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

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

[1] Kim Jaechul Graduate School of AI, KAIST, Seoul, Republic of Korea
[2] School of Electrical Engineering, KAIST, Daejeon, Republic of Korea
[3] UC Berkeley IEOR, Berkeley, CA, USA
[4] Citadel, Chicago, IL, USA
[5] Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, IL, USA
{jh_lee00, s88012, cd_yoo}@kaist.ac.kr, molfat@berkeley.edu, jhasegaw@illinois.edu

Abstract

This paper defines fair principal component analysis (PCA) as minimizing the maximum mean discrepancy (MMD) between dimensionality-reduced conditional distributions of different protected classes. The incorporation of MMD naturally leads to an exact and tractable mathematical formulation of fairness with good statistical properties. We formulate the problem of fair PCA subject to MMD constraints as a non-convex optimization over the Stiefel manifold and solve it using the Riemannian Exact Penalty Method with Smoothing (REPMS). Importantly, we provide local optimality guarantees and explicitly show the theoretical effect of each hyperparameter in practical settings, extending previous results. Experimental comparisons based on synthetic and UCI datasets show that our approach outperforms prior work in explained variance, fairness, and runtime.

Introduction

It has become increasingly evident that many widely-deployed machine learning algorithms are biased, yielding outcomes that can be discriminatory across key groupings such as race, gender and ethnicity (Mehrabi et al. 2019). As the applications of these algorithms proliferate in protected areas like healthcare (Karan et al. 2012), hiring (Chien and Chen 2008) and criminal justice (Kirchner et al. 2016), this creates the potential for further exacerbating social biases. To address this, there has recently been a surge of interest in ensuring fairness in resulting machine learning algorithms.

Working in high-dimensional spaces can be undesirable as the curse of dimensionality manifests in the form of data sparsity and computational intractability. Various dimensionality reduction algorithms are deployed to resolve these issues, and Principal Component Analysis (PCA), is one of the most popular methods (Jolliffe and Cadima 2016). One particular advantage of PCA is that there's no need to train a complex neural network.

In this work, fair PCA is defined as doing PCA while minimizing the difference in the conditional distributions of projections of different protected groups. Here, the projected data can be considered as a dimension-reduced fair representation of the input data (Zemel et al. 2013). We answer the questions of 1) how fairness should be defined for PCA and 2) how to (algorithmically) incorporate fairness into PCA in a *fast* and *efficient* manner. This work takes a different approach from prior studies on PCA fairness (Samadi et al. 2018; Olfat and Aswani 2019), which is discussed in Section and 15.

Our main contributions are as follows:

- We motivate a new mathematical definition of fairness for PCA using the maximum-mean discrepancy (MMD), which can be evaluated in a computationally efficient manner from the samples while guaranteeing asymptotic consistency. Such properties were not available in the previous definition of fair PCA (Olfat and Aswani 2019). This is discussed in detail in Section 3 and 4.
- We formulate the task of performing MMD-based fair PCA as a constrained optimization over the Stiefel manifold and propose using REPMS (Liu and Boumal 2019). For the first time, we prove two general theoretical guarantees of REPMS regarding the local minimality and feasibility. This is discussed in detail in Section 5 and 6.
- Using synthetic and UCI datasets, we verify the efficacy of our approach in terms of explained variance, fairness, and runtime. Furthermore, we verify that using fair PCA does indeed result in a fair representation, as in (Zemel et al. 2013). This is discussed in detail in Section 7.

Preliminaries

Notations

For $b \geq 1$, let $\mathcal{P}_b$ be the set of all Borel probability measures defined on $\mathbb{R}^b$. For some measurable function $\Pi : \mathbb{R}^p \to \mathbb{R}^d$ and a measure $P \in \mathcal{P}_p$, the push-forward measure of P via Π is the probability measure $\Pi_{\#}P \in \mathcal{P}_d$, defined as $(\Pi_{\#}P)(S) = P(\Pi^{-1}(S))$ for any Borel set S. Let $\mathbf{0}$ and $\mathbf{1}$ denote matrices (or vectors) of zeros and ones of appropriate size, respectively. In this work, we focus on binary cases, i.e., we assume that the protected attribute A and outcome Y are binary ($A, Y \in \{0, 1\}$). We abbreviate demographic parity, equalized opportunity, and equalized odds as DP, EOP, and EOD, respectively.

*Corresponding authors

See (Lee et al. 2021) for a more complete version of this manuscript, including the supplementary material.

Maximum Mean Discrepancy (MMD)

Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive-definite kernel function, and $\mathcal{H}_k$ be a unit-ball in the RKHS generated by k. We impose some regularity assumptions on k:

Assumption 1. *k is measurable, and bounded i.e. $K := \sup_{x,y} k(x, y) < \infty$.*

Then one can pseudo-metrize $\mathcal{P}_d$ by the following distance:

Definition 1 (Gretton et al., 2007)**.** *Given $\mu, \nu \in \mathcal{P}_d$, their* **maximum mean discrepancy (MMD)***, denoted as $\mathrm{MMD}_k(\mu, \nu)$, is a pseudo-metric on $\mathcal{P}_d$, defined as follows:*

$$\underset{k}{\mathrm{MMD}}(\mu, \nu) := \sup_{f \in \mathcal{H}_k} \left| \int_{\mathbb{R}^d} f \, d(\mu - \nu) \right| \tag{1}$$

As our fairness constraint involves exactly matching the considered distributions using MMD, we require the property of $\mathrm{MMD}_k(\mu, \nu) = 0$ implying $\mu = \nu$. Any kernel k that satisfies such property is said to be **characteristic** (Fukumizu et al. 2008) to $\mathcal{P}_d$. Furthermore, Sriperumbudur et al. (2008) defined and characterized *stationary* characteristic kernels and identified that well-known kernels such as RBF and Laplace are characteristic. Based on this fact, we set k to be the RBF kernel $k_{rbf}(x, y) := \exp\left(-\|x - y\|^2/2\sigma^2\right)$.

For the choice of bandwidth σ, the median of the set of pairwise distances of the samples after *vanilla PCA* is considered following the *median heuristic* of (Schölkopf, Smola, and Müller 1998). For simplicity, we refer to $\mathrm{MMD}_{k_{rbf}}$ as MMD.

