DISTRIBUTIONALLY ROBUST FAIR PRINCIPAL COMPONENTS VIA GEODESIC DESCENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Principal component analysis is a simple yet useful dimensionality reduction technique in modern machine learning pipelines. In consequential domains such as college admission, healthcare and credit approval, it is imperative to take into account emerging criteria such as the fairness and the robustness of the learned projection. In this paper, we propose a distributionally robust optimization problem for principal component analysis which internalizes a fairness criterion in the objective function. The learned projection thus balances the trade-off between the total reconstruction error and the reconstruction error gap between subgroups, taken in the min-max sense over all distributions in a moment-based ambiguity set. The resulting optimization problem over the Stiefel manifold can be efficiently solved by a Riemannian subgradient descent algorithm with a sub-linear convergence rate. Our experimental results on real-world datasets show the merits of our proposed method over state-of-the-art baselines.

1 INTRODUCTION

Machine learning models are ubiquitous in our daily lives and supporting the decision-making process in diverse domains. With their flourishing applications, there also surface numerous concerns regarding the fairness of the models' outputs (Mehrabi et al., 2021). Indeed, these models are prone to biases due to various reasons (Barocas et al., 2018). First, the collected training data is likely to include some demographic disparities due to the bias in the data acquisition process (e.g., conducting surveys on a specific region instead of uniformly distributed places), or the imbalance of observed events at a specific period of time. Second, because machine learning methods only care about data statistics and are objective driven, groups that are under-represented in the data can be neglected in exchange for a better objective value. Finally, even human feedback to the predictive models can also be biased, e.g., click counts are human feedback to recommendation systems but they are highly correlated with the menu list suggested previously by a potentially biased system. Real-world examples of machine learning models that amplify biases and hence potentially cause unfairness are commonplace, ranging from recidivism prediction giving higher false positive rates for African-American[1] to facial recognition systems having large error rate for women[2].

To tackle the issue, various fairness criteria for supervised learning have been proposed in the machine learning literature, which encourage the (conditional) independence of the model's predictions on a particular sensitive attribute (Dwork et al., 2012; Hardt et al., 2016b; Kusner et al., 2017; Chouldechova, 2017; Verma & Rubin, 2018; Berk et al., 2021). Strategies to mitigate algorithmic bias are also investigated for all stages of the machine learning pipelines (Berk et al., 2021). For the pre-processing steps, (Kamiran & Calders, 2012) proposed reweighting or resampling techniques to achieve statistical parity between subgroups; or in the training steps, fairness can be encouraged by adding constraints (Donini et al., 2018) or regularizing the original objective function (Kamishima et al., 2012; Zemel et al., 2013); and in the post-processing steps, adjusting classification threshold based on examining black-box models over a holdout dataset can be used (Hardt et al., 2016b; Wei et al., 2019).

Since biases may already exist in the raw data, it is reasonable to demand the machine learning pipeline to combat biases as early as possible. We focus in this paper on the Principal Component

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[2] https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

Analysis (PCA), which is a fundamental dimensionality reduction technique in the early stage of the pipelines (Pearson, 1901; Hotelling, 1933). PCA finds a linear transformation that embeds the original data into a lower-dimensional subspace that maximizes the variance of the projected data. Thus, PCA is prone to amplify biases if the data variability is different between the majority and the minority subgroups, see a toy example in Figure 1. A naive approach to promote fairness is to train one independent transformation for each subgroup. However, this requires knowing the sensitive attribute of each sample at test time, which would raise disparity concerns. On the contrary, using a single transformation for all subgroups is "group-blinded" and faces no discrimination problem (Lipton et al., 2018).

Learning a fair PCA has attracted attention from many fields from machine learning, statistics to signal process. Samadi et al. (2018) and Zalcberg & Wiesel (2021) propose to find the principal components that minimize the maximum subgroup reconstruction error; the min-max formulations can be relaxed and solved as semidefinite programs. Olfat & Aswani (2019) propose to learn a transformation that minimizes the possibility of predicting the sensitive attribute from the projected data. Apart from being a dimensionality reduction technique, PCA can also be thought of as a representation learning toolkit. Viewed in this way, we can also consider a more general family of fair representation learning methods that can be applied before any further analysis steps. There are a number of works develop towards this idea (Kamiran & Calders, 2012; Zemel et al., 2013; Calmon et al., 2017; Feldman et al., 2015; Beutel et al., 2017; Madras et al., 2018; Zhang et al., 2018; Tantipongpipat et al., 2019), which apply a multitude of fairness criteria.

In addition, we also focus on the robustness criteria for the linear transformation. Recently, it has been observed that machine learning models are susceptible to small perturbations of the data (Goodfellow et al., 2014; Madry et al., 2017; Carlini & Wagner, 2017). These observations have fuelled many defenses using adversarial training (Akhtar & Mian, 2018; Chakraborty et al., 2018) and distributionally robust optimization (Rahimian & Mehrotra, 2019; Kuhn et al., 2019).

Contributions. This paper blends the ideas from the field of fairness in artifical intelligence and distributionally robust optimization. Our contributions can be described as follows.

- We propose the fair principal components which balance between the total reconstruction error and the absolute gap of reconstruction error between subgroups. Moreover, we also add a layer of robustness to the principal components by considering a min-max formulation that hedges against all perturbations of the empirical distribution in a moment-based ambiguity set.
- We provide the reformulation of the distributionally robust fair PCA problem as a finite-dimensional optimization problem over the Stiefel manifold. We provide a Riemannian gradient descent algorithm and show that it has a sub-linear convergence rate.

Figure 1 illustrates the qualitative comparison between (fair) PCA methods and our proposed method on a 2-dimensional toy example. The majority group (blue dots) spreads on the horizontal axis, while the minority group (yellow triangles) spreads on the slanted vertical axis. The nominal PCA (red) captures the majority direction to minimize the total error, while the fair PCA of Samadi et al. (2018) returns the diagonal direction to minimize the maximum subgroup error. Our fair PCA can probe the full spectrum in between these two extremes by sweeping through our penalization parameters appropriately. If we do not penalize the error gap between subgroups, we recover the PCA method; if we penalize heavily, we recover the fair PCA of Samadi et al. (2018). Extensive numerical results on real datasets are provided in Section 5. Proofs are relegated to the appendix.

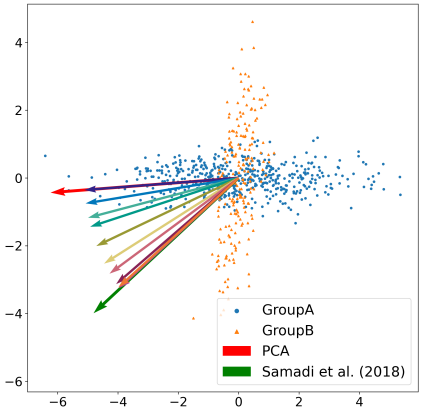

Figure 1: Nominal PCA (red arrow), fair PCA by Samadi et al. (2018) (green arrow), and our spectrum of fair PCA (shorter arrows). Arrows show directions and are not normalized to unit length.

2 FAIR PRINCIPAL COMPONENT ANALYSIS

2.1 PRINCIPAL COMPONENT ANALYSIS

We first briefly revisit the classical PCA. Suppose that we are given a collection of N i.i.d. samples $\{\hat{x}_i\}_{i=1}^N$ generated by some underlying distribution $\mathbb{P}$. For simplicity, we assume that both the empirical and population mean are zero vectors. The goal of PCA is to find a k-dimensional linear subspace of $\mathbb{R}^d$ that explains as much variance contained in the data $\{\hat{x}_i\}_{i=1}^N$ as possible, where $k < d$ is a given integer. More precisely, we parametrize k-dimensional linear subspaces by orthonormal matrices, *i.e.,* matrices whose columns are orthogonal and have unit Euclidean norm. Given any such matrix V, the associated k-dimensional subspace is the one spanned by the columns of V. The projection matrix onto the subspace is $VV^\top$, and hence the variance of the projected data is given by $\operatorname{tr}\left(VV^\top \Xi\Xi^\top\right)$, where $\Xi = [\hat{x}_1, \cdots, \hat{x}_N] \in \mathbb{R}^{d\times N}$ is the data matrix. By a slight abuse of terminology, sometimes we refer to V as the projection matrix. The problem of PCA then reads

$$\max_{V\in\mathbb{R}^{d\times k}, V^\top V = I_k} \operatorname{tr}\left(VV^\top \Xi\Xi^\top\right). \tag{1}$$

For any vector $X \in \mathbb{R}^d$ and orthonormal matrix V, denote by $\ell(V, X)$ the reconstruction error, *i.e.,*

$$\ell(V, X) = \|X - VV^\top X\|_2^2 = X^\top (I_d - VV^\top) X.$$

The problem of PCA can alternatively be formulated as a stochastic optimization problem

$$\min_{V\in\mathbb{R}^{d\times k}, V^\top V = I_k} \mathbb{E}_{\hat{\mathbb{P}}}[\ell(V, X)], \tag{2}$$

where $\hat{\mathbb{P}}$ is the empirical distribution associated with the samples $\{\hat{x}_i\}_{i=1}^N$ and $X \sim \hat{\mathbb{P}}$. It is well-known that PCA admits an analytical solution. In particular, the optimal solution to problem (2) (and also problem (1)) is given by any orthonormal matrix whose columns are the eigenvectors associated with the k largest eigenvalues of the sample covariance matrix $\Xi\Xi^\top$.

