Consistent Multigroup Low-rank Approximation

Anonymous Authors[1]

Abstract

We consider the problem of consistent low-rank approximation for multigroup data: we ask for a sequence of k basis vectors such that projecting the data onto their spanned subspace treats all groups as equally as possible, by minimizing the maximum error among the groups. Additionally, we require that the sequence of basis vectors satisfies the natural consistency property: when looking for the best k vectors, the first $d < k$ vectors are the best possible solution to the problem of finding d basis vectors. Thus, this multigroup low-rank approximation method naturally generalizes SVD and reduces to SVD for data with a single group. We give an iterative algorithm for this task that sequentially adds to the basis the vector that gives the best rank-1 projection according to the min-max criterion, and then projects the data onto the orthogonal complement of that vector. For finding the best rank-1 projection, we use primal-dual approaches or semidefinite programming. We analyze the theoretical properties of the algorithms and demonstrate empirically that the proposed methods compare favorably to existing methods for multigroup (or fair) PCA.

1. Introduction

Low-rank approximation techniques provide dimensionally reduced representations of data by expressing the data matrix as a linear combination of a small number of factors. Such methods are fundamental in machine learning and data science, due to the benefits they offer in terms of scalability, interpretability, and their strong mathematical foundation.

Among other methods, the singular value decomposition (SVD) holds a central position. A celebrated result states that the first d left or right singular vectors offer the best possible rank-d approximation to a matrix M in terms of Frobenius or spectral norm (Eckart & Young, 1936). We call this the *consistency* property of the SVD.

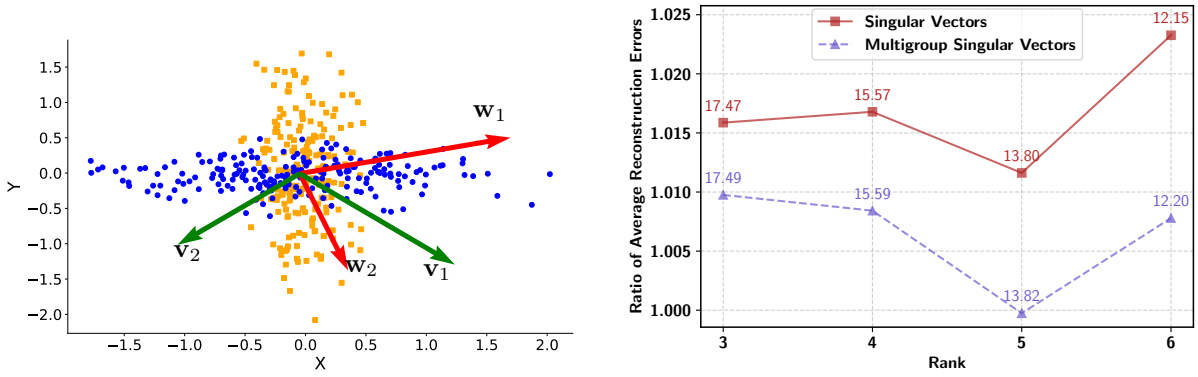
In many applications, the rows of a data matrix are divided into two or more groups according to a particular attribute, e.g., gender. In such a case, using the top k right singular vectors may not represent every group equally well, potentially resulting in inaccurate or even discriminatory outcomes. To address these concerns, previous works (Samadi et al., 2018; Tantipongpipat et al., 2019; Song et al., 2024) have studied the problem of finding a *common* projection onto a subspace that minimizes the worst-case reconstruction error of any group. This problem is typically referred to as FAIR-PCA.

While effective, previous methods (Samadi et al., 2018; Tantipongpipat et al., 2019; Song et al., 2024) do not ensure the consistency property of the SVD, i.e., given a basis of a subspace, it is not possible to readily obtain a basis of a lower-rank subspace simply by discarding some vectors.

Building on this line of work, we introduce a *multigroup* low-rank approximation formulation which, in the spirit of the SVD, imposes the consistency property. More specifically, given a data matrix $\mathbf{M}$ with rows divided into groups $\mathcal{G} = \{\mathbf{A}^1, \ldots, \mathbf{A}^k\}$, we look for an orthonormal basis $\mathbf{V}$ for a subspace of the column space with the following properties: 1) projecting onto it minimizes the maximum possible error of any group (min-max criterion), 2) is consistent: given the best r vectors, the first $d < r$ vectors from that solution are the best possible solution to the problem of finding d basis vectors. We call such vectors *multigroup singular vectors*.

Figure 1 illustrates the concept of multigroup singular vectors. Figure 1 (a) shows the standard singular vectors (in red) and multigroup singular vectors (in green) in synthetic data. While standard singular vectors clearly favor the larger group over the smaller, multigroup singular vectors seek a more balanced representation. Figure 1 (b) instead shows a comparison in the real-world compas dataset (Dua & Graff, 2017). We consider the partitioning of the data into females and males. Projecting onto the multigroup singular vectors leads to a more balanced reconstruction error than projecting onto standard singular values, while giving a similar overall reconstruction error.

We empirically evaluate our method in the task of FAIR-PCA (Samadi et al., 2018). We show that it ensures the consistency property while incurring similar reconstruction error to the previous methods (Samadi et al., 2018; Tantipongpipat et al., 2019; Song et al., 2024). In addition, an obvious advantage of the consistency property is that it confers high efficiency and scalability: we can compute



(a) **(b)**

Figure 1: Left (**a**): synthetic data partitioned in two groups, as indicated by the color of the points. Singular vectors $\{\mathbf{w}_1, \mathbf{w}_2\}$ and the multigroup singular vectors $\{\mathbf{v}_1, \mathbf{v}_2\}$ given by our method, are also shown. Right (**b**): real-world compas dataset partitioned in two groups, females and males. The y-axis indicates the ratio of the average group-wise reconstruction error incurred by standard singular vectors and the multigroup singular vectors. The x-axis indicates the number of basis vectors. We additionally report the average reconstruction error across all data instances (both males and females).

each basis vector efficiently, and once the full-rank basis is computed, we can obtain lower-dimensional representations of any rank by just discarding basis vectors, as for SVD. This is in contrast to the previous approaches (Samadi et al., 2018; Tantipongpipat et al., 2019; Song et al., 2024) which require solving an independent, computationally challenging problem for any basis dimension.

The multigroup low-rank approximation problem still presents significant challenges. To ensure the consistency property holds, we construct the k-dimensional solution through an iterative process of solving simpler rank-1 problems. The more difficult aspect is proving that this procedure also yields the optimal k-dimensional result. Notably, we demonstrate that our solution is in fact optimal in the case of two groups in the data. For scenarios with more than two groups, we show that our solutions are empirically close to optimal in practice.

The contributions of this work can be summarized as follows.

- We formalize the consistent multigroup low-rank approximation problem.
- We give an iterative procedure which selects the best basis vector according to the min-max criterion, and then projects the data onto the orthogonal complement of the previously chosen vectors. The selection of the best basis vector at each iteration represents the main algorithmic challenge that we tackle.
- We theoretically analyze the formulated problems and the proposed algorithms, focusing on the two-groups case, which exhibits interesting properties.
- We describe extensive experiments on real-world datasets to demonstrate the benefits of consistent low-rank approximation over previous work.

The rest of this paper is organized as follows. Section 2 gives an overview of related work. Section 3 gives necessary notations and definitions. Section 4 describes our overall framework. Section 5 formally introduces the multigroup singular vector problem (MG-SINGULARVECTOR), while Section 6 proposes algorithms to solve it. We present a theoretical analysis and give an algorithm for a special case in Section 7, while Section 8 contains our experimental evaluation and Section 9 presents conclusions.

2. Related Work

We assume that the reader is familiar with singular value decomposition (SVD) and principal component analysis (PCA) (see, e.g., (Van Der Maaten et al., 2009; Eckart & Young, 1936; Hotelling, 1933)).

Multigroup low-rank approximation: Fair PCA. Recently PCA has been extended to handle multigroup data. In this line of work, groups correspond to different values of a sensitive attribute (e.g., gender), and hence the proposed multigroup extension of PCA is referred to as FAIR PCA (Samadi et al., 2018; Tantipongpipat et al., 2019; Song et al., 2024). In FAIR PCA, the goal is to retrieve a low-dimensional representation of the data that maximizes and balances variance for the groups. A similar problem has also been studied by Zalcberg & Wiesel (2021) from a signal processing perspective and by Babu & Stoica (2023). Other works have instead explored a significantly different formulation of the FAIR PCA problem. For instance, some works rely on notions of fairness such as *demographic parity* or *equal opportunity* that are adapted from supervised

learning (Olfat & Aswani, 2019; Kleindessner et al., 2023). In a similar vein, Lee et al. (2022) define FAIR PCA as the problem of minimizing the maximum mean discrepancy between dimensionality-reduced conditional distributions of different classes. Instead, Pelegrina & Duarte (2023) as well as Kamani et al. (2022) formulate FAIR PCA as an optimization problem where the objective encodes the trade-off between reconstruction error and fairness.

Other multigroup dimensionality-reduction techniques. In recent years, significant attention has been devoted to algorithmic fairness and there have been efforts to extend traditional dimensionality-reduction techniques, beyond PCA. For instance, Matakos et al. (2024) and Song et al. (2024) study fair *column subset selection*, while Louizos et al. (2016) introduce the fair *variational autoencoder*.

3. Preliminaries

Notation. We denote matrices and vectors by bold uppercase and lowercase letters, respectively. The notation $\mathcal{V}_d$ indicates the set of all matrices with d orthonormal columns, i.e., $\mathcal{V}_d = \{\mathbf{V} \in \mathbb{R}^{n\times d} : \mathbf{V}^\top\mathbf{V} = \mathbf{I}_d\}$, where $\mathbf{I}_d$ is the $d \times d$ identity matrix.

