entropy

Article

A Gibbs Posterior Framework for Fair Clustering

Abhisek Chakraborty *, Anirban Bhattacharya and Debdeep Pati

Department of Statistics, Texas A&M University, College Station, TX 77843, USA; anirbanb@stat.tamu.edu (A.B.); debdeep@stat.tamu.edu (D.P.)

* Correspondence: abhisek_chakraborty@tamu.edu

Abstract: The rise of machine learning-driven decision-making has sparked a growing emphasis on algorithmic fairness. Within the realm of clustering, the notion of *balance* is utilized as a criterion for attaining *fairness*, which characterizes a clustering mechanism as fair when the resulting clusters maintain a consistent proportion of observations representing individuals from distinct groups delineated by *protected attributes*. Building on this idea, the literature has rapidly incorporated a myriad of extensions, devising fair versions of the existing frequentist clustering algorithms, e.g., k-means, k-medioids, etc., that aim at minimizing specific loss functions. These approaches lack uncertainty quantification associated with the optimal clustering configuration and only provide clustering boundaries without quantifying the probabilities associated with each observation belonging to the different clusters. In this article, we intend to offer a novel probabilistic formulation of the fair clustering problem that facilitates valid uncertainty quantification even under mild model misspecifications, without incurring substantial computational overhead. Mixture model-based fair clustering frameworks facilitate automatic uncertainty quantification, but tend to showcase brittleness under model misspecification and involve significant computational challenges. To circumnavigate such issues, we propose a generalized Bayesian fair clustering framework that inherently enjoys decision-theoretic interpretation. Moreover, we devise efficient computational algorithms that crucially leverage techniques from the existing literature on optimal transport and clustering based on loss functions. The gain from the proposed technology is showcased via numerical experiments and real data examples.

Keywords: algorithmic fairness; balance; generalized Bayes; minimum cost flow; optimal transport

Citation: Chakraborty, A.; Bhattacharya, A.; Pati, D. A Gibbs Posterior Framework for Fair Clustering. *Entropy* **2024**, *26*, 63. https://doi.org/10.3390/e26010063

Academic Editor: Dawn E. Holmes

Received: 1 December 2023
Revised: 28 December 2023
Accepted: 9 January 2024
Published: 11 January 2024

1. Introduction

Fairness in algorithmic decision-making aims to mitigate discrimination involving the unfavorable treatment of individuals based on their membership to specific demographic sub-groups identified by *protected attributes*. These protected attributes may encompass factors such as gender, race, marital status, etc., depending on the specific context, and are often delineated by local, national, or international legal frameworks. For instance, in the context of bank loan approvals, the protected attribute *marital status* might encompass labels such as married, divorced, and unmarried applicants. Perhaps unsurprisingly, early research in fairness on machine learning exclusively focused on supervised learning problems. However, there was a genuine need to understand fairness in unsupervised learning settings, especially in clustering problems.

In a seminal work, Chierichetti et al. [1] introduced the concept of *balance* as a criterion for achieving fairness in clustering, defining clustering mechanisms as fair when resulting clusters maintain a common ratio of observations representing individuals from different groups identified by protected attributes. This notion was explored in both the k-center and the k-median problems, particularly in the *two-color* case. Subsequent articles extended this framework to the more complex *multi-color* cases, addressing the situation where the protected attributes have multiple labels [2]. Esmaeili et al. [3] considered the case with imperfect knowledge of group membership through probabilistic assignments. Bera et al. [4] expanded the scope by enabling users to specify parameters controlling extent of balance in the clusters, considering the general $\mathbf{L}_p$ objective for the clustering and scenarios where individuals could belong to multiple protected groups. Fairness mechanisms have been explored in different clustering frameworks, such as spectral clustering [5], correlation clustering [6], and hierarchical clustering [7]. Additionally, researchers have investigated the notions of individual fairness [8–10] and proportional fairness [11] in the context of clustering. Fairness in clustering has also been studied in combination with other critical aspects of modern machine learning, including privacy [12] and robustness [13]. For a comprehensive review, interested readers are directed to the website on https://www.fairclustering.com/ (accessed on 30 November 2023).

While the follow-up works [2,3,12,14] greatly increased the scope of fair clustering from various aspects, uncertainty quantification associated with the optimal clustering configuration was largely illusive until recently. Chakraborty et al. [15], complementing the existing literature on fair clustering which is almost exclusively based on optimizing appropriate objective functions, took a fully Bayesian approach to tackle the problem of clustering under balance constraints to provide valid uncertainty quantification, developed a concrete notion of optimal recovery in this problem, and devised a scheme for principled performance evaluation of algorithms. In this article, we propose an alternative *generalized Bayesian* fair clustering framework, embedding common clustering loss functions at the heart of the likelihood formulation. Generalized Bayesian methodologies, as exemplified in various works [16,17], are gaining prominence due to their ability to alleviate the necessity of explicitly specifying the complete data generative mechanism. This characteristic allows us to circumnavigate the challenges related to the lack of robustness of fully Bayesian clustering approaches. Moreover, a principled selection of the *temperature parameter* in the generalized likelihood guarantees its adherence to valid decision-theoretic justifications. Finally, ardent care is exercised to devise efficient computational algorithms to carry out posterior inference, crucially leveraging techniques from the existing literature on clustering based on loss functions.

Prior to presenting the proposed methodology, we provide a concise overview of pertinent concepts in generalized Bayesian inference and fair clustering, laying the groundwork for subsequent discussions.

1.1. Generalized Bayesian Inference through Gibbs Posterior

In Bayesian inference, the need for likelihood specification introduces a notable obstacle in many practical applications, primarily stemming from apprehensions about the potential misspecification of the statistical model [18–20]. The non-parametric Bayesian approaches [21,22] enhance the adaptability of such methodologies, bolstering the robustness of the statistical inferences. Nevertheless, evaluating the trade-off between the added complexity in adopting a fully non-parametric approach and the acquired robustness necessitates meticulous consideration tailored to specific applications. Generalized Bayesian inference [16,23–25] offers an alternative model-free approach to circumvent the risk of model misspecification bias as well as excessive complexity in problem formulation. We proceed by recording the definition of generalized Bayesian posteriors, followed by a brief overview of notable contributions within this domain.

To that end, let $\mathbf{u} = (u_1, \ldots, u_N)^{\mathrm{T}}$ be the observed data, $\theta \in \Theta$ be the parameter of interest, and $\pi(\theta)$ be the prior on θ. Then, the generalized Bayesian posterior is defined as

$$\pi(\theta \mid \lambda, \mathbf{u}) \;\propto\; \pi(\theta) \exp\{-\lambda \mathcal{L}(\theta \mid \mathbf{u})\}, \tag{1}$$

where $\lambda > 0$ is a temperature parameter, and $\mathcal{L}(\theta \mid \mathbf{u}) > 0$ is a loss function of choice. The posterior in (1) is often referred to as the *Gibbs posterior*. Standard Bayesian inference is recovered from (1), when the loss function $\lambda \mathcal{L}(\theta \mid \mathbf{u})$ is a negative log-likelihood. While Gibbs posteriors have found applications in diverse contexts [16,25–29], it is only in recent times that their role in offering a rational update of beliefs, thus qualifying as genuine posterior distributions, has been established [30]. Further, Syring and Martin [31] provided sufficient conditions for establishing concentration rates for Gibbs posteriors under a sub-exponential type loss function. Martin and Syring [17], Holmes and Walker [32] presented significant developments in methodologies for selecting the temperature parameter λ. Owing to such increasing support for generalized Bayesian inference, we propose a Gibbs posterior-based framework for uncertainty quantification in fair clustering. We additionally introduce computationally efficient algorithms for point estimation and associated uncertainty quantification.

1.2. Fairness in Clustering

We shall now formally introduce the notion of *balance* in the context of fair clustering. For a positive integer t, denote $[t] := \{1, \dots, t\}$. Suppose we observe data $\{(\boldsymbol{x}_i, a_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ denotes the d-variate observation for the i-th data unit, and a_i the label of the protected attribute. For each a, let $\{\boldsymbol{x}_i^{(a)}\}_{i=1}^{N_a}$ denote the observations corresponding to the a-th level of the protected attribute, where $N_a = \sum_{i=1}^{N} \mathbf{1}(a_i = a)$ and $\sum_{a=1}^{r} N_a = N$. The goal of fair clustering is to assign the data points $\{(\boldsymbol{x}_i, a_i)\}_{i=1}^{N}$ into clusters $\boldsymbol{C} = (C_1, \dots, C_K)$, $\bigcup_{k=1}^{K} C_k = [N]$, respecting the notion of balance [1].

Definition 1 ([1])**.** *Given $\{(\boldsymbol{x}_i, a_i) \in \mathcal{X} \times [2],\ i \in [N]\}$ such that $a_i = a$ for $i = \sum_{j=1}^{a-1} N_j + 1, \dots, \sum_{j=1}^{a} N_j$ where $a \in [2]$ and $N_0 = 0$, the balance in C_k is defined as*

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

where $|C_{kj}|$ denotes the number of observations in C_k with $a = j$. The overall balance of the clustering is $Balance(\boldsymbol{C}) = \min_{k=1,\dots,K} Balance(C_k)$. The higher this measure is for a clustering configuration, the fairer the clustering is.