2.2 FAIR PRINCIPAL COMPONENT ANALYSIS

In the fair PCA setting, we are also given a discrete sensitive attribute $A \in \mathcal{A}$, where A may represent features such as race, gender or education. We consider binary attribute A and let $\mathcal{A} = \{0, 1\}$. A straightforward idea to define fairness is to require the (strict) balance of a certain objective between the two groups. For example, this is the strategy in Hardt et al. (2016a) for developing fair supervised learning algorithms. A natural objective to balance in the PCA context is the reconstruction error. It is therefore tempted to adopt the following definition.

Definition 2.1 (Fair projection)**.** Let $\mathbb{Q}$ be an arbitrary distribution of (X, A). A projection matrix $V \in \mathbb{R}^{d\times k}$ is fair relative to $\mathbb{Q}$ if the conditional expected reconstruction error is equal between subgroups, *i.e.,*

$$\mathbb{E}_{\mathbb{Q}}[\ell(V, X)|A = a] = \mathbb{E}_{\mathbb{Q}}[\ell(V, X)|A = a'] \qquad \forall (a, a') \in \mathcal{A}\times\mathcal{A}.$$

Unfortunately, Definition 2.1 is too stringent: for a general probability distribution $\mathbb{Q}$, it is possible that there exists no fair projection matrix V.

Proposition 2.2 (Impossibility result)**.** For any distribution $\mathbb{Q}$ on $\mathcal{X}\times\mathcal{A}$, let $S = \mathbb{E}_{\mathbb{Q}}[XX^\top|A = 0] - \mathbb{E}_{\mathbb{Q}}[XX^\top|A = 1]$. Then, there exists a fair projection matrix $V \in \mathbb{R}^{d\times k}$ relative to $\mathbb{Q}$ if and only if $\operatorname{rank}(S) \le k$.

One way to circumvent the impossibility result is to relax the requirement of strict balance to approximate balance. In other words, an inequality constraint of the following form is imposed:

$$|\mathbb{E}_{\mathbb{Q}}[\ell(V, X)|A = a] - \mathbb{E}_{\mathbb{Q}}[\ell(V, X)|A = a']| \le \epsilon \qquad \forall (a, a') \in \mathcal{A}\times\mathcal{A},$$

where $\epsilon > 0$ is some prescribed fairness threshold. This approach has been adopted in other fair machine learning settings, see Donini et al. (2018) and Agarwal et al. (2019) for example.

In this paper, instead of imposing the fairness requirement as a constraint, we penalize the unfairness in the objective function. Specifically, for any projection matrix V, we define the unfairness as the absolute difference between the conditional loss between two subgroups:

$$\mathbb{U}(V, \mathbb{Q}) \triangleq |\mathbb{E}_{\mathbb{Q}}[\ell(V, X)|A = 0] - \mathbb{E}_{\mathbb{Q}}[\ell(V, X)|A = 1]|.$$

We thus consider the following fairnes-aware PCA problem

$$\min_{V\in\mathbb{R}^{d\times k},\ V^\top V=I_k} \mathbb{E}_{\hat{\mathbb{P}}}[\ell(V,X)] + \lambda \mathbb{U}(V,\hat{\mathbb{P}}), \tag{3}$$

where $\lambda \geq 0$ is a penalty parameter to encourage fairness. Note that for fair PCA, the dataset is $\{(\hat{x}_i,\hat{a}_i)\}_{i=1}^N$ and hence the empirical distribution $\hat{\mathbb{P}}$ is given by $\hat{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^N \delta_{(\hat{x}_i,\hat{a}_i)}$.

3 Distributionally Robust Fair PCA

The weakness of empirical distribution-based stochastic optimization has been well-documented, see (Smith & Winkler, 2006; Homem-de Mello & Bayraksan, 2014). In particular, due to overfitting, the out-of-sample performance of the decision, prediction, or estimation obtained from such a stochastic optimization model is unsatisfactory, especially in the low sample size regime. Ideally, we could improve the performance by using the underlying distribution $\mathbb{P}$ instead of the empirical distribution $\hat{\mathbb{P}}$. But the underlying distribution $\mathbb{P}$ is unavailable in most practical situations, if not all. Distributional robustification is an emerging approach to handle this issue and has been shown to deliver promising out-of-sample performance in many applications (Delage & Ye, 2010; Namkoong & Duchi, 2017; Kuhn et al., 2019; Rahimian & Mehrotra, 2019). Motivated by the success of distributional robustification, we propose a robustified version of model (3), called the distributionally robust fairness-aware PCA:

$$\min_{V\in\mathbb{R}^{d\times k}, V^\top V=I_k} \sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\ell(V,X)] + \lambda \mathbb{U}(V,\mathbb{Q}), \tag{4}$$

where $\mathbb{B}(\hat{\mathbb{P}})$ is a set of probability distributions similar to the empirical distribution $\hat{\mathbb{P}}$ in a certain sense, called the ambiguity set. The empirical distribution $\hat{\mathbb{P}}$ is also called the nominal distribution. Many different ambiguity sets have been developed and studied in the optimization literature, see Rahimian & Mehrotra (2019) for an extensive overview.

3.1 The Wasserstein-type Ambiguity Set

To present our ambiguity set and main results, we need to introduce some definitions and notations.

Definition 3.1 (Wasserstein-type divergence)**.** The divergence $\mathbb{W}$ between two probability distributions $\mathbb{Q}_1 \sim (\mu_1,\Sigma_1) \in \mathbb{R}^d \times \mathbb{S}_+^d$ and $\mathbb{Q}_2 \sim (\mu_2,\Sigma_2) \in \mathbb{R}^d \times \mathbb{S}_+^d$ is defined as

$$\mathbb{W}\big(\mathbb{Q}_1 \parallel \mathbb{Q}_2\big) \triangleq \|\mu_1-\mu_2\|_2^2 + \operatorname{tr}\left(\Sigma_1+\Sigma_2-2\big(\Sigma_2^{\frac{1}{2}}\Sigma_1\Sigma_2^{\frac{1}{2}}\big)^{\frac{1}{2}}\right).$$

The divergence $\mathbb{W}$ coincides with the *squared* type-2 Wasserstein distance between two Gaussian distributions $\mathcal{N}(\mu_1,\Sigma_1)$ and $\mathcal{N}(\mu_2,\Sigma_2)$ (Givens & Shortt, 1984). One can readily show that $\mathbb{W}$ is non-negative, and it vanishes if and only if $(\mu_1,\Sigma_1)=(\mu_2,\Sigma_2)$, which implies that $\mathbb{Q}_1$ and $\mathbb{Q}_2$ have the same first- and second-moments.

Recall that the nominal distribution is $\hat{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^N \delta_{(\hat{x}_i,\hat{a}_i)}$. For any $a\in\mathcal{A}$, its conditional distribution given $A=a$ is given by

$$\hat{\mathbb{P}}_a = \frac{1}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}\delta_{x_i}, \quad \text{where} \quad \mathcal{I}_a \triangleq \{i\in\{1,\ldots,N\} : a_i = a\}.$$

We also use $(\hat{\mu}_a,\hat{\Sigma}_a)$ to denote the empirical mean vector and covariance matrix of X given $A=a$:

$$\hat{\mu}_a = \mathbb{E}_{\hat{\mathbb{P}}_a}[X] = \mathbb{E}_{\hat{\mathbb{P}}}[X|A=a] \quad \text{and} \quad \hat{\Sigma}_a + \hat{\mu}_a\hat{\mu}_a^\top = \mathbb{E}_{\hat{\mathbb{P}}_a}[XX^\top] = \mathbb{E}_{\hat{\mathbb{P}}}[XX^\top|A=a].$$

For any $a\in\mathcal{A}$, the empirical marginal distribution of A is denoted by $\hat{p}_a = |\mathcal{I}_a|/N$.

Finally, for any set $\mathcal{S}$, we use $\mathcal{P}(\mathcal{S})$ to denote the set of all probability distributions supported on $\mathcal{S}$. For any integer k, the k-by-k identity matrix is denoted I_k.

We then define our ambiguity set as

$$\mathbb{B}(\hat{\mathbb{P}}) \triangleq \left\{ \mathbb{Q}\in\mathcal{P}(\mathcal{X}\times\mathcal{A}) : \begin{array}{l} \exists \mathbb{Q}_a \in \mathcal{P}(\mathcal{X}) \text{ such that:} \\ \mathbb{Q}(\mathrm{d}x\times\mathrm{d}a) = \sum_{a\in\mathcal{A}} \hat{p}_a \mathbb{Q}_a(\mathrm{d}x)\delta_a(\mathrm{d}a) \\ \mathbb{W}(\mathbb{Q}_a,\hat{\mathbb{P}}_a) \leq \varepsilon_a \quad \forall a\in\mathcal{A} \end{array} \right\}, \tag{5}$$

where $\mathbb{Q}_a$ is the conditional distribution of $X|A=a$. Intuitively, each $\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})$ is a joint distribution of the random vector (X, A), formed by taking a mixture of conditional distributions $\mathbb{Q}_a$ with mixture weight $\hat{p}_a$. Each conditional distribution $\mathbb{Q}_a$ is constrained in an ε_a-neighborhood from the nominal conditional distribution $\hat{\mathbb{P}}_a$ with respect to the $\mathbb{W}$ divergence. Notice that because the loss function ℓ is a quadratic function of X, the (conditional) expected losses only involve the first two moments of X, and thus prescribing the ambiguity set using $\mathbb{W}$ would suffice for the purpose of robustification.