For a matrix $\mathbf{V} \in \mathcal{V}_d$, we denote the ordered set of columns of $\mathbf{V}$ by $\{\mathbf{V}\} = \{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$. The orthogonal complement of the span of the columns of $\mathbf{V}$ is denoted by $\mathbf{V}^\perp$. We write $\mathbf{V}_{:r} \in \mathbb{R}^{n\times r}$ for the matrix whose columns correspond to the first r columns of $\{\mathbf{V}\}$. In addition, for a matrix $\mathbf{A} \in \mathbb{R}^{a\times n}$ and matrix $\mathbf{V} \in \mathbb{R}^{n\times d}$ that has orthonormal columns, the component of $\mathbf{A}$ in $\mathbf{V}^\perp_{:r-1}$ is obtained as $\mathbf{A} - \mathbf{A}\mathbf{V}_{:r-1}\mathbf{V}^\top_{:r-1}$. Finally, the first d singular values of matrix $\mathbf{A}$, sorted in descending order, are denoted by $\sigma_1(\mathbf{A}), \ldots, \sigma_d(\mathbf{A})$. The Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as: $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$, where a_{ij} denotes the (i,j)-th element of $\mathbf{A}$.

Orthogonal Projections. We briefly recall the properties of orthogonal projections. Given a matrix with orthonormal columns $\mathbf{V} \in \mathbb{R}^{n\times d}$, and vector $\mathbf{x} \in \mathbb{R}^n$, the projection of $\mathbf{x}$ onto the column space of $\mathbf{V}$ is obtained as $\mathbf{x}^\top\mathbf{V}\mathbf{V}^\top$.

Orthogonal projections satisfy the following property.

Property 1 (Orthogonal projection). For any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and a matrix $\mathbf{V} \in \mathbb{R}^{n\times d}$ with orthonormal columns $\mathbf{v}_1, \ldots, \mathbf{v}_d$, we have $\|\mathbf{A}\mathbf{V}\mathbf{V}^\top\|_F^2 = \sum_{i=1}^d \|\mathbf{A}\mathbf{v}_i\mathbf{v}_i^\top\|_F^2$.

The proof is elementary, and we provide it in the appendix for completeness.

A fundamental property of the SVD is the *consistency* property, formally stated in the Eckart-Young-Mirsky theorem (Eckart & Young, 1936). We state this pivotal theorem next. Given a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ and its singular value decomposition $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^\top$, then for any $d = rank(\mathbf{X}_d) \leq rank(\mathbf{M})$ we have that the reconstruction error

$$\|\mathbf{M} - \mathbf{X}_d\|_\xi$$

is minimized by $\mathbf{X}_d = \mathbf{M}\mathbf{V}_{:d}\mathbf{V}^\top_{:d}$, i.e. the projection of $\mathbf{M}$ onto the first d singular vectors. Here ξ denotes either the Frobenius norm ($\xi = F$) or the spectral norm ($\xi = 2$).

4. Overview of the Method

In this section we describe the multigroup low-rank approximation method. Our fundamental building block is the concept of a *multigroup* singular vector. A multigroup singular vector is rank-1 projection of the data, that takes all groups into account. Given a method to compute such a vector, it is fairly simple to obtain a *consistent* set of multigroup singular vectors by iteratively removing the component of the data that lies in the span of that vector.

First, we formally define the concept of a multigroup singular vector. We call the problem of finding such a vector MG-SINGULARVECTOR, and it represents the main algorithmic focus of this paper. Then, we give an algorithm for computing a consistent set of such vectors.

Multigroup singular vector. We seek to find a vector, such that the resulting rank-1 projection minimizes the maximum loss incurred by any group. The idea of minimizing the maximum per-group loss is inspired by the egalitarian rule in algorithmic fairness (Martinez et al., 2020). Assume an input matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ with rows divided into groups $\mathcal{G} = \{\mathbf{A}^1, \ldots, \mathbf{A}^k\}$. A multigroup singular vector is the vector $\mathbf{v}$ minimizing the maximum loss over all groups of the difference between the largest singular value, $\sigma_1(\mathbf{A}^i)$, and the rank-1 projection of $\mathbf{A}^i$ using $\mathbf{v}$. That is, $\mathbf{v}$ minimizes the loss $\mathcal{L}(\mathbf{M}, \mathbf{v})$ defined as

$$\mathcal{L}(\mathbf{M}, \mathbf{v}) = \max_{\mathbf{A}^i \in \mathcal{G}} \{\sigma_1^2(\mathbf{A}^i) - \|\mathbf{A}^i\mathbf{v}\mathbf{v}^\top\|_F^2\}. \qquad (1)$$

Recall that $\sigma_1(\mathbf{A}^i)$, corresponds to the maximum norm of any rank-1 projection of $\mathbf{A}^i$. Subtracting from $\sigma_1(\mathbf{A}^i)$ helps avoid bad minima of the minimization problem in Equation 1, by taking into account the best achievable rank-1 representation of every group. Since this loss function is a rank-1 version of the *marginal loss* of Samadi et al. (2018), we refer to Appendix A and Samadi et al. (2018) for a broader discussion on this.

Computing a set of multigroup singular vectors. We are now ready to describe the iterative algorithm for obtaining a sequence of consistent multigroup singular vectors. The algorithm works as follows. We solve the rank-1 problem in Equation 1 to obtain the multigroup singular vector $\mathbf{v}_1$. Given $\mathbf{v}_1$, we project the groups in $\mathcal{G}$ onto $\{\mathbf{v}_1\}^\perp$, the orthogonal complement of $\mathbf{v}_1$. We repeat the same process on

Algorithm 1 Multigroup Orthonormalization

Input: Matrices $\{\mathbf{A}^1, \ldots, \mathbf{A}^k\}$, rank d.
Initialize $r \leftarrow 1$, $V \leftarrow \emptyset$
while $r <= d$ **do**
  $\mathbf{v}_r \leftarrow$ MG-SINGULARVECTOR $(\mathbf{A}_1, \ldots, \mathbf{A}_k)$
  $\mathbf{A}_i \leftarrow \mathbf{A}_i - \mathbf{A}_i \mathbf{v}_r \mathbf{v}_r^\top$
  $\mathbf{V} \leftarrow \mathbf{V} \cup \mathbf{v}_r$
  $r \leftarrow r + 1$
end while
return $\mathbf{V}$

$\{\mathbf{v}_1\}^\perp$ to obtain a new vector $\mathbf{v}_2$. Repeating this process d times, we obtain an orthonormal basis $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$.

The whole iterative process is summarized in Algorithm 1. Step 1 of the algorithm corresponds to a call to a subroutine, described in the following sections, that solves the MG-SINGULARVECTOR problem. Step 1 illustrates the projection onto the orthogonal complement of the solution vector to the MG-SINGULARVECTOR problem.

Quality of the solutions. Since our algorithm iteratively produces orthonormal vectors, Property 1 and a simple inductive argument imply that the loss for a solution of d dimensions is the sum of d losses of rank-1 solutions. Thus, the quality of our solution depends only on our ability to solve the rank-1 problem. Indeed, assume we are in step r of Algorithm 1: we have $\{\mathbf{V}\} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{r-1}\}$, and we seek $\{\mathbf{V}'\} = \{\mathbf{V}\} \cup \{\mathbf{v}_r\}$. Since $\|\mathbf{A}^i \mathbf{V}' \mathbf{V}'^\top\|_F^2$, can be decomposed as $\|\mathbf{A}^i \mathbf{V}' \mathbf{V}'^\top\|_F^2 = \|\mathbf{A}^i \mathbf{V}\mathbf{V}^\top\|_F^2 + \|\mathbf{A}^i \mathbf{v}_r \mathbf{v}_r^\top\|_F^2$, the problem reduces to solving a sequence of rank-1 problems. We will refer to the total error $\sum_{i=1}^d \mathcal{L}(\mathbf{M}, \mathbf{v}_i)$, as the *incremental error*.

In the following sections, we will show that, for the case of two groups, we can in fact efficiently solve the rank-1 problem to optimality, and for the general case we will give an approximate algorithm which works well in practice.

Complexity. The overall time complexity is $\mathcal{O}(d\ell)$, where $\mathcal{O}(\ell)$ is the complexity of solving MG-SINGULARVECTOR that is discussed later.

5. The Multigroup Singular Vector Problem

As anticipated in Section 4, solving the MG-SINGULARVECTOR problem represents the crucial algorithmic challenge to be addressed for multigroup low-rank approximation. In this section, we study the properties of the MG-SINGULARVECTOR problem, as well as of its dual problem, which is more amenable to optimization. Leveraging the insights gained in the present section, in the next section we introduce algorithms to solve the MG-SINGULARVECTOR problem.

Next, we formalize the MG-SINGULARVECTOR problem.

Problem 1 (MG-SINGULARVECTOR). Given a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ with rows divided into groups $\mathcal{G} = \{\mathbf{A}^1, \ldots, \mathbf{A}^k\}$, find the vector $\mathbf{v}$ satisfying

$$\min_{\mathbf{v} \in \mathbb{R}^n, z \in \mathbb{R}} z$$
$$\text{s.t. } \sigma_1^2(\mathbf{A}^i) - \|\mathbf{A}^i \mathbf{v}\mathbf{v}^\top\|_F^2 \leq z \text{ for all } \mathbf{A}^i \in \mathcal{G}$$
$$\text{and } \|\mathbf{v}\|_2^2 = 1.$$

We use the term *constraint functions* $h_i(\mathbf{v})$ for the left-hand sides of the constraints in the problem:

$$h_i(\mathbf{v}) = \sigma_1^2(\mathbf{A}^i) - \|\mathbf{A}^i \mathbf{v}\mathbf{v}^\top\|_F^2.$$

Convexity analysis. Problem 1 is not a convex problem. To see this, note that $-\|\mathbf{A}\mathbf{v}\mathbf{v}^\top\|_F^2 = -\mathbf{v}^\top \mathbf{A}^\top \mathbf{A}\mathbf{v}$, and since $-\mathbf{A}^\top \mathbf{A}$ is a negative semidefinite matrix, the corresponding quadratic forms are concave functions. Each $h_i(\mathbf{v})$ consists of such a quadratic form and an affine transformation (which does not impact convexity), and is thus concave. Minimization problems over concave functions are non-convex and not straightforward to solve.

We also note that the constraint functions h_i are continuous functions supported on the unit hypersphere. Importantly, all their minima are at zero, since the incremental loss attains its minimum at 0. Given these observations, we can prove that for any optimal solution $\mathbf{v}^*, z^*$, two groups attain exactly the same error, while other groups smaller or equal error.

Theorem 5.1. *For an optimal solution $\mathbf{v}^*, z^*$ to Problem 1 we have:*

$$z^* = h_i(\mathbf{v}^*) = h_j(\mathbf{v}^*) \geq h_k(\mathbf{v}^*),$$

for some $i \neq j$ and for all $k \neq i, j$.

