ORIGINAL PAPER

Entropy regularization in probabilistic clustering

Beatrice Franzolini[1] ⓘ · **Giovanni Rebaudo**[2] ⓘ

Abstract

Bayesian nonparametric mixture models are widely used to cluster observations. However, one major drawback of the approach is that the estimated partition often presents unbalanced clusters' frequencies with only a few dominating clusters and a large number of sparsely-populated ones. This feature translates into results that are often uninterpretable unless we accept to ignore a relevant number of observations and clusters. Interpreting the posterior distribution as penalized likelihood, we show how the unbalance can be explained as a direct consequence of the cost functions involved in estimating the partition. In light of our findings, we propose a novel Bayesian estimator of the clustering configuration. The proposed estimator is equivalent to a post-processing procedure that reduces the number of sparsely-populated clusters and enhances interpretability. The procedure takes the form of entropy-regularization of the Bayesian estimate. While being computationally convenient with respect to alternative strategies, it is also theoretically justified as a correction to the Bayesian loss function used for point estimation and, as such, can be applied to any posterior distribution of clusters, regardless of the specific model used.

1 Introduction

Clustering methods are used to detect patterns by partitioning observations into different groups. What are desirable characteristics of clusters depends on the specific applied problem at hand (see e.g., Hennig 2015). Nonetheless, clustering methods are typically motivated by the idea that observations are more similar within the same

✉ Beatrice Franzolini
franzolini@pm.me

Giovanni Rebaudo
giovanni.rebaudo@unito.it

1 Department of Decision Sciences, Bocconi University, Milan, Italy

2 University of Turin & Collegio Carlo Alberto, Turin, Italy

cluster than across clusters (accordingly to a certain definition of similarity). Clustering has been proven useful in a large variety of fields including but not limited to image processing, bio-medicine, marketing, and natural language processing. Clustering methods are used not only to detect sub-groups of subjects, but also for dimensionality reduction (Blei et al. 2003; Petrone et al. 2009), outlier-detection (Shotwell and Slate 2011; Ngan et al. 2015; Franzolini et al. 2023), testing for distributional homogeneity (Rodríguez et al. 2008; Camerlenghi et al. 2019; Denti et al. 2023; Beraha et al. 2021; Balocchi et al. 2023; Lijoi et al. 2023), and data pre-processing (Zhang et al. 2006).

Among clustering techniques, we can distinguish two main classes: model-based and non-model-based. The former methods are built on some assumptions about the sampling mechanism generating the observations. The latter are algorithmic procedures computing clusters' allocations without using distributional assumptions and they typically maximize a certain dissimilarity between clusters (or a measure of similarity of the points clustered together). Contrary to algorithmic clustering techniques, such as k-means or hierarchical clustering, model-based methods allow us to perform inference via rigorous probabilistic assessments, providing a natural way of quantifying uncertainty. Importantly, when the assumption about the data generating mechanism is coherently extendable to future data (for example, in infinite exchangeable models, such as the Dirichlet process mixture Ferguson 1983; Lo 1984), model-based clustering produces coherent predictions for any number of future observations, based on the available past observations. More precisely, by *coherently extendable* we mean preserving *Kolmogorov consistency*, sometimes also called *marginal invariance* (Dahl et al. 2017) or *projectivity* (Betancourt et al. 2022; Rebaudo and Müller 2023), meaning that the marginal distribution of a sample of size *n*, obtained by marginalizing out the clustering configuration, is equal to the restriction of the distribution of larger samples of size $N > n$. Thus, their statistical power is not limited to providing a summary of the observed data, as it happens with algorithmic non-model-based techniques.

Typically, model-based clustering frameworks are equivalent to the assumption that the observations $y_1, \ldots y_n$ are extracted from an infinite population following a mixture

$$y_i \overset{iid}{\sim} \sum_{h=1}^{K} w_h \, k(\cdot; \theta_h) \qquad \text{for } i = 1, \ldots, n,\, n+1, \ldots, \tag{1}$$

where the mixture components $k(\cdot; \theta_h)$ are probability kernels to be interpreted as distributions of distinct clusters in the infinite population, $(w_h, \theta_h)_{h=1}^{K}$ are unknown parameters that determine the relative proportion and the shape of such population clusters, and K is the total number of clusters in the infinite population. K can be either a fixed value or an unknown parameter. In the following, we focus on those models under which either K is unknown or $K = +\infty$. Assuming a fixed finite value for K is often restrictive because it limits the flexibility of the assumption in (1), and should be avoided unless we have strong information/preference about an upper bound in the number of clusters. Indeed, an important and typically unknown parameter is the number of clusters K_n in the observed sample, i.e., the number of occupied components in the mixture in (1). Obviously, $K_n \leq \min(K, n)$. For this reason,

in a framework in which n is let to vary, K is typically either fixed to $+\infty$ (e.g., in Dirichlet process mixtures, Ferguson 1983; Lo 1984) or it is estimated from the data (e.g., mixtures of finite mixtures, see Nobile 1994; Miller and Harrison 2018; Argiento and De Iorio 2022).

One limitation typically encountered in model-based clustering is that the clustering point estimate presents highly unbalanced cluster frequencies. Especially when the number of mixture components is not arbitrarily fixed to a finite small number, the estimated partition tends to include only a few dominating clusters and a large number of sparsely-populated ones. This problem is well-known in Bayesian discrete mixtures such as Dirichlet process mixtures, Pitman-Yor process mixtures, and mixture of finite mixtures. This feature is undesirable and poses important problems in terms of interpretability. High unbalance in the cluster frequencies typically forces us to disregard all observations assigned to small clusters and just interpret the more-populated ones, for which enough observations are available. However, the number of small clusters is often not negligible, so the total number of ignored observations in the interpretation of the cluster is not negligible as well. Disregarding observations assigned to small clusters when it comes to model-based clustering is not justified, especially in light of the fact that the unbalance in clusters' frequencies ultimately appears as a feature of the method and not of the specific data analyzed.

The unbalance of the cluster frequencies can easily be explained as the result of the interaction of the *rich that get richer* property and the unbounded number of clusters in Bayesian mixture models. See Lee et al. (2022) for a recent detailed discussion on the topic. The Bayesian learning mechanism of the *rich that get richer*, as the number of observations increases, increases the probability of observing members of clusters that have already been observed. At the same time, both in infinite mixture models and in mixture of finite mixtures with a prior on the number of components that assigns positive probability on an infinite set, the probability of observing a new cluster is always positive for any n number of observations already allocated. Thus, when new observations are collected the induced learning mechanism tends to both repopulate large existing clusters (due to the *rich that get richer* property) and to create new small clusters (due to the fact that the probability of observing new clusters is positive). The interactions between these two properties naturally reduce into unbalanced clusters. However, correcting unbalance intervening on one of these two properties is not optimal. It requires either fixing a small upper bound for the number of clusters (i.e., K) or breaking probabilistic properties of the model as Kolmogorov consistency of the law of the observable variables (see, for instance, Wallach et al. 2010; Lee et al. 2022).

Here we propose a correction of the clusters' unbalance that affects neither the Bayesian learning mechanism nor the attractive probabilistic properties of model-based clustering. Our proposal is theoretically justified as a correction to the loss function used for the Bayesian estimate that explicitly reflects the loss in which the analyst incurs when the point estimate of the clustering configuration is uninterpretable.

The content of the paper is organized as follows. Section 2 presents the study of the cost functions involved in BNP clustering models and explains the presence of

noisy and sparsely populated clusters typically observed in the posterior estimates of these models. Then, in light of this study, our computationally convenient and theoretically justified solution to reduce the number of sparsely populated clusters is presented in Sect. 3 and showcased on simulated and real data, respectively in Sects. 4 and 5. The code to reproduce all results in the paper is available at https://github.com/GiovanniRebaudo/ERC.

2 Implied costs functions in Bayesian nonparametric clustering

The main goal of clustering techniques is to estimate a partition of the observed sample, more than the distribution of the whole ideal population in (1). The partition that one wants to estimate can be encoded using a sequence of subject-specific labels $(c_1, \ldots, c_n)$ taking value in the set of natural numbers such that $c_i = c_j = c$ if and only if y_i and y_j belong to the same cluster and follow the same mixture component $k(\cdot; \theta_c)$, i.e. $y_i \mid c_i \overset{ind}{\sim} k(\cdot; \theta_{c_i})$ for $i = 1, \ldots, n$. The indicators $(c_1, \ldots, c_n)$, as just defined, are affected by the label switching problem (see, for instance, Stephens 2000; McLachlan et al. 2019; Gil-Leyva et al. 2020). In the following, we assume them to be encoded in order of appearance. This means that $c_1 = 1$, i.e. the first observation y_1 always belongs to the *first* cluster. Then either $c_2 = c_1 = 1$, if the second observation y_2 is clustered together with y_1, or $c_2 = 2$, otherwise, and so on and so forth. Note that, thanks to exchangeability, we can focus on an arbitrary order of the observations without affecting the joint law of the sample and thus posterior inference of the clustering configuration. The likelihood for $\boldsymbol{c} = (c_1, \ldots, c_n)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{K_n})$ is

$$\mathcal{L}(\boldsymbol{c}, \boldsymbol{\theta}; \boldsymbol{y}) = \prod_{c=1}^{K_n} \prod_{i: c_i = c} k(y_i; \theta_c). \tag{2}$$

When K_n is unknown, the clustering labels in (2) cannot be estimated with a standard frequentist approach. In fact, when the maximum likelihood estimator (MLE) for (2) exists, it coincides with the vector of MLEs $(\hat{\theta}_1, \ldots, \hat{\theta}_n)$, where each $\hat{\theta}_i$ is obtained considering one observation at a time and the independent models $y_i \sim k(y_i \mid \theta_i)$, for $i = 1, \ldots, n$. This result is an immediate consequence of

$$\max_{(\boldsymbol{\theta}, \boldsymbol{c})} \sum_{i=1}^{n} \log k(y_i; \theta_{c_i}) \leq \sum_{i=1}^{n} \max_{\theta_i} \log k(y_i; \theta_i)$$

and when there are no joint constraints among the parameters in $\boldsymbol{\theta}$ the equality holds.

