Fair Clustering via Hierarchical Fair-Dirichlet Prior

Abhisek Chakraborty[‡]
Global Statistical Sciences,
Eli Lilly and Company

Anirban Bhattacharya
Department of Statistics,
Texas A&M University

Debdeep Pati
Department of Statistics,
University of Wisconsin-Madison

Abstract

The advent of ML-driven decision-making has led to an increasing focus on algorithmic fairness. The widespread utility of clustering has naturally prompted proliferation of literature on fair clustering. A popular notion of fairness in clustering mandates the clusters to be balanced, i.e., each level of a protected attribute must be approximately equally represented in each cluster. In this article, we offer a novel model-based formulation of fair clustering, complementing the existing literature which is almost exclusively based on optimizing appropriate objective functions. We first rigorously define a notion of fair clustering in the population level and develop a Bayesian methodology equipped with a novel hierarchical prior specification that targets the population level objective by enforcing the notion of balance in the resulting clusters. In addition, we devise a scheme for principled performance evaluation of competing algorithms leveraging on a concrete notion of optimal recovery. An efficient collapsed Gibbs sampler is developed to sample from the posterior by integrating a novel scheme for non-uniform sampling from the space of binary matrices with fixed margin with a proposal guided by optimal transport. Superior empirical performance of the proposed methodology, compared to the state-of-the-art, is demonstrated across numerical experiments, benchmark data-sets, and gender-neutral fair clustering in distress analysis interview corpus.

[‡]Abhisek Chakraborty was formerly at the Department of Statistics, Texas A&M University.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Chierichetti et al. (2017) introduced *balance* as a criterion for fairness in clustering, in the setting where the protected attribute has two labels, commonly known as the *two-color* case. A clustering configuration is considered *fair* if each cluster preserves the same proportion of individuals from the different labels of the protected attribute. Extensions to the *multi-color* case, where the protected attribute has more than two labels, are considered in Böhm et al. (2020), and further generalized in Esmaeili et al. (2020) to allow probabilistic group membership. Bera et al. (2019) extended this framework to control the extent of fairness in clustering, use a general $\mathbf{L}_p$ norm as the clustering objective, and handle the case with individuals belonging to multiple protected groups.

The notion of fairness percolated other branches of clustering as well, such as e.g spectral clustering Kleindessner et al. (2019b), correlation clustering Ahmadian et al. (2020b), hierarchical clustering Ahmadian et al. (2020a) among others. Other notions of fairness in the context of clustering include individual fairness Kleindessner et al. (2020), proportional fairness Chen et al. (2019), etc. Perhaps unsurprisingly, fairness in clustering has been studied in conjunction with other pressing aspects of modern machine learning research as well, e.g privacy Rösner and Schmidt (2018), data summarization Kleindessner et al. (2019a), robustness Bandyapadhyay et al. (2019), to name a few.

While the follow-up works Böhm et al. (2020); Esmaeili et al. (2020); Rösner and Schmidt (2018); Kleindessner et al. (2019a) greatly increased the scope of fair clustering, a precise statistical framework is largely missing in the literature, which in turn makes it difficult to define the notion of the *true fair clustering configuration* at the *population level*, theoretically study optimal discovery/clustering consistency under fairness constraints, and device principled performance measures to compare methods. While one can device a probabilistic framework for fair clustering directly based on loss functions Bissiri et al. (2016);

Rigon et al. (2023), a model-based treatment of the problem is beneficial for simultaneous inference on number of clusters and cluster assignments, especially in presence of considerable uncertainty around cluster boundaries.

Our contributions.

Firstly, complementing the existing literature on fair clustering which is almost exclusively based on optimizing appropriate objective functions, we take a novel model-based approach to tackle the problem of clustering under approximate balance constraints. Thinking from a generative modeling perspective allows us to develop a concrete notion of *optimal recovery* in this problem, and subsequently device a scheme for principled *performance evaluation* for competing methods.

Secondly, we specify a hierarchical model equipped with a novel prior specification, termed as hierarchical fair Dirichlet prior, to encode a notion of approximate balance in resulting clusters, and provide concrete guidelines for the prior calibration that ensures a desired distribution of *balance* a-priori.

Thirdly, we develop an efficient collapsed Gibbs sampler that exploits a carefully designed gamma prior specification on the shared concentration parameter of the proposed hierarchical fair Dirichlet prior, ensuring rapid mixing and scalability. However, the most prominent computational bottleneck in our algorithm involves sampling cluster membership indices subject to balance constraints, which translates to a difficult non-uniform sampling problem in the space of binary matrices with fixed margins. This is circumnavigated via proposing a novel weighted rectangular loop scheme equipped with an binary optimal-transport based proposal mechanism. The proposed sampling scheme may be of independent importance in many applications in neurophysiology, sociology, psychometrics and ecology Brualdi and Ross (1980); Strona et al. (2014); Wang (2020).

Finally, the optimization based algorithms for fair clustering assume the knowledge of the number of clusters or estimate it a-priori using various selection criteria and subsequently estimate the clustering configuration. Ignoring the uncertainty in the first stage may lead to erroneous clustering. The proposed model-based approach enable us to estimate the number of clusters and clustering configuration simultaneously. Our approach also enjoys easy adaption to include key extensions present in the fair clustering literature, e.g, multiple colors Böhm et al. (2020), incomplete knowledge of group membership Esmaeili et al. (2020), to name a few. In summary, it inherits the usual benefits of model based clustering, e.g, simultaneous learning of the number of clusters and the cluster memberships Fraley and Raftery (2002); Ascolani et al. (2022); Miller and Harrison (2018a), and providing a framework to potentially handle pressing practical data complexities – non-isotropic covariances within clusters, mixed data type, censoring, missing values, variable selection Storlie et al. (2019), covariate adjustment, etc.

2 MODEL MISSPECIFICATION PERSPECTIVE

2.1 Probabilistic framework

Let $\mathbf{Z}_{\geqslant \mathbf{0}}$ (resp. $\mathbf{R}_{\geqslant \mathbf{0}}$) denote non-negative integers (reals), $[t] := \{1, \ldots, t\}$ for a positive integer t, and $\Delta^{t-1} = \{x \in \mathbf{R}^t_{\geqslant 0} : \langle 1_t, x\rangle = 1\}$ be the $(t-1)$-dimensional probability simplex. Given data $\{(\boldsymbol{x}_i, a_i)\}_{i=1}^N$ with d-variate observations $\boldsymbol{x}_i$ and protected attribute labels a_i, we assume $\{\boldsymbol{x}_i^{(a)}\}_{i=1}^{N_a}$ denote observations for attribute level a, where $N_a = \sum_{i=1}^N \mathbf{1}(a_i = a)$ and $\sum_{a=1}^r N_a = N$. Fair clustering seeks to partition the data into clusters $\boldsymbol{C} = (C_1, \ldots, C_K)$, $\bigcup_{k=1}^K C_k = [N]$, while respecting balance Chierichetti et al. (2017).

Definition 1 (Balance, Chierichetti et al. (2017))**.** *Given $\{(\boldsymbol{x}_i, a_i) \in \mathcal{X} \times \mathcal{A},\ i = 1 \ldots, N\}$ such that $a_i = a$ for $i = \sum_{j=1}^{a-1} N_j + 1, \ldots, \sum_{j=1}^{a} N_j$ where $a \in [r]$ and $N_0 = 0$, the balance in C_k is defined as*

$$Balance(C_k) = \min_{1 \leqslant j_1 < j_2 \leqslant r} \min \left\{ \frac{|C_{kj_1}|}{|C_{kj_2}|}, \frac{|C_{kj_2}|}{|C_{kj_1}|} \right\},$$

where $|C_{kj}|$ denote the number of observations in C_k with $a = j$. The overall balance of the clustering is

$$Balance(\boldsymbol{C}) = \min_{k=1,\ldots,K} Balance(C_k).$$

With the above definition of balance, Chierichetti et al. (2017) introduced the notion of *fairlets*, which are minimal fair sets that approximately preserve a chosen clustering objective. They showed that any fair clustering task can be approached via a two-stage procedure: first, compute an optimal fairlet decomposition of the observed data solving a *minimum cost flow* problem Edmonds and Karp (1972) and represent each fairlet by its center; second, apply standard clustering algorithms, e.g., k-means or k-center, to the fairlet centers to obtain balanced clusters. Follow-up works extended these methods both computationally and methodologically. However, a precise statistical framework is largely missing, making it difficult to define the notion of true fair clustering at the population level, study optimal discovery under fairness constraints, or develop principled measures for

performance evaluation. This motivates a probabilistic treatment of clustering with balance constraints, which we develop next.

We assume $\{(X_i, A_i)\}_{i=1}^N$ are independent copies of $(X, A) \in \mathcal{X} \times \mathcal{A}$, with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{A} = [r]$. Let $\boldsymbol{\xi} \in \Delta^{r-1}$ denote the known population proportions of the levels of the protected attribute. Let $\mathcal{P}^\star$ denote the true generative distribution. The generative mechanism is hypothesized as

$$A \sim \mathcal{P}^\star_A \equiv \text{Multinomial}(1, \boldsymbol{\xi})$$
$$(X, Z) \mid A \sim \mathcal{P}^\star_{X,Z|A} \equiv \mathcal{P}^\star_{X|Z,A} \times \mathcal{P}^\star_{Z|A},$$

where Z is the unobserved clustering index. In practice, only (X, A) are observed, and the goal is to learn $\mathcal{P}^\star_{Z|X}$ such that the clustering is *balanced* at the population level. To formalize the notion of balance at population level, let $\mathbb{P}$ denote the collection of all possible joint distributions of (X, Z, A), and define

$$\mathbb{P}^{(R)} = \{\mathcal{P}_{X,Z,A} \in \mathbb{P} : \mathcal{P}_{A|Z} = \mathcal{P}_A\}$$

as collection of data generating mechanisms with strict balance. For flexibility, we introduce

$$\mathbb{P}^{(R)}_\varepsilon = \{\mathcal{P}_{X,Z,A} \in \mathbb{P} : \text{KL}(\mathcal{P}_A \times \mathcal{P}_Z \,\|\, \mathcal{P}_{A,Z}) \leqslant \varepsilon\},$$

where $\varepsilon \geqslant 0$ controls departure from balance. We say (A, Z) is ε-*balanced* if the above KL divergence, or equivalently, the mutual information between A and Z, is at most ε. By definition, $\mathbb{P}^{(R)}_0 = \mathbb{P}^{(R)}$ corresponds to strict balance. In general, the true generative model $\mathcal{P}^\star$ may not belong to $\mathbb{P}^{(R)}_\varepsilon$. Our inferential goal is then to find the "best" approximation of $\mathcal{P}^\star$ within $\mathbb{P}^{(R)}_\varepsilon$, analogous to maximum likelihood estimation under model misspecification White (1982) or variational inference Blei et al. (2017).

Definition 2 (Pseudo True Fair Clustering Distribution, PT-FCD)**.** *For fixed $\varepsilon \geqslant 0$, the Pseudo-true clustering distribution restricted to $\mathbb{P}^{(R)}_\varepsilon$ is defined as,*

$$\mathcal{P}^\varepsilon_{Z_{\text{PT-FCD}}} = \text{argmin}_{\mathcal{P}_{A,Z,X} \in \mathbb{P}^{(R)}_\varepsilon} \text{KL}(\mathcal{P}^\star_{Z|X} \;||\; \mathcal{P}_{Z|X}),$$

where $\mathcal{P}_{Z|X}$ denotes the conditional distribution of $Z \mid X$ arising from $\mathcal{P}_{A,Z,X} \in \mathbb{P}^{(R)}_\varepsilon$. Further, the probability distribution on the size of the support of $\mathcal{P}^\varepsilon_{Z_{\text{PT-FCD}}}$ is defined as the pseudo-true distribution of number of clusters.

Note that, $\mathcal{P}^\varepsilon_{Z_{\text{PT-FCD}}}$ is the *KL projection* of the true $\mathcal{P}^\star_{Z|X}$ within the ε-*balanced* class of generative models, and the minimization problem in Definition 2 can be recasted as

$$\max_{\mathcal{P}_{A,Z,X} \in \mathbb{P}} \big[\mathbb{E}_{\mathcal{P}^\star_{Z|X}}(\log \mathcal{P}_{Z|X}) - \lambda_\varepsilon \text{KL}(\mathcal{P}_A \times \mathcal{P}_Z \;||\; \mathcal{P}_{A,Z})\big],$$

where $\lambda_\varepsilon \geqslant 0$ is a Lagrange's multiplier that depends on ε. In the following section, we develop the general template of a Bayesian procedure that targets to optimally recover the population level quantity $\mathcal{P}^\varepsilon_{Z_{\text{PT-FCD}}}$.

2.2 Restricted posterior maximization

Let $\mathbb{Z}$ denote the class of all possible clustering configurations of the observed data, and $\mathbb{Z}^{(fair)} \subset \mathbb{Z}$ be the subset where *balance* is strictly maintained. That is,

$$\mathbb{Z}^{(fair)} = \{\boldsymbol{z} \in \mathbb{Z} : \mathcal{C}_{\boldsymbol{z}} \text{ is balanced}\},$$

with $\mathcal{C}_{\boldsymbol{z}}$ being the clustering induced by $\boldsymbol{z}$. A relaxed version of the subset takes the form

$$\mathbb{Z}^{(fair)}(\varepsilon) = \{\boldsymbol{z} \in \mathbb{Z} : \text{KL}(\mathcal{P}_{\boldsymbol{A}} \times \mathcal{P}_{\boldsymbol{Z}} \,\|\, \mathcal{P}_{\boldsymbol{A},\boldsymbol{Z}}) \leqslant \varepsilon\},$$

where $\varepsilon \geqslant 0$ controls the extent of departure from strict balance. A clustering configuration is called ε-*balanced* if the above KL divergence is at most ε. The case $\mathbb{Z}^{(fair)}(0) = \mathbb{Z}^{(fair)}$ corresponds to clustering configurations with strict balance.

Definition 3 (Maximum-a-Posteriori fair cluster, MAP-FC)**.** *The modal clustering configuration obtained from the distribution of $\boldsymbol{Z} \mid \boldsymbol{X}$ restricted to $\mathbb{Z}^{(fair)}(\varepsilon)$, denoted by*

$$\mathcal{P}^{(fair)}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x}) \equiv \mathcal{P}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x}) \times \mathbf{1}_{\mathbb{Z}^{(fair)}(\varepsilon)}(\boldsymbol{z}), \quad (2.1)$$

is termed as the Maximum-a-Posteriori fair clustering configuration, i.e,

$$\boldsymbol{z}^\varepsilon_{\text{MAP-FC}} = \text{argmax}_{\boldsymbol{z} \in \mathbb{Z}^{(fair)}(\varepsilon)} \mathcal{P}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x}).$$

Further, the number of unique elements in $\boldsymbol{z}^\varepsilon_{\text{MAP-FC}}$ in Definition 3 is referred to as the Maximum-a-Posteriori number of clusters.

The maximization in Definition 3 can be equivalently written as

$$\arg\max_{\boldsymbol{z} \in \mathbb{Z}} \mathcal{P}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x}) - \lambda_\varepsilon \text{KL}(\mathcal{P}_{\boldsymbol{A}} \times \mathcal{P}_{\boldsymbol{Z}} \,\|\, \mathcal{P}_{\boldsymbol{A},\boldsymbol{Z}}),$$

where λ_ε is a Lagrange multiplier. In principle, $z^\varepsilon_{\text{MAP-FC}}$ can be targeted via a two-stage scheme:

1. **Unconstrained Bayesian clustering.** Utilize a flexible Bayesian clustering method e.g., Dirichlet process mixtures Neal (2000), mixtures of finite mixtures Miller and Harrison (2018b), and obtain posterior samples of $\boldsymbol{Z} \mid \boldsymbol{X}$;

2. **Post-processing to ensure fairness constraint.** Select the posterior sample with the highest posterior probability satisfying the ε-balance constraint.

While standard Bayesian clustering frameworks can be utilized to implement step 1, the MCMC will insufficiently explore the fair clustering configurations, making step 2 practically infeasible.

