# Efficient fair PCA for fair representation learning

Matthäus Kleindessner, Michele Donini, Chris Russell, Muhammad Bilal Zafar

## Fairness & fair representation learning

Many fairness notions—in classification, one of the most prominent ones is demographic parity:

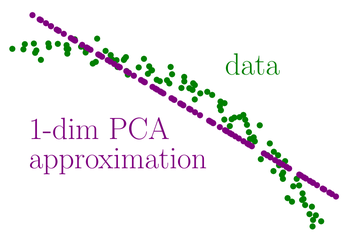
**Demographic parity (DP):** $\Pr(\hat{Y}=1|Z=z)=\Pr(\hat{Y}=1)$

Pr ... probability distribution over random variables $Y,\hat{Y}\in\{0,1\}$ and $Z\in\mathcal{Z}$
$Y$ ... ground-truth label, $\hat{Y}$ ... classifier's prediction, $Z$ ... demographic attribute (e.g., gender)

One approach to satisfy DP is **fair representation learning** (initiated by [1]):
$X\in\mathcal{X}$ ... features; learn a *fair* feature representation $f:\mathcal{X}\to\mathcal{X}'$ such that $f(X)\perp Z$
$\Rightarrow$ for any model trained on $f(X)$, predictions are independent of $Z$, and model satisfies DP

## Principal component analysis (PCA)

Among most prominent methods for dimensionality reduction.




**Goal:** find a best-approximating projection of the dataset $\{\mathbf{x}_1,\dots,\mathbf{x}_n\}\subseteq\mathbb{R}^d$ onto a $k$-dimensional linear subspace; $k$ ... given target dimension

**Formally:**

$$\operatorname{argmin}_{\mathbf{U}\in\mathbb{R}^{d\times k}:\,\mathbf{U}^\top\mathbf{U}=\mathbf{I}_{k\times k}}\sum_{i=1}^{n}\|\mathbf{x}_i-\mathbf{U}\mathbf{U}^\top\mathbf{x}_i\|_2^2\equiv\operatorname{argmax}_{\mathbf{U}\in\mathbb{R}^{d\times k}:\,\mathbf{U}^\top\mathbf{U}=\mathbf{I}_{k\times k}}\operatorname{trace}(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U})\qquad(1)$$

A solution to (1) is given by any $\mathbf{U}$ that comprises as columns orthonormal eigenvectors, corresponding to the largest $k$ eigenvalues, of $\mathbf{X}\mathbf{X}^\top$.

## Our formulation of fair PCA

**Idea:** remove demographic information when projecting the dataset onto the $k$-dimensional linear subspace

Ideally, we would like to solve

$$\operatorname{argmax}_{\mathbf{U}\in\mathcal{U}}\operatorname{trace}(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}),\quad\text{where }\mathcal{U}=\Big\{\mathbf{U}\in\mathbb{R}^{d\times k}:\mathbf{U}^\top\mathbf{U}=\mathbf{I}_{k\times k}\text{ and }\forall h:\mathbb{R}^k\to\mathbb{R},\,h(\mathbf{U}^\top\mathbf{x}_i)\text{ and }z_i\text{ are statistically independent}\Big\},\qquad(2)$$

that is no classifier can predict the demographic attribute $z_i\in\{0,1\}$ (will generalize to non-binary $z_i$ later) when getting to see only the projection of $\mathbf{x}_i$ onto the $k$-dimensional subspace.

for a given $\mathbf{X}$ and target dimension $k$ the set $\mathcal{U}$ defined in (2) may be empty $\Rightarrow$ we relax Problem (2) in two ways:

- ➢ only linear functions $h(\mathbf{x})=\mathbf{w}^\top\mathbf{x}+b$ rather than arbitrary functions $h:\mathbb{R}^k\to\mathbb{R}$
- ➢ uncorrelated rather than independent