Proof. We first prove that there exist two groups such that $z^* = h_i(\mathbf{v}^*) = h_j(\mathbf{v}^*)$. Assume for the sake of contradiction that $\mathbf{v}^*$ is an optimal solution such that $z^* = h_i(\mathbf{v}^*) > h_j(\mathbf{v}^*)$ for all j. Then, this implies that $h_i(\mathbf{v}^*) > 0$, and $\mathbf{v}^*$ cannot be a minimizer of h_i since the minima of h_l for all l, are at $h_l(\mathbf{v}) = 0$. Thus we can locally move to a nearby solution $\mathbf{v}_\epsilon$ such that $h_i(\mathbf{v}_\epsilon) < h_i(\mathbf{v}^*)$ and at the same time $h_j(\mathbf{v}_\epsilon) \leq h_i(\mathbf{v}_\epsilon)$ for all j. This contradicts the fact that $\mathbf{v}^*$ is an optimal solution. Additionally it must be that $h_k(\mathbf{v}^*) \leq h_i(\mathbf{v}^*) = h_j(\mathbf{v}^*)$ since $h_i(\mathbf{v}^*) = h_j(\mathbf{v}^*)$ attain the optimal value z^*. □

As we stated before, Problem 1 is easier when there are two groups. The proof hints at an interesting geometric intuition for why that is the case. In this setting, we have two quadratic constraint functions h_1 and h_2, and as Theorem 5.1 suggests, the candidate optimal solutions $\mathbf{v}$ lie at

the intersection points of two ellipsoids determined by the quadratic equation $h_1 = h_2$. Thus, it suffices to start from the minimum of either h_1 or h_2 (note that this minimum is found by setting $\mathbf{v}$ to the leading eigenvector of either group 1 or 2) and follow the direction of steepest descent of the objective function, to find the global minimum (due to symmetry it can be either $\mathbf{v}^*$ or $-\mathbf{v}^*$). We will formalize this intuition by characterizing the KKT points, in Section 7, where we show that in the two-group case the problem enjoys strong duality. Indeed, this is not surprising, as several strong results exist for non-convex problems with two quadratic constraints (we refer to (Boyd & Vandenberghe, 2004), Appendix B). We now proceed to define the dual problem for the general case.

The dual problem. To study Problem 1, it is useful to consider the *dual* problem. An advantage of the dual problem over the primal problem (Problem 1) is that it leads to an objective function with a gradient that is more "informative" and more convenient for gradient-based methods. Methods such as Frank-Wolfe (Frank et al., 1956), use the gradient to determine the search direction in the feasible region.

As already mentioned, we show that for $k = |\mathcal{G}| = 2$ Problem 1 exhibits strong duality (i.e., the optimal value of the primal problem equals the optimal value of the dual problem), despite being non-convex.

However, even for $|\mathcal{G}| > 2$, the dual problem is still useful in practice, since we can assess the quality of our solution by evaluating the difference between the primal and dual optimal objective values.

To formulate the dual problem we will consider the Lagrangian function corresponding to Problem 1. The Lagrangian is obtained by adding the problem constraints to the objective, along with the dual variables, which correspond to the Lagrange *multipliers*. In particular, the Lagrangian function associated with Problem 1 is:

$$\mathcal{H}(\mathbf{v}, z, \boldsymbol{\mu}, \lambda) = z + \sum_{i=1}^{k} \mu_i (h_i(\mathbf{v}) - z) + \lambda(\|\mathbf{v}\|_2^2 - 1),$$

where we denote $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]$. Further, let

$$\mathbf{A}(\boldsymbol{\mu}) = \sum_{i=1}^{k} \mu_i (\mathbf{A}^i)^\top \mathbf{A}^i,$$

and define $\mathbf{s} = [\sigma_1^2(\mathbf{A}^1), \ldots, \sigma_1^2(\mathbf{A}^k)]$. The dual problem associated with Problem 1 is the following.

Problem 2 (MG-SINGULARVECTOR-DUAL).

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^k} \boldsymbol{\mu}^\top \mathbf{s} - \lambda_{max}(\mathbf{A}(\boldsymbol{\mu}))$$
$$\text{s.t. } \mathbf{1}^\top \boldsymbol{\mu} = 1 \quad (2)$$
$$\boldsymbol{\mu} \geq 0. \quad (3)$$

Here, λ_{max} denotes the maximum eigenvalue. A detailed derivation of Problem 2 is given in the appendix.

Problem 2 is convex and has an interesting interpretation as a parametric eigenvalue problem: the solution vector $\mathbf{v}$ is the leading eigenvector of the optimal convex combination $\mathbf{A}(\boldsymbol{\mu})$, determined by the coefficients $\boldsymbol{\mu}$.

Uniqueness. Later on, we define the solution $\mathbf{v}$ which we obtain from the dual as a function of $\boldsymbol{\mu}$, which requires uniqueness. However, in general, $\mathbf{v}$ is not unique, as $\mathbf{A}(\boldsymbol{\mu})$ may have repeated eigenvalues. This is not a problem in practice since real data contain noise, which leads to distinct eigenvalues (Kato, 1966). In any case, it is always possible to slightly perturb the data to avoid ill-conditioned scenarios with repeated eigenvalues.

6. Algorithms for Multigroup Singular Vector

In this section we present two algorithms for MG-SINGULARVECTOR. The first algorithm solves the dual problem MG-SINGULARVECTOR-DUAL, which is a convex optimization problem with linear constraints, using the Frank-Wolfe algorithm. The second one solves a semidefinite programming (SDP) relaxation of the primal problem.

Frank-Wolfe. The Frank-Wolfe algorithm is a widely-used iterative algorithm for solving constrained convex optimization problems (Pokutta, 2023). In each iteration, the algorithm linearizes the objective function, and moves towards its minimizer , while staying inside the feasible region.

The Frank-Wolfe algorithm is particularly easy to use for Problem 2, as the dual constraints are almost trivial to satisfy and thus the only computationally challenging aspect for the algorithm is the computation of the gradient ∇g of the dual objective, $g(\boldsymbol{\mu}) = \boldsymbol{\mu}^\top \mathbf{s} - \lambda_{max}(\mathbf{A}(\boldsymbol{\mu}))$, which involves computing the gradient of $\lambda_{max}(\mathbf{A}(\boldsymbol{\mu}))$.

Denoting for brevity $\lambda(\boldsymbol{\mu}) = \lambda_{max}(\mathbf{A}(\boldsymbol{\mu}))$, we have that $\lambda(\boldsymbol{\mu})$ is an eigenvalue of $\mathbf{A}(\boldsymbol{\mu})$ and hence:

$$\mathbf{A}(\boldsymbol{\mu})\mathbf{v}(\boldsymbol{\mu}) = \lambda(\boldsymbol{\mu})\mathbf{v}(\boldsymbol{\mu}), \quad (4)$$

where $\mathbf{v}(\boldsymbol{\mu})$ is the eigenvector corresponding to $\lambda(\boldsymbol{\mu})$. Taking the gradient and using the product rule, we have:

$$(\mathbf{A}^i)^\top \mathbf{A}^i \mathbf{v}(\boldsymbol{\mu}) + \mathbf{A}(\boldsymbol{\mu})\nabla\mathbf{v}(\boldsymbol{\mu}) = \nabla\lambda(\boldsymbol{\mu})\mathbf{v}(\boldsymbol{\mu}) + \lambda(\boldsymbol{\mu})\nabla\mathbf{v}(\boldsymbol{\mu}). \quad (5)$$

To simplify the gradient, we use the constraint $\mathbf{v}(\boldsymbol{\mu})^\top \mathbf{v}(\boldsymbol{\mu}) = 1$. This gives:

$$\nabla\mathbf{v}(\boldsymbol{\mu})^\top \mathbf{v}(\boldsymbol{\mu}) + \mathbf{v}(\boldsymbol{\mu})\nabla\mathbf{v}(\boldsymbol{\mu}) = 0,$$

i.e., $\mathbf{v}(\boldsymbol{\mu})$ is orthogonal to its gradient. Therefore, multiplying equation 5 with $\mathbf{v}(\boldsymbol{\mu})^\top$, we obtain:

$$\mathbf{v}(\boldsymbol{\mu})^\top (\mathbf{A}^i)^\top \mathbf{A}^i \mathbf{v}(\boldsymbol{\mu}) + \lambda(\boldsymbol{\mu})\mathbf{v}(\boldsymbol{\mu})^\top \nabla\mathbf{v}(\boldsymbol{\mu})$$
$$= \nabla\lambda(\boldsymbol{\mu})\mathbf{v}(\boldsymbol{\mu})^\top \mathbf{v}(\boldsymbol{\mu}) + \lambda(\boldsymbol{\mu})\mathbf{v}(\boldsymbol{\mu})^\top \nabla\mathbf{v}(\boldsymbol{\mu}),$$

Algorithm 2 Frank-Wolfe for MG-SINGULARVECTOR-DUAL

1: **Input:** Matrices $\mathbf{A}^1, \ldots, \mathbf{A}^k$, convergence tolerance ϵ.
2: **Initialize:** Set $\boldsymbol{\mu}^{(0)} = [1, 0, \ldots, 0]$,
   $\mathbf{s} = [\sigma_1^2(\mathbf{A}^1), \ldots, \sigma_1^2(\mathbf{A}^k)]$
3: $t \leftarrow 0$
4: **repeat**
5: $\quad \mathbf{v}(\boldsymbol{\mu}^{(t)}) \leftarrow \mathbf{x}$ s.t. $\mathbf{A}(\boldsymbol{\mu}^{(t)})\mathbf{x} = \lambda_{max}\mathbf{x}$
6: $\quad \nabla g(\boldsymbol{\mu}^{(t)})_i \leftarrow \mathbf{s}_i + \mathbf{v}(\boldsymbol{\mu}^{(t)})^\top (\mathbf{A}^i)^\top \mathbf{A}^i \mathbf{v}(\boldsymbol{\mu}^{(t)})$
7: $\quad \mathbf{s}^{(t)} \leftarrow \arg\max_{\mathbf{y}:\mathbf{1}^\top\mathbf{y}=1, \mathbf{y}\geq 0} \mathbf{y}^\top \nabla g(\boldsymbol{\mu}^{(t)})$
8: $\quad \gamma_t \leftarrow \frac{2}{t+2}$
9: $\quad \boldsymbol{\mu}^{(t+1)} \leftarrow (1-\gamma_t)\boldsymbol{\mu}^{(t)} + \gamma_t \mathbf{s}^{(t)}$
10: $\quad t \leftarrow t+1$
11: **until** $\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{(t-1)}\| < \epsilon$
12: **return** $\boldsymbol{\mu}^{(t)}, \mathbf{v}(\boldsymbol{\mu}^{(t)})$

which simplifies to $(\nabla\lambda(\boldsymbol{\mu}))_i = \mathbf{v}(\boldsymbol{\mu})^\top (\mathbf{A}^i)^\top \mathbf{A}^i \mathbf{v}(\boldsymbol{\mu})$. Putting everything together, we conclude that:

$$(\nabla g)_i = \mathbf{s}_i - \mathbf{v}(\boldsymbol{\mu})^\top (\mathbf{A}^i)^\top \mathbf{A}^i \mathbf{v}(\boldsymbol{\mu}). \tag{6}$$

