A regularized-entropy estimator to enhance cluster interpretability in Bayesian nonparametric

A regularized-entropy estimator to enhance cluster interpretability in Bayesian nonparametric

Uno stimatore a entropia regolarizzata per migliorare l'interpretabilità dei cluster in Bayesiana nonparametrica

Beatrice Franzolini, Giovanni Rebaudo

Abstract Bayesian nonparametric mixture models are widely used to cluster observations. However, one of the major drawbacks of the approach is that the estimated partition often presents only few dominating clusters and a large number of sparsely-populated ones. This feature translates into results that are uninterpretable unless we accept to ignore a relevant number of observations and clusters. Here, we provide an explanation of this phenomenon through the study of the cost functions involved in the estimation of the partition. Moreover, we propose a post-processing procedure to reduce the number of sparsely-populated clusters. The procedure takes the form of entropy-regularization of posterior cluster allocations. 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 Bayesian model used.

Abstract *I modelli Bayesiani nonparametrici con misture sono ampiemente utilizzati per effettuare cluster analysis. Tuttavia, uno dei principali limiti è il fatto che spesso identifichino un ampio numero di cluster poco popolati. Questa caratteristica si traduce in risultati di difficile interpretazione a meno che non si accetti di ignorare un numero di osservazioni e cluster. In questo lavoro, presentiamo una spiegazione di questo fenomeno attraverso lo studio delle funzioni di costo coinvolte nella stima della partizione. Inoltre, proponiamo una procedura di post-processing volta a ridurre il numero di cluster scarsamente popolati. La procedura prende la forma di una regolarizzazione dell'entropia della allocazione in cluster. La proposta appare computazionalmente conveniente rispetto a strategie alternative e trova giusticazione teorica in quanto correzione della funzione di perdita Bayesiana impiegata nella stima puntuale, e, proprio per questa ragione, può essere adottata a prescindere dallo specifico modello utilizzato.*

Key words: Bayesian nonparametrics, Exchangeable partition probability function, Entropy, Clustering, Dirichlet process mixture

---

Beatrice Franzolini
Agency for Science, Technology and Research, Singapore e-mail: franzolini@pm.me
Giovanni Rebaudo
Department of Statistics and Data Sciences, University of Texas at Austin, USA e-mail: rebaudo.giovanni@gmail.com

1 Introduction

Clustering methods detect patterns assigning observations to different clusters, so that (accordingly to a certain definition of similarity) observations are more similar within the same cluster than across clusters. Clustering has been proved 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), outliers-detection (Shotwell and Slate, 2011; Ngan et al., 2015; Franzolini et al., 2022), and data pre-processing (Zhang et al., 2006). Among clustering techniques, we can distinguish two main classes: model-based and non model-based. Contrary to other popular clustering techniques, like k-means, model-based clustering is based on the assumption that the data $(y_1, \dots y_n)$ are generated by a mixture model

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

where the mixture components $k(\cdot; \theta_h)$ are probability kernels to be interpret as distributions of distinct clusters, $(w_h, \theta_h)_{h=1}^K$ are unknown parameters that determine the relative proportion and the shape of the clusters in the whole population, and K is the total number of clusters in the population. K can be either a fixed value or an unknown parameter. However, the main goal of clustering techniques is to estimate a partition of the observed sample, more than the distribution in (1) of the whole population. The partition that one wants to estimate can be encoded using a sequence of subject-specific labels $(c_1, \dots, c_n)$ 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 $i = 1, \dots, n$. The indicators $(c_1, \dots, c_n)$, as just defined, are affected by the label switching problem (see, for instance, Stephens, 2000; McLachlan et al., 2019; Gil-Leyva et al., 2020). To overcome the issue, in the following, we assume them to be encoded in order of appearance. The likelihood for $\boldsymbol{c} = (c_1, \dots, c_n)$ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{K_n})$ is

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

An important and typically unknown parameter is the number of clusters K_n observed in the sample. Obviously, $K_n \leq K$. For this reason, finite fixed values for K are usually to be avoided and K is either fixed to $+\infty$ (e.g. in Dirichlet process mixtures, see Ferguson, 1983; Lo, 1984) or it is estimated from the data (e.g. mixtures of finite mixtures, see Miller and Harrison, 2018; Argiento and De Iorio, 2019).

When K_n is unknown, the clustering labels in (2) cannot be estimated with a fully frequentist approach. In fact, if the maximum likelihood estimator (MLE) for (2) exists and is unique, it identifies a number of clusters equal to the number of dis-

tinct observed values, so that actually no information on clusters can ever be gained through MLE and overfitting is unavoidable. In this regards, notice 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, due to the fact that priors act as penalties shrinking the number of distinct clusters.