Instead, we propose a novel get-around by introducing the *hierarchical fair Dirichlet prior* (HFDP) in Section 3, which embeds the notion of *approximate balance* across clusters. The HFDP induces shrinkage toward the fairness-constrained space $\mathbb{Z}^{(\text{fair})}(\varepsilon)$, with the degree of shrinkage controlled by tunable hyperparameters. Consequently, MCMC targeting the posterior distribution under the HFDP naturally explores clustering configurations that are approximately fair with high probability, thereby enabling efficient post-processing to extract solutions that satisfy strict, user-specified fairness constraints.

Note that, $\boldsymbol{X} \mid \boldsymbol{A}$ assumes a mixture distribution of the form $\boldsymbol{X} \mid \boldsymbol{A} \sim \sum_{\boldsymbol{z}} \zeta_{\boldsymbol{a},\boldsymbol{z}} \mathcal{P}_{\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{A}}$, where $\mathcal{P}_{\boldsymbol{Z}|\boldsymbol{A}}(\boldsymbol{Z} = \boldsymbol{z} \mid \boldsymbol{A} = \boldsymbol{a}) = \zeta_{\boldsymbol{a},\boldsymbol{z}}$ and $\sum_{\boldsymbol{z}} \zeta_{\boldsymbol{a},\boldsymbol{z}} = 1 \; \forall \; \boldsymbol{a} \in [r]^N$. Then, with a slight abuse of notation, we have $\mathcal{P}^{(fair)}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x}) \propto \left[\sum_{\boldsymbol{a}} \zeta_{\boldsymbol{a},\boldsymbol{z}} \mathcal{P}_{\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{A}}(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{a}) \mathcal{P}_{\boldsymbol{A}}(\boldsymbol{a}) \right] \times \mathbf{1}_{\mathbb{Z}^{(fair)}(\varepsilon)}(\boldsymbol{z})$, with the normalizer

$$\mathcal{G}_N = \sum_{\boldsymbol{z}} \sum_{\boldsymbol{a}} [\zeta_{\boldsymbol{a},\boldsymbol{z}} \mathcal{P}_{\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{A}}(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{a}) \mathcal{P}_{\boldsymbol{A}}(\boldsymbol{a})] \times \mathbf{1}_{\mathbb{Z}^{(fair)}(\varepsilon)}(\boldsymbol{z}).$$

By definition, the posterior as defined above is maximized at $\boldsymbol{z}^{\varepsilon}_{\text{MAP}-\text{FC}}$.

In summary, Definition 2 formalizes the notion of ε-balanced clustering at the population level, under model a misspecification framework with minimal assumptions on the data-generating mechanism. At the sample level, Definition 3 introduces a general template of restricted posterior maximization framework targeting this population-level object, of which the HFDP in Section 3 is a concrete example. To reconcile these two notions, we establish an asymptotic equivalence between Definitions 2 and 3. For this, define a map $\mathcal{R} : [K]^N \to [0,1]^K$ that projects a clustering configuration $\boldsymbol{z} \in [K]^N$ to the vector of relative cluster proportions $(1/N)[\sum_{i=1}^N \mathbf{1}(z_i = 1), \ldots, \sum_{i=1}^N \mathbf{1}(z_i = K)]^{\intercal}$. The marginal posterior of $(\boldsymbol{Z} \mid \boldsymbol{X})$, reparameterized via $\boldsymbol{z} \mapsto \mathcal{R}(\boldsymbol{z})$, induces the marginal posterior of $(\mathcal{R}(\boldsymbol{Z}) \mid \boldsymbol{X})$, denoted by $P_{\mathcal{R}(\boldsymbol{Z})|\boldsymbol{X}}$.

Theorem 1. *In Definitions 2 and 3, for any fixed $\varepsilon \geqslant 0$, under some regularity conditions,* $\mathcal{P}^{\varepsilon}_{Z_{\text{PT}-\text{FCD}}} =$

$$\lim_{N \to \infty} \left[\frac{P_{\mathcal{R}(\boldsymbol{Z})|\boldsymbol{X}}(\mathcal{R}(\boldsymbol{z}^{\varepsilon}_{\text{MAP}-\text{FC}}) \mid \boldsymbol{x})}{\sum_{\boldsymbol{z}} P_{\mathcal{R}(\boldsymbol{Z})|\boldsymbol{X}}(\mathcal{R}(\boldsymbol{z}) \mid \boldsymbol{x}) \mathbf{1}_{\mathbb{Z}^{(fair)}(\varepsilon)}(\boldsymbol{z})} \right].$$

Due to space constraints, we defer the assumptions and the proof to the supplementary material, and instead discuss the key take-aways from Theorem 1. The equivalence result demonstrates that, given data $\{(\boldsymbol{x}_i, a_i)\}_{i=1}^N$, we can carefully construct probability models such that the resulting modal ε-balanced clustering configuration accurately recovers the Kullback–Liebler projection of the true population clustering distribution within the ε-balanced class, as sample size diverges to ∞. Secondly, on the practical front, Theorem 1 naturally suggests a principled measure for clustering performance comparison, that we discuss next.

2.3 Evaluating fair clustering methods

We may summarize the true data-generating process as follows: draw $A \in [r] \sim p^{\star}_A$, then $Z \in [K] \sim p^{\star}_{Z|A}$, and finally $X \sim p^{\star}_{X|A,Z}$. In general, A and Z may be dependent, so the true data generating mechanism may not satisfy the balance constraint. If we have access to the true data generating mechanism, we can explicitly compute $p^{\star}_{Z|X=x}(z) =$

$$\frac{p^{\star}_{Z,X}(z,x)}{p^{\star}_X(x)} = \frac{\sum_{a=1}^r p^{\star}_{X|A,Z}(x) \times p^{\star}_{Z|A}(z) \times p^{\star}_A(a)}{\sum_{z=1}^K \sum_{a=1}^r p^{\star}_{X|A,Z}(x) \times p^{\star}_{Z|A}(z) \times p^{\star}_A(a)},$$

for $z \in [K]$. In simulations, since the truth is known, we can use $p^{\star}_{Z|X}(z)$ to compare competing fair clustering configurations, provided they satisfy the same balance constraint.

Now suppose we have independent copies $(X_i, Z_i, A_i), i \in [N]$, where the latent clustering indices Z_i are unobserved. Under the true data-generating distribution, that does not necessarily adhere to the balance constraints, $Z_i \mid X_i$ are independent. We define

$$\text{Scaled log-score}(\mathbf{z}) = \frac{1}{N} \sum_{i=1}^N \log p^{\star}_{Z_i|X_i}(\mathbf{z}), \qquad (2.2)$$

that provides a basis for comparing fair clustering configurations that satisfy the same balance constraint. Given two such configurations $\mathbf{z}'$ and $\mathbf{z}^{\dagger}$, one prefers the configuration with higher scaled log-score under the true model, up to label switching. In real data applications, since the true data-generating mechanism is unknown, we treat the distribution of $\mathbf{Z} \mid \mathbf{X}$ estimated via a standard model-based clustering approach without balance constraints as the pseudo-true mechanism, and proceed as earlier.

3 HIERARCHICAL FAIR-DIRICHLET PRIOR

Let $\mathcal{Z}_{N,K} = \{\boldsymbol{z} = (z_1, \ldots, z_N) : z_i \in [K], i \in [N]\}$ denote the collection of all possible clustering configurations of N observations into K clusters. For any occupancy vector $\boldsymbol{m} = (m_1, \ldots, m_K) \in \mathbb{Z}^K_{\geqslant 0}$ with $\sum_{k=1}^K m_k = N$, define

$$\mathcal{Z}_{N,K,\boldsymbol{m}} = \{\boldsymbol{z} \in \mathcal{Z}_{N,K} : \sum_{i=1}^N \mathbf{1}(z_i = k) = m_k\}.$$

Equivalently, $\mathcal{Z}_{N,K,\boldsymbol{m}}$ corresponds to alll possible $N \times K$ binary membership matrices with column-sum $\boldsymbol{m}$ and row-sum $\mathbf{1}_N$, which we exploit in posterior sampling. Finally, define $\mathrm{rd} : \mathbb{N} \times \Delta^{t-1} \to \mathbb{Z}^t_{\geqslant 0}$ by $\boldsymbol{v} = \mathrm{rd}(n, \boldsymbol{u})$ with $v_i = \mathrm{round}(nu_i)$ for $i \in [t-1]$ and $v_t = n - \sum_{i=1}^{t-1} v_i \geqslant 0$, ensuring $\langle 1_t, \boldsymbol{v} \rangle = n$. This $\mathrm{rd}(\cdot)$ function underlies our novel prior on cluster occupancies.

The lowest-level hyperparameters are (K, g, b), where K is an upper bound on the number of clusters and $g, b > 0$ are shape and rate parameters. Given these, we sample a global weight vector $\boldsymbol{\beta} \in \Delta^{K-1}$ and concentration parameter $\alpha_0 > 0$ as in Equation (3.1). For each level of the protected attribute $a \in [r]$, we then draw an independent weight vector $\boldsymbol{w}^{(a)} \sim \mathrm{Dir}(\alpha_0 \boldsymbol{\beta})$:

$$\begin{aligned} &\boldsymbol{w}^{(a)} \,|\, \alpha_0, \boldsymbol{\beta} \overset{ind.}{\sim} \mathrm{Dir}(\alpha_0 \boldsymbol{\beta}), \quad a \in [r]; \\ &\boldsymbol{\beta} \,|\, g \sim \mathrm{Dir}(g/K, \ldots, g/K), \quad \alpha_0 \,|\, g, b \sim \mathrm{Gamma}(g, b). \end{aligned} \tag{3.1}$$

Given $\boldsymbol{w}^{(a)}$, we generate the cluster configuration $\boldsymbol{z}^{(a)}$ for sub-population a by first forming the occupancy vector $\boldsymbol{m}^{(a)}$ and then drawing $\boldsymbol{z}^{(a)}$:

$$\begin{aligned} &\boldsymbol{z}^{(a)} \mid \boldsymbol{m}^{(a)} \overset{ind.}{\sim} \mathrm{Unif}(\mathcal{Z}_{N_a,K,\boldsymbol{m}^{(a)}}), \\ &\boldsymbol{m}^{(a)} = \mathrm{rd}(N^{(a)}, \boldsymbol{w}^{(a)}), \quad a \in [r]. \end{aligned} \tag{3.2}$$

This layer is distinct from the standard HDP Teh et al. (2006) and enforces balanced clusters via what we call the *lifted Dirichlet* prior on $\boldsymbol{m}^{(a)}$; Figure 1 compares the lifted Dirichlet prior with the Dirichlet-Multinomial prior with $r = 2$. The lifted Dirichlet prior deterministically links cluster occupancy vector $\boldsymbol{m}^{(a)}$ to the attribute-specific weights $\boldsymbol{w}^{(a)}$, providing tighter control over cluster sizes compared to the standard Dirichlet-multinomial prior. Equations (3.1)–(3.2) define a joint distribution $\mathcal{P}_{a,z}$ over attribute labels and cluster indices. For fixed $\varepsilon \geqslant 0$, we propose to tune the hyperparameters (b, g) so that $\mathrm{KL}(\mathcal{P}_{\boldsymbol{a}} \times \mathcal{P}_{\boldsymbol{z}} \,\|\, \mathcal{P}_{\boldsymbol{a},\boldsymbol{z}}) \leqslant \varepsilon$ with 90–95% probability. Note that, the prior is not truncated to enforce this restriction strictly; rather, it concentrates mass near it. While this requires post-processing of posterior samples to ensure exact balance, the soft prior allows MCMC to explore the posterior over an unconstrained domain.

The final part of our hierarchical model specifies component-wise distributions:

$$\begin{aligned} &\boldsymbol{x}_i^{(a)} \mid z_i^{(a)} = k, \boldsymbol{\phi}_k^{(a)} \overset{ind.}{\sim} f_{obs}(\cdot \mid \boldsymbol{\phi}_k^{(a)}), \; i \in [N_a], \; a \in [r], \\ &\boldsymbol{\phi}_k^{(a)} \mid \boldsymbol{\phi}^{(a)} \overset{ind.}{\sim} f_{pop}(\cdot \mid \boldsymbol{\phi}^{(a)}), \quad k \in [K], \; a \in [r], \\ &\boldsymbol{\phi}^{(a)} \overset{i.i.d.}{\sim} f_{atom}, \quad a \in [r]. \end{aligned} \tag{3.3}$$

Here, $\boldsymbol{\phi}_k^{(a)}$ are cluster- and attribute-specific parameters drawn from a common population. Equations (3.1)–(3.3) complete the HFDP model-prior specification. For concreteness, we consider a Gaussian likelihood with conjugate NIW priors: $f_{obs}(\cdot \mid \boldsymbol{\phi}_k^{(a)}) = \mathrm{N}_d(\boldsymbol{\mu}_k^{(a)}, \Sigma_k^{(a)})$ and $f_{pop}(\cdot \mid \boldsymbol{\phi}^{(a)}) = \mathrm{NIW}(\boldsymbol{\mu}_0^{(a)}, \lambda_0^{(a)}, \Lambda_0^{(a)}, \nu_0^{(a)})$, with hyperparameters suitably fixed. We collect $\boldsymbol{\mu}^{(a)} = \{\boldsymbol{\mu}_1^{(a)}, \ldots, \boldsymbol{\mu}_K^{(a)}\}$ and $\Sigma^{(a)} = \{\Sigma_1^{(a)}, \ldots, \Sigma_K^{(a)}\}$ for each $a \in [r]$. Under this model, our goal is to learn the clustering indices $\{\boldsymbol{z}^{(a)}, a \in [r]\}$ respecting approximate balance, reducing the need for posterior sample post-processing, as described in Section 2.1. Extensions to non-Gaussian likelihoods are straightforward.

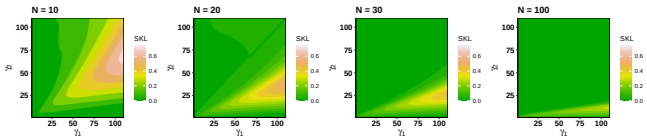

Figure 1: *Symmetrized KL between lifted Beta and Beta-Binomial distribution for varying γ_1 and γ_2. As N increases, the symmetrized KL uniformly gets smaller, implying the distributions get closer.*

We next the highlight key features of the hierarchical specification. The concentration parameter α_0 in Equation (3.1) controls how tightly the attribute-specific weights $\{\boldsymbol{w}^{(a)}\}_{a=1}^r$ cluster around the global weight $\boldsymbol{\beta}$, which is crucial for enforcing balance. The values of (b, g) that yield higher α_0 with high probability produce more concentrated $\{\boldsymbol{w}^{(a)}\}_{a=1}^r$ and hence more balanced clusters. For $r = 2$, the KL divergence $\mathrm{KL}(w^{(1)} \,\|\, w^{(2)})$ measures discrepancy between attribute-specific weights. Fixing (K, g, b), prior draws of $\{\boldsymbol{w}^{(a)}\}$ allow numerical estimation of this divergence; Figure 2 illustrates the distribution for varying (b, g) with $K = 2$. For fixed g, smaller b yields a larger α_0 with high probability, hence enhances balance. Analysts can control the expected balance of clusters a priori by selecting (b, g), which effectively calibrates the prior concentration of α_0.

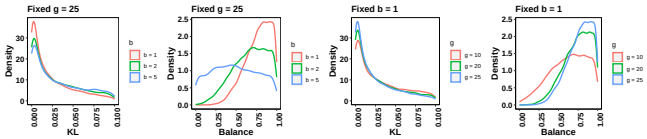

Figure 2: *Visual recipe for prior calibration with $r = 2$. The first two plots, respectively, present the distribution of $KL(w^{(1)} \,||\, w^{(2)})$, and balance between $(\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)})$, for varying b with fixed g. The final two plots present the same quantities, now with varying g with fixed b. In practice, we vary (b, g) until we obtain the desired distribution of balance a-priori.*

In summary, the prior specification in Equations (3.1)–(3.3) strongly favors fairness *a priori*. Even when the true latent clusters are imbalanced, the posterior tends to yield fair clusters, driven by the attribute-specific cluster proportions induced by the model-prior and consistent with the principles in Section 2. This soft enforcement of fairness is intentional, allowing the MCMC algorithm to explore the posterior over an unconstrained domain while still inclining toward balanced partitions. In practice, a simple post-processing step ensures that the desired fairness cut-off is met, with the key advantage being computational: the proposed HFDP methodology more efficiently identifies fair clustering configurations than a standard hierarchical Dirichlet process by construction.