Algorithm 1 contains the pseudocode of the Frank-Wolfe method for Problem 2. The algorithm proceeds as follows. It starts with an initial feasible solution $\boldsymbol{\mu}^{(0)}$. In each step, line 5 solves the maximum eigenvalue problem associated with the given parameter vector $\boldsymbol{\mu}^{(t)}$. Line 6 computes the gradient. Line 7 solves a linear maximization problem over the simplex defined by constraints 2 and 3 in Problem 2. Lines 8-9 describe standard parameter update steps of the Frank-Wolfe algorithm. Finally, the returned solution is the $\mathbf{v}(\boldsymbol{\mu})$ for the final update and the corresponding dual solution $\boldsymbol{\mu}$. The complexity of the algorithm is dominated by the maximum eigenvalue step (Line 5) which can be handled using a fast Lanczos implementation. Thus the overall complexity is $\mathcal{O}(tn^2)$ where t is the number of iterations until convergence.

Algorithm 2 solves Problem 2 optimally, as it is a convex problem. However, the value $g(\boldsymbol{\mu})$ of the dual objective is only a lower bound on the primal objective (i.e., Problem 1) i.e., there can be a non-zero duality gap.

Semidefinite programming. We also solve MG-SINGULARVECTOR through a semidefinite programming (SDP) relaxation (Boyd & Vandenberghe, 2004). Since SDP solvers come with an $\mathcal{O}(n^6)$ running time, this algorithm is expected to be significantly slower than Algorithm 2. However, since for $|\mathcal{G}| > 2$ we are not guaranteed to solve MG-SINGULARVECTOR exactly, the SDP relaxation may offer a solution that is close to rank-1, and hence close to optimal. As a consequence, this approach can be useful in settings where accuracy is more important than efficiency. The pseudocode of the SDP for solving Problem 1 is provided in the appendix (Algorithm 3) where we also present experiments demonstrating more accurate approximation of the primal optimum compared to Algorithm 2.

7. Algorithm and Analysis for Two Groups

Often, the data are divided into exactly two groups, e.g., on the basis of sex. As mentioned, in this particular case, we are able to solve Problem 1 optimally and, moreover, the optimal solution equalizes the loss.

Algorithm. We give a novel algorithm dedicated to the two-group case, which outperforms competitors (such as the Frank-Wolfe algorithm) in this setting, but cannot be conveniently extended to address the setting of more than two groups. The algorithm relies on the observation that for $|\mathcal{G}| = 2$ there exists a unique feasible $\boldsymbol{\mu}$ for the dual, which satisfies Theorem 5.1. We can find such a $\boldsymbol{\mu}$ using a fast root-finding approach, such as Brent's method (Brent, 1971). This algorithm is evaluated in our experiments and its details are in the appendix (see Lemma E.1).

Theoretical Analysis. The case of $|\mathcal{G}| = 2$ has interesting theoretical properties, which we present here. All the proofs can be found in the appendix. First, we observe that, as a consequence of Theorem 5.1, it holds that $h_1(\mathbf{v}^*) = h_2(\mathbf{v}^*)$ for any optimal solution $\mathbf{v}^*$ to Problem 1.

Furthermore, leveraging the KKT conditions (see, e.g., (Kuhn & Tucker, 2013)) to characterize the optimal solutions to Problem 2 leads to the following theorem.

Theorem 7.1. *For $|\mathcal{G}| = 2$, the optimal solution to* MG-SINGULARVECTOR *can be computed in polynomial time.*

The proof relies on a simple idea. Since Problem 2 is convex, it has a unique maximum, which can be found in polynomial time (for example, using the approach based on the Frank-Wolfe algorithm). Such a unique maximum can be characterized by the KKT conditions. Then, to complete the proof, it suffices to show that Problems 1 and 2 attain strong duality, i.e., $f(\mathbf{v}^*) = z^* = g(\boldsymbol{\mu}^*)$, where $\mathbf{v}^*$ and $\boldsymbol{\mu}^*$ are optimal solutions to Problems 1 and 2, respectively.

Our analysis reveals interesting properties which are described in the following lemmas and proved in the appendix.

Lemma 7.2. *For $|\mathcal{G}| = 2$, the SDP relaxation in Algorithm 3 is tight.*

Finally, the following lemma related to Algorithm 1 follows.

Lemma 7.3. *For $|\mathcal{G}| = 2$, an optimal solution of Algorithm 1 is such that the total error for the two groups is equal.*

8. Experiments

This section presents our experimental evaluation, which aims at assessing the performance of our method (Algorithm

Table 1: Dataset statistics. For each dataset, we report the number of columns (n), the number of groups ($|\mathcal{G}|$), and the number of rows and rank by group.

Dataset	Columns (n)	$\lvert\mathcal{G}\rvert$	Group Rows	Group Ranks
heart	14	2	201, 96	13, 13
german	63	2	690, 310	49,47
credit	25	2	18 112, 11 888	24, 24
student	58	2	383, 266	42, 42
adult	109	2	21 790, 10 771	98, 98
compas	189	2	619, 100	165, 71
communities	104	2	1 685, 309	101, 101
recidivism	227	2	1 923, 310	175, 113
compas-3	189	3	241, 240, 238	115, 110, 97
communities-4	104	4	90, 1 571, 218, 115	90, 99, 103, 103

1) in the FAIR-PCA task. Exploring other applications is left to future work. We refer to the multigroup singular vectors output by Algorithm 1 as MULTIGROUP SVS.

The experiments consider both the two-group case, where our methods are supported by optimality guarantees, and the case of more than two groups, where the optimality guarantees no longer hold, but we observe that, in practice, the gap between primal and dual solutions is consistently small and hence they are close to optimal (see Appendix D). The results show that our method can offer significant advantages over recent methods for FAIR-PCA.

8.1. Settings

Next, we illustrate the datasets, metrics, baselines, parameter settings and experimental setup used in our experiments.

Datasets.

- **Datasets with two groups,** We use the juvenile recidivism data (recidivism) from Catalunya (Tolan et al., 2019) and various datasets from the UCI machine learning repository (Dua & Graff, 2017): "heart-cleveland" (heart), "german-credit" (german), "credit-card" (credit), "student performance" (student), "adult" (adult), "compas-recidivism" (compas), "communities" (communities). Group membership is based on sex, except for "communities " where groups determined by racial composition (caucasian majority or not).
- **Datasets with more than two groups.** We consider the "compas-recidivism" dataset partitioned into three groups according to age (compas-3), and the "communities" dataset partitioned into four groups, namely "blacks", "hispanics", "asians" and "caucasians", according to the dominant ethnicity (communities-4).

Data are processed by removing protected attributes, one-hot encoding categorical variables, and standardizing columns. Table 1 shows summary statistics of the datasets.

Baselines. We compare against the FAIR-PCA-SDP algorithm based on semi-definite programming (Tantipongpipat et al., 2019) and against the BICRITERIA algorithm (Song et al. (2024), Algorithm 3). Given target rank d, FAIR-PCA-SDP and BICRITERIA return a rank-d projection matrix $\mathbf{P} = \mathbf{U}\Lambda\mathbf{U}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times d}$ is obtained through SVD. In our experiments, we evaluate the consistency property by comparing the loss when using $\mathbf{U}_{:r}$ of FAIR-PCA-SDP and BICRITERIA against $\mathbf{V}_{:r}$ retrieved by Algorithm 1, for all $r < d$.

Metrics and parameters. To evaluate the performance of our method and FAIR-PCA-SDP, we monitor the marginal loss (introduced in Samadi et al. (2018)), incremental loss, (see Section 4), and the standard L_2 reconstruction loss.

Both the marginal and incremental losses quantify the deviation from the optimal reconstruction, whereas the L_2 reconstruction loss does not account for such optimal reconstruction. As the BICRITERIA algorithm (Song et al., 2024) is designed to optimize the L_2 reconstruction loss only, it is not competitive with our method and FAIR-PCA-SDP in terms of marginal and incremental loss.

We show the variation of each loss in the groups as a function of the (target) reconstruction rank d, which we vary from 1 to 8. Finally, we measure runtimes in seconds.

Experimental setup. Our implementation is written in `python`. In the two-groups case, the singular vectors for multigroup data are obtained by the tailored algorithm based on the root-finding procedure, while for more than two groups, they are obtained by the Frank-Wolfe algorithm.

Experiments are executed on a compute node with 32 cores and 256GB of RAM. The (anonymized) source code is available online [1].

