Fair Soft Clustering

Rune D. Kjærsgaard[*,1] **Pekka Parviainen**[2] **Saket Saurabh**[2,3]

Madhumita Kundu[2] **Line K. H. Clemmensen**[1]

[1]DTU Compute, Technical University of Denmark, Denmark
[2]Department of Informatics, University of Bergen, Norway
[3]Theoretical Computer Science Group, The Institute of Mathematical Sciences, India
[*]Correspondence to: rdokj@dtu.dk

Abstract

Scholars in the machine learning community have recently focused on analyzing the fairness of learning models, including clustering algorithms. In this work we study fair clustering in a probabilistic (soft) setting, where observations may belong to several clusters determined by probabilities. We introduce new probabilistic fairness metrics, which generalize and extend existing non-probabilistic fairness frameworks and propose an algorithm for obtaining a fair probabilistic cluster solution from a data representation known as a fairlet decomposition. Finally, we demonstrate our proposed fairness metrics and algorithm by constructing a fair Gaussian mixture model on three real-world datasets. We achieve this by identifying balanced micro-clusters which minimize the distances induced by the model, and on which traditional clustering can be performed while ensuring the fairness of the solution.

1 INTRODUCTION

Decision making systems based on machine learning (ML) applications have demonstrated unwanted consequences as a result of biased data (Phillips et al., 2011; Z. Obermeyer and Mullainan, 2019). This has fostered efforts towards artificial intelligence (AI) alignment, wherein ML systems are aligned with their intended objectives. This includes ensuring decisions are fair and do not show bias against or for certain population sub-groups. Many of these fairness interventions are based on the Disparate Impact (DI) doctrine (Rutherglen, 1987), which prohibits discrimination between different groups of protected attributes such as race or sex. For clustering, this type of non-discrimination is denoted group-level fairness (Chhabra et al., 2021).

Clustering algorithms are an unsupervised ML approach used to partition a dataspace into clusters. These algorithms are widely used, particularly in settings where data labels are scarce. Here, clustering may be used as a feature engineering tool to supplement points with cluster assignments in an effort to increase expressive power of downstream models. If the underlying training data is unfair, this may propagate into the generated features and ultimately cause biased predictions. Fair clustering aims to prevent this.

The topic of fairness for clustering was initiated in a seminal work by Chierichetti et al. (2017), which considered group-level fairness obtained by modifying the input data for traditional hard clustering algorithms like k-center and k-median. The literature on fair clustering is largely focused on such non-probabilistic algorithms, where point assignments are deterministic (Chhabra et al., 2021). However, for a number of applications soft clustering is more appropriate. In our work, we consider group-level fair clustering in a probabilistic setting, where equal representation is ensured for protected groups in clusters found using soft clustering algorithms. As an example, a bank might use a dataset containing information about educational attainment and wages of individuals to train a model with the goal of identifying potential customers and offering them loans or credit opportunities. The bank then trains a soft clustering algorithm to group customers into low or high risk candidates, where the soft assignments

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

could imply the probability (risk) that a given customer will default their loan. It should be pointed out, that a wage gap has been identified for women and people-of-color, who usually earn lower wages than White males (Patten, 2016), and that people-of-color often face additional adversities that lead to educational disparities as compared to White individuals (Sablich, 2016). Thus, a clustering algorithm trained on this data would be prone to group White males as better prospective candidates and correspondingly deny people-of-color and women the potential for improvement, thus propagating the systemic bias from the training data to downstream decisions. Ensuring group-level fairness from a probabilistic cluster solution could prevent such decision-making systems from adversely affecting specific groups and thus ensure that the models adhere to the DI doctrine.

Fair probabilistic clustering has previously been studied by Esmaeili et al. (2020), where they considered probabilistic fairness in a setting of imperfect group membership knowledge. This considers the protected group membership in a probabilistic setting while considering the cluster assignments in a deterministic setting. We on the other hand, consider the case of deterministic protected group membership and probabilistic cluster assignments. In Anderson et al. (2020) they consider individual-level fairness in a probabilistic setting, where no protected groups exist, and fairness is achieved by ensuring similar individuals are treated similarly by the algorithm. Corresponding probabilistic assignments have been studied by Brubach et al. (2020) and Brubach et al. (2021). In our work, we consider group-level fairness under the same conditions as Chierichetti et al. (2017), where protected groups exist and the goal is to construct a representation of the data, on which a traditional algorithm can be trained to obtain a balanced cluster solution. We generalize this to the probabilistic setting. No metrics for group-level fairness under probabilistic cluster assignments have been established (Chhabra et al., 2021). To this end, we propose probabilistic fairness metrics, which generalize current definitions for deterministic cluster assignment. Moreover, we demonstrate an algorithm for obtaining a fair cluster solution from a fairlet decomposition in the probabilistic setting. Finally, we demonstrate our metrics and algorithm by applying them on a fairlet decomposition constructed for a Gaussian mixture model (McLachlan and Basford, 1988).

Our contributions are:

- Probabilistic generalizations of metrics for group-level fairness.
- An algorithm for obtaining a fair probabilistic cluster solution from a fairlet decomposition.
- An approach for generating a fairlet decomposition for a GMM.

2 CLUSTER FAIRNESS

In most work, group-level fair clustering is defined in terms of balance or relaxations thereof but may also be defined in terms of entropy (Chhabra et al., 2021). These metrics measure the representation equality of protected groups described by an attribute vector $\boldsymbol{A}$. In this work, we denote protected groups by colors $p \in P$.

2.1 Deterministic Assignment Fairness

Balance measures algorithmic fairness of a cluster solution by considering the degree of balance between protected groups within each cluster. This fairness definition complies with the the DI doctrine, and the goal is to obtain a balanced representation (similar fraction) of all groups within all clusters.

