Fairness-aware Dimensionality Reduction

O. Deniz Kose
Electrical Engineering and Computer Science
University of California Irvine
Irvine, USA
okose@uci.edu

Yanning Shen
Electrical Engineering and Computer Science
University of California Irvine
Irvine, USA
yannings@uci.edu

***Abstract*—In this era of digitalization, the massive increase in available data leads to great potential for advancing various domains. However, the available data, such as images, videos, and speech signals, generally lie in high-dimensional space, which calls for efficient dimensionality reduction techniques to process them efficiently. Furthermore, while the fairness of algorithms is essential for their deployment in real-world systems, the effect of dimensionality reduction on fairness is an under-explored research area. Motivated by this, this paper puts forth a fairness-aware dimensionality reduction framework that is capable of properly compressing the data while mitigating bias. Specifically, our design targets at reducing the correlation between the compressed data and sensitive attributes, while projecting the data into a new coordinate system where most of its variation can be described. Experimental results on the CelebA dataset demonstrate that the proposed dimensionality reduction framework can improve group fairness measures for image classification while providing comparable utility to the conventional techniques.**

***Index Terms*—dimension reduction, fairness, PCA, image classification**

I. Introduction

During the last decade, the accelerated deployment of high technologies has led to the accumulation of large volumes of high-dimensional data. While processing and learning from such data can provide significant understanding and advancements for several systems, conventional data processing tools cannot cope with it due to its high dimension. This motivates the recently reignited attention on dimensionality reduction techniques. For a number of machine learning (ML) tasks, including classification, regression, and clustering, dimensionality reduction is a critical pre-processing step that allows learning from high-dimensional big data. Several dimensionality reduction strategies have been proposed so far in both ML and signal processing domains [1]–[6], among which principal component analysis (PCA) [2] serves as a seminal work. PCA projects data into a new coordinate system that preserves most of its variance.

Due to their success, learning algorithms have widespread use in our everyday lives to make life-changing decisions, which showcases the importance of preventing any discriminatory behaviour in these algorithms towards under-represented groups. Group fairness concerns the performance gap incurred by the learning algorithms with respect to certain sensitive/protected attributes (e.g., gender, ethnicity) [7], which is the main focus of this work. Furthermore, within the context of this paper, algorithmic bias corresponds to the stereotypical correlations the learning models encode and propagate with respect to the sensitive attributes.

It has been shown that ML models propagate the bias in training data, which may lead to discriminatory decisions in the ensuing applications [8]–[11]. As a result, fairness-aware ML has emerged as a growing line of research, where different fairness metrics are introduced for different learning tasks together with novel algorithmic tools that can help to satisfy these measures [12]–[18]. While fairness is a well-studied research area for supervised learning, it is still under-explored in the context of unsupervised learning [19], specifically for dimensionality reduction. Since dimensionality reduction is typically an essential pre-processing step for supervised learning algorithms, examining the fairness aspect of dimensionality reduction techniques can bring significant advancements in ensuring fairness for both unsupervised and supervised learning algorithms.

This work develops a fairness-aware adaptation of the PCA algorithm to mitigate bias while compressing data. While there have been several attempts for designing a fair PCA implementation [19]–[23], these works have certain limitations. First, [19], [20], [22], [23] all define fairness based on the reconstruction errors resulting from PCA for different sensitive groups, where the goal is to balance the reconstruction error-based measures for different sensitive groups. Such a fairness definition is limited in preventing the propagation of correlations with the sensitive attributes to the compressed data. To exemplify, making the PCA reconstruction error zero for each sensitive group would imply that the original correlations of data and sensitive attributes would also be reflected in the compressed form, which can lead to discrimination in the ensuing applications. Second, previous works are typically designed for a binary sensitive attribute [20], [21], [23], while sensitive attributes can have multiple values in many real-world applications (e.g., ethnicity). Finally, none of the aforementioned studies leads to a closed-form solution, which can affect the run-time efficiency of the algorithms. To overcome these limitations, we propose a fairness-aware PCA design which aims to lower the correlation between the compressed data and sensitive attributes, is applicable to non-binary sensitive attributes, and admits a closed-form solution.