8.2. Results for Two-group Data

Figure 2 (top) shows the different losses incurred by our method and the baselines in the compas datasets as a function of the target rank d. Due to the space limitations, analogous results for all the other datasets are presented in the appendix (Figure 3). The figure highlights the crucial advantage of MULTIGROUP SVS: the incremental loss is the same in both groups for all values of the rank parameter lower than the input target rank (8), meaning that fairness is also pursued in the lower-dimensional subspaces. In particular, the incremental loss is considerably smaller for MULTIGROUP SVS than for FAIR-PCA-SDP. On the other hand, the marginal loss optimized by FAIR-PCA-SDP is never significantly smaller for FAIR-PCA-SDP than for MULTIGROUP SVS, but tends to be smaller for MULTIGROUP SVS. Finally, the reconstruction loss is consistently comparable for FAIR-PCA-SDP and MULTIGROUP SVS, but tends

[1] https://anonymous.4open.science/r/multigroupSVs-F716/

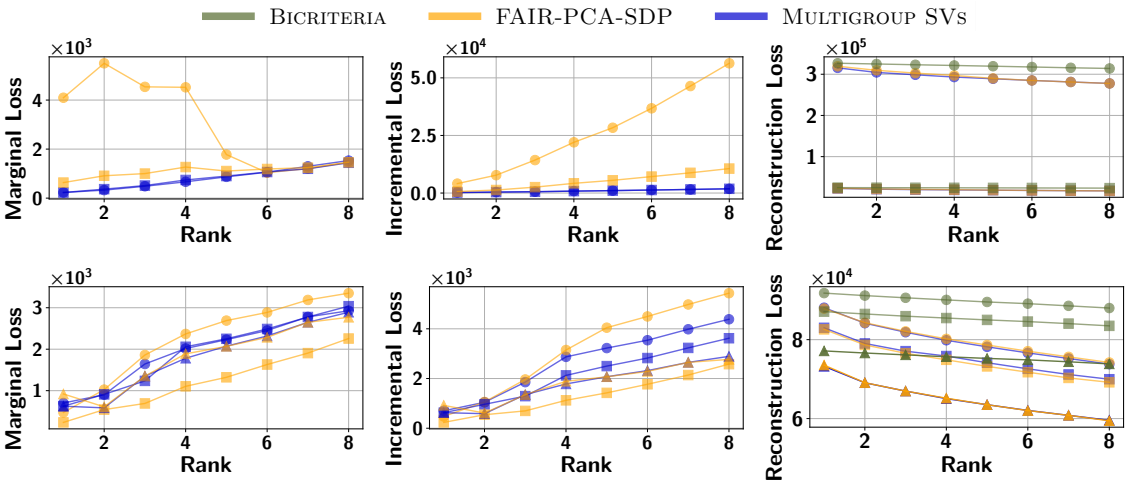


Figure 2: compas dataset with two groups (top) and compas-3 dataset with three groups (bottom). Marginal, incremental, and reconstruction loss by rank. Different marker symbols indicate different groups.

Table 2: Runtimes of MULTIGROUP SVS, FAIR-PCA-SDP and BICRITERIA (in seconds) in all datasets ($d = 8$).

Dataset	MULTIGROUP SVS	FAIR-PCA-SDP	BICRITERIA
heart	0.009	0.022	0.016
german	0.1	0.9	0.021
credit	0.23	0.084	0.053
student	0.067	0.64	0.031
adult	2.16	9.13	0.2
compas	0.71	143.15	0.053
communities	0.28	8.62	0.035
recidivism	1.28	357.59	0.061
compas-3	2.54	124.11	0.019
communities-4	1.23	11.16	0.024

to be larger for BICRITERIA. Unlike the incremental and marginal losses, the reconstruction loss can be highly unbalanced since both MULTIGROUP SVS and FAIR-PCA-SDP do not seek to balance the reconstruction loss, but rather the distance to the best possible approximation.

In addition, Table 2 shows the runtime of MULTIGROUP SVS and the baselines in the different datasets. BICRITERIA is typically the fastest algorithm. However, it is not competitive with MULTIGROUP SVS and FAIR-PCA-SDP in terms of performance, even for reconstruction loss. On the other hand, FAIR-PCA-SDP becomes slow as dataset size increases, and MULTIGROUP SVS is generally faster that FAIR-PCA-SDP, often by orders of magnitude. MULTIGROUP SVS always deliver high-quality results in terms of all the metrics under consideration in less than three seconds.

8.3. Results for more than two groups

In case there are more than two groups, the problems solved by the algorithms under comparison become NP-hard, and the algorithms drop the optimality guarantees.

As Figure 2 (bottom) shows for the compas-3 dataset, MULTIGROUP SVS consistently yield a more balanced low-dimensional data representation than FAIR-PCA-SDPand BICRITERIA as the rank increases. This observation suggests that MULTIGROUP SVS provides an effective heuristic for multigroup dimensionality reduction with an arbitrary number of groups $|\mathcal{G}| > 2$. Results of the same experiments for the communities-4 dataset, given in Figure 4 in the appendix, lead to analogous observations. Figures 2 (bottom) and 4 also show that, as discussed, when there are more than two groups in the data, there is no guarantee of equality in the marginal and incremental losses associated with different groups.

Nevertheless, in the appendix (Figures 5 and 6), we present experimental results demonstrating that the gap between primal and dual solutions in practice tends to be negligible, and thus our solutions tend to be close to optimal.

To conclude, Table 2 also reports the runtimes in the experiments with more than two groups, which confirm the trends observed in the two-group case.

9. Conclusion

We have introduced the problem of consistent multigroup low-rank approximation that, given a dataset partitioned into groups, asks for a sequence of orthonormal vectors such that projecting the data onto their spanned subspace minimizes the maximum error across groups, and such that any subsequence is also an optimal solution of smaller length.

We have proposed efficient and theoretically well-founded methods to compute the desired sequence of vectors. Extensive experiments highlight the advantages of our methods over existing approaches.

References

Babu, P. and Stoica, P. Fair principal component analysis (pca): minorization-maximization algorithms for fair pca, fair robust pca and fair sparse pca. *arXiv preprint arXiv:2305.05963*, 2023.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Brent, R. P. An algorithm with guaranteed convergence for finding a zero of a function. *The computer journal*, 14(4): 422–425, 1971.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Ehiwario, J. and Aghamie, S. Comparative study of bisection, newton-raphson and secant methods of root-finding problems. *IOSR Journal of Engineering*, 4(4):01–07, 2014.

Frank, M., Wolfe, P., et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956.

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

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

Kato, T. *Perturbation Theory for Linear Operators*. 1966.

Kleindessner, M., Donini, M., Russell, C., and Zafar, M. B. Efficient fair pca for fair representation learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 5250–5270. PMLR, 2023.

Kuhn, H. W. and Tucker, A. W. Nonlinear programming. In *Traces and emergence of nonlinear programming*, pp. 247–258. Springer, 2013.

Lee, J., Kim, G., Olfat, M., Hasegawa-Johnson, M., and Yoo, C. D. Fast and efficient mmd-based fair pca via optimization over stiefel manifold. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7363–7371, 2022.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. S. The variational fair autoencoder. In *4th International Conference on Learning Representations, ICLR, San Juan, Puerto Rico*, 2016.

Martinez, N., Bertran, M., and Sapiro, G. Minimax pareto fairness: a multi objective perspective. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Matakos, A., Ordozgoiti, B., and Thejaswi, S. Fair column subset selection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 2189–2199, New York, NY, USA, 2024. Association for Computing Machinery.

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

Pelegrina, G. D. and Duarte, L. T. A novel approach for fair principal component analysis based on eigendecomposition. *IEEE Transactions on Artificial Intelligence*, 2023.

Pokutta, S. The frank-wolfe algorithm: A short introduction. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 2023.

Samadi, S., Tantipongpipat, U., Morgenstern, J., Singh, M., and Vempala, S. The price of fair pca: One extra dimension. In *NeuRIPS*, NIPS'18, pp. 10999–11010. Curran Associates Inc., 2018.

Schiffer, M. Hadamard's formula and variation of domain-functions. *American Journal of Mathematics*, 68(3):417–448, 1946.

Song, Z., Vakilian, A., Woodruff, D., and Zhou, S. On socially fair low-rank approximation and column subset selection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Tantipongpipat, U. T., Samadi, S., Singh, M., Morgenstern, J., and Vempala, S. Multi-criteria dimensionality reduction with applications to fairness. In *NIPS*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Tolan, S., Miron, M., Gómez, E., and Castillo, C. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *International Conference on Artificial Intelligence and Law*, pp. 83–92, 2019.

Van Der Maaten, L., Postma, E. O., Van Den Herik, H. J., et al. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(66-71):13, 2009.

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

A. Loss Functions

In this section, we discuss widely used loss functions.

An equivalence that we will frequently use is $\sigma_i(\mathbf{M}) = \|\mathbf{M}\mathbf{v}_i\mathbf{v}_i^\top\|_F$, where $\mathbf{v}_i$ is the i-th singular vector of $\mathbf{M}$.

A.1. Reconstruction Error

A commonly used and natural loss function for a given group is the reconstruction error.

Definition A.1 (Reconstruction Error)**.** Given matrix $\mathbf{A} \in \mathbb{R}^{a\times n}$ and an $n \times d$ matrix $\mathbf{V} \in \mathcal{V}_d$, the reconstruction error of $\mathbf{A}$ using $\mathbf{V}$ is

$$\mathcal{L}_{\text{rec}}(\mathbf{A}, \mathbf{V}) \triangleq \|\mathbf{A} - \mathbf{A}\mathbf{V}\mathbf{V}^\top\|_F^2 = \sum_{i=1}^{n} \sigma_i^2(\mathbf{A}) - \|\mathbf{A}\mathbf{V}\mathbf{V}^\top\|_F^2,$$

where the equivalence holds due to the properties of projection matrices.

However, the reconstruction error has a serious limitation when considering multiple groups (Samadi et al., 2018). To explain this, imagine that we are given a data matrix with two groups $\mathbf{M} = \{\mathbf{A}, \mathbf{B}\}$, and $\mathbf{W}^A$, $\mathbf{W}^B$ the corresponding minimizers of $\mathcal{L}_{\text{rec}}(\mathbf{A}, \cdot\,)$ and $\mathcal{L}_{\text{rec}}(\mathbf{B}, \cdot\,)$, for some rank d. We can obtain $\mathbf{W}^A$ and $\mathbf{W}^B$ from the SVD of $\mathbf{A}$ and $\mathbf{B}$ accordingly. Now consider that $\mathcal{L}_{\text{rec}}(\mathbf{A}, \mathbf{W}^A) = \ell_A >> \mathcal{L}_{\text{rec}}(\mathbf{B}, \mathbf{W}^B) = \ell_B$, i.e., the best possible rank-d reconstruction error for $\mathbf{B}$ is much better than the best possible reconstruction error for $\mathbf{A}$. We can see that this puts a lower bound of ℓ_A to the loss. This means that any improvement to the reconstruction error of $\mathbf{B}$, beyond ℓ_A, cannot improve the objective. This may be considered unfair to group B, since it suffers from a high reconstruction error only due to the fact that group A cannot be reconstructed well enough in a rank d subspace.

