Bayesian Statistics

Bastien Billiot and Rémy Deshayes

January 2021

1 Introduction

Our project builds on the 2006 paper *A Predictive View of Bayesian Clustering* by Fernando A. Quintana.

In his paper, Quintana reviews various probabilistic methods for set partitioning, and their interactions. In other words, models given as conditional probabilities of either joining an already existing cluster or forming a new one.

Firstly, we recall the paper theoretical framework. Then, we review Quintana's implementations and results - notably, the various Model-Based Clustering (MBC) initializations.
Secondly, we move away from the paper per se and introduce two MBC implementations before comparing them with traditional clustering methods.

2 Framework and motivation

Quintana is interested in models described as a sequence of conditional probabilities of either joining an already existing cluster or forming a new one. We call cluster the elements of a partition of a finite set of objects $E = \{1, \dots, n\}$.
The author's theoretical motivation is to highlight the differences between parametric and non-parametric examples. On the numerical side, motivation is to contrast the introduced models.

Now, let's assume that $E_1, \dots, E_K$ are sorted in ascending order, which means that:

$$\min\{x : x \in E_1\} < \cdots < \min\{x : x \in E_K\}$$

We denote the cluster membership as follows: $x_i = j$ if $i \in E_j$, which by construction gives $x_1 = 1$. We add a «no gap»restriction, i.e $x_i = \alpha > 1$ for some $i \in E$ implies that there exist $i_1, \dots, i_{\alpha-1} \in E$ so that $s_{i_j} = j$ for $j \leq \alpha - 1$

This gives us the form of the problem we are interested in - namely models giving probability distributions on the space of partitions $\mathcal{P}$ of E, which amounts to finding the probability distribution of a vector $(x_1, \dots, x_n)$ - this is equivalent to (chain rule and $p(x_1 = 1) = 1$):

$$p(x_1, \dots, x_n) = \prod_{j=2}^{n} p(x_j | x_{j-1}, x_{j-2}, \dots, x_1) \tag{1}$$

With (1), we introduce a distribution on the vector of cluster memberships in terms of conditional probabilities $p(x_j | x_{j-1}, x_{j-2}, \dots, x_1)$ for $j = 2, \dots, n$.
Each successive choice on the conditional probabilities represents a probabilistic belief as to how the clusters are formed, indeed these choices characterize how the elements of E are successively added to an already existing cluster or become the first element of a new cluster.

The interesting question then lies in the model choice for those conditional probabilities.

Finally, let k_α be the number of clusters formed with the elements $1, \ldots, \alpha$ and $m_{1,\alpha}, \ldots, m_{k_\alpha,\alpha}$ be the corresponding cluster sizes.

2.1 Introducing key conditional probabilities models

2.1.1 Dirichlet Process Partitioning

The Dirichlet Process Partitioning is a non-parametric model. The conditional probability to join cluster i is:

$$P(x_{\alpha+1} = i | x_\alpha, \ldots, x_1) = \begin{cases} \frac{m_{i,\alpha}}{c+\alpha} \text{ si } 1 \geq i \geq k_\alpha \\ \frac{c}{c+\alpha} \text{ si } i = k_\alpha + 1 \end{cases}$$

with c a weight parameter in Ferguson's model.

2.1.2 Model-Based Clustering

Model Based Clustering is a mixture model, i.e a probabilistic model for representing the presence of sub-populations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs. This method was originally introduced by Banfield and Raftery (1993) and relies on independent variables $X_1, \ldots, X_n$ and the following mixture model:

$$p(X_1, \ldots, X_n | K, \theta) = \prod_{i=1}^{n} \sum_{j=1}^{K} \tau_{K,j} f_j(X_i | \theta_j)$$

where K is the number of components in the mixture, and $\tau_{K,j}$ the weights given to the j^{th} element - those weights are non-negative and $\sum_j \tau_{K,j} = 1$

The structure of the MBC implies that the probability of any cluster assignment is of multinomial type, so clusters are not constrained to be nonempty and no order relation exists between them.

Quintana alternatively introduces two other models that we are not reviewing here: the Species Sampling Model and the Product Partition Model.

3 A new criterion : the MBC initialization

In what follows, we especially focus on MBC. The idea is based off of the above-mentioned mixture model maximized by the EM algorithm (Dempster et al. - 1977).

The EM algorithm is used to estimate the model parameters by maximum likelihood. The MBC selects both the model and the number of mixture elements (clusters). The number of clusters is therefore established - various values are then tested through Schwarz's Bayesian Information Criterion (BIC). In other words, for each chosen value K, the log-likelihood maximizing partition is found. The partition and the corresponding number of final selected mixture components are those with the highest BIC criterion (among the K candidates and their respective optimal partition).