Work in this paper was partially supported by Google Research Scholar Award and Hellman Fellowship Award (PI: Y. Shen).

Overall, our contributions can be summarized as follows:

- This work proposes a novel dimensionality reduction framework which is capable of mitigating bias while successfully compressing the data.
- A novel fairness measure is introduced for dimensionality reduction, which can eliminate the propagation of bias towards the compressed data.
- The proposed algorithm admits a closed-form solution and can be flexibly employed for non-binary sensitive attributes.
- Experimental results on a real-world dataset demonstrates the efficacy of the proposed algorithm in mitigating bias while protecting the utility for image classification.

II. Preliminaries and Problem Statement

Consider a dataset with N samples of dimension D written as $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, where the ith column $\mathbf{x}_i \in \mathbb{R}^D$ denotes ith data sample. Without loss of generality, the sample mean $\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$ is assumed to be removed from each sample $\mathbf{x}_i$. The trace and singular value decomposition (SVD) of a matrix $\mathbf{X}$ are denoted by $\operatorname{tr}(\mathbf{X})$ and $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, respectively. The main goal of this study is to find a set of d-dimensional vectors $\{\mathbf{y}_i\}_{i=1}^{N}$, with $d < D$, that preserve certain properties of $\{\mathbf{x}_i\}$ while reducing the correlations between $\{\mathbf{y}_i\}$ and sensitive attributes $\mathbf{s} \in \mathbb{R}^N$. Similar to the definition of data matrix $\mathbf{X}$, $\{\mathbf{y}_i\}$ vectors are the columns of the compressed data matrix $\mathbf{Y} := [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^{d\times N}$.

From the least-squares perspective, PCA searches for a linear subspace of dimension $d < D$ such that the reconstruction error incurred by the dimensionality reduction is minimized. Specifically, PCA deals with the following optimization problem:

$$\min_{\mathbf{U}_d, \{\mathbf{y}_i\}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{U}_d\mathbf{y}_i\|_2^2 \quad \text{s. to} \quad \mathbf{U}_d^\top\mathbf{U}_d = \mathbf{I}.$$

It is shown in [24] that $\mathbf{y}_i^* = (\mathbf{U}_d^*)^\top\mathbf{x}_i$ is the optimal solution for this problem, where $\mathbf{U}_d^*$ consists of the eigenvectors of $\mathbf{X}\mathbf{X}^\top = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top$ corresponding to the d largest eigenvalues. Herein, the reconstructed data vectors $\{\tilde{\mathbf{x}}_i\}$ based on the latent representations $\{\mathbf{y}_i\}$ can be recovered as $\tilde{\mathbf{x}}_i = \mathbf{U}_d\mathbf{y}_i$.

PCA can also be interpreted as a linear transformation that projects the original data into a new coordinate system where its variance is maximized. Therefore, PCA has the potential of amplifying algorithmic bias, if the data variability is mainly resulted from the different sensitive attribute values. In this case, the latent representations created by PCA will be highly correlated with the sensitive attributes, which may lead to discriminatory results in the ensuing tasks inputting these representations. Note that previous works have also empirically demonstrated that PCA can amplify algorithmic bias and result in disparate performance for different sensitive groups [23].

III. Fairness-aware Dimensionality Reduction

A. Fairness Metric for Dimensionality Reduction

Although the design of different fairness metrics has been thoroughly studied for supervised learning, this area is rather under-explored in the context of unsupervised learning. In the previous fairness-aware PCA formulations, fairness metrics are typically designed based on the reconstruction errors incurred for different sensitive groups [19], [20], [22], [23]. Specifically, a loss that depends on the reconstruction error is defined and this loss is aimed to be balanced across groups. However, if there are certain proxy features for the sensitive attributes within the data samples $\{\mathbf{x}_i\}$, such a fairness measure does not prevent the propagation of sensitive information to the compressed samples $\{\mathbf{y}_i\}$. For example, zip code is a proxy feature for the ethnicity under certain conditions. Thus, for the case where reconstruction error is low for every sensitive group (which is the main fairness target of previous fair PCA formulations), the zip code can be inferred from the latent representations $\{\mathbf{y}_i\}$, which can lead to discriminatory results with respect to ethnicity in the ensuing applications that input these latent representations.