Consider a set of points D partitioned into a set of clusters C. Balance may be measured by comparing two fractions $r_{D,p}$ and $r_{c,p}$, where $r_{D,p}$ is the fraction of a color p in D and $r_{c,p} = \frac{|N_{c,p}|}{n_c}$ is the fraction of a color p in cluster c, where n_c is the number of observations in cluster c, and $N_{c,p}$ is the set of observations in the dataset belonging to both color p and cluster c. Now construct a fraction $R_{c,p} = \frac{r_{D,p}}{r_{c,p}}$ and define the balance by:

$$B = \min_{c \in C, p \in P} \min \left(R_{c,p}, \frac{1}{R_{c,p}} \right), \tag{1}$$

where B is the balance and $R_{c,p} = \frac{r_{D,p}}{r_{c,p}}$ is a fraction for a given cluster c and color p (Chhabra et al., 2021). Balance is bounded in $B \in [0, 1]$ with higher balance being more fair. This metric measures the overall fairness of the cluster solution through the minimum balance across all clusters $c \in C$ and colors $p \in P$. Optimal balance ($B = 1$) is found when all clusters share the same color fraction $r_{c,p} = r_{D,p}\ \forall c, p$, while worst case balance ($B = 0$) is found when a cluster contains no members of a protected group $r_{c,p} = 0$.

Contrary to the balance metric, entropy does not measure the worst case fairness of all clusters, but rather quantifies the overall fairness through an information-theoretic perspective across all clusters simultaneously:

$$H = \min_{p \in P} \left(-\sum_{c=1}^{C} r_{c,p} \log r_{c,p} \right), \tag{2}$$

where H is the entropy (Chhabra et al., 2021). The entropy fairness is the level of information entropy across all clusters. Higher entropy equates to a more

fair cluster solution. Optimal entropy fairness is found when all clusters share the same color fraction $r_{c,p}$ while worst case entropy fairness is found when all clusters are monochromatic.

2.2 Probabilistic Assignment Fairness

In soft clustering algorithms the point assignments are probabilistic and determined by a responsibility vector $\boldsymbol{\gamma}_c$ for each cluster c. The entries $\gamma_{i,c}$ in this vector describe the probability that the i^{th} data point is generated by component c. We use the responsibilities to construct a measure for weighted color contribution:

$$w_{c,p} = \frac{\sum_{i=1}^{N} \gamma_{i,c}\alpha_{i,p}}{\sum_{i=1}^{N} \gamma_{i,c}}, \tag{3}$$

where $w_{c,p}$ is the weighted contribution of color p to cluster c, $\gamma_{i,c}$ is the i^{th} entry in the responsibility vector $\boldsymbol{\gamma_c}$ and $\alpha_{i,p}$ is the i^{th} entry in a color vector $\boldsymbol{\alpha}_p$ constructed by setting $\alpha_{i,p} = 1$ if observation $\boldsymbol{x}_i \in p$ and 0 otherwise. The numerator represents the total color mass (weighted color contribution) for the given cluster, while the denominator represents the total mass of the cluster. Note that the weighted color contribution $w_{c,p}$ reduces to $r_{c,p}$ if $\boldsymbol{\gamma_c}$ dictates hard assignments (probabilities either 1 or 0). Thus $w_{c,p}$ generalizes the unweighted color contribution $r_{c,p}$ from the deterministic assignment setting.

We propose to substitute the weighted color contribution $w_{c,p}$ into the established fairness frameworks for deterministic assignment fairness in Eqs. 1 and 2:

$$B_{\text{soft}} = \min_{c\in C, p\in P} \min\left(W_{c,p}, \frac{1}{W_{c,p}}\right), \tag{4}$$

where B_{soft} is the soft assignment balance and $W_{c,p} = \frac{r_{D,p}}{w_{c,p}}$.

Equivalently we define the soft assignment entropy fairness by:

$$H_{\text{soft}} = \min_{p\in P}\left(-\sum_{c=1}^{C} w_{c,p}\ \log\ w_{c,p}\right), \tag{5}$$

2.3 Entropy Ratio

Unlike balance, entropy is not bounded in $H \in [0,1]$, but we can normalize it by comparing the information entropy of the cluster solution to the optimal entropy of the cluster configuration:

$$H_{\text{ratio}} = H_{\text{soft}}/H_{\text{OPT}}, \tag{6}$$

where H_{OPT} is a cluster solution with optimal (largest) entropy under the given number of clusters. H_{OPT} is found when all clusters share the same color fraction. Thus $H_{\text{ratio}} \in [0,1]$, where $H_{\text{ratio}} = 1$ when $w_{c,p} = r_{D,p}\ \forall c,p$ and $H_{\text{ratio}} = 0$ when all clusters are monochromatic with respect to color p.

3 OBTAINING FAIR CLUSTERS

The first part of our contribution relates to defining what fairness entails in the soft setting, and similarly how we may measure this. This is a key issue and prerequisite for the future study of fair soft clustering in general, as noted in Chhabra et al. (2021). The second part of our contribution relates to constructing fair solutions in terms of these definitions.

Standard cluster algorithms optimize an objective function and ignore the distribution of protected attributes. This may end up propagating inherent bias from the training data to the final model solution. To avoid this, the data can be modified by constructing a balanced representation. Fair cluster solutions can be found by generating a fair representation through a fairlet decomposition and subsequently performing clustering with a traditional color blind algorithm on the decomposition. The decomposition is constructed by identifying micro-clusters, called fairlets, which preserve balance.

For a binary protected attribute consisting of two colors, a decomposition can be specified as a (p_1, p_2)-fairlet decomposition (assuming $p_1 < p_2$) with balance parameters p_1 and p_2 indicating that all fairlets have a color fraction $r_{c,p} \geq \frac{p_1}{p_1+p_2}$. For a perfectly balanced dataset ($N_{p_1} = N_{p_2}$) it is possible to obtain a $(1,1)$-fairlet decomposition, where each fairlet consists of exactly one point of each color. For this setting $r_{c,p} = r_{D,p}\,\forall c,p$ in the decomposition, which results in a balance of $B = 1$. To construct a fair clustering from the decomposition, centers are assigned for each fairlet and a traditional clustering is performed on the centers. Since the union of balanced micro-clusters is necessarily also balanced, this will ensure a fair clustering. This procedure has been constructed for deterministic assignments in the literature. We demonstrate an algorithm for constructing a fair probabilistic clustering from any fairlet decomposition by modifying existing framework for deterministic clustering (Backurs et al., 2019). This is shown in Algorithm 1. The fair cluster solution is found by generating a fairlet decomposition, applying a traditional soft clustering algorithm on the fairlet centers and subsequently assigning appropriate responsibilities to the fairlet members. Algorithm 1 provides the same theoretical fairness bounds as previous works in the hard assignment setting (see Appendix A for details).