Moreover, note that under typical mixture model assumptions for clustering, we have that $\hat{\theta}_1 \neq \ldots \neq \hat{\theta}_n$. For instance, when k is a multivariate Gaussian density and θ is the pair of mean vector and variance matrix of the Gaussian component, such the MLE entails a number of clusters equal to the number of distinct observed values, that by model's assumptions equals n with probability 1. Thus, no information on clusters can ever be gained through MLE and overfitting is unavoidable unless

one relies on strong restrictions of the parameter space (cfr. also with Theorem 1 and 2 in Casella et al. 2014, where the use of a uniform prior over all possible partitions is considered). In this regard, note that maximizing (2) is not the same as computing the nonparametric maximum likelihood estimator (Lindsay 1995; Polyanskiy and Wu 2020; Saha and Guntuboyina 2020) for the mixture model in (1).

Differently, Bayesian models, and in particular Bayesian nonparametric (BNP) models, are largely used for model-based clustering, since priors act as penalties, shrinking the number of distinct clusters. The vast majority of Bayesian models for clustering rely on a prior for c and K_n defined through an exchangeable partition probability function (EPPF) (see, Pitman 1996) and, independently, a prior P is used for the unique values $(\theta_1, \dots, \theta_{K_n})$. Recall that an EPPF characterizes the distribution of an exchangeable partition, with $\mathrm{EPPF}(n_1, \dots, n_{K_n})$ being the probability of observing a particular (unordered) partition of n observations into K_n subsets of cardinalities $\{n_1, \dots, n_{K_n}\}$.

Therefore, the corresponding posterior distribution is

$$p(K_n, c, \theta \mid y) \propto \prod_{c=1}^{K_n} \prod_{i:c_i=c} k(y_i; \theta_c) \times \mathrm{EPPF}(n_1, \dots, n_{K_n}) \times P(d\theta), \qquad (3)$$

which can be equivalently represented as the cost function $-\log(p(K_n, c, \theta \mid y))$, i.e.

$$C(K_n, c, \theta; y) = C_{\mathrm{lik}}(K_n, c, \theta; y) + C_{\mathrm{part}}(K_n, c; \alpha) + C_{\mathrm{base}}(K_n, \theta),$$

which is the sum of three terms, that in the following are named respectively likelihood cost, partition cost, and base cost.

As already mentioned, the minimum likelihood cost

$$C_{\mathrm{lik}}(K_n, c, \theta; y) = -\sum_{c=1}^{K_n} \sum_{i:c_i=c}^{n} \log k(y_i; \theta_c)$$

typically corresponds to K_n equal to the number of distinct observed values. The remaining two costs are those defined by the prior of the model and their marginal behavior is described here below. Clearly, any inference result has to be derived based on the whole posterior distribution in (3), which is the result of the joint, and not marginal, effects of all three costs. Nonetheless considering one cost at a time allows us to gain insights regarding the estimation procedure and the frequentist penalties induced by the prior.

2.1 Base cost

A lot of attention in the literature has been devoted to the choice of the EPPF and many alternatives are available (see, for example, Antoniak 1974; Green and Richardson 2001; Lijoi et al. 2007; Lijoi and Prünster 2010; De Blasi et al. 2015; Camerlenghi et al. 2018; Miller and Harrison 2018; Greve et al. 2022), while, except for few cases, mainly within repulsive mixtures (Petralia et al. 2012; Xu et al. 2016;

Bianchini et al. 2020; Xie and Xu 2020; Beraha et al. 2022), the role of the base cost appears partially overlooked within the Bayesian methodology literature.

However, when BNP clustering methods are applied in practice, the choice of an appropriate base distribution is known to be crucial. The most common choice is to use an independent prior on the unique values so that $\theta_c \overset{iid}{\sim} P_0$ and

$$C_{\text{base}}(K_n, \boldsymbol{\theta}) = -\sum_{c=1}^{K_n} \log P_0(d\theta_c),$$

where the variance of the distribution P_0 is known to play an important role in the estimation process and, typically, the higher the variance of P_0 the lower the number of clusters identified by the posterior (cfr., e.g. Gelman et al. 2013, p. 535). This phenomenon can be explained by looking at the joint distribution induced by P_0 on the unique value. Higher values of the variance correspond to a joint distribution with a smaller mass around the main diagonal and, therefore, a higher base cost for those vectors $(\theta_1, \ldots, \theta_{K_n})$ whose components are similar, thus ultimately favoring the variability of the unique values and penalizing many overlapping clusters. Consider for instance the case of P_0 set to a univariate normal distribution centered in μ and with variance σ^2, we have

$$C_{\text{base}}(K_n, \boldsymbol{\theta}) = \frac{K_n}{2} \log(2\pi) + \frac{K_n}{2} \log \sigma^2 + \frac{1}{2} \sum_{c=1}^{K_n} \frac{(\theta_c - \mu)^2}{\sigma^2}.$$

When the variance is increased from σ^2 to λ^2, the base cost increases closer to the K_n-dimensional vector $(\mu, \ldots, \mu)$. More formally, defining the K_n-sphere $\boldsymbol{\theta} \in \mathbb{R}^{K_n}$ such that $\sum_{c=1}^{K_n} (\theta_c - \mu)^2 = K_n \frac{\log(\lambda^2/\sigma^2)\sigma^2\lambda^2}{\lambda^2 - \sigma^2}$, we have that the cost increases for vectors $(\theta_1, \ldots, \theta_{K_n})$ corresponding to points inside the sphere and decreases for those vectors corresponding to points outside the sphere. This causes also a reduction in the relative cost of those vectors located far from the main diagonal compared to the cost of the vectors closer to the main diagonal. To clarify this point, Fig. 1 shows the cost function shift caused by an increase in variance from 1 to 3 in the case of $K_n = 2$ and P_0 univariate normal centered in 0. In Fig. 1, the number of cluster is fixed to $K_n = 2$ and the cost associated to different clusters' locations $(\theta_1, \theta_2) \in [-3, 3]^2$ is considered. Figure 1a, b show the base cost in the whole plane $[-3, 3]^2$, while Figs. 1c, d show how the base cost changes based on the distance between θ_1 and θ_2, i.e, $|\theta_1 - \theta_2|$. Figures 1c, d are obtained considering a grid of equally spaced points in the plane $[-3, 3]^2$. Comparing the two scenarios of variance equal 1 and 3, it is evident as the increase in variance results in a smaller penalization of the distance between cluster locations. In practice, P_0 is usually set to be a continuous scale mixture, where the mixed density is conjugate to the kernel k for computational convenience, while the mixing density is used to increase appropriately the marginal scale of the mixture P_0.
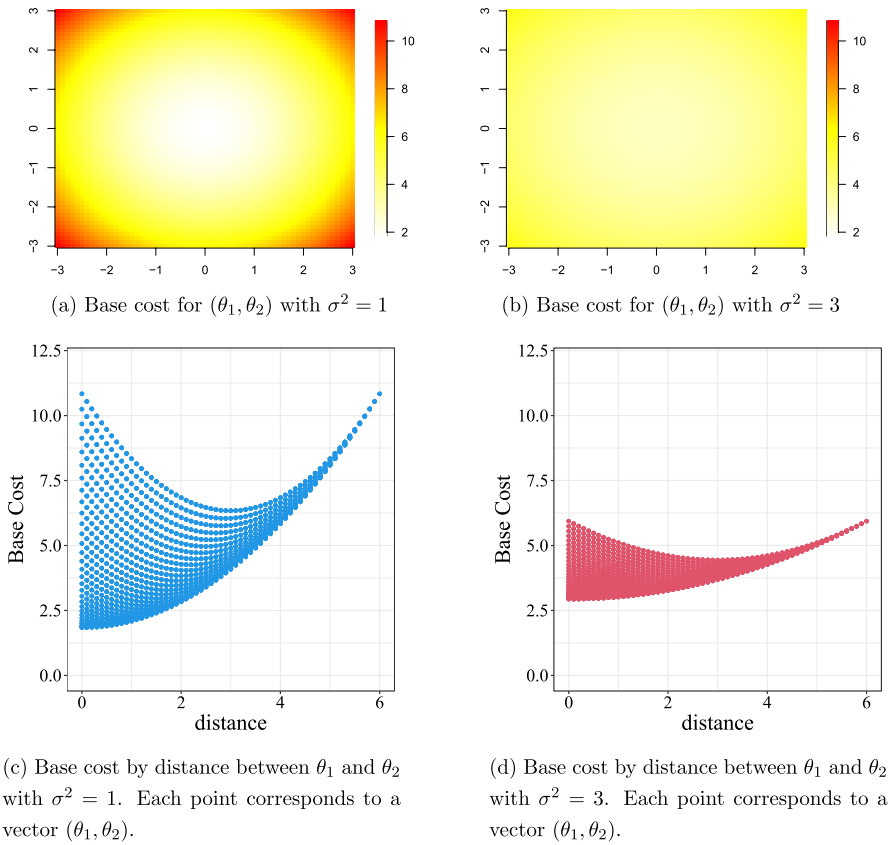
(a) Base cost for (θ_1, θ_2) with $\sigma^2 = 1$

(b) Base cost for (θ_1, θ_2) with $\sigma^2 = 3$

(c) Base cost by distance between θ_1 and θ_2 with $\sigma^2 = 1$. Each point corresponds to a vector (θ_1, θ_2).

