A clustering view on ESS measures of political interest: an EM-MML approach

Keywords: Official statistics, Categorical data, Clustering, EM-MML, European Social Survey, Finite Mixture Models.

1. INTRODUCTION

In this work, we perform the clustering of European regions, based on their citizens' political interests and electoral participation, as expressed in data from the European Social Survey (ESS). Clustering is applied to sets of questions referring to whether the citizens were involved in "different ways of trying to improve things in their country or help prevent things from going wrong" – e.g., signed a petition or worked in a political organisation or association.

We used data from the two most recent ESS surveys - 2012 (round 6) and 2014 (round 7) - referring to 20 countries and 240 regions (considered by both ESS surveys). The citizens' responses are aggregated by region and ESS sampling weights are taken into account.

We resort to a new clustering approach, named EM-MML [1], which clusters categorical data and simultaneously determines the number of clusters. This approach is particularly relevant when aggregated data is considered, as is often the case in official statistics. The approach assumes that the data originates from a finite mixture of multinomials and uses a *minimum message length* (MML) criterion to estimate the number of clusters [2]. We compare the results of the EM-MML approach with results from the classical EM approach combined with several information criteria. The comparisons address the number of clusters selected (parsimony), their cohesion-separation, and also their temporal stability.

2. THE EM-MML ALGORITHM

The *expectation-maximization* (EM) algorithm [3] (and its many variants) is commonly used to estimate the parameters of a finite mixture model, thus determining the distributional characteristics of each component or cluster. In general, to determine the number of clusters, a complementary analysis is conducted, resorting to information criteria such as BIC, AIC, AIC3, CAIC or ICL (see [4] for a comprehensive review).

For clustering categorical data and simultaneously determining the number of clusters, we use an embedded approach - the EM-MML algorithm - which is an EM variant that integrates estimation and model selection in a single procedure. The EM-MML algorithm relies on an MML criterion. It is initialized with a maximum number of clusters (mixture components) and automatically removes those deemed unnecessary.

From an information theory point of view, MML-type criteria choose the model that provides the shortest description of the observations and the parameters [5]. The rationale of this criterion is that a good model is one that achieves a short description of the underlying data.

3. CLUSTERING RESULTS

The clustering of European regions is based on the variables in Table 1. They all require "yes/no" answers, except for the first question, which admits an additional category: "not eligible to vote" (VotedNE_ne). EM-MML is used to obtain the clusters, by estimating a mixture of multinomials.

Table 1 – Clustering base variables

Voted last national election	VotedNE
Contacted politician or government official last 12 months	ContPOL
Worked in political party or action group last 12 months	WorkPOL
Worked in another organisation or association last 12 months	WorkASS
Worn or displayed campaign badge/sticker last 12 months	DispCAMP
Signed petition last 12 months	SignPET
Taken part in lawful public demonstration last 12 months	PartDEM
Boycotted certain products last 12 months	BoycotPROD
Feel closer to a particular party than all other parties	ClosePARTY

Table 2 – Evaluation of clustering solutions

	BIC; CAIC; ICL	AIC; AIC3	EM-MML
Number of clusters	7	7	2
Silhouette index	0.213	0.191	**0.361**
Calinski-Harabasz	83.327	74.977	**190.825**
Computation time (seconds)	109	109	**2**
Number of clusters	7	8	2
Silhouette index	0.152	0.164	**0.367**
Calinski-Harabasz	80.766	78.477	**189.552**
Computation time (seconds)	91	91	**2**
Adjusted Rand	0.377	0.499	**0.707**
Normalized mutual information	0.523	0.591	**0.598**

The number of segments selected by EM-MML is much lower than that of the alternative methods considered, with AIC and AIC3 being the least conservative criteria - Table 2. This increased parsimony avoids estimation problems associated with very small segments and also improves the interpretability of the clustering solution. In addition, it can favour better clustering results, improving the cohesion-separation and stability of the clusters. In fact, according to the *clustering validation indices* (CVI) Silhouette and Calinski-Harabasz (see [6] for an extensive CVI comparison), the results of EM-MML clearly surpass the others. The EM-MML clusters are also more stable according to comparisons between the 2012 and 2014 clusterings – see the results of the adjusted Rand index of (paired) agreement [7] and of the adjusted *normalized mutual information* [8]. Furthermore, the EM-MML (average) time of computation is a clear winner.

The two segments of European citizens yielded by EM-MML in 2012 are summarized in Figure 1. The results in 2014 are very similar - e.g., in the first segment of 2012 (2014) 33.63% (36.13%) "Feel closer to a particular party than all other parties"; in the second segment, these values are 55.74% in 2012, and 58.26% in 2014. The first segment includes 195 regions and the second one 285.