Algorithm 1 SOFTCLUSTERFAIRLET(Q)

Input: $Q = \{q_1, q_2, \cdots, q_\ell\}$ where every q_i is a fairlet with center c_i

Output: The algorithm returns a fair probabilistic clustering of D given a fairlet decomposition Q of D

multiset $\bar{D} \leftarrow \emptyset$ (initilization)
for all fairlets $q_i \in Q$ **do**
  $\bar{D} \leftarrow \bar{D} + \{|q_i|$ copies of $c_i\}$ (sum of two multisets)
end for
$C \leftarrow$ Traditional probabilistic clustering of $\bar{D}$
$C^* \leftarrow \gamma_i$ (assign each fairlet member the responsibility vector of its center in C)
return C^*

However, the algorithm does not ensure optimality of the decomposition and may result in a sub-optimal cost of the studied clustering objective depending on the spatial location of the points selected for each fairlet. To obtain a solution which maintains a fair representation of protected groups and simultaneously minimizes a clustering objective function, it is necessary to take the cost of the decomposition into account. Fairlet decompositions are tailored to specific objective functions like the k-median and k-means objective (Bercea et al., 2018):

$$\mathcal{L}_{\mathrm{k}}(D, Q) = \sum_{x \in D} d(\boldsymbol{x}, \beta_{\mathrm{Q}}(\boldsymbol{x})), \tag{7}$$

where $d(\cdot)$ is a metric (distance function) and $\beta_{\mathrm{Q}}(\boldsymbol{x})$ denotes the center location of the fairlet to which the data point $\boldsymbol{x}$ is mapped. For k-median clustering $\beta_{\mathrm{Q}}(\boldsymbol{x}) \in D$ and for k-means clustering $\beta_{\mathrm{Q}}(\boldsymbol{x}) \in \mathbb{R}^m$ for $D \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$. The distance metric in k-means is $d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||^2$.

Chierichetti et al. (2017) define the total cost of the overall fair clustering assignment from D to C^* (Lemma 6) as:

$$\mathcal{L}_{\mathrm{k-tot}}(D, C^*) = \mathcal{L}_{\mathrm{k}}(D, Q) + \mathcal{L}_{\mathrm{k}}(\bar{D}, C^*), \tag{8}$$

where $\mathcal{L}_{\mathrm{k}}(D, Q)$ is the fairlet decomposition cost, and $\mathcal{L}_{\mathrm{k}}(\bar{D}, C^*)$ is the cost on a transformed dataset $\bar{D}$, where for each fairlet q_i the fairlet center c_i appears $|q_i|$ times.

The cost on the transformed dataset $\mathcal{L}_{\mathrm{k}}(\bar{D}, C^*)$ is the sum of distances of each point in $\bar{D}$ to their assigned cluster center:

$$\mathcal{L}_{\mathrm{k}}(\bar{D}, C^*) = \sum_{\boldsymbol{x} \in \bar{D}} d(\boldsymbol{x}, \alpha_{C^*}(\boldsymbol{x})), \tag{9}$$

where $\alpha_{C^*}(\boldsymbol{x})$ is location of the center for which the data point $\boldsymbol{x}$ is mapped by the clustering C^*.

The goal of the fair clustering is to construct a fairlet decomposition which minimizes the cost in Eq. 8. Chierichetti et al. (2017) propose solving the problem by transforming it into a minimum cost flow (MCF) problem, where a directed graph is constructed. This graph may be modified to suit different cluster objective functions. To generate a decomposition, the weights on the edges between nodes are represented by a distance function between points. The objective is to minimize the sum of distances from fairlet members to fairlet centers. The MCF approach has super-quadratic time in dataset size and becomes computationally expensive for large datasets. Alternative scalable approaches have been introduced, where the optimal fairlet decomposition is approximated and found in nearly linear time in dataset size (Backurs et al., 2019). In our results we illustrate that the scalable k-median fairlet decomposition introduced by Backurs et al. (2019) can be fed as input to Algorithm 1 to produce a fair probabilistic clustering, which can be assessed by our proposed metrics in Eqs. 4 and 5. This clustering may however have sub-optimal cost.

3.1 Probabilistic Model Fairlet Decomposition

To demonstrate our proposed metrics and algorithm, we directly translate the fair cost defined by Chierichetti et al. (2017) and construct a fairlet decomposition to minimize this cost for a probabilistic model known as a Gaussian mixture model (GMM). A GMM describes the data distribution through a mixture of multivariate normal distributions $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$, covariance structure $\boldsymbol{\Sigma}$ and component weights $\boldsymbol{\pi}$. The distribution parameters can be inferred through the expectation maximization (EM) algorithm (Moon, 1996), which iterates between updating the parameters (maximization step) and computing the responsibility $\gamma_{i,c}$ for all i, c (expectation step) until the likelihood converges. The responsibility can be computed by:

$$\gamma_{i,c} = \frac{\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\pi_c}{\sum_{j=1}^{K} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j} \tag{10}$$

where $\gamma_{i,c}$ is the probability that the data point $\boldsymbol{x}_i$ is generated by component c. Note that the total mass of a mixture component is $N_c = \sum_{i=1}^{N} \gamma_{i,c}$ and that the sum of total component masses is the number of data points $N = \sum_{c=1}^{C} N_c$.

To construct a fairlet decomposition which is simultaneously fair and minimizes the distances between fairlet members in the space modelled by the GMM, we need a distance metric which takes into account the mixture model. The natural distance function for data modeled by a single multivariate Gaussian probability distribution $\mathcal{N}$ with covariance matrix $\boldsymbol{\Sigma}$ and mean $\boldsymbol{\mu}$

is the Mahalanobis distance:

$$d^2_{\mathrm{M}}(\boldsymbol{x}, \mathcal{N}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \tag{11}$$

where $d^2_{\mathrm{M}}(\boldsymbol{x}, \mathcal{N})$ is the squared Mahalanobis distance of a point $\boldsymbol{x}$ from the distribution $\mathcal{N}$.

The likelihood of a GMM is directly related to the Mahalanobis distance between observed points and presumed distributions. The log-likelihood of a data point belonging to a multivariate normal distribution is given by the logarithm of the probability density function of distribution $\mathcal{N}$:

$$\log L(\boldsymbol{x}) = -\frac{1}{2}\left[\log(|\boldsymbol{\Sigma}|) + \log(d^2_{\mathrm{M}}(\boldsymbol{x}, \mathcal{N})) + m \cdot \log(2\pi)\right], \tag{12}$$

where m is the multivariate dimension of $\mathcal{N}$.

