGroups partition of L

Data distribution for computing $\mathcal{P}$	$L \sim \mathcal{U}(0, 100)$		$L \sim \mathcal{N}(50, 20)$ truncated on $[0, 100]$	
Metric	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$	Rand Index (with ground truth)	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$	Rand Index (with ground truth)
$\mathcal{P}$: Ground Truth	**0.068**	1	**0.032**	1
$\mathcal{P}$: FairGroups	**0.068**	0.99	**0.032**	0.97

ARTEFACT RESEARCH CENTER

Remark: Skin color space

CIELAB (L^*, a^*, b^*) color space. Euclidean distance between two colors is proportional to their difference as perceived by humans (not verified in the RGB space!).

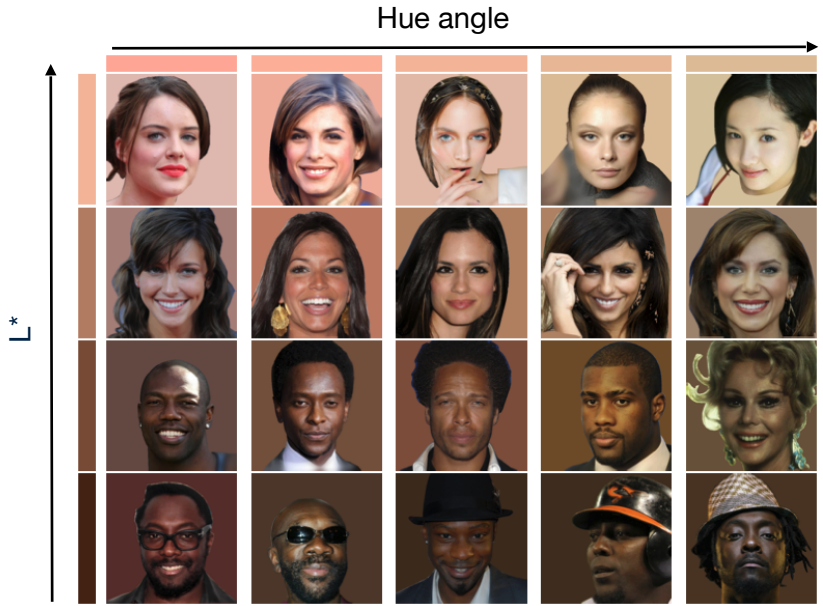
- $L^* \in [0,100]$ perceptual lightness (black to white);
- $a^* \in [-128,127]$ green to red colors;
- $b^* \in [-128,127]$ blue to yellow colors.

Individual Typology Angle (ITA)

$$ITA = arctan\left(\frac{L^* - 50}{b^*}\right) \times \frac{180^\circ}{\pi}$$

Hue angle (h^*)

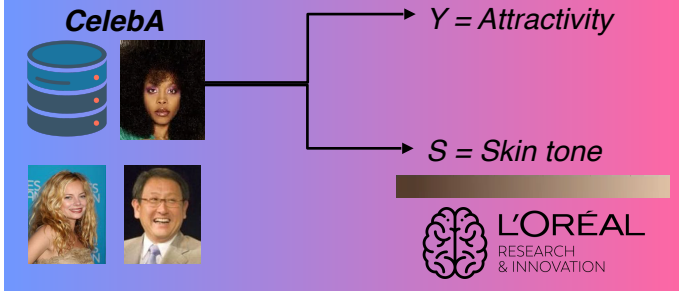
$$h^* = arctan\left(\frac{b^*}{a^*}\right) \times \frac{180^\circ}{\pi}$$