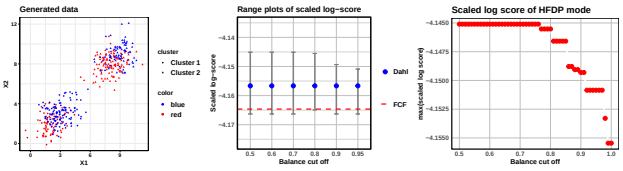

Figure 3: ***Left panel.*** *Generated data with* $N_{11} = 80$, $N_{21} = 120$, $N_{12} = 120$, $N_{22} = 80$. ***Middle panel.*** *Range of the scaled log-score of the posterior clustering configurations via HFDP that has balance* $\geqslant$ *cut-off* $\in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$, *and the scaled log-score of the modal clustering configuration from HFDP, computed via Dahl's method Dahl (2006). The scaled log-score for Chierichetti et al. (2017) (termed FCF) is marked in red.* ***Right panel.*** *Scaled log-scores for the modal posterior clustering configurations, where the balance satisfies* $\geqslant$ *cut-off* $\in \{0.5, 0.51, \ldots, 1.00\}$.

Comparison via scaled log-score. We generate a dataset with two protected attribute labels and two clusters. In cluster i, N_{ji} individuals with $a = j$ are drawn from $N_2(\mu_{ji}, S)$, with equal representation of the two attributes and total sample size $N = \sum_{i=1}^{n}\sum_{j=1}^{n} N_{ji} = 400$. We set $N_{11} = 80, N_{21} = 120, N_{12} = 120, N_{22} = 80$, $\mu_{11} = (2,2)^\top$, $\mu_{21} = (3,3)^\top$, $\mu_{12} = (8,8)^\top$, $\mu_{22} = (9,9)^\top$, and $S = \rho 11^\top + (1-\rho)I_2$, $\rho = 0.3$. The middle panel in Figure 3 reveals that the proposed methodology often explores clustering configurations that are more likely under the true generative mechanism than the optimal fair clustering via fairlets with $\mathcal{L}_2$ loss Chierichetti et al. (2017). In particular, the modal clustering configuration arising from the proposed methodology, computed by Dahl's method Dahl (2006), yields higher scaled log-score under the true data generative model, compared to the optimal fair clustering via fairlets, owing to potentially sub-optimal exchanges of observations between clusters that take place to maintain balance in the optimal fair clustering via fairlets.

4 POSTERIOR COMPUTATION

Sampling from the posterior arising from the hierarchical model-prior in Equation (3.1)-(3.3) via Gibbs sampling encounters two challenges: (a) the lack of conjugacy in the shared weight vector β, and concentration parameter α_0, and (b) the combinatorial complexity of sampling $[\boldsymbol{z}^{(a)} \mid \boldsymbol{m}^{(a)}, \{\phi_k^{(a)}\}_{k=1}^{K}, \beta, \alpha_0\}]$, $a \in [r]$. The bottleneck (a) is alleviated by noting that a well-designed Gamma prior on α_0 allows us to break the dependence in β, and permits independent updates of certain log-concave random variables in a block. The second bottleneck is carefully circumnavigated via a novel scheme for non-uniform sampling in the space of binary matrices with fixed margins, that utilizes techniques from integer-valued optimal transport (OT) Peyré and Cuturi (2019) towards constructing the Metropolis–Hastings proposals. Due to space limitations, full details of the Gibbs sampler are deferred to the supplementary material and we only present details of (b) in the sequel.

4.1 Posterior mode of $[\boldsymbol{z}^{(a)} \mid \cdot]$ via discrete optimal transport

To compute the mode of $[\boldsymbol{z}^{(a)} \mid \cdot]$, $a \in [r]$, (i) we first calculate component specific means and variances $(\boldsymbol{\mu}_{k,\star}^{(a)}, \Sigma_{k,\star}^{(a)})$ for all $a \in [r], k \in [K]$ at the current Gibbs sampler iteration. (ii) Next, we define the $N_a \times K$ cost matrix

$$\mathrm{L}^{(a)} = ((l_{ik})) = \big(\big(-\log \mathrm{N}_d(\boldsymbol{x}_i^{(a)} \mid \boldsymbol{\mu}_{k,\star}^{(a)}, \Sigma_{k,\star}^{(a)})\big), \quad (4.1)$$

the column sum vectors $\boldsymbol{c}^{(a)} = \boldsymbol{m}^{(a)}$ and row sum vectors $\boldsymbol{r}^{(a)} = \mathbf{1}_{N_a}$, where $\mathbf{1}_{N_a}$ is a vector of N_a 1s. (iii) Next, given $\boldsymbol{r}^{(a)}, \boldsymbol{c}^{(a)}$, we define the polytope of $K \times N_a$ binary cluster membership matrices with fixed margins

$$\mathrm{U}(\boldsymbol{r}^{(a)}, \boldsymbol{c}^{(a)}) := \{\mathrm{B} \mid \mathrm{B}\mathbf{1}_{N_a} = \boldsymbol{r}^{(a)};\ \mathrm{B}^{\mathrm{T}} 1_K = \boldsymbol{c}^{(a)}\},$$

and solve the constrained binary optimal transport problem

$$\mathrm{B}^{(a)} = \mathrm{argmin}_{\mathrm{B} \in \mathrm{U}(\boldsymbol{r}^{(a)},\ \boldsymbol{c}^{(a)})} \langle \mathrm{B}, \mathrm{L}^{(a)} \rangle,$$

where $\langle \mathrm{B}, \mathrm{L}^{(a)} \rangle = \mathrm{tr}(\mathrm{B}^{\mathrm{T}} \mathrm{L}^{(a)})$, via the branch and bound algorithm Land and Doig (1960). Finally, we obtain the modal clustering indices $\boldsymbol{z}^{(a)}$ from the binary cluster membership matrices $\mathrm{B}^{(a)}, a \in [r]$.

4.2 Sampling of fixed margin binary matrices

Both exact Miller and Harrison (2013) and approximate Brualdi and Ross (1980); Strona et al. (2014);

Wang (2020) uniform sampling methods in the space of binary matrices with fixed margins are now well-established. However, sampling from $[\boldsymbol{z}^{(a)} \mid \cdot]$, $a \in [r]$ requires sampling according to a non-uniform probability distribution defined by a weight matrix. To that end, we adapt Wang (2020) to introduce novel rectangular loop moves in the non-uniform case to improve mixing. Recall that, given fixed row sums $\mathbf{r} = (r_1, \ldots, r_u)^{\mathrm{T}}$ and column sums $\mathbf{c} = (c_1, \ldots, c_v)^{\mathrm{T}}$, we denote the collection of all $u \times v$ binary matrices by $\mathrm{U}(\mathbf{r}, \mathbf{c})$. Further, we denote by $\Omega = (\omega_{ij}) \in [0, \infty)$ a non negative weight matrix representing the relative probability of observing a count of 1 at the (i, j)-th cell. Let

$$\mathrm{U}'(\mathbf{r}, \mathbf{c}) = \{H \in \mathrm{U}(\mathbf{r}, \mathbf{c}) : \ P(H) > 0\}$$

denote the subset of matrices in $\mathrm{U}(\mathbf{r}, \mathbf{c})$ with positive probability. Then, the likelihood associated with the observed binary matrix $H \in \mathrm{U}'(\mathbf{r}, \mathbf{c})$ is $\mathrm{P}(H) \propto \prod_{i,j} \omega_{ij}^{h_{ij}}$. We denote the identity matrix of order 2, and the 2×2 matrix with zero on the diagonals and one on the off-diagonals as *checker-board* matrices. With these notation, we introduce a *Weighted Rectangular Loop Algorithm* (W-RLA) for non-uniform sampling from the space of fixed margin binary matrices $H \in \mathrm{U}'(\mathbf{r}, \mathbf{c})$, given the weight matrix $\Omega = (\omega_{ij}) \in [0, \infty)$.

W-RLA is described as follows.

1. At iteration $t = 0$, provide an initial binary matrix A_0 that respects the margin constraints, and total number of iterations T.

2. Increment $t \to t + 1$, and *(a)* choose one row and one column (r_1, c_1) uniformly at random. *(b) If* $A_{t-1}(r_1, c_1) = 1$, choose a column c_2 at random among all the 0 entries in r_1, and a row r_2 at random among all the 1 entries in c_2. *Else* choose a row r_2 at random among all the 1 entries in c_1, and a column c_2 at random among all the 0 entries in r_2.

3. (a) *If* the sub-matrix extracted from r_1, r_2, c_1, c_2 is a checkerboard unit – obtain B_t from A_{t-1} by swapping the checker-board; calculate

$$p_t = \mathrm{P}(B_t)/[\mathrm{P}(B_t) + \mathrm{P}(A_{t-1})];$$

draw $r_t \sim \text{Bernoulli}(p_t)$; and *If* $r_t = 1$, then set $A_t = B_t$; *else* set $A_t = A_{t-1}$.

(b) *If* the sub-matrix extracted from r_1, r_2, c_1, c_2 is *not* a checkerboard unit, set $A_t = A_{t-1}$.

This completes the description of W-RLA; the full algorithm is provided in the supplement. Theorem 2 establishes that the W-RLA Markov chain converges to the correct stationary distribution as $T \to \infty$, with the proof deferred to the supplementary material.

Theorem 2. *Given an* Ω *and* $(\boldsymbol{r}, \boldsymbol{c})$*, the W-RLA generates a Markov chain with stationary distribution given by* $P(H) \propto \prod_{i,j} \omega_{ij}^{h_{ij}}$.

Sampling from $[\boldsymbol{z}^{(a)} \mid \cdot]$. In the blocked Gibbs sampler for the posterior in (3.1)–(3.3), sampling $[\boldsymbol{z}^{(a)} \mid -]$, $a \in [r]$, proceeds by first generating a proposal cluster membership matrix $\mathrm{B}^{(a)}_{\mathrm{prop}}$ via the constrained binary optimal transport problem

$$\mathrm{B}^{(a)}_{\mathrm{prop}} = \underset{\mathrm{B} \in \mathrm{U}(\boldsymbol{r}^{(a)}, \boldsymbol{c}^{(a)})}{\arg\min} \ \langle \mathrm{B}, \mathrm{L}^{(a)} \rangle,$$

see Eq. (4.1). The proposal is then mutated using the weighted rectangular loop introduced above to obtain the M–H proposal $\mathrm{B}^{(a)}_{\mathrm{rec-loop}}$ to sample from $[\boldsymbol{z}^{(a)} \mid -]$, setting $\Omega = \mathrm{L}^{(a)}$, $\boldsymbol{c}^{(a)} = \boldsymbol{m}^{(a)}$ and $\boldsymbol{r}^{(a)} = \mathbf{1}_{N_a}, a \in [r]$.

5 NUMERICAL EXPERIMENTS

5.1 Simulation study 1 (two-color case)

Assume that there are two labels of the protected attribute $a \in \{1, 2\}$. The goal is to compare the perfectly balanced modal clustering configuration arising from the proposed methodology and the optimal fair clustering configuration via fairlets with $\mathcal{L}_2$ loss. Chierichetti et al. (2017), in replicated simulations. To that end, we fix the true number of clusters $K_{true} = 2$, set the attribute specific sample sizes $N^{(1)} = N^{(2)} = 200$, and generate synthetic data from the hierarchical model described in Equations (3.1)-(3.3) with (b, g) fixed such that we observe clusters that are inherently unbalanced. We consider three different choices for the within cluster attribute specific densities.

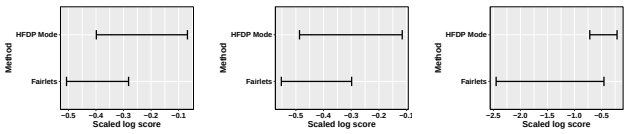

Figure 4: ***Two-color case using the within cluster attribute specific densities as (left panel)*** *multivariate normal distribution,* ***(middle panel)*** *multivariate Student's t, and* ***(right panel)*** *multivariate skew normal. From replicated simulations, we plot the range of the scaled log-score, yielded by the HFDP modal clustering configurations and the optimal clustering configurations arising from fair clustering via fairlets.*

(A.1) Well-specified case (multivariate normal components). In the i-th cluster, N_{ji} individuals with $a = j$ are generated from $N_2(\mu_{ji}, S)$, where $i \in \{1,2\}$ and $j \in \{1,2\}$. Finally, we set $\mu_{11} = (4,4)'$, $\mu_{21} = (2,2)'$, $\mu_{12} = (10,10)'$, $\mu_{22} = (8,8)'$, and $S = 3 \times [\rho 11^{\mathrm{T}} + (1-\rho) I_2]$ with $\rho = 0.3$.

(A.2) Mis-specified case 1 (multivariate t components). We follow the same scheme as before, except for simulating from multivariate t-distributions with centers $(\mu_{11}, \mu_{21}, \mu_{12}, \mu_{22})$, scale S, and degrees of freedom 4.

(A.3) Mis-specified case 2 (multivariate skew normal components). We follow the same scheme as before, except for simulating from multivariate skew normal distributions Azzalini and Valle (1996) with centers $(\mu_{11}, \mu_{21}, \mu_{12}, \mu_{22})$, scale S, and the skewness parameter $\alpha = (1,1)^{\mathrm{T}}$.

In Figure 4, across replicated simulations, we plot the the range of the scaled log-scores, defined in (2.2), yielded by the HFDP modal clustering configurations and the optimal clustering configurations arising from fair clustering via fairlets, in both well-specified and mis-specified set ups. Across all the numerical experiments, the HFDP modal clustering configuration fares substantially better in terms of the scaled log-score. This underscores that the proposed methodology recovers a perfectly balanced clustering configuration of the observed data, that is more probable under the true data generating mechanism, compared to the optimal fair clustering via fairlets. Further, HFDP estimates the number of clusters automatically, but the number of clusters for the competitor is set deterministically at the value estimated by the the HFDP modal clustering configuration.

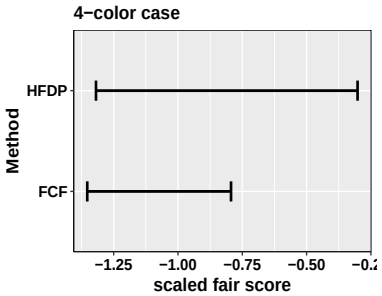

Figure 5: ***Multi-color case.*** *The HFDP versus optimal fair clustering configuration via Böhm et al. (2020).*

.

5.2 Simulation study 2 (multiple color case)

We assume that there are more than two labels of the protected attribute $a \in \{1, 2, \ldots, r\}$. We generate data from the model in Equations (3.1)-(3.3), with number of colors $r = 4$, sample sizes $N^{(1)} = N^{(2)} = N^{(3)} = N^{(4)} = 200$, true number of clusters $K_{true} = 2$. In the first cluster, the individuals with $a = 1, 2, 3$ and 4 are generated from $N_2((4,4)', S)$, $N_2((2,2)', S)$, $N_2((0,0)', S)$, and $N_2((-2,-2)', S)$, respectively. In the second cluster, individuals with $a = 1, 2, 3$ and 4 are generated from $N_2((10,10)', S)$, $N_2((8,8)', S)$, $N_2((6,6)', S)$, and $N_2((4,4)', S)$, respectively. We set $S = 3 \times [\rho 11^{\mathrm{T}} + (1-\rho) I_2]$ with $\rho = 0.3$. In Figure 5, the method in Böhm et al. (2020) with $\mathcal{L}_2$ loss (termed as K-Means) is compared with HFDP via repeated simulations. Across all the numerical experiments, the HFDP modal clustering fares substantially better in terms of the scaled log-score.

We defer the results under imperfect knowledge of the protected attribute Esmaeili et al. (2020), and three benchmark datasets to the supplementary material.

6 DISTRESS ANALYSIS INTERVIEW CORPUS

The Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) is a popular dataset to create automated agents for assessment of psychological distress. Interviews of participants are transcribed and annotated for verbal and non-verbal markers, with depression severity labeled using the PHQ-8 ranging between $0 - 24$.