When the data are modelled by a mixture of multiple Gaussians the covariance matrix $\boldsymbol{\Sigma}$ is not unique. To extend the notion of distance between points to this setting, the data space can be interpreted as a Riemannian manifold with metric $\boldsymbol{G}(\boldsymbol{x})$. This metric can be approximated leading to a model-weighted distance (MWD) (Tipping, 1999):

$$d^2_{\mathrm{MWD}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{G} (\boldsymbol{x}_i - \boldsymbol{x}_j), \tag{13}$$

where $d^2_{\mathrm{MWD}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the model-weighted distance between points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and $\boldsymbol{G}$ is given by:

$$\boldsymbol{G} = \frac{\sum_{k=1}^{K} \boldsymbol{\Sigma}_k^{-1} \pi_k \int_{x_i}^{x_j} p(\boldsymbol{x}|k) d\boldsymbol{x}}{\sum_{k=1}^{K} \pi_k \int_{x_i}^{x_j} p(\boldsymbol{x}|k) d\boldsymbol{x}}, \tag{14}$$

where π_k is the mixing proportion of the k^{th} mixture component and $\int_{x_i}^{x_j} p(\boldsymbol{x}|k)\ d\boldsymbol{x}$ is the unidimensional integral of the probability density of the k^{th} component along the straight path between point $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Computing the distance in this manner assumes a constant metric $\boldsymbol{G}$ along the path between the points. This metric can be interpreted as a probabilistically-weighted average of the inverse covariances of the different components in the mixture model. The integral is analytically tractable and is given by:

$$\int_{x_i}^{x_j} p(\boldsymbol{x}|k)\ d\boldsymbol{x} = \sqrt{\frac{\pi b^2}{2}} e^{-Z/2} \times \left[\operatorname{erf}\left(\frac{1-a}{\sqrt{2b^2}}\right) - \operatorname{erf}\left(\frac{-a}{\sqrt{2b^2}}\right)\right], \tag{15}$$

where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ is the error function and

$$b^2 = (\boldsymbol{v}^\intercal \boldsymbol{\Sigma}_k^{-1} \boldsymbol{v})^{-1}, \tag{16}$$

$$a = b^2 \boldsymbol{v}^\intercal \boldsymbol{\Sigma}_k^{-1} \boldsymbol{u}, \tag{17}$$

$$Z = \boldsymbol{u}^\intercal \boldsymbol{\Sigma}_k^{-1} \boldsymbol{u} - b^2 (\boldsymbol{v}^\intercal \boldsymbol{\Sigma}_k^{-1} \boldsymbol{u})^2, \tag{18}$$

with $\boldsymbol{u} = \boldsymbol{\mu}_k - \boldsymbol{x}_j$ and $\boldsymbol{v} = \boldsymbol{x}_i - \boldsymbol{x}_j$.

Equipped with a metric for computing distances between points we define the GMM fairlet decomposition cost by:

$$\mathcal{L}_{\mathrm{GMM}}(D, Q) = \sum_{\boldsymbol{x} \in D} d_{\mathrm{MWD}}(\boldsymbol{x}, \beta_{\mathrm{Q}}(\boldsymbol{x})), \tag{19}$$

where the metric $\boldsymbol{G}$ describing data manifold is computed from a GMM on the original dataspace D. Similarly we define the GMM cost on the transformed dataset $\bar{D}$ as:

$$\mathcal{L}_{\mathrm{GMM}}(\bar{D}, C^*) = \sum_{\boldsymbol{x} \in \bar{D}} d_{\mathrm{MWD}}(\boldsymbol{x}, \Gamma_{C^*}(\boldsymbol{x})), \tag{20}$$

where $\Gamma_{C^*}(\boldsymbol{x})$ denotes the mean locations $\boldsymbol{\mu}$ of the components to which $\boldsymbol{x}$ is mapped. We restrict the distance of the k^{th} mixture component to be based on a $\boldsymbol{G}$ metric for the k^{th} component and it thus reduces to a weighted sum of Mahalanobis distances in the transformed dataspace $\bar{D}$, where the weights are dictated by the component responsibilities. This choice gives a more robust cost measure of the GMM fit. The direct translation of the total cost of the fair solution defined by Chierichetti et al. (2017) is then:

$$\mathcal{L}_{\mathrm{GMM-tot}}(D, C^*) = \mathcal{L}_{\mathrm{GMM}}(D, Q) + \mathcal{L}_{\mathrm{GMM}}(\bar{D}, C^*) \tag{21}$$

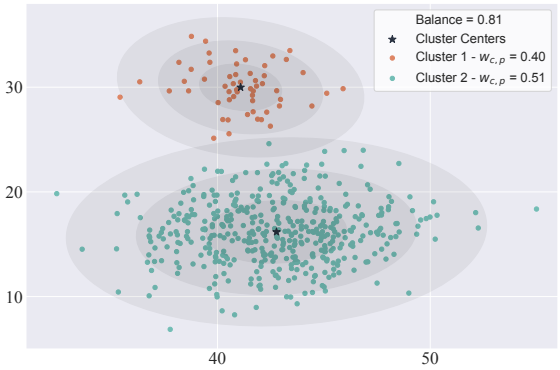
We generate a GMM fairlet decomposition by minimizing the GMM cost through a MCF algorithm[1]. We utilize the approach described in Chierichetti et al. (2017) for the k-median cost and change the weights on the edges of the graph to the MWD between points. Prior to running the algorithm the metric space is instantiated by fitting a traditional GMM with the desired number of components on the original data. The distribution parameters of these components are then used to generate the metric $\boldsymbol{G}$ and compute the model-weighted distances. The fairlet centers are then generated as the mean of the members in each fairlet.

Fig. 1 presents a visualisation of the approach on simulated data of 500 points in $\mathbb{R}^2$ with 250 red and 250 blue points. Fig 1(a) illustrates the original data with points colored according to their protected attribute. We apply a traditional GMM on the data to obtain a color blind solution shown in Fig. 1(c).

[1] Our code is publicly available at `https://github.com/RuneDK93/fair-soft-clustering`

(a) Original data with colors.

(b) Fairlet decomposition.



(c) Traditional GMM.

