
# An MML Embedded Approach for Estimating the Number of Clusters

Cláudia Silvestre, Margarida G. M. S. Cardoso, and Mário Figueiredo

**Abstract** Assuming that the data originate from a finite mixture of multinomial distributions, we study the performance of an integrated *Expectation Maximization* (EM) algorithm considering *Minimum Message Length* (MML) criterion to select the number of mixture components. The referred EM-MML approach, rather than selecting one among a set of pre-estimated candidate models (which requires running EM several times), seamlessly integrates estimation and model selection in a single algorithm. Comparisons are provided with EM combined with well-known information criteria – e.g. the Bayesian information Criterion. We resort to synthetic data examples and a real application. The EM-MML computation time is a clear advantage of this method; also, the real data solution it provides is more parsimonious, which reduces the risk of model order overestimation and improves interpretability.

**Keywords:** finite mixture model, EM algorithm, model selection, minimum message length, categorical data

## 1 Introduction

Clustering is a technique commonly used in several research and application areas. Most of the clustering techniques are focused on numerical data. In fact, clustering

Cláudia Silvestre (✉)
Escola Superior de Comunicação Social, Campus de Benfica do IPL 1549-014 Lisboa, Portugal,
e-mail: csilvestre@escs.ipl.pt

Margarida G. M. S. Cardoso
BRU-UNIDE, ISCTE-IUL, Av. das Forças Armadas, 1649-026 Lisboa, Portugal,
e-mail: margarida.cardoso@iscte-iul.pt

Mário Figueiredo
Instituto de Telecomunicações, Portugal, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal,
e-mail: mario.figueiredo@tecnico.ulisboa.pt



P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_38

methods for categorical data are more challenging [12] and there are fewer techniques available [11].

In order to determine the number of clusters, model-based approaches commonly resort to information-based criteria e.g., the *Bayesian Information Criterion* (BIC) [15] or the *Akaike Information Criterion* (AIC) [1]. These criteria look for a balance between the model's fit to the data (which corresponds to maximizing the likelihood function) and parsimony (using penalties associated with measures of model complexity), thus trying to avoid over-fitting. The use of information criteria follows the estimation of candidate finite mixture models for which a predetermined number of clusters is indicated, generally resorting to an EM (*Expectation Maximization*) algorithm [7]. In this work, we focus on determining the number of clusters while clustering categorical data, using an EM embedded approach to estimate the number of clusters. This approach does not rely on selecting among a set of pre-estimated candidate models, but rather integrates estimation and model selection in a single algorithm. Our new implementation to deal with categorical variables by estimating a finite mixture of multinomials, follows a previous version described in [16]. We capitalized on the work of Figueiredo and Jain [9] for clustering continuous data and extended it for dealing with categorical data. The embedded method is thus based on a *Minimum Message Length* (MML) criterion to select the number of clusters and on an EM algorithm to estimate the model parameters.

## 2 Clustering with Finite Mixture Models

The literature on finite mixture models and their application is vast, including some books covering theory, geometry, and applications [8, 13, 3]. When applying finite mixture models to social sciences, the analyst is often confronted with the need to uncover sub-populations based on qualitative indicators.

### 2.1 Definitions and Concepts

Let $\mathbf{Y} = \{\underline{y}_i,\ i = 1, \ldots, n\}$ be a set of $n$ independent and identically distributed (i.i.d.) sample of observations of a random vector, $\underline{Y} = [Y_1, \ldots, Y_L]'$. We assume $\underline{Y}$ follows a mixture of $K$ components densities, $f(\underline{y}|\underline{\theta}_k)$ $(k = 1, \ldots, K)$, with probabilities $\{\alpha_1, \ldots, \alpha_K\}$, where $\underline{\theta}_k$ are the distributional parameters defining the $k$-th component and $\Theta = \{\underline{\theta}_1, \ldots, \underline{\theta}_K, \alpha_1, \ldots, \alpha_K\}$ the set of all the parameters of the model. The $\alpha$ values, also called *mixing probabilities*, are subject to the usual constraints: $\sum_{k=1}^{K} \alpha_k = 1$ and $\alpha_k \geq 0$, $k = 1, \ldots, K$. The log-likelihood of the observed set of sample observations is

$$\log f(\mathbf{Y}|\Theta) = \log \prod_{i=1}^{n} f(\underline{y}_i|\Theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k f(\underline{y}_i|\underline{\theta}_k). \tag{1}$$

