Fast and Efficient MMD-based Fair PCA via Optimization over Stiefel Manifold

Junghyun Lee[1], Gwangsu Kim[2], Matt Olfat[3,4], Mark Hasegawa-Johnson[5], Chang D. Yoo[2]

[1]KAIST AI [2]KAIST EE [3]IEOR, UC Berkeley [4]Citadel [5]Dept. of ECE, UIUC

December 21, 2022

Outline

1. Introduction
2. Review of FPCA
   - Adversarial Definition [Olfat and Aswani, 2019]
   - Problems with FPCA
3. MbF-PCA
   - New Definition: Δ-fairness
   - Manifold Optimization for MbF-PCA
4. Experiments
5. Conclusion

Fair Machine Learning

- An active area of research with enormous societal impact
  - cf. Machine Bias [Angwin et al., 2016] - Black vs White Defendant's recidivism scores
- Machine learning algorithms should not be dependent on specific (sensitive) variables such as gender, age, race...etc.

	White	Black
Higher risk, yet didn't re-offend	23.5%	44.9%
Lower risk, yet did re-offend	47.7%	28.0%

Fair Machine Learning

- There are multiple frameworks on how to do this:
  - Fair supervised learning
  - Fair unsupervised learning
  - **Fair representation learning** [Zemel et al., 2013, Cisse and Koyejo, 2019]
  - Fair data preprocessing
  - ...etc.
- Some useful resources:
  - `https://fairmlbook.org/pdf/fairmlbook.pdf`
  - `https://dl.acm.org/doi/pdf/10.1145/3457607`

Mathematically speaking, (in my humble opinion), many of the algorithmic fair ML problems can be formulated as *(constrained) optimizations*! (i.e. optimizationists(?)' roles are very important)

Problem Setting

- $\{x_i\}_{i=1}^n \subset \mathbb{R}^p$: original given data points (as row vectors)
  - $X \in \mathbb{R}^{n \times p}$: data matrix
  - Σ: empirical covariance matrix
- X is composed of *two* groups, which correspond to the protected classes (e.g. gender, age)
- $d < p$: dimension to which we want to reduce to
- $V \in \mathbb{R}^{p \times d}$: linear projection matrix (in case of PCA, $V^\intercal V = \mathbb{I}_d$)

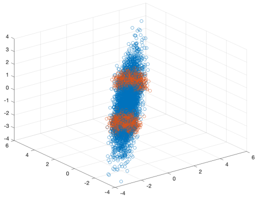
Problem Setting

- $\{x_i\}_{i=1}^n \subset \mathbb{R}^p$: original given data points (as row vectors)
  - $X \in \mathbb{R}^{n \times p}$: data matrix
  - Σ: empirical covariance matrix
- X is composed of *two* groups, which correspond to the protected classes (e.g. gender, age)
- $d < p$: dimension to which we want to reduce to
- $V \in \mathbb{R}^{p \times d}$: linear projection matrix (in case of PCA, $V^\intercal V = \mathbb{I}_d$)
- Main objectives:

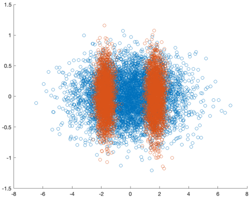
> - Maximize $\langle \Sigma, VV^\intercal \rangle$: *explained variance* of X after applying (linear) PCA using V.
> - Minimize fairness: *to be defined/discussed*

Problem setting

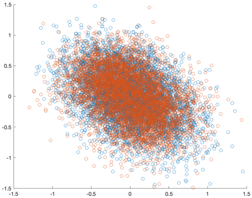
Fair PCA: the problem of maximizing the explained variance while imposing *distribution similarity after projection*!



(a) Original data



(b) Vanilla PCA



(c) Fair PCA

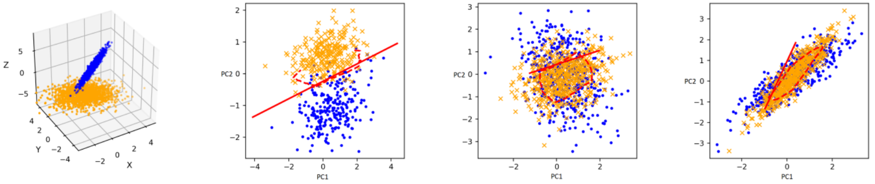
Outline

Adversarial Definition: FPCA

- To the best of our knowledge, [Olfat and Aswani, 2019] is the *only* prior work that considered this notion of fair PCA, in which they proposed the following adversarial definition, referred to as *FPCA*:

Definition (Δ_A-fairness, [Olfat and Aswani, 2019] (Informal))

The dimensionality reduction $\Pi : \mathbb{R}^p \to \mathbb{R}^d$ is $\Delta_A(h)$-fair if adversarial classifiers that try to classify the protected class perform poorly in the projected space; the fairness metric is defined in terms of the difference between true positive and false positive.