(d) Base cost by distance between θ_1 and θ_2 with $\sigma^2 = 3$. Each point corresponds to a vector (θ_1, θ_2).

Fig. 1 $K_n = 2$, bivariate normals with $\sigma^2 = 1$ and $\sigma^2 = 3$

2.2 Partition cost

Finally, let us comment on the partition cost C_{part}. Its behavior is less straightforward and we consider here only two important and widely used cases: Dirichlet process mixtures (DPM) and Pitman-Yor process (Pitman and Yor 1997) mixtures (PYPM). With a DPM model, up to an additive constant, we have

$$C_{\text{part}}(K_n, \boldsymbol{c}; \alpha) = -K_n \log \alpha - \sum_{c=1}^{K_n} \log \Gamma(n_c),$$

where α is the concentration parameter of the Dirichlet Process. The DPM partition cost tends to favor parsimonious values of K_n (with respect to the likelihood cost that in general tends to favor $K_n = n$). However, contrary to the base cost, it depends also on clusters' frequencies.

Figure 2 showcases the partition cost of DPM for different values of what we refer henceforth to as the entropy of the frequencies $(n_1, \ldots, n_{K_n})$, i.e.
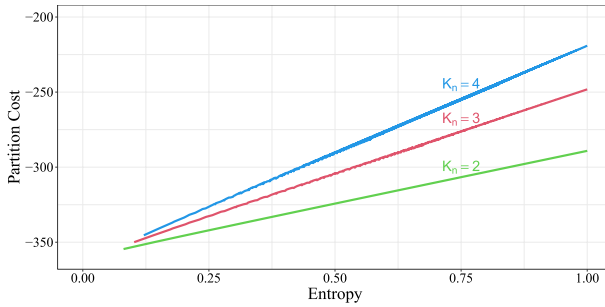
Fig. 2 Average partition cost as a function of entropy of the partition in a DPM model with $\alpha = 1$ for $n = 100$ observations clustered into 2, 3, and 4 clusters. Plotted values are obtained analytically: for each possible partition, the value of the entropy and the cost are computed and, then, the cost is averaged across partitions with the same entropy, keeping the number of clusters K_n fixed

$$S(n_1, \ldots, n_{K_n}) = -\sum_{c=1}^{K_n} \frac{n_c}{n} \log_{K_n} \frac{n_c}{n}.$$

Overall the EPPF acts favoring frequencies $(n_1, \ldots, n_{K_n})$ with low entropy and thus, roughly speaking, higher sample variance of the frequencies. However, this feature ultimately results in two distinct effects: one acting on the total number of occupied clusters K_n and another acting on the variance of the clusters' frequencies $(n_1, \ldots, n_{K_n})$. Even though these two features both favor a reduced entropy, they entail very different scenarios in terms of estimated clustering structure, especially from an applied and practical point of view. Penalizing large numbers of clusters is typically desirable in applications because an elevated number of clusters may be difficult to interpret. However, a partition with few dominating clusters and many sparsely populated clusters is often highly undesirable because it is hard to interpret unless one decides to ignore all the information contained in the small clusters and focus only on the dominating ones. See also Green and Richardson (2001) for a study of the posterior entropy in the Dirichlet process mixture and Greve et al. (2022) for more details on entropy in mixtures of finite mixture models. In the case of a PYPM the partition cost, up to an additive constant, equals

$$C_{\text{part}}(K_n, \boldsymbol{c}; \alpha, \sigma) = -\sum_{c=1}^{K_n} \log(\alpha + \sigma(c-1)) - \sum_{c=1}^{K_n} \log \Gamma(n_c - \sigma) + K_n \log \Gamma(1 - \sigma).$$

Despite that the EPPFs are different, Figs. 2 and 3 show in both processes a closely similar behavior in terms of entropy penalization. This tendency is coherent with the fact that the posterior unbalance of cluster frequencies is typically observed in practice under both models, although they are built on different EPPFs.

Note that Figs. 2 and 3 provide us with insights into the behavior of the EPPFs evaluated (analytically from the aforementioned expressions) in correspondence of different vectors of clusters' frequencies $(n_1, \ldots, n_{K_n})$, i.e., the probability of a specific clustering configuration with unordered frequencies
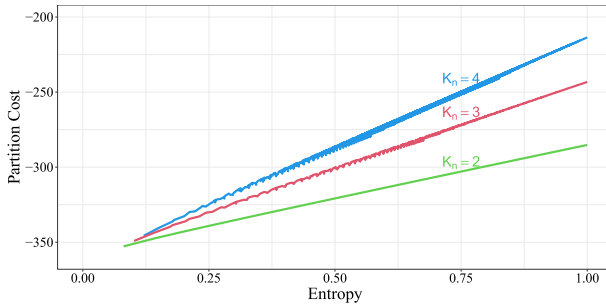
Fig. 3 Average partition cost as a function of entropy of the partition in a PYPM model with $\alpha = 1$ and $\sigma = 0.5$ for $n = 100$ observations clustered into 2, 3, and 4 clusters. Plotted values are obtained analytically: for each possible partition, the value of the entropy and the cost are computed and, then, the cost is averaged across partitions with the same entropy, keeping the number of clusters K_n fixed

$\{n_1, \ldots, n_{K_n}\}$. In particular, they show how the EPPF associates different levels of penalty with different values of entropy. In this regard is important to stress that the vectors $(n_1, \ldots, n_{K_n})$ are not in a one-to-one correspondence with the partitions, and the number of partitions corresponding to certain frequencies varies across vectors. For instance, when $n = 100$, there exist $\binom{100}{50, 50} \approx 1.01e + 29$ distinct partitions corresponding to the vector of frequencies $(50, 50)$ and $\binom{100}{25, 25, 25, 25} \approx 1.61e + 57$ distinct partitions corresponding to the vector of frequencies $(25, 25, 25, 25)$. The number of partitions per different vectors of frequencies, which is not depicted in Figs. 2 and 3, does affect both estimates of marginal quantities, such as the number of clusters K_n, as well as point estimates of the clustering configuration that are different from the MAP (maximum a posteriori).

If we are interested in estimating the number of clusters K_n, we should note that the number of possible partitions rapidly changes with K_n accordingly to Stirling numbers of the second kind. More precisely, the Stirling number of the second kind counts the number of different partitions of n objects into K_n non-empty unordered subsets and can be computed as

$$\frac{1}{K_n!} \sum_{i=0}^{K_n} (-1)^{K_n - i} \binom{K_n}{i} i^n.$$

This information must be combined with the partition cost, as represented in Figs. 2 and 3, if we are interested in fully understanding the impact of the EPPF on the marginal prior and posterior distributions of K_n. Combining the two features (i.e., the partition cost per each vector of frequencies and the number of partitions per each vector of frequencies) the typical partition cost strongly penalized too many clusters suggested by the likelihood costs, i.e. $K_n = n$, but still favors a small (higher than

1) number of clusters that adaptively increases with the sample size n (see e.g., De Blasi et al. 2015).

3 Regularized-entropy estimator

Once the posterior distribution $\mathbb{P}(c \mid y_{1:n})$ over the space of partitions is obtained, typically thanks to a Markov Chain Monte Carlo algorithm, a point estimate $\hat{c}$ of the partition can be obtained accordingly to the decision-theoretic approach of Bayesian analysis. More precisely, $\hat{c}$ is obtained by minimizing the Bayesian risk, i.e, the expected value of a loss function $L(c, \hat{c})$ with respect to the posterior

$$c^* = \underset{\hat{c}}{\operatorname{argmin}} \, \mathbb{E}[L(c, \hat{c}) \mid y_{1:n}] = \underset{\hat{c}}{\operatorname{argmin}} \sum_c L(c, \hat{c})\mathbb{P}(c \mid y_{1:n}),$$

where $L(c, \hat{c})$ is the loss in which we incur using $\hat{c}$ as estimates when the partition takes the value c. How to interpret and elicit the loss in practice can change according to the philosophical point of view. See, for instance, Robert (2007). Often in parameter estimation, the loss is interpreted as the cost of choosing $\hat{c}$ instead of the ideally optimal parameter value c (sometimes interpreted as the *truth*). In a more subjective Bayesian framework, it can be interpreted, together with the model and prior, in terms of the preferences implied on the possible parameter values c via the Bayesian risk (Savage 1972). Finally, also in a more frequentist framework, the loss can be chosen in terms of the implied properties of the estimator $\hat{c}$ of the unknown true parameter.