Given the aforementioned definition of *balance*, Chierichetti et al. [1] introduced the notion of *fairlets* as minimal fair sets that approximately maintain the selected clustering objective. The authors illustrated that addressing any fair clustering problem involves initially obtaining a fairlet decomposition of the data through the solution of a *minimum cost flow* problem. Subsequently, classical clustering algorithms, such as k-means or k-center, can be employed for further processing. As we eluded to earlier, the follow-up literature in fair clustering [2,3,12,14] involves devising fair versions of the existing frequentist clustering algorithms, e.g., k-means, k-medioids, etc., that aim at minimizing specific loss functions. These approaches lack the quantification of uncertainty linked to the optimal fair clustering configuration, i.e., these approaches only provide clustering boundaries but do not quantify the probabilities of each of the observations belonging to the different clusters. While model-based clustering approaches are specifically designed to answer such questions, they routinely fall prey to even minor model misspecification [19,20]. The generalized Bayesian approach provides a convenient middle ground that not only is immune to such minor model misspecifications, but also provides valid uncertainty quantification associated with the clustering. This article introduces a generalized Bayesian fair clustering framework with inherent uncertainty quantification and presents efficient computational algorithms for posterior inference, leveraging techniques from the clustering literature based on loss functions.

The rest of the article is arranged as follows. In Section 2, we introduce the proposed generalized Bayesian fair clustering procedure. Section 3 presents an efficient MCEM and a MCMC scheme to carry out posterior analysis under the proposed framework. Sections 4 and 5 provide detailed numerical experiments and real data examples to delineate the key gain from the proposed methodology over competing methods. Finally, we conclude with a discussion.

2. Methodology

2.1. Preliminaries

Chierichetti et al. [1] introduced the notion of *fairlets*, minimal sets that adhere to fair representation while approximately preserving a clustering objective. Given the observed data $\{(\boldsymbol{x}_i, a_i) \in \mathcal{X} \times [2],\ i \in [N]\}$, fair clustering via fairlets [1] involves first decomposing data into a set of m fairlets, and calculate the m fairlet centers. Let $\mathcal{U}(\subset \mathcal{X}^m)$ denote the class of all such "m fairlet centers". Let $\mathcal{L}_f : \mathcal{U} \to \mathbf{R}^+$ denote loss function utilized for the fairlet decomposition. The optimal fairlet decomposition $\mathbf{u}^\star \in \mathcal{U}$ is then expressed as

$$\mathbf{u}^\star = \operatorname*{arg\,min}_{\mathbf{u} \in \mathcal{U}} \mathcal{L}_f(\mathbf{u}). \tag{2}$$

This optimization problem in (2) is often recognized as the *minimum cost flow* problem [33].

Prior to progressing further, it is imperative to conduct a comprehensive examination of the constituents comprising the loss function $\mathcal{L}_f$. To that end, suppose the two labels of the protected attribute is represented in a $1 : t$ ratio in the observed data, i.e., $t \times N_1 = N_2$. We wish to find a perfectly balanced clustering of the observed data, i.e., the two labels of the protected attribute should be represented in a $1 : t$ ratio in each of the clusters. The construction of an optimal $(1, t)$-fairlet decomposition of the observed data involves solution of a constrained binary optimal transport problem. First, we define the $N_1 \times N_2$ cost matrix

$$\mathrm{L} = ((l_{ik})) = \big((\mathcal{D}(x_i, x_{N_1+k}))\big), \quad i \in [N_1],\ j \in [N_2],$$

where $\mathcal{D}(w, v) \geq 0$ quantifies the discrepancy between w and $v \in \mathcal{X}$. We introduce a column sum vector $\boldsymbol{c} = t \times \mathbf{1}_{N_1}$ and a row sum vector $\boldsymbol{r} = \mathbf{1}_{N_2}$, where $\mathbf{1}_s$ is a vector of s 1 s. Given the two fixed vectors $\boldsymbol{r}, \boldsymbol{c}$, we define a polytope of $N_1 \times N_2$ binary matrices

$$\mathrm{U}(\boldsymbol{r}, \boldsymbol{c}) := \{\mathrm{B} \mid \mathrm{B}\mathbf{1}_{N'} = \boldsymbol{r};\ \mathrm{B}^{\mathrm{T}}\mathbf{1}_{N'} = \boldsymbol{c}\},$$

with fixed margin $\boldsymbol{r}, \boldsymbol{c}$ and solve the constrained binary optimal transport problem [34]

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

where $\langle \mathrm{B}, \mathrm{L} \rangle = \mathrm{tr}(\mathrm{B}^{\mathrm{T}}\mathrm{L})$. The matrix $\mathrm{B}' = ((b'_{ik}))$ describes an optimal $(1, t)$-fairlet decomposition. That is, if $b'_{ik_1} = \ldots = b'_{ik_t} = 1$ for some $i \in [N_1]$ and $1 \leq k_1 < \ldots < k_t \leq N_2$, then $(x_i, x_{N_1+k_1}, \ldots, x_{N_1+k_t})^T$, $i \in [N]_1$ defines the fairlets. The fairlet centers are represented as $\mathbf{u}^\star = (u_1^\star, \ldots, u_m^\star)^{\mathrm{T}}$, acquired through averaging observations within the respective fairlets. Finally, we define a map

$$\xi : \mathrm{U}(\boldsymbol{r}, \boldsymbol{c}) \to \mathcal{U},$$

that takes a binary matrix B with fixed margins $(\boldsymbol{r}, \boldsymbol{c})$, representing a fairlet decomposition of $\{(\boldsymbol{x}_i, a_i) \in \mathcal{X} \times [2],\ i \in [N]\}$ to m-fairlet centers $\mathbf{u} \in \mathcal{U}$. Then, loss function $\mathcal{L}_f$ is represented as

$$\mathcal{L}_f(\xi(\mathrm{B})) = \langle \mathrm{B}, \mathrm{L} \rangle, \qquad \mathrm{B} \in \mathrm{U}(\boldsymbol{r}, \boldsymbol{c}).$$

Efficient off-the-shelf algorithms for optimizing the loss function are routinely available.

Next, given the optimal fairlet decomposition of the observed data $\mathbf{u}^\star = (u_1^\star, \ldots, u_m^\star)^{\mathrm{T}} \in \mathcal{U}$, Chierichetti et al. [1] proposed to invoke existing machinery for traditional clustering algorithms, e.g., k-means, k-center, etc., to cluster the m-fairlet centers into K groups. We focus on the flexible class of clustering mechanisms characterized by a factorized loss $\mathcal{L}_c(\mathbf{C} \mid \mathbf{u}^\star)$. Throughout the article, we assume that the number of clusters K is fixed, e.g., it is known or has been selected in an exploration phase. Let $\mathbf{u}^\star_{(k)}$ denote the center of the fairlet centers $\{u_i^\star : i \in C_k\}$ belonging to cluster C_k, for $k \in [K]$. Then, the clustering is characterized by factorized loss takes the form

$$\mathcal{L}_c(\mathbf{C} \mid \mathbf{u}^\star) = \sum_{k=1}^{K} \sum_{i \in C_k} \mathcal{D}(u_i^\star, \mathbf{u}_{(k)}^\star), \qquad \mathbf{C} : |\mathbf{C}| = K, \tag{3}$$

where $\mathcal{D}(u_i^\star, \mathbf{u}_{(k)}^\star) \geq 0$ is a function of $u_i^\star$ and $\mathbf{u}_{(k)}^\star$ which quantifies the discrepancy of the i-th unit from the k-th cluster. The formulation in (3) encapsulates a large class of common clustering costs. For example, suppose we assume that $\mathbf{u}_{(k)}^\star$ represents the arithmetic means of the vectors $u_i^\star$, $i \in C_K$ for any $k = 1, \ldots, K$. Then, the k-means loss function takes the form

$$\mathcal{L}_c(\mathbf{C} \mid \mathbf{u}^\star) = \sum_{k=1}^{K} \sum_{i \in C_k} ||u_i^\star - \mathbf{u}_{(k)}^\star||_2^2.$$

Importantly, efficient off-the-shelf algorithms [35] for solution of the clustering problem in (3) are routinely available. In summary, Chierichetti et al. [1] critically exploits the existing tools in minimum cost flow problem (2) and factorized loss-based clustering (3) to obtain the optimal fair clustering configuration. In a subsequent work, we shall integrate this methodology within the generalized Bayesian inference framework to quantify uncertainty associated with the optimal fair clustering configuration.