The structure of the paper is as follow. Section 2 presents the study of the cost functions involved in BNP clustering models and provide an explanation for the presence of sparsely populated clusters, typically observed in the posterior estimates of these models. Then, a computationally convenient and theoretically justified solution to reduce the number of sparsely populated clusters is presented in Section 3 and showcased on simulated data and real data, respectively in Section 4 and 5.

2 Implied costs functions in Bayesian nonparametric clustering

The vast majority of Bayesian models for clustering relies on a prior for $\boldsymbol{c}$ 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 $(\theta_1,\dots,\theta_{K_n})$. Therefore, the corresponding posterior distribution is

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

and the maximum a posteriori (MAP) estimates is obtained minimizing the cost function $-\log(p(K_n,\boldsymbol{c},\boldsymbol{\theta}\mid\boldsymbol{y}))$, i.e.

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

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

$$C_{\text{lik}}(K_n,\boldsymbol{c},\boldsymbol{\theta};\boldsymbol{y}) = -\sum_{c=1}^{K_n}\sum_{i:c_i=c}^{n}\log k(y_i;\theta_c) \tag{4}$$

corresponds to K_n at least equal to the number of distinct observed values, that, if data comes from a non-atomic distribution, means $K_n \overset{a.s.}{=} n$.

The remaining two costs are those defined by the prior of the model. A lot of attention in the literature have been posed on the choice of the EPPF and many alternatives are available (see, for example, Lijoi et al., 2007, 2010; De Blasi et al., 2013), while, except for few cases (Petralia et al., 2012; Quinlan et al., 2018, 2021), the role of the base cost appears partially overlooked within the Bayesian methodol-

ogy 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 $\theta_c \overset{iid}{\sim} P_0$. Firstly notice that, when atoms are i.i.d., the resulting base cost

$$C_{\text{base}}(K_n, \boldsymbol{\theta}) = -\sum_{c=1}^{K_n} \log P_0(d\theta_c) \tag{5}$$

does not depend on the frequencies $(n_1, \dots, n_{K_n})$ and the incremental base cost for a new cluster, $\min_\theta(C_{\text{base}}(K_n+1, \boldsymbol{\theta})) - \min_\theta(C_{\text{base}}(K_n, \boldsymbol{\theta})) = -\max_\theta \log(P_0(d\theta))$ is constant in K_n. For example, when P_0 is set to be a univariate normal distribution centered in 0 and with variance σ^2, we have

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

and the incremental base cost for a new cluster is

$$-\max_\theta \log(P_0(d\theta)) = \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\sigma^2$$

from which it is clear that higher values of σ^2 result in a smaller number of clusters (cfr., e.g. Gelman et al., 2014, p. 535). In practice, P_0 is usually set to be a continuous mixture distribution, where the mixed density is conjugate to the kernel k and guarantees computational convenience, while the mixing distribution, usually placed on the scale parameter of the mixed, provides further sparsity on cluster locations. See Zhang et al. (2012) for a comprehensive account of scale mixtures and sparsity.

Finally, let us comment also 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 mixtures (PYPM, Pitman and Yor, 1997) . With a DPM model, we have

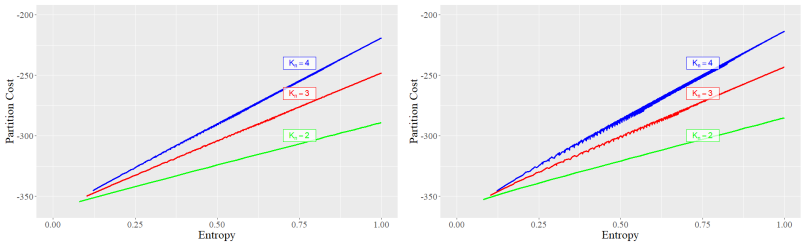
$$C_{\text{part}}(K_n, \boldsymbol{c}; \alpha) = -K_n \log\,\alpha - \sum_{c=1}^{K_n}\log\Gamma(n_c).$$

where α is the concentration parameter of the Dirichlet Process. The DPM partition cost tends to favor parsimonious values of K_n. However, contrary to the base cost, it depends also on the cluster allocations.