It has been demonstrated that features which are correlated with the sensitive attribute lead to intrinsic bias, even when the sensitive attribute is not utilized in learning [25]. Therefore, the correlation between the inputs to a learning algorithm and sensitive attributes is a measure for the resulting bias. Motivated by this, the linear correlation between the sensitive attribute $\mathbf{s} \in \mathbb{R}^N$ and the rows of compressed data matrix $\mathbf{Y} \in \mathbb{R}^{d\times N}$ (each feature in the representations created by PCA) is considered as a bias measure in this study, since the latent representations in $\mathbf{Y}$ are generally utilized as inputs to learning algorithms. Specifically, inspired by the total correlation measure introduced in [26], $\|\mathbf{Y}\mathbf{s}\|_2^2$ is employed as the fairness metric in this study, which reflects the linear correlation between the sensitive attribute $\mathbf{s}$ and the rows of compressed data matrix $\mathbf{Y}$. Overall, this work aims to design a PCA formulation that aims to lower $\|\mathbf{Y}\mathbf{s}\|_2^2$ while compressing data.

B. Fairness-aware PCA Formulation

Based on the fairness metric introduced in Subsection III-A, the fairness-aware PCA formulation in this study solves the following problem:

$$\begin{aligned}&\min_{\mathbf{U}_d,\mathbf{Y}} \operatorname{tr}((\mathbf{X}-\mathbf{U}_d\mathbf{Y})^\top(\mathbf{X}-\mathbf{U}_d\mathbf{Y})) + \beta(\mathbf{Y}\mathbf{s})^\top(\mathbf{Y}\mathbf{s})\\ &\text{s. to} \quad \mathbf{U}_d^\top\mathbf{U}_d = \mathbf{I}.\end{aligned} \tag{1}$$

In this formulation, β is utilized as a hyperparameter that adjusts the focus on the fairness regularizer and provides a trade-off between the utility and fairness. Following [24], we first solve (1) for $\mathbf{Y}$, which leads to the following equivalent formulation:

$$\begin{aligned}&\min_{\mathbf{Y}} \mathcal{L}(\mathbf{Y}, \mathbf{U}_d), \text{ where}\\ &\mathcal{L}(\mathbf{Y}, \mathbf{U}_d) := -2\operatorname{tr}(\mathbf{X}^\top\mathbf{U}_d\mathbf{Y}) + \operatorname{tr}(\mathbf{Y}^\top\mathbf{U}_d^\top\mathbf{U}_d\mathbf{Y})\\ &+\beta(\mathbf{Y}\mathbf{s})^\top(\mathbf{Y}\mathbf{s}).\end{aligned} \tag{2}$$

The optimal solution for (2), $\mathbf{Y}^*$, can be obtained by solving $\frac{\nabla\mathcal{L}}{\nabla\mathbf{Y}} = \mathbf{0}$, which leads to $\mathbf{Y}^* = \mathbf{U}_d^\top\mathbf{X}(\mathbf{I}_N + \mathbf{s}\mathbf{s}^\top)^{-1}$. Here, $\mathbf{I}_N \in \mathbb{R}^{N\times N}$ denotes the identity matrix. Let $\mathbf{C} := (\mathbf{I}_N + \mathbf{s}\mathbf{s}^\top)^{-1}$.