(a) Original data (b) PCA (c) FPCA - Mean con. (d) FPCA - Both con.

Figure 1: Comparison of PCA and FPCA on synthetic data. In each plot, the thick red line is the optimal linear SVM separating

SDP formulation of FPCA

- [Olfat and Aswani, 2019] provided an SDP formulation of fair PCA[1]:

$$\max \langle X^\top X, P\rangle - \mu t \tag{7a}$$
$$\text{s.t. } \operatorname{trace}(P) \le d,\ \mathbb{I} \succeq P \succeq 0 \tag{7b}$$
$$\langle P, ff^\top\rangle \le \delta^2 \tag{7c}$$
$$\begin{bmatrix} t\mathbb{I} & PM_+ \\ M_+^\top P & \mathbb{I} \end{bmatrix} \succeq 0, \tag{7d}$$
$$\begin{bmatrix} t\mathbb{I} & PM_- \\ M_-^\top P & \mathbb{I} \end{bmatrix} \succeq 0 \tag{7e}$$

where $M_i M_i^\top$ is the Cholesky decomposition of $iQ + \varphi\mathbb{I}$ $(i \in \{-, +\})$, $\varphi \ge \|\widehat{\Sigma}_+ - \widehat{\Sigma}_-\|_2$, (7c) is called the *mean constraint* and denotes the use (5), and (7d) and (7e) are called the *covariance constraints* and are the SDP reformulation of (6). Our convex formulation for FPCA consists of solving (7) and then extracting the d largest eigenvectors from the optimal P^*.

Figure: δ: bound for mean difference, μ: bound for covariance difference

[1] This was heavily inspired from the SDP formulation of vanilla PCA [Arora et al., 2013].

Problems with the Definition of FPCA

$\widehat{\Delta}_A(\mathcal{F})$ can**not** be computed exactly nor efficiently.

$$\widehat{\Delta}_A(\mathcal{F}_c) := \sup_{h \in \mathcal{F}_c} \sup_{t} \left| \frac{1}{|P|} \sum_{i \in P} l_i(\Pi, h_t) - \frac{1}{|N|} \sum_{i \in N} l_i(\Pi, h_t) \right|$$

Problems with the Definition of FPCA

$\widehat{\Delta}_A(\mathcal{F})$ can**not** be computed exactly nor efficiently.

$$\widehat{\Delta}_A(\mathcal{F}_c) := \sup_{h \in \mathcal{F}_c} \sup_{t} \left| \frac{1}{|P|} \sum_{i \in P} l_i(\Pi, h_t) - \frac{1}{|N|} \sum_{i \in N} l_i(\Pi, h_t) \right|$$

$\widehat{\Delta}_A(\mathcal{F})$ may be asymptotically **inconsistent**.

Proposition ([Olfat and Aswani, 2019])

Consider a fixed family of classifiers $\mathcal{F}_c$. Then for any $\delta > 0$, with probability at least $1 - \exp\left(-\frac{(n+m)\delta^2}{2}\right)$ the following holds:

$$\left| \Delta_A(\mathcal{F}_c) - \widehat{\Delta}_A(\mathcal{F}_c) \right| \leq 8 \sqrt{\frac{VC(\mathcal{F}_c)}{m+n}} + \delta.$$

Problems with the SDP Formulation of FPCA

- The SDP is **inscalable** to high dimensional input data.
- The resulting solution is **suboptimal** due to the SDP relaxations

- Instead of dealing with V directly, [Olfat and Aswani, 2019] optimize w.r.t. $P = VV^\intercal \in \mathbb{R}^{p \times p}$
- The orthogonality constraint $V^\intercal V = \mathbb{I}_d$ becomes $\text{rank}(P) \leq d$, which was then *relaxed*[2] to $\text{tr}(P) \leq d$.