Figure 1 (a) showcases the partition cost of DPM for different values of entropy of the frequencies $(n_1, \dots, n_{K_n})$, i.e.

$$S(n_1, \dots, n_{K_n}) = -\sum_{c=1}^{K_n}\frac{n_c}{n}\log_{K_n}\frac{n_c}{n} \tag{6}$$

Overall the EPPF acts favoring frequencies $(n_1, \dots, n_{K_n})$ with low entropy. However, this feature can be interpreted as resulting into two distinct effects: one acting



(a) (b)

Fig. 1: Partition cost as function of the entropy in a DPM model with $\alpha = 1$ (panel a) and in a PYPM model with $\alpha = 1$ and $\sigma = 0.5$ (panel b). Values are computed for $n = 100$ observations clustered into 2 (blue line), 3 (red line), and 4 (greeen line) clusters.

on the total number of occupied clusters K_n and another acting on the skewness of the clusters' distribution $(n_1, \ldots, n_{K_n})$. Even though theoretically these two features both coincides with a reduced entropy in (6), they coincide with very different scenarios in terms of estimated clustering structure, especially from an applied and practical point of view. Penalizing the number of clusters is typically desirable, because an elevated number of clusters maybe difficult to interpret, however a partition with few dominating clusters and many sparsely populated clusters is highly undesirable, because it is hard to interpret, unless one decide to ignore all the information contained in the small clusters and focus only on the dominating ones.

In case of a PYPM the partition cost 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 the EPPFs are different, Figure 1 shows for both processes a closely similar behavior in terms of entropy penalization.

3 Regularized-entropy estimator

Once a posterior distribution $\mathbb{P}(\boldsymbol{c} \mid y_{1:n})$ over the possible partitions is obtained, typically thanks to a Markov Chain Monte Carlo algorithm, a point estimate $\hat{\boldsymbol{c}}$ of the partition is obtained minimizing the expected value of a loss function $L(\boldsymbol{c}, \hat{\boldsymbol{c}})$ with respect to the posterior, i.e.

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

Algorithm 1 Entropy-regularized estimates

Input: MCMC chain of partitions $\{\boldsymbol{c}_m, m = 1,\dots,M\}$, λ
Output: point estimate $\boldsymbol{c}^*$

1: Compute $S(\boldsymbol{c}_m)$ for $m = 1,\dots,M$
2: Compute $w_m = \exp\{\lambda S(\boldsymbol{c}_m)\}$ for $m = 1,\dots,M$
3: $\bar{w}_m \leftarrow w_m / \sum_m w_m$ for $m = 1,\dots,M$
4: Generate $\{\tilde{\boldsymbol{c}}_m, m = 1,\dots,M\}$, sampling with replacement from $\{\boldsymbol{c}_1,\dots,\boldsymbol{c}_M\}$ with prob. $\{\bar{w}_m, m = 1,\dots,M\}$
5: $\boldsymbol{c}^* \leftarrow \text{argmin} \sum_{m=1}^{M} \sum_{\hat{\boldsymbol{c}}} L(\tilde{\boldsymbol{c}}_m, \hat{\boldsymbol{c}})$

Even though in the previous section we focused our attention on the MAP estimator and, thus, adopt a 0-1 loss function, rarely in Bayesian clustering models the MAP estimator is employed due to the large support of the posterior and the fact that the 0-1 loss function does not take into account the distance between two partitions (see, Wade and Ghahramani, 2018). A widely used alternative is for instance the Binder loss.

We already stressed how a large presence of noisy clusters is undesirable in practice and our claim is 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 make this point even clearer: the idea is that wrongly estimating a low entropy partition with a high entropy partition is preferable wrt to wrongly estimating a high entropy partition with a low entropy partition.

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 wrt 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 clusters $K_n \geq 2$. Clearly, when λ is positive, for any candidate estimate $\hat{c}$, the loss function associates higher loss to partitions $\boldsymbol{c}$ with low 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 a entropy-regularized version $\bar{\mathbb{P}}[\boldsymbol{c} \mid y_{1:n}]$ of 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 results, while immediate to prove, is highly desirable, because it allows to implement the entropy-correction in a very straightforward and computationally feasible way which is described in Algorithm 1.

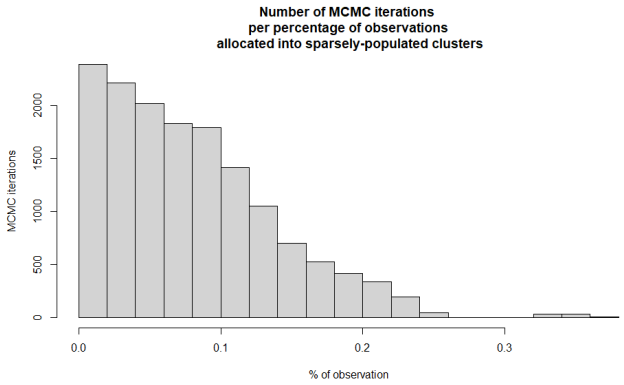
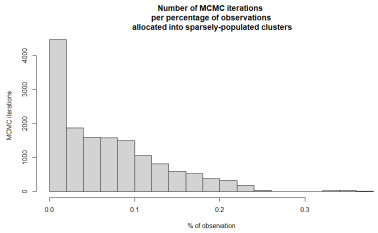
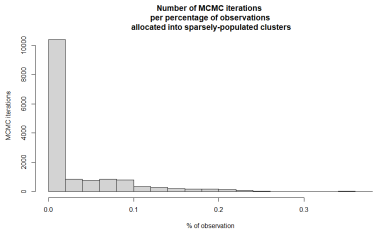