3.2 Reformulation

We now present the reformulation of problem (4) under the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$.

Theorem 3.2 (Reformulation)**.** Suppose that either of the following two conditions holds:

(i) $0 \leq \lambda \leq \min\{\hat{p}_a, \hat{p}_{a'}\}$,

(ii) for any $a \in \mathcal{A}$, the empirical second moment matrix $\hat{M}_a = \frac{1}{N_a}\sum_{i\in\mathcal{I}_a} \hat{x}_i\hat{x}_i^\top$ satisfies $\sum_{j=1}^{d-k} \sigma_j(\hat{M}_a) \geq \varepsilon_a$, where $\sigma_j(\hat{M}_a)$ is the j-th smallest eigenvalues of $\hat{M}_a$.

Then problem (4) is equivalent to

$$\min_{V\in\mathbb{R}^{d\times k}, V^\top V = I_k} \max\{J_0(V), J_1(V)\}, \tag{6a}$$

where for each $(a, a') \in \{(0,1),(1,0)\}$, the function J_a is defined as

$$J_a(V) = \kappa_a + \theta_a\sqrt{\langle I_d - VV^\top, \hat{M}_a\rangle} + \vartheta_{a'}\sqrt{\langle I_d - VV^\top, \hat{M}_{a'}\rangle} + \langle I_d - VV^\top, C_a\rangle, \tag{6b}$$

and the parameters $\kappa \in \mathbb{R}$, $\theta \in \mathbb{R}$, $\vartheta \in \mathbb{R}$ and $C \in \mathbb{S}_+^d$ are defined as

$$\begin{aligned} \kappa_a &= (\hat{p}_a + \lambda)\varepsilon_a + (\hat{p}_{a'} - \lambda)\varepsilon_{a'}, \quad \theta_a = 2|\hat{p}_a + \lambda|\sqrt{\varepsilon_a}, \quad \vartheta_{a'} = 2|\hat{p}_{a'} - \lambda|\sqrt{\varepsilon_{a'}}, \\ C_a &= (\hat{p}_a + \lambda)\hat{M}_a + (\hat{p}_{a'} - \lambda)\hat{M}_{a'}. \end{aligned} \tag{6c}$$

We now briefly explain the steps that lead to the results in Theorem 3.2. Letting

$$\begin{aligned} J_0(V) &= \sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} (\hat{p}_0 + \lambda)\mathbb{E}_{\mathbb{Q}}[\ell(V,X)|A=0] + (\hat{p}_1 - \lambda)\mathbb{E}_{\mathbb{Q}}[\ell(V,X)|A=1], \\ J_1(V) &= \sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} (\hat{p}_0 - \lambda)\mathbb{E}_{\mathbb{Q}}[\ell(V,X)|A=0] + (\hat{p}_1 + \lambda)\mathbb{E}_{\mathbb{Q}}[\ell(V,X)|A=1], \end{aligned}$$

then by expanding the term $\mathbb{U}(V, \mathbb{Q})$ using its definition, problem (4) becomes

$$\min_{V\in\mathbb{R}^{d\times k}, V^\top V = I_k} \max\{J_0(V), J_1(V)\}.$$

By the definition the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$, for any pair $(a, a') \in \{(0,1),(1,0)\}$, we can decompose J_a into two separate supremum problems as follows

$$J_a(V) = \sup_{\mathbb{Q}_a:\mathbb{W}(\mathbb{Q}_a,\hat{\mathbb{P}}_a)\leq\varepsilon_a} (\hat{p}_a + \lambda)\mathbb{E}_{\mathbb{Q}_a}[\ell(V,X)] + \sup_{\mathbb{Q}_1:\mathbb{W}(\mathbb{Q}_{a'},\hat{\mathbb{P}}_{a'})\leq\varepsilon_{a'}} (\hat{p}_{a'} - \lambda)\mathbb{E}_{\mathbb{Q}_{a'}}[\ell(V,X)].$$

The next proposition asserts that each individual supremum in the above expression admits an analytical expression.

Proposition 3.3 (Reformulation)**.** Fix $a \in \mathcal{A}$. For any $\upsilon \in \mathbb{R}$, $\varepsilon_a \in \mathbb{R}_+$, it holds that

$$\begin{aligned} &\sup_{\mathbb{Q}_a:\mathbb{W}(\mathbb{Q}_a,\hat{\mathbb{P}}_a)\leq\varepsilon_a} \upsilon\mathbb{E}_{\mathbb{Q}_a}[\ell(V,X)] \\ &= \begin{cases} \upsilon\left(\sqrt{\langle I_d - VV^\top, \hat{M}_a\rangle} + \sqrt{\varepsilon_a}\right)^2 & \text{if } \upsilon \geq 0, \\ \upsilon\left(\sqrt{\langle I_d - VV^\top, \hat{M}_a\rangle} - \sqrt{\varepsilon_a}\right)^2 & \text{if } \upsilon < 0 \text{ and } \langle I_d - VV^\top, \hat{M}_a\rangle \geq \varepsilon_a, \\ 0 & \text{if } \upsilon < 0 \text{ and } \langle I_d - VV^\top, \hat{M}_a\rangle < \varepsilon_a. \end{cases} \end{aligned}$$

The proof of Theorem 3.2 now follows by applying Proposition 3.3 to each term in J_a, and balance the parameters to obtain (6c). A detailed proof is relegated to the appendix. In the next section, we study an efficient algorithm to solve (6a).

Remark 3.4 (Recovery of the nominal PCA)**.** If $\lambda = 0$ and $\varepsilon_a = 0\ \forall a \in \mathcal{A}$, our formulation (4) becomes the standard PCA problem (2). In this case, our robust fair principal components reduce to the standard principal components. On the contrary, existing fair PCA methods such as Samadi et al. (2018) and Olfat & Aswani (2019) cannot recover the standard principal components.

4 Riemannian Gradient Descent Algorithm

Using Theorem 3.2, our distributionally robust fairness-aware PCA problem (4), which is an infinite-dimensional minimax problem, is reduced to the simpler finite-dimensional minimax problem (6a), where the inner problem is only a maximization over two points. Problem (6a) is, however, still challenging as it is a non-convex optimization problem over a non-convex feasible region defined by the orthogonality constraint $V^\top V = I_d$. The purpose of this section is to devise an efficient algorithm for solving problem (6a) to local optimality based on Riemannian optimization.

4.1 Reparametrization

As mentioned above, the non-convexity of problem (6a) comes from both the objective function and the feasible region. It turns out that we can get rid of the non-convexity of the objective function via a simple change of variables. To see that, we let $U \in \mathbb{R}^{d\times(d-k)}$ be an orthonormal matrix complement to V, that is, U and V satisfy $UU^\top + VV^\top = I_d$. Thus, we can express the objective function J via

$$J(V) = F(U) \triangleq \max\{F_0(U), F_1(U)\},$$

where for $(a, a') \in \{(0,1),(1,0)\}$, the function F_a is defined as

$$F_a(U) \triangleq \kappa_a + \theta_a\sqrt{\langle UU^\top, \hat{M}_a\rangle} + \vartheta_{a'}\sqrt{\langle UU^\top, \hat{M}_{a'}\rangle} + \langle UU^\top, C_a\rangle.$$

Moreover, letting $\mathcal{M} \triangleq \{U \in \mathbb{R}^{d\times(d-k)} : U^\top U = I_{d-k}\}$, we can re-express problem (6a) as

$$\min_{U\in\mathcal{M}}\ F(U). \tag{7}$$

The feasible region $\mathcal{M}$ of problem (7) is a Riemannian manifold, called the Stiefel manifold (Absil et al., 2007, Section 3.3.2). It is then natural to solve problem (7) by using Riemannian optimization algorithms (Absil et al., 2007). In fact, problem (6a) itself (before the change of variables) can also be seen as a Riemannian optimization problem over another Stiefel manifold. The change of variables above might seem unnecessary. Nonetheless, the upshot of problem (7) is that the objective function F is convex (in the traditional sense). This faciliates the application of the theoretical and algorithmic framework developed in Li et al. (2019) for (weakly) convex optimization over Stiefel manifolds.

4.2 The Riemannian Subgradient

Note that the objective function F is non-smooth since it is defined as the maximum of two functions F_0 and F_1. To apply the framework in Li et al. (2019), we need to compute the Riemannian subgradient of the objective function F. Since the Stiefel manifold $\mathcal{M}$ is an embedded manifold in Euclidean space, the Riemannian subgradient of F at any point $U \in \mathcal{M}$ is given by the orthogonal projection of the usual Euclidean subgradient onto the tangent space of the manifold $\mathcal{M}$ at the point U, see Absil et al. (2007, Section 3.6.1) for example.

Lemma 4.1. For any point $U \in \mathcal{M}$, let[3] $a_U \in \arg\max_{a\in\{0,1\}} F_a(U)$ and $a'_U = 1 - a_U$. Then, a Riemannian subgradient of the objective function F at the point U is given by

$$\mathrm{grad}F(U) = (I_d - UU^\top)\left(\frac{\theta_{a_U}}{\sqrt{\langle UU^\top, \hat{M}_{a_U}\rangle}}\hat{M}_{a_U}U + \frac{\vartheta_{a'_U}}{\sqrt{\langle UU^\top, \hat{M}_{a'_U}\rangle}}\hat{M}_{a'_U}U + 2C_{a_U}U\right).$$

[3]It is possible that the maximizer is not unique. In that case, choosing a_U to be either 0 or 1 would work.