[2] This is exact when there's no additional constraints [Olfat and Aswani, 2019]

Problems with the SDP Formulation of FPCA

- The SDP is **inscalable** to high dimensional input data.
- The resulting solution is **suboptimal** due to the SDP relaxations

- Instead of dealing with V directly, [Olfat and Aswani, 2019] optimize w.r.t. $P = VV^\intercal \in \mathbb{R}^{p \times p}$
- The orthogonality constraint $V^\intercal V = \mathbb{I}_d$ becomes $\mathrm{rank}(P) \leq d$, which was then *relaxed*[2] to $\mathrm{tr}(P) \leq d$.

As the fairness constraints were derived under **Gaussian assumption**, they do *not* ensure an exact distribution equality.

- Their SDP assumes that the underlying datas are Gaussian.
  - Two *projected* sensitive groups have different distributions, yet have the same first and second moments.

[2] This is exact when there's no additional constraints [Olfat and Aswani, 2019]

Outline

Maximum mean discrepancy (MMD)

- We need a new definition of fairness in PCA that can
  - directly lead to a tractable and exact optimization
  - intuitive and be more easily interpretable

Maximum mean discrepancy (MMD)

- We need a new definition of fairness in PCA that can
  - directly lead to a tractable and exact optimization
  - intuitive and be more easily interpretable

Definition ([Gretton et al., 2007])

Given $\mu, \nu \in \mathcal{P}_d$ and a positive-definite kernel k, their **maximum mean discrepancy (MMD)** is a pseudo-metric on $\mathcal{P}_d$, defined as follows[a]:

$$MMD_k(\mu, \nu) := \sup_{f \in \mathcal{H}_k} \left| \int_{\mathbb{R}^d} f \; d(\mu - \nu) \right|$$

[a] $\mathcal{P}_d$ is the set of all possible probability measures defined on $\mathbb{R}^d$; $\mathcal{H}_k$ is the Reproducing Kernel Hilbert Space (RKHS) generated by k

With characteristic kernels (ex. RBF kernel), MMD_k becomes a *metric* on $\mathcal{P}_d$ [Fukumizu et al., 2008].

Maximum mean discrepancy (MMD)

- We need a new definition of fairness in PCA that can
  - directly lead to a tractable and exact optimization
  - intuitive and be more easily interpretable

Definition ([Gretton et al., 2007])

Given $\mu, \nu \in \mathcal{P}_d$ and a positive-definite kernel k, their **maximum mean discrepancy (MMD)** is a pseudo-metric on $\mathcal{P}_d$, defined as follows[a]:

$$MMD_k(\mu, \nu) := \sup_{f \in \mathcal{H}_k} \left| \int_{\mathbb{R}^d} f \ d(\mu - \nu) \right|$$

[a] $\mathcal{P}_d$ is the set of all possible probability measures defined on $\mathbb{R}^d$; $\mathcal{H}_k$ is the Reproducing Kernel Hilbert Space (RKHS) generated by k

With characteristic kernels (ex. RBF kernel), MMD_k becomes a *metric* on $\mathcal{P}_d$ [Fukumizu et al., 2008].

From hereon and forth, we only consider MMD with the RBF kernel.

Contribution #1. New Fair PCA Definition: Δ-fairness

- Motivated from previous discussions, we propose a new definition for fair PCA based on MMD, referred to as MbF-PCA:

Contribution #1. New Fair PCA Definition: Δ-fairness

- Motivated from previous discussions, we propose a new definition for fair PCA based on MMD, referred to as MBF-PCA:

Definition (Δ-fairness (informal))

The dimensionality reduction $\Pi : \mathbb{R}^p \to \mathbb{R}^d$ is Δ-fair with Δ being the MMD of projected distributions, which is precisely the fairness metric.

Contribution #1. New Fair PCA Definition: Δ-fairness

- Motivated from previous discussions, we propose a new definition for fair PCA based on MMD, referred to as MbF-PCA:

Definition (Δ-fairness (informal))

The dimensionality reduction $\Pi : \mathbb{R}^p \to \mathbb{R}^d$ is Δ-fair with Δ being the MMD of projected distributions, which is precisely the fairness metric.

