1 Introduction

Despite the success of machine learning in informing policies and automating decision-making, there is growing concern about the fairness (with respect to protected classes like race or gender) of the resulting policies and decisions (Miller 2015; Rudin 2013; Angwin et al. 2016; Munoz, Smith, and Patil 2016). Hence, several groups have studied how to define fairness for supervised learning (Hardt, Price, and Srebro 2016; Calders, Kamiran, and Pechenizkiy 2009; Dwork et al. 2012; Zliobaite 2015) and developed supervised learners that maintain high prediction accuracy while reducing unfairness (Berk et al. 2017; Chouldechova 2017; Hardt, Price, and Srebro 2016; Zafar et al. 2017; Olfat and Aswani 2017).

However, fairness in the context of unsupervised learning has not been well-studied to date. One reason is that fairness is easier to define in the supervised setting, where positive predictions can often be mapped to positive decisions (e.g., an individual who is predicted to not default on a loan maps to the individual being offered a loan). Such notions of fairness cannot be used for unsupervised learning, which does not involve making predictions. A second reason is that it is not obvious why fairness is an issue of relevance to unsupervised learning, since predictions are not made.

1.1 Relevance of fairness to unsupervised learning

Fairness is important to unsupervised learning: First, unsupervised learning is often used to generate qualitative insights from data. Examples include visualizing high-dimensional data through dimensionality-reduction and clustering data to identify common trends or behaviors. If such qualitative insights are used to generate policies, then there is an opportunity to introduce unfairness in the resulting policies if the results of the unsupervised learning are unequal for different protected classes (e.g., race or gender). We present such an example in Section 6 using individual health data.

Second, unsupervised learning is often used as a preprocessing step for other learning methods. For instance, dimensionality reduction is sometimes performed prior to clustering, and hence fair dimensionality reduction could indirectly provide methods for fair clustering. Similarly, there are no fairness-enhancing versions of most supervised learners. Consequently, techniques for fair unsupervised learning could be combined with state-of-the-art supervised learners to develop new fair supervised learners. In fact, the past work most related to this paper concerns techniques that have been developed to generate fair data transformations that maintaining high prediction accuracy. For instance, the past work most related to this paper concerns techniques that have been developed to generate fair data transformations that maintaining high prediction accuracy for classifiers that make predictions using the transformed data (Dwork et al. 2012; Zemel et al. 2013; Feldman et al. 2015); however, these past works are most accurately classified as supervised learning because the data transformations are computed with respect to a label used for predictions.

We briefly review this work. Dwork et al. (2012) propose a linear program that maps individuals to probability distributions over possible classifications such that similar individuals are classified similarly. Zemel et al. (2013) and Calmon et al. (2017) generate an intermediate representation for fair clustering using a non-convex formulation that is difficult to solve. Feldman et al. (2015) propose an algorithm that scales data points such that the distributions of features, conditioned on the protected attribute, are matched; however, this approach makes the restrictive assumption that predictions are monotonic with respect to each dimension. Chierichetti et al. (2017) directly perform fair clustering by approximating an NP-hard preprocessing step; however, this approach only applies to specific clustering techniques whereas the approach we develop can be used with arbitrary
clustering techniques. Finally, a series of work has emerged using auto-encoders in the the context of deep classification. This area is promising, but suffers from a lack of theoretical guarantees and is further oriented almost entirely around an explicit classification task (Beutel et al. 2017; Zhang, Lemoine, and Mitchell 2018). In contrast, our method has applications in both supervised and unsupervised learning tasks, and well-defined convergence and optimality guarantees.

1.2 Outline and novel contributions

This paper studies fairness for principal component analysis (PCA), and we make three main contributions: First, in Section 3 we propose and motivate a novel quantitative definition of fairness for dimensionality reduction. Second, in Section 5 we develop convex optimization formulations for fair PCA and fair kernel PCA. Third, in Section 6 we demonstrate the efficacy of our semidefinite programming (SDP) formulations using several datasets, including using fair PCA as preprocessing to perform fair (with respect to age) clustering of health data that can impact health insurance rates.