4.3 RETRACTIONS

Another important instrument required by the framework in Li et al. (2019) is a retraction of the Stiefel manifold $\mathcal{M}$. At each iteration, the point $U - \gamma\Delta$ obtained by moving from the current iterate U in the opposite direction of the Riemannian gradient Δ may not lie on the manifold in general, where $\gamma > 0$ is the stepsize. In Riemannian optimization, this is circumvented by the concept of retraction. Given a point $U \in \mathcal{M}$ on the manifold, the Riemannian gradient $\Delta \in T_U\mathcal{M}$ (which must lie in the tangent space $T_U\mathcal{M}$) and a stepsize γ, the retraction map Rtr defines a point $\mathrm{Rtr}_U(-\gamma\Delta)$ which is guaranteed to lie on the manifold $\mathcal{M}$. Roughly speaking, the retraction $\mathrm{Rtr}_U(\,\cdot\,)$ approximates the geodesic curve through U along the input tangential direction. For a formal definition of retractions, we refer the readers to (Absil et al., 2007, Section 4.1). In this paper, we focus on the following two commonly used retractions for Stiefel manifolds. The first one is the QR decomposition-based retraction

$$\mathrm{Rtr}_U^{\mathrm{qf}}(\Delta) = \mathrm{qf}(U + \Delta), \quad U \in \mathcal{M}, \Delta \in T_U\mathcal{M},$$

where $\mathrm{qf}(\,\cdot\,)$ is the Q-factor in the QR decomposition. The second one is the polar decomposition-based retraction

$$\mathrm{Rtr}_U^{\mathrm{polar}}(\Delta) = (U + \Delta)(I_{d-k} + \Delta^\top\Delta)^{-\frac{1}{2}}, \quad U \in \mathcal{M}, \Delta \in T_U\mathcal{M}. \tag{8}$$

4.4 ALGORITHM AND CONVERGENCE GUARANTEES

Associated with any choice of retraction Rtr is a concrete instantiation of the Riemannian subgradient descent algorithm for our problem (7), which is presented in Algorithm 1.

Algorithm 1 Riemannian Subgradient Descent for (7)

1: **Input:** An initial point U_0, a number of iterations τ and a retraction $\mathrm{Rtr} : (U, \Delta) \mapsto \mathrm{Rtr}_U(\Delta)$.
2: **for** $t = 0, 1, \ldots, \tau - 1$, **do**
3: Find $a_t \triangleq \arg\max_{a\in\{0,1\}}\{F_a(U_t)\}$.
4: Compute the Riemannian subgradient $\Delta_t = \mathrm{grad}F(U_t)$ using the formula

$$\Delta_t = (I - U_tU_t^\top)\left(\frac{\theta_{a_t}}{\sqrt{\langle U_tU_t^\top, \hat{M}_{a_t}\rangle}}\hat{M}_{a_t}U_t + \frac{\vartheta_{a_t'}}{\sqrt{\langle U_tU_t^\top, \hat{M}_{a_t'}\rangle}}\hat{M}_{a_t'}U_t + 2C_{a_t}U_t\right).$$

5: Set $U_{t+1} = \mathrm{Rtr}_{U_t}(-\gamma_t\Delta_t)$, where the step-size $\gamma_t \equiv \frac{1}{\sqrt{\tau+1}}$ is constant.
6: **end for**
7: **Output:** U_τ.

The specific choice of the stepsizes γ_t is motivated by the theoretical results of (Li et al., 2019).

We now study the convergence guarantee of Algorithm 1. The following lemma shows that the objective function F is Lipschitz continuous (with respect to the Riemannian metric on the Stiefel manifold $\mathcal{M}$) with an explicit Lipschitz constant L.

Lemma 4.2 (Lipschitz continuity)**.** The function F is L-Lipschitz continuous on $\mathcal{M}$, where $L > 0$ is given by

$$\begin{aligned} L \triangleq \max\Bigg\{&\theta_0\frac{\sigma_{\max}(\hat{M}_0)}{\sqrt{\sigma_{\min}(\hat{M}_0)}}, \theta_1\frac{\sigma_{\max}(\hat{M}_1)}{\sqrt{\sigma_{\min}(\hat{M}_1)}}, \vartheta_0\frac{\sigma_{\max}(\hat{M}_0)}{\sqrt{\sigma_{\min}(\hat{M}_0)}}, \vartheta_1\frac{\sigma_{\max}(\hat{M}_1)}{\sqrt{\sigma_{\min}(\hat{M}_1)}}, \\ &2\sqrt{d-k}\sigma_{\max}(C_0), 2\sqrt{d-k}\sigma_{\max}(C_1)\Bigg\}. \end{aligned} \tag{9}$$

We now proceed to show that Algorithm 1 enjoys a sub-linear convergence rate. To state the result, we define the Moreau envelope

$$F_\mu(U) \triangleq \min_{U'\in\mathcal{M}}\left\{F(U') + \frac{1}{2\mu}\|U' - U\|_F^2\right\},$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Also, to measure the progress of the algorithm, we need to introduce the proximal mapping on the Stiefel manifold (Li et al., 2019):

$$\text{prox}_{\mu F}(U) \in \underset{U' \in \mathcal{M}}{\arg\min} \left\{ F(U') + \frac{1}{2\mu} \|U' - U\|_F^2 \right\}.$$

From Li et al. (2019, Equaton (22)), we have that

$$\|\text{grad}F(U)\|_F \leq \frac{\left\|\text{prox}_{\mu F}(U) - U\right\|_F}{\mu} \triangleq \text{gap}_\mu(U).$$

Therefore, the number $\text{gap}_\mu(U)$ is a good candidate to quantify the progress of optimization algorithms for solving problem (7).

Theorem 4.3 (Convergence guarantee)**.** Let $\{U_t\}_{t=1,\ldots,\tau}$ be the sequence of iterates generated by Algorithm 1. Suppose that $\mu = 1/4L$, where L is the Lipschitz constant of F in (9). Then, we have

$$\min_{t=0,\ldots,\tau} \text{gap}_\mu(U_t) \leq \frac{2\sqrt{F_\mu(U_0) - \min_U F_\mu(U) + 2L^3(L+1)}}{(\tau+1)^{1/4}}.$$

5 Numerical Experiments

We compare our proposed method, denoted `RFPCA`, against two state-of-the-art methods for fair PCA: 1) `FairPCA` from Samadi et al. (2018)[4], and 2) `CFPCA` from Olfat & Aswani (2019)[5] with both cases: only mean constraint, and both mean and covariance constraints. We consider a wide variety of datasets from UC Irvine's online Machine Learning Repository (Dua & Graff, 2017) with ranging sample sizes and number of features. Further details about the datatasets can be found in Appendix B. The code for all experiments is available in supplementary materials.

We include here some details about the hyper-parameters that we search in the cross-validation steps.

- `RFPCA`. We notice that the neighborhood size ε_a should be inversely proportional to the size of subgroup a. Indeed, a subgroup with large sample size is likely to have more reliable estimate of the moment information. Then we parameterize the neighborhood size ε_a by a common scalar α, and we have $\varepsilon_a = \alpha/N_a$, where N_a is the number of samples in group a. We search $\alpha \in \{0.05, 0.1, 0.15\}$ and $\lambda \in \{0., 0.5, 1., 1.5, 2.0, 2.5\}$. For better convergence quality, we set the number of iteration for our subgradient descent algorithm to $\tau = 1000$ and also repeat the Riemannian descent for 20 randomly generated initial point U_0.
- `FairPCA`. According to Samadi et al. (2018), we only need tens of iterations for the multiplicative weight algorithm to provide good-quality solution; however, to ensure a fair comparison, we set the number of iterations to be 1000 for the convergence guarantee. We search the learning rate η of the algorithm from set of 17 values evenly spaced in $[0.25, 4.25]$ and $\{0.1\}$.
- `CFPCA`. Followed Olfat & Aswani (2019), for the mean-constrained version of `CFPCA`, we search δ from $\{0., 0.1, 0.3, 0.5, 0.6, 0.7, 0.8, 0.9\}$, and for both mean and covariance constrained version, we fix $\delta = 0$ while searching μ in $\{0.0001, 0.001, 0.01, 0.05, 0.5\}$.

Trade-offs. First, we examine the trade-off between the total reconstruction error and the gap between the subgroup error. In this experiment, we only compare our model with `FairPCA` and `CFPCA` mean-constraint version. We plot a pareto curve for each of them over the two criteria with different hyper-parameters (hyper-parameters test range are mentioned above). The whole datasets are used for training and evaluation. The results averaged over 5 runs are shown in Figure 2.

In testing methods with different principal components, we first split each dataset into training set and test set with equal size (50% each), the projection matrix of each method is learned from training set and tested over both sets. In this case, we only compare our method with traditional PCA and `FairPCA` method. We fix one set hyper-parameters for each method. For `FairPCA`, we set $\eta = 0.1$ and for `RFPCA` we set $\alpha = 0.15, \lambda = 0.5$, others hyper-parameters are kept as discussed before. The results are averaged over 5 different splits. Figure 3 shows the consistence of our method performing fair projections over different values of k. Our method (cross) exhibits smaller gap of subgroup errors. More results can be found in Appendix C.2.