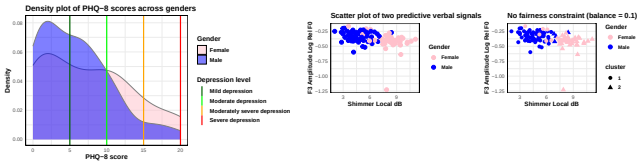

Figure 6: ***(Left panel.)*** *Gender-specific densities of the PHQ-8 scores, revealing clear differences between genders.* ***(Middle panel.)*** *Scatter plot of two verbal features predictive of PHQ-8 score reveals differences across biological genders.* ***(Right panel.)*** *Clustering the two verbal features without fairness constraints results in highly imbalanced clusters.*

When self-assessed PHQ-8 scores are unavailable, agents pre-trained using annotated verbal signals are used for mental health diagnostics, including in employment decisions, raising concerns about fairness with respect to gender. We predict PHQ-8 scores from verbal features derived via Fourier analysis of speech

signals in clinical interviews of $n = 107$ participants (44 females, 63 males; balance ≈ 0.69). Figure 6 shows gender-specific score distributions with clear differences, which may partly reflect biological factors, but fairness constraints can be imposed to ensure equitable treatment across genders in downstream decisions.

To evaluate HFDP clustering under gender based balance constraints, we first fit a linear regression with PHQ-8 as the response and 17 verbal features as covariates. Two features emerged as most important, which we use for visualization, though the method can handle more features. The second panel in Figure 6 shows clear gender differences verbal signals. Clustering the verbal signal into $K = 2$ without fairness constraints produces highly imbalanced clusters; see the right panel of Figure 6.

Since the true data-generating mechanism is unknown, we treat the clustering distribution without balance constraints as the pseudo-true mechanism as described in section 2. The attribute probabilities are estimated as $\hat{p}^{\star}_{A}(\text{male}) = 44/107$ and $\hat{p}^{\star}_{A}(\text{female}) = 63/107$, and the attribute-specific cluster probabilities $\hat{p}^{\star}_{Z|A}(z)$ are based on observed proportions. The attribute- and cluster-specific densities $\hat{p}^{\star}_{X|A=a,Z=z}(x)$ are modeled as bivariate normals, with parameters estimated from observations with attribute a in cluster z. These estimates allow computation of scaled log-scores for any clustering configuration using equation (2.2). Under the above pseudo-true data-generating mechanism, we compute the balance and scaled log-score for each posterior clustering configuration obtained from HFDP, and the perfectly balanced clustering configuration generated by fair clustering via fairlets.

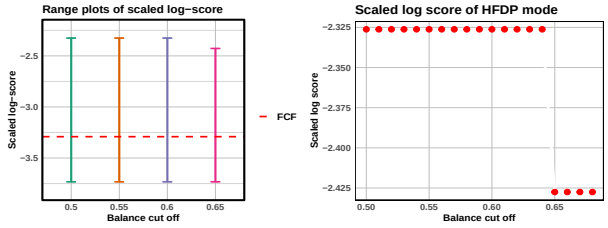

Figure 7: ***(Left panel.)*** *Range of the scaled log-score of the posterior clustering configurations via HFDP with balance* $\geqslant$ *cut-off* $\in \{0.5, 0.55, 0.6, 0.65\}$. *The horizontal dashed red line represents scaled log-score of the optimal clustering configurations via fair clustering via fairlets.* ***(Right panel.)*** *Scaled log-scores for the modal posterior clustering configurations obtained from HFDP, where the balance satisfies* $\geqslant$ *cut-off* $\in \{0.50, 0.51, \ldots, 0.69\}$.

The right panel of Figure 7 shows scaled log-scores for HFDP modal posterior clustering as a function of the balance cut-off, with a sharp decline at balance $= 0.63$. This indicates that a slight relaxation of the balance constraint can substantially improve the clustering and demonstrates that HFDP provides data-driven guidance on balance choice when flexibility exists. The left panel of Figure 7 shows that, as the balance cut-off increases, the scaled log-score of HFDP modal clustering decreases monotonically, as expected. Compared to the optimal fair clustering via fairlets, it also shows that HFDP recovers a perfectly balanced clustering that is more probable under the pseudo-true data-generating mechanism.

7 CONCLUSION

In this article, complementing the predominantly optimization-driven approaches in existing literature, we introduced a novel model-based formulation for the fair clustering problem, and rigorously defined a notion of fair clustering at the population level, enabling us to propose a principled framework for evaluating the performance of fair clustering algorithms. The proposed Bayesian methodology incorporates a hierarchical prior to enforce a generalized notion of balance in the resulting clusters.

Much of the prior work on fair clustering assumes complete knowledge of group membership, with Esmaeili et al. (2020) extending this to probabilistic assignments. Our fully probabilistic framework inherently accommodates such scenarios by integrating an additional sampling step into each computational cycle. Moreover, the proposed model-based approach effectively handles practical data complexities, such as mixed data types, censoring, and missing values, in a seamless manner (Storlie et al., 2019). We discuss such extensions in Supplementary Section S4.

Future research could explore how a generative modeling perspective, akin to that presented in this article, can be extended to address other variants of fair clustering and alternative notions of fairness in clustering. Developing a robust non-parametric Bayesian framework to address the fair clustering problem under misspecified cluster-specific densities also presents a compelling direction for future research.

References

Ahmadian, S., Epasto, A., Knittel, M., Kumar, R., Mahdian, M., Moseley, B., Pham, P., Vassilvitskii, S., and Wang, Y. (2020a). Fair hierarchical clustering. *CoRR*, abs/2006.10221. 1

Ahmadian, S., Epasto, A., Kumar, R., and Mahdian,

M. (2020b). Fair correlation clustering. *CoRR*, abs/2002.02274. 1

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*. 23

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2022). Clustering consistency with Dirichlet process mixtures. *Biometrika*, 110(2):551–558. 2

Azzalini, A. and Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726. 8

Bandyapadhyay, S., Inamdar, T., Pai, S., and Varadarajan, K. R. (2019). A constant approximation for colorful k-center. *CoRR*, abs/1907.08906. 1

Bera, S., Chakrabarty, D., Flores, N., and Negahbani, M. (2019). Fair algorithms for clustering. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. 1

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130. 1

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877. 3

Brualdi, R. A. and Ross, J. A. (1980). On ryser's maximum term rank formula. *Linear Algebra and its Applications*, 29:33–38. Special Volume Dedicated to Alson S. Householder. 2, 6, 18

Böhm, M., Fazzone, A., Leonardi, S., and Schwiegelshohn, C. (2020). Fair clustering with multiple colors. 1, 2, 8, 12, 24

Chen, X., Fain, B., Lyu, C., and Munagala, K. (2019). Proportionally fair clustering. *CoRR*, abs/1905.03674. 1

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. (2017). Fair clustering through fairlets. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. 1, 2, 6, 7, 12, 15, 24

Dahl, D. (2006). Model-based clustering for expression data via a dirichlet process mixture model, in bayesian inference for gene expression and proteomics. *Cambridge University Press*. 6, 15, 22

Dua, D. and Graff, C. (2017). UCI machine learning repository. 12, 24

Edmonds, J. and Karp, R. M. (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM*, 19(2):248–264. 2

Esmaeili, S., Brubach, B., Tsepenekas, L., and Dickerson, J. (2020). Probabilistic fair clustering. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12743–12755. Curran Associates, Inc. 1, 2, 8, 9, 12, 22, 23, 24

Fout, A., Fosdick, B. K., and Hitt, M. P. (2020). Non-uniform sampling of fixed margin binary matrices. 20

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631. 2

Kleindessner, M., Awasthi, P., and Morgenstern, J. (2019a). Fair k-center clustering for data summarization. 1

Kleindessner, M., Awasthi, P., and Morgenstern, J. (2020). A notion of individual fairness for clustering. 1

Kleindessner, M., Samadi, S., Awasthi, P., and Morgenstern, J. (2019b). Guarantees for spectral clustering with fairness constraints. 1

Land, A. H. and Doig, A. G. (1960). An automatic method for solving discrete programming problems. *Econometrica*, 28(3):497–520. 6

Levine, R. A. and Casella, G. (2001). Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439. 19

Leydold, J. (1998). A rejection technique for sampling from log-concave multivariate distributions. *ACM Transactions on Modeling and Computer Simulation*, 8. 18

Miller, J. W. and Harrison, M. T. (2013). Exact sampling and counting for fixed-margin matrices. *The Annals of Statistics*, 41(3):1569 – 1592. 6, 18

Miller, J. W. and Harrison, M. T. (2018a). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356. PMID: 29983475. 2

Miller, J. W. and Harrison, M. T. (2018b). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356. 3

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of*

Computational and Graphical Statistics, 9(2):249–265. 3

Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607. 6

Rigon, T., Herring, A. H., and Dunson, D. B. (2023). A generalized bayes framework for probabilistic clustering. *Biometrika*, 110:559–578. 2

Rösner, C. and Schmidt, M. (2018). Privacy preserving clustering with constraints. *CoRR*, abs/1802.02497. 1

Shorack, G. and Wellner, J. (1986). Empirical Processes with Applications to Statistics. *Wiley-Interscience, New York*. 14

Storlie, C. B., Myers, S. M., Katusic, S. K., Weaver, A. L., Voigt, R. G., Croarkin, P. E., Stoeckel, R. E., and Port, J. D. (2019). Clustering and variable selection in the presence of mixed variable types and missing data. *Statistics in Medicine*. 2, 9, 23

Strona, G., Nappo, D., Boccacci, F., Fattorini, S., and San-Miguel-Ayanz, J. (2014). A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature Communications*, 5:4114. 2, 6, 18

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581. 5, 13

Villani, C. (2008). Optimal transport: Old and new. 18, 19

Wang, G. (2020). A fast MCMC algorithm for the uniform sampling of binary matrices with fixed margins. *Electronic Journal of Statistics*, 14(1):1690 – 1706. 2, 7, 20, 21, 22

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25. 3

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]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable]
   (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). [Not Applicable]

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. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

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]

Supplementary material to "Fair Clustering via Hierarchical Fair-Dirichlet Prior"

Content

1. Section 8 contains the detailed proof of Theorem 1, presented in Section 2.2 in the main document.
2. Section 9 contains additional examples on utility of the scaled log-score to compare two competing fair clustering configurations, as presented in Section 3 in the main document.
3. Sub-section 10.1 contains the detailed derivation of the posterior computation steps, presented in section 4 in the main document.
4. Sub-section 10.2 records the Weighted Rectangular Loop algorithm in complete details, presented in section 4 in the main document.
5. In Section 11.1, we extend the proposed methodology to the case where only imperfect knowledge of the protected attribute is available Esmaeili et al. (2020), as indicated in Section 5 in the main document.
6. In Section 11.2, we extend the proposed methodology to address data complexities, such as mixed data types and censoring.
7. In Section 12, we compare the performance of HFDP relative to the existing methodologies on popular bench mark data sets from the UCI repository Dua and Graff (2017), considered before in the literature Chierichetti et al. (2017); Böhm et al. (2020); Esmaeili et al. (2020), as indicated in Section 5 in the main document.

8 Assumptions and Proof of Theorem 1 in Section 2.2 in the main document

Let $\{(\boldsymbol{x}_i, a_i)\}_{i=1}^N$ be i.i.d. samples from a true distribution $\mathcal{P}^{\star}_{X,A,Z}$ with latent cluster labels $\boldsymbol{Z} \in [K]^N$. We assume the following.

Assumption 1. ***(Finite and non-degenerate clusters.)*** *The number of clusters K is fixed, and each cluster has strictly positive probability, i.e., $\pi^{\star}_{\cdot,k} > 0$ for all $k \in [K]$.*

Assumption 2. ***(Cluster identifiability.)*** *Distinct cluster configurations correspond to distinct conditional distributions of $\boldsymbol{X}$, up to label switching.*

Assumption 3. ***(Population-level feasibility of balance constraints.)*** *For any $\varepsilon \geqslant 0$ of interest, there exists a non-trivial population distribution*

$$\mathcal{P}_{Z|X,A} \in \mathbb{P}^{(R)} = \{\mathcal{P}_{X,Z,A} \in \mathbb{P} : \mathcal{P}_{A|Z} = \mathcal{P}_A\},$$

such that the ε-balance condition is satisfied.

Assumption 4. ***(Finite moments..)*** *The data $(\boldsymbol{x}_i, a_i)$ are independent and identically distributed with finite first moments.*

Assumption 5. ***Population-level feasibility of balance constraints:*** *For any $\varepsilon \geqslant 0$ of interest, there exists a non-trivial population distribution*

$$\mathcal{P}_{Z|X,A} \in \mathbb{P}^{(R)} = \{\mathcal{P}_{X,Z,A} \in \mathbb{P} : \mathcal{P}_{A|Z} = \mathcal{P}_A\},$$

such that the ε-balance condition is satisfied.

Proof. The proof proceeds via three key steps. **Step 1** deals with sample quantities in the left hand side of the statement, **Step 2** simplifies the population quantities in the right hand side of the statement, and **Step 3** presents the limiting arguments to prove the equivalence result.

Step 1 (sample quantities)

We observe data $(\boldsymbol{X}, \boldsymbol{A})$, and want to impute the clustering indices $\boldsymbol{Z} \in [K]^N$. Suppose

$$\mathcal{P}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x}) \propto \mathbf{m}_N(\boldsymbol{z})$$

is the conditional posterior distribution of $\boldsymbol{Z} \mid \boldsymbol{X}$. Under common hierarchical model specifications, e.g Teh et al. (2006), the marginal posterior of clustering configuration $\mathbf{m}_N(\boldsymbol{z})$ often does not factorise over $k \in [K]$, owing to marginalization with respect to lower level parameters. For the sake of this proof, we assume that the lower level parameters are set at reasonable values, say, obtained via an empirical Bayesian procedure. Then, we have

$$\mathcal{P}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x}) \propto \mathbf{m}_N(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_{\cdot,k}^{N_{\cdot,k}},$$

where $N_{\cdot,k} = \sum_{i=1}^{N} \mathbf{1}(z_i = k)$ for all $k \in [K]$.

We fix $\varepsilon \geqslant 0$. Assume that the conditional distribution of $\boldsymbol{Z} \mid \boldsymbol{X}$, restricted to the ε-balanced class of clustering configurations, denoted by $\mathcal{P}'_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{A}}$, takes the form

$$\mathcal{P}'_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x}) \propto \frac{\mathbf{m}_N(\boldsymbol{z})}{W^{(N)}} \times \mathbf{1}[\boldsymbol{z}^{(N)} \in \mathbb{Z}^{(fair)}(\varepsilon)], \tag{8.1}$$

where

$$W^{(N)} = \sum_{\boldsymbol{z}^{(N)} \in \mathbb{Z}^{(fair)}(\varepsilon)} \mathbf{m}_N(\boldsymbol{z}).$$

We can sample from $\mathcal{P}_{\boldsymbol{Z}|\boldsymbol{X}}$ (e.g., via MCMC), and retain the sample if it satisfies $\boldsymbol{Z} \in \mathbb{Z}^{(fair)}(\varepsilon)$. For the retained sample, we calculate $\pi_{a,k}^{(N)}(\varepsilon) = (N_{a,k}/N)$ and collect in a matrix $\boldsymbol{\pi}^{(N)}(\varepsilon) = ((\pi_{a,k}^{(N)}(\varepsilon)))$. Assumption 1 ensures true K is finite and $N_{\cdot,k} > 0$, and the proportions $N_{\cdot,k}/N$ are well-defined.

Further, we note that the dual objective function for maximization of the restricted likelihood of $\mathcal{P}'_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x})$ takes the form

$$\begin{aligned} \mathcal{T}(\boldsymbol{z}) &:= (1/N) \log \mathbf{m}_N(\boldsymbol{z}) - \lambda \ \mathrm{KL}(\mathcal{P}_{\boldsymbol{A}} \times \mathcal{P}_{\boldsymbol{Z}} \mid\mid \mathcal{P}_{\boldsymbol{A},\boldsymbol{Z}}) \\ &= (1/N) \sum_{k=1}^{K} N_{\cdot,k} \log \pi_{\cdot,k} - \lambda \ \mathrm{KL}(\mathcal{P}_{\boldsymbol{A}} \times \mathcal{P}_{\boldsymbol{Z}} \mid\mid \mathcal{P}_{\boldsymbol{A},\boldsymbol{Z}}), \end{aligned} \tag{8.2}$$

where $\lambda \geqslant 0$ is the Lagrange multiplier that depends on ε.