2 Notation

Let \([n] = \{1, \ldots, n\}\), \(1(u)\) be the Heaviside function, and let \(e\) be the vector whose entries are all 1. A positive semidefinite matrix \(U\) with dimensions \(q \times q\) is denoted \(U \in \mathbb{S}_+^q\) (or \(U \succeq 0\) when dimensions are clear). We use the notation \((\cdot, \cdot)\) to denote the inner product and \(\|\cdot\|\) to denote the norm.

Our data consists of 2-tuples \((x_i, z_i)\) for \(i = 1, \ldots, n\), where \(x_i \in \mathbb{R}^p\) are a set of features, and \(z_i \in \{-1, 1\}\) label a protected class. For a matrix \(W\), the \(i\)-th row of \(W\) is denoted \(W_i\). Let \(X \in \mathbb{R}^{n \times p}\) and \(Z \in \mathbb{R}^n\) be the matrices so that \(X_i = (x_i - \bar{x})^T\) and \(Z_i = z_i\), where \(\bar{x} = \frac{1}{n} \sum x_i\). Also, we use the notation \(\Pi : \mathbb{R}^p \to [0, 1]\) to refer to a function that performs dimensionality reduction on the \(x_i\) data, where \(d\) is the dimension of the dimensionality-reduced data.

Let \(P = \{i : z_i = +1\}\) be the set of indices where the protected class is positive, and similarly let \(N = \{i : z_i = -1\}\) be the set of indices where the protected class is negative. We use \#\(P\) and \#\(N\) for the cardinality of these sets. Furthermore, we define \(X_+\) to be the matrix whose rows are \(x_i^T\) for \(i \in P\), and we similarly define \(X_-\) to be the matrix whose rows are \(x_i^T\) for \(i \in N\). Next, let \(\Sigma_+\) and \(\Sigma_-\) be the sample covariances matrices of \(X_+\) and \(X_-\), respectively.

For a kernel function \(k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+\), let \(K(X, X') = [k(x_i, x_j')]_{ij}\) be the transformed Gram matrix. Since the kernel trick involves replacing \(x_i^T x_j\) with \(K(x_i, x_j)\), the benefit of the above notation is it allows us to replace \(X(X')^T\) with \(K(X, X')\) as part of applying the kernel trick.

3 Fairness for dimensionality reduction

Definitions of fairness for supervised learning (Hardt, Price, and Srebro 2016; Dwork et al. 2012; Calders, Kamiran, and Pechenizkiy 2009; Zliobaite 2015; Feldman et al. 2015; Chouldechova 2017; Berk et al. 2017) specify that predictions conditioned on the protected class are roughly equivalent. However, these fairness notions cannot be used for dimensionality reduction because predictions are not made in unsupervised learning. This section discusses fairness for dimensionality reduction. We first provide and motivate a general quantitative definition of fairness, and then present several important cases of this definition.

3.1 General definition

Consider a fixed classifier \(h(u, t) : \mathbb{R}^d \times \mathbb{R} \to \{-1, +1\}\) that inputs features \(u \in \mathbb{R}^d\) and a threshold \(t\), and predicts the protected class \(z \in \{-1, +1\}\). We say that a dimensionality reduction \(\Pi : \mathbb{R}^p \to \mathbb{R}^d\) is \((\Delta(h))-\text{fair}\) if

\[
\left| \mathbb{P} \left[ h(\Pi(x), t) = +1 \mid z = +1 \right] - \mathbb{P} \left[ h(\Pi(x), t) = +1 \mid z = -1 \right] \right| \leq \Delta(h), \forall t \in \mathbb{R}.
\]

Moreover, let \(\mathcal{F}\) be a family of classifiers. Then we say that a dimensionality reduction \(\Pi : \mathbb{R}^p \to \mathbb{R}^d\) is \((\Delta(\mathcal{F}))-\text{fair}\) if it is \((\Delta(h))-\text{fair}\) for all classifiers \(h \in \mathcal{F}\).