[4]https://github.com/samirasamadi/Fair-PCA

[5]https://github.com/molfat66/FairML

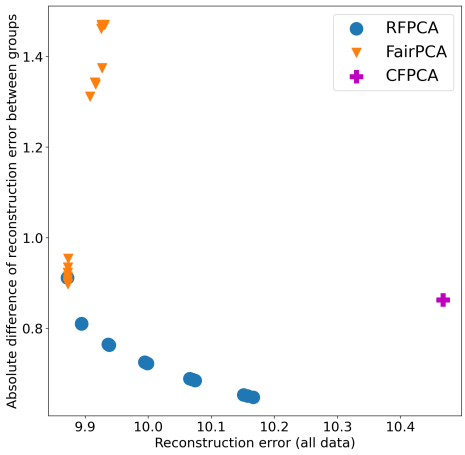

Figure 2: Pareto curves on Default Credit dataset (all data) with 3 principal components

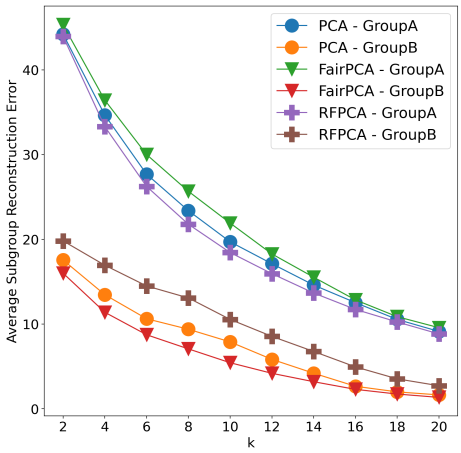

Figure 3: Subgroup average error with different k on Biodeg dataset (Out-of-sample).

Cross-validations. Next, we report the performance of all methods based on four criteria: absolute difference between average reconstruction error between groups (ABDiff.), average reconstruction error of all data (ARE.), and the fairness criterion defined by Olfat & Aswani (2019) with respect to a linear SVM's classifier family ($\triangle\mathcal{F}_{Lin}$) [6]. Due to the space constraint, we only include the first two criteria in the main text, see Appendix 4 for full results. In Each dataset, one feature is selected as a sensitive attribute, other features is used as the input for algorithms. To emphasize the the generalization capacity of each algorithm, we split each dataset into a training set and a test set with ratio of $30\% - 70\%$ respectively, and only extract top three principal components from the training set. We find the best hyper-parameters by 3-fold cross validation, and prioritize the one giving minimum value of the summation (ABDiff. + ARE.). The results are averaged over 10 different training-testing splits. We report the performance on both training set (In-sample data) and test set (Out-of-sample data). The details results for Out-of-sample data is given in Table 1 while one for In-sample data is reported in the appendix at Table 3.

Results. Our proposed `RFPCA` method outperforms on 10 out of 14 datasets in terms of the subgroup error gap ABDiff, and 8 out of 14 with the totall error ARE. criterion. There are 4 datasets that `RFPCA` gives the best results for both criteria, and for the remaining datasets, `RFPCA` has small performance gaps compared with the best method.

Table 1: Out-of-sample errors on real datasets. Bold indicates the lowest error for each dataset.

	`RFPCA`		`FairPCA`		`CFPCA`-Mean Con.		`CFPCA` - Both Con.	
Dataset	ABDiff.	ARE.	ABDiff.	ARE.	ABDiff.	ARE.	ABDiff.	ARE.
Default Credit	0.9483	**10.3995**	1.4401	10.4439	**0.9367**	10.9451	3.3359	22.0310
Biodeg	**23.0066**	**33.8571**	27.5159	34.6184	29.1728	37.6052	37.9533	50.7090
E. Coli	1.1500	**1.7210**	1.5280	2.4799	**1.1005**	2.9466	5.1275	5.6674
Energy	**0.0125**	0.2238	0.0138	**0.2225**	0.1229	2.7318	0.1001	7.9511
German Credit	2.0588	**43.9032**	**1.3670**	44.0064	1.7845	43.9648	1.4955	49.5014
Image	**0.7522**	**6.0199**	1.6129	10.2616	1.1499	14.3725	4.7013	19.3356
Letter	**0.1712**	**7.4176**	1.2489	7.4470	0.4427	8.7445	0.5743	15.1779
Magic	**1.8314**	3.9094	2.9405	**3.3815**	5.5790	4.2105	8.7810	9.0064
Parkinsons	**0.3273**	5.0597	0.8678	**4.9044**	3.3804	5.7260	18.3312	19.7001
SkillCraft	**0.7669**	8.2828	0.7771	**8.2494**	1.0283	9.9484	1.2849	15.9751
Statlog	**0.0838**	**3.0998**	0.3356	7.9734	0.4476	10.8263	13.8437	35.8268
Steel	**1.1472**	12.5944	1.2208	**12.3096**	4.8710	16.4015	3.8084	25.8953
Taiwan Credit	**0.5523**	10.9845	0.5710	**10.9415**	0.5744	13.0437	0.9535	21.8963
Wine Quality	0.6359	**4.2801**	**0.3046**	6.0936	1.5020	6.1118	3.0451	10.1001

[6]code for estimate this quantity is provided at the author's repository

REFERENCES

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

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pp. 120–129. PMLR, 2019.

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning. fairmlbook. org, 2019, 2018.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3995–4004, 2017.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

C.R. Givens and R.M. Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, pp. 3315–3323, 2016a.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016b.

Tito Homem-de Mello and Güzin Bayraksan. Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85, 2014.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2012.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pp. 130–166. INFORMS, 2019.

Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man Cho So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *arXiv preprint arXiv:1911.05047*, 2019.

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ML's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pp. 8125–8135, 2018.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems 30*, pp. 2971–2980, 2017.

Viet Anh Nguyen. *Adversarial Analytics*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2019.

Matt Olfat and Anil Aswani. Convex formulations for fair principal component analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 663–670, 2019.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems*, pp. 10976–10987, 2018.

James E Smith and Robert L Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

Uthaipon Tantipongpipat, Samira Samadi, Mohit Singh, Jamie Morgenstern, and Santosh Vempala. Multi-criteria dimensionality reduction with applications to fairness. *arXiv preprint arXiv:1902.11281*, 2019.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (fairware)*, pp. 1–7. IEEE, 2018.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio du Pin Calmon. Optimized score transformation for fair classification. *arXiv preprint arXiv:1906.00066*, 2019.

Gad Zalcberg and Ami Wiesel. Fair principal component analysis and filter design. *IEEE Transactions on Signal Processing*, 69:4835–4842, 2021.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

A PROOFS

A.1 PROOFS OF SECTION 2

Proof of Proposition 2.2. We first prove the "only if" direction. Suppose that there exists a fair projection matrix $V \in \mathcal{M}_k$ relative to $\mathbb{Q}$. Let $U \in \mathcal{M}_{d-k}$ be a complement matrix of V. Then, Definition 2.1 can be rewritten as

$$\langle UU^\top, S\rangle = 0,$$

which implies that the null space of S has a dimension at least $d-k$. By the rank-nullity duality, we have $\operatorname{rank}(S) \leq k$.

Next, we prove the "if" direction. Suppose that $\operatorname{rank}(S) \leq k$. Then, the matrix S has at least $d-k$ (repeated) zero eigenvalues. Let $U \in \mathcal{M}_{d-k}$ be an orthonormal matrix whose columns are any $d-k$ eigenvectors corresponding to the zero eigenvalues of S and $V \in \mathcal{M}_k$ be a complement matrix of U. Then,

$$\langle I_d - VV^\top, S\rangle = \langle UU^\top, S\rangle = 0.$$

Therefore, V is a fair projection matrix relative to $\mathbb{Q}$. This completes the proof. □

A.2 PROOF OF SECTION 3

Proofs of Proposition 3.3. By exploiting the definition of the loss function ℓ, we find

$$\begin{aligned}
&\sup_{\mathbb{Q}_a:\mathbb{W}(\mathbb{Q}_a,\hat{\mathbb{P}}_a)\leq\varepsilon_a} \upsilon\mathbb{E}_{\mathbb{Q}_a}[\ell(V,X)] \\
&= \left\{\begin{array}{ll} \sup\limits_{\mu_a,\Sigma_a} & \operatorname{tr}\left(\upsilon(I-VV^\top)(\Sigma_a+\mu_a\mu_a^\top)\right) \\ \text{s.t.} & \|\mu_a-\hat{\mu}_a\|_2^2 + \operatorname{tr}\left(\Sigma_a+\hat{\Sigma}_a - 2\big(\hat{\Sigma}_a^{\frac{1}{2}}\Sigma_a\hat{\Sigma}_a^{\frac{1}{2}}\big)^{\frac{1}{2}}\right) \leq \varepsilon_a \end{array}\right. \\
&= \left\{\begin{array}{ll} \inf & \gamma(\varepsilon_a - \operatorname{tr}\left(\hat{\Sigma}_a\right)) + \gamma^2 \operatorname{tr}\left((\gamma I - \upsilon(I-VV^\top))^{-1}\hat{\Sigma}_a\right) + \tau \\ \text{s.t.} & \begin{bmatrix} \gamma I - \upsilon(I-VV^\top) & \gamma\hat{\mu}_a \\ \gamma\hat{\mu}_a^\top & \gamma\|\hat{\mu}_a\|_2^2+\tau \end{bmatrix} \succeq 0, \quad \gamma I \succ \upsilon(I-VV^\top), \quad \gamma \geq 0, \end{array}\right.
\end{aligned}$$

where the last equality follows from Nguyen (2019, Lemma 3.22). By the Woodbury matrix inversion, we have

$$(\gamma I - \upsilon(I-VV^\top))^{-1} = \gamma^{-1}I - \frac{\upsilon}{\gamma(\upsilon-\gamma)}(I-VV^\top).$$