This leaves us with the following problem:

$$\operatorname{argmax}_{\mathbf{U}\in\mathcal{U}'}\operatorname{trace}(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}),\quad\text{where }\mathcal{U}'=\Big\{\mathbf{U}\in\mathbb{R}^{d\times k}:\mathbf{U}^\top\mathbf{U}=\mathbf{I}_{k\times k}\text{ and }\forall\mathbf{w}\in\mathbb{R}^k,b\in\mathbb{R},\,\mathbf{w}^\top\mathbf{U}^\top\mathbf{x}_i+b\text{ and }z_i\text{ are uncorrelated, that is }\operatorname{Cov}(\mathbf{w}^\top\mathbf{U}^\top\mathbf{x}_i+b,z_i)=0\Big\}\qquad(3)$$

Problem (3) can be solved analytically: with $\bar{z}=\frac{1}{n}\sum_{i=1}^n z_i$ and $\mathbf{z}=(z_1-\bar{z},\dots,z_n-\bar{z})^\top\in\mathbb{R}^n$,

$$\forall\mathbf{w}\in\mathbb{R}^k,b\in\mathbb{R}:\mathbf{w}^\top\mathbf{U}^\top\mathbf{x}_i+b\text{ and }z_i\text{ are uncorrelated}\;\Leftrightarrow\;\mathbf{z}^\top\mathbf{X}^\top\mathbf{U}=\mathbf{0}.$$

With a change of variables, this leads to the following algorithm with running time $\mathcal{O}(nd^2+d^3)$ (same as the running time of standard PCA):

**Algorithm 1** Fair PCA (for two demographic groups)

**Input:** data matrix $\mathbf{X}\in\mathbb{R}^{d\times n}$; demographic attribute $z_i\in\{0,1\}$, $i\in\{1,\dots,n\}$; target dimension $k\in\{1,\dots,d-1\}$

**Output:** a solution $\mathbf{U}$ to Problem (3)

- set $\mathbf{z}=(z_1-\bar{z},\dots,z_n-\bar{z})^\top$ with $\bar{z}=\frac{1}{n}\sum_{i=1}^n z_i$
- compute an orthonormal basis of the nullspace of $\mathbf{z}^\top\mathbf{X}^\top$ and build matrix $\mathbf{R}$ comprising the basis vectors as columns
- compute orthonormal eigenvectors, corresponding to the largest $k$ eigenvalues, of $\mathbf{R}^\top\mathbf{X}\mathbf{X}^\top\mathbf{R}$ and build matrix $\mathbf{\Lambda}$ comprising the eigenvectors as columns
- return $\mathbf{U}=\mathbf{R}\mathbf{\Lambda}$

Alternative interpretation of the constraint:

$$\mathbf{z}^\top\mathbf{X}^\top\mathbf{U}=\mathbf{0}\quad\Leftrightarrow\quad\frac{1}{|\{i:z_i=0\}|}\sum_{i:z_i=0}\mathbf{U}^\top\mathbf{x}_i=\frac{1}{|\{i:z_i=1\}|}\sum_{i:z_i=1}\mathbf{U}^\top\mathbf{x}_i$$

Hence, fair PCA finds a best-approximating projection such that the projected data's group-conditional means coincide.

## Variants & extensions

**Trading Off Accuracy vs. Fairness:** We can trade off accuracy vs. fairness by using the representation $(\mathbf{U}_{\text{fair}}^\top\mathbf{x};\lambda\cdot\mathbf{U}_{\text{st}}^\top\mathbf{x})\in\mathbb{R}^{2k}$, where $\mathbf{U}_{\text{fair}}^\top$ and $\mathbf{U}_{\text{st}}^\top$ are the projection matrices of fair PCA and standard PCA, respectively, and $\lambda\in[0,1]$. For $\lambda\ll 1$, any ML model trained with weight regularization will have troubles to exploit the standard PCA representation.

**Adaptation to Equal Opportunity:** We can use fair PCA to aim for equality of opportunity (requiring $\Pr(\hat{Y}=1|Z=z,Y=1)=\Pr(\hat{Y}=1|Y=1)$) in a specific downstream task by fitting the transformation (i.e., applying Alg. 1) only on datapoints $\mathbf{x}_i$ with $y_i=1$.

**Kernelized Version:** Using the representer theorem, we can kernelize fair PCA.

**Multiple Groups or Multiple Demographic Attributes:** By replacing the vector $\mathbf{z}$ in Algorithm 1 with an appropriate matrix $\mathbf{Z}$, we can adapt fair PCA to obfuscate demographic information for multiple demographic groups (e.g., defined by race) or for multiple demographic attributes simultaneously (e.g., gender *and* race).