(d) Fair GMM.

Figure 1: Illustration of our approach on simulated data of 500 points in $\mathbb{R}^2$. The original data points are shown colored according to their protected attribute in (a). A GMM fit on the original data is shown in (c). (b) shows a (1,1)-fairlet MWD decomposition of the data, while (d) shows the resulting fair solution from fitting a GMM on the fairlet centers in (b) and mapping the responsibilities $\gamma_{i,c}$ according to Algorithm 1. The points in (c) and (d) are colored according to the cluster index in $\gamma_{i,c}$ with highest probability. The weighted cluster color fractions $w_{c,p}$ in (c) and (d) are shown for the red class. Notice that the resulting balance is $B = 0.81$ for the traditional GMM fit in (c) and $B = 1.00$ for the fair GMM fit in (d).

The balance of red and blue points allows us to construct a $(1, 1)$-fairlet decomposition of the data through a perfect matching on the bichromatic graph. We construct the decomposition using the MCF approach by utilizing the distribution parameters of the colorblind solution to instantiate the $\boldsymbol{G}$ metric and use these distances on the edges of the graph. This results in a MWD fairlet decomposition Q of the data illustrated in Fig. 1(b). The fairlet decomposition is then fed as input to Algorithm 1 to obtain the final fair clustering C^* shown in Fig. 1(d). The fairness of both solutions is assessed with our proposed soft balance fairness metric (Eq. 4).

4 RESULTS

We demonstrate our approach on real-world data by performing experiments on three widely used datasets in the fair clustering community. The datasets are Census[2], Bank[3] and Diabetes[4]. We select numerical features for the dimensions in the data point space and use 'sex' and 'marital status' as protected attributes. See Tab. 1 for an overview of the datasets.

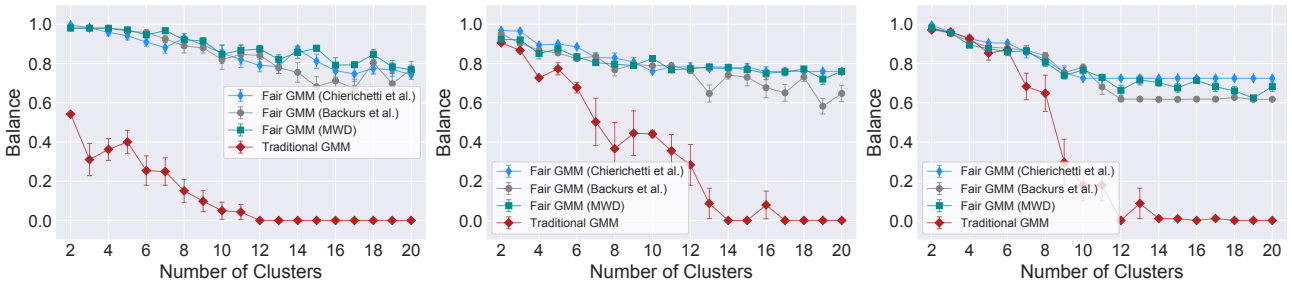
Census: The dataset collects records of the 1994 US Census and presents an income prediction task based on various attributes of individuals. We select 'age', fnlwgt', 'education-num', 'capital-gain' and 'hours-per-week' as features representing the spacial dimensions of the data. We select 'sex' as the protected attribute.

Bank: The dataset (Moro et al., 2014) is from a Por-

[2] https://archive.ics.uci.edu/ml/datasets/adult

[3] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[4] https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008



(a) Census (b) Bank (c) Diabetes

Figure 2: Fairness in terms of soft balance on the three datasets. Data points are the mean values from five iterations of different random seeds. Error bars indicate standard errors. The traditional GMM cluster approach obtains monochromatic cluster solutions ($B = 0$) for all datasets while the balance of the fair solutions are bounded. For instance the fair solutions of the Diabetes dataset are bounded at balance $B \geq 0.62$ (see Appendix A for details).

tuguese phone call based bank marketing campaign. We select 'age', balance' and 'duration-of-account' as features representing the spacial dimensions of the data. We select 'marital-status' as the protected attribute.

Diabetes: The dataset (Strack et al., 2014) spans 10 years of information and outcomes of diabetes across 130 US hospitals. We select 'age' and 'time-in-hospital' as features representing the spacial dimensions of the data and select 'sex' as the protected attribute.

Table 1: Overview of the datasets used for our experiments. The table shows the number of spacial dimensions, the type of protected attribute and the color fraction for the three datasets.

Dataset	Dimension	Protected Att.	$r_{D,p}$
Census	5	sex	0.67
Bank	3	marital-status	0.62
Diabetes	2	sex	0.54

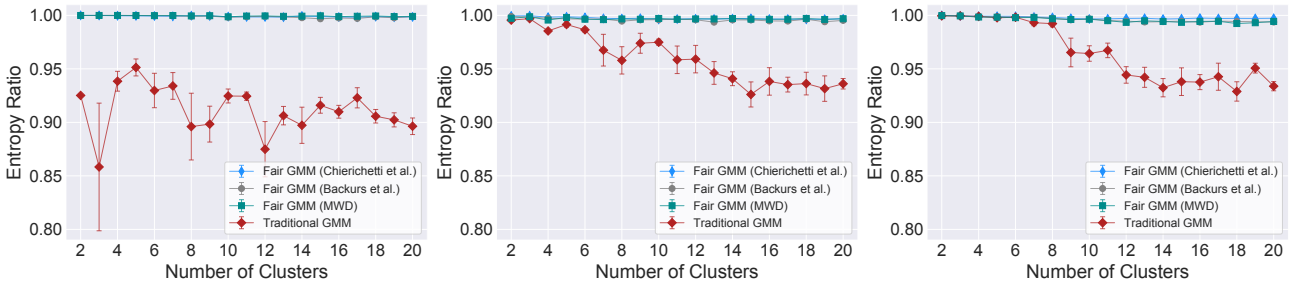
Similarly to Chierichetti et al. (2017) we sub-sample each dataset to 500 observations and preserve the protected attribute fraction from the original data. These fractions are $r_{D,p} = 0.67$ (Census), $r_{D,p} = 0.62$ (Bank) and $r_{D,p} = 0.54$ (Diabetes).