Moreover, using the Schur complement, the semidefinite constraint is equivalent to

$$\gamma\|\hat{\mu}_a\|_2^2 + \tau \geq \gamma^2\hat{\mu}_a^\top(\gamma I - \upsilon(I-VV^\top))^{-1}\hat{\mu}_a,$$

which implies that at optimality, we have

$$\tau = \frac{\upsilon\gamma}{\gamma-\upsilon}\hat{\mu}_a^\top(I-VV^\top)\hat{\mu}_a.$$

At the same time, the constraint $\gamma I \succ \upsilon(I - VV^\top)$ is equivalent to $\gamma > \upsilon$. Combining all previous equations, we have

$$\sup_{\mathbb{Q}_a:\mathbb{W}(\mathbb{Q}_a,\hat{\mathbb{P}}_a)\leq\varepsilon_a} \upsilon\mathbb{E}_{\mathbb{Q}_a}[\ell(V,X)] = \inf_{\gamma>\max\{0,\upsilon\}} \gamma\varepsilon_a + \frac{\gamma\upsilon}{\gamma-\upsilon}\big\langle I_d - VV^\top, \hat{M}_a\big\rangle.$$

The dual optimal solution $\gamma^\star$ is given by

$$\gamma^\star = \begin{cases} \upsilon\left(1+\sqrt{\frac{\langle I_d - VV^\top, \hat{M}_a\rangle}{\varepsilon_a}}\right) & \text{if } \upsilon \geq 0, \\ \upsilon\left(1-\sqrt{\frac{\langle I_d - VV^\top, \hat{M}_a\rangle}{\varepsilon_a}}\right) & \text{if } \upsilon < 0 \text{ and } \big\langle I_d - VV^\top, \hat{M}_a\big\rangle \geq \varepsilon_a, \\ 0 & \text{if } \upsilon < 0 \text{ and } \big\langle I_d - VV^\top, \hat{M}_a\big\rangle < \varepsilon_a. \end{cases}$$

Note that $\gamma^\star \geq \max\{0,\upsilon\}$ in all the cases. Therefore, we have

$$\begin{aligned} &\sup_{\mathbb{Q}_a:\mathbb{W}(\mathbb{Q}_a,\hat{\mathbb{P}}_a)\leq\varepsilon_a} \upsilon\mathbb{E}_{\mathbb{Q}_a}[\ell(V,X)] \\ &= \begin{cases} \upsilon\left(\sqrt{\varepsilon_a}+\sqrt{\langle I_d - VV^\top, \hat{M}_a\rangle}\right)^2 & \text{if } \upsilon \geq 0, \\ \upsilon\left(\sqrt{\varepsilon_a}-\sqrt{\langle I_d - VV^\top, \hat{M}_a\rangle}\right)^2 & \text{if } \upsilon < 0 \text{ and } \big\langle I_d - VV^\top, \hat{M}_a\big\rangle \geq \varepsilon_a, \\ 0 & \text{if } \upsilon < 0 \text{ and } \big\langle I_d - VV^\top, \hat{M}_a\big\rangle < \varepsilon_a. \end{cases} \end{aligned}$$

This completes the proof. □

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2. By expanding the absolute value, problem (4) is equivalent to

$$\min_{V\in\mathbb{R}^{d\times k}, V^\top V = I_k} \max\{J_0(V), J_1(V)\},$$

where for each $(a,a') \in \{(0,1),(1,0)\}$, we can re-express J_a as

$$J_a(V) = \sup_{\mathbb{Q}_a:\mathbb{W}(\mathbb{Q}_a,\hat{\mathbb{P}}_a)\leq\varepsilon_a} (\hat{p}_a+\lambda)\mathbb{E}_{\mathbb{Q}_a}[\ell(V,X)] + \sup_{\mathbb{Q}_{a'}:\mathbb{W}(\mathbb{Q}_{a'},\hat{\mathbb{P}}_{a'})\leq\varepsilon_{a'}} (\hat{p}_{a'}-\lambda)\mathbb{E}_{\mathbb{Q}_{a'}}[\ell(V,X)]$$

Using Proposition 3.3 to reformulate the two individual supremum problems, we have

$$\begin{aligned} J_a(V) &= (\hat{p}_a+\lambda)\varepsilon_a + 2|\hat{p}_a+\lambda|\sqrt{\varepsilon_a\big\langle I_d - VV^\top, \hat{M}_a\big\rangle} + (\hat{p}_a+\lambda)\big\langle I_d - VV^\top, \hat{M}_a\big\rangle \\ &\quad + (\hat{p}_{a'}-\lambda)\varepsilon_{a'} + 2|\hat{p}_{a'}-\lambda|\sqrt{\varepsilon_{a'}\big\langle I_d - VV^\top, \hat{M}_{a'}\big\rangle} + (\hat{p}_{a'}-\lambda)\big\langle I_d - VV^\top, \hat{M}_{a'}\big\rangle. \end{aligned}$$

By defining the necessary parameters κ, θ, ϑ and C as in the statement of the theorem, we arrive at the postulated result. □

A.3 Proofs of Section 4

Proof of Lemma 4.1. Let $a_U \in \arg\max_{a\in\{0,1\}} F_a(U)$ and $a'_U = 1 - a_U$. Then, an Euclidean subgradient of F is given by

$$\nabla F(U) = \frac{\theta_{a_U}}{\sqrt{\langle UU^\top, \hat{M}_{a_U}\rangle}}\hat{M}_{a_U}U + \frac{\vartheta_{a'_U}}{\sqrt{\langle UU^\top, \hat{M}_{a'_U}\rangle}}\hat{M}_{a'_U}U + 2C_{a_U}U \in \mathbb{R}^{d\times(d-k)}.$$

The tangent space of the Stiefel manifold $\mathcal{M}$ at U is given by

$$T_U\mathcal{M} = \{\Delta \in \mathbb{R}^{d\times(d-k)} : \Delta^\top U + U^\top\Delta = 0\},$$

whose orthogonal projection (Absil et al., 2007, Example 3.6.2) can be computed explicitly via

$$\mathrm{Proj}_{T_U\mathcal{M}}(D) = (I_d - UU^\top)D + \frac{1}{2}U(U^\top D - D^\top U), \quad D \in \mathbb{R}^{d\times(d-k)}.$$

Therefore, a Riemannian subgradient of F at any point $U \in \mathcal{M}$ is given by

$$\begin{aligned}\mathrm{grad}F(U) &= \mathrm{Proj}_{T_U\mathcal{M}}(\nabla F(U)) \\ &= (I_d - UU^\top)\left(\frac{\theta_{a_U}}{\sqrt{\langle UU^\top, \hat{M}_{a_U}\rangle}}\hat{M}_{a_U}U + \frac{\vartheta_{a'_U}}{\sqrt{\langle UU^\top, \hat{M}_{a'_U}\rangle}}\hat{M}_{a'_U}U + 2C_{a_U}U\right).\end{aligned}$$

In the last line, we have used the fact that, if $D = SU$ for some symmetric matrix S, then

$$U^\top D - D^\top U = U^\top SU - U^\top S^\top U = 0.$$

This completes the proof. □

The proof of Lemma 4.2 relies on the following preliminary result.

Lemma A.1. Let $M \in \mathbb{R}^{(d-k)\times(d-k)}$ be a positive definite matrix. Then,

$$\left|\langle UU^\top, M\rangle - \langle U'U'^\top, M\rangle\right| \leq 2\sqrt{d-k}\sigma_{\max}(M)\|U - U'\|_F \quad \forall U, U' \in \mathcal{M}, \tag{10}$$

and

$$\left|\sqrt{\langle UU^\top, M\rangle} - \sqrt{\langle U'U'^\top, M\rangle}\right| \leq \frac{\sigma_{\max}(M)}{\sqrt{\sigma_{\min}(M)}}\|U - U'\|_F \quad \forall U, U' \in \mathcal{M}, \tag{11}$$

where $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ denote the maximum and minimum eigenvalues of the matrix M.

Proof of Lemma A.1. For inequality (10),

$$\begin{aligned}\left|\langle UU^\top, M\rangle - \langle U'U'^\top, M\rangle\right| &\leq \left|\langle UU^\top, M\rangle - \langle UU'^\top, M\rangle\right| + \left|\langle UU'^\top, M\rangle - \langle U'U'^\top, M\rangle\right| \\ &\leq \left|\langle U, M(U-U')\rangle\right| + \left|\langle U', M(U-U')\rangle\right| \\ &\leq \|U\|_F\|M(U-U')\|_F + \|U'\|_F\|M(U-U')\|_F \\ &= 2\sqrt{d-k}\sigma_{\max}\|U-U'\|_F.\end{aligned}$$

For inequality (11), we first note that the function $x \mapsto \sqrt{x}$ is $1/(2\sqrt{x_{\min}})$-Lipschitz on $[x_{\min}, +\infty)$ and that

$$\langle UU^\top, M\rangle \geq (d-k)\sigma_{\min}(M) \quad \forall U \in \mathcal{M}.$$

Therefore,

$$\begin{aligned}\left|\sqrt{\langle UU^\top, M\rangle} - \sqrt{\langle U'U'^\top, M\rangle}\right| &\leq \frac{1}{2\sqrt{(d-k)\sigma_{\min}(M)}}\left|\langle UU^\top, M\rangle - \langle U'U'^\top, M\rangle\right| \\ &\leq \frac{\sigma_{\max}(M)}{\sqrt{\sigma_{\min}(M)}}\|U-U'\|_F,\end{aligned}$$

where the last inequality follows from (10). This completes the proof. □

We are now ready to prove Lemma 4.2.

