Fair Streaming Principal Component Analysis: Statistical and Algorithmic Viewpoint

Junghyun Lee*, Hanseul Cho*, Se-Young Yun, Chulhee Yun | { jh_lee00, jhs4015 }@kaist.ac.kr

KAIST AI Kim Jaechul Graduate School

OptiML Optimization & Machine Learning

OSI Optimization and Statistical Inference LAB

NEURAL INFORMATION PROCESSING SYSTEMS

Contribution

- **Statistical Viewpoint**:
  - New formulation of fair PCA ("Null It Out" formulation)
  - New statistical framework (PAFO-Learnability)
- **Algorithmic Viewpoint**:
  - New problem setting (Fair Streaming PCA)
  - New algorithm (Fair Noisy Power Method, FNPM)
- Combining these, we rigorously prove a sample complexity guarantee of FNPM for fair streaming PCA

Problem Setting : Fair PCA

<u>Given.</u>

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 $\boldsymbol{\Sigma}$

<u>Our Goal.</u>

Output a loading matrix $\boldsymbol{V} \in \mathbb{R}^{d\times k}, \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_k$ $(k < d)$ such that

• **Explained variance (PCA)**: maximize $\mathrm{tr}(\boldsymbol{V}^\top \boldsymbol{\Sigma} \boldsymbol{V})$

• **Representation fairness**: make the (conditional) distributions after PCA *indistinguishable*

[Olfat & Aswani, AAAI'19; Lee et al., AAAI'22, Kleindessner et al., AISTATS'23]

"Null It Out" Formulation of Fair PCA

Directions to be <u>nullified</u> [Rafovgel et al., ACL'20]:

1. Mean difference $\boldsymbol{f} \coloneqq \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$
2. Top m eigenvectors $\boldsymbol{P}_m$ of the covariance difference $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0$

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

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

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

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

Probably Approximately Fair and Optimal Learnability

Definition. Let d, k, m be given $(1 \le k < d,\ m < d)$. A collection $\mathcal{F}_d$ of tuples $(\mathcal{D}_0, \mathcal{D}_1, p)$ is said to be ***PAFO-learnable for PCA*** if there exists a $N_{\mathcal{F}_d}: (0,1)^3 \to \mathbb{N}$ and an algorithm $\mathcal{A}$ satisfying: for any accuracy levels $\varepsilon_o, \varepsilon_f \in (0,1)$ and confidence level $\delta \in (0,1)$, with <u>sufficiently many samples</u> ($\ge N_{\mathcal{F}_d}(\varepsilon_o, \varepsilon_f, \delta)$) from

