Fair Streaming Principal Component Analysis: Statistical and Algorithmic Viewpoint

Junghyun Lee*, Hanseul Cho*, Se-Young Yun, Chulhee Yun

Kim Jaechul Graduate School of AI, KAIST

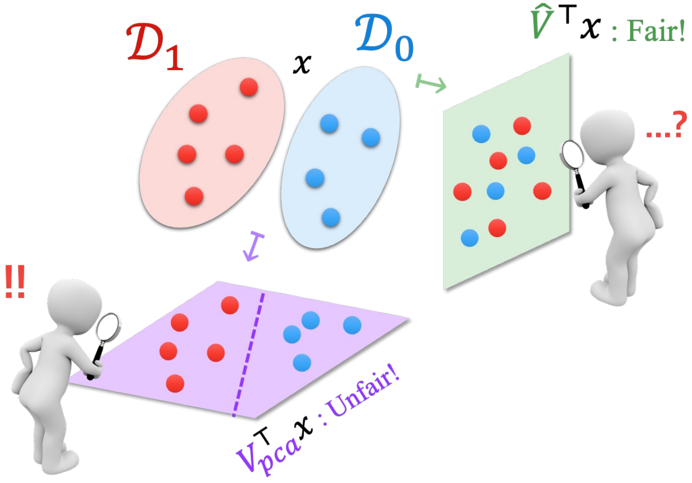
NEURAL INFORMATION PROCESSING SYSTEMS

Fair PCA: Problem Setting

- Group fairness scenario, with binary[1] sensitive attribute $a \in \{0,1\}$
  - e.g., {young, old}, {rich, poor}, {male, female}

Given.

- Samples from a mixture of $\mathcal{D}_0$ and $\mathcal{D}_1$ of the form (a, x)
  - $\mathcal{D}_a$'s covariance is $\boldsymbol{\Sigma}_a$; the total covariance is $\boldsymbol{\Sigma}$



Our Goal.

- Output a loading matrix $\boldsymbol{V} \in \mathbb{R}^{d\times k}, \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_k$ such that
  - **Explained variance (PCA)**: maximize $\mathrm{tr}(\boldsymbol{V}^T\boldsymbol{\Sigma}\boldsymbol{V})$
  - **Representation fairness**: make the (conditional) distributions after PCA *indistinguishable*
    [Olfat & Aswani, AAAI'19; Lee et al., AAAI'22, Kleindessner et al., AISTATS'23]

[1]In our paper, we provide discussions on how to extend this to multiple sensitive groups and non-binary attributes

Unsolved Problems in Fair PCA

Statistical Viewpoint

- No statistical framework
  - PAC-type definition
  - Sample complexity guarantee
- Use of several relaxations without theoretical justifications
  [Olfat & Aswani, AAAI’19; Kleindessner et al., AISTATS’23]

Algorithmic Viewpoint

- Too much memory requirement
  - Require loading the whole data
  - Require computing the entire (empirical) covariance matrix
- Streaming setting?
  [Mitliagkas et al., NIPS’13]

Contribution #1. **Statistical Viewpoint**

“Null It Out” Formulation of Fair PCA

- We define the directions to be *nullified* *[Rafovgel et al., ACL’20]* as follows:
  1. mean difference $\boldsymbol{f} := \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$
  2. top m eigenvectors $\boldsymbol{P}_m$ of the covariance difference $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0$

$$\max_{\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}_k} \operatorname{tr}(\boldsymbol{V}^T\boldsymbol{\Sigma}\boldsymbol{V}), \qquad \text{subject to } \boldsymbol{V} \perp \boldsymbol{f} \text{ and } \boldsymbol{V} \perp \boldsymbol{P}_m$$

$$\Leftrightarrow \max_{\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}_k} \operatorname{tr}\left(\boldsymbol{V}^T\boldsymbol{\Pi}_{\boldsymbol{U}}^{\perp}\boldsymbol{\Sigma}\boldsymbol{\Pi}_{\boldsymbol{U}}^{\perp}\boldsymbol{V}\right)$$

where $\boldsymbol{\Pi}_{\boldsymbol{U}}^{\perp} := \boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^T$ and $\boldsymbol{U}$ is a semi-*orthogonal* matrix whose columns form a basis of $\operatorname{col}([\boldsymbol{P}_m|\boldsymbol{f}])$.

$\boldsymbol{V}^\star$ is the solution to the above.

PAFO-Learnability

- We propose a learnability framework for fair PCA!