Now, we introduce the model implementation for the estimation of an univariate density function. Here we consider a Gaussian mixture model with specific mean and variance for each element. We have, for each given K:

$$p(y_1, \ldots, y_n) = l(\{y_i\}, \{\tau_{K,j}\}, \{\mu_j\}, \{\sigma_j^2\}) = \prod_{i=1}^{n} \sum_{j=1}^{K} \tau_{K,j} \frac{1}{\sigma_j} \phi\left(\frac{y_i - \mu_j}{\sigma_j}\right)$$

The expectation step of the EM algorithm amounts to finding the estimated posterior probability that the i^{th} observation joins the j^{th} cluster. Expected likelihood is then maximized. Both steps are reiterated until convergence, yielding a partition for each K.

We now discuss Quintana's results on the *Galaxy dataset* - 82 galaxy velocities. Quintana compares the MBC obtained through the above-mentioned steps with a Dirichlet process based clustering with associated clusters number and size varying according to 4 parameters (not reviewed in this project). However, we notice that the author does not modify the initial partition of the MBC used in the EM algorithm. Varying this initialization to observe possible changes in the clusters (and the number of clusters yielded) is definitely interesting. These variations could lead to MBC models with potentially lower associated BIC but more meaningful clusters depending on the problem (an additional cluster or an aggregation of two clusters can highlight behaviors differences or similarities).

3.1 Initializations

We will now study the impact of varying the initialization. We introduce 2 different initialization: a *sub-sampling* method - we use a sub-sample as the initial clustering - and an initialization based on a hierarchical model-based clustering. We compare our results to Quintana's ones.

[Table 1] *Galaxy dataset* partitioning according to various initializations

Initialization	**K**	**BIC**	**Partitioning**
Quintana	3	~ -441	{1 - 7}, {8 - 79}, { 80 - 82 }
sub-sampling	3	-441.6122	{1 - 7}, {8 - 79}, { 80 - 82 }
hierarchical	5	-446.36	{1 - 7}, {8 - 9}, { 10 - 44 }, { 45 - 77 }, { 78 - 82 }

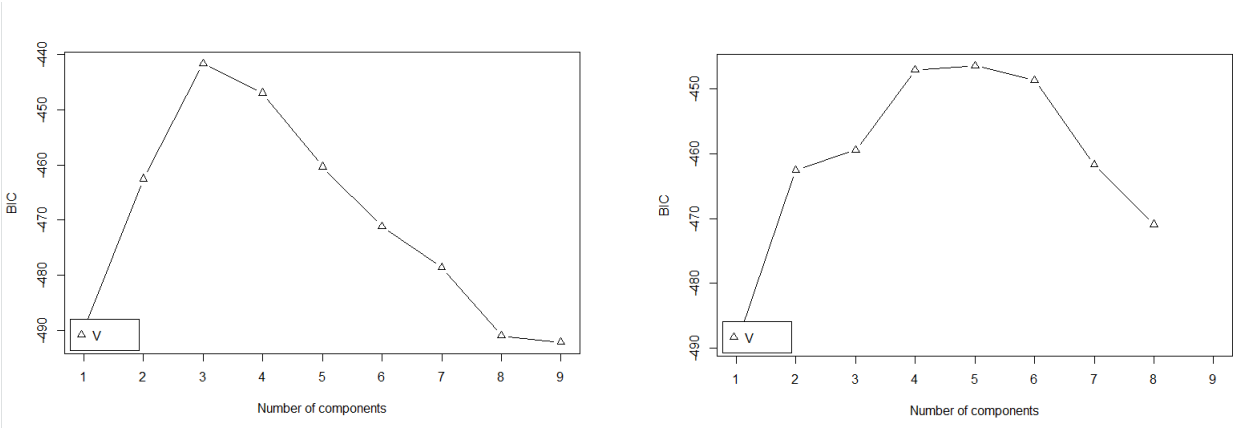

Figure 1: BIC values for the MBC : sub-sampling (LHS) and hierarchical (RHS)

We notice that the MBC reacts to the partition initialization. In the Galaxy dataset example (Table 1), even if the 5-cluster partition is consistent with the data histogram, it is an unrealistic approach to the density estimation problem. In contrast, the 3- and 4-cluster models allow estimation at a different level of granularity. As shown in Figure 1, the 3-cluster estimation allows the central cluster to be well differentiated from the two smaller side clusters.
An initialization of the MBC leading to 4 clusters allows to study the differences between the 2 central-cluster sub-clusters even though the model is not optimal according to the BIC criterion (we go beyond density estimation here, we rather talk about cluster characteristics).