$$\mathcal{D} = p\mathcal{D}_1 + (1-p)\mathcal{D}_0,$$

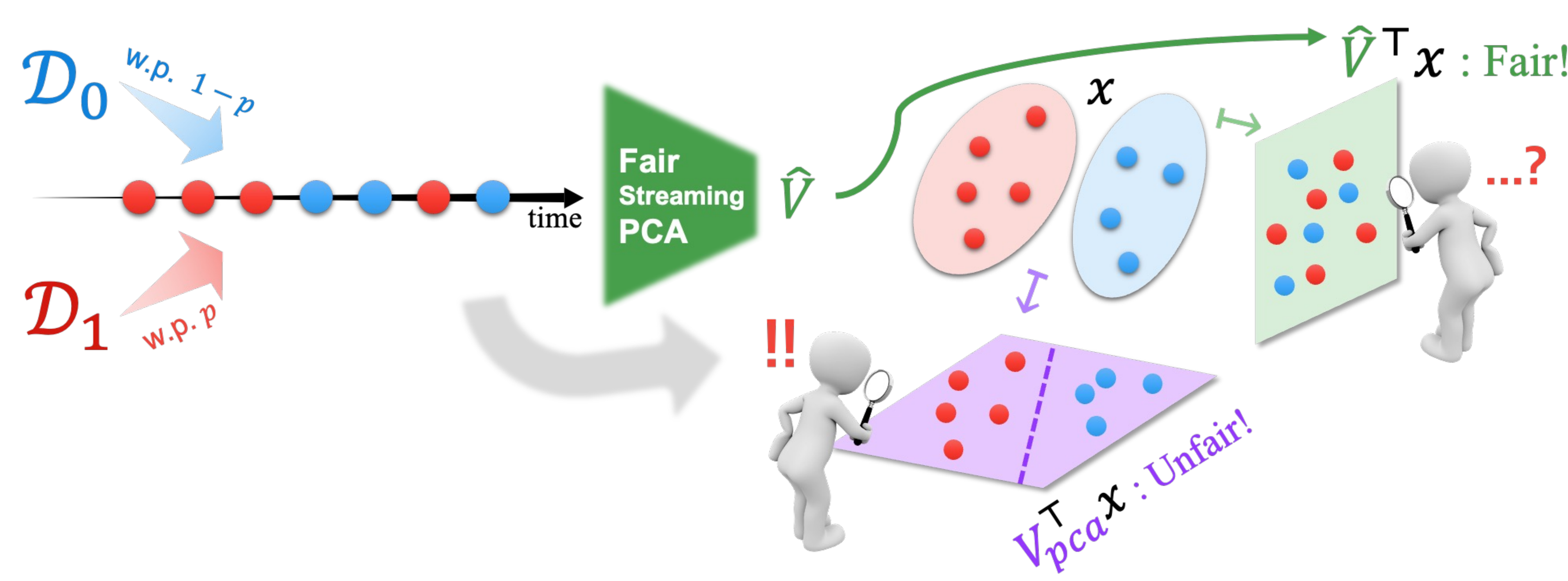
$\mathcal{A}$ returns $\widehat{\boldsymbol{V}} \in \mathbb{R}^{d\times k}, \widehat{\boldsymbol{V}}^\top \widehat{\boldsymbol{V}} = \boldsymbol{I}_k$ s.t. with probability at least $1-\delta$:

$$\mathrm{tr}\left(\widehat{\boldsymbol{V}}^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{V}}\right) \ge \mathrm{tr}\left(\boldsymbol{V}^{\star\top} \boldsymbol{\Sigma} \boldsymbol{V}^\star\right) - \varepsilon_o, \qquad \left\|\boldsymbol{\Pi}_{\boldsymbol{U}} \widehat{\boldsymbol{V}}\right\| \le \varepsilon_f.$$

Optimality **Fairness**

Fair Streaming PCA

Here, the learner can use only $o(d^2)$ memory!



Fair Noisy Power Method (FNPM)

<u>Phase 1. Estimate $\boldsymbol{U}$:</u>

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

<u>Phase 2. Obtain the final $\widehat{\boldsymbol{V}}$:</u>

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

A two-phase algorithm based on the noisy power method [Hardt & Price, NIPS'14]

Sample Complexity Guarantee

Assumption. (Informal) The collection $\mathcal{F}_d \subset \mathcal{P}_d \times \mathcal{P}_d \times (0,1)$ satisfies the following: for any $(\mathcal{D}_0, \mathcal{D}_1, p) \in \mathcal{F}_d$,

- k^{th} and m^{th} singular value gap of $\boldsymbol{\Pi}_{\boldsymbol{U}}^\perp \boldsymbol{\Sigma} \boldsymbol{\Pi}_{\boldsymbol{U}}^\perp$ and $\boldsymbol{S}$ are bounded below by $\Delta_{\boldsymbol{\Sigma},k}$ and $\Delta_{\boldsymbol{S},m}$ (resp.).
- $\|(\boldsymbol{I} - \boldsymbol{P}_m \boldsymbol{P}_m^\top)\boldsymbol{f}\| \in \{0\} \cup [g_{\min}, f_{\max}], \quad p \in [p_{\min}, 1 - p_{\min}]$