Definition 2. A collection $\mathcal{F}_d$ of tuples $(\mathcal{D}_0, \mathcal{D}_1, p)$ is ***PAFO**** ***-learnable for PCA*** if for any accuracy levels $\varepsilon_{\mathrm{o}}, \varepsilon_{\mathrm{f}} \in (0,1)$ and confidence level $\delta \in (0,1)$, with sufficiently many samples** from $\mathcal{D} = p\mathcal{D}_1 + (1-p)\mathcal{D}_0$, we can obtain $\widehat{\boldsymbol{V}}$ satisfying the following with probability at least $1-\delta$:

$$\mathrm{tr}\left(\widehat{\boldsymbol{V}}^T \boldsymbol{\Sigma} \widehat{\boldsymbol{V}}\right) \geq \mathrm{tr}\left(\boldsymbol{V}^{\star T} \boldsymbol{\Sigma} \boldsymbol{V}^{\star}\right) - \varepsilon_{\mathrm{o}}, \qquad \left\|\boldsymbol{\Pi}_{\mathbf{U}} \widehat{\boldsymbol{V}}\right\| \leq \varepsilon_{\mathrm{f}}.$$

Optimality   Fairness

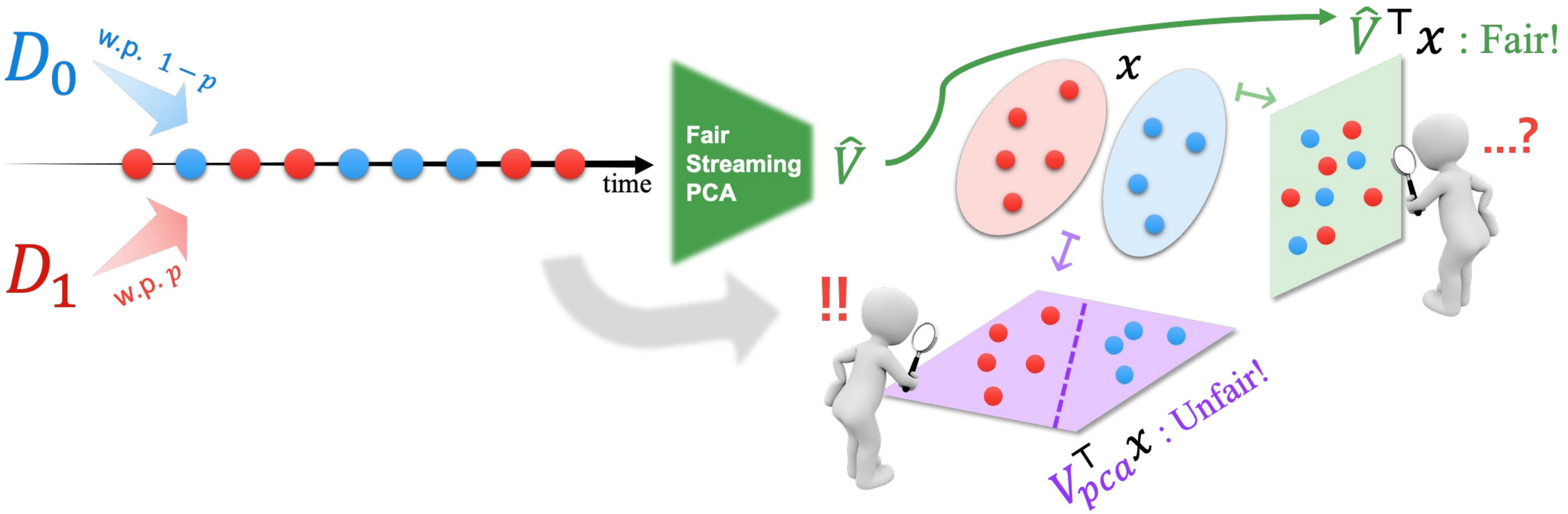
* *Probably Approximately Fair and Optimal*

** *sample complexity* depends on $\varepsilon_{\mathrm{o}}, \varepsilon_{\mathrm{f}}, \delta$, and distribution-dependent quantities.

Contribution #2. **Algorithmic Viewpoint**

Fair *Streaming* PCA

- A new problem setting called ***fair streaming PCA*** that accounts for memory limitation common in big data regimes:



- Here, the learner can use only $o(d^2)$ memory!
  - To be precise, $O(d \max(k, m))$ memory, where k is the target dimension and m is the nullifying dimension.