2.2. Generalized Bayesian Fair Clustering

Given the observed data $\{(\boldsymbol{x}_i, a_i) \in \mathcal{X} \times [2],\ i \in [N]\}$, we recall that $\mathcal{U}(\subset \mathcal{X}^m)$ denote the class of all "m fairlet centers". We placed a uniform prior on the space of all possible fairlet decompositions $\mathcal{U}$, i.e., we assume

$$\pi(\mathbf{u}) = \frac{1}{|\mathcal{U}|}, \quad \mathbf{u} \in \mathcal{U}. \tag{4}$$

Our framework is easily modified to consider more elaborate priors, but we focus on the uniform case throughout the paper. Then, the *generalized Bayes posterior* for *fairlet decomposition* takes the form

$$\pi(\mathbf{u} \mid \lambda_f, \{(\boldsymbol{x}_i, a_i)\}_{i=1}^N) \ \propto \ \frac{\exp\Big\{-\lambda_f \mathcal{L}_f(\mathbf{u})\Big\}}{\sum_{\mathbf{u} \in \mathcal{U}} \exp\Big\{-\lambda_f \mathcal{L}_f(\mathbf{u})\Big\}}, \tag{5}$$

where λ_f is a temperature parameter.

Given a fairlet decomposition $\mathbf{u} \in \mathcal{U}$, which may be different from the optimal fairlet decomposition $\mathbf{u}^\star$, a typical Bayesian model for clustering [36–38] is based on the assumption that observations follows from

$$(u_i \mid \theta_k, i \in C_k) \overset{\text{ind}}{\sim} \pi(u_i \mid \theta_k), \qquad k \in [K], \tag{6}$$

where $\theta_k \overset{\text{iid}}{\sim} \pi(\theta)$ for $k \in [K]$. Under the above model and prior specification, the posterior distributions of clustering configurations take the form

$$\pi(\mathbf{C} \mid \mathbf{u}) \ \propto \ \pi(\mathbf{C}) \prod_{k=1}^{K} \left[\int_\Theta \prod_{i \in C_k} \pi(u_i \mid \theta) \pi(\theta) d\theta \right], \tag{7}$$

where $\pi(\mathbf{C})$ is the prior probability of $\mathbf{C}$, $\pi(u \mid \theta)$ is the within-cluster likelihood, and $\pi(\theta)$ is the prior distribution on the cluster-specific parameters. While Equation (7) serves as the foundation for an extensive body of literature on Bayesian clustering, it gives rise to significant practical challenges. The integral often does not admit a closed form expression, introducing computational complexities. Furthermore, the posterior of clustering configurations is highly sensitive to the precise specifications of data generating mechanism $\pi(u \mid \theta)$. Such model-based clustering frameworks (6) are routinely criticized for various aspects. Firstly, clustering may just serve as a convenient preprocessing step, and there might not be distinct groups present in the data. Moreover, even if such groups exist, the distribution of the data within each cluster is unlikely to precisely adhere to the chosen distribution

$\pi(u \mid \theta)$. In terms of computational aspects, even when the data accurately conform to the selected mixture distribution, we frequently encounter substantial computational bottlenecks, especially in high dimensions [39–41].

To address these brittleness issues, Rigon et al. [29] implemented a Gibbs posterior framework tailored for clustering, aiming to navigate around these challenges. Let the loss function $\mathcal{L}(\mathbf{C} \mid \mathbf{u})$ be as in (3). In this article, our attention is directed towards employing uniform clustering priors of the form

$$\pi(\mathbf{C}) = \frac{1}{\mathcal{S}(n,K)}, \qquad \mathbf{C} : |\mathbf{C}| = K, \tag{8}$$

where $\mathcal{S}(n,K) = 1/K! \sum_{k=0}^{K} (-1)^{K-k} K! \{(K-k)!k!\}^{-1} k^n$ is the Stirling number of the second kind. Prior (8) is uniform over partitions having K components. Although the framework readily accommodates more intricate clustering priors, our emphasis in this paper remains on the uniform case. The generalized Bayes posterior under a *generalized Bayes product partition model* has the form

$$\pi(\mathbf{C} \mid \lambda_c, \mathbf{u}) \propto \prod_{k=1}^{K} \exp\left\{ -\lambda_c \sum_{i \in C_k} \mathcal{D}(u_i, \mathbf{u}_{(k)}) \right\}, \qquad \mathbf{C} : |\mathbf{C}| = K,$$

where λ_c is a temperature parameter and $\{\mathbf{u}_{(k)},\ k \in [K]\}$ are the K cluster centers. We have now assembled all the necessary components to introduce a generalized Bayesian posterior framework for the fair clustering problem.

In this article, we still focus on the uniform clustering prior in (8), and utilize the uniform prior on the space of all possible fairlet decompositions $\mathcal{U}$, introduced in (4). Then, the *generalized Bayes posterior* for *fair clustering* takes the form

$$\pi(\mathbf{C}, \mathbf{u} \mid (\lambda_f, \lambda_c), \{(\boldsymbol{x}_i, a_i)\}_{i=1}^{N}) \propto \frac{\exp\left\{-\lambda_f \mathcal{L}_f(\mathbf{u})\right\}}{\sum_{\mathbf{u} \in \mathcal{U}} \exp\left\{-\lambda_f \mathcal{L}_f(\mathbf{u})\right\}} \times \exp\{-\lambda_c \mathcal{L}_c(\mathbf{C} \mid \mathbf{u})\}.$$

Under the assumption of factorised clustering loss, the posterior simplifies to

$$\pi(\mathbf{C}, \mathbf{u} \mid (\lambda_f, \lambda_c), \{(\boldsymbol{x}_i, a_i)\}_{i=1}^{N}) \propto \frac{\exp\left\{-\lambda_f \mathcal{L}_f(\mathbf{u})\right\}}{\sum_{\mathbf{u} \in \mathcal{U}} \exp\left\{-\lambda_f \mathcal{L}_f(\mathbf{u})\right\}} \times \prod_{k=1}^{K} \exp\left\{ -\lambda_c \sum_{i \in C_k} \mathcal{D}(u_i, \mathbf{u}_{(k)}) \right\}, \tag{9}$$

such that $\mathbf{C} : |\mathbf{C}| = K$. We employ the methodology proposed in [32] for the selection of the temperature parameters (λ_f, λ_c).

A natural competitor for the proposed methodology based on the joint posterior in (9) is the fair clustering with fairlets [1]. Subsequent to [1], many follow up articles proposed suggestions for improved computational scalability of the fair clustering task [42,43]. However, from a methodological perspective, [1] still serves as the primary go-to method.

Proposition 1. *Let $\pi(\mathbf{C}, \mathbf{u} \mid (\lambda_f, \lambda_c), \{(\boldsymbol{x}_i, a_i)\}_{i=1}^{N})$ denote the joint posterior of the fairlet decompositions of the observed data and clustering configurations in (9). Then, we denote the posterior mode by*

$$(\mathbf{C}^{\mathrm{MAP}}, \mathbf{u}^{\mathrm{MAP}}) = \arg\max_{\mathbf{C}, \mathbf{u}} \left[\pi(\mathbf{C}, \mathbf{u} \mid (\lambda_f, \lambda_c), \{(\boldsymbol{x}_i, a_i)\}_{i=1}^{N}) \right].$$

The clustering configuration $\mathbf{C}^{\mathrm{MAP}}$ does not coincide with the optimal clustering obtained via fair clustering via fairlets [1].

Proof. From Equation (9), we note that the quantities $(\mathbf{C}^{\mathrm{MAP}}, \mathbf{u}^{\mathrm{MAP}})$ are computed via solving

$$\arg\max_{\mathbf{C},\mathbf{u}}[\pi(\mathbf{C},\mathbf{u} \mid (\lambda_f,\lambda_c),\{(\boldsymbol{x}_i,a_i)\}_{i=1}^N)] = \arg\min_{\mathbf{C},\mathbf{u}}[\lambda_f\mathcal{L}_f(\mathbf{u}) + \lambda_c\sum_{k=1}^K\sum_{i\in C_k}\mathcal{D}(u_i,\mathbf{u}_{(k)})],$$

where $\lambda_f \geq 0$ and $\lambda_c \geq 0$. On the other hand, the optimal clustering obtained via fair clustering via fairlets is calculated in a two-step process: first, the optimal fairlet decomposition $\mathbf{u}^\star \in \mathcal{U}$ is obtained by

$$\mathbf{u}^\star = \arg\min_{\mathbf{u}\in\mathcal{U}}\mathcal{L}_f(\mathbf{u}),$$

and then, the optimal clustering is obtained by minimizing the factorized loss given $\mathbf{u}^\star$,

$$\mathbf{C}^{\text{FCF}} = \arg\min_{\mathcal{C}}\mathcal{L}_c(\mathbf{C} \mid \mathbf{u}^\star) = \arg\min_{\mathcal{C}}\sum_{k=1}^K\sum_{i\in C_k}\mathcal{D}(u_i^\star,\mathbf{u}_{(k)}^\star), \qquad \mathbf{C} : |\mathbf{C}| = K.$$

This completes the proof. □

In Proposition 1, we prove that the maximum a posteriori fair clustering configuration obtained via maximizing the joint posterior in (9) is different from the optimal clustering configuration obtained using fair clustering via fairlets. For the purpose of point estimation, we argue that $\mathbf{C}^{\text{MAP}}$ should be preferred over $\mathbf{C}^{\text{FCF}}$, since it is obtained via maximizing the posterior with respect to the fairlet decomposition and clustering configuration simultaneously. The $\mathbf{C}^{\text{FCF}}$ estimator, on the other hand, should be considered as an approximation of our $\mathbf{C}^{\text{MAP}}$ estimator since $\mathbf{C}^{\text{FCF}}$ is practically obtained via first solving

$$\mathbf{u}^\star = \arg\max_{\mathbf{u}\in\mathcal{U}}\left[\exp\left\{-\lambda_f\mathcal{L}_f(\mathbf{u})\right\}\right],$$

then, given the optimal fairlet decomposition $\mathbf{u}^\star$ of the observed data, solving

$$\mathbf{C}^{\text{FCF}} = \arg\max_{\mathbf{C}}\left[\prod_{k=1}^K\exp\left\{-\lambda_c\sum_{i\in C_k}\mathcal{D}(u_i^\star,\mathbf{u}_{(k)}^\star)\right\}\right].$$

In numerical studies, we observed that the maximum a posteriori estimate of the proposed fair k-means clustering via Gibbs posterior turns out to be very similar to the fair clustering with fairlets estimator (refer to Figures 1 and 2). The proposed Gibbs posterior-based approach, unlike a traditional optimization-based approach [3], also provides uncertainty quantification associated with the clustering configurations along with the maximum a posteriori estimate.