A.2. Marginal Loss

Tantipongpipat et al. (2019) consider a family of problems under the term *multicriteria dimensional reduction*, where the task is to find a subspace that takes into account various groups present in the data, in a balanced manner.

Problem 3 ((f, g)-Multicriteria dimension reduction). For each group $\mathbf{A}^i$, associate a function $f_i : \mathcal{V}_d \to \mathbb{R}$ that denotes the group's objective value for a particular $\mathbf{V} \in \mathbb{R}^{n\times d}$, and an aggregation function $g : \mathbb{R}^k \to \mathbb{R}$. Find $\mathbf{V} \in \mathcal{V}_d$ which optimizes

$$\min_{\mathbf{V}\in\mathcal{V}_d} g(f_1(\mathbf{V}\mathbf{V}^\top), f_2(\mathbf{V}\mathbf{V}^\top), \ldots, f_k(\mathbf{V}\mathbf{V}^\top)).$$

Samadi et al. (2018) introduced the marginal loss, described next. Assume that we are given a matrix $\mathbf{M}$ with groups $\{\mathbf{A}^1, \ldots, \mathbf{A}^k\}$. For some group $\mathbf{A}^g$, the singular values are $\sigma_1(\mathbf{A}^g), \ldots, \sigma_n(\mathbf{A}^g)$. Given an $n \times d$ matrix $\mathbf{V} \in \mathcal{V}_d$, the *marginal* error of group $\mathbf{A}^g$ using $\mathbf{V}$ is as follows.

Definition A.2 (FAIR-PCA loss)**.**

$$\mathcal{L}_{\text{marg}}(\mathbf{A}^g, \mathbf{V}) \triangleq \sum_{i=1}^{d} \sigma_i^2(\mathbf{A}^g) - \|\mathbf{A}^g\mathbf{V}\mathbf{V}^\top\|_F^2.$$

For more information on the marginal loss, we refer the reader to Samadi et al. (2018) and Tantipongpipat et al. (2019).

A.3. Consistency Makes Parity More Challenging

A motivating factor for using the marginal error objective in FAIR-PCA is that it ensures equal loss, when two groups are present in the data, i.e. $\mathcal{L}_{\text{marg}}(\mathbf{A}, \mathbf{V}^*) = \mathcal{L}_{\text{marg}}(\mathbf{B}, \mathbf{V}^*)$ (see Theorem 4.5 in Samadi et al. (2018))

However, the consistency requirement means neither the reconstruction error nor the marginal loss can guarantee parity of loss while meeting the consistency requirements.

As already noticed, we are interested in minimizing the loss of projecting the groups in $\mathcal{G}$ using the common projection $\mathbf{V}_{:d}\mathbf{V}_{:d}^\top$ for all values of d. Observe that $\mathbf{V}_{:d}\mathbf{V}_{:d}^\top$ is an orthogonal projection.

Observation 1. Assume that Algorithm 1 is executed on an instance with two groups $\mathbf{A} \in \mathbb{R}^{a\times n}$ and $\mathbf{B} \in \mathbb{R}^{b\times n}$, where the loss function $\mathcal{L}$ is instead either $\mathcal{L}_{\text{rec}}$ or $\mathcal{L}_{\text{marg}}$. Then for optimal solution $\mathbf{V}^* \in \mathbb{R}^{n\times d}$ it may hold that

$$\mathcal{L}(\mathbf{A}, \mathbf{V}^*(\mathbf{V}^*)^\top) \neq \mathcal{L}(\mathbf{B}, \mathbf{V}^*(\mathbf{V}^*)^\top)$$

To see why this holds for $\mathcal{L} = \mathcal{L}_{\text{rec}}$, assume that we have a solution of MULTIGROUP SVS of rank d, and that $\mathcal{L}(\mathbf{A}, \mathbf{V}_{:d}) = \mathcal{L}(\mathbf{B}, \mathbf{V}_{:d})$. The vector $\mathbf{v}_{d+1}$ lies in the orthogonal complement of $\mathbf{V}_{:d}$, $\mathbf{V}_{:d}^{\perp}$. We denote the component of $\mathbf{A}$ in the orthogonal complement of $\mathbf{V}_{:d}$ as $\mathbf{A}_{d+1}$.

If:

$$\|\mathbf{A}\|_F^2 - \|\mathbf{A}_{d+1}\mathbf{x}\mathbf{x}^\top\|_F^2 < \|\mathbf{B}\|_F^2 - \sigma_1^2(\mathbf{B}_{d+1}) \quad \forall \|\mathbf{x}\|_2^2 = 1,$$

or vice versa, then necessarily either $\mathcal{L}(\mathbf{A}, \mathbf{V}^*) < \mathcal{L}(\mathbf{B}, \mathbf{V}^*)$ or $\mathcal{L}(\mathbf{A}, \mathbf{V}^*) > \mathcal{L}(\mathbf{B}, \mathbf{V}^*)$,

For the marginal error $\mathcal{L}_{\text{marg}}$, assume again that $\mathcal{L}(\mathbf{A}, \mathbf{V}_{:d}) = \mathcal{L}(\mathbf{B}, \mathbf{V}_{:d})$, and we are seeking a vector $\mathbf{v}_{d+1}$ in $\mathbf{V}_{:d}^{\perp}$. In order for $\mathcal{L}(\mathbf{A}, \mathbf{V}_{d+1}) = \mathcal{L}(\mathbf{B}, \mathbf{V}_{d+1})$ to hold, according to Property 1, we must have:

$$\sum_{i=1}^{d+1}(\sigma_i^2(\mathbf{A}) - \|\mathbf{A}\mathbf{v}_i\mathbf{v}_i^\top\|_F^2) = \sum_{i=1}^{d+1}(\sigma_i^2(\mathbf{B}) - \|\mathbf{B}\mathbf{v}_i\mathbf{v}_i^\top\|_F^2).$$

Since by hypothesis the equality holds for the summands up to the d-th, then the equality needs to hold also for $i = d + 1$. However if:

$$\sigma_{d+1}^2(\mathbf{A}) - \|\mathbf{A}_{d+1}\mathbf{x}\mathbf{x}^\top\|_F^2 < \sigma_{d+1}^2(\mathbf{B}) - \sigma_1^2(\mathbf{B}_{d+1}) \quad \forall \mathbf{x} : \|\mathbf{x}\|_2^2 = 1$$

or vice versa, then again either $\mathcal{L}(\mathbf{A}, \mathbf{V}^*) < \mathcal{L}(\mathbf{B}, \mathbf{V}^*)$ or $\mathcal{L}(\mathbf{A}, \mathbf{V}^*) > \mathcal{L}(\mathbf{B}, \mathbf{V}^*)$.

B. Derivation of the Dual of Problem 1

The dual objective is obtained as:

$$g(\boldsymbol{\mu}, \lambda) = \inf_{\mathbf{v}, z} \mathcal{H}(\mathbf{v}, z, \boldsymbol{\mu}, \lambda).$$

First, notice that, grouping the terms containing z together we can see that the coefficient of z is $1 - \sum_{i \in \mathcal{G}} \mu_i$. The infimum of $\mathcal{H}$ involves taking the derivative of $\mathcal{H}$ with respect to z and setting to zero.

$$\frac{\partial \mathcal{H}}{\partial z} = 0 \implies \sum_{i \in \mathcal{G}} \mu_i = 1$$

Since its coefficient is zero, we can effectively delete z from the lagrangian without changing the optimal solution. However, the infimum of the lagrangian w.r.t. $\mathbf{v}$ is particularly interesting. Rearranging the terms, we observe that the infimum involves the quadratic form: $\mathbf{v}^\top(\sum_{i \in \mathcal{G}} -\mu_i(\mathbf{A}^i)^\top\mathbf{A}^i + \lambda\mathbf{I})\mathbf{v}$. In general, the infimum of this expression is $-\infty$, unless the matrix $(\sum_{i \in \mathcal{G}} -\mu_i(\mathbf{A}^i)^\top\mathbf{A}^i + \lambda\mathbf{I})$ is positive semi-definite. We set $\mathbf{A}(\boldsymbol{\mu}) = \sum_{i \in \mathcal{G}} \mu_i(\mathbf{A}^i)^\top\mathbf{A}^i$ and thus equivalently we write:

$$-\mathbf{A}(\boldsymbol{\mu}) + \lambda\mathbf{I} \succeq 0$$

We observe that the matrix $\mathbf{A}(\boldsymbol{\mu})$ is a convex combination (since $0 \leq \mu_i$ and $\sum_i \mu_i = 1$) of positive semidefinite matrices, thus its negation is negative semidefinite. It follows that the primal minimization problem is bounded from below when $\lambda = \lambda_{max}(\mathbf{A}(\boldsymbol{\mu}))$. We define $\mathbf{s} = [\sigma_1^2(\mathbf{A}^1), \ldots, \sigma_1^2(\mathbf{A}^k)]$. Putting everything together, we obtain the dual problem:

$$\begin{aligned} \max_{\boldsymbol{\mu} \in \mathbb{R}^k} \ & \boldsymbol{\mu}^\top\mathbf{s} - \lambda_{max}(\mathbf{A}(\boldsymbol{\mu})) \\ \text{s.t. } & \mathbf{1}^\top\boldsymbol{\mu} = 1 \qquad (7) \\ & \boldsymbol{\mu} \geq 0. \qquad (8) \end{aligned}$$

C. SDP

Algorithm 3 contains the pseudocode of SDP to solve Problem 1.

D. Additional Experiment Results

In this section, we present additional experiments.

Algorithm 3 MG-SINGULARVECTOR-SDP

1: **Input:** Matrices $[\mathbf{A}^1, \ldots, \mathbf{A}^k]$
2: $\mathbf{X} \in \mathbb{R}^{n \times n} \leftarrow$ **Solve**:

$$\begin{aligned} \min_{z \in \mathbb{R}} \quad & z \\ \text{s.t. } & \sigma_1^2(\mathbf{A}^i) - \text{Tr}(\mathbf{A}^i \mathbf{X}) \leq z \quad \text{for} \quad \mathbf{A}^i \in \mathcal{G} \\ & \begin{bmatrix} \mathbf{X} & \mathbf{x} \\ \mathbf{x}^\top & 1 \end{bmatrix} \succeq 0 \quad ,\text{Tr}(\mathbf{X}) \leq 1, \end{aligned} \tag{9}$$

3: $\mathbf{X} = \sum_{j=1}^{n} \lambda_j \mathbf{x}_j \mathbf{x}_j^\top$
4: **Output:** $\mathbf{x}_1 \in \mathbb{R}^n$

D.1. Two groups

Figure 3 shows the different metrics being monitored in our experiments (i.e., the marginal loss, the incremental loss and the reconstruction loss) as a function of reconstruction (target) rank in all considered two-group datasets except the compas dataset, for which results are provided in Figure 2 in the main text.