Benefits of MMD There are several reasons for using MMD as the distance on a space of probability measures. First, it can act as a distance between distributions with different, or even disjoint, supports. This is especially crucial as the empirical distributions are often discrete and completely disjoint. Such a property is not generally true, one prominent example being the KL-divergence. Second, since many problems in fairness involve comparing two distributions, MMD has already been used in much of the fairness literature as a metric (Madras et al. 2018; Adel et al. 2019) and as an explicit constraint/penalty (Prost et al. 2019; Oneto et al. 2020; Jung et al. 2021), among other usages.

Fairness for Supervised Learning

The fair PCA discussed above should ultimately lead to fairness in supervised learning tasks based on the dimension-reduced data with minimal loss in performance. Let us now review three of the most widely-used definitions of fairness in supervised learning, as formulated in (Madras et al. 2018). Let $(Z, Y, A) \in \mathbb{R}^d \times \{0, 1\} \times \{0, 1\}$ be the joint distribution of the dimensionality-reduced data, (downstream task) label, and protected attribute. Furthermore, let $g : \mathbb{R}^d \to \{0, 1\}$ be a classifier that outputs prediction $\hat{Y}$ for Y from Z. We want to determine the fairness of a well-performing classifier g w.r.t. protected attribute A.

First, let D_s be the probability measure of $Z_s \triangleq Z|A = s$ for $s \in \{0, 1\}$:

Definition 2 (Feldman et al., 2015)**.** *g is said to satisfy* **demographic parity (DP) up to** Δ_{DP} *w.r.t. A with $\Delta_{DP} \triangleq |\mathbb{E}_{x \sim D_0}[g(x)] - \mathbb{E}_{x \sim D_1}[g(x)]|$.*

Now, let $D_{s,y}$ be the probability measure of $Z_s \triangleq Z|A = s, Y = y$ for $s, y \in \{0, 1\}$.

Definition 3 (Hardt, Price, and Srebro, 2016)**.** *g is said to satisfy* **equalized opportunity (EOP) up to** Δ_{EOP} *w.r.t. A and Y with $\Delta_{EOP} \triangleq \left|\mathbb{E}_{x \sim D_{0,1}}[g(x)] - \mathbb{E}_{x \sim D_{1,1}}[g(x)]\right|$.*

Definition 4 (Hardt, Price, and Srebro, 2016)**.** *g is said to satisfy* **equalized odds (EOD) up to** Δ_{EOD} *w.r.t. A and Y with $\Delta_{EOD} \triangleq \max_{y \in \{0,1\}} \left|\mathbb{E}_{x \sim D_{0,y}}[g(x)] - \mathbb{E}_{x \sim D_{1,y}}[g(x)]\right|$.*

From hereon, we refer to such $\Delta_f(g)$ as the **fairness metric of** $f \in \{DP, EOP, EOD\}$ **w.r.t.** g, respectively.

New Definition of Fairness for PCA

For $p > d$, let $\mathbb{R}^d$ be the space onto which data will be projected. A dimensionality reduction is a map $\Pi : \mathbb{R}^p \to \mathbb{R}^d$, and PCA is defined as $\Pi(x) = V^\intercal x$ for some $V \in \mathbb{R}^{p \times d}$ satisfying[1] $V^\intercal V = \mathbb{I}_d$ i.e. PCA is a linear, orthogonal dimensionality-reduction. From hereon, we denote a linear PCA as the mapping Π. The definition for fairness that we will be following throughout is given as follows.

Definition 5 (Δ-fairness)**.** *Let P_s be the probability measure of $X_s \triangleq X|A = s$ for $s \in \{0, 1\}$, and let $Q_s := \Pi_\# P_s \in \mathcal{P}_d$. Then Π is said to be* **Δ-fair** *with $\Delta := \mathrm{MMD}(Q_0, Q_1)$, and we refer to Δ as the* **fairness metric***.*

In other words, for lower Δ-fairness, the discrepancy between the dimensionality-reduced conditional distributions of different protected classes, measured in a non-parametric manner using MMD, while retaining as much variance as possible, should be minimized.

Furthermore, Definition 5 ensures that a downstream classification task using Δ-fair dimensionality-reduced data will be fair, as formalized below[2]:

Proposition 1 (Oneto et al., 2020)**.** *Up to a constant factor, $\mathrm{MMD}(Q_0, Q_1)$ bounds the MMD of the push-forward measures of Q_0, Q_1 via the weight vector of any given downstream task classifier g.*

Remark 1. *The above discussions easily generalize to equal opportunity and equalized odds.*

Relation with Other Definitions of Fair PCA

The notion of fairness proposed by Olfat and Aswani (2019) is similar to ours in that it measures the predictability of protected group membership in dimensonality-reduced data. However, unlike ours, their definition is explicitly adversarial, which can be a problem.

Definition 6 (Δ_A-fairness; Olfat and Aswani, 2019)**.** *Consider a fixed classifier $h(u, t) : \mathbb{R}^d \times \mathbb{R} \to \{0, 1\}$ that inputs*

[1] The benefits of pursuing orthogonality in the loading matrix, and thus the resulting PCs, are already well-studied; for example, see (Qi, Luo, and Zhao 2013; Benidis et al. 2016).

[2] See Lemma 3 of (Oneto et al. 2020) for the precise statement.

features $u \in \mathbb{R}^d$ and a threshold t, and predicts the protected class $s \in \{0, 1\}$. Then, Π is $\Delta_A(h)$-fair if

$$\begin{aligned}\sup_{t\in\mathbb{R}} \Big| &\mathbb{P}\big[h(\Pi(x), t) = 1 | s = 1\big] \\ &- \mathbb{P}\big[h(\Pi(x), t) = 1 | s = 0\big]\Big| \leq \Delta_A(h)\end{aligned} \tag{2}$$

Moreover, for a family of classifiers $\mathcal{F}_c$, if Π is $\Delta_A(h)$-fair for $\forall h \in \mathcal{F}_c$, we say that Π is $\Delta_A(\mathcal{F}_c)$-fair.