Step 2. (population quantities)

Next, we uncover the quantities in Definition of pseudo true fair clustering distribution $\mathcal{P}^{\varepsilon}_{Z_{\mathrm{PT-FCD}}}$ in sub-section (2.1) in the main document. For the true generative model $\mathcal{P}^{\star}_{A,Z,X}$, we denote

$$T = Z \mid X, A = a \sim \mathrm{Categorical}(\pi^{\star}_{a,1}, \ldots, \pi^{\star}_{a,K}),$$

and collect $\boldsymbol{\pi}^{\star} = ((\pi^{\star}_{a,k}))$. For any $\mathcal{P}_{A,Z,X} \in \mathbb{P}^{(R)}_{\varepsilon}$, we denote

$$V_{\varepsilon} = Z \mid X, A = a \sim \mathrm{Categorical}(\tilde{\pi}_{a,1}(\varepsilon), \ldots, \tilde{\pi}_{a,K}(\varepsilon))$$

and collect $\tilde{\boldsymbol{\pi}}(\varepsilon) = ((\tilde{\pi}_{a,k}(\varepsilon)))$. Assumption 5 guarantees $\mathbb{P}^{(R)}_{\varepsilon}$ is non-empty. To minimize

$$\mathrm{KL}(\mathcal{P}^{\star}_{Z|X} \mid\mid \mathcal{P}_{Z|X})$$

with respect to $\mathcal{P}_{A,Z,X} \in \mathbb{P}_{\varepsilon}^{(R)}$, we can equivalently maximize

$$\sum_{k=1}^{K} \pi^{\star}_{\cdot,k} \log \tilde{\pi}_{\cdot,k}(\varepsilon) - \lambda \ \mathrm{KL}(\mathcal{P}_A \times \mathcal{P}_Z \ || \ \mathcal{P}_{A,Z}), \tag{8.3}$$

with respect to $\tilde{\boldsymbol{\pi}}(\varepsilon)$. Here, we obtain $(\pi^{\star}_{\cdot,1}, \ldots, \pi^{\star}_{\cdot,K})$ and $(\tilde{\pi}_{\cdot,1}(\varepsilon), \ldots, \tilde{\pi}_{\cdot,K}(\varepsilon))$ via marginalization of $(\pi^{\star}_{a,1}, \ldots, \pi^{\star}_{a,K})$ and $(\tilde{\pi}_{a,1}(\varepsilon), \ldots, \tilde{\pi}_{a,K}(\varepsilon))$, respectively, over A. Assumption 2 ensures the population maximizer $\tilde{\boldsymbol{\pi}}(\varepsilon)$ is well-defined and unique up to label switching.

Step 3 (equivalence)

First, we uncover the quantities in Definition of maximum-a-posteriori fair cluster $\boldsymbol{z}^{\varepsilon}_{\mathrm{MAP-FC}}$ in sub-section (2.2) in the main document. We observe that,

$$\begin{aligned} \pi^{(N)}_{a|k} = \frac{\sum_{i=1}^{N} \mathbf{1}(z_i = k, a_i = a)}{\sum_{i=1}^{N} \mathbf{1}(z_i = k)} &= \frac{[\sum_{i=1}^{N} \mathbf{1}(z_i = k, a_i = a)]/N}{[\sum_{i=1}^{N} \mathbf{1}(z_i = k)]/N} \\ &\xrightarrow{a.s} \frac{\mathcal{P}_{Z,A}(Z = k, A = a)}{\mathcal{P}_Z(Z = k)} = \mathcal{P}_{A|Z}(A = a \mid Z = k), \end{aligned}$$

$\forall\ a \in \mathcal{A},\ \forall\ k$ as $N \to \infty$, by the *Strong Law of Large Numbers*. Assumption 4 justifies application of the Strong Law of Large Numbers (SLLN) to the empirical cluster proportions. Then, by Equation (8.4), there exists a $\mathcal{P}_{X,Z,A} \in \mathbb{P}_{\varepsilon}^{(1)}$ such that

$$\mathcal{P}'_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{A}} \xrightarrow{a.s} \mathcal{P}_{X,Z,A}$$

as $N \to \infty$. Recall that we defined,

$$\mathbb{Z}^{(fair)}(\varepsilon) = \big\{\boldsymbol{Z} \in \mathbb{Z} : \mathrm{KL}(\mathcal{P}_{\boldsymbol{A}} \times \mathcal{P}_{\boldsymbol{Z}} \ || \ \mathcal{P}_{\boldsymbol{A},\boldsymbol{Z}}) \leqslant \varepsilon\big\}.$$

Then for any $\boldsymbol{Z} \in \mathbb{Z}^{(fair)}(\varepsilon)$, we have

$$\mathrm{KL}(\mathcal{P}_{\boldsymbol{A}} \times \mathcal{P}_{\boldsymbol{Z}} \ || \ \mathcal{P}_{\boldsymbol{A},\boldsymbol{Z}}) \leqslant \varepsilon \equiv \sum_{a,z} \pi^{(N)}_a \pi^{(N)}_z \ \log\left[\frac{\pi^{(N)}_a \pi^{(N)}_z}{\pi^{(N)}_{a,z}}\right] \leqslant \varepsilon \ \forall\ N, \tag{8.4}$$

which by the *Strong Law of Large Numbers* yields

$$\mathrm{KL}(\mathcal{P}_A \times \mathcal{P}_Z \ || \ \mathcal{P}_{A,Z}) \leqslant \varepsilon$$

as $N \to \infty$, $(\star)$. Assumption 4 justifies application of the Strong Law of Large Numbers (SLLN) to the empirical cluster proportions.

Next, under Assumption 3, the unconstrained log marginal likelihood of $\boldsymbol{z}$, $(1/N) \log \mathbf{m}_N(\boldsymbol{z}) \xrightarrow{a.s} \sum_{k=1}^{K} \pi^{\star}_{\cdot,k} \log \tilde{\pi}_{\cdot,k}(\varepsilon)$. This together with $\star$ implies

$$(1/N) \log \mathbf{m}_N(\boldsymbol{z}) - \lambda \ \mathrm{KL}(\mathcal{P}_{\boldsymbol{A}} \times \mathcal{P}_{\boldsymbol{Z}} \ || \ \mathcal{P}_{\boldsymbol{A},\boldsymbol{Z}}) \xrightarrow{a.s} \sum_{k=1}^{K} \pi^{\star}_{\cdot,k} \log \tilde{\pi}_{\cdot,k}(\varepsilon) - \lambda \ \mathrm{KL}(\mathcal{P}_A \times \mathcal{P}_Z \ || \ \mathcal{P}_{A,Z}), \tag{8.5}$$

as $N \to \infty$.

The objective function $\mathcal{T}(\boldsymbol{z})$ in Equation (8.2) is maximised with respect to $\boldsymbol{z}$. The maxima is attained at a $\boldsymbol{z}$ with proportion of $(k, a) \in [K] \times [r]$ equal to $\boldsymbol{\pi}^{(N)}(\varepsilon)$, by definition. Suppose the objective function in Equation (8.5) is maximised with respect to $\tilde{\boldsymbol{\pi}}(\varepsilon)$ at $\tilde{\boldsymbol{\pi}}(\varepsilon) = \tilde{\boldsymbol{\pi}}_m(\varepsilon)$. Next, we invoke the *argmax continuous mapping theorem* Shorack and Wellner (1986), and we have $\boldsymbol{\pi}^{(N)}(\varepsilon) \xrightarrow{P} \tilde{\boldsymbol{\pi}}_m(\varepsilon)$. Recall that the map $\mathcal{R} : [K]^N \to [0,1]^K$ takes a clustering configuration $\boldsymbol{z} \in [K]^N$ as input and outputs the relative proportion of cluster indices $(1/N)[\sum_{i=1}^{N} \mathbf{1}(z_i = 1), \ldots, \sum_{i=1}^{N} \mathbf{1}(z_i = K)]^{\intercal}$. The marginal posterior of $(\boldsymbol{Z} \mid \boldsymbol{X})$ along with the reparameterization $\boldsymbol{z} \to \mathcal{R}(\boldsymbol{z})$, yields the marginal posterior of $(\mathcal{R}(\boldsymbol{Z}) \mid \boldsymbol{X})$, denoted by $P_{\mathcal{R}(\boldsymbol{Z})|\boldsymbol{X}}$. Hence, we have the proof.

□

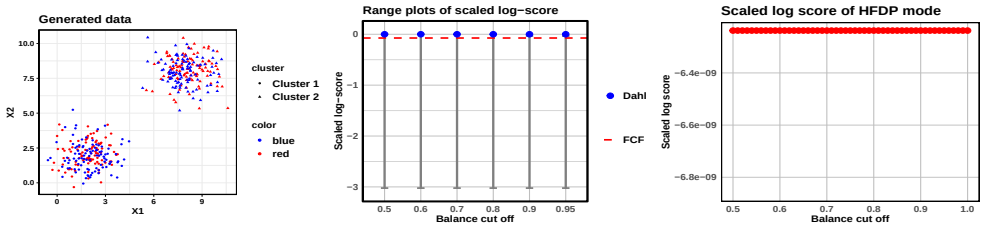

Figure 8: ***Left panel*** *presents the generated data with* $N_{11} = N_{21} = N_{12} = N_{22} = 100$ *and* $\mu_{11} = \mu_{21} = (2,2)'$, $\mu_{12} = \mu_{22} = (8,8)'$, *and* $S = I_2$. ***Middle panel*** *presents the range of the scaled log-score of the posterior clustering configurations that has balance* $\geqslant$ *cut-off* $\in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$, *and the scaled log-score of the modal clustering configuration computed via Dahl's method (Dahl, 2006).* ***Right panel*** *displays that the scaled log-scores for the modal posterior clustering configurations, where the balance satisfies* $\geqslant$ *cut-off* $\in \{0.5, 0.51, \ldots, 1.00\}$.

9 Remaining Examples from Section 3.2 in the main document

In each of the examples, we generate a data set with two labels of the protected attribute and two clusters. In the i-th cluster, N_{ji} individuals with $a = j$ are generated from $N_2(\mu_{ji}, S)$, where $i \in \{1,2\}$ and $j \in \{1,2\}$. Further, we assume that the individuals with the two labels of the protected attribute A, i.e. $a = 1$ and $a = 2$, are equally represented in the observed sample of size $N = \sum_{i=1}^{2}\sum_{j=1}^{2} N_{ji} = 400$.

Example 1. Here, we generate data with $N_{11} = N_{21} = N_{12} = N_{22} = 100$ and $\mu_{11} = \mu_{21} = (2,2)'$, $\mu_{12} = \mu_{22} = (8,8)'$, and $S = I_2$, as presented in the left panel in Figure 8. This generating mechanism has two key features. Firstly, any reasonable clustering approach to cluster the observed dataset into $K = 2$ clusters, even without balance constraints, will yield two well-separated and perfectly balanced clusters. Secondly, within each cluster, the two attribute specific densities are identical and are isotropic normal densities. Recall, fair clustering via fairlets Chierichetti et al. (2017) is computed via a two stage procedure - first finding an optimal fairlet decomposition of the data and representing each of the fairlets via a fairlet center, and then resorting to popular clustering algorithms, e.g k-means, to cluster the fairlet centers to get balanced clusters. The data generative mechanism in this example ensures that the fairlet centers are approximately spherically distributed. As a result, both the fairlet construction and the clustering of the fairlet centers via k-means clustering based on the L_2 metric are appropriate in this example.

Now, suppose the goal is indeed to obtain $K = 2$ clusters, that are completely balanced. To that end, while we could have simply resorted to a clustering mechanism without a balance constraint, we first employ the fair clustering methodology to be proposed in sequel, and compute the balance and the scaled log-score at each posterior draw of clustering configurations. In Figure 8, as we increase the balance cut-off, the scaled log-score of the modal clustering configuration remains constant, since the methodology does not produce clusters with low balance, under the specific data generating mechanism. Further, both the perfectly balanced modal clustering configuration and the optimal fair clustering configuration via fairlets yield similar scaled log-score under the true data generating mechanism.

Example 2. Here, we generate data with $N_{11} = N_{21} = N_{12} = N_{22} = 100$ and $\mu_{11} = (2,2)', \mu_{21} = (3,3)'$, $\mu_{12} = (8,8)^{\mathrm{T}}, \mu_{22} = (9,9)'$, and $S = I_2$. This generating mechanism is similar to Example 1, except that if one adopts the fair clustering via fairlets approach, the fairlet centers will not be spherically distributed. As a result, the clustering of the fairlet centers via k-means clustering based on the L_2 metric is not appropriate in this example. However, since the method we propose in the paper adapts to the shape of the cluster, the corresponding perfectly balanced modal clustering configuration yield a slightly higher scaled log-score under the true data generating mechanism, compared to the optimal fair clustering configuration via fairlets. Refer to Figure 9 for details.

Example 3. Here, we generate data with $N_{11} = N_{21} = N_{12} = N_{22} = 100$ and $\mu_{11} = \mu_{21} = (2,2)'$, $\mu_{12} = \mu_{22} = (8,8)'$, and $S = [\rho 11^{\mathrm{T}} + (1-\rho)I_2]$, with $\rho = 0.3$. This generating mechanism is similar to Example 2, except that the within cluster attribute specific densities are anisotropic. As a result, both the fairlet construction and the clustering of the fairlet centers via k-means clustering based on the L_2 metric are not appropriate in this

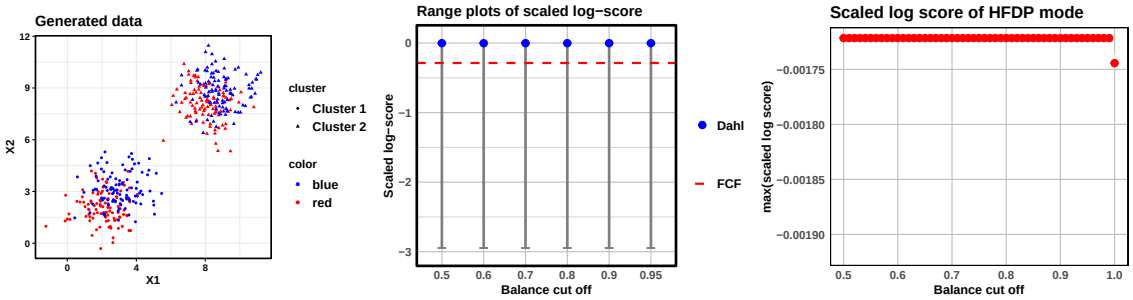

Figure 9: ***Left panel*** *presents the generated data with* $N_{11} = N_{21} = N_{12} = N_{22} = 100$ *and* $\mu_{11} = (2,2)', \mu_{21} = (3,3)'$, $\mu_{12} = (8,8)^{\mathrm{T}}, \mu_{22} = (9,9)'$, *and* $S = I_2$. ***Middle panel*** *and* ***Right panel*** *displays same quantities as in Figure 1 in the main text.*

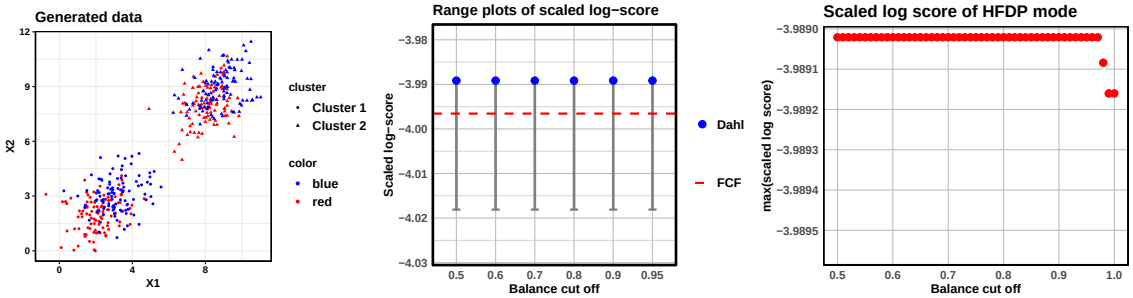

Figure 10: ***Left panel*** *presents the generated data with* $N_{11} = N_{21} = N_{12} = N_{22} = 100$ $\mu_{11} = (2,2)', \mu_{21} = (3,3)'$, $\mu_{12} = (8,8)^{\mathrm{T}}, \mu_{22} = (9,9)'$, *and* $S = [\rho 11^{\mathrm{T}} + (1-\rho)I_2]$, *with* $\rho = 0.3$. ***Middle panel*** *and* ***Right panel*** *displays same quantities as in Figure 1 in the main text.*

example. Consequently, the perfectly balanced modal clustering configuration obtained via the method proposed in the paper yield a higher scaled log-score under the true data generating mechanism, compared to the optimal fair clustering configuration via fairlets. Refer to Figure 10 for details.