Despite the different philosophical justifications, rarely, in applied Bayesian clustering analysis, a 0–1 loss function and the resulting MAP estimator are employed due to the large support of the posterior and the fact that the 0–1 loss function does not reflect different levels of distance between two non-coinciding partitions. Widely used alternatives in applications are Binder loss (Binder 1978) or variation of information (VI) loss (see, Meilă 2007; Wade and Ghahramani 2018; Dahl et al. 2022b).

Algorithm 1 Entropy-regularized estimate

Inputs: chain of partitions $\{c_m, m = 1, \dots, M\}$ sampled from the posterior, λ
Output: point estimate c^*

1: Compute $S(c_m)$ for $m = 1, \dots, M$

2: Compute $w_m = \exp\{\lambda S(c_m)\}$ for $m = 1, \dots, M$

3: $\bar{w}_m \leftarrow w_m / \sum_m w_m$ for $m = 1, \dots, M$

4: Generate $\{\tilde{c}_m, m = 1, \dots, M\}$, sampling with replacement from $\{c_1, \dots, c_M\}$ with prob. $\{\bar{w}_m, m = 1, \dots, M\}$

5: $c^* \leftarrow \operatorname{argmin} \sum_{m=1}^M L(\tilde{c}_m, \hat{c})$

We have already stressed how a large presence of noisy clusters is typically undesirable in practice and we claim that this aspect should be reflected in the loss function used for point estimation so that the loss of each partition is proportional to its entropy. To do so, consider any possible loss function $L(\boldsymbol{c}, \hat{\boldsymbol{c}})$ one would like to use to derive the estimate, we can define a new loss function, that we named entropy-regularized, as

$$\bar{L}(\boldsymbol{c}, \hat{\boldsymbol{c}}) = \exp\{\lambda S(\boldsymbol{c})\} L(\boldsymbol{c}, \hat{\boldsymbol{c}}),$$

where, with a little abuse of notation w.r.t. the previous section, $S(\boldsymbol{c})$ is the entropy of the partition identified by $\boldsymbol{c}$ and $\lambda \in \mathbb{R}$. Recall that the base of the logarithm involved in the computation of $S(\boldsymbol{c})$ changes with the argument $\boldsymbol{c}$ and it is equal to the number of unique values in $\boldsymbol{c}$ so that $S(\boldsymbol{c}) = 1$ can be obtained for any number of non-empty clusters $K_n \geq 2$ (provided that $n/K_n \in \mathbb{N}$). Clearly, when λ is positive, for any candidate estimate $\hat{c}$, the loss function is inflated in correspondence of partitions $\boldsymbol{c}$ with high entropy, as desired.

Minimizing the expected entropy-regularized loss function $\bar{L}(\boldsymbol{c}, \hat{\boldsymbol{c}})$ with respect to the posterior is equivalent to minimizing the original loss function $L(\boldsymbol{c}, \hat{\boldsymbol{c}})$ with respect to an entropy-regularized version $\bar{\mathbb{P}}[\boldsymbol{c} \mid y_{1:n}]$ of the posterior distribution, i.e.

$$\bar{\mathbb{P}}[\boldsymbol{c} \mid y_{1:n}] \propto \exp\{\lambda S(\boldsymbol{c})\} \mathbb{P}[\boldsymbol{c} \mid y_{1:n}].$$

This result, while immediate to prove, is highly desirable, because it allows implementation of the entropy-correction in a very straightforward and computationally feasible way which is described in Algorithm 1. Before computing summaries of the posterior, a resampling step is applied.

More precisely, each sample from the posterior is resampled with probability proportional to $\exp\{\lambda S(\boldsymbol{c})\}$ so that an entropy-regularized version of the whole posterior distribution is obtained, thanks to a sampling importance resampling step. Then, in the last step of the algorithm the original loss function $L(\boldsymbol{c}, \hat{\boldsymbol{c}})$ is minimized with respect to the entropy-regularized version of the posterior. Note that to solve such an optimization step we can rely on any of the effective algorithms available in the literature for the optimization of non-entropy regularized losses. See e.g. Rastelli and Friel (2018); Dahl et al. (2022a). In particular, we use the greedy algorithm described in Dahl et al. (2022b) as implemented in the R library salso (Dahl et al. 2022a) to perform the analysis presented in this work.

Thanks to the properties of the importance sampling procedure, the point estimate obtained minimizing $\sum_{m=1}^{M} L(\tilde{\boldsymbol{c}}_m, \hat{\boldsymbol{c}})$ is asymptotically equivalent to the solution of the minimization problem $\sum_{m=1}^{M} \bar{L}(\boldsymbol{c}_m, \hat{\boldsymbol{c}})$. However, there is a potential drawback of Algorithm 1, which stems from the finite dimension M of the original sample from the posterior, $\boldsymbol{c}_m, m = 1, \ldots, M$. Although Algorithm 1 is easy to implement, it may significantly reduce the number of MCMC iterations considered in the minimization problem. To overcome this issue, one possible solution may be to monitor the effective sample size (ESS) of the importance sampling step of Algorithm 1 (see e.g., Liu 1996) that can be approximated as

$$\text{ESS} = \frac{1}{\sum_{m=1}^{M} w_m^2} \tag{4}$$

where w_m are defined in Algorithm 1 and M is the number of initial draws for the posterior. When the ESS is below a certain threshold, it can be increased by increasing the number of initial draws M from the posterior. It is important to note that this use of the ESS indicator deviates from the conventional practice. The ESS is typically employed to measure the mixing performance of sampling algorithms, having as optimal value for the relative effective sample size ESS/M the value of 1. In this standard use, the ESS can be interpreted as the approximate number of independent draws obtained from a target distribution. However, this is not the case in our context. In fact, we should always expect a relative sample size ESS/M lower than 1 to ensure that the entropy regularization has the desired effect on the estimates, the lower *ESS/M* the higher the effect of the regularization. Moreover, here there is no target distribution we are referring to while computing the ESS. Roughly speaking, the entropy regularization shifts the *importance* (i.e., the posterior density) towards specific areas in the support of the posterior and the ESS in (4) may serve only as a practical indicator of how well those areas have been previously explored by the original chain.

Finally, note that the choice of λ plays an important role in defining the clustering estimator (as well as the choice of the not-regularized loss and the probabilistic clustering model assumptions). The hyperparameter λ can be elicited jointly with the rest of the prior settings, (e.g., prior, likelihood, and loss) in a Bayesian decision framework according to the preference on the point estimate of the clustering. In particular, we recommend choosing λ large if we want a stronger regularization. How large depends on the specific analysis and the other model and prior choices. In practice, if the goal is to use clustering just as a data summary can be easy and meaningful to try different values of λ on a grid and see what produces more interpretable results in a cross-validation spirit.

4 Simulation studies

4.1 Univariate Gaussian mixtures

We provide here a simulation study, where $n = 1000$ observations are sampled from 3 distinct and well-separated univariate Gaussian distributions centered in -4, 0, and 4 and with unitary variance. Here we refer to "ground true" clustering as the one implied by the membership indicators of the Gaussian kernels under the data-generating truth. We employ a Normal-Normal DPM, with the base distribution centered at 0 and variance equal to 1. We compare the posterior estimates obtained by minimizing the Binder loss function and the entropy-regularized Binder loss function. We set the concentration parameter $\alpha = 1$, perform 20 000 MCMC simulations, and use the first 5000 as burn-in. See Sect. 2 for the results in the exact same setting but with a Gamma hyperprior for the concentration parameter.
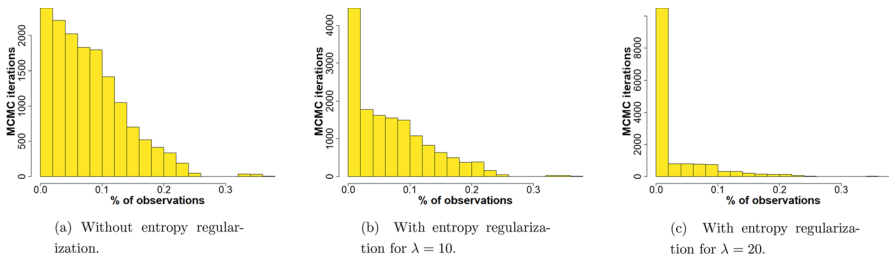
Fig. 4 Gaussian simulation study. Percentage of observations in sparsely-populated clusters before and after entropy-regularization. Sparsely populated clusters are here defined as clusters containing 10% or less of observations. The horizontal axis denotes the percentage of observations (out of 1000) that are assigned to those clusters. The y-axis represents the number of MCMC samples (out of 15 000)

Defining as sparsely populated clusters those clusters containing 10% or less of observations, we found that in almost a third (4755 out 15 000) of the MCMC iterations, 10% or more of the observations are allocated into sparsely populated clusters, while in almost two thirds (9306 out of 15 000) of MCMC iterations, 5% or more of the observations are allocated into sparsely populated clusters, see Fig. 4a. The same counts after entropy-regularization of the posterior (as described in the previous section) are, with $\lambda = 10$, 4088 and 7888 out 15 000, see Fig. 4b, and, with $\lambda = 20$, 1375 and 3290 out 15 000, see Fig. 4c. Notice that coherently with the interpretation of the regularization in terms of the loss function, the regularized posterior should be intended only as a computational tool to provide summaries of the posterior distribution (e.g., point estimates and credible balls) and not as a posterior distribution itself.