In clustering, the identity of the component that generated each sample observation is unknown. The observed data $\mathbf{Y}$ is therefore regarded as incomplete, where the missing data is a set of indicator variables $\mathbf{Z} = \{\underline{z}_1, ..., \underline{z}_n\}$, each taking the form $\underline{z}_i = [z_{i1}, ..., z_{iK}]'$, where $z_{ik}$ is a binary indicator: $z_{ik}$ takes the value 1 if the observation $\underline{y}_i$ was generated by the k-th component, and 0 otherwise. It is usually assumed that the $\{\underline{z}_i,\ i = 1, \ldots, n\}$ are i.i.d., following a multinomial distribution of $K$ categories, with probabilities $\{\alpha_1, \ldots, \alpha_K\}$. The log-likelihood of complete data $\{\mathbf{Y}, \mathbf{Z}\}$ is given by

$$\log f(\mathbf{Y}, \mathbf{Z}|\Theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left[\alpha_k f(\underline{y}_i|\underline{\theta}_k)\right]. \tag{2}$$

## 2.2 Discrete Finite Mixture Models

Consider that each variable in $\underline{Y}$, $Y_l$ ($l = 1, \ldots, L$) can take one of $C_l$ categories. Conditionally on having been generated by the k-th component of the mixture, each $Y_l$ is thus modeled by a multinomial distribution with $n_l$ trials, $C_l$ categories, and non-negative parameters $\underline{\theta}_{kl} = \{\theta_{klc},\ c = 1, \ldots, C_l\}$, with $\sum_{c=1}^{C_l} \theta_{klc} = 1$. For a sample $y_{il} (i = 1, \ldots, n)$ of $Y_l$, we denote as $y_{ilc}$ the number of outcomes in category $c$, which is a sufficient statistic; naturally, $\sum_{c=1}^{C_l} y_{ilc} = n_l$. Thus, with $\underline{\theta}_k = \{\underline{\theta}_{k1}, \ldots, \underline{\theta}_{kL}\}$ and $\Theta = \{\underline{\theta}_1, \ldots, \underline{\theta}_K, \alpha_1, \ldots, \alpha_k\}$, the log-likelihood function, for a set of observations corresponding to a discrete finite mixture model (mixture of multinomials). This log-likelihood can be seen as corresponding to a missing-data problem, where the missing data has exactly the same meaning and structure as above. The log-likelihood of the complete data $\{\mathbf{Y}, \mathbf{Z}\}$ is thus given by

$$\log p(\mathbf{Y}, \mathbf{Z}|\Theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left( \alpha_k \prod_{l=1}^{L} \left[ n_l! \prod_{c=1}^{C_l} \frac{(\theta_{klc})^{y_{ilc}}}{y_{ilc}!} \right] \right). \tag{3}$$

To obtain a *maximum-likelihood* (ML) or *maximum a posteriori* (MAP) estimate of the parameters of a multinomial mixture, the well-known EM algorithm is usually the tool of choice [7].

# 3 Model Selection for Categorical Data

Model selection is an important problem in statistical analysis [6]. In model-based clustering, the term *model selection* usually refers to the problem of determining the number of clusters, although it may also refer to the problem of selecting the structure of the clusters. Model-based clustering provides a statistical framework to solve this problem usually resorting to *information criteria*. Among the best-known information criteria we find BIC and AIC, their modifications - namely the consistent

AIC, (CAIC) and the Modified AIC (MAIC) - and also the Integrated Completed Likelihood (ICL) [14, 4]. They are all easily implemented, the final model being selected according to a compromise between its fit to data and its complexity. In this work, we use the *Minimum Message Length* (MML) criterion to choose the number of components of a mixture of multinomials. MML is based on the information-theoretic view of estimation and model selection, according to which an adequate model is one that allows a short description of the observations. MML-type criteria evaluate statistical models according to their ability to compress a message containing the data, looking for a balance between choosing a simple model and one that describes the data well. According to Shannon's information theory, if $Y$ is some random variable with probability distribution $p(y|\Theta)$, the optimal code-length (in an expected value sense) for an outcome $y$ is $l(y|\Theta) = -\log_2 p(y|\Theta)$, measured in bits (from the base-2 logarithm). If $\Theta$ is unknown, the total code-length function has two parts: $l(y, \Theta) = l(y|\Theta) + l(\Theta)$; the first part encodes the outcome $y$, while the second part encodes the parameters of the model. The first part corresponds the fit of the model to the data (better fit corresponds to higher compression), while the second part represents the complexity of the model. The message length function for a mixture of distributions (as developed in [2]) is:

$$l(y, \Theta) = -\log p(\Theta) - \log p(y|\Theta) + \frac{1}{2}\log |I(\Theta)| + \frac{C}{2}\left(1 - \log(12)\right), \qquad (4)$$

where $p(\Theta)$ is a prior distribution over the parameters, $p(y|\Theta)$ is the likelihood function of mixture, $|I(\Theta)| \equiv |-E\left[\frac{\partial^2}{\partial\Theta^2}\log p(Y|\Theta)\right]|$ is the determinant of the expected Fisher information matrix, and $C$ is the the number of parameters of the model that need to be estimated. For example, for the $K$ mixture multinomial distributions presented in (3), $C = (K-1) + K\left(\sum_{l=1}^{L}(C_l - 1)\right)$. The expected Fisher information matrix of a mixture leads to a complex analytical form of MML which cannot be easily computed. To overcome this difficulty, Figueiredo and Jain [9] replace the expected Fisher information matrix by its complete-data counterpart $I_c(\Theta) \equiv -E\left[\frac{\partial^2}{\partial\theta^2}\log p(Y, Z|\Theta)\right]$. Also, they adopt independent Jeffreys' *priors* for the mixture parameters that is proportional to the square root of the determinant of the Fisher information matrix. The resulting message length function is

$$l(y, \Theta) = \frac{M}{2}\sum_{k:\,\alpha_k>0}\log\left(\frac{n\,\alpha_k}{12}\right) + \frac{k_{nz}}{2}\log\frac{n}{12} + \frac{k_{nz}(M+1)}{2} - \log p(y, \Theta) \qquad (5)$$

where $M$ is the number of parameters specifying each component (the dimension of each $\underline{\theta}_k$) and $k_{nz}$ the number of components with non zero probability (for more details on the derivation of (5), see [9, 2]).

## 4 The MML Based EM Algorithm

In order to estimate a mixture of multinomials, we use a variant of the EM algorithm (herein termed EM-MML), which integrates both estimation and model selection, by directly minimizing (5). The algorithm results from observing that (5) contains, in addition to the log-likelihood term, an explicit penalty on the number of components (the two terms proportional to $k_{nz}$), and a term (the first one) that can be seen as a log-prior on the $\alpha_k$ parameters of $\Theta$, that will directly affect the M-step.

**E-step:** The E-step of the EM-MML is precisely the same as in the case of ML or MAP estimation, since the generative model for the data is the same. Since we are dealing with a multinomial mixture, we simply have to plug the corresponding multinomial probability function yielding

$$\bar{z}_{ik}^{(t)} = \frac{\alpha_k \prod_{l=1}^{L} \left[ n_l! \prod_{c=1}^{C_l} \frac{(\widehat{\theta}_{klc}^{(t)})^{y_{ilc}}}{y_{ilc}!} \right]}{\sum_{j=1}^{K} \alpha_j \prod_{l=1}^{L} \left[ n_l! \prod_{c=1}^{C_l} \frac{(\widehat{\theta}_{jlc}^{(t)})^{y_{ilc}}}{y_{ilc}!} \right]}, \tag{6}$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, K$.

**M-step:** For the M-step, noticing that the first term in (5) can be seen as the negative log-prior $-\log p(\alpha_k) = \frac{C-K+1}{2K} \log \alpha_k$ (plus a constant), and enforcing the conditions that $\alpha_k \geq 0$, for $k = 1, ..., K$ and that $\sum_{k=1}^{K} \alpha_k = 1$, yields the following updates for the estimates of the $\alpha_k$ parameters:

$$\widehat{\alpha}_k^{(t+1)} = \frac{\max\left\{0, \sum_{i=1}^{n} \bar{z}_{ik}^{(t)} - \frac{C-K+1}{2K}\right\}}{\sum_{j=1}^{K} \max\left\{0, \sum_{i=1}^{n} \bar{z}_{ij}^{(t)} - \frac{C-K+1}{2K}\right\}}, \tag{7}$$

for $k = 1, ..., K$. Notice that, some $\widehat{\alpha}_k^{(t+1)}$ may be zero; in that case, the $k$-th component is excluded from the mixture model. The multinomial parameters corresponding to components with $\widehat{\alpha}_k^{(t+1)} = 0$ need not be further calculated, since these components do not contribute to the likelihood. For the components with non-zero probability, $\widehat{\alpha}_k^{(t+1)} > 0$, the estimates of multinomial parameters are updated to their standard weighted ML estimates:

$$\widehat{\theta}_{klc}^{(t+1)} = \frac{\sum_{i=1}^{n} \bar{z}_{ik}^{(t)} y_{ilc}}{n_l \sum_{i=1}^{n} \bar{z}_{ik}^{(t)}}, \tag{8}$$

for $k = 1, \ldots, K$, $l = 1, \ldots, L$, and $c = 1, \ldots, C_l$. Notice that, in accordance with the meaning of the $\theta_{klc}$ parameters, $\sum_{c=1}^{C_l} \widehat{\theta}_{klc}^{(t+1)} = 1$.