As Δ_A can't be computed exactly, an estimator of the following form was used:

$$\widehat{\Delta}_A(\mathcal{F}_c) := \sup_{h\in\mathcal{F}_c} \sup_t \left| \frac{1}{|P|} \sum_{i\in P} I_i(\Pi, h_t) - \frac{1}{|N|} \sum_{i\in N} I_i(\Pi, h_t) \right| \tag{3}$$

where $\{x_i\}_{i=1}^n$ are the data points, (P, N) is a partition of the index set $\{1, 2, \ldots, n\}$ into two protected groups, $I_i(\Pi, h_t) = \mathbf{1}(h(\Pi(x_i), t) = +1)$, and $\mathbf{1}(\cdot)$ is the indicator function.

Remark 2. *It can be argued that, for some choice of $\mathcal{F}_c$, Definition 5 and 6 are equivalent: in effect, that these are dual notions. Recognizing this, we proceed with Definition 5, as it has two main advantages in the context of our work:*

- *It ties more directly and intuitively into our optimization formulation; see Section .*
- *It can be represented non-variationally which allows for tighter statistical guarantees.*

Statistical Properties of Δ

Consistent and Efficient Estimation of Δ

As defined in Definition 5, let $Q_0, Q_1 \in \mathcal{P}_d$ be the probability measures with respect to the samples of which we want to estimate $\mathrm{MMD}(\cdot, \cdot)$. Let $\{X_i\}_{i=1}^m$ and $\{Y_j\}_{j=1}^n$ be these samples, respectively. Accordingly, we consider the following estimator:

$$\widehat{\Delta} := \mathrm{MMD}(\hat{Q}_0, \hat{Q}_1) \tag{4}$$

where $\hat{Q}_s$ is the usual empirical distribution, defined as the mixture of Dirac measures on the samples.

Unlike other statistical distances (e.g. total variation distance), $\widehat{\Delta}_k$ has several theoretical properties that have important practical implications; see Sriperumbudur et al. (2010) for more details.

First, it can be computed exactly and efficiently:

Lemma 1 (Gretton et al., 2007)**.** *$\widehat{\Delta}$ is computed as follows:*

$$\begin{aligned}\widehat{\Delta} = \Bigg[&\frac{1}{m^2} \sum_{i,j=1}^m k(X_i, X_j) + \frac{1}{n^2} \sum_{i,j=1}^n k(Y_i, Y_j) \\ &- \frac{2}{mn} \sum_{i,j=1}^{m,n} k(X_i, Y_j) \Bigg]^{1/2}.\end{aligned} \tag{5}$$

Moreover, it is asymptotically consistent with a convergence rate, depending only on m and n:

Proposition 2 (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:*

$$\left|\Delta - \widehat{\Delta}\right| \leq 2\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) + \delta \tag{6}$$

Advantages over Δ_A

$\widehat{\Delta}_A$ is known to satisfy the following high probability bound:

Proposition 3 (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:*

$$\left|\Delta_A(\mathcal{F}_c) - \widehat{\Delta}_A(\mathcal{F}_c)\right| \leq 8\sqrt{\frac{VC(\mathcal{F}_c)}{m+n}} + \delta \tag{7}$$

where $VC(\cdot)$ is the VC dimension.

Remark 3. *If $\mathcal{F}_c$ is too expressive in terms of VC-dimension, then the above bound may become void. This is the case, for instance, when $\mathcal{F}_c$ is the set of RBF-kernel SVMs.*

In addition, computing $\widehat{\Delta}_A$ requires considering all possible classifiers in the designated family $\mathcal{F}_c$. This is computationally infeasible, and it forces one to use another approximation (e.g. discretization of $\mathcal{F}_c$), which incurs additional error that may further inhibit asymptotic consistency.

As exhibited in the previous subsection, our MMD-based approach suffers from none of these issues.

Manifold Optimization for MBF-PCA

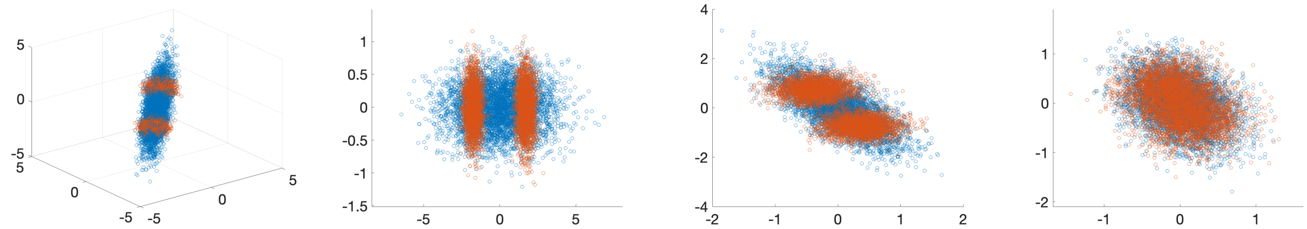
Improvements over FPCA

Olfat and Aswani (2019) proposed FPCA, an SDP formulation of fair PCA[3], in which matching the first and second moments of the protected groups after dimensionality-reduction are approximated as convex constraints. However, this has several shortcomings, which we discuss here and empirically exhibit in a later section.

First, there are cases in which matching the mean and covariance alone is not enough. The simplest "counterexample" would be when two protected groups have the same mean and covariance, yet they have different distributions. This is illustrated in Figure 1. While this previous point can be countered by the application of the kernel trick to FPCA, this raises a second issue: their formulation requires solving [4] a $p \times p$-dimensional SDP, motivated by the reparameterization $P = VV^\intercal$ (Arora, Cotter, and Srebro 2013). Since SDP is known to become inefficient (or even computationally infeasible) in high dimensions, this quickly becomes intractable for high-dimensional data (for linear or polynomial kernels) or for any moderate to large size datasets (for the RBF kernel). Finally, their approach involves a relaxation of a rank constraint ($\mathrm{rank}(P) \leq d$) to a trace constraint ($\mathrm{tr}(P) \leq d$), yielding sub-optimal outputs in presence of (fairness) constraints, even to substantial order in some cases. In Section C

[3] See Section B of the supplementary material for its precise description.

[4] In their approach, the final solution V is obtained by taking the first d eigenvectors of P.



(a) Original data

(b) PCA

(c) FPCA (Olfat and Aswani 2019)

(d) MBF-PCA (ours)

Figure 1: 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.

of the SP, we discuss in detail why FPCA may lead to such degraded explained variance.