Finding optimal $\mathbf{Y}^*$ leads to finding the orthonormal matrix $\mathbf{U}_d$ that satisfies the following optimization problem:

$$\begin{aligned}\min_{\mathbf{U}_d} \operatorname{tr}((\mathbf{X} - \mathbf{U}_d\mathbf{U}_d^\top\mathbf{X}\mathbf{C})^\top(\mathbf{X} - \mathbf{U}_d\mathbf{U}_d^\top\mathbf{X}\mathbf{C})) \\ + \beta(\mathbf{U}_d^\top\mathbf{X}\mathbf{C}\mathbf{s})^\top(\mathbf{U}_d^\top\mathbf{X}\mathbf{C}\mathbf{s}) \\ \text{s. to} \quad \mathbf{U}_d^\top\mathbf{U}_d = \mathbf{I}.\end{aligned} \tag{3}$$

which is equivalent to:

$$\begin{aligned}\min_{\mathbf{U}_d}(-2\operatorname{tr}(\mathbf{U}_d^\top\mathbf{X}\mathbf{C}\mathbf{X}^\top\mathbf{U}_d) + \operatorname{tr}(\mathbf{U}_d^\top\mathbf{X}\mathbf{C}\mathbf{C}^\top\mathbf{X}^\top\mathbf{U}_d) \\ + \beta\mathbf{s}^\top\mathbf{C}^\top\mathbf{X}^\top\mathbf{U}_d\mathbf{U}_d^\top\mathbf{X}\mathbf{C}\mathbf{s}) \\ \text{s. to} \quad \mathbf{U}_d^\top\mathbf{U}_d = \mathbf{I}.\end{aligned} \tag{4}$$

The problem in (4) can be reformulated as:

$$\begin{aligned}\min_{\mathbf{U}_d}(-2\operatorname{tr}(\mathbf{U}_d^\top\mathbf{X}\mathbf{C}\mathbf{X}^\top\mathbf{U}_d) + \operatorname{tr}(\mathbf{U}_d^\top\mathbf{X}\mathbf{C}\mathbf{C}^\top\mathbf{X}^\top\mathbf{U}_d^\top) \\ + \beta\operatorname{tr}(\mathbf{U}_d^\top\mathbf{X}\mathbf{C}\mathbf{s}\mathbf{s}^\top\mathbf{C}^\top\mathbf{X}^\top\mathbf{U}_d)) \\ \text{s. to} \quad \mathbf{U}_d^\top\mathbf{U}_d = \mathbf{I}.\end{aligned} \tag{5}$$

Finally, let $\mathbf{D} := -2\mathbf{X}\mathbf{C}\mathbf{X}^\top + \mathbf{X}\mathbf{C}\mathbf{C}^\top\mathbf{X}^\top + \beta\mathbf{X}\mathbf{C}\mathbf{s}\mathbf{s}^\top\mathbf{C}^\top\mathbf{X}^\top$, the optimization problem in (5) can be rewritten as

$$\begin{aligned}\min_{\mathbf{U}_d} \quad & \operatorname{tr}(\mathbf{U}_d^\top\mathbf{D}\mathbf{U}_d) \\ \text{s. to} \quad & \mathbf{U}_d^\top\mathbf{U}_d = \mathbf{I}.\end{aligned} \tag{6}$$

The problem in (6) is a typical truncated singular value decomposition formulation. Therefore, the columns of the optimal solution $\mathbf{U}_d^* \in \mathbb{R}^{D\times d}$ is composed of d eigenvectors of $\mathbf{D}$ associated with the d smallest eigenvalues, which provides a closed-form solution to the optimization problem formulated for fairness-aware PCA in (1).

Remark. It is important to note that while this work focuses on the conventional PCA that captures and protects the linear relationships within the data, the fairness metric introduced herein can also be utilized for a fairness-aware Kernel PCA [27] design. Such a formulation can be valuable when the data resides on highly nonlinear manifolds, where Kernel PCA can better capture nonlinear relationships within the data for a pre-selected kernel function compared to PCA.

```
Algorithm 1: Fair Dimensionality Reduction
  Data: $\mathbf{X}, \mathbf{s}$, $\beta$, $d$
  Result: $\mathbf{Y}^*$
  1. Calculate $\mathbf{D}$, where
     $\mathbf{D} := -2\mathbf{X}\mathbf{C}\mathbf{X}^\top + \mathbf{X}\mathbf{C}\mathbf{C}^\top\mathbf{X}^\top + \beta\mathbf{X}\mathbf{C}\mathbf{s}\mathbf{s}^\top\mathbf{C}^\top\mathbf{X}^\top$,
     and $\mathbf{C} := (\mathbf{I}_N + \mathbf{s}\mathbf{s}^\top)^{-1}$.
  2. Apply orthogonal eigen-decomposition to $\mathbf{D}$, i.e.
     $\mathbf{D} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$.
  3. For $\mathbf{V} := [\mathbf{v}_1, \dots, \mathbf{v}_D]$, sort the eigenvalues and
     corresponding eigenvectors such that $\mathbf{v_1}$ and $\mathbf{v_D}$
     correspond to the smallest and largest eigenvalues,
     respectively.
  4. Build $\mathbf{U}_d^* := [\mathbf{v}_1, \dots, \mathbf{v}_d]$.
  5. Calculate $\mathbf{Y}^* = (\mathbf{U}_d^*)^\top\mathbf{X}\mathbf{C}$.
```

