Proceedings of the World Congress on Engineering 2010 Vol III
WCE 2010, June 30 - July 2, 2010, London, U.K.

Assessing the Number of Clusters From a Mixture of Von Mises-Fisher

Wafia Parr Bouberima [1,2], Mohamed Nadif [1], Yamina Khemal Bencheikh [2] *

Abstract—**We consider the clustering problem of directional data and specifically the choice of the number of clusters. Setting this problem under the mixture approach, we perform a comparative study of different criteria. Monte Carlo simulations are performed taking into account the overlap degree of clusters and the size of data.**

Keywords: clustering, model selection, mixture models, information criteria, Von Mises-Fisher distribution

1 Introduction

Clustering is a key form of scientific research utilized within a variety of different scientific disciplines. Principally the classification method is used to produce g different clusters of wide distinctions. It should be noted that the optimum number of clusters g leading to the greatest separation is not known a priori and must be computed from the data, this is an heuristic problem in the classification topic; and this paper will be mainly concerned with this issue. In its main usage (Mainly), Clustering supports two approaches: a geometric one where the quality of the clustering depends on the chosen distance, and a probabilistic approach which is considered as a standard approach [13]. The latter covers the most widely used clustering methods. In this approach, data iare presumed to come from a sampled mixture of g components which are modeled by a distribution of probability. This approach can support several situations, depending on the parameters of the model, to obtain a best description of a heterogeneous population considering a selected model which is in itself another problem.

The clustering problem can be resolved by mixture modeling and we can, for this, consider two approaches: the Maximum Likelihood (ML) and the Classification Maximum Likelihood (CML) approaches. The former is based on the maximization of the Likelihood , and the latter one is based on the maximization of the Classification (or complete data) Likelihood. These maximizations can be performed respectively by the EM algorithm and by the Classification EM (CEM) [9]. The model selection problem is to find the most appropriate and concise model to express given data.

Here, we merely examine some criteria from a practical point of view and in the context of the directional data utilizing a suitable distribution for mixture of directional data. The von MisesFisher distributions (VMF) are defined on the hypersphere $S^{(d-1)}$ [2] and appear adapted in this context. We consider Mont Carlo simulations and examine through numerical experiments on "real data" to see the validity of the proposed criteria for our main goal to estimate the number of clusters in a mixture.

This paper is organized as follows. Section 2 is devoted to describe the VMF mixture model. Section 3 begins with a review of the ML and CML approaches and a description of the EM and CEM algorithms. In Section 3, we review several criteria used in the determination of the number of clusters, and we evaluate these criteria. Finally, the last section summarizes the main points of this paper.

Notation Along this work, we assume that the data matrix $\mathbf{x}$ is a contingency table, crossing, for example, n documents (rows) and d words (columns). In this case, each document is represented by $\mathbf{x}_i = (x_i^1, \cdots, x_i^d) \in \mathbf{R}^d$, with $\|\mathbf{x}_i\| = 1$ ($\|.\|$ denotes the standard L_2). Each value x_i^j corresponds to the frequency of a word j in a document i. A clustering of n documents provides a partition z into g classes.

2 Clustering via the von Mises-Fisher mixture models

2.1 Finite Mixture Model

Finite mixture models underpin a variety of techniques in major areas of statistics including cluster analysis; see for instance [13]. With a mixture model-based approach clustering, it is assumed that the data to be clustered are generated by a mixture of underlying probability distributions in which each component represents a different cluster. Given observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, let $\varphi_k(\mathbf{x}_i; \alpha_k)$ be the density of an observation $\mathbf{x}_i$ from the kth component, where the α_k's are the corresponding parameters and let g be the number of components in the

*1. LIPADE EA 2517, UFR Maths-Info, Paris Descartes University, 45, rue des Saints Peres 75006 Paris, France. 2. LMNF laboratory, Department of Mathematics, Faculty of sciences, Ferhat Abbas university, Setif. Algeria. Emails: wboub@yahoo.fr, mohamed.nadif@univ-paris5.fr, bencheikh_00@yahoo.fr

ISBN: 978-988-18210-8-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

mixture. The probability density function is

$$f(\mathbf{x}_i;\theta)=\sum_{k=1}^{g}\pi_k\varphi_k(\mathbf{x}_i;\alpha_k), \quad (1)$$

where π_k is the probability that an observation belongs to the kth component and θ is the vector of the unknown parameters $(\pi_1,\ldots,\pi_g;\alpha_1,\ldots,\alpha_g)$.