- Well-known properties of MMD [Gretton et al., 2007] already make it superior over the previous adversarial definition:

- $\widehat{\Delta}$ can be computed exactly and efficiently.
- $\widehat{\Delta}$ is asymptotically consistent.
- As it is a metric over $\mathcal{P}_d$, no assumption on the datas is necessary; $MMD = 0$ is itself the naturally induced fairness constraint!

Computational Efficiency

- We consider the following estimator:

$$\widehat{\Delta} := MMD(\hat{Q}_0(\Pi), \hat{Q}_1(\Pi)) \tag{1}$$

where $\hat{Q}_s(V)$ is the (nonparametric) empirical measure[3] of the projected distribution corresponding sensitive variable s.

- Unlike $\widehat{\Delta}_A$ [Olfat and Aswani, 2019], $\widehat{\Delta}$ can be computed exactly and efficiently:

Lemma ([Gretton et al., 2007])

$\widehat{\Delta}$ *is computed as follows:*

$$\widehat{\Delta} = \left[\frac{1}{m^2} \sum_{i,j=1}^{m} k(X_i, X_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(Y_i, Y_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(X_i, Y_j) \right]^{1/2}. \tag{2}$$

[3] the mixture of Dirac measures

Asymptotic Consistency

- Unlike $\widehat{\Delta}_A$ [Olfat and Aswani, 2019], $\widehat{\Delta}$ is asymptotic convergent, with the rate depending only on m and n with no function class complexity involved:

Theorem ([Gretton et al., 2007])

For any $\delta > 0$, with probability at least $1 - 2\exp\left(-\frac{\delta^2 mn}{2(m+n)}\right)$ the following holds:

$$\left|\Delta - \widehat{\Delta}\right| \leq 2\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) + \delta \tag{3}$$

Contribution #2. Fair PCA as Manifold Optimization

- All of the aformentioned problems of FPCA [Olfat and Aswani, 2019] were because *the optimization(SDP) was not directly w.r.t.* V
  - The SDP was solved w.r.t. $P = VV^\intercal \in \mathbb{R}^{p \times p}$; the final solution is obtained by the eigendecomposition of the resulting P^*.

Contribution #2. Fair PCA as Manifold Optimization

- All of the aformentioned problems of FPCA [Olfat and Aswani, 2019] were because *the optimization(SDP) was not directly w.r.t.* V
  - The SDP was solved w.r.t. $P = VV^\intercal \in \mathbb{R}^{p \times p}$; the final solution is obtained by the eigendecomposition of the resulting P^*.

Instead of trying to transform our problem into some surrogate optimization problem (ex. SDP), let us optimize **directly** for V!

$$\begin{aligned} \underset{V \in \mathbb{R}^{p \times d}}{\text{maximize}} \quad & \langle \Sigma, VV^\intercal \rangle \\ \text{subject to} \quad & V^\intercal V = \mathbb{I}_d, \\ & h(V) := MMD^2(\hat{Q}_0(V), \hat{Q}_1(V)) = 0. \end{aligned} \tag{4}$$

- Above is a smooth, nonconvex **Euclidean** optimization with *two* constraints.

Fair PCA as Manifold Optimization

- We utilize the *manifold structure of PCA*, namely, that the set of all V's with $V^\intercal V = \mathbb{I}_d$ forms the Stiefel manifold, denoted as $St(p, d)$.

Fair PCA as Manifold Optimization

- We utilize the *manifold structure of PCA*, namely, that the set of all V's with $V^\intercal V = \mathbb{I}_d$ forms the Stiefel manifold, denoted as $St(p, d)$.
- Then the previous problem can be formulated as a smooth, nonconvex **manifold (Riemannian)** with a *single* constraint, which we refer to as MbF-PCA:

$$\begin{aligned} \underset{V \in St(p,d)}{\text{maximize}} \quad & \langle \Sigma, VV^\intercal \rangle \\ \text{subject to} \quad & h(V) := MMD^2(\hat{Q}_0, \hat{Q}_1) = 0. \end{aligned} \tag{5}$$

- This has several advantages:

> - No relaxation!
> - One less constraint!
> - Avoids (partly) the inscalability issue in high dimensions!

REPMS for MbF-PCA

- To solve this optimization, we use REPMS [Liu and Boumal, 2019], a Riemannian counterpart for the exact penalty method:

Algorithm 1: REPMS for MbF-PCA

Input: X, K, $\epsilon_{min}, \epsilon_0 > 0$, $\theta_\epsilon \in (0, 1)$, $\rho_0 > 0$, $\theta_\rho > 1$, $\rho_{max} \in (0, \infty)$, $\tau > 0$, $d_{min} > 0$.