Formulating MBF-PCA

Observing that the shortcomings of FPCA stem from the reparametrization of $P = VV^\intercal$, we propose a new formulation of fair PCA that solves *directly* for V. This allows for an effective and efficient approach.

We start by noting that the set of all V with orthonormal columns has the intrinsic geometric structure of a *manifold*:

Definition 7. *For $p > d$, the Stiefel manifold, denoted as $St(p, d)$, is an embedded Riemannian sub-manifold of $\mathbb{R}^{p\times d}$ such that each element of $St(p, d)$ has orthonormal columns i.e. $V^\intercal V = I_d$ for all $V \in St(p, d)$.*

$St(p, d)$ has several desirable properties such as compactness, completeness and smoothness, which we present in Section D of the SP. As $St(p, d)$ is prevalent in various fields of machine learning (most notably PCA), much work has been done that focuses on exploiting this geometric structure for efficient optimization (Hu et al. 2020).

Based on our MMD-based formulation and letting Σ be the sample covariance matrix of the full dataset, we formulate our fair PCA as follows, which we refer to as MBF-PCA:

$$\begin{aligned} &\underset{V\in St(p,d)}{\text{minimize}} && f(V) := -\langle \Sigma, VV^\intercal \rangle \\ &\text{subject to} && h(V) := \overset{2}{\text{MMD}}(Q_0, Q_1) = 0. \end{aligned} \tag{8}$$

Here, Q_0 and Q_1 are defined as in definition 5. Observe how our definition of fairness *directly* incorporates itself into the optimization problem as a constraint.

Remark 4. *Under an assumption of normality, our* MMD*-based formulation amounts to the same constraints as* FPCA *since* $\text{MMD}^2(\cdot, \cdot)$ *is a metric and a Gaussian distribution is completely characterized by its first and second moments.*

REPMS for MBF-PCA, with New Theoretical Guarantees

Description of Algorithm 1

One crucial observation is that the constraint function h is always non-negative[5] and smooth. This motivates the use of the exact penalty method (Han and Mangasarian 1979), recently extended to manifold optimization as the Riemannian Exact Penalty Method via Smoothing (REPMS; Liu and Boumal, 2019). Note that smoothing tricks (Liu and Boumal 2019), which were required to smooth out possible non-differentiable functions emerging from the ℓ_1-penalty, are *not* necessary. Moreover, by leveraging the kernel trick, there is a *closed* form for $\nabla_V h(V)$, thus alleviating the need for computationally expensive numerical approximations; see Section G of the SP for the derivation. The pseudo-code for the algorithm is shown in Algorithm 1.

For practical concerns that will be addressed in the following subsection, we've set the *fairness tolerance level*, τ, to be a fixed and sufficiently small, non-negative value. Formally, we consider the following definition:

Definition 8. *For fixed $\tau \geq 0$, $V \in St(p, d)$ is τ-***approximate fair** *if it satisfies $h(V) \leq \tau$. If $\tau = 0$, we simply say that V is* **fair**.

New Theoretical Guarantees for Algorithm 1

We start by observing that Eq. (9) in Algorithm 1 is smooth, unconstrained manifold optimization problem, which can be solved using conventional algorithms; these include first-order methods like line-search methods (Absil, Mahony, and Sepulchre 2007), or second-order methods like the Riemannian Trust Region (RTR; Absil, Baker, and Gallivan, 2007) method. 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, Baker, and Gallivan 2007; Absil, Mahony, and Sepulchre 2007) for more detailed discussions.

Motivated by this, we consider the following assumption:

[5] This is due to our choice of estimator for MMD inducing a *non-negative* estimate of MMD^2; see Section 2 of Gretton et al. (2012) for more detailed discussions.

Algorithm 1: REPMS for MbF-PCA

```
Input: $X, K, \epsilon_{min}, \epsilon_0 > 0, \theta_\epsilon \in (0,1), \rho_0 > 0,$
       $\theta_\rho > 1, \rho_{max} \in (0, \infty), \tau > 0, d_{min} > 0.$
1  Initialize $V_0$;
2  for $k = 0, 1, \dots, K$ do
3      Compute an approximate solution $V_{k+1}$ for the
       following sub-problem, with a warm-start at $V_k$,
       until $\|\text{grad } \mathcal{Q}\| \leq \epsilon_k$:
           $$\min_{V \in St(p,d)} \mathcal{Q}(V, \rho_k) \quad (9)$$
       where
           $$\mathcal{Q}(V, \rho_k) = f(V) + \rho_k h(V)$$
4      if $\|V_{k+1} - V_k\|_F \leq d_{min}$ and $\epsilon_k \leq \epsilon_{min}$ then
5          if $h(V_{k+1}) \leq \tau$ then
6              return $V_{k+1}$;
7          end
8      end
9      $\epsilon_{k+1} = \max\{\epsilon_{min}, \theta_\epsilon \epsilon_k\}$;
10     if $h(V_{k+1}) > \tau$ then
11         $\rho_{k+1} = \min(\theta_\rho \rho_k, \rho_{max})$;
12     else
13         $\rho_{k+1} = \rho_k$;
14     end
15 end
```

Assumption 2 (informal; locality assumption). *Each V_{k+1} is sufficiently close to a local minimum of Eq.* (9).

Lastly, we consider the following auxiliary optimization problem:

$$\min_{V \in St(p,d)} h(V) \tag{10}$$

The following theorem, whose proof is deferred to Section F of the SP, provides an *exact* theoretical convergence guarantee of MBF-PCA under the *ideal* hyperparameter setting:

Theorem 1. *Let $K = \infty$, $\rho_{max} = \infty$, $\epsilon_{min} = \tau = 0$, $\{V_k\}$ be the sequence generated by Alg. 1 under Assumption 2, and $\overline{V}$ be any limit point of $\{V_k\}$, whose existence is guaranteed. Then the following holds:*

(A) $\overline{V}$ is a local minimizer of Eq. (10)*, which is a necessary condition for $\overline{V}$ to be fair.*

(B) If $\overline{V}$ is fair, then $\overline{V}$ is a local minimizer of Eq. (8)

The assumption of $\overline{V}$ being fair, which is used in (B), is at least partially justified in (A) in the following sense: the ideal hyperparameter setting of $\rho_{max} = \infty, \tau = 0, \epsilon_{min} = 0$ implies the *exact* local minimality of $\overline{V}$ for Eq. (10), which is in turn a *necessary condition* for $\overline{V}$ to be fair.