## 5 Data Analysis and Results

First, we evaluate the performance of the EM-MML algorithm on 10 synthetic data sets, over 50 runs. The data sets were originated from a mixture of 3 categorical variables (with 2, 3 and 4 levels) and 2 components. The correponding Sihouette index values illustrate the structures diversity: 0.099; 0.216; 0.217; 0.230; 0.713; 0.733; 0.746; 0.778; 0.805; 0.817. The obtained results are compared with those obtained from a standard EM algorithm combined with BIC, AIC, CAIC, MAIC, and ICL criteria.

The comparison resorts to a cohesion-separation measure and a concordance measure: the Fuzzy Silhouette index [5] of the clustering structure obtained and the Adjust Rand [10] between the same clustering structure and the original one. In Table 1 we can verify there are no significant differences between the EM-MML and the other criteria, except ICL which only recovers the very well separated structures. Regarding the number of clusters, EM-MML and MAIC are tied, recovering this number correctly for all data sets.The same is not true for the other criteria: AIC identifies 3 clusters in 3 data sets and 4 clusters once; in addition, BIC and CAIC could not find any cluster structure once and ICL was unable to do it for 4 data sets. In terms of computation time, since EM-MML does not require a sequential approach, it becomes clearly faster than the other criteria (Friedman test yields $\chi^2(5)$=2500 and p-value<0.01; Post hoc tests, with Bonferroni correction, only reveal statistically significant differences between the EM-MML and the other criteria).

**Table 1** Criteria performance.

| Criterion | Number of data sets | Fuzzy Silhouette: 95% CI Lower ; Upper Limits[a] | Adjusted Rand: 95% CI Lower ; Upper Limits[a] |
|---|---|---|---|
| AIC | 10 | 0.430 ; 0.741 | 0.545 ; 0.867 |
| BIC | 9 | 0.622 ; 0.935 | 0.728 ; 1.000 |
| CAIC | 9 | 0.616 ; 0.931 | 0.732 ; 1.000 |
| ICL | 6 | 0.917 ; 0.948 | 1.000 ; 1.000 |
| MAIC | 10 | 0.568 ; 0.887 | 0.623 ; 0.950 |
| **EM-MML** | **10** | **0.561 ; 0.891** | **0.594 ; 0.955** |

[a] 1000 bootstrap samples were used to estimate the Confidence Intervals (CI).

Additional insight into the performance of EM-MML is obtained by applying it to a real data set referring to the 6th European Working Conditions Survey (2015), Eurofound working conditions survey. Note that these data are the most recent.

For the purpose of our experiment, we consider the aggregate data referring to 305 European regions and the answers to the following questions: Are you able to

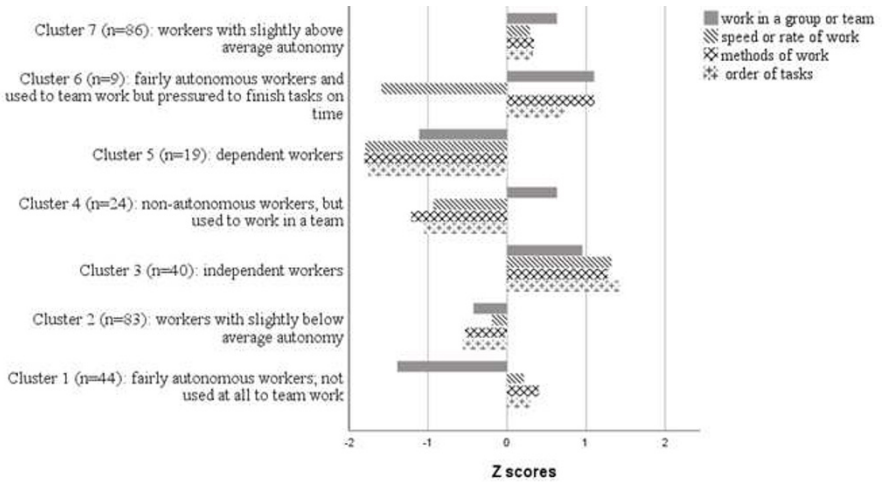


**Fig. 1** Clusters' profile and their dimensions ($n$).

choose or change: a) your order of tasks; b) your methods of work; c) your speed or rate of work. Do you work in a group or team that has common tasks and can plan its work?