Initialize V_0;

for $k = 0, 1, \ldots, K$ **do**

  Compute an approximate solution V_{k+1} for the following sub-problem, with a warm-start at V_k, until $\|\text{grad } \mathcal{Q}\| \leq \epsilon_k$:

$$\min_{V \in St(p,d)} \mathcal{Q}(V, \rho_k) \tag{9}$$

  where

$$\mathcal{Q}(V, \rho_k) = f(V) + \rho_k h(V)$$

  if $\|V_{k+1} - V_k\|_F \leq d_{min}$ *and* $\epsilon_k \leq \epsilon_{min}$ **then**

    if $h(V_{k+1}) \leq \tau$ **then**

      return V_{k+1};

    end

  end

  $\epsilon_{k+1} = \max\{\epsilon_{min}, \theta_\epsilon \epsilon_k\}$;

  if $h(V_{k+1}) > \tau$ **then**

    $\rho_{k+1} = \min(\theta_\rho \rho_k, \rho_{max})$;

  else

    $\rho_{k+1} = \rho_k$;

  end

end

Figure: Pseudocode of REPMS

New Theoretical Guarantees

- Under some mild conditions (see the paper for more details), we derive two *new* theoretical guarantees for REPMS.

Theorem

Let $K = \infty$, $\rho_{max} = \infty$, $\epsilon_{min} = \tau = 0$, $\{V_k\}$ *be the sequence generated by REPMS, and* $\overline{V}$ *be any limit point of* $\{V_k\}$, *whose existence is guaranteed. Then the following holds:*

- $\overline{V}$ *always satisfies a necessary condition for* $\overline{V}$ *to be fair.*
- *If* $\overline{V}$ *is fair, then* $\overline{V}$ *is a local maximizer of Eq.* (5)

Theorem (Informal)

Let $K = \infty$, $\rho_{max} < \infty$, $\epsilon_{min}, \tau > 0$. *Then above holds approximately in the following sense: as* $\rho_{max} \to \infty$ *and* $\epsilon_{min}, \tau \to 0$, *we recover the previous exact guarantees.*

Novelty of our theoretical guarantees

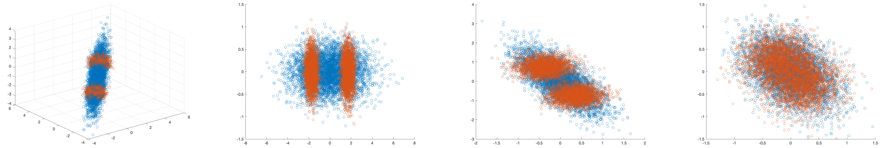
- Existing optimality guarantee of REPMS (Proposition 4.2; [Liu and Boumal, 2019]):
  - $\epsilon_{min} = \tau = 0$, ρ is *not* updated (i.e. line 10-14 is ignored)
  - "If the resulting limit point is fair, then that limit point satisfies the Riemannian KKT condition [Yang et al., 2014]".
- Our theoretical analyses[4]:
  - $\epsilon_{min}, \tau \geq 0$, ρ *is* updated
  - If the resulting limit point is (approximately) fair, then that limit point is (approximately) local maximizer.

[4]We've incorporated a new, yet reasonable assumption; see our paper for more details.

Outline

Synthetic data #1

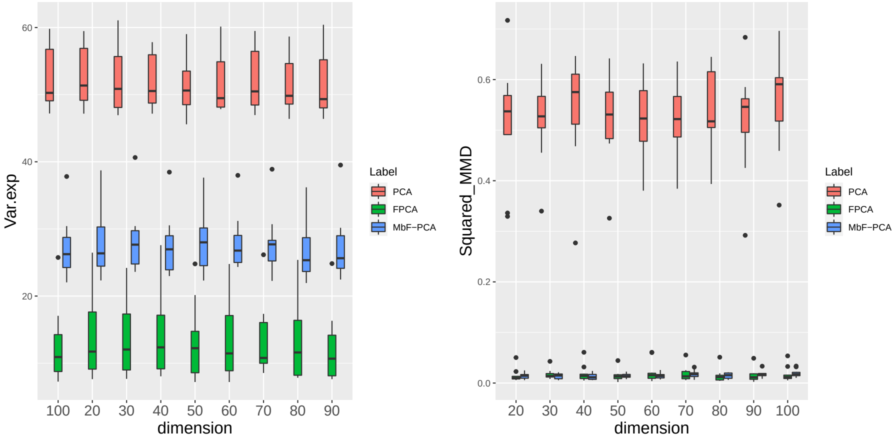
- Due to the Gaussian assumption, FPCA cannot cover the case when two sensitive distributions, that are different, have the same first two moments (mean, covariance):