The next theorem, whose proof is also deferred to Section F of the SP, asserts that with small τ, ϵ_{min} and large ρ_{max}, the above guarantee can be approximated in rigorous sense:

Theorem 2. *Let $K = \infty$, $\rho_{max} < \infty$, $\epsilon_{min}, \tau > 0$, $\{V_k\}$ be the sequences generated by Alg. 1 under Assumption 2 and $\overline{V}$ be any limit point of $\{V_k\}$, whose existence is guaranteed. Then for any sufficiently small ϵ_{min} and $\tilde{r} = \tilde{r}(\epsilon_{min}) > 0$, the following hold:*

(A) $\overline{V}$ is an approximate local minimizer of Eq. (10) *in the sense that*

$$h(\overline{V}) \leq h(V) + \beta\|V - \overline{V}\| + (\beta + L_h) g(\epsilon_{min}) \tag{11}$$

for all $V \in B_{\tilde{r}}(\overline{V}) \cap St(p,d)$, where $\beta = \beta(\rho_{max}, \tau)$ is a function that satisfies the following:

- *$0 < \beta \leq \frac{2\|\Sigma\|}{\rho_0}$*
- *$\beta(\rho_{max}, \tau)$ is increasing in ρ_{max} and decreasing in τ.*

(B) If $\overline{V}$ is fair, then it is an approximate local minimizer of Eq. (8) *in the sense that it satisfies*

$$f(\overline{V}) \leq f(V) + 2\|\Sigma\| g(\epsilon_{min}) \tag{12}$$

for all fair $V \in B_{\tilde{r}}(\overline{V}) \cap St(p,d)$.

In both (A) and (B), g is some continuous, decreasing function that satisfies $g(0) = 0$, and $\tilde{r}(\epsilon_{min}) = r - g(\epsilon_{min})$ for some fixed constant $r > 0$.

Existing optimality guarantee of REPMS (Proposition 4.2; Liu and Boumal, 2019) states that when $\epsilon_{min} = 0$, ρ is *not* updated (i.e. line 10-14 is ignored), and the resulting limit point is feasible, then that limit point satisfies the KKT condition (Yang, Zhang, and Song 2014). Comparing Theorem 1 and 2 to the previous result, we see that ours extend the previous result in several ways:

- Our theoretical analyses are much closer to the actual implementation, by incorporating the ρ-update step (line 11) and the *practical* hyperparameter setting.
- Our theoretical analyses are much more stronger in the sense that 1) by *introducing* a reasonable, yet novel locality assumption, we go beyond the existing KKT conditions and prove the *local minimality* of the limit point, and 2) we provide a partial justification of the feasibility assumption in (A) by proving a necessary condition for it.

Practical Implementation

In line 4 in Algorithm 1, we implemented the termination criteria: sufficiently small distance between iterates and sufficiently small tolerance for solving Eq. (9). However, such a heuristic may return some point $\overline{V}$ that is not τ-approximate fair for user-defined level τ in practical hyperparameter setting. To overcome this issue, we've additionally implemented line 5 that forces the algorithm to continue on with the loop until the desired level of fairness is achieved.

Related Work

Fairness in ML

A large body of work regarding fairness in the context of supervised learning (Feldman et al. 2015; Calders, Kamiran, and Pechenizkiy 2009; Dwork et al. 2012; Hardt, Price, and Srebro 2016; Zafar et al. 2017) has been published. This includes key innovations in quantifying algorithmic bias, notably the concepts of *demographic parity* and *equalized odds (opportunity)* that have become ubiquitous in fairness research (Barocas and Selbst 2016; Hardt, Price, and Srebro 2016). More recently, fair machine learning literatures have branched out into a variety of fields, including deep learning

(Beutel et al. 2017), regression (Calders et al. 2013), and even hypothesis testing (Olfat et al. 2020).

Among these, one line of research has focused on learning fair representations (Kamiran and Calders 2011; Zemel et al. 2013; Feldman et al. 2015; Calmon et al. 2017), which aims to learn a representation of the given data on which various fairness definitions are ensured for downstream modeling. A growing number of inquiries have been made into highly specialized algorithms for specific unsupervised learning problems like clustering (Chierichetti et al. 2017; Bera et al. 2019), but these lack the general applicability of key dimensionality reduction algorithms such as PCA.

To the best of our knowledge, Olfat & Aswani (2019) is the **only** work on incorporating fair representation to PCA, making it the sole comparable approach to ours. Another line of work (Samadi et al. 2018; Tantipongpipat et al. 2019) considers a completely *orthogonal* definition of fairness for PCA: minimizing the discrepancy between reconstruction errors over protected attributes. This doesn't ensure the fairness of downstream tasks, rendering it incomparable to our definition of fairness; see Section A of the SP for more details.

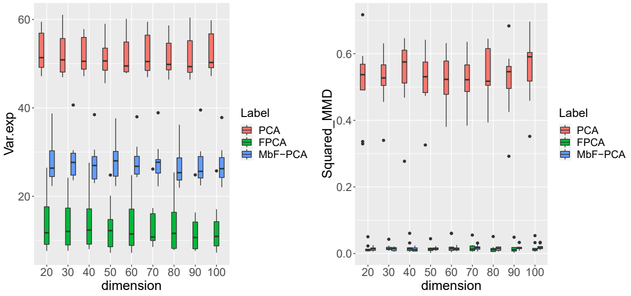
Manifold Optimization

A constrained problem over Euclidean space can be transformed to an unconstrained problem over a manifold (or at least manifold optimization with less constraints). Many algorithms for solving Euclidean optimization problems have direct counterparts in manifold optimization problems that includes Riemannian gradient descent and Riemannian BFGS. By making use of the geometry of lower dimensional manifold structure, often embedded in potentially very high dimensional ambient space, such Riemannian counterparts are much more computationally *efficient* than algorithms that do not make use of manifold structure. This is shown in numerous literatures (Liu and Boumal 2019; Alsharif et al. 2021), including this work. We refer interested readers to the standard textbooks Absil, Mahony, and Sepulchre (2007) and Boumal (2022) on this field, along with a survey by Hu et al. (2020).