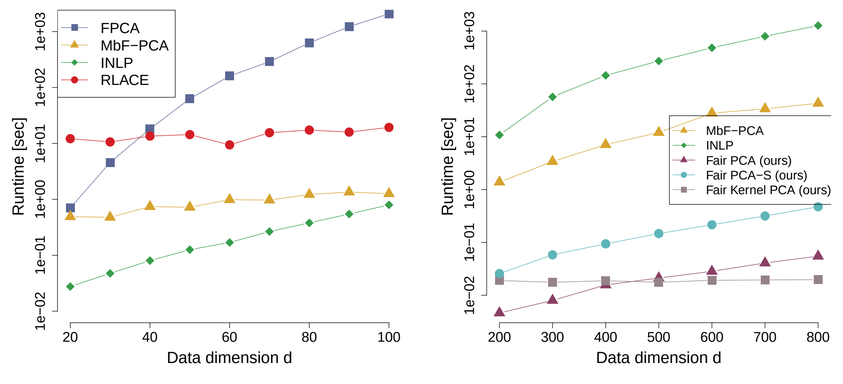
**Equalizing Group-Conditional Covariance Matrices:** Fair PCA equalizes the projected data's group-conditional means. We provide a simple strategy to also approximately equalize the projected data's group-conditional covariance matrices (requires $k$ to be small).

## Experiments

### Comparison with existing methods

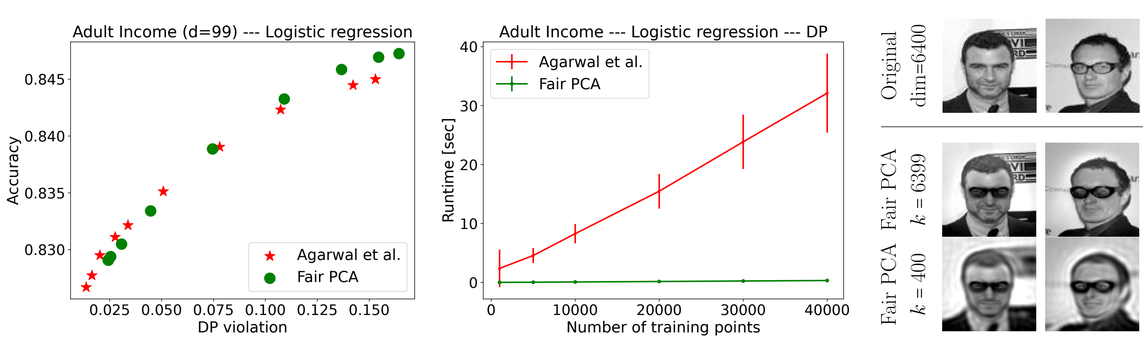
**Adult Income** [feature dim = 97, $\Pr(Y=1)=0.2489$] with **target dimension** $k=10$