Our fairness definition can be interpreted via classification: Observe that the first term in the left-hand-side of (1) is the true positive rate of the classifier \(h\) in predicting the protected class using the dimensionality-reduced variable \(\Pi(x)\) at threshold \(t\), and the second term is the corresponding false positive rate. Thus, \((\Delta(h))\) in our definition (1) can be interpreted as bounding the accuracy of the classifier \(h\) in predicting the protected class using the dimensionality-reduced variable \(\Pi(x)\).

Note that eq. (1) is analogous to disparate impact for classifiers (Calders, Kamiran, and Pechenizkiy 2009; Feldman et al. 2015), where we require that treatment not vary at all between protected classes. This has often been criticized as too strict of a notion in classification, and so alternate notions of fairness have been developed, such as equalized odds and equalized opportunity (Hardt, Price, and Srebro 2016). Instead of equalizing all treatment across protected classes, these notions instead focus on equalizing error rates; for example, in the case of lending, equalized odds would require nondiscrimination among all applicants of similar FICO scores, whereas disparate impact would require nondiscrimination among all applicants. This may be preferred in cases where \(y\) and \(z\) are strongly correlated. In any case, it can easily be incorporated into our model by simply further conditioning the two terms on the left-hand-side of eq. (1) on the main label, \(y\).

3.2 Motivation

The above is a meaningful definition of fairness for dimensionality reduction because it implies that a supervised learner using fair dimensionality-reduced data will itself be fair. This is formalized below:

Proposition 1. Suppose we have a family of classifiers \(\mathcal{F}\) and a dimensionality reduction \(\Pi\) that is \((\Delta(\mathcal{F}))-\text{fair}\). Then any classifier that is selected from \(\mathcal{F}\) to predict a label \(y \in \{-1, +1\}\) using \(\Pi(x)\) as features will have disparate impact less than \((\Delta(\mathcal{F}))\).

Proposition 1 follows directly from our definition of fairness. We anticipate that in most situations the goal of the
dimensionality reduction would not be to explicitly predict the protected class. Thus, our approach of bounding intentional discrimination on $z$ represents a conservative bound on any discrimination that may incidentally arise when performing classification using the family $F$ or when deriving qualitative insights form the results of unsupervised learning.

3.3 Special cases

An important special case of our definition occurs for the family $F_c = \{h(u,t) = 1(u \leq w + t) : w \in \mathbb{R}^d\}$, where the inequality in this expression should be interpreted element-wise. In this case, our definition can be rewritten as $\sup_u |F_{\Pi(x)}|_{z=+1}(u) - F_{\Pi(x)}|_{z=-1}(u)| \leq \Delta(F_c)$, where $F$ is the cumulative distribution function (c.d.f.) of the random variable in the subscript. Restated, for this family our definition is equivalent to saying $\Delta(F)$ is a bound on the Kolmogorov distance between $\Pi(x)$ conditioned on $z = \pm 1$ (i.e., the left-hand side of the above equation).

Other important cases are the family of linear support vector machines (SVM's) $F_k = \{h(u,t) = 1(w^T u - t \leq 0) : w \in \mathbb{R}^d\}$ and the family of kernel SVM's $F_k$ for a fixed kernel $k$. These important cases are used in Section 5 to propose formulations for fair PCA and fair kernel PCA.

Next, we briefly discuss empirical estimation of $\Delta(F)$. An empirical estimate of $\Delta(h)$ is given by $\Delta(h) = \sup_1 \frac{1}{n} \sum_{i \in p} 1(h(\Pi(x), t) = +1) - \sup_1 \frac{1}{n} \sum_{i \in n} 1(h(\Pi(x), t) = +1)$. Similarly, we define $\hat{\Delta}(F) = \sup_1 \hat{\Delta}(h)$. Last, note that we can provide high probability bounds of the actual fairness level in terms of these empirical estimates:

Proposition 2. Consider a fixed family of classifiers $F$. If the samples $(x_i, z_i)$ are i.i.d., then for any $\delta > 0$ we have with probability at least $1 - \exp(-n\delta^2/2)$ that $\Delta(F) \leq \hat{\Delta}(F) + 8\sqrt{\mathbb{V}(F)/n} + \delta$, where $\mathbb{V}(F)$ is the VC dimension of the family $F$.