Experiments

MBF-PCA was implemented using ROPTLIB (Huang et al. 2018), a state-of-the-art manifold optimization framework on MATLAB. For solving Eq. (9), we use the cautious Riemannian BFGS method (RBFGS; Huang, Absil, and Gallivan, 2018), a quasi-Newton method that is much more memory-efficient. We've set $K = 100, \epsilon_{min} = 10^{-6}, \epsilon_0 = 10^{-1}, \theta_\epsilon = (\epsilon_{min}/\epsilon_0)^{1/5}, \rho_{max} = 10^{10}, \theta_\rho = 2, d_{min} = 10^{-6}$. For FPCA, we use the same Python MOSEK(ApS 2021) implementation as provided by (Olfat and Aswani 2019). (μ, δ) are the hyperparameters of FPCA; see Section B of the SP. Codes are available in our Github repository[6].

All data is pre-processed to be standardized such that each covariate has zero mean and unit variance. For all experiments, we considered 10 different $70-30$ train-test splits.



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

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

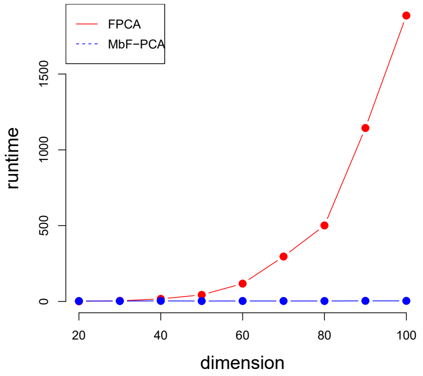


Figure 3: Synthetic data #2: Comparison of runtimes of FPCA, and MbF-PCA.

Synthetic Data #1

We consider synthetic data composed of two groups, each of size $n = 150$; one is sampled from $\mathcal{N}_3(\mathbf{0}, 0.1 I_3 + \mathbf{1})$ and one is sampled from a (balanced) mixture of $\mathcal{N}_3(\mathbf{1}, 0.1 I_3)$ and $\mathcal{N}_3(-\mathbf{1}, 0.1 I_3)$. Note how the two groups follow different distributions, yet have the same mean and covariance. Thus, we expect FPCA to project in a similar way as vanilla PCA, while MBF-PCA should find a fairer subspace such that the projected distributions are exactly the same. Hyperparameters are set as follows: $\delta = 0, \mu = 0.01$ for FPCA and $\tau = 10^{-5}$ for MBF-PCA. We've set $d = 2$ and Figure 1 displays the results of each algorithm using the top two principal components. Indeed, only MBF-PCA successfully obfuscates the protected group information by merging the two orange clusters with the blue cluster.

Synthetic Data #2

We consider a series of synthetic datasets of dimension p. For each p, the dataset is composed of two groups, each of size $n = 240$ and sampled from two different p-variate normal distributions. We vary $p \in \{20, 30, \ldots, 100\}$; see

[6] https://github.com/nick-jhlee/fair-manifold-pca

		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 $(0.1, 0.01)$	$\mathbf{35.06_{5.16}}$	$61.65_{1.17}$	$0.012_{0.007}$	$0.10_{0.07}$	$7.43_{0.59}$	$72.17_{1.09}$	$0.017_{0.010}$	$0.03_{0.02}$	$4.05_{0.98}$	$77.44_{2.96}$	$0.016_{0.011}$	$0.04_{0.04}$
	FPCA $(0, 0.01)$	$34.43_{5.02}$	$60.86_{1.09}$	$0.011_{0.006}$	$0.10_{0.06}$	$7.33_{0.57}$	$71.77_{1.60}$	$\mathbf{0.015_{0.010}}$	$0.03_{0.03}$	$3.65_{0.97}$	$77.05_{3.18}$	$\mathbf{0.005_{0.004}}$	$\mathbf{0.01_{0.01}}$
	MBF-PCA (10^{-3})	$33.95_{5.01}$	$\mathbf{65.37_{1.11}}$	$0.005_{0.002}$	$0.12_{0.07}$	$\mathbf{10.17_{0.57}}$	$\mathbf{74.53_{1.92}}$	$0.018_{0.014}$	$0.05_{0.04}$	$\mathbf{6.03_{0.61}}$	$\mathbf{79.50_{1.22}}$	$\mathbf{0.005_{0.004}}$	$0.03_{0.02}$
	MBF-PCA (10^{-6})	$11.83_{3.59}$	$57.73_{1.50}$	$\mathbf{0.002_{0.002}}$	$\mathbf{0.06_{0.08}}$	$9.36_{0.33}$	$74.10_{1.56}$	$0.016_{0.010}$	$\mathbf{0.02_{0.02}}$	$5.83_{0.57}$	$79.12_{1.14}$	$\mathbf{0.005_{0.004}}$	$\mathbf{0.01_{0.01}}$
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 $(0.1, 0.01)$	$\mathbf{87.79_{1.27}}$	$72.25_{0.93}$	$0.015_{0.003}$	$\mathbf{0.16_{0.06}}$	$29.85_{0.87}$	$\mathbf{99.93_{0.14}}$	$0.020_{0.005}$	$0.12_{0.08}$	$15.75_{1.20}$	$91.94_{0.88}$	$0.006_{0.003}$	$0.13_{0.02}$
	FPCA $(0, 0.1)$	$87.44_{1.35}$	$\mathbf{72.32_{0.93}}$	$0.015_{0.002}$	$\mathbf{0.16_{0.07}}$	$29.79_{0.89}$	$\mathbf{99.93_{0.14}}$	$0.020_{0.006}$	$0.12_{0.08}$	$15.52_{1.18}$	$91.66_{0.97}$	$0.004_{0.002}$	$0.13_{0.02}$
	MBF-PCA (10^{-3})	$87.75^{*}_{1.36}$	$72.16^{*}_{0.90}$	$\mathbf{0.014^{*}_{0.002}}$	$\mathbf{0.16^{*}_{0.07}}$	$\mathbf{34.10_{1.00}}$	$\mathbf{99.93_{0.14}}$	$0.020_{0.008}$	$0.12_{0.08}$	$\mathbf{18.71_{1.47}}$	$\mathbf{92.81_{0.84}}$	$0.005_{0.002}$	$0.14_{0.01}$
	MBF-PCA (10^{-6})	$87.75^{*}_{1.36}$	$72.16^{*}_{0.90}$	$\mathbf{0.014^{*}_{0.002}}$	$\mathbf{0.16^{*}_{0.07}}$	$16.95_{1.52}$	$92.70_{3.00}$	$\mathbf{0.013_{0.007}}$	$\mathbf{0.06_{0.05}}$	$15.49^{*}_{6.44}$	$86.36^{*}_{3.77}$	$\mathbf{0.003^{*}_{0.002}}$	$\mathbf{0.07^{*}_{0.03}}$

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. Lastly, starred(*) results are those such that the maximum iteration is reached before passing the termination criteria in MBF-PCA.