We conclude this section by discussing some other salient features of the posterior in (9). Firstly, sampling in the space of all possible fairlet decompositions, rather than only focusing on the optimal fairlet decomposition, enables us to take into account the uncertainties arising from this step. This, in turn, crucially enables us to conduct joint inference on $(\mathbf{C},\mathbf{u})$. Secondly, the suggested formulation for clustering the fairlet centers avoids specifying an underlying data generation mechanism and adeptly circumvents the computation of integrals in high-dimensional spaces. Finally, the assumed factorized loss for the clustering significantly simplifies the posterior computation (refer to Section 3 for details).

As we eluded to earlier, ardent care is required to ensure efficient sampling from the joint posterior of fairlet decompositions and clustering configurations in (9). To that end, we first develop an intermediate Monte Carlo expectation maximization (MCEM) scheme, followed by a subsequent full Markov Chain Monte Carlo (MCMC) scheme to sample from the posterior.

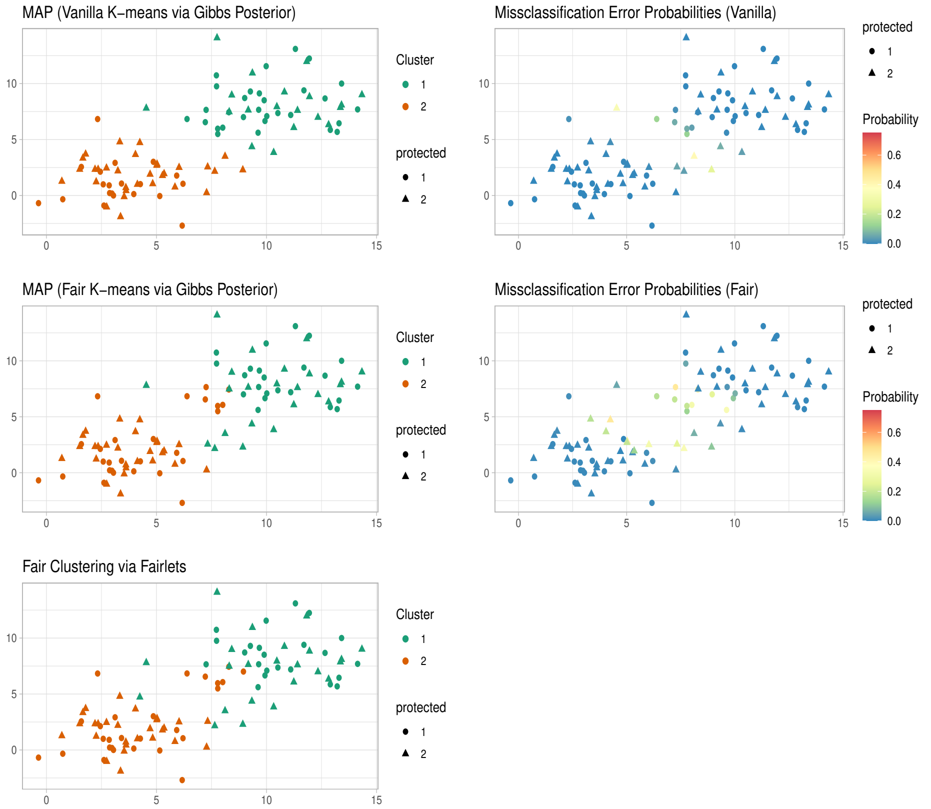

Figure 1. Well-specified case. K-means and fair k-means through Gibbs posterior with $K = 2$. We plot the maximum a posteriori clustering configurations and misclassification error probabilities obtained from the posterior samples, via the scheme in Section 3.2.

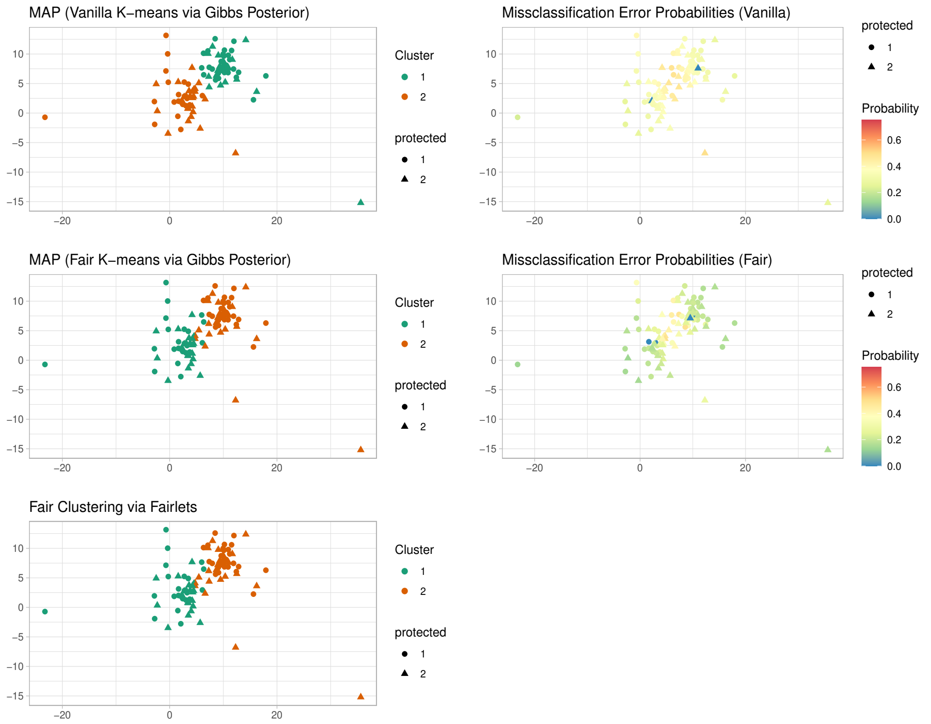

Figure 2. Misspecified case (L_2 loss). K-means and fair k-means through Gibbs posterior with $K = 2$. We plot the maximum a posteriori clustering configurations and misclassification error probabilities obtained from the posterior samples, via the scheme in Section 3.2.

3. Posterior Analysis

3.1. Sampling Scheme

We develop a Gibbs sampling scheme to sample from the joint posterior of fairlet decompositions and clustering configurations in (9). We achieve this via iterating over a step-by-step scheme, as per the common practice. We first draw a potentially non-optimal fairlet decomposition of the data and then sample the clustering indices given the specific fairlet decomposition of the data. We cycle through the steps until convergence.

We now put our computational strategy in concrete terms. Suppose the two labels of the protected attribute are represented in a 1:t ratio in the observed data, where t is an integer and assume $t \times N_1 = N_2$. Our goal is to explore the space of the perfectly balanced clustering configurations, compute the maximum a posteriori perfectly balanced clustering configuration, and quantify the uncertainty attached to it. This task can be accomplished via iterating over two steps: **Step 1.** sampling in the space of the all possible $(1, t)$-fairlet decomposition of the observed data; and **Step 2.** given a potentially non-optimal $(1, t)$-fairlet decomposition of the observed data, sample in the space of all possible clustering configurations of the fairlet centers. We cycle through **Step 1** and **Step 2** until convergence.

Step 1 (Sampling the Fairlets). Sampling in the space of the all possible $(1, t)$-fairlet decomposition of the observed data is further carried out in two steps. In **(i)**, we simply obtain the optimal $(1, t)$-fairlet decomposition of the observed data. However, to accomplish our goal of quantifying the uncertainty associated with the fair clustering, we first need to take into account the uncertainty associated with the $(1, t)$-fairlet decomposition of the observed data. To that end, in step **(ii)**, we utilize an innovative Metropolis step to explore other potentially non-optimal $(1, t)$-fairlet decompositions *near* the optimal $(1, t)$-fairlet decomposition of the observed data. The two steps follow.

