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.

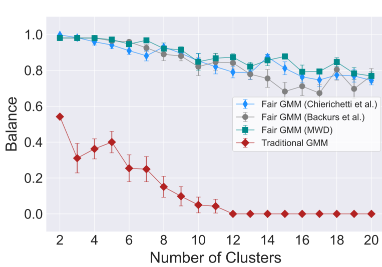
Consider a dataset D with a binary protected attribute. A fairlet decomposition Q of this data can be specified as a (p_1, p_2)-fairlet decomposition with parameters p_1 and p_2 (where $p_1 < p_2$) indicating that all fairlets have a color fraction $r \geq \frac{p_1}{p_1+p_2}$. The color fraction obtained 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 $r(C_1 \cup C_2) \geq min(r(C_1), r(C_2))$.

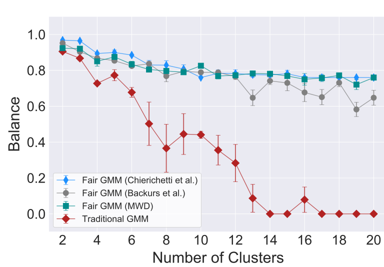
Algorithm 1 in the main paper combines the micro-clusters $q_1, q_2, \cdots, q_\ell$ (fairlets) into the probabilistic clustering C^* through a weighted combination dictated by the responsibilities $\boldsymbol{\gamma}$ of the fairlet centers. The weighted color fraction w of this combination is given 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 $w(C_1 \cup C_2) \geq min(w(C_1), w(C_2))$.

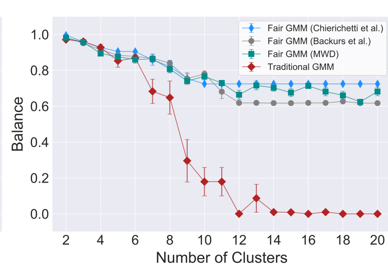
This means that the weighted color fractions w_c of the final mixture components in C^* are bounded by $w_c \geq \frac{p_1}{p_1+p_2}$ $\forall$ c. To take a concrete example consider the Diabetes dataset from our experiments. We perform a $(1, 2)$-fairlet decomposition on the dataset and the weighted 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

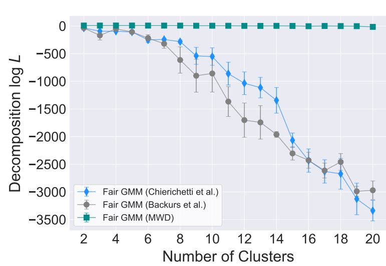
Our GMM fairlet decomposition is constructed by adapting the cost introduced by Chierichetti et al. (2017), which involves generating a fairlet decomposition through minimization of distances induced by the clustering objective. In our GMM fairlet decomposition we operate with the Mahalanobis and model-weighted 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:

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

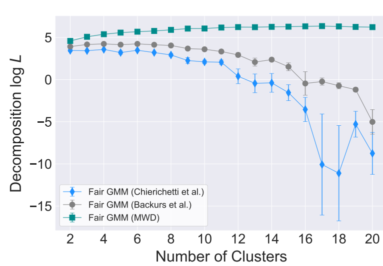
where $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix, $d^2_{\mathrm{M}}(\boldsymbol{x},\mathcal{N})$ is the squared Mahalanobis distance between data point $\boldsymbol{x}$ and distribution $\mathcal{N}$ and m is the multivariate dimension of $\mathcal{N}$. The model weighted distance 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:

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

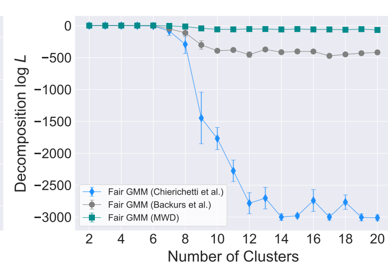
where $d^2_{\mathrm{MWD}}(\boldsymbol{x},\beta_{\mathrm{Q}}(\boldsymbol{x}))$ is the model-weighted distance from fairlet member $\boldsymbol{x}$ to fairlet center $\beta_{\mathrm{Q}}(\boldsymbol{x})$ and $|\boldsymbol{G}^{-1}|$ is the determinant of the associated inverse model-weighted distance metric. 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.