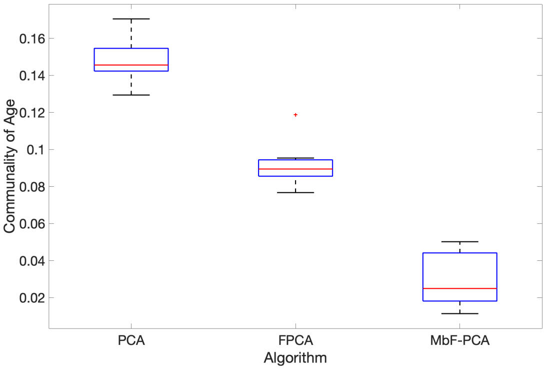


Figure 4: Comparison of communality of "age" of German credit dataset for PCA, FPCA, and MBF-PCA.

Section H of the SP for a full description of the setting. For the hyperparameters, we've set $\delta = 0, \mu = 0.01$ for FPCA and $\tau = 10^{-5}$ for MBF-PCA.

Figure 2 plots the explained variance and fairness metric values. Observe how MBF-PCA achieves better explained variance, while achieving similar level of fairness. In addition, Figure 3 shows a clear gap in runtime between FPCA and MBF-PCA; the runtime of FPCA explodes for even moderate problem sizes, while MBF-PCA scales well. For higher dimensions, conventional computing machine will not be able to handle such computational burden.

UCI Datasets

For target dimensions $d \in \{2, 10\}$, we compare the performance of FPCA and MBF-PCA on 3 datasets from the UCI Machine Learning Repository (Dua and Graff 2017); COMPAS dataset (Kirchner et al. 2016), Adult income dataset, and German credit dataset. See Section I of the SP for complete description of the pre-processing steps. For both algorithms, we consider two different hyperparameter settings, such that one simulates the relaxed fairness while the other simulates a stricter fairness constraints. For computing $\Delta_{DP}(g)$, we trained a RBF SVM g to be the *downstream task classifier* that best classifies the target attribute in the *dimensionality-reduced* data. Table 1 displays the results, in which among the fair algorithms considered, results with the best mean values are **bolded**. Several observations can be made:

- Across all considered datasets, MBF-PCA is shown to outperform FPCA in terms of fairness (both MMD^2 and Δ_{DP}) with low enough τ.
- For GERMAN CREDIT and ADULT INCOME, MBF-PCA shows a clear trade-off between explained variance and fairness; by relaxing τ, we see that MBF-PCA outperforms FPCA in terms of explained variance and downstream task accuracy.

In addition, to see how correlated are the PCs with the protected attribute, we examine the communalities. For clarity of exposition, we consider the German credit dataset, whose protected attribute is age, and $d = 10$. Here, we again consider PCA, FPCA $(0, 0.01)$, and MBF-PCA (10^{-3}). For PCA, communality of a feature is its variance contributed by the PCs (Johnson and Wichern 2008), which is computed as the sum of squares of the loadings of the considered feature. Larger value of communality implies that the correlations between the considered feature and the PCs are strong. Figure 4 displays the boxplot of communality of considered 10 splits. Indeed the amount of variance in age that is accounted for from the loadings of MBF-PCA is much smaller than that of PCA or FPCA i.e. the PCs resulting from MBF-PCA have the *least* correlations with age, the protected attribute.

Conclusion and Future Works

We present a MMD-based definition of fair PCA, and formulate it as a constrained optimization over the Stiefel manifold. Through both theoretical and empirical discussions, we show that our approach outperforms the previous approach (Olfat and Aswani 2019) in terms of explained variance, fairness, and runtime. Many avenues remain for future research: statistical characterizations of our fair PCA in asymptotic regime, as well as incorporation of sparsity (Johnstone et al. 2009); incorporating stochastic optimization-type modifications (Shamir 2015), as such modifications are expected to result in better scalability and performance.

Acknowledgements

This work was partly supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2019-0-01396, Development of framework for analyzing, detecting, mitigating of bias in AI model and training data), and partly supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2021-0-01381, Development of Causal AI through Video Understanding).

References

Absil, P.-A.; Baker, C. G.; and Gallivan, K. A. 2007. Trust-Region Methods on Riemannian Manifolds. *Foundations of Computational Mathematics*, 7(3): 303–330.

Absil, P.-A.; Mahony, R.; and Sepulchre, R. 2007. *Optimization Algorithms on Matrix Manifolds*. USA: Princeton University Press.

Adel, T.; Valera, I.; Ghahramani, Z.; and Weller, A. 2019. One-Network Adversarial Fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2412–2420.

Alsharif, M. H.; Douik, A.; Ahmed, M.; Alnaffouri, T.; and Hassibi, B. 2021. Manifold Optimization for High Accuracy Spacial Location Estimation Using Ultrasound Waves. *IEEE Transactions on Signal Processing*, 1–1.

ApS, M. 2021. MOSEK Optimizer API for Python. Version 9.2.36. *MOSEK*.

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., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Barocas, S.; and Selbst, A. D. 2016. Big Data's Disparate Impact. *104 California Law Review 671*.

Benidis, K.; Sun, Y.; Babu, P.; and Palomar, D. P. 2016. Orthogonal Sparse PCA and Covariance Estimation via Procrustes Reformulation. *IEEE Transactions on Signal Processing*, 64(23): 6211–6226.