(i) We demonstrate how we can utilize discrete optimal transport to obtain the optimal $(1, t)$-fairlet decomposition of the observed data. We undertake the following steps. First, we define the $N_1 \times N_2$ cost matrix

$$\mathrm{L} = ((l_{ik})) = \big((\mathcal{D}(x_i, x_{N_1+k}))\big), \quad i \in [N_1],\ j \in [N_2],$$

column sum vector $\boldsymbol{c} = \mathbf{1}_{N_1}$, and row sum vectors $\boldsymbol{r} = \mathbf{1}_{N_2}$, where $\mathbf{1}_s$ is a vector of s 1 s. Next, given the two vectors $\boldsymbol{r}, \boldsymbol{c}$, we define the polytope of $N_1 \times N_2$ binary matrices

$$\mathrm{U}(\boldsymbol{r}, \boldsymbol{c}) := \{\mathrm{B} \mid \mathrm{B}\mathbf{1}_{N'} = \boldsymbol{r};\ \mathrm{B}^{\mathrm{T}}\mathbf{1}_{N'} = \boldsymbol{c}\}$$

and solve the constrained binary optimal transport problem [34]

$$\mathrm{B}' = \mathrm{argmin}_{\mathrm{B} \in \mathrm{U}(\boldsymbol{r}, \boldsymbol{c})} \langle \mathrm{B}, \mathrm{L} \rangle, \tag{10}$$

where $\langle \mathrm{B}, \mathrm{L} \rangle = \mathrm{tr}(\mathrm{B}^{\mathrm{T}}\mathrm{L})$. The matrix $\mathrm{B}' = ((b'_{ik}))$ describes an optimal $(1, t)$-fairlet decomposition. That is, for any $i \in [N_1]$, if $b'_{ik_1} = \ldots = b'_{ik_t} = 1$ for some $1 \leq k_{i,1} < \ldots < k_{i,t} \leq N_2$, then

$$\Big\{ (x_i, x_{N_1+k_{i,1}}, \ldots, x_{N_1+k_{i,t}})^T,\ i \in [N_1] \Big\}$$

defines a fairlet decomposition of the observed data.

(ii) We shall see that a weighted rectangular loop [15] update $\mathrm{B}'' = ((b''_{ik}))$ on the B' matrix provides an alternative, but potentially non-optimal $(1, t)$-fairlet decomposition of the observed data. Then, for any $i \in [N_1]$, if $b''_{ik_1} = \ldots = b''_{ik_t} = 1$ for some $1 \leq k_{i,1} < \ldots < k_{i,t} \leq N_2$, then $\{(x_i, x_{N_1+k_{i,1}}, \ldots, x_{N_1+k_{i,t}})^T,\ i \in [N_1]\}$ defines a fairlet decomposition of the observed data.

To describe the weighted rectangular loop scheme, let us denote a non-negative weight matrix representing the relative probability of observing a count of 1 at the (i, j)-th cell as $\Omega = (\omega_{ij}) = ((\exp -\lambda_c l_{ik})) \in [0, \infty)$. Then, the likelihood associated with the observed binary matrix $H \in \mathrm{U}(\mathbf{r}, \mathbf{c})$ is

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

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

$$\frac{\mathrm{P}(H_1)}{\mathrm{P}(H_2)} = \frac{\prod_{\{i,j:h_{1,ij}=1,h_{2,ij}=0\}} \omega_{ij}^{h_{1,ij}}}{\prod_{\{i,j:h_{1,ij}=0,h_{2,ij}=1\}} \omega_{ij}^{h_{2,ij}}}. \tag{11}$$

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, the 2×2 matrix with all zero diagonal entries, and all one off-diagonal entries are referred to as *checker-board* matrices. W-RLA is then described in Algorithm 1 in complete generality. The validity of the W-RLA scheme is established in Chakraborty et al. [15].

Algorithm 1: Weighted rectangular loop algorithm [15]

```
Input: An initial binary matrix $A_0 = \mathrm{B}'$, and total number of iterations $T$.
for $t = 1, \ldots, T$ do
    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 $r_2$.
    end
    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 checkerboard.
        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 $A_t = A_{t-1}$
        end
    else
        $A_t = A_{t-1}$
    end
end
Output: The final binary matrix $\mathrm{B}'' = A_T$.
```

Step 2 (Sampling Clustering Indices). The fairlet decomposition of the observed data,

$$\left\{ (x_i, x_{N_1+k_{i,1}}, \ldots, x_{N_1+k_{i,t}})^T, \; i \in [N_1] \right\},$$

induced by the binary matrix B'', is summarized by the set of fairlet centers $\mathbf{u} = (u_1, \ldots, u_m)^{\mathrm{T}} \in \mathcal{U}$, acquired through averaging observations within the respective fairlets. Given a set of fairlet centers $\mathbf{u} = (u_1, \ldots, u_m)^{\mathrm{T}} \in \mathcal{U}$, we adopt a set of strategies developed in Rigon et al. [29] for sampling the clustering indices. Implementation is generally simpler and more efficient compared to mixture models. A straightforward Gibbs sampler implementation showcases favorable mixing properties.

Here and elsewhere, we refer to each of the fairlet centers as units. Suppose $\mathbf{c}_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n)$ denotes the set of clustering indices without the i-th unit, and let $\{C_{1,-i}, \ldots, C_{K,-i}\}$ be the induced partition of the fairlet centers. Suppose $\mathbf{u}_{(k),-i}$ denotes the center of the units $\{u_i : i \in C_{k,-i}\}$. In Gibbs sampling we cyclically reallocate the indicators c_i by sampling from their full conditionals. Then, the conditional distribution of c_i given $\mathbf{c}_{-i}$ is

$$\mathbf{P}(c_i = k \mid \mathbf{c}_{-i}, \lambda_f, \mathbf{u}) \propto \exp\left\{ -\lambda_f \left[\sum_{i' \in C_k} \mathcal{D}(u_{i'}, \mathbf{u}_{(k)}) - \sum_{i' \in C_{k,-i}} \mathcal{D}(u_{i'}, \mathbf{u}_{(k),-i}) \right] \right\}, \tag{12}$$

for $k = 1, \ldots, K$ and for any partition $\mathbf{C} : |\mathbf{C}| = K$.

In summary, an **MC-EM** algorithm to sample from the posterior in Equation (9) involves first finding the optimal fairlet decomposition via (2) and then sampling the clustering indices via the scheme described in (12). A complete **MCMC** scheme to sample from the posterior needs a bit more work. First, given the observed data $\{(\boldsymbol{x}_i, a_i) \in \mathcal{X} \times [2],\ i \in [N]\}$, the optimal fairlet decomposition is obtained via the optimization problem in (10). We propose a non-optimal fairlet decomposition of the data-weighted rectangular loop scheme in Algorithm 1. For given fairlet decomposition, sample the clustering indices via the scheme described in (12). We repeat the steps until convergence.

The proposed Metropolis within Gibbs sampling algorithm automatically enable us to obtain asymptotically exact samples from the target joint posterior [44], and hence, we prefer it over common optimization-based algorithms, such as variational inference [45], which does not enjoy such guarantees and often provides inadequate uncertainty quantification. Further, while gradient-based MCMC algorithms for discrete spaces [46,47] are becoming increasingly popular in the literature, development of of such schemes under the proposed setup involve significant work, since exploring the space of all possible fairlet decompositions require sampling in constrained spaces.

3.2. Posterior Summaries

Suppose $S_T = \{s_1, \ldots, s_T\}$ is 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 yields 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 [48] to obtain $s_{\text{LS}} = \operatorname{argmin}_{s \in S_T} \sum_{i=1}^{N} \sum_{j=1}^{N} (\eta_{ij}(s) - \bar{\eta}_{ij})^2$. Since $s_{\text{LS}} \in S_T$ by construction, the notion of balance is retained in the resulting s_{LS}.

We can also the obtain misclassification probabilities $1 - \bar{\eta}_{ii^\star}$ of the observations, where $i^\star$ denotes the medioid of the cluster to which the i-th observation was allocated in the maximum a posteriori (MAP) clustering. The quantity $1 - \bar{\eta}_{ii^\star}$ approximates the probability that the i-th unit is allocated to a cluster different from the cluster in the MAP clustering.

3.3. Hyperparameter Tuning

We shall now delve into the considerations regarding the selection of the number of clusters K and the temperature parameter $\lambda = (\lambda_f, \lambda_c)^{\mathrm{T}}$ within the joint posterior of clustering indices and fairlet decompositions, as expressed in Equation (9). Firstly, aligning with the approach presented in [29], we conceive the number of clusters denoted as K, not as an intrinsic attribute of the data to be estimated but rather as a parameter reflecting the desired level of granularity for partitioning observations. Hence, we propose the subjective specification of K as intrinsic to the specified loss function, rather than inferring it from the data. Alternatively, during an exploratory phase, conventional techniques such as the "elbow" rule may be employed to determine a suitable K.