This result follows from the triangle inequality, bounding $\Delta(F)$ with $\hat{\Delta}(F)$ plus a generalization error, for which there are standard bounds via Dudley's entropy integral (Wainwright 2017).

Remark 1. Recall that $\mathbb{V}(F_c) = d + 1$ (Shorack and Wellner 2009), and that $\mathbb{V}(F_v) = d + 1$ (Wainwright 2017). This means $\hat{\Delta}(F_c)$ and $\hat{\Delta}(F_v)$ will be accurate when $n$ is large relative to $d$.

4 Projection defined by PCA

Our approach to designing an algorithm for fair PCA will begin by first studying the convex relaxation of a non-convex optimization problem whose solution provides the projection defined by PCA. First, note that computation of the first $d$ PCA components $v_i$ for $i = 1, \ldots, d$ can be written as the following non-convex optimization problem:

$$\max \{|X^T X v_i| : \|v_i\|_2 \leq 1, v_i^T v_j = 0, \forall i \neq j\}.$$  

Now suppose we define the matrix $P = \sum_{i=1}^d v_i v_i^T$, and note $\sum_{i=1}^d v_i^T X^T X v_i = \sum_{i=1}^d \langle X^T X, v_i v_i^T \rangle = \langle X^T X, P \rangle$. Thus, we can rewrite the above optimization problem as

$$\max \{|X^T X | \text{ rank}(P) \leq d, I \geq P \geq 0\}. \quad (2)$$

In the above problem, we should interpret the optimal $P^*$ to be the projection matrix that projects $x \in \mathbb{R}^p$ onto the $d$ PCA components (still in the original $p$-dimensional space). Next, we consider a convex relaxation of (2). Since $I - P \succeq 0$, the usual nuclear norm relaxation is equivalent to the trace (Recht, Fazel, and Parrilo 2010). So our convex relaxation is

$$\max \{|X^T X | \text{ trace}(P) \leq d, I \geq P \geq 0\}. \quad (3)$$

Note that this base model is the same as that used by (Arora, Cotter, and Srebro 2013). The following result shows that we can recover the first $d$ PCA components from any $P^*$ that solves (3).

Theorem 1. Let $P^*$ be an optimal solution of (3), and consider its diagonalization: $P^* = \sum_{i=1}^p \lambda_i^* v_i v_i^T$, where $v_i$ is an orthonormal basis, and (without loss of generality) the $\lambda_i^*$ are in non-increasing order. Then the positive semidefinite $P^{**} = \sum_{i=1}^d \lambda_i^* v_i v_i^T$ is an optimal solution to (2).

Proof. We consider two cases. First, if $\rank(P^*) \leq d$ then $\lambda_i^* \in \{0, 1\}$ or $v_i^T X^T v_i = 0$ for all $i$, since otherwise we could increase $\lambda_i^*$ if $v_i^T X^T v_i > 0$ (or vice versa) to improve the objective while maintaining feasibility. It follows that $\langle X^T X, P^* \rangle = \langle X^T X, P^{**} \rangle$. This means that $P^{**}$ is optimal for (3); since it is also feasible for (2), we are done. Second, if $\rank(P^*) > d$ then $0 < \lambda^*_d < 1$ since the $\lambda^*_d$ are ordered. Consider $\tilde{P} = \frac{P^* - c P^{**}}{1-c}$, $c = \min\{\lambda^*_d, 1 - \lambda^*_d\}$. Note that $\tilde{P}$ is feasible for (3), and that $P^*$ is a strict convex combination of $P^{**}$ and $\tilde{P}$. All points between $P$ and $P^{**}$ are feasible by convexity, and so the optimality of $P^*$ implies that $P^{**}$ and $\tilde{P}$ must also be optimal for (3) by linearity of the objective (i.e., at least one must have objective value no less than that of $P^*$, but if one had a strictly better objective value than the other, no strict convex combination of the two could be optimal). The result then follows from the optimality of $P^{**}$ for (3) and feasibility for (2).