Proof of Lemma 4.2. Let $U, U' \in \mathcal{M}$ be two arbitrary points. We have

$$\begin{aligned}&|F(U) - F(U')| \\ &= |\max\{F_0(U), F_1(U)\} - \max\{F_0(U'), F_1(U')\}| \\ &\leq \max_{a\in\{0,1\}}|F_a(U) - F_a(U')| \\ &\leq \max_{a\in\{0,1\}}\max\left\{\theta_a\frac{\sigma_{\max}(\hat{M}_a)}{\sqrt{\sigma_{\min}(\hat{M}_a)}}, \vartheta_{1-a}\frac{\sigma_{\max}(\hat{M}_{1-a})}{\sqrt{\sigma_{\min}(\hat{M}_{1-a})}}, 2\sqrt{d-k}\sigma_{\max}(C_a)\right\}\|U-U'\|_F,\end{aligned}$$

where the last inequality follows from the definition of F_a and Lemma A.1. This completes the proof. □

Proof of Theorem 4.3. The proof follows from the fact that F is convex on the Euclidean space $\mathbb{R}^{d\times(d-k)}$, Lemma 4.2 and Li et al. (2019, Theorem 2) (and the remarks following it). □

B INFORMATION ON DATASETS

Table 2: Number of observations N and dimensions d of various UIC datasets

	Default Credit	Biodeg	E. Coli	Energy	German Credit	Image	Letter
N	30000	1055	333	768	1000	660	20000
d	22	40	7	8	48	18	16
	Magic	Parkinsons	SkillCraft	Statlog	Steel	Taiwan Credit	Wine Quality
N	19020	5875	3337	3071	1941	29623	6497
d	10	20	17	36	24	22	11

C ADDITIONAL RESULTS

C.1 DETAIL PERFORMANCES

Table 3 shows the performances of four examined methods with two criteria $\triangle\mathcal{F}_{Lin}$ and ARE.. It is clear that our method achieves the best results over all 14 datasets w.r.t ABDiff., and 7 datasets on ARE., which is equal to the number of datasets `FairPCA` out-perform others.

Table 4 complements Table 1 from the main text, from which we can see that two versions of `CFPCA` out-perform others over all datasets w.r.t. $\triangle\mathcal{F}_{Lin}$, which is the criteria they optimize for.

Table 3: In-sample performance over two criteria

	RFPCA		FairPCA		CFPCA-Mean Con.		CFPCA - Both Con.	
Dataset	ABDiff.	ARE.	ABDiff.	ARE.	ABDiff.	ARE.	ABDiff.	ARE.
Default Credit	**0.9457**	9.9072	1.5821	**9.9049**	0.9949	10.5164	3.2827	21.4523
Biodeg	**9.4093**	**23.1555**	14.2587	23.8227	15.5545	26.6540	24.8706	39.8737
E. Coli	**0.5678**	**1.4804**	0.9191	2.0840	0.9539	2.8360	4.5225	5.2155
Energy	**0.0094**	0.2295	0.0153	**0.2273**	0.2658	2.7893	0.2136	7.8768
German Credit	**1.6265**	**40.1512**	2.9824	40.3393	2.6109	40.1860	2.8741	47.1006
Image	**0.1320**	**5.0924**	0.7941	9.0437	0.6910	13.4491	3.0118	18.0000
Letter	**0.1121**	**7.4088**	1.2560	7.4375	0.4572	8.7764	0.5301	15.2234
Magic	**1.7405**	3.8766	2.8679	**3.3500**	5.5405	4.1938	8.7963	8.9695
Parkinsons	**0.1238**	5.0471	0.6702	**4.8760**	3.9470	5.9379	17.8122	19.9788
SkillCraft	**0.4231**	8.1569	0.5576	**8.1096**	0.7156	9.7755	0.9334	15.8245
Statlog	**0.1972**	**3.0588**	0.3315	7.9980	0.3857	10.9358	13.0725	35.9214
Steel	**0.6943**	11.0396	1.8015	**10.7653**	2.8933	14.5680	1.9322	23.9906
Taiwan Credit	**1.1516**	10.5136	1.3362	**10.4478**	1.3158	12.5867	2.2720	21.4365
Wine Quality	**0.1125**	**4.1491**	0.1705	5.8999	1.1359	5.9117	2.5852	9.8959

Table 4: Out-of-sample performance over $\triangle\mathcal{F}_{Lin}$

	RFPCA	FairPCA	CFPCA-Mean Con.	CFPCA - Both Con.
Default Credit	0.1596	0.2236	0.0574	**0.0413**
Biodeg	0.4892	0.4759	0.2014	**0.1371**
E. Coli	0.8556	0.7444	0.4455	**0.2532**
Energy	0.0580	0.0554	**0.0502**	0.0736
German Credit	0.1997	0.1737	0.1408	**0.1093**
Image	0.9996	0.9498	**0.1874**	0.2013
Letter	0.0954	0.0942	0.0556	**0.0455**
Magic	0.2195	0.2531	0.1561	**0.0882**
Parkinson's	0.1459	0.1061	0.1805	**0.0480**
SkillCraft	0.1126	0.1141	**0.0721**	0.0742
Statlog	0.9804	0.6309	0.1359	**0.0669**
Steel	0.2288	0.2240	0.1418	**0.0875**
Taiwan Credit	0.0604	0.0535	0.0391	**0.0370**
Wine Quality	0.9699	0.4639	0.2192	**0.0817**

C.2 VISUALIZATION

C.2.1 PARETO CURVES

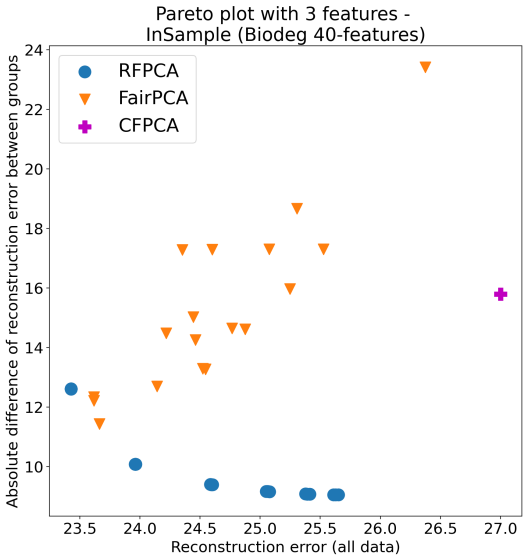

Figure 4: Pareto curves on Biodeg dataset (all data) with 3 principal components

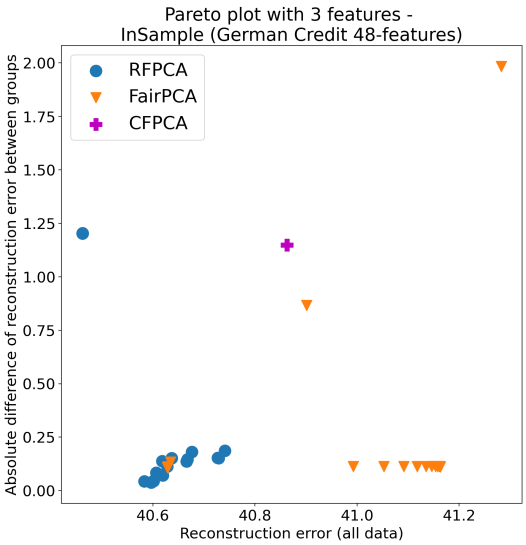

Figure 5: Pareto curves on German Credit dataset (all data) with 3 principal components