It is important to recall that a significant advantage of mixture model-based Bayesian clustering lies in its capacity to automatically deduce the optimal number of clusters (K)

from the available data. Nonetheless, akin to various model-based methodologies, model-based clustering is susceptible to even minor misspecifications in the component specific parametric distributions. Consequently, the derived estimates for the number of clusters (K) through these procedures exhibit inconsistency [19,20]. Conversely, a notable disadvantage inherent in Gibbs posterior-based clustering approaches, including the methodology introduced in this article, is the imperative requirement to fix the number of clusters, designated as K, before initiating the clustering procedure. In developing a Gibbs posterior-based clustering framework with an undetermined number of clusters, the complexities escalate concerning the solicitation of loss, the selection of an appropriate temperature parameter, etc. The direct application of generalized Bayesian clustering with a variable K may result in undesirable properties in the clustering posterior [29]. This presents a challenging yet captivating avenue for prospective investigation.

Next, our attention shifts to the calibration of the temperature parameters $\lambda = (\lambda_f, \lambda_c)^{\mathsf{T}}$. Determining the parameter λ_f within the Gibbs posterior of the fairlet decomposition in (5) is executed by employing the general principles discussed in Holmes and Walker [32] for assigning a value to a power likelihood in generalized Bayesian models. With regard to tuning λ_c in the Gibbs posterior associated with the clustering loss given a specific fairlet decomposition $\mathbf{u} \in \mathcal{U}$, we note that [29] highlighted the association between the generalized Bayes product partition model and mixtures of exponential dispersion models. These models represent a generalization of regular exponential families, characterized by a *dispersion parameter*. It was further illustrated in [29] that λ_c in the definition of the loss function coincides with this dispersion parameter. This alternative probabilistic representation aids in interpreting the loss and facilitates the elicitation of λ_c, an otherwise challenging task as acknowledged in the literature [30,32]. We are now equipped with all of the necessary components to demonstrate the efficacy of the proposed methodology through numerical experiments and real data examples.

Noteworthy, the proposed fair clustering approach and computational strategies work for any factorized clustering loss, e.g., Minkowski's loss, Bregman k-means loss, etc. For the sake of demonstration, we present Gibbs posterior with k-means clustering loss in the following numerical experiments and real data examples.

4. Experiments

4.1. Well-Specified Case

In this simulation, we consider a generating mechanism such that there are two distinct groups present in the data. Moreover, each of the attribute specific components precisely follows an isotropic bivariate normal distribution. Specifically, we consider the following setup. In the first cluster, 20 individuals with $a = 1$ are generated from $N_2(\mu_{11}, S)$ and 30 individuals with $a = 2$ are generated from $N_2(\mu_{21}, S)$. In the second cluster, 30 individuals with $a = 1$ are generated from $N_2(\mu_{12}, S)$, and 20 individuals with $a = 2$ are generated from $N_2(\mu_{22}, S)$, where $\mu_{11} = (4, 4)'$, $\mu_{21} = (2, 2)'$, $\mu_{12} = (10, 10)'$, $\mu_{22} = (8, 8)'$, and $S = 4\mathrm{I}_2$ where I_2 is the two-dimensional identity matrix. This data generation mechanism ensures that individuals with $a = 1$ and $a = 2$ are equally represented in the observed sample. The goal is to obtain two (i.e., we assume $K = 2$ is known) completely balanced clusters along with uncertainty quantification associated with the clustering indices.

The maximum a posteriori estimate of the vanilla k-means clustering via Gibbs posterior [29] in Figure 1 coincides with the k-means estimator and provides clusters with balance 0.62. The maximum a posteriori estimate of the proposed fair k-means clustering via Gibbs posterior in Figure 1 turns out to be very similar to the fair clustering with fairlets estimator [1] and provides clusters with balance 1. The Gibbs posterior-based approaches, unlike the optimization-based approaches Table 1, also provide uncertainty quantification associated with the clustering configurations along with the maximum a posteriori estimate.

Table 1. Overview of Methods.

Method	Fairness	Uncertainty Quantification
K-means	✗	✗
K-means via Gibbs posterior	✗	✓
Fair clustering via fairlets	✓	✗
Fair clustering via Gibbs posterior	✓	✓

4.2. Misspecified Case

In this simulation, we consider a generating mechanism such that there are two distinct groups present in the data. Moreover, each of the attribute specific components does not follow an isotropic bivariate normal distribution. Specifically, we consider the following setup. In the first cluster, 20 individuals with $a = 1$ are generated from $t_2(\mu_{11}, S)$ and 30 individuals with $a = 2$ are generated from $t_2(\mu_{21}, S)$. In the second cluster, 30 individuals with $a = 1$ are generated from $t_2(\mu_{12}, S)$ and 20 individuals with $a = 2$ are generated from $t_2(\mu_{22}, S)$, where $\mu_{11} = (4, 4)'$, $\mu_{21} = (2, 2)'$, $\mu_{12} = (10, 10)'$, $\mu_{22} = (8, 8)'$, $S = 3\mathrm{I}_2$ where I_2 is the two-dimensional identity matrix, $t_\nu(\mu, S)$ is the bivariate Student's t-distribution with mean μ, scale matrix S, and degrees of freedom ν. This data generation mechanism ensures that individuals with $a = 1$ and $a = 2$ are equally represented in the observed sample. Note that for $\nu = 2$, the variance of the $t_\nu(\mu, S)$ distribution does not exist, and under this data generation scheme outliers are expected. The goal is to obtain two (i.e., we assume $K = 2$ is known) completely balanced clusters along with uncertainty quantification of the clustering indices.

The maximum a posteriori estimate of the vanilla k-means clustering via the Gibbs posterior [29] in Figure 2 coincides with the k-means estimator and provides clusters with balance 0.67. The maximum a posteriori estimate of the proposed fair k-means clustering via Gibbs posterior in Figure 2 turns out to be very similar to the fair clustering with fairlets estimator [1] and provides clusters with balance 1. The Gibbs posterior-based approaches, unlike the k-means and the fair clustering with fairlets estimators, also provide uncertainty quantification associated with the clustering configurations along with point estimates.

Under the above misspecified data generative mechanism, we further try fair clustering via Gibbs posteriors based on two distinct clustering losses—k-means loss and Manhattan dissimilarities—to compare them with respect to robustness to outliers. It is important to underscore that neither of these clustering costs aligns with the underlying generative process. To ensure the comparability of outcomes, we establish K = 2 for both methodologies. Subsequently, we compute the corresponding co-clustering matrices, which are presented graphically in Figure 3. The visual representation strongly suggests superior performance of the clustering approach based on Manhattan pairwise dissimilarities when contrasted with the one relying on squared Euclidean loss. In the k-means scenario, the presence of outliers unreliable uncertainty quantification. These observations align with expectations, as the historical use of absolute deviations in lieu of squared losses has been a strategy to enhance the robustness of clustering.

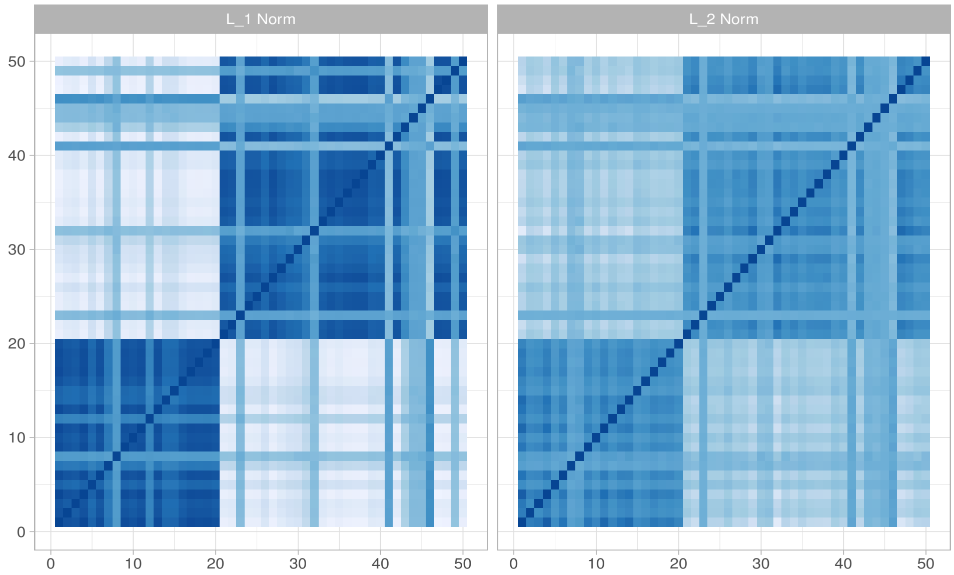

Figure 3. Misspecified case (fair clustering via Gibbs posteriors with L_1 versus L_2 loss). Fair k-means through Gibbs posterior with $K = 2$. We plot the co-clustering probability matrix obtained from the posterior samples. The colors are indicative of probabilities, ranging from white indicating low probability to deep blue indicating a high probability.

5. Benchmark Data Sets

We assess the effectiveness of the proposed approach in comparison to established methodologies using well-known benchmark data sets from the UCI repository [49]. These data sets have been previously examined in the fair clustering literature [1–3].