The findings of the experiments presented in Figure 3 largely corroborate the findings presented in the main text (Figure 2) for the compas dataset.

D.2. More than two groups

Figure 4 displays marginal, incremental and reconstruction loss by rank in the communities-4 dataset partitioned into four groups. Again, the results for the communities-4 are consistent with and confirm the results seen in in Figure 2 for the compas-3 dataset.

Empirical Duality gap. In the case of more than two groups, the proposed methods are heuristics as they are not guaranteed to retrieve an optimal solution. In particular, there can be a discrepancy between the optimum of the primal and the one of the dual. Such discrepancy is known as *duality gap*. We note that we can compare the value of the dual objective g, at the obtained solutions for Algorithms 2 and 3, and also the primal objective for the corrsponing solution vector $\mathbf{v} = \mathbf{v}(\boldsymbol{\mu})$ from Algorithm 2 and $\mathbf{v} = \mathbf{x}_1$ from Algorithm 3, by computing $f = \max(h_1(\mathbf{v}), \ldots, h_k(\mathbf{v}))$. We call the difference $|f - g|$, *empirical duality gap*, as it gives us an empirical estimate of how far away from optimality are our solutions (a zero empirical duality gap means that the particular primal-dual solution pair is optimal).

In practice, as shown in Figure 5, such empirical duality gap is typically narrow. In particular, Figure 5 shows the value of the primal and dual objective in the compas-3 dataset with three groups, communities-4 dataset with four groups as well as in a synthetic dataset (gaussian-3) consisting of three groups, each of size 50×10 and with entries independent and identically distributed according to a standard Gaussian distribution. The difference between the primal and dual objective is generally limited and often negligible.

The results presented in Figure 5 are obtained by resorting to the Frank-Wolfe procedure to solve the dual problem, i.e., Algorithm 2.

The Frank-Wolfe algorithm is the algorithm of choice because of its simplicity and efficiency. However, solving MG-SINGULARVECTOR by the semidefinite programming relaxation (Algorithm 3) yields an even smaller duality gap, as demonstrated in Figure 6 for the same datasets considered in Figure 5.

E. Proofs

All the proofs of our results omitted from the main text due to space constraints are detailed in this section.

E.1. Proof of Property 1

Proof. We have that:

$$\|\mathbf{A}\mathbf{V}\mathbf{V}^\top\|_F^2 = \|\mathbf{A}\mathbf{v}_1\mathbf{v}_1^\top + \ldots + \mathbf{A}\mathbf{v}_d\mathbf{v}_d^\top\|_F^2.$$

The result follows from orthogonality, i.e., $\mathbf{v}_i^\top\mathbf{v}_j = 0$ for all $i, j \in [1, \ldots, d]$. This implies that:

$$\|\mathbf{A}\mathbf{v}_1\mathbf{v}_1^\top + \ldots + \mathbf{A}\mathbf{v}_d\mathbf{v}_d^\top\|_F^2 = \|\mathbf{A}\mathbf{v}_1\mathbf{v}_1^\top\| + \ldots + \|\mathbf{A}\mathbf{v}_d\mathbf{v}_d^\top\|_F^2 = \sum_{i=1}^{d}\|\mathbf{A}\mathbf{v}_i\mathbf{v}_i^\top\|_F^2$$

□

E.2. Proof of Orthonormalization Argument

Proof. Following an inductive argument (where the induction is on d), we can prove that $V = \{v_1, \ldots, v_d\}$ is indeed an orthonormal basis.

Base case. For $d = 1$, we can choose an arbitrary unit vector v_1. Note that v_1 is in the orthogonal complement of the subspace spanned by the $\mathbf{0}$. Since v_1 is a unit vector, it forms an orthormal basis of its span $\{v_1\}$.

Inductive hypothesis At step $k-1$, we have a $k-1$-dimensional orthonormal basis $V_{k-1} = \{v_1, \ldots, v_{k-1}\}$.

Inductive step At step k, we project the data onto the orthogonal complement of v_{k-1} and we select v_k in such subspace. The orthogonal complement of v_{k-1} is which also orthogonal to the space spanned by v_{k-2}, and so on. Thus, v_k is orthogonal to all vectors v_j for $j < k$ and $V = \{v_1, \ldots, v_k\}$ must be an orthonormal basis, which completes the proof.

□

E.3. Proof of Theorem 7.1

Proof. Since we are in the case $|\mathcal{G}| = 2$, we can consider a simplified formulation. We notice that $\mu_2 = 1 - \mu_1$ and set $\mu_1 = \mu$ and $\mu_2 = 1 - \mu$. We also set $\mathbf{A}^1 = \mathbf{A}$, $\mathbf{A}^2 = \mathbf{B}$ and $\mathbf{C}(\mu) = \mu\mathbf{A}^\top\mathbf{A} + (1-\mu)\mathbf{B}^\top\mathbf{B}$. Thus, Problem 2 becomes:

$$\max_{\mu\in\mathbb{R}} \mu s_1 + (1-\mu)s_2 - \lambda_{max}(\mathbf{C}(\mu)), \quad \mu \in [0, 1]. \tag{10}$$

We can now perform the standard KKT analysis. The dual lagrangian is:

$$\mathcal{H}_D(\mu, \xi_1, \xi_2) = g(\mu) + \xi_1\mu + \xi_2(1-\mu).$$

The stationarity condition is:

$$\frac{\partial}{\partial\mu}\mathcal{H}_D(\mu^*, \xi_1, \xi_2) = \frac{\partial}{\partial\mu}g(\mu^*) + \xi_1 - \xi_2 = 0.$$

Additionally, the complementary slackness condition requires that $\xi_1\mu = 0$ and $\xi_2(1-\mu) = 0$. To see this, first recall the duality between MG-SINGULARVECTOR and MG-SINGULARVECTOR-DUAL, from which we know that $\mu_1 = \mu$ and $\mu_2 = 1 - \mu$ are the associated multipliers with constraints $h_A - z$ and $h_B - z$ of MG-SINGULARVECTOR. From Theorem 5.1 we know that $h_A - z = 0$ and $h_B - z = 0$ and thus from complementary slackness we can infer that μ can be neither 0 or 1. Similarly, complementary slackness between μ and ξ_1 and ξ_2 indicates that $\xi_1 = \xi_2 = 0$.

Thus, stationarity simply reduces to $\frac{\partial}{\partial\mu}g(\mu^*) = 0$. From this and using equation 6, it follows that:

$$s_1 - s_2 - \mathbf{v}^\top(\mu^*)(\mathbf{A}^\top\mathbf{A} - \mathbf{B}^\top\mathbf{B})\mathbf{v}(\mu^*) = 0. \tag{11}$$

Therefore, $\mathbf{v}(\mu^*)$ leads to equal loss between the two groups. Additionally, this stationary point is a global maximum of g. To see this, we take the second derivative of g:

$$\frac{\partial^2 g}{\partial\mu^2} = -\frac{\partial^2}{\partial\mu^2}\lambda_{max}(\mathbf{C}(\mu)).$$

The Hadamard second variation formula (Schiffer, 1946), gives us an analytical expression for the second derivative of λ_{max}:

$$\frac{\partial^2}{\partial \mu^2}\lambda_{max}(\mathbf{C}(\mu)) = \\ \mathbf{v}(\mu)^\top \frac{\partial^2 \mathbf{C}(\mu)}{\partial \mu^2}\mathbf{v}(\mu) + 2\sum_{j\neq max}\frac{|\mathbf{v}(\mu)^\top \frac{\partial \mathbf{C}(\mu)}{\partial \mu}\mathbf{v}_j(\mu)|}{\lambda_{max}-\lambda_j(\mu)}. \tag{12}$$

where $\lambda_j, \mathbf{v}_j$ are eigenvalue-eigenvector pairs corresponding to smaller eigenvalues. The first term of Equation 12 vanishes ($\mathbf{C}(\mu)$ is only linearly dependent on μ), while the numerator and denominator in the second term are trivially positive (since $\mathbf{C}(\mu)$ is positive semidefinite and $\lambda_{max} > \lambda_j$. An important thing to note is that we have assumed simple spectrum. From this we can conclude that $\frac{\partial^2 g}{\partial \mu^2} < 0$, i.e., the function is concave, and thus has a unique maximum, at μ^*. At μ^*, we have that:

$$\begin{aligned} g(\mu^*) &= s_1 - \mathbf{v}(\boldsymbol{\mu}^*)^\top \mathbf{A}^\top \mathbf{A}\mathbf{v}(\boldsymbol{\mu}^*) \\ &= s_2 - \mathbf{v}(\boldsymbol{\mu}^*)^\top \mathbf{B}^\top \mathbf{B}\mathbf{v}(\boldsymbol{\mu}^*). \end{aligned}$$

As $\mathbf{v}(\boldsymbol{\mu}^*)$ is also a feasible point of Problem 1, with some value $\overline{z}$, we have that $g(\mu^*) = \overline{z}$ and since the primal is always lower bounded by the dual, we conclude that strong duality holds. □

Lemma E.1. *Define $q(\mu) = s_1 - s_2 - \mathbf{v}^\top(\mu)(\mathbf{A}^\top\mathbf{A} - \mathbf{B}^\top\mathbf{B})\mathbf{v}(\mu)$. Then, μ^* is a root of $q(\mu)$ and additionally $q(\mu)$ is monotone with respect to μ*

The fact that μ^* is a root of $q(\mu)$ follows directly from Equation 11. The monotonicity follows from $\frac{\partial q}{\partial \mu} = -\frac{\partial^2}{\partial \mu^2}\lambda_{max}(\mathbf{C}(\mu)) > 0$. This has an interesting consequence for the problem under investigation when $|\mathcal{G}| = 2$. The fact that a unique root exists in $\mu \in (0, 1)$ and the monotonicity mean that we can resort to a root-finding algorithm (such as Brent's method (Brent, 1971) or the bisection method (Ehiwario & Aghamie, 2014)) to locate the optimal μ^*. In fact, as we show in the experiments, such an algorithm is highly effective for MG-SINGULARVECTOR, when $|\mathcal{G}| = 2$. By default, we use the aforementioned Brent's method for finding the unique root $\mu \in (0, 1)$.