ARTEFACT RESEARCH CENTER

Experiments: CelebA and FFHQ

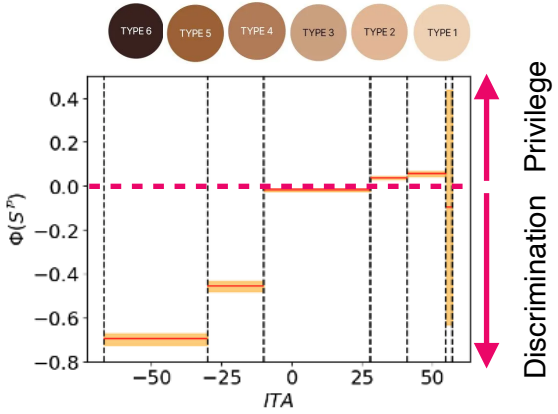
Experimental Setting



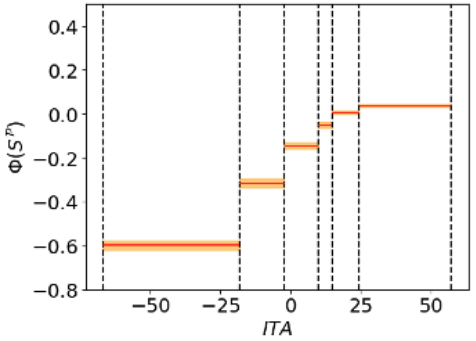
Takeaways

- Fitzpatrick: confidence intervals for certain groups are too large to draw meaningful conclusions regarding fairness.
- FairGroups: partition with varying levels of discrimination and tight confidence intervals.

Comparison of partitions of ITA based on $\Phi(S^{\mathcal{P}}) = \mathbb{P}(\hat{Y} = 1 \mid S^{\mathcal{P}}) - \mathbb{P}(\hat{Y} = 1), K = 6$



Fitzpatrick (default) partition of ITA



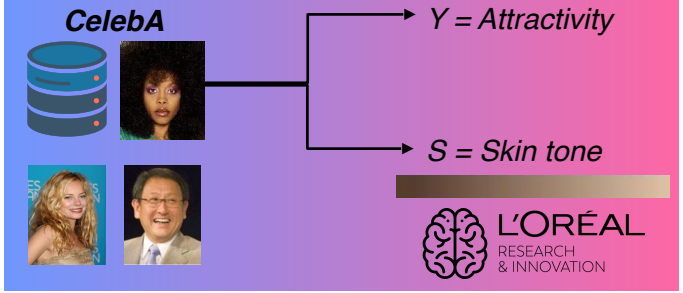
FairGroups partition of ITA

Data for computing $\mathcal{P}$	CelebA		FFHQ	CelebA & FFHQ
Metric	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on CelebA	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on FFHQ	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on FFHQ	Rand Index between both partitions, on FFHQ
$\mathcal{P}$: Fitzpatrick	0.078	0.079	0.079	1
$\mathcal{P}$: FairGroups	**0.092**	**0.09**	**0.093**	0.921