5.1. Credit Card Data

We opted for numerical attributes, including age and credit limit, to characterize data points within the Euclidean space. Marital status (categorized as married or unmarried) is designated as the sensitive dimension. We conducted a sub-sampling of 120 individuals from the data set, ensuring a target balance of 1. The objective is to achieve three completely balanced clusters (assuming $K = 3$ is known) while simultaneously quantifying uncertainty associated with the clustering indices.

The maximum a posteriori (MAP) estimate for vanilla k-means clustering through a Gibbs posterior aligns with the conventional k-means estimator, yielding clusters with a balance of 0.27. Conversely, the MAP estimate for fair k-means clustering via a Gibbs posterior roughly aligns with the fair clustering with fairlets estimator, resulting in clusters with a balance of 1. Notably, the Gibbs posterior-based approaches in Figure 4, in contrast to both the k-means and fair clustering with fairlets estimators, additionally provide uncertainty quantification pertaining to the clustering configurations alongside MAP estimates.

5.2. Diabetes Data

We analyze a data set sourced from the UCI repository consisting of the health outcomes of patients in relation to diabetes. Numerical attributes such as age and time spent in the hospital serve as points in Euclidean space, while gender is designated as the sensitive dimension. Through a sub-sampling process involving 965 individuals from the data set, a gender ratio of 4:5 is maintained, achieving a target balance of 0.8. The objective is to derive four completely balanced clusters (assuming $K = 4$ is known) while concurrently quantifying uncertainty associated with the clustering indices.

The maximum a posteriori (MAP) estimate for vanilla k-means clustering using the Gibbs posterior aligns with the conventional k-means estimator, yielding clusters with a balance of 0.23. In contrast, the MAP estimate for fair k-means clustering through Gibbs posterior roughly aligns with the fair clustering with fairlets estimator, resulting in clusters with a balance of 0.8. Notably, the Gibbs posterior-based approaches in Figure 5, unlike both the k-means and fair clustering with fairlets estimators, additionally furnish uncertainty quantification related to the clustering configurations alongside the MAP estimators.

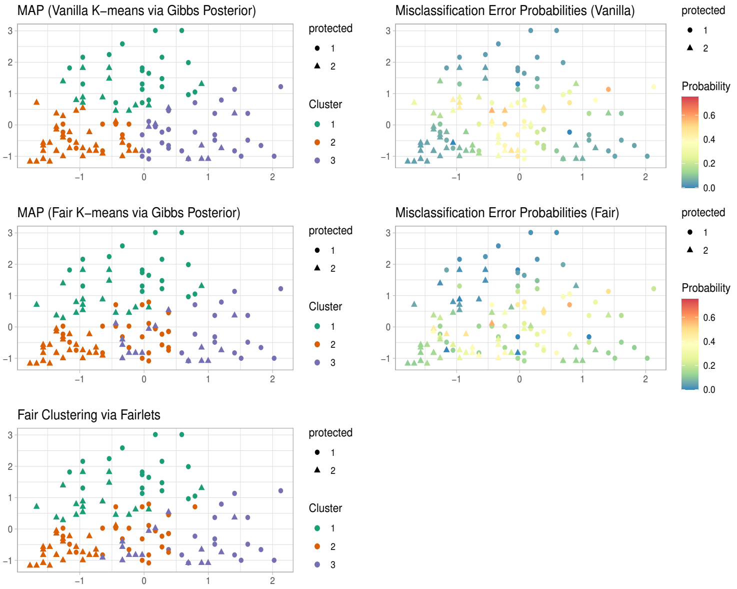

Figure 4. Credit Card Data. K-means and fair k-means through Gibbs posterior with $K = 3$. We plot the maximum a posteriori clustering configurations and misclassification error probabilities obtained from the posterior samples via the scheme in Section 3.2.

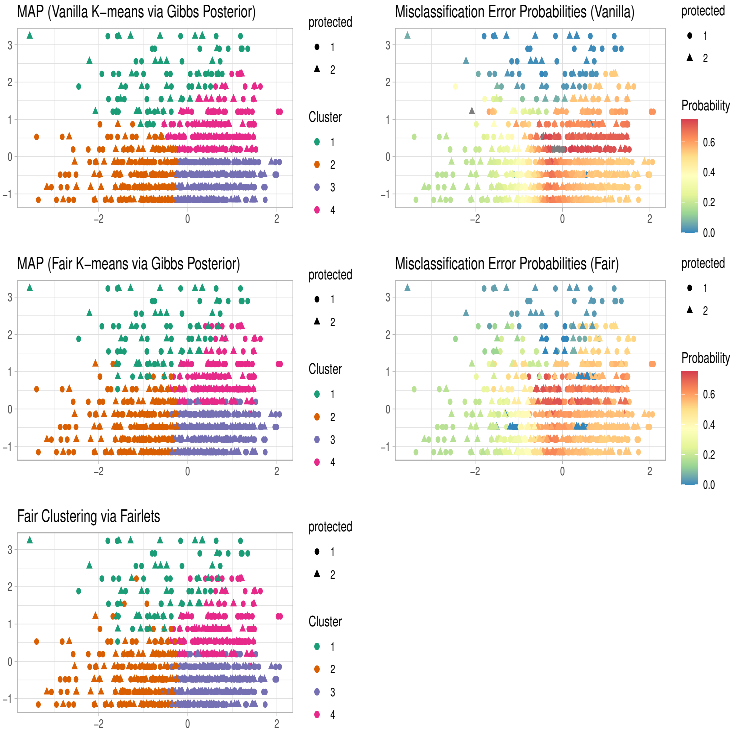

Figure 5. Diabetes Data. Comparison of k-means, fair clustering via fairlets, and fair k-means through a Gibbs posterior with $K = 4$. We plot the maximum a posteriori clustering configurations and misclassification error probabilities obtained from the posterior samples, via the scheme in Section 3.2.

5.3. Portuguese Banking Data

Subsequently, we turn our attention to the Portuguese banking data set, which comprises individual records corresponding to each phone call conducted during a marketing campaign by a Portuguese banking institution. Each record encapsulates information about the client engaged by the institution. For the representation of points in the Euclidean space, we have selected numerical attributes such as age, balance, and duration. Our clustering objective involves achieving balance between married and unmarried clients. Through a sub-sampling process, we reduced the data set to 939 records, maintaining a married-to-not-married client ratio of $(2, 1)$, thereby establishing a target balance of $1/2$. The primary aim is to derive four completely balanced clusters (assuming $K = 4$ is known), coupled with an assessment of uncertainty in the clustering indices.

The maximum a posteriori (MAP) estimate for vanilla k-means clustering using the Gibbs posterior aligns with the conventional k-means estimator, resulting in clusters with a balance of 0.06. Conversely, the MAP estimate for fair k-means clustering through the Gibbs posterior roughly aligns with the fair clustering with fairlets estimator, yielding clusters with a balance of $1/2$. Importantly, the Gibbs posterior-based approaches, in contrast to both the k-means and fair clustering with fairlets estimators, offer additional insight by providing uncertainty quantification associated with the clustering configurations alongside the MAP estimators.

6. Discussion

Despite recent advancements [2,3,12,14] that significantly expanded the scope of fair clustering, uncertainty quantification associated with the optimal clustering configuration remained elusive until recently. In a recent contribution, Chakraborty et al. [15] extended the current body of literature on fair clustering by adopting a novel model-based approach to address clustering under balance constraints. Adopting a generative modeling perspective enabled them to offer valid uncertainty quantification linked to the optimal fair clustering configuration. However, fair clustering frameworks based on such naive mixture models often exhibit fragility in the presence of model misspecification and usually entail notable computational challenges. The main contribution of the current article is the proposed generalized Bayesian fair clustering framework, that inherently provides valid uncertainty quantification while avoiding significant complexities in problem formulation. Secondly, we develop efficient computational algorithms for posterior inference, leveraging techniques from the prevailing literature on clustering based on loss functions as well as computational optimal transport.

A significant limitation of Gibbs posterior-based clustering approaches, including the method proposed in this article, lies in the need to fix the number of clusters, denoted as K, prior to the clustering process. When considering an indeterminate number of clusters within a Gibbs posterior-based clustering framework, the challenges intensify with regard to loss elicitation, the selection of a suitable temperature parameter, and associated considerations. A straightforward implementation of generalized Bayesian clustering with a variable K may lead to undesirable behaviors in the resulting posterior [29]. This provides a challenging yet intriguing avenue for future enquiry.

In conclusion, we reiterate that the generalized Bayesian inference framework presents a potentially advantageous middle ground in trustworthy machine learning applications, bridging the gap between traditional Bayesian inference and inference based on loss functions. This framework allows for valid uncertainty quantification even in the presence of mild model misspecification, all while potentially maintaining computational scalability. Important directions for future investigation involve the development of Gibbs posterior frameworks for various other fair clustering paradigms, including correlation clustering [6], hierarchical clustering [7], functional data clustering [50], etc. The practical significance of generalized Bayesian inference in fair clustering in conjunction with other pivotal facets of modern machine learning such as privacy [12] and robustness [13] is also noteworthy.