Fair Noisy Power Method (FNPM)

NEURAL INFORMATION PROCESSING SYSTEMS

- We then propose a new algorithm, the ***Fair Noisy Power Method*** *(FNPM)*
  - A two-phase algorithm based on the noisy power method [Hardt & Price, NIPS'14]

Phase 1. Estimate $\boldsymbol{U}$:

for $t \in [T]$ **do**
  Sample b data points;
  $\boldsymbol{W}_t \leftarrow \mathrm{QR}\big((\widehat{\boldsymbol{\Sigma}}_{1,t} - \widehat{\boldsymbol{\Sigma}}_{0,t})\boldsymbol{W}_{t-1}\big)$;
end
$\widehat{\boldsymbol{f}}$ ←MLE estimator of $\boldsymbol{f}$;
$\widehat{\boldsymbol{g}} \leftarrow \frac{\boldsymbol{\Pi}^{\perp}_{\boldsymbol{W}_T}\widehat{\boldsymbol{f}}}{\left\|\boldsymbol{\Pi}^{\perp}_{\boldsymbol{W}_T}\widehat{\boldsymbol{f}}\right\|}$;
return $\widehat{\boldsymbol{U}} = [\boldsymbol{W}_T \mid \widehat{\boldsymbol{g}}]$

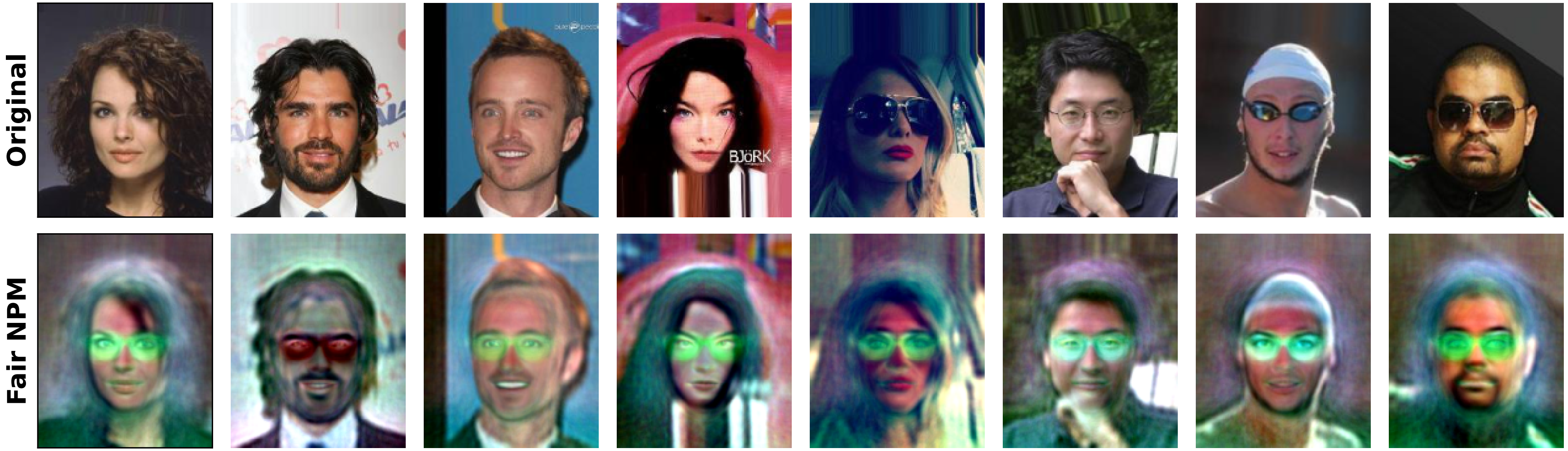
Phase 2. Obtain the final $\widehat{\boldsymbol{V}}$:

for $\tau \in [\mathcal{T}]$ **do**
  Sample B data points;
  $\boldsymbol{V}_\tau \leftarrow \mathrm{QR}\big(\boldsymbol{\Pi}^{\perp}_{\widehat{\boldsymbol{U}}}\widehat{\boldsymbol{\Sigma}}_\tau\boldsymbol{\Pi}^{\perp}_{\widehat{\boldsymbol{U}}}\boldsymbol{V}_{\tau-1}\big)$;
end
return $\widehat{\boldsymbol{V}} = \boldsymbol{V}_{\mathcal{T}}$

- We also provide a sample complexity guarantee of FNPM
  - the first of its kind in the fair PCA literature!

Experiments

- ***Full-color, original resolution*** CelebA Dataset
    - All 162,770 images can*not* be loaded into the memory of a moderate-sized computer
- Transform the setting to *streaming* and apply our FNPM!
- The most scalable fair PCA algorithm to date!



Sensitive attribute: Eyeglasses

See you at Poster Session #1! (Dec 12 Tue)

Location: Great Hall & Hall B1+B2 #1600

Full paper (arXiv)

GitHub link