IV. Experiments

This section evaluates the proposed fairness-aware dimensionality reduction framework for image classification on a real-world dataset.

A. Dataset

In the experiments, a real-world dataset, CelebFaces Attributes Dataset (CelebA) [28], is utilized for the ensuing task of image classification. CelebA is a large-scale dataset that includes approximately $200,000$ celebrity images together with 40 binary face attribute annotations. Background clutter and poses vary significantly across images leading to a large diversity within the dataset.

In the experiments, we sample this dataset and utilize a total number of 40519 images. A pre-processing operation is also applied to the images with the original dimensions of $218 \times 178 \times 3$ to resize them down to 28×28. Specifically, the images are cropped to remove parts that do not include the face, and then they are transformed to be grayscale. Afterward, the 28×28 matrix are flattened to a vector and concatenated with 38 of the binary face attributes to create data vectors $\{\mathbf{x}_i\}$. Labels for the image classification task are the attractiveness of the faces in the images (i.e., attractive or not attractive), while gender information is utilized as the sensitive attribute. Statistical information for the utilized dataset is presented in Table I.

Female	Male	Attractive	Not Attractive
23580	16939	20874	19645

TABLE I: CelebA statistics.

B. Performance metrics

For the utility metric of image classification, accuracy is utilized. For fairness, two quantitative measures of group fairness metrics are reported in terms of **statistical parity** (Δ_{SP}) **[29]** and **equal opportunity**(Δ_{EO}) **[30]**:
$\Delta_{SP} := |P(\hat{c} = 1 \mid s = 0) - P(\hat{c} = 1 \mid s = 1)|$
$\Delta_{EO} := |P(\hat{c} = 1 \mid c = 1, s = 0) - P(\hat{c} = 1 \mid c = 1, s = 1)|$
where c represents the ground truth class label ($c = 1$ if attractive, $c = 0$ otherwise), and $\hat{c}$ is the predicted class. Note that sensitive attributes $s = 1$ and $s = 0$ correspond to being male and female, respectively. Lower values for Δ_{SP} and Δ_{EO} indicate better fairness performance and are more desirable.

C. Experimental settings

In the experiments, the designed fairness-aware PCA is employed as a pre-processing operator on the data vectors $\{\mathbf{x}_i\}$, where the compressed data is then input to a three-layer multi-layer perceptron (MLP). In training, weights of the MLP are initialized with Glorot initialization [31] and ReLU activation is applied after the hidden layer. The utilized MLP model is trained for the log loss function for 1000 epochs by utilizing Adam optimizer [32], where the learning rate is chosen as 0.01. Note that training is early-stopped, if the loss does not improve for 10 epochs.