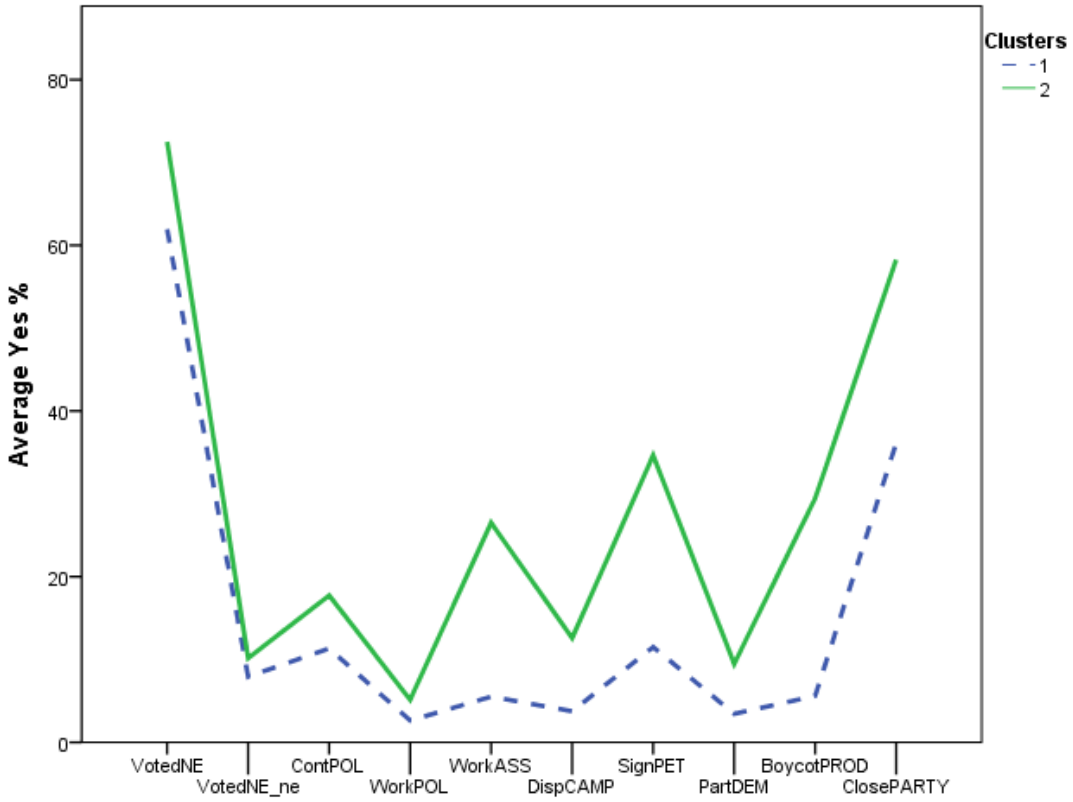

Figure 1 - The clusters' profiles

4. CONCLUSIONS

We propose using the EM-MML algorithm to cluster categorical aggregated data from the European Social Survey. This approach (simultaneously performing model estimation and identifying the number of clusters) provides more parsimonious and robust solutions than those obtained by standard EM combined with several information criteria. The EM-MML approach is also faster than the other methods considered, which is especially relevant when dealing with large data sets.

In future research we intend to explore the possibility of estimating finite mixtures of mixed variables – continuous and categorical – using an EM-MML variant.

REFERENCES

[1] C. Silvestre, M. Cardoso and M. Figueiredo, Clustering with finite mixture models and categorical variables, Preceedings of COMPSTAT2008 - 18-th nternational Conference on Computational Statistics, in Brito, Physics-Verlang (2008), 109-116.

[2] M. A. T. Figueiredo and A.K. Jain, Unsupervised learning of finite mixture models, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002), 381-396.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood for incomplete data via the EM algorithm, Journal of the Royal Statistical Society 39 (B), (1977), 1-38.

[4] J. R. Fonseca and M. G. Cardoso, Mixture-model cluster analysis using information theoretical criteria. Intelligent Data Analysis 11(2) (2007), 155-173.

[5] C. Wallace and D. Boulton, An Information Measure for Classification, Computer Journal 11 (1968), 195-209.

[6] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez and I. Perona, An extensive comparative study of cluster validity indices, Pattern Recognition 46**(**1), (2013), 243-256

[7] L. Hubert and P.Arabie, Comparing partitions, Journal of classification 2 (1), (1985), 243-256.

[8] A. Strehl and J. Ghosh, Cluster ensembles-a knowledge reuse framework for combining partitionings, AAAI/IAAI (2002), 93-99.