(a) Original data (b) PCA (c) FPCA [Olfat and Aswani, 2019] (d) MbF-PCA (ours)

Figure: Synthetic data #1: Comparison of PCA, FPCA, and MbF-PCA on data composed of two groups with same mean and covariance, but different distributions. Blue and orange represent different protected groups.

Synthetic data #2



(a) Variance explained (%) (b) MMD^2

Figure: Synthetic data #2: Comparison of PCA, FPCA, and MBF-PCA on the synthetic datasets of increasing dimensions.

Synthetic data #2

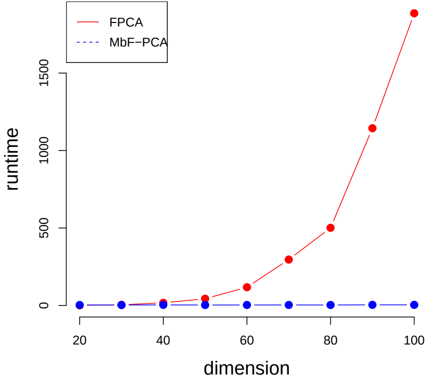


Figure: FPCA represents the SDP algorithm for fair PCA, and MbF-PCA represents our manifold-based framework. Note the drastic difference in scalability!

UCI Datasets

Table 1: Comparison of PCA, FPCA, MBF-PCA for UCI datasets. Number in parenthesis for each dataset is its dimension. Also, the parenthesis for each fair algorithm is its hyperparameter setting; (μ, δ) for FPCA and τ for MBF-PCA. Among the fair algorithms considered, results with the best mean values are **bolded**. Results in which our approach terminates improperly in the sense that the maximum iteration is reached before passing the termination criteria are highlighted.