10 Posterior Computation details

10.1 Full Conditionals in section 4 in the main document

For Dirichlet process mixture models, it is well known that blocked Gibbs sampling has significant improvements in terms of both mixing and scalability over the Chinese restaurant process (CRP) based collapsed samplers. The blocked Gibbs algorithm replaces the infinite dimensional Dirichlet process prior by its finite dimensional approximation allowing the model parameters to be expressed entirely in terms of a finite number of random variables, which are updated in blocks from simple multivariate distributions. Instead of adapting the CRF-based samplers to make them efficient and scalable, we adopt a blocked Gibbs sampler for our model in this paper, which will adhere to both improved mixing and scaling over large sample sizes. Starting with the joint distribution of all parameters described in equations (4)-(6) in the main document, along with specific choices, we intend to analytically integrate out cluster specific parameters $\{\boldsymbol{\phi}_k,\ k \in [K]\}$, and intermediate level concentration parameter α_0 , wherever possible, to develop an efficient Collapsed Gibbs sampler, or an MC-EM algorithm, to sample from the posterior of the remaining parameters. The details of the derivation of the sampler are presented here.

The joint posterior of all parameters in equations (4)-(6) in the main document is as follows

$$
\begin{aligned}
&\pi(\boldsymbol{\beta},\ \alpha_0, \{\boldsymbol{w}^{(a)},\ \boldsymbol{z}^{(a)},\ \boldsymbol{\mu}_{(a)},\ \Sigma_{(a)}\ \}_{a=1}^{r} \mid \{(\boldsymbol{x}_i, a_i)\}_{i=1}^{N}) \\
\propto\ & \left\{ \prod_{a=1}^{r} \prod_{i=1}^{N_a} [\boldsymbol{x}_{(\sum_{j=1}^{a-1} n_j)+i} \mid \boldsymbol{\mu}^{(a)}, \Sigma^{(a)}, \boldsymbol{z}^{(a)}] \right\} \left\{ \prod_{a=1}^{r} \prod_{k=1}^{K} [\boldsymbol{\mu}_k^{(a)}, \Sigma_k^{(a)} \mid \boldsymbol{\mu}_0^{(a)}, \Sigma_0^{(a)}] \right\} \\
& \left\{ \prod_{a=1}^{r} [\boldsymbol{\mu}_0^{(a)}, \Sigma_0^{(a)}] \right\} \left\{ \prod_{a=1}^{r} [\boldsymbol{z}^{(a)} \mid \boldsymbol{m}^{(a)}] \right\} \left\{ \prod_{a=1}^{r} [\boldsymbol{w}^{(a)} \mid \alpha_0, \boldsymbol{\beta}] \right\} [\alpha_0 \mid g]\ [\boldsymbol{\beta}] \\
\propto\ & \left\{ \prod_{a=1}^{r} \prod_{i=1}^{N_a} w_{z_i^{(a)}}^{(a)} \mathrm{N}_d\big(\boldsymbol{x}_{(\sum_{j=1}^{a-1} n_j)+i} \mid \boldsymbol{\mu}_{z_i^{(a)}}^{(a)}, \Sigma_{z_i^{(a)}}^{(a)}\big) \right\} \\
& \left\{ \prod_{a=1}^{r} \prod_{k=1}^{K} \mathrm{NIW}(\boldsymbol{\mu}_k^{(a)}, \Sigma_k^{(a)} \mid \mu_0^{(a)}, \lambda_0^{(a)}, \Lambda_0^{(a)}, \nu_0^{(a)}) \right\} \left\{ \prod_{a=1}^{r} \frac{\prod_{k=1}^{K} m_k^{(a)}!}{n_a!} \right\} \\
& \left\{ \prod_{a=1}^{r} \mathrm{Dir}(\boldsymbol{w}^{(a)} \mid \alpha_0 \boldsymbol{\beta}) \right\} \mathrm{Gamma}(\alpha_0 \mid g, b)\ \mathrm{Dir}(\boldsymbol{\beta} \mid g/K, \ldots, g/K).
\end{aligned}
$$

In order to develop a computationally efficient blocked Gibbs sampler, we marginalise out the joint posterior with respect to the component specific parameters $\boldsymbol{\mu}^{(a)}, \Sigma^{(a)}, a \in [r]$, and obtain the marginal posterior:

$$
\begin{aligned}
&\pi(\boldsymbol{\beta},\ \alpha_0, \{\boldsymbol{w}^{(a)},\ \boldsymbol{z}^{(a)}\}_{a=1}^{r} \mid \{(\boldsymbol{x}_i, a_i)\}_{i=1}^{N}) \\
\propto\ & \left\{ \prod_{a=1}^{r} \prod_{k=1}^{K} \frac{\Gamma_d(\nu_k^{(a)}/2)\ (\lambda_0^{(a)})^{d/2}\ |\Lambda_0^{(a)}|^{\nu_0^{(a)}/2}}{\Gamma_d(\nu_0^{(a)}/2)\ (\lambda_k^{(a)})^{d/2}\ |\Lambda_k^{(a)}|^{\nu_k^{(a)}/2}} \right\} \\
& \left\{ \frac{(\Gamma(\alpha_0))^r}{\prod_{k=1}^{K} (\Gamma(\alpha_0 \beta_k))^r} \prod_{a=1}^{r} \prod_{k=1}^{K} (w_k^{(a)})^{\alpha_0 \beta_k + m_k^{(a)} - 1} \right\} \left\{ \prod_{k=1}^{K} \beta_k^{\frac{g}{K} - 1} \right\} \alpha_0^{g-1} e^{-b\alpha_0}
\end{aligned}
$$

where we have

$$
\begin{aligned}
&\lambda_k^{(a)} = \lambda_0^{(a)} + m_k^{(a)}; \quad \nu_k^{(a)} = \nu_0^{(a)} + m_k^{(a)}; \\
&\boldsymbol{\mu}_k^{(a)} = \frac{\lambda_0^{(1)} \boldsymbol{\mu}_0^{(a)}}{\lambda_0^{(a)} + m_k^{(a)}} + \frac{\sum_{i=1}^{N_a} \delta(z_i^{(a)} = k)\ \boldsymbol{x}_{(\sum_{j=1}^{a-1} n_j)+i}}{\lambda_0^{(a)} + m_k^{(a)}}; \\
&\Lambda_k^{(a)} = \Lambda_0^{(a)} + \sum_{i=1}^{N_a} \delta(z_i^{(a)} = k)\ (\boldsymbol{x}_{(\sum_{j=1}^{a-1} n_j)+i} - \bar{\boldsymbol{x}}_k^{(a)})(\boldsymbol{x}_{(\sum_{j=1}^{a-1} n_j)+i} - \bar{\boldsymbol{x}}_k^{(a)})^{\mathrm{T}} + \\
&\frac{\lambda_0^{(a)} m_k^{(a)}}{\lambda_0^{(a)} + m_k^{(a)}} (\bar{\boldsymbol{x}}_k^{(a)} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_k^{(a)} - \bar{\boldsymbol{x}})^{\mathrm{T}}; \quad k \in [K],\ a \in [r].
\end{aligned}
$$

The blocked parameter updates, and details of the computational strategies are enlisted next. Letting $\theta \mid -$ denote the full conditional distribution of a parameter θ given other parameters and the data, the Gibbs sampler (or an **MC-EM** algorithm) cycles through the following steps, sampling parameters from their full conditional distributions:

Step 1: To sample from $[\alpha_0 \mid \cdot]$, we note that

$$
\begin{aligned}
[\alpha_0 \mid \cdot] &\propto [\alpha_0 \mid \boldsymbol{\beta}, \{\boldsymbol{w}^{(a)}\}_{a=1}^{r}] \\
&\propto \left\{ \frac{(\Gamma(\alpha_0))^r}{\prod_{k=1}^{K} (\Gamma(\alpha_0 \beta_k))^r} \prod_{a=1}^{r} \prod_{k=1}^{K} (w_k^{(a)})^{\alpha_0 \beta_k} \right\} \alpha_0^{g-1} e^{-b\alpha_0}
\end{aligned}
$$

where $0 < \alpha_0 < \infty$. We can evaluate the density over a grid and sample from the resulting discrete distribution. In practice, our choice of the hyper-parameter b dictates how far we need to go in the positive real line to

construct our grid. Noteworthy, α_0 is one-dimensional, and it can be easily sampled from its posterior via an Metropolis–Hastings update with independent proposals.

Step 2: To sample from $[\boldsymbol{\beta} \mid \cdot]$, we note that

$$[\boldsymbol{\beta} \mid \cdot] \propto [\boldsymbol{\beta} \mid \alpha_0, \{\boldsymbol{w}^{(a)}\}_{a=1}^r] \propto (\Gamma(\alpha_0))^r \frac{\prod_{k=1}^K (\beta_k^{\frac{g}{K}-1} \prod_{a=1}^r (w_k^{(a)})^{\alpha_0\beta_k})}{\prod_{k=1}^K (\Gamma(\alpha_0\beta_k))^r}.$$

The dimension K of the global weights can be potentially large in practice since the distinct global atoms are shared by all the groups, which poses challenges for sampling β. To tackle that, we marginalise out α_0 from the joint prior on $(\alpha_0, \beta, \{\boldsymbol{w}\}_{a=1}^r)$, which yields

$$[\boldsymbol{\beta}, \{\boldsymbol{w}^{(a)}\}_{a=1}^r] \propto \int_0^\infty t^{K-1} \, [\Gamma(t)]^r \frac{\prod_{k=1}^K [e^{-bt\beta_k} (t\beta_k)^{\frac{g}{K}-1} \prod_{a=1}^r (w_k^{(a)})^{t\beta_k}]}{\prod_{k=1}^K [\Gamma(t\beta_k)]^r} \, dt.$$

The parameters β can be observed to be the normalized version of the random vector $\boldsymbol{t} = (t_1, \ldots, t_K)$,

$$[\boldsymbol{t}, \{\boldsymbol{w}^{(a)}\}_{a=1}^r] \propto t^{K-1} \left[\Gamma\Big(\sum_{k=1}^K t_k\Big)\right]^r \frac{\prod_{k=1}^K [e^{-bt_k} t_k^{\frac{g}{K}-1} \prod_{a=1}^r (w_k^{(a)})^{t_k}]}{\prod_{k=1}^K [\Gamma(t_k)]^r}.$$

Note that, the term $[\Gamma(\sum_{k=1}^K t_k)]^r$ above, prevents the posterior density of $\boldsymbol{t}$ from factorizing over k. To deal the inconvenience, we observe that

$$\left[\Gamma(\sum_{k=1}^K t_k)\right]^r = \prod_{i=1}^r \left[\int_0^\infty e^{-u_i} u_i^{\sum_{k=1}^K t_k - 1} du_i\right],$$

and introduce a slice using auxiliary random variables $\boldsymbol{u} = (u_1, u_2, \ldots, u_r)$ that yield the full conditionals:

$$[\boldsymbol{u} \mid \cdot] \propto \prod_{i=1}^r \text{Gamma}\Big(u_i \Big| \sum_{k=1}^K t_k, 1\Big); \quad [\boldsymbol{t} \mid \cdot] \propto \prod_{k=1}^K f_k(t_k),$$

where

$$f_k(t) \propto \frac{1}{[\Gamma(t)]^r} t^{\frac{g}{K}-1} e^{-t(b - \log p_k - \sum_{i=1}^r \log u_i)} \text{ with } p_k = \prod_{a=1}^r w_k^{(a)}, \; k \in [K].$$

We can obtain exact samples from the above log-concave density via a simple rejection sampler equipped with a well-designed covering density Leydold (1998). Note that, sampling from the full conditional distributions of $\{t_k, \; k \in [K]\}$ are inherently parallelizable. Finally, we simply obtain $\beta_k = \frac{t_k}{\sum_{k=1}^K t_k}$, $k \in [K]$.

Step 3: To sample from $[\boldsymbol{w}^{(a)} \mid \cdot]$, $a \in [r]$ independently, we note that

$$[\boldsymbol{w}^{(a)} \mid \cdot] \sim \text{Dir}(\alpha_0\beta_1 + m_1^{(a)}, \ldots, \alpha_0\beta_K + m_K^{(a)}),$$

and set $\boldsymbol{m}^{(a)} = \text{rd}(n^{(a)} \times \boldsymbol{w}^{(a)}), \quad a \in [r]$.

Step 4. The marginal conditional of clustering indices $[\boldsymbol{z}^{(a)} \mid -]$, $a \in [r]$, integrating out population parameters, is

$$[\boldsymbol{z}^{(a)} \mid -] \propto \prod_{k=1}^K \frac{\Gamma_d(\nu_k^{(a)}/2) \; (\lambda_0^{(a)})^{d/2} \; |\Lambda_0^{(a)}|^{\nu_0^{(a)}/2}}{\Gamma_d(\nu_0^{(a)}/2) \; (\lambda_k^{(a)})^{d/2} \; |\Lambda_k^{(a)}|^{\nu_k^{(a)}/2}}, \quad \boldsymbol{z}^{(a)} \in \mathcal{Z}_{N_a, K, \boldsymbol{m}^{(a)}} \quad a \in [r]. \tag{10.1}$$

Sampling (10.1) poses a difficult combinatorial problem and is the most substantial computational bottleneck in our algorithm. We circumnavigate this issue via recasting the problem as a non-uniform sampling task from the space of binary matrices with fixed margin Miller and Harrison (2013); Brualdi and Ross (1980); Strona et al. (2014), and propose a novel sampling scheme equipped with an MH proposal based on integer-valued optimal transport Villani (2008).

As an intermediate step, we first develop a computationally convenient MC-EM algorithm Levine and Casella (2001), where instead of sampling from $[\boldsymbol{z}^{(a)} \mid -]$, $a \in [r]$, we update the chain with the posterior mode of $[\boldsymbol{z}^{(a)} \mid -]$, $a \in [r]$. In particular, to compute the posterior mode of $[\boldsymbol{z}^{(a)} \mid -]$, $a \in [r]$, we go over the following steps. **(i)** First, for $a \in [r]$, we calculate the current component specific means and variances $\boldsymbol{\mu}_{k,\star}^{(a)}, \Sigma_{k,\star}^{(a)}$ for all $a \in [r], k \in [K]$. **(ii)** Next, we define the $N_a \times K$ cost matrix $\mathrm{L}^{(a)} = ((l_{ik})) = \big(\big(-\log \mathrm{N}_d(\boldsymbol{x}_i^{(a)} \mid \boldsymbol{\mu}_{k,\star}^{(a)}, \Sigma_{k,\star}^{(a)})\big)$, column sum vectors $\boldsymbol{c}^{(a)} = \boldsymbol{m}^{(a)}$ and row sum vectors $\boldsymbol{r}^{(a)} = \mathbf{1}_{N_a}$, where $\mathbf{1}_{N_a}$ is a vector of N_a 1s. **(iii)** Next, given the two vectors $\boldsymbol{r}^{(a)}, \boldsymbol{c}^{(a)}$, we define the polytope of $K \times N_a$ binary cluster membership matrices $\mathrm{U}(\boldsymbol{r}^{(a)}, \boldsymbol{c}^{(a)}) := \{\mathrm{B} \mid \mathrm{B}\mathbf{1}_{N_a} = \boldsymbol{r}^{(a)};\ \mathrm{B}^{\mathrm{T}} 1_K = \boldsymbol{c}^{(a)}\}$, and solve the constrained binary optimal transport problem Villani (2008) $\mathrm{B}^{(a)} = \mathrm{argmin}_{\mathrm{B} \in \mathrm{U}(\boldsymbol{r}^{(a)},\ \boldsymbol{c}^{(a)})} \langle \mathrm{B}, \mathrm{L}^{(a)} \rangle$ where $\langle \mathrm{B}, \mathrm{L}^{(a)} \rangle = \mathrm{tr}(\mathrm{B}^{\mathrm{T}} \mathrm{L}^{(\mathrm{a})})$. Finally, we obtain the modal clustering indices $\boldsymbol{z}^{(a)}$ from the binary cluster membership matrices $\mathrm{B}^{(a)}, a \in [r]$, completing the MC-EM algorithm.