Theorem. (Informal) FNPM guarantees PAFO-learnability with the sample complexity $N_{\mathcal{F}_d}(\varepsilon_f, \varepsilon_o, \delta) = N_1(\varepsilon_f, \varepsilon_o, \delta) + N_2(\varepsilon_o, \delta)$, with

$$N_1(\varepsilon_f, \varepsilon_o, \delta) = \Omega\left(\frac{1}{p_{\min}\Delta_{\boldsymbol{S},m}^3}\left(\frac{dm}{\delta^2} + \alpha\left(\frac{1}{\varepsilon_f^2} + \Delta_{\boldsymbol{\Sigma},k} N_2(\varepsilon_o, \delta)\right)\right)\right),$$

$$N_2(\varepsilon_o, \delta) = \Omega\left(\frac{1}{\Delta_{\boldsymbol{\Sigma},k}^3}\left(\frac{dk}{\delta^2} + \frac{k^2}{\varepsilon_o^2}\right)\right),$$

where $\alpha \coloneqq 1 + \frac{1[\boldsymbol{g} \neq \boldsymbol{0}] f_{\max}^2}{g_{\min}^2}$.

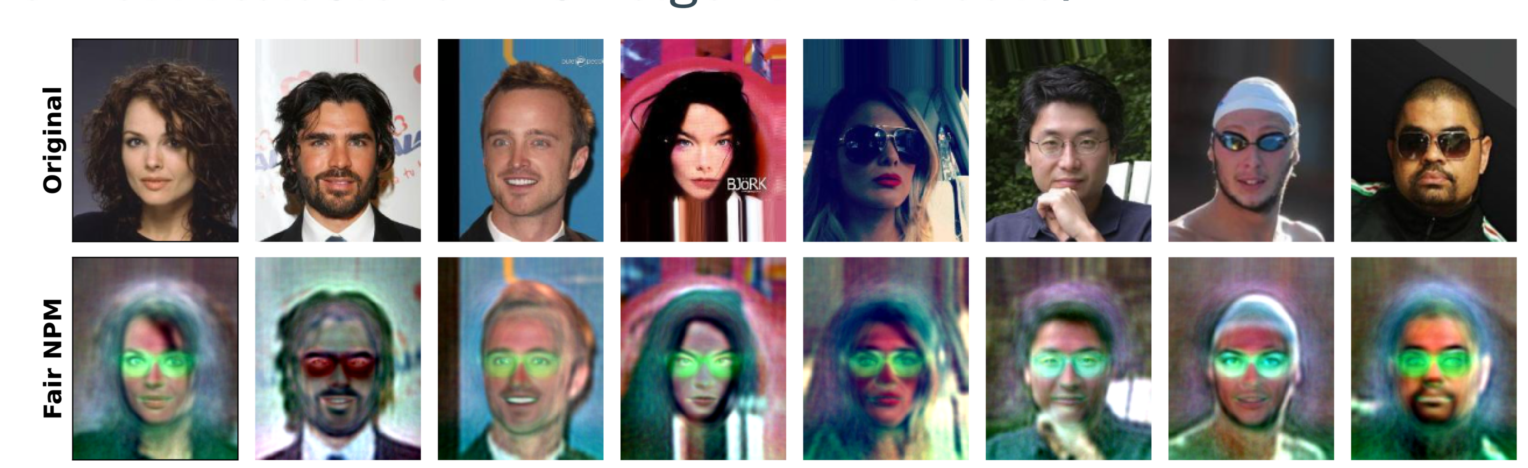
Experiment: Full-color CelebA dataset

Full-color, original resolution CelebA Dataset

- All 162,770 images can *not* be loaded onto a moderate-sized computer

Transform the setting to *streaming* and apply our FNPM!

The most scalable fair PCA algorithm to date!



Sensitive attribute: Eyeglasses

References

M. Olfat & A. Aswani. "Convex Formulations of Fair Principal Component Analysis." In: *AAAI* 2019.

J. Lee, G. Kim, M. Olfat, M. Hasegawa-Johnson, and C. D. Yoo. "Fast and Efficient MMD-based Fair PCA via Optimization over Stiefel Manifold." In: *AAAI* 2022.

M. Kleindessner, M. Donini, C. Russel, and M. B. Zafar. "Efficient fair PCA for fair representation learning." In: *AISTATS* 2023.

S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. "Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection." In: *ACL* 2020.

M. Hardt & E. Price. "The Noisy Power Method: A Meta-Algorithm with Applications." In: *NIPS* 2014.