$d = 25$	Accuracy (%)	Δ_{SP} (%)	Δ_{EO} (%)
$D = d$	$\mathbf{77.98} \pm 0.60$	54.26 ± 1.47	43.82 ± 6.20
PCA	72.04 ± 0.71	43.96 ± 1.67	33.92 ± 2.38
FairPCA [20]	72.17 ± 0.57	40.84 ± 1.82	30.29 ± 3.98
FairDR	70.80 ± 0.47	$\mathbf{37.98} \pm 2.51$	$\mathbf{28.53} \pm 3.18$
$d = 50$	Accuracy (%)	Δ_{SP} (%)	Δ_{EO} (%)
$D = d$	$\mathbf{77.98} \pm 0.60$	54.26 ± 1.47	43.82 ± 6.20
PCA	73.02 ± 0.75	42.48 ± 1.16	30.75 ± 2.66
FairPCA [20]	72.17 ± 0.57	$\mathbf{40.84} \pm 1.82$	30.29 ± 3.98
FairDR	72.71 ± 0.19	$\mathbf{40.84} \pm 0.37$	$\mathbf{29.13} \pm 1.50$

TABLE II: Fairness-aware dimensionality reduction on CelebA.

$d = 25$	Accuracy (%)	Δ_{SP} (%)	Δ_{EO} (%)
PCA	$\mathbf{72.04} \pm 0.71$	43.96 ± 1.67	33.92 ± 2.38
$\beta = 0.1$	70.13 ± 0.32	38.25 ± 1.02	$\mathbf{28.06} \pm 2.27$
$\beta = 1$	70.94 ± 0.60	39.45 ± 2.19	30.14 ± 4.50
$\beta = 10$	70.80 ± 0.47	$\mathbf{37.98} \pm 2.51$	28.53 ± 3.18
$d = 50$	Accuracy (%)	Δ_{SP} (%)	Δ_{EO} (%)
PCA	$\mathbf{73.02} \pm 0.75$	42.48 ± 1.16	30.75 ± 2.66
$\beta = 0.1$	72.86 ± 0.37	42.03 ± 1.59	31.95 ± 2.49
$\beta = 1$	$\mathbf{73.01} \pm 0.30$	41.13 ± 1.43	29.88 ± 2.68
$\beta = 10$	72.71 ± 0.19	$\mathbf{40.84} \pm 0.37$	$\mathbf{29.13} \pm 1.50$

TABLE III: Sensitivity analysis for β.

The model is trained over 90% of the images, while the remaining images contribute to the test set. The hyperparameter β is selected via grid search among the values $\{0.1, 1, 10\}$. Furthermore, FairPCA [20] is employed as the fairness-aware dimensionality reduction baseline, where FairPCA aims to balance the reconstruction loss incurred by PCA for different sensitive groups. The min-max formulations in FairPCA are relaxed and solved as semidefinite programs. The hyperparameters η and T in FairPCA are tuned via grid searches among the values $\{1, 20\}$ and $\{5, 10\}$, respectively. Note that the candidate values for the hyperparameters η and T are the suggested values for them in the corresponding study. For all experiments, results are collected for five random data splits, and the average of them together with standard deviations are presented.

D. Experimental Results

Comparative results for image classification are presented in Table II for the cases where $d = 25$ and $d = 50$. The proposed fairness-aware dimensionality reduction scheme is denoted by FairDR in this table. For the proposed scheme, the natural baseline is to employ the conventional PCA algorithm. In addition to PCA, the results are also obtained for the case where no dimensionality reduction is employed ($D = d$ in Table II) and for the fairness-aware baseline FairPCA [20].

In Table II, all dimensionality reduction techniques are observed to lead better fairness measures compared to the case where no dimensionality reduction is employed. For the utilized dataset, inferral of the sensitive information may become more challenging based on the latent representations output by these dimensionality reduction algorithms compared to the original data vectors, which can justify this observation. The results in Table II demonstrate that the proposed fairness-aware dimensionality reduction framework herein consistently achieves better fairness measures compared to the fairness-aware baseline, along with similar utility values. Furthermore, it can be observed that the improvements provided by FairDR in terms of fairness increase, as d gets smaller, which makes FairDR more advantegous for the cases where compression rate needs to be high. Overall, the results corroborate the efficacy of the proposed fairness-aware dimensionality reduction design in mitigating bias while also providing similar utility measures compared to the state-of-the-art fairness-aware baseline.