For each dataset we apply a standard GMM on the original data and compare this to our proposed fair probabilistic clustering constructed by first finding a MWD (1,2)-fairlet decomposition in a MCF setting and then applying Algorithm 1 on the decomposition. The final clustering outcome is dependent on the initialization of the GMM. We initialize the GMM using k-means clustering and repeat the overall clustering and fairlet decomposition five times with different random seeds for the initialization parameters to generate mean values and associated standard errors.

We compare our proposed solution to two fair GMM baselines. We implement the first baseline using the k-median fairlet decomposition algorithm from Chierichetti et al. (2017) to construct a k-median fairlet decomposition. The fairlet centers from this decomposition are then assigned probabilities using the traditional GMM fit on the original data. We implement the second baseline by finding a scaleable k-median (1,2)-fairlet decomposition using the method described in Backurs et al. (2019). This fairlet decomposition is then fed as input to Algorithm 1 to construct a fair probabilistic clustering. These baselines provide the same theoretical fairness bounds as our proposed MWD solution but may incur a high cost by enjoining points that are far apart in the GMM space.

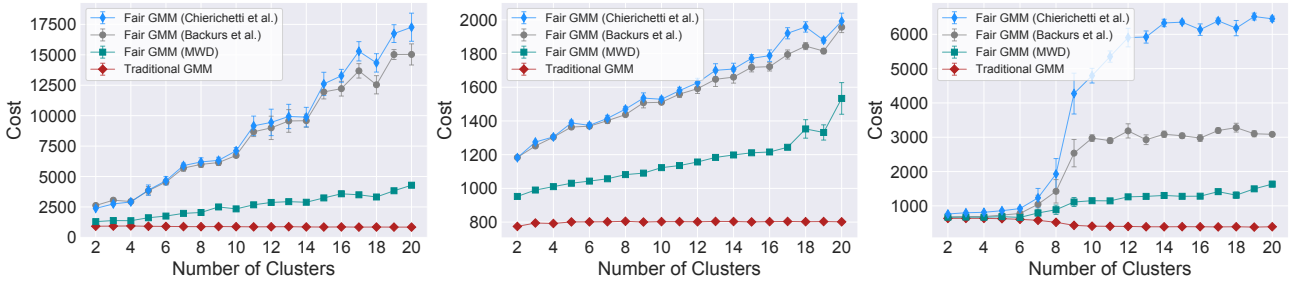
Fig. 2 shows the resulting soft balance according to Eq. 4 and Fig. 3 shows the soft entropy ratio according to Eq. 6. For all datasets the fairness disparity between the traditional and fair solutions increases sharply with the number of clusters. Observe that for a large number of mixture components, the colorblind model has a balance of zero for all datasets. The optimal GMM solution to the data thus requires monochromatic clusters. Additionally, the fair GMM shows less fairness variance and is thus more robust to the initialization. This is especially the case for entropy fairness, where the fair GMM is highly robust while the color blind GMM is much more sensitive to the initialization parameters. The fair baseline solutions generated from k-median decompositions and our MWD decomposition achieve equivalent fairness scores.

The price of fairness is quantified by the negative impact on the cost of the solutions. This is illustrated in Fig. 4, where the costs of the different cluster solutions



(a) Census (b) Bank (c) Diabetes

Figure 3: Fairness in terms of soft entropy ratio on the three datasets.



(a) Census (b) Bank (c) Diabetes

Figure 4: Total cost on the three datasets. The costs for the fair solutions are computed as $\mathcal{L}_{\text{GMM}-\text{tot}}(D, C^*)$ according to Eq. 21 while the cost of the standard GMM is computed according to Eq. 20 on the original data as $\mathcal{L}_{\text{GMM}}(D, C)$, where C is a traditional GMM clustering on D.

are shown for the three datasets. The costs for the fair solutions are computed by Eq. 21 while the cost of the traditional solution is computed by Eq. 20 on the original data as $\mathcal{L}_{\text{GMM}}(D, C)$, where C is a traditional GMM clustering on D. The traditional GMM solution has the lowest cost, while the fair MWD GMM has a larger cost, which increases with the number of cluster components. The fair baseline solutions from Euclidean k-median fairlet decompositions have the highest costs, which are significantly higher than both the traditional GMM and fair MWD GMM for large k. We note that for the diabetes dataset, the cost of the fair and traditional cluster solutions are similar for low k. Likewise, the difference in cost for this data between the solutions from the Euclidean and MWD decompositions is small here.

For our results we have used a direct translation of the fairlet decomposition cost from previous work based on minimization of distances induced by the clustering objective. A GMM cluster fit can also be evaluated in terms of likelihood. In Appendix B we present an approach for evaluating the likelihood of a GMM fairlet decomposition, which supports the results shown in Fig. 4.

5 DISCUSSION

Our results demonstrate that the generalized fairness metrics can be used to assess fairness of probabilistic cluster solutions and that such fair solutions can be obtained through a fairlet decomposition of the data fed as input to Algorithm 1. We observe that the fair GMM ensures high fairness even for a large number of mixture components, whereas the fairness of the traditional GMM becomes progressively worse and ultimately dictates monochromatic clusters ($B = 0$). Additionally, we note that fair GMMs generated with the Euclidean k-median objective produces similarly fair solutions as a fair GMM found from a MWD decomposition. Generally, the fair solutions demonstrate much less variance on the entropy metric than on the balance metric across the different number of cluster components. This is because the balance measure is determined by the least balanced cluster, while the entropy measure takes the fairness of all clusters into account. However, the balance measure has an intuitive appeal, as it measures the worst case fairness among all clusters, and consequently it may be better suited for ensuring adherence to the DI doctrine. From our experiments we also observe that the fair

GMM solutions increase the cost over a traditional GMM. The cost increase is significantly lower for the MWD decomposition than for the Euclidean k-median decomposition baselines, especially for a large number of mixture components in higher-dimensional spaces. This is due to the fact that k-median decompositions are designed to locate data points close in Euclidean space, which may be located far apart in the non-Euclidean space induced by the GMM. On the other hand, the MWD decomposition specifically connects points which lie close on the data manifold dictated by the GMM. For a low number of mixture components and for data in a low-dimensional space (like the diabetes dataset), the Euclidean k-median decomposition baselines do not increase the cost significantly over the MWD decomposition. This indicates than in such settings, the Euclidean k-median decomposition approach introduced by Backurs et al. (2019) can be used as a highly scaleable alternative to the MWD decomposition for obtaining fair GMM solutions without significantly increasing the cost.