Finally, Fig. 5 shows the ground truth and the estimated clusters with and without entropy regularization. They highlight how the regularization allocates observations from noisy clusters into dominating ones.

The main purpose of regularization is to provide a more interpretable and possibly more parsimonious representation of the dataset at hand without disregarding observations. Therefore, in general, the procedure prioritizes interpretability over
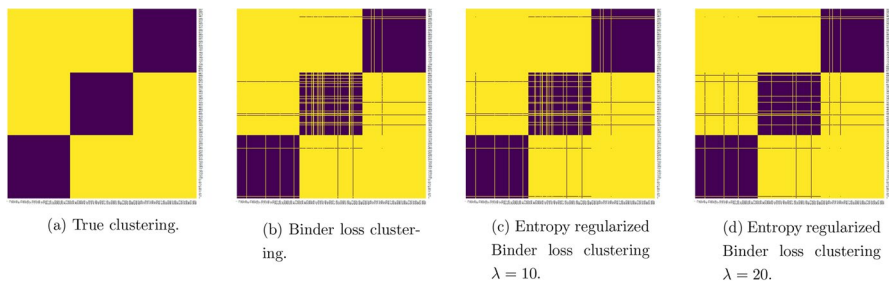

Fig. 5 Gaussian simulation study. Estimated clustering for the simulation study darker squares denote couples of observations clustered together. Panel (**a**) shows the ground true clustering. Panel (**b**) shows the clustering minimizing the Binder loss. Panel (**c**) shows the clustering minimizing the entropy-regularized Binder loss for $\lambda = 10$ and panel (**d**) for $\lambda = 20$

the recovery of a frequentist truth. Note also that even in the ideal frequentist situation of knowing the true data simulation density the misclassification rate will be typically low, but not zero, if the mixture kernels have overlapping supports as in the Gaussian scenario. However, monitoring misclassification errors in simulation studies can still be useful as it provides insights into how entropy regularization redistributes observations to achieve more balanced cluster frequencies. In this study, the application of regularization results in a reduction in misclassification errors, which is consistent with the fact that the highly unbalanced clusters are induced by the learning mechanism of BNP mixtures rather than the data itself. In particular, the number of observations misclassified (with respect to the simulation truth) with the Binder loss point estimate ($\lambda = 0$) is 61, with the regularization with $\lambda = 10$ is 54 and with $\lambda = 20$ is 30.

Figure 6 shows the cluster frequencies for the three point-estimates. Note that in this simple univariate Gaussian kernel simulation scenario we can obtain the correct number of occupied components (i.e., clusters) and thus avoid sparsely populated clusters in the point estimate also using the VI loss (with the default parameter $a = 1$ as implemented in SALSO) instead of considering the Binder loss (with the default parameter $a = 1$). However, without entropy regularization, both the VI loss and the Binder loss entail sparsely populated clusters in more complex scenarios, such as the multivariate simulation study presented in the next Sect. 4.2 and real-world dataset considered in the Sect. 5, respectively.

4.2 Multivariate Bernoulli mixtures and latent class analysis

In this section, we discuss the results of a synthetic numerical experiment involving multivariate Bernoulli data. Results are obtained employing a DPM with independent Bernoulli kernels, such that the likelihood is

$$y_i \mid (w_h)_{h \geq 1}, (p_{1,h}, \dots, p_{J,h})_{h \geq 1} \overset{iid}{\sim} \sum_{h=1}^{\infty} w_h \left(\prod_{j=1}^{J} p_{j,h}^{y_i} (1 - p_{j,h})^{1-y_i} \right)$$

where the hth component has weight w_h and it is characterized by the vector of probabilities $(p_{1,h}, \dots, p_{J,h})$. The model allows for the estimation of latent classes, which
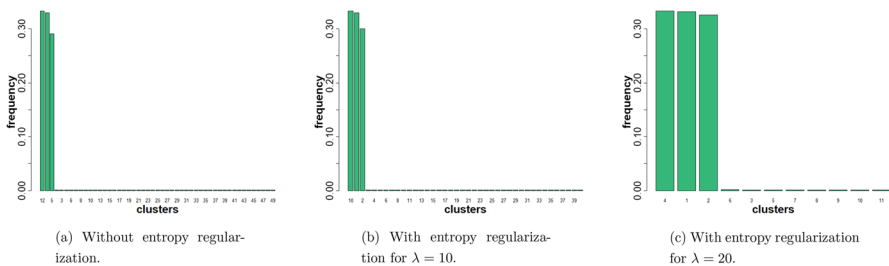

(a) Without entropy regularization.

(b) With entropy regularization for $\lambda = 10$.

(c) With entropy regularization for $\lambda = 20$.

Fig. 6 Gaussian simulation study. Estimated number of clusters and clusters' frequencies

are commonly used in latent class analysis (LCA) to analyze multivariate discrete outcomes, often binary in nature.

In LCA, each latent class is represented by a mixture component, and the observations within each class are assumed to be independent. This assumption holds for LCA even though, typically, the observed variables are assumed to be statistically dependent. This is a crucial aspect of LCA: the classes are indeed used to represent the observed dependence (see, for instance, McCutcheon 1987). The rationale behind the approach is that estimating the dependence across J binary outcomes is often challenging, particularly when J is large, as a J-variate Bernoulli distribution requires $J^2 - 1$ parameters to be estimated. The goal of LCA is to explain and approximate the observed dependence in the data by introducing latent classes. Thus, this model serves the purpose of approximating complex dependent J-variate binary distributions through the identification of patterns in the data that can explain the observed dependence in a more concise and interpretable manner than estimating the entire set of $J^2 - 1$ parameters. This method is also referred to as "the categorical data analog of factor analysis" (McCutcheon 1987). For more details on classical LCA we refer to Lazarsfeld (1955), McCutcheon (1987), and Andersen (1982), for Bayesian LDA to White and Murphy (2014) and Li et al. (2018), and for recent Bayesian nonparametric generalizations to Bartolucci et al. (2017), Koo and Kim (2020), Franzolini et al. (2023), and Qiu et al. (2023).

In this simulation, we generate data for $n = 250$ subjects and $p = 50$ positively correlated binary outcomes, with pairwise correlations ranging from 0.0871 to 0.5014. We fit a multivariate Bernoulli DPM, with a J-variate product of Beta $(0.2, 0.2)$ as base distribution and a Gamma $(1, 1)$ prior on the concentration parameter α, perform 20 000 MCMC simulations, and use the first 5 000 as burn-in.

Defining, as in the previous section, as sparsely populated clusters those clusters containing 10% or less of observations, we have that in 11 788 out 15 000 of the MCMC iterations, 10% or more of the observations are allocated into sparsely populated clusters, while in 13 815 out of 15 000 of MCMC iterations, 5% or more of the observations are allocated into sparsely populated clusters, see Fig. 7a. The
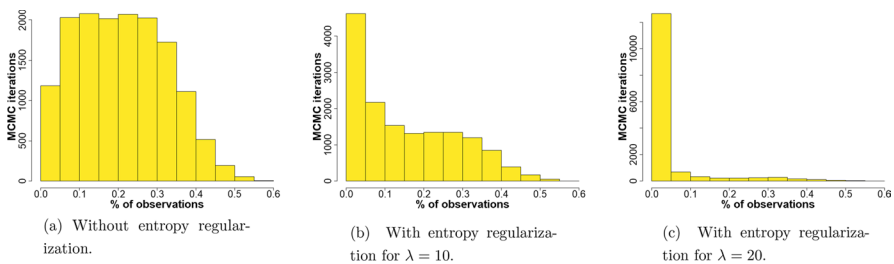

(a) Without entropy regularization.

(b) With entropy regularization for $\lambda = 10$.

(c) With entropy regularization for $\lambda = 20$.

Fig. 7 Multivariate Bernoulli simulation study. Percentage of observations in sparsely-populated clusters before and after entropy-regularization. Sparsely populated clusters are here defined as clusters containing 10% or less of observations. The horizontal axis denotes the percentage of observations (out of 1000) that are assigned to those clusters. The y-axis represents the number of MCMC samples (out of 15 000)
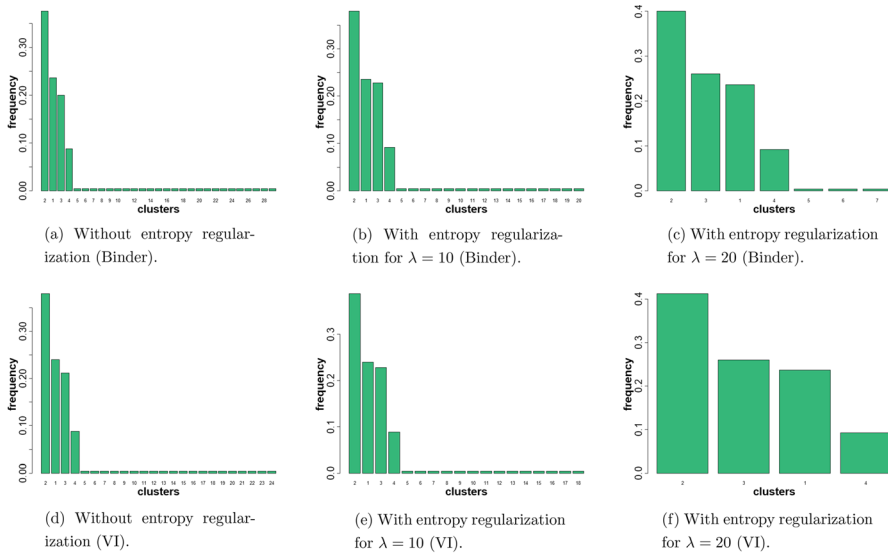
(a) Without entropy regularization (Binder).