		COMPAS (11)				German Credit (57)				Adult Income (97)			
d	Alg.	%Var	%Acc	MMD^2	Δ_{DP}	%Var	%Acc	MMD^2	Δ_{DP}	%Var	%Acc	MMD^2	Δ_{DP}
2	PCA	$39.28_{5.17}$	$64.53_{1.45}$	$0.092_{0.010}$	$0.29_{0.09}$	$11.42_{0.47}$	$76.87_{1.39}$	$0.147_{0.049}$	$0.12_{0.06}$	$7.78_{0.82}$	$82.03_{1.15}$	$0.349_{0.027}$	$0.20_{0.05}$
	FPCA $(0.1, 0.01)$	**$35.06_{5.16}$**	$61.65_{1.17}$	$0.012_{0.007}$	$0.10_{0.07}$	$7.43_{0.59}$	$72.17_{1.09}$	$0.017_{0.010}$	$0.03_{0.02}$	$4.05_{0.98}$	$77.44_{2.96}$	$0.016_{0.011}$	$0.04_{0.04}$
	FPCA $(0, 0.01)$	$34.43_{5.02}$	$60.86_{1.09}$	$0.011_{0.006}$	$0.10_{0.06}$	$7.33_{0.57}$	$71.77_{1.60}$	**$0.015_{0.010}$**	$0.03_{0.03}$	$3.65_{0.97}$	$77.05_{3.18}$	**$0.005_{0.004}$**	**$0.01_{0.01}$**
	MBF-PCA (10^{-3})	$33.95_{5.01}$	**$65.37_{1.11}$**	$0.005_{0.002}$	$0.12_{0.07}$	**$10.17_{0.57}$**	**$74.53_{1.92}$**	$0.018_{0.014}$	$0.05_{0.04}$	**$6.03_{0.61}$**	**$79.50_{1.22}$**	**$0.005_{0.004}$**	$0.03_{0.02}$
	MBF-PCA (10^{-6})	$11.83_{3.59}$	$57.73_{1.50}$	**$0.002_{0.002}$**	**$0.06_{0.08}$**	$9.36_{0.33}$	$74.10_{1.56}$	$0.016_{0.010}$	**$0.02_{0.02}$**	$5.83_{0.57}$	$79.12_{1.14}$	**$0.005_{0.004}$**	**$0.01_{0.01}$**
10	PCA	$100.00_{0.00}$	$73.14_{1.22}$	$0.241_{0.005}$	$0.21_{0.07}$	$38.25_{0.98}$	$99.93_{0.14}$	$0.130_{0.019}$	$0.12_{0.08}$	$21.77_{2.06}$	$93.64_{0.92}$	$0.195_{0.007}$	$0.16_{0.01}$
	FPCA $(0.1, 0.01)$	**$87.79_{1.27}$**	$72.25_{0.93}$	$0.015_{0.003}$	**$0.16_{0.06}$**	$29.85_{0.87}$	**$99.93_{0.14}$**	$0.020_{0.005}$	$0.12_{0.08}$	$15.75_{1.20}$	$91.94_{0.88}$	$0.006_{0.003}$	$0.13_{0.02}$
	FPCA $(0, 0.1)$	$87.44_{1.35}$	**$72.32_{0.93}$**	$0.015_{0.002}$	**$0.16_{0.07}$**	$29.79_{0.89}$	**$99.93_{0.14}$**	$0.020_{0.006}$	$0.12_{0.08}$	$15.52_{1.18}$	$91.66_{0.97}$	$0.004_{0.002}$	$0.13_{0.02}$
	MBF-PCA (10^{-3})	$87.75_{1.36}$	$72.16_{0.90}$	**$0.014_{0.002}$**	**$0.16_{0.07}$**	**$34.10_{1.00}$**	**$99.93_{0.14}$**	$0.020_{0.008}$	$0.12_{0.08}$	**$18.71_{1.47}$**	**$92.81_{0.84}$**	$0.005_{0.002}$	$0.14_{0.01}$
	MBF-PCA (10^{-6})	$87.75_{1.36}$	$72.16_{0.90}$	**$0.014_{0.002}$**	**$0.16_{0.07}$**	$16.95_{1.52}$	$92.70_{3.00}$	**$0.013_{0.007}$**	**$0.06_{0.05}$**	$15.49_{6.44}$	$86.36_{3.77}$	**$0.003_{0.002}$**	**$0.07_{0.03}$**

- Across all considered datasets, MBF-PCA is shown to outperform FPCA in terms of fairness (MMD^2 and Δ_{DP}) with low enough τ.
  - Δ_{DP}: measure of demographic parity [Feldman et al., 2015] w.r.t. the downstream task
- For German Credit and Adult Income, controlling τ shows a good trade-off between explained variance and fairness

UCI Datasets

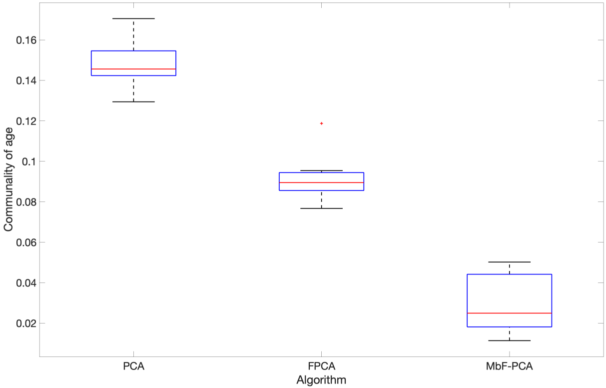


Figure: Comparison of communality of "age" of German credit dataset for PCA, FPCA, and MbF-PCA.

Outline

Conclusion

Our contributions:

- MbF-PCA: a new framework for fair PCA, with several advantages over the previous approach [Olfat and Aswani, 2019]
  - **New definition** for fair PCA based on **MMD**.
  - Utilization of **manifold optimization framework**.
- **Improved guarantees** for REPMS [Liu and Boumal, 2019].
- Empirical verification of our algorithm on synthetic and UCI datasets in explained variance, fairness, and runtime.

Check out our paper for more details!

Paper Code (GitHub)

Absil, P.-A., Baker, C. G., and Gallivan, K. A. (2007a).
Trust-region methods on riemannian manifolds.
Foundations of Computational Mathematics, 7(3):303–330.