Author Contributions: Conceptualization, A.C., A.B. and D.P.; methodology, A.C., A.B. and D.P.; software, A.C.; validation, A.C.; formal analysis, A.C.; investigation, A.C.; resources, A.C.; data curation, A.C.; writing—original draft preparation, A.C., A.B. and D.P.; writing—review and editing, A.C., A.B. and D.P.; visualization, A.C.; supervision, A.B. and D.P.; project administration, A.C.; funding acquisition, A.B. and D.P. All authors have read and agreed to the published version of the manuscript.

Funding: There was no external or internal funding for this work.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: We use well-known benchmark data sets from the UCI repository [49], that have been previously examined in the fair clustering literature.

Acknowledgments: The authors thank the two referees for their comments, which improved the presentation of the material in the paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chierichetti, F.; Kumar, R.; Lattanzi, S.; Vassilvitskii, S. Fair Clustering Through Fairlets. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
2. Böhm, M.; Fazzone, A.; Leonardi, S.; Schwiegelshohn, C. Fair Clustering with Multiple Colors. *arXiv* **2020**, arXiv:2002.07892 [CrossRef]
3. Esmaeili, S.; Brubach, B.; Tsepenekas, L.; Dickerson, J. Probabilistic Fair Clustering. In *Proceedings of the Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 12743–12755.
4. Bera, S.; Chakrabarty, D.; Flores, N.; Negahbani, M. Fair Algorithms for Clustering. In *Proceedings of the Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
5. Kleindessner, M.; Samadi, S.; Awasthi, P.; Morgenstern, J. Guarantees for Spectral Clustering with Fairness Constraints. *arXiv* **2019**, arXiv:1901.08668. [CrossRef]
6. Ahmadian, S.; Epasto, A.; Kumar, R.; Mahdian, M. Fair Correlation Clustering. *arXiv* **2020**, arXiv:2002.02274.
7. Ahmadian, S.; Epasto, A.; Knittel, M.; Kumar, R.; Mahdian, M.; Moseley, B.; Pham, P.; Vassilvitskii, S.; Wang, Y. Fair Hierarchical Clustering. *arXiv* **2020**, arXiv:2006.10221.
8. Kleindessner, M.; Awasthi, P.; Morgenstern, J. A Notion of Individual Fairness for Clustering. *arXiv* **2020**, arXiv:2006.04960. [CrossRef]
9. Mahabadi, S.; Vakilian, A. Individual Fairness for *k*-Clustering. *arXiv* **2020**, arXiv:2002.06742. [CrossRef]
10. Chakrabarty, D.; Negahbani, M. Better Algorithms for Individually Fair k-Clustering. *arXiv* **2021**, arXiv:2106.12150.
11. Chen, X.; Fain, B.; Lyu, C.; Munagala, K. Proportionally Fair Clustering. *arXiv* **2019**, arXiv:1905.03674.
12. Rösner, C.; Schmidt, M. Privacy preserving clustering with constraints. *arXiv* **2019**, arXiv:1802.02497.
13. Bandyapadhyay, S.; Inamdar, T.; Pai, S.; Varadarajan, K.R. A Constant Approximation for Colorful k-Center. *arXiv* **2019**, arXiv:1907.08906.
14. Kleindessner, M.; Awasthi, P.; Morgenstern, J. Fair k-Center Clustering for Data Summarization. *arXiv* **2019**, arXiv:1901.08628. [CrossRef]
15. Chakraborty, A.; Bhattacharya, A.; Pati, D. Fair Clustering via Hierarchical Fair-Dirichlet Process. *arXiv* **2023**, arXiv:2305.17557.
16. Jiang, W.; Tanner, M.A. Gibbs Posterior for Variable Selection in High-Dimensional Classification and Data Mining. *Ann. Stat.* **2008**, *36*, 2207–2231. [CrossRef]
17. Martin, R.; Syring, N. Direct Gibbs Posterior Inference on Risk Minimizers: Construction, Concentration, and Calibration. *arXiv* **2023**, arXiv:2203.09381.
18. Berger, J.O. An Overview of Robust Bayesian Analysis. *Test* **1994**, *3*, 5–124. [CrossRef]
19. Miller, J.W.; Dunson, D.B. Robust Bayesian Inference via Coarsening. *J. Am. Stat. Assoc.* **2019**, *114*, 1113–1125. [CrossRef]
20. Chakraborty, A.; Bhattacharya, A.; Pati, D. Robust probabilistic inference via a constrained transport metric. *arXiv* **2023**, arXiv:2303.10085.
21. Robert, C.P.; Rousseau, J. Nonparametric Bayesian Clay for Robust Decision Bricks. *Stat. Sci.* **2016**, *31*, 506–510. [CrossRef]
22. Ghosal, S.; van der Vaart, A. *Fundamentals of Nonparametric Bayesian Inference*; Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: Cambridge, UK, 2017; Volume 44.
23. Chernozhukov, V.; Hong, H. An MCMC Approach to Classical Estimation. *SSRN* **2002**. [CrossRef]
24. Hong, L.; Martin, R. Model Misspecification, Bayesian versus Credibility Estimation, and Gibbs Posteriors. *Scand. Actuar. J.* **2020**, *2020*, 634–649. [CrossRef]

25. Syring, N.; Martin, R. Robust and Rate-Optimal Gibbs Posterior Inference on the Boundary of a Noisy Image. *Ann. Stat.* **2020**, *48*, 1498–1513. [CrossRef]
26. Wang, Z.; Martin, R. Gibbs posterior inference on a Levy density under discrete sampling. *arXiv* **2021**, arXiv:2109.06567.
27. Bhattacharya, I.; Martin, R. Gibbs posterior inference on multivariate quantiles. *J. Stat. Plan. Inference* **2022**, *218*, 106–121. [CrossRef]
28. Syring, N.; Hong, L.; Martin, R. Gibbs Posterior Inference on Value-at-Risk. *Scand. Actuar. J.* **2019**, *2019*, 548–557. [CrossRef]
29. Rigon, T.; Herring, A.H.; Dunson, D.B. A Generalized Bayes Framework for Probabilistic Clustering. *Biometrika* **2023**, *110*, 559–578. [CrossRef]
30. Bissiri, P.G.; Holmes, C.C.; Walker, S.G. A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B Methodol.* **2016**, *78*, 1103–1130. [CrossRef] [PubMed]
31. Syring, N.; Martin, R. Gibbs Posterior Concentration Rates under Sub-Exponential Type Losses. *Bernoulli* **2023**, *29*, 1080–1108. [CrossRef]
32. Holmes, C.C.; Walker, S.G. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* **2017**, *104*, 497–503.
33. Ahuja, R.K.; Magnanti, T.L.; Orlin, J.B. *Network Flows: Theory, Algorithms, and Applications*; Prentice Hall: Hoboken, NJ, USA, 1993.
34. Villani, C. *Optimal Transport: Old and New*; Springer: Berlin/Heidelberg, Germany, 2008.
35. Lloyd, S.P. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
36. Richardson, S.; Green, P.J. On Bayesian analysis of mixtures with an unknown number of components. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1997**, *59*, 731–792. [CrossRef]
37. Stephens, M. Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2000**, *62*, 795–809. [CrossRef]
38. Celeux, G.; Govaert, G. A classification EM algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.* **1992**, *14*, 315–332. [CrossRef]
39. Banfield, J.D.; Raftery, A.E. Model-based Gaussian and non-Gaussian clustering. *Biometrics* **1993**, *49*, 803–821. [CrossRef]
40. Ghahramani, Z.; Hinton, G.E. *The EM Algorithm for Mixtures of Factor Analyzers*; University of Toronto: Toronto, ON, Canada, 1996.
41. Maugis, C. Model-based clustering of high-dimensional data: A review. *Comput. Stat. Data Anal.* **2014**, *71*, 52–78.
42. Backurs, A.; Indyk, P.; Onak, K.; Schieber, B.; Vakilian, A.; Wagner, T. Scalable Fair Clustering. *arXiv* **2019**, arXiv:1902.03519.
43. Ziko, I.M.; Granger, E.; Yuan, J.; Ayed, I.B. Variational Fair Clustering. *arXiv* **2020**, arXiv:1906.08207.
44. Robert, C.P.; Casella, G. *Monte Carlo Statistical Methods*; Springer Texts in Statistics; Springer: New York, NY, USA, 2004.
45. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
46. Rhodes, B.; Gutmann, M. Enhanced gradient-based MCMC in discrete spaces. *arXiv* **2022**, arXiv:2208.00040.
47. Zanella, G. Informed proposals for local MCMC in discrete spaces. *arXiv* **2017**, arXiv:1711.07424.
48. Dahl, D. *Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model, in Bayesian Inference for Gene Expression and Proteomics*; Cambridge University Press: Cambridge, UK, 2006.
49. Dua, D.; Graff, C. *UCI Machine Learning Repository*; UCI: Irvine, CA, USA, 2017.
50. Chakraborty, A.; Chakraborty, A. Scalable Model-Based Gaussian Process Clustering. *arXiv* **2023**, arXiv:2309.07882.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.