Bera, S. K.; Chakrabarty, D.; Flores, N.; and Negahbani, M. 2019. Fair Algorithms for Clustering. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, 4955–4966. Vancouver, BC, Canada.

Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Boumal, N. 2022. An Introduction to Optimization on Smooth Manifolds. http://www.nicolasboumal.net/book, note = Accessed: 2022-01-25.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building Classifiers with Independency Constraints. In *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops*, 13–18. Miami, Florida, USA.

Calders, T.; Karim, A.; Kamiran, F.; Ali, W.; and Zhang, X. 2013. Controlling attribute effect in linear regression. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 71–80. IEEE.

Calmon, F. P.; Wei, D.; Vinzamuri, B.; Ramamurthy, K. N.; and Varshney, K. R. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 3992–4001. Long Beach, CA, USA.

Chien, C.-F.; and Chen, L.-F. 2008. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with applications*, 34(1): 280–290.

Chierichetti, F.; Kumar, R.; Lattanzi, S.; and Vassilvitskii, S. 2017. Fair Clustering Through Fairlets. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 5029–5037. Long Beach, CA, USA.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. *University of California, Irvine, School of Information and Computer Sciences*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. S. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012*, 214–226. Cambridge, MA, USA.

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*, 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., eds., *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., eds., *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25): 723–773.

Han, S. P.; and Mangasarian, O. L. 1979. Exact penalty functions in nonlinear programming. *Mathematical Programming*, 17(1): 251–269.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, 3315–3323. Barcelona, Spain.

Hu, J.; Liu, X.; Wen, Z.-W.; and Yuan, Y.-X. 2020. A Brief Introduction to Manifold Optimization. *Journal of Operations Research Society of China*, 8: 199–248.

Huang, W.; Absil, P.-A.; and Gallivan, K. A. 2018. A Riemannian BFGS Method Without Differentiated Retraction for Nonconvex Optimization Problems. *SIAM Journal on Optimization*, 28(1): 470–495.

Huang, W.; Absil, P.-A.; Gallivan, K. A.; and Hand, P. 2018. ROPTLIB: An Object-Oriented C++ Library for Optimization on Riemannian Manifolds. *ACM Transactions on Mathematical Softwares*, 44(4).

Johnson, R. A.; and Wichern, D. W. 2008. *Applied Multivariate Statistical Analysis*. Pearson, 6 edition.

Johnstone, I. M.; Lu, A. Y.; Nadler, B.; Witten, D. M.; Hastie, T.; Tibshirani, R.; and Ramsay, J. O. 2009. On Consistency and Sparsity for Principal Components Analysis in High Dimensions [with Comments]. *Journal of the American Statistical Association*, 104(486): 682–703.

Jolliffe, I. T.; and Cadima, J. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065).

Jung, S.; Lee, D.; Park, T.; and Moon, T. 2021. Fair Feature Distillation for Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12115–12124.

Kamiran, F.; and Calders, T. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1): 1–33.

Karan, O.; Bayraktar, C.; Gümüşkaya, H.; and Karlık, B. 2012. Diagnosing diabetes using neural networks on small mobile devices. *Expert Systems with Applications*, 39(1): 54–60.

Kirchner, L.; Larson, J.; Mattu, S.; and Angwin, J. 2016. Machine Bias. ProPublica.

Lee, J.; Kim, G.; Olfat, M.; Hasegawa-Johnson, M.; and Yoo, C. D. 2021. Fast and Efficient MMD-based Fair PCA via Optimization over Stiefel Manifold. *arXiv preprint arXiv:2109.11196*.

Liu, C.; and Boumal, N. 2019. Simple Algorithms for Optimization on Riemannian Manifolds with Constraints. *Applied Mathematics and Optimization*, 82: 949–981.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning Adversarially Fair and Transferable Representations. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 3384–3393. PMLR.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2019. A Survey on Bias and Fairness in Machine Learning. *arXiv e-prints*, arXiv:1908.09635.

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*, 663–670.

Olfat, M.; Sloan, S.; Hespanhol, P.; Porter, M.; Vasudevan, R.; and Aswani, A. 2020. Covariance-robust dynamic watermarking. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 3793–3799. IEEE.

Oneto, L.; Donini, M.; Luise, G.; Ciliberto, C.; Maurer, A.; and Pontil, M. 2020. Exploiting MMD and Sinkhorn Divergences for Fair and Transferable Representation Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 15360–15370. Curran Associates, Inc.

Prost, F.; Qian, H.; Chen, Q.; Chi, E. H.; Chen, J.; and Beutel, A. 2019. Toward a better trade-off between performance and fairness with kernel-based distribution matching. In *NeurIPS 2019 Workshop on Machine Learning with Guarantees*. Vancouver, BC, Canada.

Qi, X.; Luo, R.; and Zhao, H. 2013. Sparse principal component analysis by choice of norm. *Journal of Multivariate Analysis*, 114: 127 – 160.

Samadi, S.; Tantipongpipat, U. T.; Morgenstern, J. H.; Singh, M.; and Vempala, S. S. 2018. The Price of Fair PCA: One Extra dimension. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 10999–11010. Montréal, Canada.

Schölkopf, B.; Smola, A.; and Müller, K. 1998. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5): 1299–1319.

Shamir, O. 2015. A Stochastic PCA and SVD Algorithm with an Exponential Convergence Rate. In Bach, F.; and Blei, D., eds., *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 144–152. Lille, France: PMLR.

Sriperumbudur, B. K.; Gretton, A.; Fukumizu, K.; Lanckriet, G. R.; and Schölkopf, B. 2008. Injective Hilbert Space Embeddings of Probability Measures. In Servedio, R.; and Zhang, T., eds., *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 111–222.

Sriperumbudur, B. K.; Gretton, A.; Fukumizu, K.; Schölkopf, B.; and Lanckriet, G. R. 2010. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11(50): 1517–1561.

Tantipongpipat, U.; Samadi, S.; Singh, M.; Morgenstern, J. H.; and Vempala, S. S. 2019. Multi-Criteria Dimensionality Reduction with Applications to Fairness. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, 15135–15145. Vancouver, BC, Canada.

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.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017*, 1171–1180. Perth, Austrailia.

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*, 325–333. Atlanta, GA, USA.