(b) With entropy regularization for $\lambda = 10$ (Binder).

(c) With entropy regularization for $\lambda = 20$ (Binder).

(d) Without entropy regularization (VI).

(e) With entropy regularization for $\lambda = 10$ (VI).

(f) With entropy regularization for $\lambda = 20$ (VI).

Fig. 8 Multivariate Bernoulli simulation study. Estimated number of clusters and clusters' frequencies

same counts after entropy-regularization of the posterior are, with $\lambda = 10$, 8 203 and 10 373 out 15 000, see Fig. 7b, and, with $\lambda = 20$, 1 660 and 2 342 out 15 000, see Fig. 7c. Figure 8 shows the cluster frequencies for the three point-estimates obtained with both Binder and VI losses, without and with regularization.

5 Results for the wine dataset

We test the performance of our method also on the wine dataset available on R, where data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. Here we refer to the clustering identified by the three types of wines as "ground truth".

We use the 13 constituents to estimate a Dirichlet process mixture model with a multivariate Gaussian kernel, and we try to recover the three groups of types of wine through the estimated clustering. The concentration parameter is set to 0.1 to further favor a small number of clusters (see Sect. 1 for the results in the exact same setting but with a Gamma hyperprior for the concentration parameter). Data have been scaled before estimating the clustering configuration. After running the MCMC for 20 000 iterations and using the first 5000 as burnin, both the Binder loss and the VI functions identify a partition of five clusters, while our estimator for $\lambda = 50$ correctly identifies three clusters. Note that both the point estimates obtained with regularizing (with $\lambda = 50$) the Binder loss and the VI
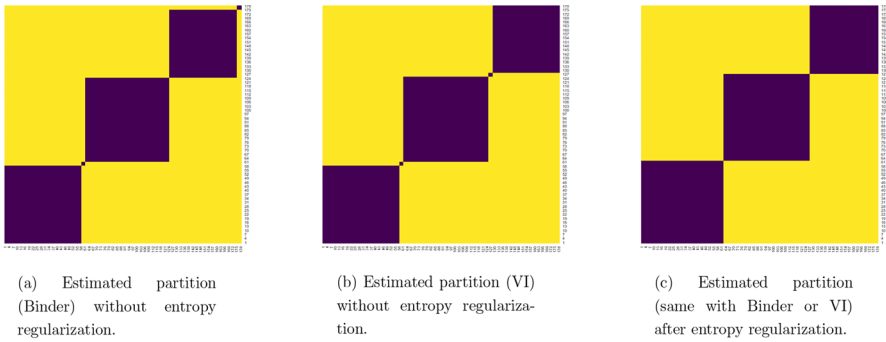
(a) Estimated partition (Binder) without entropy regularization.

(b) Estimated partition (VI) without entropy regularization.

(c) Estimated partition (same with Binder or VI) after entropy regularization.

Fig. 9 Estimated partitions for the wine dataset. Darker squares denote couples of observations clustered together, observations are ordered based on co-clustering
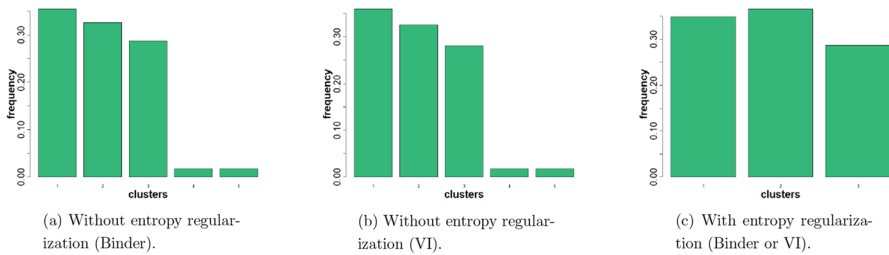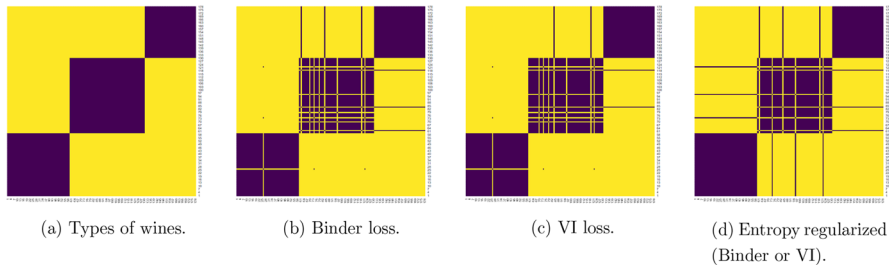

(a) Without entropy regularization (Binder).

(b) Without entropy regularization (VI).

(c) With entropy regularization (Binder or VI).

Fig. 10 Estimated number of clusters and clusters' frequencies for the wine dataset

(a) Types of wines.

(b) Binder loss.

(c) VI loss.

(d) Entropy regularized (Binder or VI).

Fig. 11 Estimated clustering for the wine dataset. Darker squares denote couples of observations clustered together, observations are ordered based on three groups of types of wine

loss are identical in this analysis (contrary to their not regularized estimates). See Figs. 9 and 10.

Lastly, Fig. 11 compares the clustering based on three groups of types of wine with the three estimates. The number of wrongly allocated wines, which equals 9 in the Binder loss point estimate and 8 in the VI point estimates, is reduced to 6 in the entropy regularized point estimate.

6 Conclusions

As highlighted in the recent literature, common posterior point estimates of the clustering obtained from Bayesian discrete mixture models suffer from unbalanced clusters' frequencies with only a few dominating clusters and a large number of sparsely-populated ones. In contrast, we introduced a general entropy-regularization of the existing losses that reduces the number of sparsely-populated clusters and enhances the interpretability of the Bayesian point estimate. Importantly, our proposal is theoretically justified and does not break the projectivity of the Bayesian model. We have further devised a simple and general computational scheme allowing for efficient computation of such entropy-regularized clustering estimate. This work paves the way for future intriguing research directions that we plan to address in forthcoming works.

From a theoretical perspective, it is interesting to study the connection with a wise recent probabilistic clustering model that induces less unbalanced clusters via breaking Kolmogorov consistency (Lee et al. 2022). Indeed we note that the entropy penalization introduced in this work can be incorporated in the prior instead of being applied to the loss function, ultimately constituting a sparsity penalized random partition model. However, such a random partition model breaks Kolmogorov consistency, contrary to our coherent loss-based approach discussed in Sect. 3. The decision to use a Kolmogorov consistent or inconsistent model ultimately depends on the specific applied problem. One of the strengths of our loss-based approach lies in the fact that, regardless of the choice made in this regard by the analyst, the Kolmogorov consistency of the model would not be affected by adopting our technique. Nonetheless, this duality of our proposal allows bridging connection with other recent interesting non-Kolmogorov consistent random partition models that are built by modifying existing EPPFs. See, e.g., Dahl et al. (2017); Paganin et al. (2021); Zanella et al. (2016).

From modeling and applied perspectives, it is natural to move beyond the exchangeable case and extend our regularized loss-based estimator to perform probabilistic clustering for dependent random partition models that allow considering covariates (see e.g., Teh et al. 2006; Müller et al. 2011; Page et al. 2022) or time-dependent random partition models such as those proposed, for example, in Page et al. (2022); Franzolini et al. (2023). Finally, the general loss-penalization that we have introduced and Algorithm 1 seem an appropriate tool for performing joint probabilistic clustering of different entities like in separate exchangeable partition models that allow performing bi-clustering in matrix data (Lee et al. 2013; Lin et al. 2021) and nested random partition model that allows us to jointly cluster populations and observations such as in nested partial exchangeable partition models (Rodríguez et al. 2008; Zuanetti et al. 2018) and recent extensions.
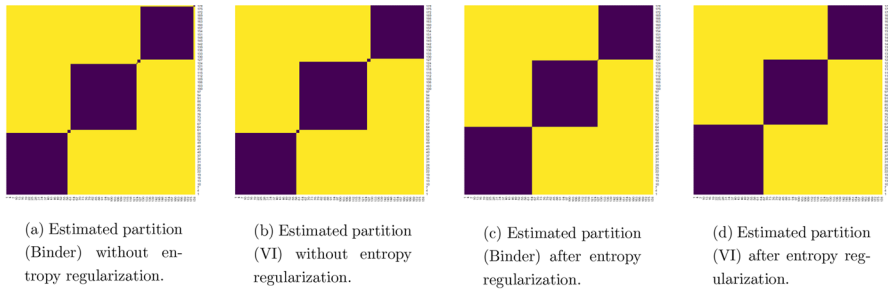
(a) Estimated partition (Binder) without entropy regularization.

(b) Estimated partition (VI) without entropy regularization.

(c) Estimated partition (Binder) after entropy regularization.