Fig. 2: Percentage of observations in sparsely-populated clusters before entropy-regularization



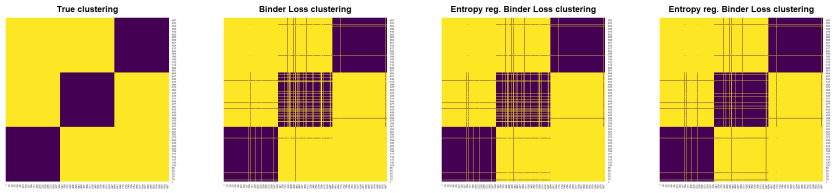
(a) Percentage of observations in sparsely-populated clusters after entropy-regularization with $\lambda = 10$



(b) Percentage of observations in sparsely-populated clusters after entropy-regularization with $\lambda = 20$

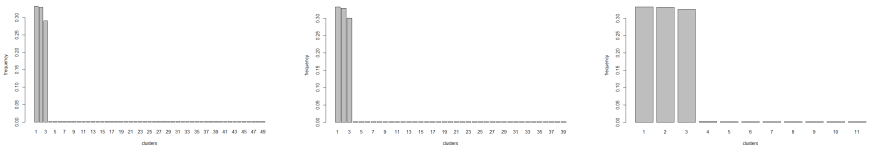
4 Simulation study

We provide here a simulation study, where $n = 1000$ observations are sampled from 3 univariate normally distributed clusters, we employ a normal-normal DPM and we compare the posterior estimates obtained minimizing the Binder loss function and the entropy-regularized Binder loss function. We set the concentration parameter $\alpha = 1$, we perform 20 000 MCMC simulations, and use the first 5000 as burnin. 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% of more of the observations are allocated into sparsely populated clusters, while in almost two third (9306 out of 15 000) of MCMC iterations, 5% of more of the observations are allocated into sparsely populated clusters, see Figure 2. The same counts after entropy-regularization of the posterior (as described in the previous section) are, with $\lambda = 10$, 3981 and 7825 out 15 000, see Figure 3a, and, with $\lambda = 20$, 1393 and 3366 out 15 000, see Figure 3b.



(a) (b) (c) (d)

Fig. 4: estimated clustering



(a) Clusters' estimated frequencies without entropy regularization

(b) Clusters' estimated frequencies with entropy regularization for $\lambda = 10$

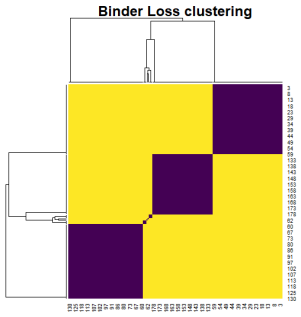
(c) Clusters' estimated frequencies with entropy regularization for $\lambda = 20$

Fig. 5: estimated clusters

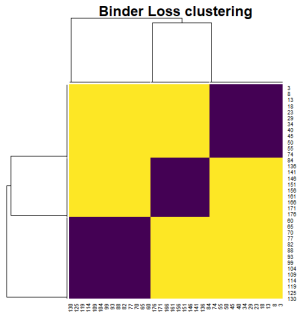
Finally, Figure 4 shows the true and the estimated clustering with and without entropy regularization. While Figure 5 shows the cluster frequencies for the three point estimates.

5 Results for the wine dataset

We test the performace of our estimator 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. We use the 13 constituents to estimate a Dirichlet process mixture model with multivariate Gaussian kernel, and we try to recover the three groups of types of wine through the estimated clustering. After running the MCMC for 10000 iterations and using the first 2000 as burnin, the Binder loss function identifies a partition of seven clusters, while our estimator for $\lambda = 20$ identifies three clusters. See Figure 6 and Figure 7.