Note that a similar approach based on root-finding algorithms cannot be applied to the case of more than two groups and there is no obvious way to extend this approach to the general case.

E.4. Proof of Lemma 7.2

Proof. Using a Schur complement (Boyd & Vandenberghe, 2004), we can rewrite Problem 2 as:

$$\begin{aligned} \max_{\boldsymbol{\mu}\in\mathbb{R}^k}\ & \gamma \\ \text{s.t. } & \begin{bmatrix} -\mathbf{A}(\boldsymbol{\mu}) + \lambda\mathbf{I} & 0 \\ 0 & \boldsymbol{\mu}^\top\mathbf{s} - \gamma \end{bmatrix} \succeq 0 \\ & \mathbf{1}^\top\boldsymbol{\mu} = 1 \\ & \boldsymbol{\mu} \geq 0. \end{aligned}$$

To complete the proof, it suffices to notice that the SDP relaxation illustrated in Algorithm 3 is the dual problem to this problem (with dual variable $\mathbf{X}$). From our previous duality results it follows that strong duality exists between these two SDPs. Then, we can conclude that the SDP in Algorithm 3 solves Problem 1 to optimality. □

E.5. Proof of Lemma 7.3

Proof. Observe that $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is a matrix with orthonormal columns since it is constructed using Algorithm 1. Hence, we can invoke Property 1 along with Theorem 5.1 to obtain the result. Namely, after running Algorithm 1, we obtain $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$, which gives a total error of $\sum_{i=1}^d \mathcal{L}(\mathbf{A}, \mathbf{v}_i)$ for group A and a total error of $\sum_{i=1}^d \mathcal{L}(\mathbf{B}, \mathbf{v}_i)$ for group B. We know that $\mathcal{L}(\mathbf{A}, \mathbf{v}_i) = \mathcal{L}(\mathbf{B}, \mathbf{v}_i)$ for any $i \in \{1, \ldots, d\}$ due to Theorem 5.1. The lemma then follows.

As for time complexity, it suffices to consider that the optimal rank-1 solutions of MG-SINGULARVECTOR for two groups can be obtained in polynomial time $\mathcal{O}(\ell)$, as stated in Theorem 7.1. Then Property 1 implies that we need total time $\mathcal{O}(d\ell)$ to obtain an optimal solution. □

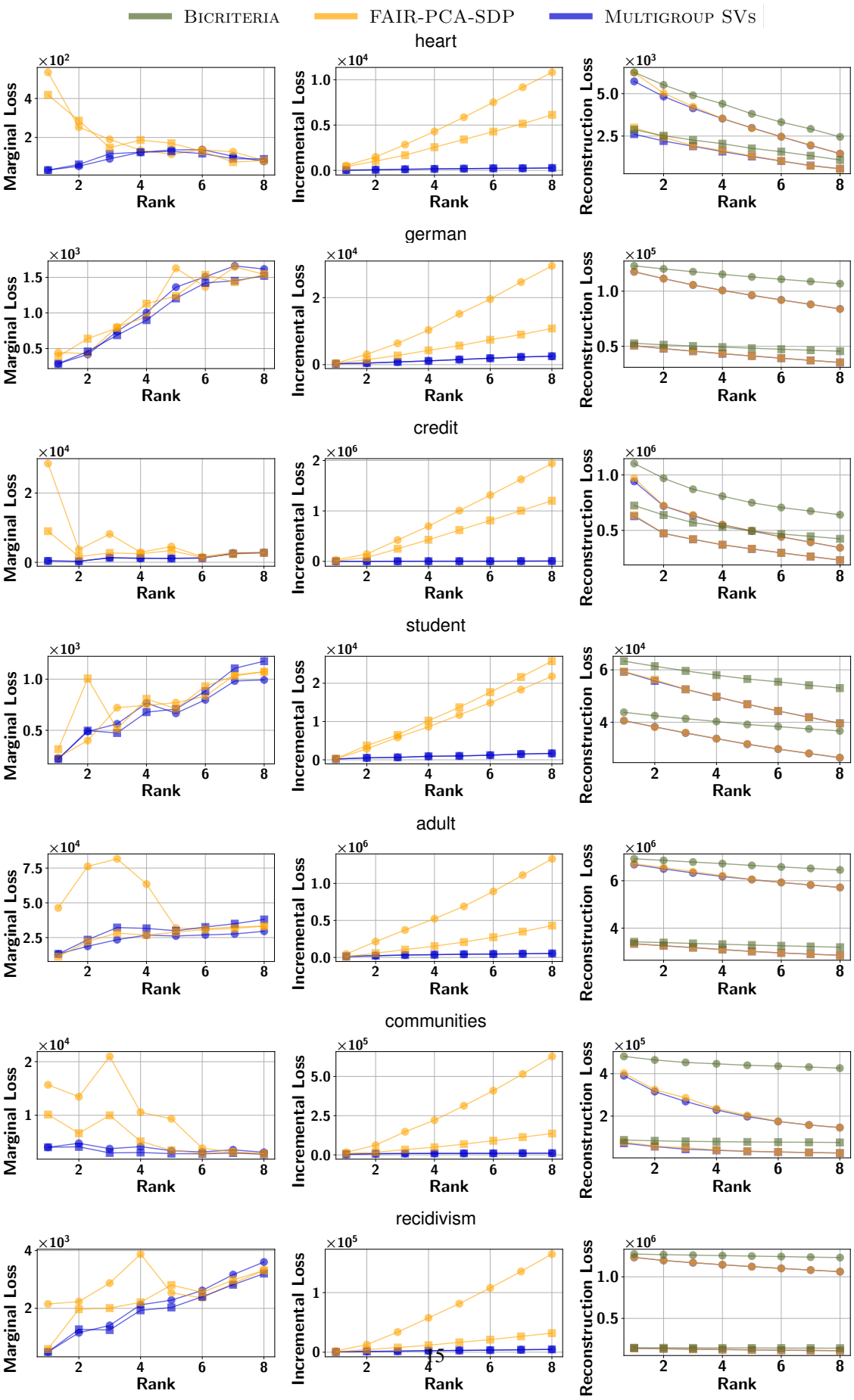


Figure 3: Real-world datasets with two groups. Marginal, incremental and reconstruction loss by rank. Different marker symbols indicate different groups.

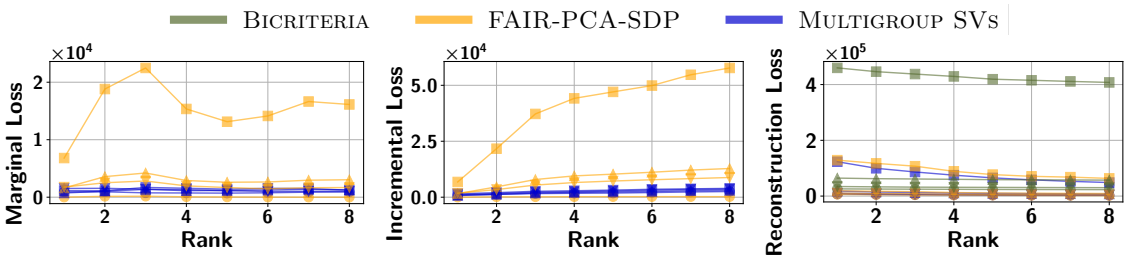


Figure 4: communities-4 dataset with four groups. Marginal, incremental and reconstruction loss by rank. Different marker signs indicate different groups.

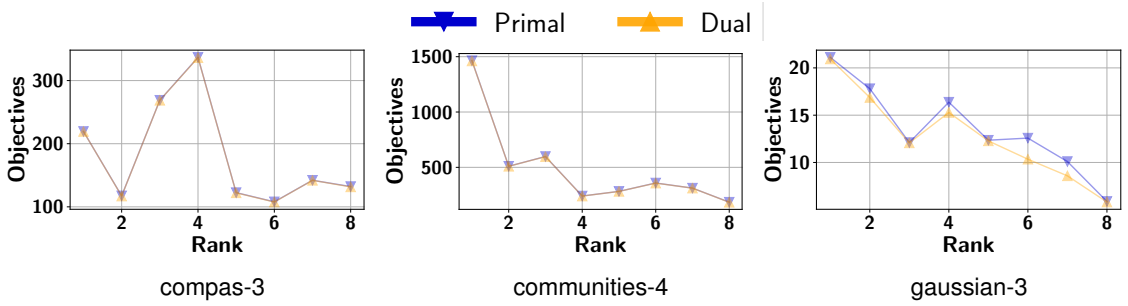


Figure 5: Real-world and syntethic data. Primal and dual optimal objective values as a function of rank for the solution relying on the Frank-Wolfe algorithm.

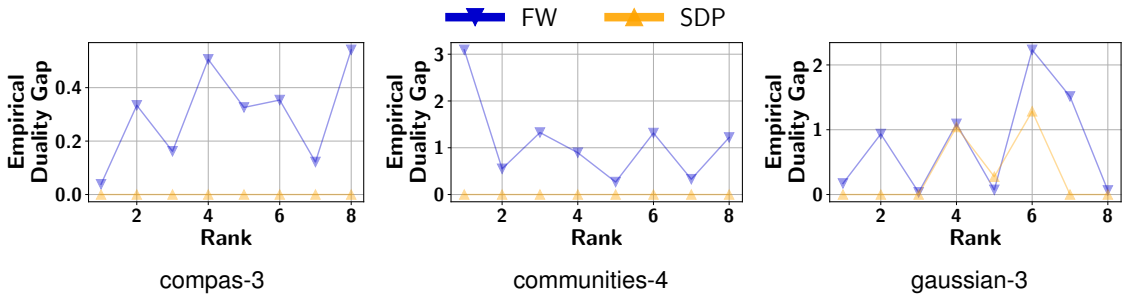


Figure 6: Real-world and synthetic data. Duality gap as a function of rank for the solutions relying on the Frank-Wolfe (FW) and semidefinite programming solver (SDP).