We conclude this section with two useful results on the spectral norm $\| \cdot \|_2$ of a symmetric matrix.

Theorem 2. Let $Q$ be a symmetric matrix, and suppose $\varphi \geq \|Q\|_2$. Then $\|Q\|_2 = \max(\|Q + \varphi I\|_2, \|Q - \varphi I\|_2) - \varphi$.

Proof. First diagonalize $Q = \sum_{i=1}^p \lambda_i v_i v_i^T$, with orthonormal basis $v_i$ and (without loss of generality) $\lambda_i$ in non-increasing order. Then $Q + \varphi I = \sum_{i=1}^p (\lambda_i + \varphi) v_i v_i^T$, $\|Q + \varphi I\|_2 = \sum_{i=1}^p (\lambda_i + \varphi)$. But by construction $\lambda_1 + \varphi \geq 0$ and $-\lambda_i + \varphi \geq 0$ for all $i = 1, \ldots, p$. Thus $\|Q + \varphi I\|_2 = \lambda_1 + \varphi$ and $\|Q - \varphi I\|_2 = -\lambda_p + \varphi$. The result follows since $\|Q\|_2 = \max(\lambda_1, -\lambda_p)$.

Corollary 1. Let $Q$ be a symmetric matrix, and suppose $\varphi \geq \|Q\|_2$. If $V$ is such that $V^T V = I$, then $\|V^T Q V\|_2 = \max(\|V^T (Q + \varphi I) V\|_2, \|V^T (Q - \varphi I) V\|_2) - \varphi$.
Proof. First note that $V^T(Q + \varphi \mathbb{I})V = V^TQV + \varphi \mathbb{I}$ and that $V^T(-Q + \varphi \mathbb{I})V = -V^TQV + \varphi \mathbb{I}$. Since the spectral norm is submultiplicative, this means $\|V^TQV\|_2 \leq \|V^T\|_2 \|Q\|_2 \|V\|_2 \leq \|Q\|_2$. So $\varphi \geq \|V^TQV\|_2$, and the result follows by applying Theorem 2 to $V^TQV$.

Recall that using the Schur complement allows representation of $\|VRV^T\|_2$ as a positive semidefinite matrix constraint when $R$ is positive semidefinite (Boyd et al. 1994). So the above corollary is useful because we can represent $\|VQV\|_2$ using positive semidefinite matrix constraints since $(Q + \varphi \mathbb{I})$ and $(-Q + \varphi \mathbb{I})$ are positive semidefinite by construction.

5 Designing formulations for fair PCA

Consider the linear dimensionality reduction $\Pi(x) = V^T x$ for $V \in \mathbb{R}^{p \times d}$ such that $V^TV = \mathbb{I}$. Then for linear classifier $h(u,t) = 1(w^T u - t \leq 0)$, definition (1) simplifies to $\Delta(h) = \sup_x |\mathbb{P}[w^T V^T x \leq t | z = \pm 1] - \mathbb{P}[w^T V^T x \leq t | z = \mp 1]|$. But the right-hand side is the Kolmogorov distance between $w^T V^T x$ conditioned on $z = \pm 1$, which is upper bounded (as can be seen trivially from its definition) by the total variation distance. Consequently, applying Pinsker’s inequality (Massart 2007) gives $\Delta(h) \leq \sqrt{\frac{1}{2} \mathbb{KL}(w^T V^T X_+ \parallel w^T V^T X_-)}$, where $\mathbb{KL}(\cdot)\parallel\cdot$ is the Kullback-Leibler divergence, $X_+$ is the random variable $[x | z = \pm 1]$, and $X_-$ is the random variable $[x | z = \mp 1]$. For the special case $X_+ \sim \mathcal{N}(\mu_+, \Sigma_+)$ and $X_- \sim \mathcal{N}(\mu_-, \Sigma_-)$, we have (Kullback 1997):