4 Comparing traditional clustering methods and the MBC

In this section we use the *Italian Wine dataset* from the PGMM R package. Dataset contains 178 observations for 3 wine types (Barolo, Grignolilo and Barbera) and 27 corresponding variables. The clustering task is to form clusters corresponding to the 3 wine types.
We review the performance of the hierarchical clustering, which allows to create a cluster hierarchy without

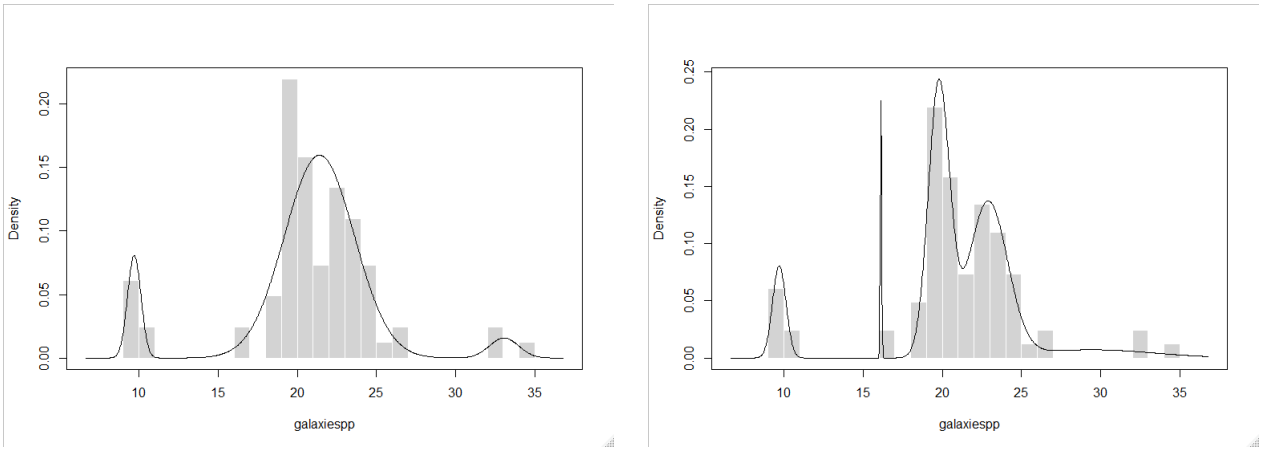

Figure 2: Densities for the MBC : sub-sampling (LHS) and hierarchical (RHS)

specifying the initial number of clusters. This method uses the distance between observations as a measure. On the traditional side, we review the k-means associated results. k-means uses the number of clusters as input and the partition yielding is such that each observation is assigned to the cluster that has the closest mean.
For this multivariate clustering problem , the MBC chosen is an EVI (diagonal, equal volume, varying shape) version of the *EM* algorithm - we review 14 shapes of the *EM* algorithm, we compute the BIC criterion values for a number of clusters between 1 and 7 for each of these shape. The selected model is the one with shape and number of clusters maximizing the BIC criterion.
Models evaluation can be seen in Figure 3, performance comparison with the models introduced for this task are shown in Table 2.

[Table 2]

Model	**% classification error for K = 3**	**computational time**
MBC (EVI)	5.61 %	32.5 seconds
k-means	25 %	0.007 seconds
hierarchical	60 %	0.005 seconds

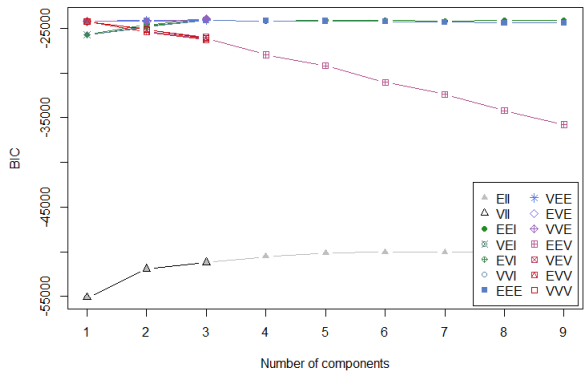

Figure 3: BIC values for the *Italian Wine dataset* clustering - MBC model selection

We notice that the MBC outperforms the other methods by a significant margin - results of the hierarchical clustering are only very weak. In our simulation example, although a change of initialization can lead to a

different optimal models, the advantages of the MBC are numerous. Indeed, compared to the hierarchical model, it is more efficient and in comparison with the k-mean whose performances are closer, it does not require to input a pre-determined number of clusters. On the other hand, the computational cost is the main drawback of the MBC.

5 Conclusion

Fernando Quintana's paper theoretically and numerically compare different models specified through the sequence of conditional probabilities of joining an existing cluster or forming a new one. We have discussed the initialization parameter for one of those models: the MBC. Studying its performances shows that this method can be more efficient than some traditional clustering models. However, the model selection and associated cluster number must always be evaluated in accordance to the specific challenges of the involved task. Finally, to deepen our comparison between traditional methods and MBC methods, we could review Model-Based Classification techniques, in particular Mixture Discriminant Analysis, which performances as a classifier can be compared with methods such as Linear or Quadratic Discriminant Analysis or more recently developed algorithms such as neural networks.