Algorithm 1: Gibbs sampler for HFDP and post-processing

Input: Data $\mathcal{D} = \{(\boldsymbol{x}_i, a_i) \in \mathcal{X} \times \mathcal{A},\ i = 1 \ldots, N\}$,
Number of categorical values the protected attribute a can take $= r$,
Initial values $\{\alpha_0^{(0)}, \boldsymbol{\beta}^{(0)}, \boldsymbol{w}^{(a,0)}, \boldsymbol{z}^{(a,0)}, \{\phi_k^{(a,0)}\}_{k=1}^K\},\ a \in [r]$,
Number of iterations T, Number of burn in $T_{\mathrm{burn-in}}$,
Target balance $\mathrm{B}_{\mathrm{target}}$ and/or discrepancy ε
Output: Posterior clustering configurations over T iterations, final post-processed clustering $\boldsymbol{z}^{\mathrm{final}}$

Hyper-parameter tuning. Set the hyper-parameters (b, g) of the prior on α such that $\mathrm{KL}(\mathcal{P}_{\mathbf{a}} \times \mathcal{P}_{\mathbf{z}} \,\|\, \mathcal{P}_{\mathbf{a},\mathbf{z}}) \leqslant \varepsilon$ with probability between $90 - 95\%$. For $r = 2$, alternatively, set (b, g) such that $\mathrm{KL}(\mathcal{P}_{\mathbf{w}^{(1)}} \times \mathcal{P}_{\mathbf{w}^{(2)}} \,\|\, \mathcal{P}_{\mathbf{w}^{(1)},\mathbf{w}^{(2)}}) \leqslant \varepsilon$ with probability between $90 - 95\%$.

Gibbs sampler.

for $t = 1$ **to** $T_{\mathrm{burn-in}} + T$ **do**
  Sample $\alpha_0^{(t)} \sim [\alpha_0 \mid \boldsymbol{\beta}^{(t-1)}, \boldsymbol{w}^{(a,t-1)}, \boldsymbol{z}^{(a,t-1)}, \{\phi_k^{(a,t-1)}\}_{k=1}^K],\ a \in [r]$
  Sample $\boldsymbol{\beta}^{(t)} \sim [\boldsymbol{\beta} \mid \alpha_0^{(t)}, \boldsymbol{w}^{(a,t-1)}, \boldsymbol{z}^{(a,t-1)}, \{\phi_k^{(a,t-1)}\}_{k=1}^K],\ a \in [r]$
  for $a = 1$ **to** r **do**
    Sample $\boldsymbol{w}^{(a,t)} \sim [\boldsymbol{w}^{(a)} \mid \alpha_0^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{z}^{(a,t-1)}, \{\phi_k^{(a,t-1)}\}_{k=1}^K]$
    Sample $\boldsymbol{z}^{(a,t)} \sim [\boldsymbol{z}^{(a)} \mid \boldsymbol{w}^{(a,t)}, \{\phi_k^{(a,t-1)}\}_{k=1}^K]$
    for $k = 1$ **to** K **do**
      Sample $\phi_k^{(a,t)} \sim [\phi_k^{(a)} \mid \boldsymbol{z}^{(a,t)}, \mathcal{D}]$
    end
  end
end

Post-processing to ensure strict balance constraint.

1. Compute the balance and scaled log-score for posterior clustering configurations $\{z^{(a,t)},\ a \in [r]\},\ t = T_{\mathrm{burn-in}} + 1, \ldots,\ T + T_{\mathrm{burn-in}}$.
2. Select the posterior clustering configuration that maximizes the log-score satisfying the balance constraint $\geqslant \mathrm{B}_{\mathrm{target}}$, denoted by $\boldsymbol{z}^{\mathrm{final}}$.

10.2 Weighted Rectangular Loop Algorithm (W-RLA) and Proof of Theorem 2 in the main document

Given fixed row sums $\mathbf{r} = (r_1, \ldots, r_u)^{\mathrm{T}}$ and column sums $\mathbf{c} = (c_1, \ldots, c_v)^{\mathrm{T}}$, we denote the collection of all $u \times v$ binary matrices by $\mathrm{U}(\mathbf{r}, \mathbf{c})$. Further, we denote by $\Omega = (\omega_{ij}) \in [0, \infty)$ a non negative weight matrix representing the relative probability of observing a count of 1 at the (i, j)-th cell. Then the likelihood associated with the observed binary matrix $H \in \mathrm{U}(\mathbf{r}, \mathbf{c})$ as

$$\mathrm{P}(H) = (1/\kappa) \prod_{i,j} \omega_{ij}^{h_{ij}},\ \kappa = \sum_{H \in \mathrm{U}(\mathbf{r},\mathbf{c})} \prod_{i,j} \omega_{ij}^{h_{ij}}.$$

Let $\mathrm{U}'(\mathbf{r},\mathbf{c}) = \{H \in \mathrm{U}(\mathbf{r},\mathbf{c}) : \ P(H) > 0\}$ denote the subset of matrices in $\mathrm{U}(\mathbf{r},\mathbf{c})$ with positive probability. Then, for $H_1, H_2 \in \mathrm{U}'(\mathbf{r},\mathbf{c})$, the relative probability of the two observed matrices is $\mathrm{P}(H_1)/\mathrm{P}(H_2) = \prod_{\{i,j:h_{1,ij}=1,h_{2,ij}=0\}} \omega_{ij}^{h_{1,ij}} / \prod_{\{i,j:h_{1,ij}=0,h_{2,ij}=1\}} \omega_{ij}^{h_{2,ij}}$. With these notations, we are all set to introduce a *Weighted Rectangular Loop Algorithm* (W-RLA) for non-uniform sampling from the space of fixed margin binary matrices $H \in \mathrm{U}(\mathbf{r},\mathbf{c})$, given the weight matrix $\Omega = (\omega_{ij}) \in [0,\infty)$. To that end, let us first record that the identity matrix of order 2, and the 2×2 matrix with all zero diagonal entries and all one off-diagonal entries, are referred to as *checker-board* matrices.

Algorithm 2: Weighted rectangular loop algorithm

Input: An initial binary matrix A_0, and total number of iterations T.
for $t = 1, \ldots, T$ **do**
  ; `/* Try to find a 2 × 2 checker-board. */`
  Choose one row and one column (r_1, c_1) uniformly at random.
  if $A_{t-1}(r_1, c_1) = 1$ **then**
    Choose one column c_2 at random among all the 0 entries in r_1.
    Choose one row r_2 at random among all the 1 entries in c_2.
  else
    Choose one row r_2 at random among all the 1 entries in c_1.
    Choose one column c_2 at random among all the 0 entries in r_2.
  ; `/* If we find a checker-board, we may want to swap. */`
  if *The sub-matrix extracted from* r_1, r_2, c_1, c_2 *is a checkerboard unit* **then**
    Obtain B_t from A_{t-1} by swapping the checker-board.
    Calculate $p_t = \frac{P(B_t)}{P(B_t)+P(A_{t-1})}$.
    Draw $r_t \sim \text{Bernoulli}(p_t)$.
    if $r_t = 1$ **then**
      Set $A_t = B_t$.
    else
      Set $A_t = A_{t-1}$
  else
    $A_t = A_{t-1}$

The Theorem 2 in the main document states that, given an weight matrix Ω with $\omega_{ij} > 0$, and fixed row-sum and column sum vectors $(\boldsymbol{r},\boldsymbol{c})$, the Weighted Rectangular Loop Algorithm (W-RLA) generates a Markov chain with stationary distribution given by $\mathrm{P}(H) = (1/\kappa)\prod_{i,j}\omega_{ij}^{h_{ij}}$, $\kappa = \sum_{H\in\mathrm{U}(\mathbf{r},\mathbf{c})} \prod_{i,j}\omega_{ij}^{h_{ij}}$.

Proof of Theorem 2.

Proof. We write the proof by combining the argument of Theorem 1 in Fout et al. (2020) and the proof of Theorem 2 in Wang (2020).

Step 1. Proposal kernel.
The Weighted Rectangle Loop Algorithm first proposes a move exactly as in the uniform Rectangle Loop Algorithm. That is, starting from a configuration $A \in \mathcal{U}'(\mathbf{r},\mathbf{c})$, the proposal distribution $Q(A,\cdot)$ is supported only on configurations B that differ from A on a 2×2 submatrix with a checkerboard pattern. Such pairs (A,B) are called *swappable configurations*.

The mechanism for choosing B is unchanged from the uniform case: one picks a row and column uniformly, identifies a checkerboard, and attempts to flip it. Consequently, the proposal probability $Q(A,B)$ coincides with $P_r(A,B)$ from Wang (2020). Concretely, if A and B differ in rows i_1, i_2 and columns j_1, j_2, then

$$Q(A,B) = P_r(A,B) = \frac{1}{mn}\left(\frac{1}{n-r_{i_1}}\cdot\frac{1}{c_{j_2}} + \frac{1}{c_{j_2}}\cdot\frac{1}{n-r_{i_2}} + \frac{1}{n-r_{i_2}}\cdot\frac{1}{c_{j_1}} + \frac{1}{c_{j_1}}\cdot\frac{1}{n-r_{i_1}}\right),$$

where r_i is the sum of row i and c_j the sum of column j. The form of this expression reflects the fact that the move can be initiated from any of the four vertices of the checkerboard, and each contributes a probability term.

A key fact, proved in Wang (2020), is that this proposal is *symmetric*: for all swappable A, B,

$$Q(A, B) = Q(B, A).$$

This symmetry will simplify the Metropolis–Hastings acceptance ratio below.

Step 2. Target distribution.
We target the non-uniform distribution

$$\pi(H) \propto \prod_{i,j} \omega_{ij}^{h_{ij}}, \qquad H \in \mathcal{U}'(\mathbf{r}, \mathbf{c}),$$

where $\Omega = (\omega_{ij}) \geqslant 0$ is the given weight matrix. When $\omega_{ij} \equiv 1$ this reduces to the uniform law on $\mathcal{U}'(\mathbf{r}, \mathbf{c})$.

Step 3. Ratio of target densities for a single checkerboard flip.

To make the acceptance step concrete, compute the ratio $\dfrac{\pi(B)}{\pi(A)}$ when A, B differ only on the 2×2 submatrix given by rows i_1, i_2 and columns j_1, j_2. Without loss of generality assume the 2×2 submatrix in A is the checkerboard

$$A|_{\{i_1,i_2\}\times\{j_1,j_2\}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and the flipped matrix B has

$$B|_{\{i_1,i_2\}\times\{j_1,j_2\}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

All other entries are identical in A and B. Hence only four weights appear in the ratio:

$$\frac{\pi(B)}{\pi(A)} = \frac{\omega^1_{i_1 j_2}\, \omega^1_{i_2 j_1}}{\omega^1_{i_1 j_1}\, \omega^1_{i_2 j_2}} = \frac{\omega_{i_1 j_2}\, \omega_{i_2 j_1}}{\omega_{i_1 j_1}\, \omega_{i_2 j_2}}.$$

If the checkerboard pattern in A is the other orientation, or if the roles of A and B are reversed, the same algebra applies with indices permuted. Thus, the acceptance probability depends only on these four weights.

Step 4. Acceptance probability and transition kernel.
Using the standard Metropolis–Hastings rule with symmetric proposal Q, the acceptance probability from A to B is

$$\alpha(A, B) = \min\Big(1, \frac{\pi(B)}{\pi(A)}\Big) = \min\Big(1, \frac{\omega_{i_1 j_2}\, \omega_{i_2 j_1}}{\omega_{i_1 j_1}\, \omega_{i_2 j_2}}\Big).$$

Therefore, the transition probability from A to B in one step of W-RLA is

$$P(A, B) = Q(A, B)\, \alpha(A, B) = Q(A, B) \cdot \frac{\pi(B)}{\pi(A) + \pi(B)},$$

and the probability of staying at A is

$$P(A, A) = 1 - \sum_{B \neq A} P(A, B).$$

Note that, the expression $P(A, B) = Q(A, B)\, \frac{\pi(B)}{\pi(A)+\pi(B)}$ is algebraically equivalent to $Q(A, B) \min(1, \pi(B)/\pi(A))$ but is convenient for symmetric presentation and for verifying detailed balance.

Step 5. Detailed balance and stationarity.
For any swappable A, B the following holds:

$$\begin{aligned} \pi(A)P(A, B) &= \pi(A)Q(A, B)\, \frac{\pi(B)}{\pi(A) + \pi(B)} = Q(A, B)\, \frac{\pi(A)\pi(B)}{\pi(A) + \pi(B)} \\ &= Q(B, A)\, \frac{\pi(B)\pi(A)}{\pi(B) + \pi(A)} = \pi(B)Q(B, A)\, \frac{\pi(A)}{\pi(B) + \pi(A)} = \pi(B)P(B, A). \end{aligned}$$

Hence, the W-RLA transition kernel satisfies detailed balance with respect to π. Detailed balance implies reversibility, and therefore π is a stationary distribution of the Markov chain defined by W-RLA.

Step 6. Irreducibility and aperiodicity.
To conclude that π is the unique stationary distribution and that the chain converges to π from any starting state in $\mathcal{U}'(\mathbf{r}, \mathbf{c})$, we check irreducibility and aperiodicity.

Irreducibility. It is a standard fact in the literature on binary matrices with fixed margins (see e.g. Ryser-type swap/connectivity results and the discussion in Wang (2020)) that the set $\mathcal{U}(\mathbf{r}, \mathbf{c})$ is connected under the class of 2×2 checkerboard swaps whenever the margins are feasible: any two matrices with the same margins can be transformed one into the other by a finite sequence of such checkerboard flips. Since $\mathcal{U}'(\mathbf{r}, \mathbf{c}) \subseteq \mathcal{U}(\mathbf{r}, \mathbf{c})$ contains all configurations with positive π-mass, and every allowed checkerboard flip proposed by Q stays within $\mathcal{U}'(\mathbf{r}, \mathbf{c})$, the W-RLA chain can reach any state in $\mathcal{U}'(\mathbf{r}, \mathbf{c})$ from any other via a finite sequence of positive-probability transitions. Thus, the chain is irreducible on $\mathcal{U}'(\mathbf{r}, \mathbf{c})$.

Aperiodicity. Aperiodicity follows because for any state A there is positive probability of remaining at A in a single step. Indeed, since some proposed swaps may be rejected by the Metropolis–Hastings rule, we have

$$P(A, A) \;=\; 1 - \sum_{B \neq A} Q(A, B)\, \alpha(A, B) \;>\; 0$$

whenever not all proposals are accepted with probability one. If every possible proposal were accepted with probability one for some pathological weight configuration, then we simply need to augment the algorithm by allowing a small probability of proposing a null move to guarantee self-loops; in practice for non-trivial weights $P(A, A) > 0$. Hence, the chain is aperiodic.

Irreducibility together with aperiodicity and detailed balance implies the chain is ergodic and has the unique stationary distribution π; moreover, from any starting configuration in $\mathcal{U}'(\mathbf{r}, \mathbf{c})$ the distribution of the chain converges to π as the number of iterations tends to infinity.

Combining Steps 1–6, we have shown that the Weighted Rectangle Loop Algorithm defines an ergodic Metropolis–Hastings Markov chain on $\mathcal{U}'(\mathbf{r}, \mathbf{c})$ with unique stationary distribution

$$\pi(H) \propto \prod_{i,j} \omega_{ij}^{h_{ij}}.$$

This completes the proof. □

10.3 Posterior summary

Suppose $S_T = \{s_1, \ldots, s_T\}$ be T post burn-in draws from the marginal posterior of clustering distribution. For each clustering configuration $s \in S_T$, one can obtain the association matrix $\eta(s)$ of order $N \times N$, whose (i, j)-th element is an indicator whether the i-th and the j-th observation are clustered together or not. Element wise averaging of these T association matrices yield the pairwise clustering probability matrix, denoted by $\bar{\eta} = \frac{1}{T} \sum_{s \in S_T} \eta(s)$. To summarize the MCMC draws, we adopt the least square model-based clustering introduced in Dahl (2006) to obtain $s_{\text{LS}} = \text{argmin}_{s \in S_T} \sum_{i=1}^{N} \sum_{j=1}^{N} (\eta_{ij}(s) - \bar{\eta}_{ij})^2$. Since $s_{\text{LS}} \in S_T$ by construction, the notion of balance is retained in the resulting s_{LS}.