$$\Delta(h) \leq \sqrt{\frac{1}{4} \mathbb{KL}(w^T V^T X_+ \parallel w^T V^T X_-) \leq \frac{\|w^T V^T \Sigma_+ V w\|_2 + \|w^T V^T \Sigma_- V w\|_2}{\|w^T V^T \mu_+\|_2 + \|w^T V^T \mu_-\|_2}}$$

where $s_+ = w^T V^T \Sigma_+ V w$, $s_- = w^T V^T \Sigma_- V w$, $m_+ = w^T V^T \mu_+$, and $m_- = w^T V^T \mu_-$. The key observation here is that (4) is minimized when $s_+ = s_-$ and $m_+ = m_-$, and we will use this insight to propose constraints for FPCA. If $X_+$ and $X_-$ are not Gaussian, the three-point property may be used to obtain a similar bound with a couple extra terms involving the divergence between $X_+$ and a normal distribution with the same mean and variance (and the analog for $X_-$).

We first design constraints for the non-convex formulation (2) so that $\hat{\mu}_+ - \hat{\mu}_- = w^T V^T f$ has small magnitude, where $f = \hat{\mu}_+ - \hat{\mu}_- = \frac{1}{\#N} \sum_{i \in N} x_i - \frac{1}{\#N} \sum_{i \in N} x_i$. Note we make the identification $P = V^T V$ because of the properties of $P$ in (2) and since $V^TV = \mathbb{I}$. Observe that $w^T V^T f$ is small if $V^T f$ is small, which can be formulated as

$$\|V^T f\|^2 = \langle VV^T, f f^T \rangle = \langle P, f f^T \rangle \leq \delta^2,$$ (5)

where $\| \cdot \|$ is the $\ell_2$-norm, and $\delta$ is a bound on the norm. This (5) is a linear constraint on $P$.

We next design constraints for the non-convex formulation (2) so that $\hat{s}_+ - \hat{s}_- = w^T V^T (\hat{\Sigma}_+ - \hat{\Sigma}_-) V w$ has small magnitude. Recall the identification $P = V^T V$ because of the properties of $P$ in (2) and since $V^TV = \mathbb{I}$. Next observe that $w^T V^T (\hat{\Sigma}_+ - \hat{\Sigma}_-) V w$ is small if $V^T (\hat{\Sigma}_+ - \hat{\Sigma}_-) V$ is small. Let $Q = \hat{\Sigma}_+ - \hat{\Sigma}_-$, then using Corollary 1 gives

$$\mu + \varphi \geq \|V^T Q V\|_2 + \varphi$$

$$= \max\{\|V^T (Q + \varphi \mathbb{I}) V\|_2, \|V^T (-Q + \varphi \mathbb{I}) V\|_2\}$$

$$= \max\{\|VV^T (Q + \varphi \mathbb{I})\|_2, \|VV^T (-Q + \varphi \mathbb{I})\|_2\}$$

$$= \max\{\|P (Q + \varphi \mathbb{I})\|_2, \|P (-Q + \varphi \mathbb{I})\|_2\},$$ (6)

where $\varphi \geq \|\hat{\Sigma}_+ - \hat{\Sigma}_-\|_2$, and $\mu$ is a bound on the norm. Note (6) can be rewritten as SDP constraints using a standard reformulation for the spectral norm (Boyd et al. 1994). We design an SDP formulation for FPCA by combining the above elements. Though (2) with constraint (5) and (6) is a non-convex problem for FPCA, we showed in Theorem 1 that (3) was an exact relaxation of (2) after extracting the $d$ largest eigenvectors of the solution of (3). Thus, we propose...
We next consider a selection of datasets from UC Irvine’s online Machine Learning Repository (Lichman 2013). For each of the datasets, one attribute was selected as a protected class, and the remaining attributes were considered part of the feature space. After splitting each dataset into separate training (70%) and testing (30%) sets, the top five principal components were then found for the training sets of each of these datasets three times: once unconstrained, once with (7) with only the mean constraint, and once with (7) with both the mean and covariance constraints with $\delta = 0$ and $\mu = 0.01$. The test data was then projected onto these vectors. All data was normalized to have unit variance in each feature. For each instance, we estimated $\Delta(F)$ using the test set and for the families of linear SVM’s $F_0$ and Gaussian kernel SVM’s $F_\nu$. Finally, for each set of principal components $V$, the proportion of variance explained by the components was calculated as $\text{trace}(V\Sigma V^T)/\text{trace}(\Sigma)$, where $\Sigma$ is the centered sample covariance matrix of training set $X$.