EM-MML selected 7 clusters, which is a smaller number than for the remaining criteria (ICL, BIC, CAIC, AIC and MAIC select 10, 12, 12, 15 and 15 respectively). This fact avoids estimation problems associated with very small segments and also improves the interpretability of the clustering solution.

The segments selected by EM-MML criterion are presented in Figure 1. Workers with slightly above average autonomy (cluster 7) live in several countries, but Ireland stands out, as well as Belgium, Germany, Netherlands, Switzerland, and the UK regions. Denmark, Estonia, Malta, and Norway are the countries where the most independent workers are found (cluster 3). The smallest cluster, 6, includes Sweden and a region of Greece and Kriti and Açores, a Greek and a Portuguese region, respectively. The cluster 5, where workers claim they have no autonomy, includes regions from many countries.

## 6 Discussion and Perspectives

In this work, a model selection criterion and method for finite mixture models of categorical observations was studied - EM-MML. This algorithm simultaneously performs model estimation and selects the number of components/clusters. When compared to information criteria, which are commonly associated with the use of the EM algorithm, the EM-MML method exhibits several advantages: 1) it easily recovers the true number of clusters in synthetic data sets with various degrees of

separation; 2) its computations times are significantly lower than those required by standard approaches resorting to the sequential use of EM and an information criterion; 3) when applied to a real data set it produces a more parsimonious solution, thus easier to interpret. An additional advantage of this approach that stems from obtaining more parsimonious solutions is that such solutions have a higher number of observations per cluster, thus helping to overcome eventual estimation problems.

The performance of the EM-MML is encouraging for selecting the number of clusters, and the same criterion was already used for feature selection [17]. However, future research is required, namely considering data sets with different numbers of clusters and high dimensional data.

**Acknowledgements** This work was supported by Fundação para a Ciência e Tecnologia, grant UIDB /00315/2020.

## References

1. Akaike, H.: Maximum likelihood identification of Gaussian autorregressive moving average models. Biometrika. **60**, 255–265 (1973)
2. Baxter, R. A., Olivier, J. J.: Finding overlapping components with MML. Stat. Comput. **10**(1), 5–16 (2000)
3. Bouguila, N., Fan, W.: Mixture Models and Applications (Unsupervised and Semi-Supervised Learning). Springer Nature Switzerland AG, Switzerland (2020)
4. Bozdogan, H.: Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In: Bozdogan, H. (eds.) Proceedings of the First US/Japan Conf. Frontiers of Stat. Modeling, pp.69–113. Boston: Kluwer Academic Publishers (1994)
5. Campello, R. J., Hruschka, E. R.: A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Set. Syst., **157**(21), 2858–2875 (2006)
6. Celeux, G., Martin-Magniette, M. L., Maugis-Rabusseau, C., Raftery, A. E.: Comparing model selection and regularization approaches to variable selection in model-based clustering. J. Soc. Fr. Statistique. **155**(2), 57–71 (2014)
7. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood estimation from incomplete data via the EM Algorithm. J. R. Stat. Soc. **39**, 1–38 (1997)
8. Everitt, B. S., Hand, D.: Finite Mixture Distributions. Chapman and Hall, New York (1981)
9. Figueiredo, M. A. T., Jain, A. K.: Unsupervised learning of finite mixture models. IEEE T. Pattern Anal. **24**, 381–396 (2002)
10. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 93–218 (1985).
11. Kumar, P., Kanavalli, A.: A similarity based K-means clustering technique for categorical data in data mining application. Int. J. Intell. Eng. Syst. (2021) doi: 10.22266/ijies2021.0430.05
12. Lee, C., Jung, U.: Context-based geodesic dissimilarity measure for clustering categorical data. Appl. Sciences (2021) doi: 10.3390/app11188416
13. McLachlan, G. J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
14. Novais, L., Faria, S.: Selection of the number of components for finite mixtures of linear mixed models. J. Int. Math. **24**(8), 2237–2268 (2021)
15. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
16. Silvestre, C., Cardoso, M. G. M. S. and Figueiredo, M.: A clustering view on ESS measures of political interest: an EM-MML approach. NTTS - New Techniques and Technologies for Statistics (2017).
17. Silvestre, C., Cardoso, M. G. M. S. and Figueiredo, M.: Feature selection for clustering categorical data with an embedded modeling approach. Expert Syst. **32**(3), 444–453 (2014).