11 Extensions

11.1 Extension: Imperfect knowledge on the protected attribute

Much of the prior work on fair clustering assumes complete knowledge of group membership. Esmaeili et al. (2020) generalized this by assuming imperfect knowledge of group membership through probabilistic assignments, where it is assumed that, for individual $i \in [N]$, the protected label $a_i = a$ with probability $p_i^{(a)}$ (known) such that $\sum_{a=1}^{r} p_i^{(a)} = 1,\ i \in [N]$. The goal is to balance the *expected color* in each cluster. Our fully probabilistic set up enables us seamlessly achieve this via augmenting every cycle of our computational scheme by an additional sampling step as follows $a_i = a \quad$ with probability $p_i^{(a)},\ a \in [r],\ i \in [N]$.

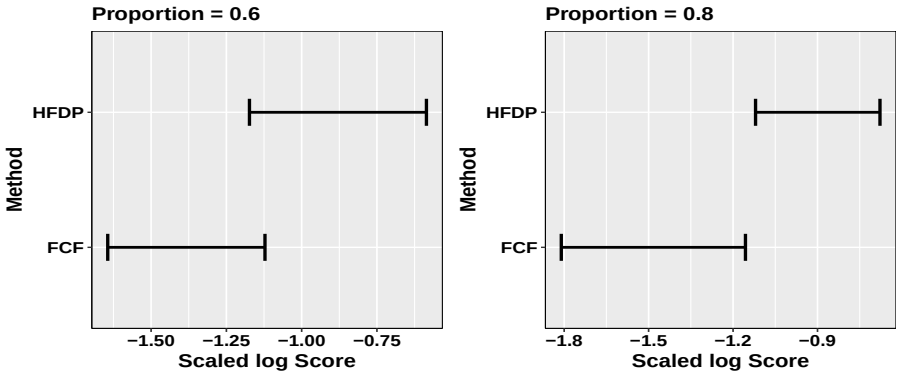

Figure 11: ***Imperfect knowledge of protected attribute.*** *The HFDP modal clustering configuration fares better compared to optimal fare clustering configuration obtained via probabilistic fair clustering (Esmaeili et al., 2020) (termed as K-means) (termed as K-Means), in terms of the scaled log score, in repeated simulations with varying* $p_{acc} \in \{0.6, 0.8\}$.

.

In this simulation study, we assume that we only have imperfect knowledge of group membership of the individuals expressed through probabilistic assignments. That is, for individual $i \in [N]$, we know that the protected label $a_i = a$ with probability $p_i^{(a)}$ (known) and $\sum_{a=1}^{r} p_i^{(a)} = 1,\ i \in [N]$. The objective is to produce a clustering configurations of the observed data that balances the expected color in each cluster. Our fully probabilistic approach enables us seamlessly achieve this via augmenting every cycle of our MCMC scheme by an additional sampling step. At each iteration of the MCMC, for the individual $i \in [N]$, we set $a_i = a$ with probability $p_i^{(a)},\ a \in [r],\ i \in [N]$, then proceed as described earlier. The goal in this simulation is to compare this slight variation of the HFDP procedure, and the optimal fair clustering configuration obtained via the method proposed in Esmaeili et al. (2020), in repeated simulations. To that end, we simulate synthetic data as follows: first generate data exactly as described in simulation study 1-(A.1), then retain the label of the protected attribute A with probability p_{acc} for each of the individuals, and swap the label with probability $1 - p_{acc}$, where we vary $p_{acc} \in \{0.6, 0.8\}$. In Figure 11, we compare HFDP with probabilistic fair clustering (Esmaeili et al., 2020) (termed as K-means) via repeated simulations with varying $p_{acc} \in \{0.6, 0.8\}$. Across all the numerical experiments, the HFDP modal clustering fares substantially better in terms of the scaled log-score, as well as automatically estimates the number of clusters.

Model-based approaches often enables us handle practical complexities (Storlie et al., 2019) e.g mixed data type, censoring, missing values, in a seamless manner. Next, we present the case with mixed data type where some variables are ordinal, and rest are continuous and potentially censored.

11.2 Extension: Presence of mixed variable types or censoring

Model-based approaches often enables us handle practical complexities Storlie et al. (2019) e.g mixed data type, censoring, missing values, in a seamless manner. For simplicity in exposition, we shall present the case with mixed data type where some variables are ordinal, and rest are continuous and potentially censored. Recall that, we have observed data $\{(\boldsymbol{x}_i, a_i)\}_{i=1}^{N}$ with $(\boldsymbol{x}_i, a_i) \in \mathcal{X} \times \mathcal{A}$, where $\boldsymbol{x}_i$ denotes the d-variate observation for the ith data unit, and a_i the corresponding level of the protected attribute. Here, we shall further assume $\mathcal{X} = \mathcal{X}_C \times \mathcal{X}_D$, where $\mathcal{X}_C$ and $\mathcal{X}_D$ are the supports of the potentially censored continuous variables and ordinal discrete variables respectively. We introduce d-variate Gaussian latent variables $\{u_i\}_{i=1}^{N}$ to operate in this set-up:

$$x_{ij} = \begin{cases} \sum_{l=1}^{L_j} d_{j,l}\ \mathbf{1}(a_{j,l-1} < u_{ij} \leqslant a_{j,l}), \text{ if } \ \text{support}(x_{ij}) \subset \mathcal{X}_D, \\ u_{ij}\ \mathbf{1}(b_j \leqslant u_{ij} \leqslant c_j) + b_j\ \mathbf{1}(u_{ij} < b_j) + c_j\ \mathbf{1}(u_{ij} > c_j), \text{ if } \ \text{support}(x_{ij}) \subset \mathcal{X}_C, \end{cases}$$

where $i \in [N],\ j \in [d],\ \mathbf{1}(\cdot)$ is the indicator function. It is straightforward to use the latent Gaussian variable approach to allow for unordered categorical variables Albert and Chib (1993), is omitted here to avoid redundancies.

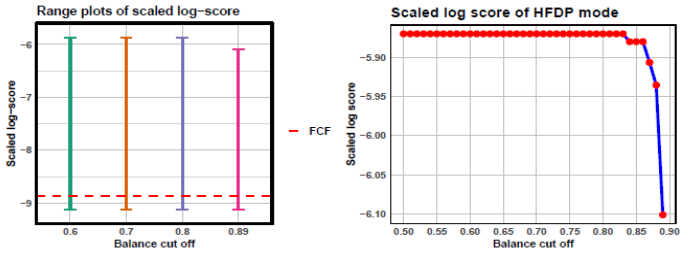

Figure 12: ***Diabetes data. (Left panel.)*** *The plot presents the range of the scaled log-score of the posterior clustering configurations arising from HFDP that has balance* $\geqslant$ *cut-off* $\in \{0.6, 0.7, 0.8, 0.89\}$. *The horizontal dashed red line represents scaled log-score of the optimal clustering configurations obtained from fair clustering via fairlets.* ***(Right panel.)*** *The plot displays the scaled log-scores for the modal posterior clustering configurations obtained from HFDP, where the balance satisfies* $\geqslant$ *cut-off* $\in \{0.50, 0.51, \ldots, 0.9\}$.

In Section S6, we compare the performance of HFDP relative to the existing methodologies on popular bench mark data sets from the UCI repository (Dua and Graff, 2017), considered before in the literature (Chierichetti et al., 2017; Böhm et al., 2020; Esmaeili et al., 2020).

Next, we present a case study on gender-neutral fair clustering in distress analysis interview corpus.

12 Benchmark datasets

We compare the performance of HFDP relative to the existing methodologies on popular bench mark data sets from the UCI repository Dua and Graff (2017), considered before in the literature Chierichetti et al. (2017); Böhm et al. (2020); Esmaeili et al. (2020).

12.1 Diabetes data

We consider the diabetes dataset available in the UCI Machine Learning Repository, that contains demographic information and medical records related to diabetes patient hospitalizations from 130 U.S. hospitals over a 10-year period. For the purposes of our analysis, we chose the numeric attributes such as age, time in hospital, to represent points in the feature space and gender as the sensitive dimension. We sub-sampled 1000 individuals from the data-set, with proportion of two genders equal to $(0.474, 0.526)$. Consequently, the target balance is 0.90. As discussed in Section 7 in the main document, we assume that the density derived from the clustering configuration without balance constraints serves as the pseudo-true data-generating mechanism.

The left panel of Figure 12 illustrates the range of scaled log-scores for posterior clustering configurations generated by HFDP, subject to balance thresholds $\geqslant$ cut-off $\in \{0.6, 0.7, 0.8, 0.89\}$. As the balance threshold increases, the scaled log-score of the modal clustering configuration produced by HFDP is a monotonic non-increasing, as expected. HFDP estimates the optimal number of clusters to be $K = 5$. For comparison, fair clustering with fairlets is applied with a fixed $K = 5$ clusters. A comparison of the perfectly balanced modal clustering configuration from HFDP with the optimal fair clustering configuration via fairlets reveals that HFDP achieves a higher scaled log-score under the pseudo-true data-generating mechanism. The right panel of Figure 12 presents the scaled log-scores for the modal posterior clustering configurations obtained from HFDP, where the balance satisfies $\geqslant$ cut-off $\in \{0.60, 0.61, \ldots, 0.90\}$. The plot exhibits a sharp decline in scaled log-scores at a balance threshold of 0.88, highlighting that a slight relaxation of the balance constraint can result in a modal posterior clustering configuration with substantially improved scaled log-scores under the pseudo-true data-generating mechanism.

12.2 Portuguese banking data

Next, we analyze the Portuguese banking dataset, from the UCI Machine Learning Repository. This dataset comprises records from a telemarketing campaign conducted by a Portuguese banking institution. Each record encapsulates detailed information about individual clients contacted during the campaign, including demographic

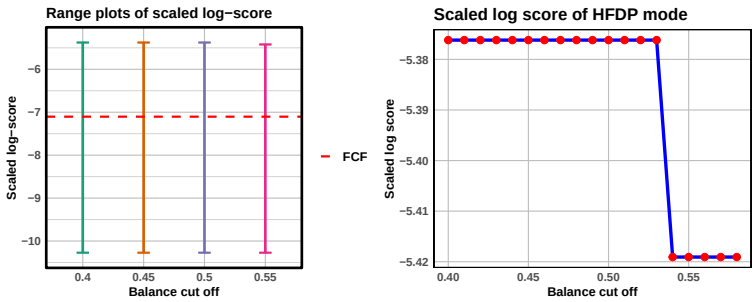

Figure 13: ***Portuguese banking data. (Left panel.)*** *The plot presents the range of the scaled log-score of the posterior clustering configurations arising from HFDP that has balance* $\geqslant$ *cut-off* $\in \{0.4, 0.45, 0.5, 0.55\}$. *The horizontal dashed red line represents scaled log-score of the optimal clustering configurations obtained from fair clustering via fairlets.* ***(Right panel.)*** *The plot displays the scaled log-scores for the modal posterior clustering configurations obtained from HFDP, where the balance satisfies* $\geqslant$ *cut-off* $\in \{0.4, 0.41, \ldots, 0.6\}$.

and behavioral attributes. We selected numeric attributes, including age, bank balance, and account duration, to represent points in the feature space, while marital status was designated as the protected attribute. The dataset was sub-sampled to include 1000 records, with the proportion of married and unmarried clients set to $(0.626, 0.374)$, corresponding to a target balance of 0.60. As discussed in Section 7 in the main document, we assume that the density derived from the clustering configuration without balance constraints serves as the pseudo-true data-generating mechanism.

The left panel of Figure 13 illustrates the range of scaled log-scores for posterior clustering configurations generated by HFDP, subject to balance thresholds $\geqslant$ cut-off $\in \{0.4, 0.45, 0.5, 0.55\}$. As the balance threshold increases, the scaled log-score of the modal clustering configuration produced by HFDP is a monotonic non-increasing, as expected. HFDP estimates the optimal number of clusters to be $K = 6$. For comparison, fair clustering with fairlets is applied with a fixed $K = 6$ clusters. A comparison of the perfectly balanced modal clustering configuration from HFDP with the optimal fair clustering configuration via fairlets reveals that HFDP achieves a higher scaled log-score under the pseudo-true data-generating mechanism. The right panel of Figure 13 presents the scaled log-scores for the modal posterior clustering configurations obtained from HFDP, where the balance satisfies $\geqslant$ cut-off $\in \{0.40, 0.41, \ldots, 0.60\}$. The plot exhibits a sharp decline in scaled log-scores at a balance threshold of 0.53, highlighting that a slight relaxation of the balance constraint can result in a modal posterior clustering configuration with substantially improved scaled log-scores under the pseudo-true data-generating mechanism.

12.3 Credit card data

The UCI Default of Credit Card Clients Dataset contains information on 30,000 credit card holders, including demographic attributes such as age and marital status, and financial data such as credit limit and payment history. We chose numeric attributes such as age, credit limit, to represent points in the feature space and marital status (married, unmarried, others) as the sensitive dimension. We sub-sampled 600 individuals from the data-set, such that the target balance is 0.99. As discussed in Section 7 in the main document, we assume that the density derived from the clustering configuration without balance constraints serves as the pseudo-true data-generating mechanism.

The left panel of Figure 14 illustrates the range of scaled log-scores for posterior clustering configurations generated by HFDP, subject to balance thresholds $\geqslant$ cut-off $\in \{0.7, 0.8, 0.9, 0.99\}$. As the balance threshold increases, the scaled log-score of the modal clustering configuration produced by HFDP is a monotonic non-increasing, as expected. HFDP estimates the optimal number of clusters to be $K = 2$. For comparison, fair clustering with fairlets is applied with a fixed $K = 2$ clusters. A comparison of the perfectly balanced modal clustering configuration from HFDP with the optimal fair clustering configuration via fairlets reveals that HFDP achieves a higher scaled log-score under the pseudo-true data-generating mechanism. The right panel of Figure 14 presents the scaled log-scores for the modal posterior clustering configurations obtained from HFDP, where the balance satisfies $\geqslant$ cut-off $\in \{0.7, 0.71, \ldots, 0.99\}$. The plot exhibits a sharp decline in scaled log-scores at a balance threshold of 0.97, highlighting that a slight relaxation of the balance constraint can result in a modal posterior clustering

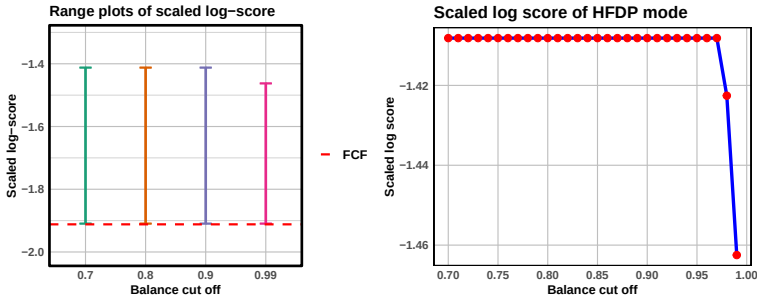

Figure 14: ***Credit card data. (Left panel.)*** *The plot presents the range of the scaled log-score of the posterior clustering configurations arising from HFDP that has balance* $\geqslant$ *cut-off* $\in \{0.7, 0.8, 0.9, 0.99\}$. *The horizontal dashed red line represents scaled log-score of the optimal clustering configurations obtained from fair clustering via fairlets.* ***(Right panel.)*** *The plot displays the scaled log-scores for the modal posterior clustering configurations obtained from HFDP, where the balance satisfies* $\geqslant$ *cut-off* $\in \{0.7, 0.71, \ldots, 0.99\}$.

configuration with substantially improved scaled log-scores under the pseudo-true data-generating mechanism.