Our proposed soft fairness metrics are both designed to generalize to the multiple non-overlapping color fairness setting. Similarly, Algorithm 1, which constructs the fair soft solution from any fairlet decomposition of the data, also generalizes to this setting. This means that inputting a fairlet decomposition of multiple non-overlapping groups (colors) would provide the same theoretical fairness bounds as the binary color setting. However, our minimum cost flow approach for constructing the input GMM fairlet decomposition is designed to only accommodate data points of two colors. The construction of fair clustering for multiple and overlapping protected groups has been studied in the deterministic setting by Bera et al. (2019) but remains an open question of potential future research in the probabilistic fairness setting.

6 CONCLUSION

Previous work on fair clustering has focused on deterministic hard clustering algorithms like k-means and k-median, where data points belong to specific clusters in a binary sense. In this work we study fair soft clustering by proposing generalizations of group-level fairness metrics. These generalizations allow the fairness metrics to be used in the presence of soft clustering algorithms by reflecting the underlying probabilistic nature. Furthermore, we have demonstrated an approach for obtaining a fair probabilistic cluster solution from a fairlet decomposition of the data. This approach may be applied on decompositions tailored specifically to mixture models, but can also be used to modify fairlet decompositions from previous work on hard clustering algorithms. Ultimately, the resulting solutions are costlier than their traditional counterparts, but in turn provide guaranteed bounds on their fairness.

Acknowledgements

We thank the reviewers for their insightful feedback and efforts towards improving our paper. Authors Rune D. Kjærsgaard and Madhumita Kundu were funded by a university alliance scholarship between UiB (University of Bergen) and DTU (Technical University of Denmark).

References

Anderson, N., Bera, S. K., Das, S., and Liu, Y. (2020). Distributional individual fairness in clustering. *arXiv preprint arXiv:2006.12589*.

Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., and Wagner, T. (2019). Scalable fair clustering. In *International Conference on Machine Learning*, pages 405–413. PMLR.

Bera, S., Chakrabarty, D., Flores, N., and Negahbani, M. (2019). Fair algorithms for clustering. *Advances in Neural Information Processing Systems*, 32.

Bercea, I. O., Groß, M., Khuller, S., Kumar, A., Rösner, C., Schmidt, D. R., and Schmidt, M. (2018). On the cost of essentially fair clusterings. *arXiv preprint arXiv:1811.10319*.

Brubach, B., Chakrabarti, D., Dickerson, J., Khuller, S., Srinivasan, A., and Tsepenekas, L. (2020). A pairwise fair and community-preserving approach to k-center clustering. In *International Conference on Machine Learning*, pages 1178–1189. PMLR.

Brubach, B., Chakrabarti, D., Dickerson, J. P., Srinivasan, A., and Tsepenekas, L. (2021). Fairness, semi-supervised learning, and more: A general framework for clustering with stochastic pairwise constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6822–6830.

Chhabra, A., Masalkovaitė, K., and Mohapatra, P. (2021). An overview of fairness in clustering. *IEEE Access*.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. (2017). Fair clustering through fairlets. *Advances in Neural Information Processing Systems*, 30.

Esmaeili, S., Brubach, B., Tsepenekas, L., and Dickerson, J. (2020). Probabilistic fair clustering. *Advances in Neural Information Processing Systems*, 33:12743–12755.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60.

Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31.

Patten, E. (2016). Racial, gender wage gaps persist in us despite some progress.

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., and O'Toole, A. J. (2011). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):1–11.

Rutherglen, G. (1987). Disparate impact under title vii: an objective theory of discrimination. *Va. L. Rev.*, 73:1297.

Sablich, L. (2016). 7 findings that illustrate racial disparities in education. *Brookings. edu.(Brookings (https://www. brookings. edu/blog/brown-center-chalkboard/2016/06/06/7-findings-that-illustrate-racial-disparities-in-education/))*.

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., and Clore, J. N. (2014). Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014.

Tipping, M. E. (1999). Deriving cluster analytic distance functions from gaussian mixture models.

Z. Obermeyer, B. Powers, C. V. and Mullainan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] All details for the empirical results are included in the supplemental code material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes] All used assets are publicly available and the required licenses are included in the supplementary code.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes] The three datasets are publicly available. The datasets are anonymized so that no identifiable information is available. This is also mentioned in the supplementary code and experiment instructions.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Fair Soft Clustering: Appendix

A FAIRNESS BOUND

This section explains the theoretical bound on the fairness of the solution C^* provided by Algorithm 1 in the main paper.

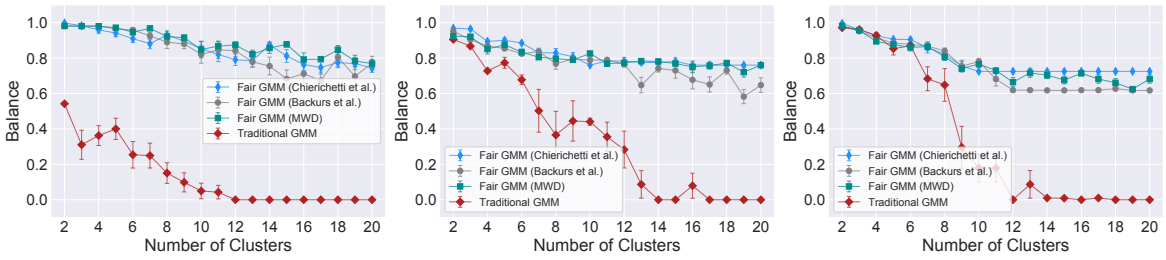
Consider a dataset D with a binary protected attribute. A fairlet decomposition Q of this data can be specified as a (p_1, p_2)-fairlet decomposition with parameters p_1 and p_2 (where $p_1 < p_2$) indicating that all fairlets have a color fraction $r \geq \frac{p_1}{p_1+p_2}$. The color fraction obtained from the union of these fairlets is bounded according to Lemma 1 (analogous to Lemma 2 from Chierichetti et al. (2017)).

Lemma 1 (Combination):
Let $Y_1, Y_2 \subseteq D$ be disjoint. If C_1 is a clustering of Y_1 and C_2 is a clustering of Y_2, then $r(C_1 \cup C_2) \geq min(r(C_1), r(C_2))$.