(d) Estimated partition (VI) after entropy regularization.

Fig. 12 Estimated (α random) partitions for the wine dataset. Darker squares denote couples of observations clustered together, observations are ordered based on co-clustering
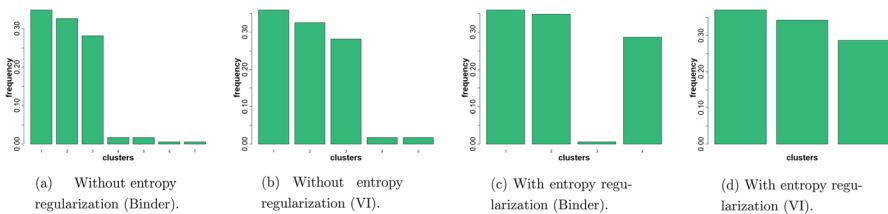

(a) Without entropy regularization (Binder).

(b) Without entropy regularization (VI).

(c) With entropy regularization (Binder).

(d) With entropy regularization (VI).

Fig. 13 Estimated (α random) number of clusters and clusters' frequencies for the wine dataset

Appendix: Prior on the concentration parameter

In Sect. 2 we report the partition cost associated with a DPM. It is clear that the choice of the concentration parameter α is relevant in controlling the number of clusters. Thus, in order to have a more flexible distribution on the clustering of the data, in many implementations of the Dirichlet process mixture a prior for α is specified, leading to a mixing measure that is itself a mixture in the sense of Antoniak (1974). Ascolani et al. (2023) also show that introducing such a prior can have a major impact on the asymptotic behavior of the number of clusters, as Dirichlet process mixtures can be consistent for the number of clusters. In these sections, we show the results obtained by repeating the analyses performed in the main paper where we use a Gamma prior (Escobar and West 1995) on the concentration parameter of the Dirichlet process. More precisely, we change the prior such that

$$\alpha \sim \text{Gamma}(1, 1)$$

and all the remaining prior specifications and MCMC settings (e.g., also the number of iterations) are set equal to the previous one. We show that also in such a case the proposed entropy-regularization is still crucial to enhance the interpretability of the clustering point estimate.
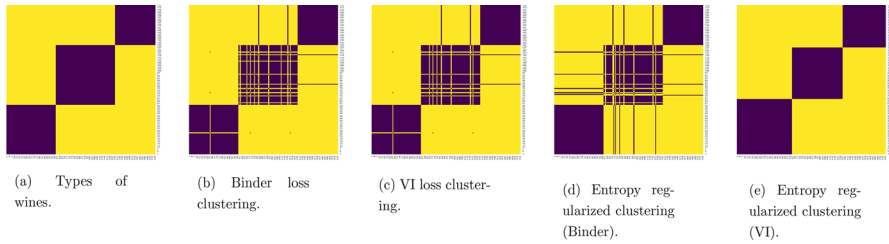
(a) Types of wines.

(b) Binder loss clustering.

(c) VI loss cluster-ing.

(d) Entropy reg-ularized clustering (Binder).

(e) Entropy reg-ularized clustering (VI).

Fig. 14 Estimated clustering for the wine dataset (α random). Darker squares denote couples of observations clustered together, observations are ordered based on three groups of types of wine

Wine dataset

We note that the point estimates obtained with regularizing (with $\lambda = 50$) the Binder loss and the VI loss are different in this analysis (contrary to the not regularized estimates with fixed concentration parameter). See Figs. 12 and 13 for the estimated partitions and related frequencies, respectively. Figure 14 compares the clustering based on three groups of types of wine with the four estimates. The number of wrongly allocated wines is equal to 10 in the Binder loss point estimate and 8 in the VI point estimates, while 8 and 10 with their regularized versions. Moreover, the number of clusters in the point estimate of the partition is 7 with the Binder loss and

(a) Without entropy regular-ization.

(b) With entropy regulariza-tion for $\lambda = 10$.

(c) With entropy regulariza-tion for $\lambda = 20$.
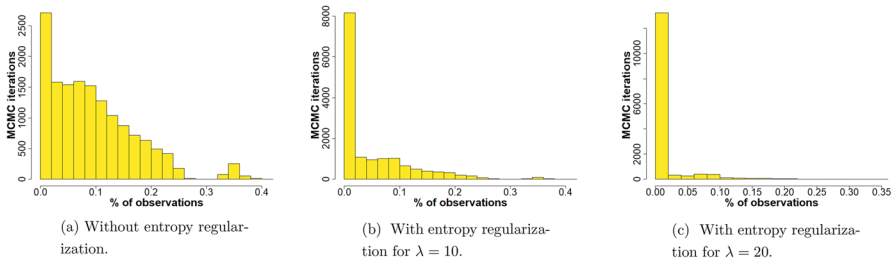
Fig. 15 Gaussian simulation study (α random). Percentage of observations in sparsely-populated clusters before and after entropy-regularization. Sparsely populated clusters are here defined as clusters containing 10% or less of observations
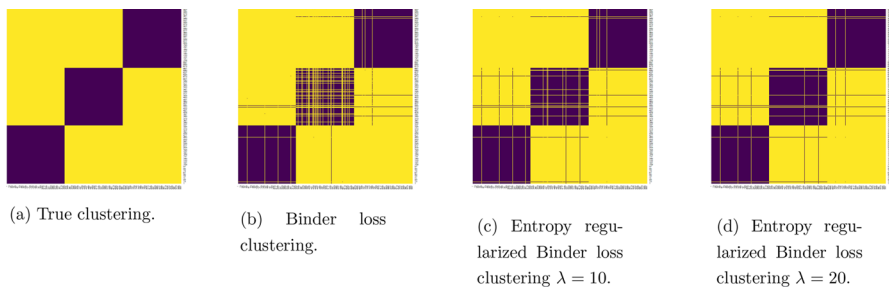

(a) True clustering.

(b) Binder loss clustering.

(c) Entropy regu-larized Binder loss clustering $\lambda = 10$.

(d) Entropy regu-larized Binder loss clustering $\lambda = 20$.

Fig. 16 Gaussian simulation study (α random). Estimated clustering for the simulation study darker squares denote couples of observations clustered together
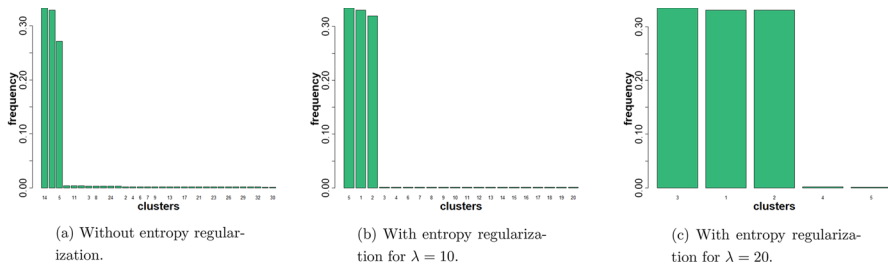
Fig. 17 Gaussian simulation study (α random). Estimated number of clusters and clusters' frequencies

5 with the VI loss, while 4 and 3 with their regularized versions (where the ground-truth is 3).

Gaussian simulation scenario

Recalling that we defined as sparsely populated clusters those clusters containing 10% or less of observations, we found that 6 040 out 15 000) of the MCMC iterations, 10% or more of the observations are allocated into sparsely populated clusters, while in 9 908 out of 15 000 of MCMC iterations, 5% or more of the observations are allocated into sparsely populated clusters, see Fig. 15a. The same counts after entropy-regularization of the posterior are, with $\lambda = 10$, 2772 and 5269 out 15 000, see Fig. 15b, and, with $\lambda = 20$, 417 and 1302 out 15 000, see Fig. 15c. Figure 16 shows the ground truth and the estimated clusters with and without entropy regularization. The number of observations misclassified (with respect to the simulation truth) with the Binder loss point estimate ($\lambda = 0$) is 80, with the regularization with $\lambda = 10$ is 37 and with $\lambda = 20$ is 25, showing an improvement also due to regularization also with this model. Figure 17 shows the cluster frequencies for the three point-estimates. Finally, we note that also with α random in this simple univariate Gaussian kernel simulation scenario we can obtain the correct number of occupied components (i.e., clusters) and thus avoid sparsely populated clusters in the point estimate using the more parsimonious VI loss (with the default parameter $a = 1$ as implemented in SALSO) instead of considering the Binder loss (with the default parameter $a = 1$).