| Algorithm | %Var(↑) | MMD²(↓) | %Acc(↑) Kernel SVM | $\Delta_{DP}$(↓) Kernel SVM | %Acc(↑) Linear SVM | $\Delta_{DP}$(↓) Linear SVM | %Acc(↑) MLP | $\Delta_{DP}$(↓) MLP |
|---|---|---|---|---|---|---|---|---|
| PCA | $\mathbf{21.77_{1.95}}$ | $\mathbf{0.195_{0.006}}$ | $\mathbf{93.64_{0.87}}$ | $\mathbf{0.16_{0.01}}$ | $\mathbf{82.68_{0.96}}$ | $\mathbf{0.18_{0.02}}$ | $\mathbf{89.06_{2.07}}$ | $\mathbf{0.2_{0.03}}$ |
| FPCA (0.1, 0.01) [2] | $15.75_{1.14}$ | $0.006_{0.003}$ | $91.94_{0.84}$ | $0.13_{0.02}$ | $78.1_{2.15}$ | $0.03_{0.02}$ | $\mathbf{87.17_{1.1}}$ | $0.11_{0.04}$ |
| FPCA (0, 0.01) [2] | $15.52_{1.12}$ | $0.004_{0.002}$ | $91.66_{0.92}$ | $0.13_{0.02}$ | $77.72_{2.06}$ | $0.03_{0.02}$ | $85.38_{2.08}$ | $0.09_{0.03}$ |
| MbF-PCA ($10^{-3}$) [3] | $\mathbf{18.86_{1.46}}$ | $0.005_{0.002}$ | $\mathbf{93.06_{0.85}}$ | $0.15_{0.01}$ | $\mathbf{80.53_{1.31}}$ | $0.03_{0.02}$ | $86.83_{2.05}$ | $0.08_{0.03}$ |
| MbF-PCA ($10^{-6}$) [3] | $12.36_{4.15}$ | $\mathbf{0.002_{0.001}}$ | $83.58_{3.58}$ | $\mathbf{0.05_{0.02}}$ | $75.11_{1.66}$ | $\mathbf{0.0_{0.0}}$ | $80.27_{3.55}$ | $\mathbf{0.04_{0.04}}$ |
| INLP [4] | $10.79_{0.84}$ | $0.004_{0.001}$ | $89.01_{1.19}$ | $0.1_{0.03}$ | $75.11_{1.66}$ | $\mathbf{0.0_{0.0}}$ | $85.24_{1.2}$ | $0.11_{0.03}$ |
| RLACE [5] | $10.3_{0.49}$ | $0.007_{0.005}$ | $90.96_{1.04}$ | $0.12_{0.04}$ | $75.11_{1.66}$ | $\mathbf{0.0_{0.0}}$ | $86.61_{1.69}$ | $0.11_{0.05}$ |
| Fair PCA (ours) | $\mathbf{19.62_{1.73}}$ | $0.014_{0.003}$ | $\mathbf{93.42_{0.8}}$ | $0.16_{0.01}$ | $\mathbf{81.31_{1.23}}$ | $0.04_{0.02}$ | $\mathbf{88.16_{1.45}}$ | $0.16_{0.03}$ |
| Fair Kernel PCA (ours) | *n/a* | *n/a* | $79.9_{1.54}$ | $\mathbf{0.04_{0.03}}$ | $78.34_{1.21}$ | $0.02_{0.02}$ | $80.52_{2.05}$ | $\mathbf{0.06_{0.03}}$ |
| Fair PCA-S (0.5) (ours) | $12.75_{1.31}$ | $\mathbf{0.004_{0.001}}$ | $86.85_{1.79}$ | $0.09_{0.02}$ | $75.11_{1.66}$ | $\mathbf{0.0_{0.0}}$ | $82.89_{2.01}$ | $0.07_{0.04}$ |
| Fair PCA-S (0.85) (ours) | $15.79_{1.05}$ | $0.005_{0.001}$ | $91.81_{1.05}$ | $0.15_{0.02}$ | $75.11_{1.66}$ | $\mathbf{0.0_{0.0}}$ | $87.07_{1.4}$ | $0.15_{0.02}$ |




*Running time of the various methods as a function of the data dimension $d$. The target dimension $k$ is 5 independent of $d$.*

### Comparison with the state-of-the-art in-processing bias mitigation method of Agarwal et al. [6] and fair PCA applied to CelebA




*Left: Accuracy vs. fairness curves for the method of Agarwal et al. [6] and the trade-off variant of fair PCA. Middle: Runtime as a function of the number of training points. Right: Fair PCA applied to the CelebA dataset [7] to erase the concept of "glasses".*

## References

[1] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML*, 2013.

[2] M. Olfat and A. Aswani. Convex formulations for fair principal component analysis. In *AAAI*, 2019.

[3] J. Lee, G. Kim, M. Olfat, M. Hasegawa-Johnson, and C. Yoo. Fast and efficient MMD-based fair PCA via optimization over Stiefel manifold. In *AAAI*, 2022.

[4] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

[5] S. Ravfogel, M. Twiton, Y. Goldberg, and R. Cotterell. Linear adversarial concept erasure. In *ICML*, 2022.

[6] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In *ICML*, 2018.

[7] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.