Algorithm 1 in the main paper combines the micro-clusters $q_1, q_2, \cdots, q_\ell$ (fairlets) into the probabilistic clustering C^* through a weighted combination dictated by the responsibilities $\boldsymbol{\gamma}$ of the fairlet centers. The weighted color fraction w of this combination is given by Eq. 3 in the main paper and is bounded according to Lemma 2.

Lemma 2 (Weighted combination):
Let $Y_1, Y_2 \subseteq D$ be disjoint. If C_1 is a weighted clustering of Y_1 and C_2 is a weighted clustering of Y_2, then $w(C_1 \cup C_2) \geq min(w(C_1), w(C_2))$.

This means that the weighted color fractions w_c of the final mixture components in C^* are bounded by $w_c \geq \frac{p_1}{p_1+p_2}\ \forall\ c$. To take a concrete example consider the Diabetes dataset from our experiments. We perform a $(1, 2)$-fairlet decomposition on the dataset and the weighted color fraction for any of the final mixture components is thus bounded by $w_c \geq \frac{1}{3}\ \forall\ c$. This dataset has an overall color fraction of $r_D = 0.54$. The soft balance of the final cluster solution is then bounded by $B \geq \frac{w_c}{r_D}$, i.e. $B \geq \frac{1/3}{0.54}$. This can be verified by inspecting Fig. A.1 (Fig. 2 in the main paper), where the balance for the fair solution on the Diabetes dataset never drops below $\frac{1/3}{0.54} = 0.62$.



(a) Census (b) Bank (c) Diabetes

Figure A.1: Fairness in terms of soft balance on the three datasets. Data points are the mean values from 5 iterations of different random seeds. Error bars indicate standard errors. The traditional GMM cluster approach obtains monochromatic cluster solutions ($B = 0$) for all datasets while the balance of the fair solutions are bounded. For instance the fair solutions of the Diabetes dataset are bounded at balance $B \geq 0.62$.

B GMM DECOMPOSITION LIKELIHOOD

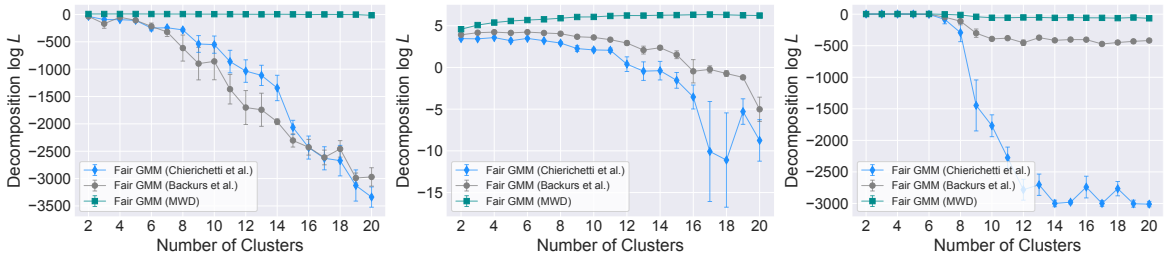
Our GMM fairlet decomposition is constructed by adapting the cost introduced by Chierichetti et al. (2017), which involves generating a fairlet decomposition through minimization of distances induced by the clustering objective. In our GMM fairlet decomposition we operate with the Mahalanobis and model-weighted distance. The cost of a GMM is typically not evaluated based on distances, but rather in terms of the log-likelihood. The log-likelihood of a data point belonging to a multivariate normal distribution is directly related to the Mahalanobis distance and is given by:

$$\log\, L(\boldsymbol{x}) = -\frac{1}{2}\left[\log(|\boldsymbol{\Sigma}|) + \log(d^2_{\mathrm{M}}(\boldsymbol{x}, \mathcal{N})) + m \cdot \log(2\pi)\right], \tag{1}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix, $d^2_{\mathrm{M}}(\boldsymbol{x}, \mathcal{N})$ is the squared Mahalanobis distance between data point $\boldsymbol{x}$ and distribution $\mathcal{N}$ and m is the multivariate dimension of $\mathcal{N}$. The model weighted distance is a generalization of the Mahalanobis distance to the Gaussian mixture setting. The model-weighted distance reduces to the Mahalanobis distance in settings with a single Gaussian, or in regions of space where only a single component density $p(\boldsymbol{x}|k)$ is non-zero along the path between the points Tipping (1999). While the Mahalanbis distance is directly related to the likelihood of a GMM solution, the connection between the model-weighted distance and the likelihood is less clear, and consequently the likelihood of the fairlet decomposition is harder to evaluate. However, we propose to estimate the likelihood by substituting the covariance matrix $\boldsymbol{\Sigma}$ in Eq. 1 with the model-weighted distance metric $\boldsymbol{G}$, and consequently the Mahalanobis distance with the model-weighted distance. The log-likelihood of a data point (fairlet member) belonging to a fairlet is then estimated as:

$$\log\, L_{\mathrm{Fairlet}}(\boldsymbol{x}) = -\frac{1}{2}[\log(|\boldsymbol{G}^{-1}|) + \log(d^2_{\mathrm{MWD}}(\boldsymbol{x}, \beta_{\mathrm{Q}}(\boldsymbol{x}))) + m \cdot \log(2\pi)], \tag{2}$$

where $d^2_{\mathrm{MWD}}(\boldsymbol{x}, \beta_{\mathrm{Q}}(\boldsymbol{x}))$ is the model-weighted distance from fairlet member $\boldsymbol{x}$ to fairlet center $\beta_{\mathrm{Q}}(\boldsymbol{x})$ and $|\boldsymbol{G}^{-1}|$ is the determinant of the associated inverse model-weighted distance metric. Under this view Eq. 2 evaluates the likelihood that a fairlet member was generated by the fairlet it is assigned to. Fig. B.1 shows the log-likelihood of the fairlet decompositions for the different datasets in our experiments.



(a) Census (b) Bank (c) Diabetes

Figure B.1: Per observation average log-likelihood of the fairlet decompositions of the three datasets. The log-likelihood is evaluated with Eq. 2. Data points are mean values from 5 iterations of different random seeds. Error bars indicate standard errors.

References

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. (2017). Fair clustering through fairlets. *Advances in Neural Information Processing Systems*, 30.

Tipping, M. E. (1999). Deriving cluster analytic distance functions from gaussian mixture models.