6.2 Real data

We may observe that our additional constraints are largely helpful in ensuring fairness by all definitions. Furthermore,
Table 1: $\Delta$-fairness for both linear and Gaussian kernel SVM for PCA and FPCA. Best results for each fairness metric are bolded.

<table>
<thead>
<tr>
<th>DATA SET</th>
<th>UNCONSTRAINED</th>
<th>FPCA - MEAN CON.</th>
<th>FPCA - BOTH CON.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%VAR</td>
<td>LIN.</td>
<td>GAUS.</td>
</tr>
<tr>
<td>ADULT INCOME</td>
<td>11.41</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>BIODEG $^1$</td>
<td>31.16</td>
<td>0.2</td>
<td>0.35</td>
</tr>
<tr>
<td>E. COLL $^2$</td>
<td>65.01</td>
<td>0.65</td>
<td>0.80</td>
</tr>
<tr>
<td>ENERGY $^3$</td>
<td>84.08</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>GERMAN CREDIT</td>
<td>11.19</td>
<td>0.21</td>
<td>0.31</td>
</tr>
<tr>
<td>IMAGE</td>
<td>62.68</td>
<td>0.18</td>
<td>0.32</td>
</tr>
<tr>
<td>LETTER</td>
<td>42.33</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>MAGIC $^4$</td>
<td>61.91</td>
<td>0.32</td>
<td>0.33</td>
</tr>
<tr>
<td>PIMA $^5$</td>
<td>49.00</td>
<td>0.30</td>
<td>0.37</td>
</tr>
<tr>
<td>RECIDIVISM $^6$</td>
<td>56.28</td>
<td>0.24</td>
<td>0.26</td>
</tr>
<tr>
<td>SKILLCRAFT $^7$</td>
<td>40.62</td>
<td>0.15</td>
<td>0.19</td>
</tr>
<tr>
<td>STATLOG</td>
<td>87.80</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>STEEL</td>
<td>46.05</td>
<td>0.64</td>
<td>0.71</td>
</tr>
<tr>
<td>TAIW. CREDIT $^8$</td>
<td>45.52</td>
<td>0.11</td>
<td>0.17</td>
</tr>
<tr>
<td>WINE QUALITY $^9$</td>
<td>50.21</td>
<td>0.97</td>
<td>0.96</td>
</tr>
</tbody>
</table>

in many cases, this increase in fairness comes at minimal loss in the explanatory power of the principal components. There are a few datasets for which (7d) appear superfluous. In general, gains in fairness are stronger with respect to $F_\delta$, this is to be expected, as $F_\delta$ is a highly sophisticated set, and thus more robust to linear projections. Kernel FPCA may be a better approach to tackling this issue, but we leave this for future work. Additional experiments and a comparison to the method of Calmon et al. are shown in the appendix. We find that our method leads to more fairness on almost all datasets.

6.3 Hyperparameter sensitivity

Next, we consider the sensitivity of our results to hyperparameters $\delta, \mu$, for the Wine Quality dataset. The data was split into training (70%) and testing (30%) sets, and the top three fair principle components were found using (7) with only the mean constraint for each candidate $\delta$ and using (7) with both constraints for all combinations of candidate $\delta$ and $\mu$. All data was normalized to have unit variance in each independent feature. We calculated the percentage of the variance explained by the resulting principle components, and we estimated the fairness level $\Delta(F_\delta, \mu)$ for the family of linear SVM’s. This process was run 10 times for random data splits, and the averaged results are plotted in Figure 2. Here, the solid red line represents (7) with only the mean constraint. On the other hand, the dotted blue lines represent the (7) with both constraints, for the indicated $\mu$.