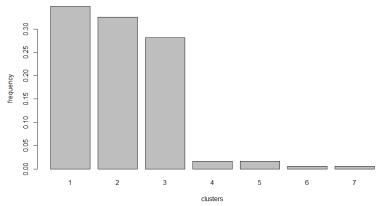


(a) Estimated partition without entropy-regularization

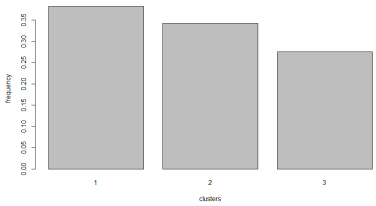


(b) Estimated partition after entropy-regularization

Fig. 6: Estimated partitions for the wine dataset, dark squares denote couples of observations clustered together, observations are ordered based on co-clustering.

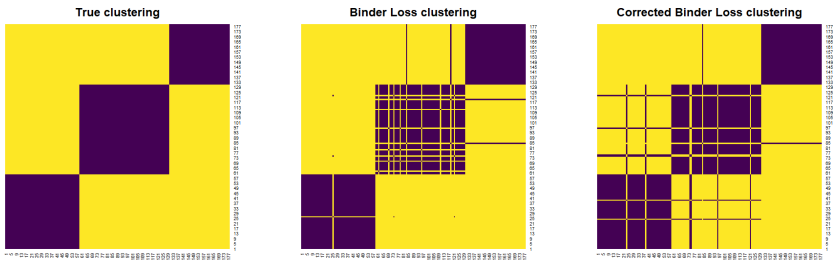


(a) Clusters' estimated frequencies without entropy-regularization



(b) Clusters' estimated frequencies after entropy-regularization

Fig. 7



(a) (b) (c)

Fig. 8: estimated clustering

References

Argiento, R. and M. De Iorio (2019). Is infinity that far? a bayesian nonparametric perspective of finite mixture models. *arXiv preprint arXiv:1904.09733*.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research 3*(Jan), 993–1022.

De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero (2013). Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE transactions on pattern analysis and machine intelligence 37*(2), 212–229.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics*, pp. 287–302. Elsevier.

Franzolini, B., A. Lijoi, and I. Prünster (2022). Model selection for maternal hypertensive disorders with symmetric hierarchical Dirichlet processes. *The Annals of Applied Statistics*, forthcoming.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2014). *Bayesian data analysis. vol. 2 CRC press*.

Gil-Leyva, M. F., R. H. Mena, and T. Nicoleris (2020). Beta-binomial stick-breaking non-parametric prior. *Electronic Journal of Statistics 14*(1), 1479–1507.

Lijoi, A., R. H. Mena, and I. Prünster (2007). Controlling the reinforcement in bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 715–740.

Lijoi, A., I. Prünster, et al. (2010). Models beyond the dirichlet process. *Bayesian nonparametrics 28*(80), 342.

Lindsay, B. G. (1995). Mixture models: theory, geometry, and applications. Ims.

Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, 351–357.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). Finite mixture models. *Annual review of statistics and its application 6*, 355–378.

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

Ngan, H. Y., N. H. Yung, and A. G. Yeh (2015). Outlier detection in traffic data based on the dirichlet process mixture model. *IET intelligent transport systems 9*(7), 773–781.

Petralia, F., V. Rao, and D. Dunson (2012). Repulsive mixtures. *Advances in neural information processing systems 25*.

Petrone, S., M. Guindani, and A. E. Gelfand (2009). Hybrid dirichlet mixture models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(4), 755–782.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. *Lecture Notes-Monograph Series*, 245–267.

Pitman, J. and M. Yor (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 855–900.

Polyanskiy, Y. and Y. Wu (2020). Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*.

Quinlan, J. J., G. L. Page, and F. A. Quintana (2018). Density regression using repulsive distributions. *Journal of Statistical Computation and Simulation 88*(15), 2931–2947.

Quinlan, J. J., F. A. Quintana, and G. L. Page (2021). On a class of repulsive mixture models. *TEST 30*(2), 445–461.

Saha, S. and A. Guntuboyina (2020). On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *The Annals of Statistics 48*(2), 738–762.

Shotwell, M. S. and E. H. Slate (2011). Bayesian outlier detection with dirichlet process mixtures. *Bayesian Analysis 6*(4), 665–690.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

Wade, S. and Z. Ghahramani (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis 13*(2), 559–626.

Zhang, C., Y. Qin, X. Zhu, J. Zhang, and S. Zhang (2006). Clustering-based missing value imputation for data preprocessing. In *2006 4th IEEE International Conference on Industrial Informatics*, pp. 1081–1086. IEEE.

Zhang, Z., S. Wang, D. Liu, M. I. Jordan, and N. Lawrence (2012). Ep-gig priors and applications in bayesian sparse learning. *Journal of Machine Learning Research 13*(6).