ARTEFACT RESEARCH CENTER

Experiments: CelebA and FFHQ

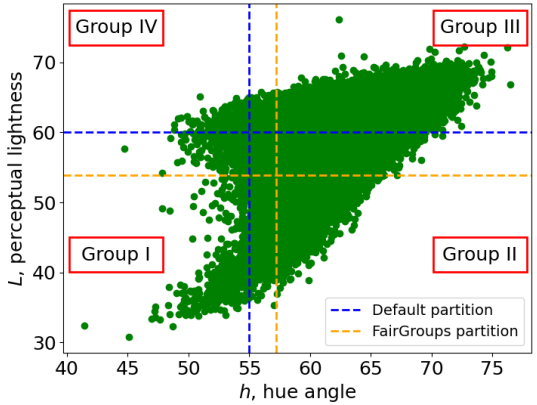
Experimental Setting



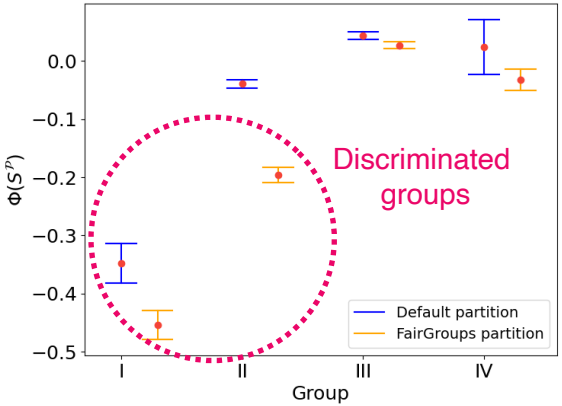
Takeaways

- Default: identified only one discriminated group underestimating the level of discrimination; confidence intervals are larger.
- FairGroups: found two discriminated groups showing the worst case discrimination levels; confidence intervals are tighter.

Comparison of partitions of (L, h) in 2D based on $\Phi(S^{\mathcal{P}}) = \mathbb{P}(\hat{Y} = 1 \mid S^{\mathcal{P}}) - \mathbb{P}(\hat{Y} = 1)$, $K = 4$



Clusters of Default and FairGroups partitions of (L, h)



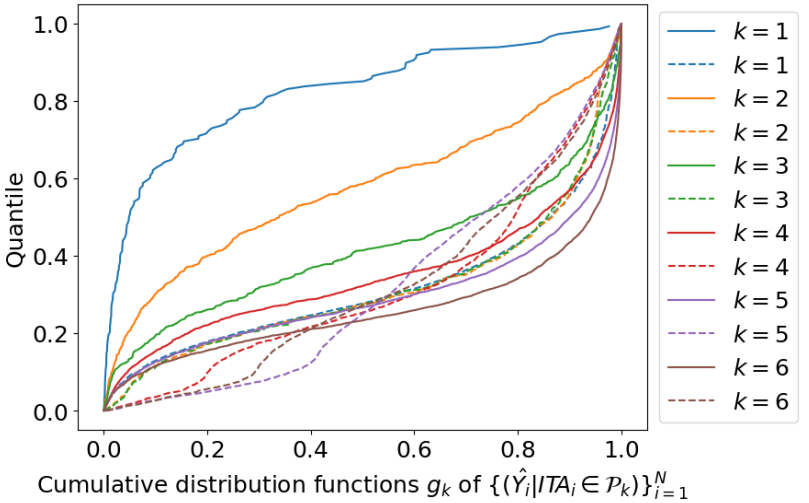
Confidence intervals for $\Phi(S^{\mathcal{P}})$

Data for computing $\mathcal{P}$	CelebA		FFHQ	CelebA & FFHQ
Metric	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on CelebA	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on FFHQ	$\mathbb{V}ar\left(\Phi(S^{\mathcal{P}})\right)$ on FFHQ	Rand Index between both partitions, on FFHQ
$\mathcal{P}$: Default	0.054	0.082	0.082	1
$\mathcal{P}$: FairGroups	**0.084**	**0.102**	**0.105**	0.883

ARTEFACT RESEARCH CENTER

Post-processing using FairGroups

Optimal transport-based post-processing: the minimum classification error achievable under Statistical Parity (SP) constraints corresponds to the solution of a Wasserstein barycenter problem



	Accuracy	PR-AUC	HGR$(\hat{Y}, ITA)$
No post-processing	0.793	0.824	0.126
$\mathcal{P}$: Fitzpatrick	0.782	0.815	0.095
$\mathcal{P}$: K-Means	0.783	0.813	0.065
$\mathcal{P}$: FairGroups	0.781	0.811	**0.039**

Conclusion

1. We proposed novel data-driven approaches to segment a continuous, possibly multidimensional sensitive variable into groups being homogeneous w.r.t. the discrimination present in a dataset or a model;
2. We showed that data-driven grouping reveals more discrimination patterns than predefined groups and remains robust to population shifts.
3. We used the groups identified by our data-driven method to mitigate the bias of a predictive model and obtained the best trade-off between model fairness and accuracy.

Code is available at https://github.com/artefactory/fair_groups/