Adding the covariance constraints and further tightening $\mu$ generally improves fairness and decreases the proportion of variance explained. However, observe that the relative sensitivity of fairness to $\delta$ is higher than that of the variance explained, at least for this dataset. Similarly, increasing $\mu$ decreases the portion of variance explained while resulting in a less discriminatory dataset after the dimensionality reduction. We note that increasing $\mu$ past a certain point does not provide much benefit, and so smaller values of $\mu$ are to be preferred. We found that increasing $\mu$ past 0.1 did not substantively change results further, so the largest $\mu$ that we consider is 0.1. In general, hyperparameters may be set with cross-validation, although (4) may serve as guidance.

6.4 Fair clustering of health data

Health insurance companies are considering the use of patterns of physical activity as measured by activity trackers in order to adjust health insurance rates of specific individuals (Sallis, Bauman, and Pratt 1998; Paluch and Tuzovic 2017). In fact, a recent clustering analysis found that different patterns of physical activity are correlated with different health outcomes (Fukuoka et al. 2018). The objective of a health insurer in clustering activity data would be to find qualitative trends in an individual’s physical activity that help categorize the risks that that customer portends. That is, individuals within these activity clusters are likely to incur similar levels of medical costs, and so it would be beneficial to engineer easy-to-spot features that can help insurers bucket customers. However, health insurance rates must satisfy a number of legal fairness considerations with respect to gender, race, and age. This means that an insurance company may be found legally liable if the patterns used to adjust rates result in an unreasonably-negative impact on individuals of a specific gender, race, or age. Thus, an insurer may be interested in a feature engineering method to bucket customers while minimizing discrimination on protected attributes. Motivated by this, we use FPCA to perform a fair clustering of physical activity. Our goal is to find discernible qualitative trends in activity which are indicative of an individual’s activity patterns, and thus health risks, but fair with respect to age.

We use minute-level data from the the National Health and Nutrition Examination Survey (NHANES) from 2005–2006 (Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). 2018), on the intensity levels of the physical activity of about 6000 women, mea-

Figure 3: The mean physical activity intensities, plotted throughout a day, of the clusters generated after dimensionality reduction through PCA, FPCA with the mean constraint, and FPCA with both constraints. In each plot, each line represents the average activity level of the members of one cluster.

PCA is sometimes used as a preprocessing step prior to clustering in order to expedite runtime. In this spirit, we find the top five principal components through PCA, FPCA with mean constraint, and FPCA with both constraints, with $\delta = 0$ and $\mu = 0.1$ throughout. Then we conduct $k$-means clustering (with $k = 3$) on the dimensionality-reduced data for each case. Figure 3 displays the averaged physical activity patterns for each of the clusters in each of the cases. Furthermore, Table 2 documents the proportion of each cluster comprised of examinees over 40. We note that the clusters found under an unconstrained PCA are most distinguishable after 3:00 PM, so an insurer interested in profiling an individual’s risk would largely consider their activity in the evenings. However, we may observe in Table 2 that this approach results in notable age discrimination between buckets, opening the insurer to the risk of illegal price discrimination. On the hand, the second and third plots in Figure 3 and columns in Table 2 suggest that clustering customers based on their activity during the workday, between 8:00 AM and 5:00 PM, would be less prone to discrimination.

### 7 Conclusion

In this paper, we proposed a quantitative definition of fairness for dimensionality reduction, developed convex SDP formulations for fair PCA, and then demonstrated its effectiveness using several datasets. Many avenues remain for future research on fair unsupervised learning. For instance, we believe that our formulations in this paper may have suitable modifications that can be used to develop deflation and regression approaches for fair PCA analogous to those for sparse PCA (d’Aspremont et al. 2007; Zou, Hastie, and Tibshirani 2006).

### References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There’s software used across the country to predict future criminals, and it’s biased against blacks. ProPublica, May 23.


Boyd, S.; El Ghaoui, L.; Feron, E.; and Balakrishnan, V. 1994. Linear matrix inequalities in system and control theory, volume 15. SIAM.


Lichman, M. 2013. UCI machine learning repository.