Absil, P.-A., Mahony, R., and Sepulchre, R. (2007b).
Optimization Algorithms on Matrix Manifolds.
Princeton University Press, USA.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016).
Machine bias.
ProPublica.

Arora, R., Cotter, A., and Srebro, N. (2013).
Stochastic optimization of pca with capped msg.
In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26, pages 1815–1823. Curran Associates, Inc.

Cisse, M. and Koyejo, S. (2019).
Nips 2019 tutorial: Fairness and representation learning.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015).
Certifying and removing disparate impact.
In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, Sydney, NSW, Australia.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008).
Kernel measures of conditional dependence.
In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2007).
A kernel method for the two-sample-problem.
In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Liu, C. and Boumal, N. (2019).

Simple algorithms for optimization on riemannian manifolds with constraints.
Applied Mathematics and Optimization, 82:949–981.

Olfat, M. and Aswani, A. (2019).
Convex formulations for fair principal component analysis.
In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 663–670.

Yang, W. H., Zhang, L.-H., and Song, R. (2014).
Optimality conditions for the nonlinear programming problems on riemannian manifolds.
Pacific Journal of Optimization, 10:415–434.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013).
Learning fair representations.
In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, pages 325–333, Atlanta, GA, USA.

Quick Intuition behind Manifold Optimization

- Consider $\mathcal{M}$, an embedded Riemannian sub-manifold of $\mathbb{R}^{p \times d}$.
- Suppose we want to minimize some function $f : \mathbb{R}^{p \times d} \to \mathbb{R}$ *over* $\mathcal{M}$.
- If $\mathcal{M}$ is simply viewed as a subset of $\mathbb{R}^{p \times d}$, then this is a constrained optimization problem:

$$\begin{aligned} \underset{V}{\text{minimize}} \quad & f(V) \\ \text{subject to} \quad & V \in \mathcal{M}. \end{aligned} \tag{6}$$

- In this case, the optimization algorithm will make use of the canonical gradients and Hessians of $\mathbb{R}^{p \times d}$.

Quick Intuition behind Manifold Optimization

- If $\mathcal{M}$ is "all there is", then this problem is an unconstrained optimization problem over $\mathcal{M}$.
  - Consider an ant living on $\mathcal{M}$. From the universe ($\mathbb{R}^{p \times d}$), the ant is constrained on $\mathcal{M}$. But from the ant's perspective, $\mathcal{M}$ is all they have i.e. he/she would feel *unconstrained*!
- In this case, the optimization algorithm will make use of the *Riemannian* gradients and Hessians of $\mathcal{M}$.
- By making use of the intrinsic geometry of $\mathcal{M}$, the optimization becomes much more efficient!

Quick Intuition behind Manifold Optimization

- A very straightforward way to think of this is by considering the simplest Riemannian manifold[5], $\mathbb{R}^{p\times d}$.
- When we write the optimization as

$$\begin{aligned} \underset{V}{\text{minimize}} \quad & f(V) \\ \text{subject to} \quad & V \in \mathbb{R}^{p\times d}, \end{aligned} \tag{7}$$

technically this is a "constrained" optimization because we're "constraining" V to be in $\mathbb{R}^{p\times d}$.

- However, gradients and Hessian (and other geometric concepts) are derived directly from the intrinsic geometry of $\mathbb{R}^{p\times d}$ i.e. $V \in \mathbb{R}^{p\times d}$ **isn't considered as a constraint.**

[5]inner product is the Frobenius product: $\langle X, Y\rangle := \mathrm{tr}(X^\intercal Y)$

Extra Comments for Our New Theoretical Guarantees

- Our problem is non-convex in V, which naturally brings up the question of convergence and optimality guarantees.
- First, from various Riemannian optim literatures, we motivate the following assumption, which is to the best of our knowledge, new:

Assumption (informal; locality assumption)

Each V_{k+1} is sufficiently close to a local minimum of Eq. (9).

  - It is known that, pathological examples excluded, most conventional *unconstrained* manifold optimization solvers produce iterates whose limit points are local minima, and not other stationary points such as saddle point or local maxima: see [Absil et al., 2007a, Absil et al., 2007b] for more detailed discussions.
  - Many theoretical results have also emerged (ex. "First-order methods almost always avoid strict saddle points" Lee et al., Math. Prog. 2019)