References

Andersen EB (1982) Latent structure analysis: a survey. Scand J Stat 9:1–12

Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat 2:1152–1174

Argiento R, De Iorio M (2022) Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. Ann Stat 50:2641–2663

Ascolani F, Lijoi A, Rebaudo G, Zanella G (2023) Clustering consistency with Dirichlet process mixtures. Biometrika 110:551–558

Balocchi C, George EI, Jensen ST (2023) Clustering areal units at multiple levels of resolution to model crime incidence in Philadelphia. Preprint at arXiv 2112:02059

Bartolucci F, Farcomeni A, Scaccia L (2017) A nonparametric multidimensional latent class IRT model in a Bayesian framework. Psychometrika 82:952–978

Beraha M, Guglielmi A, Quintana FA (2021) The semi-hierarchical Dirichlet process and its application to clustering homogeneous distributions. Bayesian Anal 16:1187–1219

Beraha M, Argiento R, Möller J, Guglielmi A (2022) MCMC computations for Bayesian mixture models using repulsive point processes. J Comput Graph Stat 31:422–435

Betancourt B, Zanella G, Steorts RC (2022) Random partition models for microclustering tasks. J Am Stat Assoc 117:1215–1227

Bianchini I, Guglielmi A, Quintana FA (2020) Determinantal point process mixtures via spectral density approach. Bayesian Anal 15:187–214

Binder DA (1978) Bayesian cluster analysis. Biometrika 65:31–38

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Camerlenghi F, Lijoi A, Prünster I (2018) Bayesian nonparametric inference beyond the Gibbs-type framework. Scand J Stat 45:1062–1091

Camerlenghi F, Dunson DB, Lijoi A, Prünster I, Rodríguez A (2019) Latent nested nonparametric priors (with discussion). Bayesian Anal 14:1303–1356

Casella G, Moreno E, Girón FJ (2014) Cluster analysis, model selection, and prior distributions on models. Bayesian Anal 9:613–658

Dahl DB, Day R, Tsai JW (2017) Random partition distribution indexed by pairwise information. J Am Stat Assoc 112:721–732

Dahl DB, Johnson DJ, Müller P (2022a) Salso: search algorithms and loss functions for Bayesian clustering. R package version 0.3.29

Dahl DB, Johnson DJ, Müller P (2022b) Search algorithms and loss functions for Bayesian clustering. J Comput Graph Stat 31:1189–1201

De Blasi P, Favaro S, Lijoi A, Mena RH, Prünster I, Ruggiero M (2015) Are Gibbs-type priors the most natural generalization of the Dirichlet process? IEEE Trans Pattern Anal Mach Intell 37:212–229

Denti F, Camerlenghi F, Guindani M, Mira A (2023) A common atom model for the Bayesian nonparametric analysis of nested data. J Am Stat Assoc 118:405–416

Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90:577–588

Ferguson TS (1983) Bayesian density estimation by mixtures of normal distributions. In Recent Advances in Statistics, pages 287–302. Elsevier

Franzolini B, Cremaschi A, van den Boom W, De Iorio M (2023) Bayesian clustering of multiple zero-inflated outcomes. Philos Trans R Soc A 381:1–16

Franzolini B, De Iorio M, Eriksson J (2023) Conditional partial exchangeability: a probabilistic framework for multi-view clustering. Preprint at arXiv 2307:01152

Franzolini B, Lijoi A, Prünster I (2023) Model selection for maternal hypertensive disorders with symmetric hierarchical Dirichlet processes. Ann Appl Stat 17:313–332

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian Data Anal. Chapman and Hall/CRC

Gil-Leyva MF, Mena RH, Nicoleris T (2020) Beta-Binomial stick-breaking non-parametric prior. Electron J Stat 14:1479–1507

Green PJ, Richardson S (2001) Modelling heterogeneity with and without the Dirichlet process. Scand J Stat 28:355–375

Greve J, Grün B, Malsiner-Walli G, Frühwirth-Schnatter S (2022) Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis. Aust N Z J Stat 64:205–229

Hennig C (2015) What are the true clusters? Pattern Recognit. Letters 64:53–62

Koo W, Kim H (2020) Bayesian nonparametric latent class model for longitudinal data. Stat Methods Med Res 29:3381–3395

Lazarsfeld PF (1955) Recent developments in latent structure analysis. Sociometry 18:391–403

Lee, C. J. and Sang, H. (2022). Why the rich get richer? On the balancedness of random partition models. In Int Conf Mach Learn, pages 12521–12541

Lee J, Müller P, Zhu Y, Ji Y (2013) A nonparametric Bayesian model for local clustering with application to proteomics. J Am Stat Assoc 108:775–788

Li Y, Lord-Bessen J, Shiyko M, Loeb R (2018) Bayesian latent class analysis tutorial. Multivar Behav Res 53:430–451

Lijoi A, Prünster I (2010) Models beyond the Dirichlet process. In: Hjort NL, Holmes C, Müller P, Walker SG (eds) Bayesian Nonparametrics. Cambridge Univ, Press

Lijoi A, Mena RH, Prünster I (2007) Controlling the reinforcement in Bayesian non-parametric mixture models. J. R. Stat Soc Ser B Stat Methodol 69:715–740

Lijoi A, Prünster I, Rebaudo G (2023) Flexible clustering via hidden hierarchical Dirichlet priors. Scand J Stat 50:213–234

Lin Q, Rebaudo G, Müller P (2021) Separate exchangeability as modeling principle in Bayesian nonparametrics. Preprint at arXiv 2112:07755

Lindsay BG (1995) Mixture models: theory, geometry, and applications. In NSF-CBMS Regional Conf. Ser Prob Stat 5:1–165

Liu JS (1996) Metropolized independent sampling with comparisons to rejection sampling and importance sampling. Stat Comput 6:113–119

Lo AY (1984) On a class of Bayesian nonparametric estimates: I. density estimates. Ann Stat 12:351–357

McCutcheon AL (1987) Latent Class Analysis. SAGE PublicationS

McLachlan GJ, Lee SX, Rathnayake SI (2019) Finite mixture models. Annu Rev Stat Appl 6:355–378

Meilă M (2007) Comparing clusterings-an information based distance. J Multivar Anal 98:873–895

Miller JW, Harrison MT (2018) Mixture models with a prior on the number of components. J Am Stat Assoc 113:340–356

Müller P, Quintana FA, Rosner GL (2011) A product partition model with regression on covariates. J Comput Graph Stat 20:260–278

Ngan HYT, Yung NHC, Yeh AGO (2015) Outlier detection in traffic data based on the Dirichlet process mixture model. IET Intell Transp Syst 9:773–781

Nobile, A. (1994). Bayesian Analysis of Finite Mixture Distributions. Ph.D. thesis, Carnegie Mellon Univ

Paganin S, Herring AH, Olshan AF, Dunson DB (2021) Centered partition processes: informative priors for clustering (with discussion). Bayesian Anal 16:301–370

Page GL, Quintana FA, Müller P (2022) Clustering and prediction with variable dimension covariates. J Comput Graph Stat 31:466–476

Page GL, Quintana FA, Dahl DB (2022) Dependent modeling of temporal sequences of random partitions. J Comput Graph Stat 31:614–627

Petralia F, Rao V, Dunson DB (2012) Repulsive mixtures. In Adv Neural Inf Process Syst 25:1889–1897

Petrone S, Guindani M, Gelfand AE (2009) Hybrid Dirichlet mixture models for functional data. J. R. Stat Soc Ser B Stat Methodol 71:755–782

Pitman J (1996) Some developments of the Blackwell-MacQueen urn scheme. Lect Notes-Monogr Ser 30:245–267

Pitman J, Yor M (1997) The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann Probab 25:855–900

Polyanskiy Y, Wu Y (2020) Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. Preprint at arXiv 2008:08244

Qiu M, Paganin S, Ohn I, Lin L (2023) Bayesian nonparametric latent class analysis for different item types. Multivar Behav Res 58:156–157

Rastelli R, Friel N (2018) Optimal Bayesian estimators for latent variable cluster models. Stat Comput 28:1169–1186

Rebaudo G, Müller P (2023) Graph-aligned random partition model (GARP). Preprint at arXiv 2306:08485

Robert CP (2007) The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation, vol 2. Springer, Berlin

Rodríguez A, Dunson DB, Gelfand AE (2008) The nested Dirichlet process (with discussion). J Am Stat Assoc 103:1131–1154

Saha S, Guntuboyina A (2020) On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. Ann Stat 48:738–762

Savage LJ (1972) The Foundations of Statistics. Dover Publications, New York

Shotwell MS, Slate EH (2011) Bayesian outlier detection with Dirichlet process mixtures. Bayesian Anal 6:665–690

Stephens M (2000) Dealing with label switching in mixture models. J. R. Stat Soc Ser B Stat Methodol 62:795–809

Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. J Am Stat Assoc 101:1566–1581

Wade S, Ghahramani Z (2018) Bayesian cluster analysis: point estimation and credible balls (with discussion). Bayesian Anal 13:559–626

Wallach H, Jensen S, Dicker L, Heller K (2010) An alternative prior process for nonparametric Bayesian clustering. In Proc. Int Conf Artif Intell Stat 9:892–899

White A, Murphy TB (2014) BayesLCA: an R package for Bayesian latent class analysis. J Stat Softw 61:1–28

Xie F, Xu Y (2020) Bayesian repulsive Gaussian mixture model. J Am Stat Assoc 115:187–203

Xu Y, Müller P, Telesca D (2016) Bayesian inference for latent biologic structure with determinantal point processes (DPP). Biometrics 72:955–964

Zanella, G., Betancourt, B., Wallach, H., Miller, J., Zaidi, A., and Steorts, R. C. (2016). Flexible models for microclustering with application to entity resolution. In Advanced Neural Information and Process Syst, pages 1417–1425

Zhang, C., Qin, Y., Zhu, X., Zhang, J., and Zhang, S. (2006). Clustering-based missing value imputation for data preprocessing. In IEEE Int. Conf Industr Inform, pages 1081–1086

Zuanetti DA, Müller P, Zhu Y, Yang S, Ji Y (2018) Clustering distributions with the marginalized nested Dirichlet process. Biometrics 74:584–594