In order to examine the influence of hyperparameter selection on the performance, the sensitivity analysis for the hyperparameter β is presented in Table III for d values 25 and 50. The results in Table III show that the fairness metrics generally improve as β value increases, which is expected, since β adjusts the focus on the fairness regularizer in (1). Overall, the results in Table III demonstrate that the proposed fairness-aware dimensionality reduction scheme typically leads to better fairness measures than the natural baseline, PCA, for a large range of β values.

V. Conclusion and Future Work

This study presents a fairness-aware dimensionality reduction technique that projects the data into a lower dimensional space where most of its variation can be preserved. For bias mitigation, the proposed scheme employs a fairness regularizer whose design is based on a novel fairness notion introduced in this work. Specifically, the presented fairness notion aims to reduce the correlation between the compressed data and sensitive attributes, which can alleviate the propagated bias from the original representations towards the compressed data. Differing from previous fairness-aware PCA formulations, the proposed approach can be directly employed with non-binary sensitive attributes and admits a closed-form solution that can improve the run-time complexity. Experimental results on a real-world dataset show the efficacy of the proposed dimensionality reduction framework in mitigating bias while providing similar utility to a state-of-the-art fairness-aware baseline for image classification.

This work opens up a number of possible future directions: (i) extension of the present framework and analysis to non-linear dimensionality reduction techniques; (ii) the consideration of the case where multiple sensitive attributes are available; (iii) exploration of novel fairness measures for dimensionality reduction based on non-linear correlations between $\mathbf{s}$ and $\mathbf{Y}$.

References

[1] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[2] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[3] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[4] B. Schölkopf, A. Smola, and K. R. Müller, "Kernel principal component analysis," in *International Conference on Artificial Neural Networks, (ICANN)*. Springer Verlag, 1997, pp. 583–588.

[5] J. Zabalza, J. Ren, and S. Marshall, "'on the fly'dimensionality reduction for hyperspectral image acquisition," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 749–753.

[6] C. Hellings, P. Gogler, and W. Utschick, "Composite real principal component analysis of complex signals," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 2216–2220.

[7] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 560–568.

[8] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. Innovations in Theoretical Computer Science (ITCS)*, January 2012, pp. 214–226.

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

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[11] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions," *Annual Review of Statistics and Its Application*, vol. 8, pp. 141–163, 2021.

[12] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[13] M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller, "From parity to preference-based notions of fairness in classification," *Advances in neural information processing systems*, vol. 30, 2017.

[14] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE international conference on data mining workshops*. IEEE, 2009, pp. 13–18.

[15] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data mining and knowledge discovery*, vol. 21, pp. 277–292, 2010.

[16] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 643–650.

[17] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*. PMLR, 2018, pp. 60–69.

[18] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[19] M. M. Kamani, F. Haddadpour, R. Forsati, and M. Mahdavi, "Efficient fair principal component analysis," *Machine Learning*, pp. 1–32, 2022.

[20] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala, "The price of fair pca: One extra dimension," *Advances in neural information processing systems*, vol. 31, 2018.

[21] M. Olfat and A. Aswani, "Convex formulations for fair principal component analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 663–670.

[22] G. Zalcberg and A. Wiesel, "Fair principal component analysis and filter design," *IEEE Transactions on Signal Processing*, vol. 69, pp. 4835–4842, 2021.

[23] H. Vu, T. Tran, M.-C. Yue, and V. A. Nguyen, "Distributionally robust fair principal components via geodesic descents," in *International Conference on Learning Representations*, 2022.

[24] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[25] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1445–1459, July 2013.

[26] O. D. Kose and Y. Shen, "Fair contrastive learning on graphs," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 8, pp. 475–488, 2022.

[27] T. Jin, J. Yu, J. You, K. Zeng, C. Li, and Z. Yu, "Low-rank matrix factorization with multiple hypergraph regularizer," *Pattern Recognition*, vol. 48, no. 3, pp. 1011–1022, 2015.

[28] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[29] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. Innovations in Theoretical Computer Science Conf.*, Jan. 2012, pp. 214–226.

[30] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Adv. in Neural Inf. Processing Systems (NeurIPS)*, vol. 29, pp. 3315–3323, Dec. 2016.

[31] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, May 2010, pp. 249–256.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.