C.2.2 PERFORMANCE WITH DIFFERENT PRINCIPAL COMPONENTS

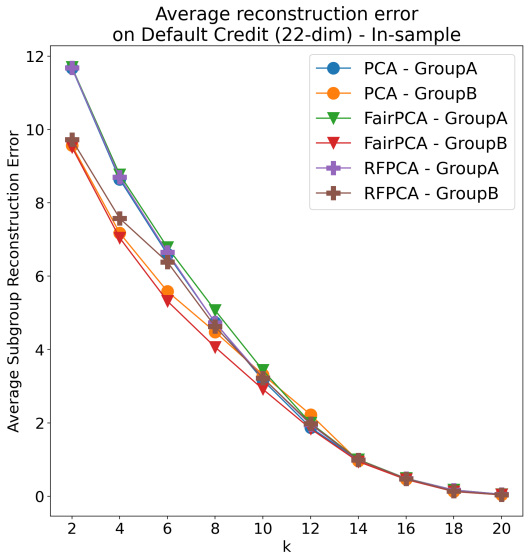

Figure 6: Subgroup average error with different k on Default Credit dataset

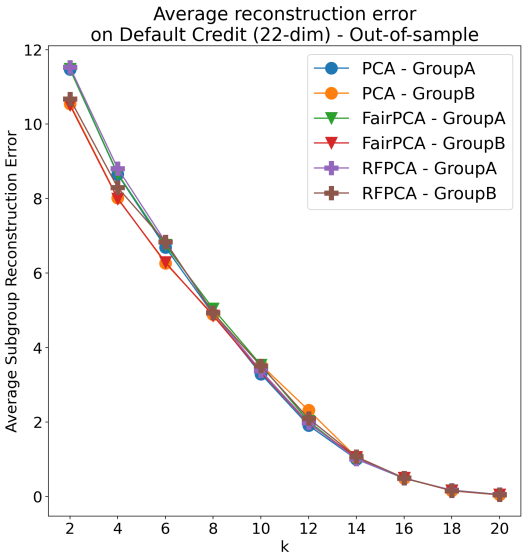

Figure 7: Subgroup average error with different k on Default Credit dataset

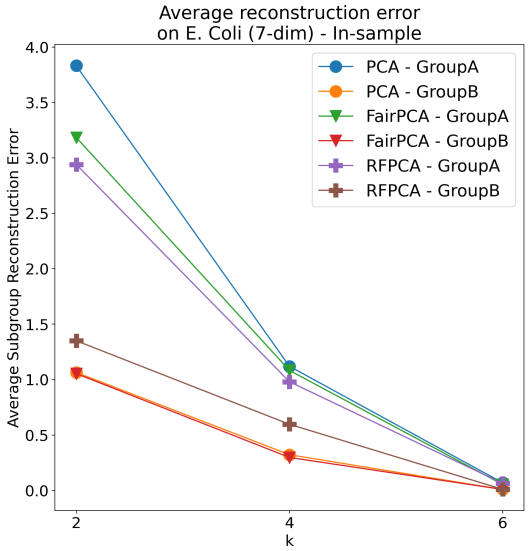

Figure 8: Subgroup average error with different k on E. Coli dataset

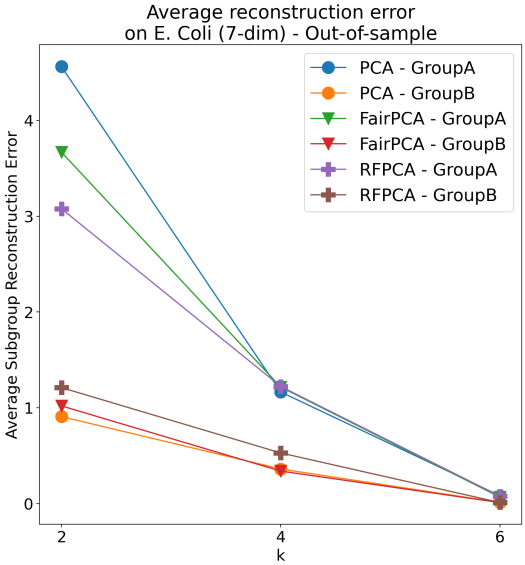

Figure 9: Subgroup average error with different k on E. Coli dataset

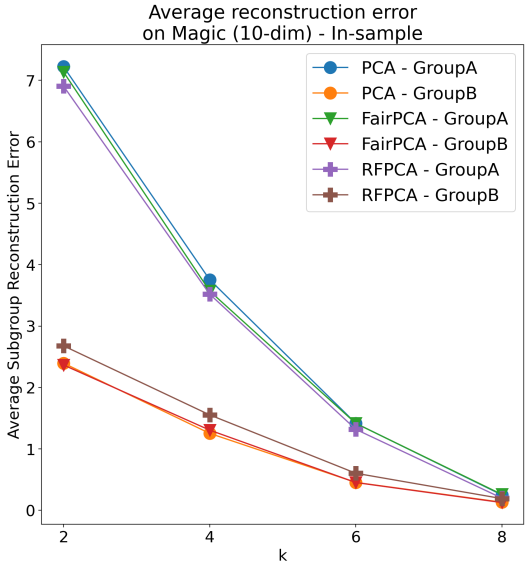

Figure 10: Subgroup average error with different k on Magic dataset

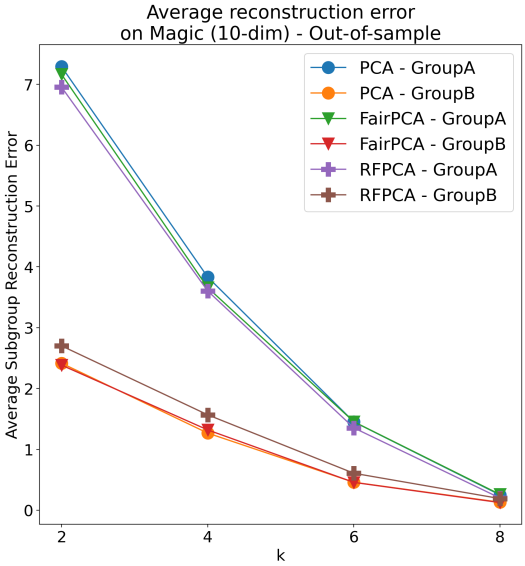

Figure 11: Subgroup average error with different k on Magic dataset

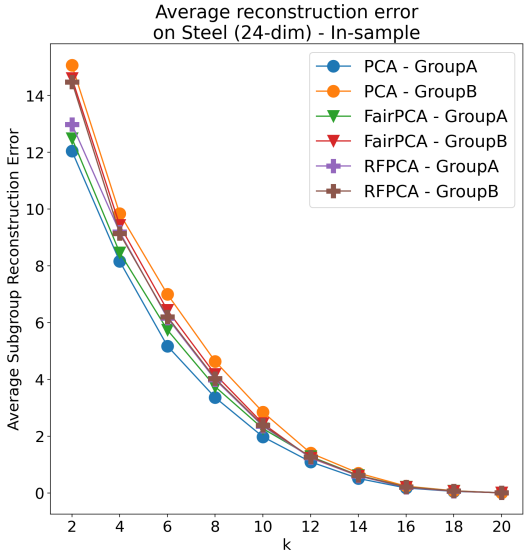

Figure 12: Subgroup average error with different k on Steel dataset

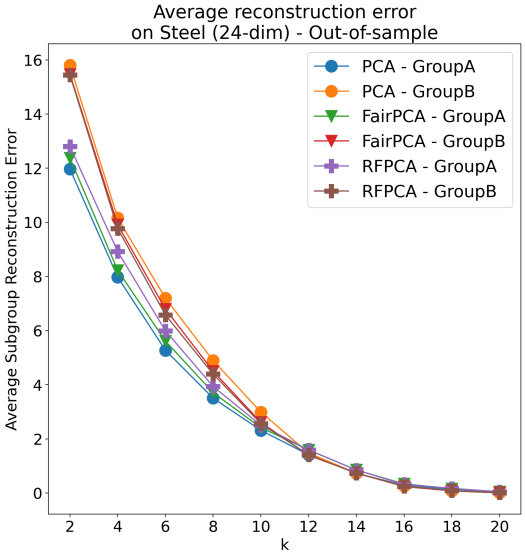

Figure 13: Subgroup average error with different k on Steel dataset

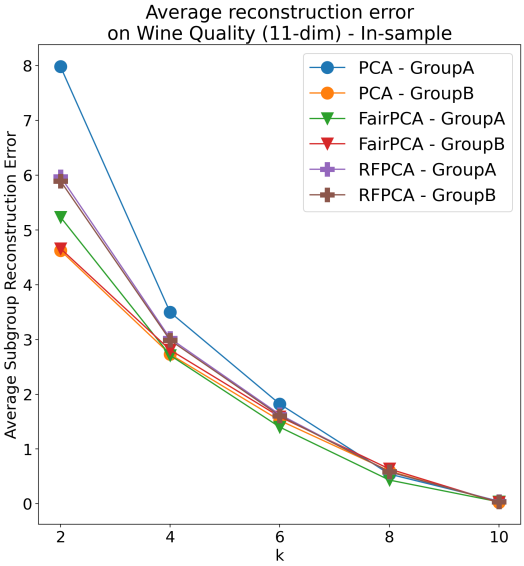

Figure 14: Subgroup average error with different k on Wine Quality dataset

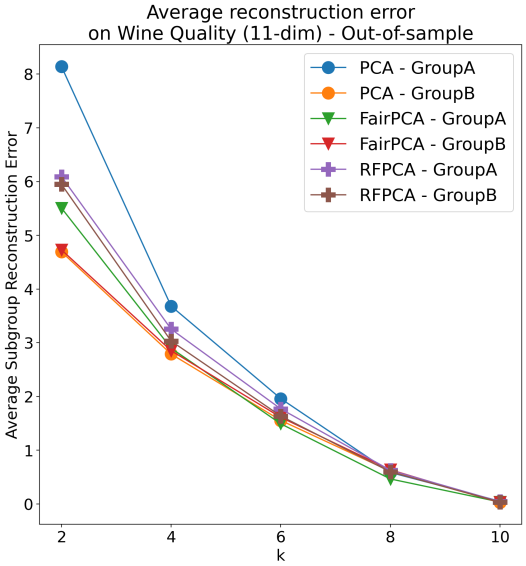

Figure 15: Subgroup average error with different k on Wine Quality dataset