Setting the clustering problem of directional data under the mixture model approach, we assume that x is generated from a von Mises-Fisher mixture of g components. In this case

$$\varphi_k(\mathbf{x}_i;\alpha_k)=c_d(\xi_k)\exp{\xi_k}^T\mu_k\mathbf{x}_i$$

where $\alpha_k=(\mu_k,\xi_k)$; μ_k is the center, ξ_k is the concentration of the kth cluster and $c_d(\xi)=\frac{\xi^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}}I_{\frac{d}{2}-1}(\xi)}$ with $I_{\frac{d}{2}}(\xi)$ is the modified Bessel function of the 1^{st} type and of order $\frac{d}{2}$: $I_d(\xi)=\frac{1}{2\pi}\int_0^{2\pi}e^{\xi\cos\theta}\cos(d\theta)d\theta$.

Note that we can consider different parsimonious models by imposing constraints on π_k and ξ_k.

1. the proportions π_k of clusters and the concentrations ξ_k are supposed not equal, this model is noted $[\pi_k,\xi_k]$,
2. the concentrations ξ_k are supposed equal, this model is noted $[\pi_k,\xi]$,
3. the proportions π_k of clusters are supposed equal, this model is noted $[\pi,\xi_k]$,
4. the proportions π_k of clusters and the concentrations ξ_k are supposed equal, this model is noted $[\pi,\xi]$.

Next we focus on the general model $[\pi_k,\xi_k]$.

2.2 ML and CML approaches

The problem of clustering can be studied in the mixture model using the ML approach. This one, by maximizing the likelihood

$$L(\theta)=\prod_i\sum_{k=1}^{g}\pi_k\varphi_k(\mathbf{x}_i;\alpha_k),$$

has been by far the most commonly used approach to the fitting of mixture distribution and is appropriate to tackle this problem. It estimates the parameters of the mixture, and the partition of I is derived from these parameters using the maximum a posteriori principle (MAP). Classical optimization techniques such as Newton-Raphson or gradient methods can be used but, in mixture context, the EM algorithm [10] has been successfully applied and is one of the most widely used procedures.

2.2.1 EM and CEM Algorithms

The EM algorithm is a method for maximizing the log-likelihood $L(\theta)$ iteratively, using the maximization of the conditional expectation of the complete-data log-likelihood given a previous current estimate $\theta^{(c)}$ and the observed data $\mathbf{x}$. In mixture model, we take the complete-data to be the vector $(\mathbf{x},\mathbf{z})$ where the unobservable vector $\mathbf{z}$ is the label data; the complete-data log-likelihood $L_c(\theta;\mathbf{x},\mathbf{z})$ noted also $L_c(\mathbf{z};\theta)$ is

$$L_c(\mathbf{z};\theta)=\sum_{i,k}z_{ik}\log\pi_k\varphi_k(\mathbf{x}_i;\alpha_k) \quad (2)$$

and its conditional expectation can be written

$$\begin{aligned}Q(\theta,\theta^{(c)}) &= \sum_{i,k}s_{ik}^{(c)}\log(\pi_k\varphi_k(\mathbf{x}_i;\alpha_k)\\ &= \sum_{i,k}s_{ik}^{(c)}\log\left(\pi_k c_d(\xi_k)e^{{\xi_k}^T\mu_k x_i}\right)\end{aligned}$$

where $s_{ik}^{(c)}=P(z_{ik}=1|\mathbf{x},\theta^{(c)})=\frac{\pi_k^{(c)}\varphi_k(\mathbf{x}_i;\alpha_k^{(c)})}{\sum_{k'=1}^{g}\pi_{k'}^{(c)}\varphi_{k'}(\mathbf{x}_i;\alpha_{k'}^{(c)})}$ denotes the conditional probability, given $\mathbf{x}$ and $\theta^{(c)}$, that $\mathbf{x}_i$ arises from the mixture component with density $\varphi_k(\mathbf{x}_i;\alpha_k)$. Each iteration of EM has two steps: an E-step and a M-step. The $(c+1)$st E-step finds the conditional expectation of the complete-data log-likelihood. Note that in the mixture case this step reduces to the computation of the conditional density of the $s_{ik}^{(c)}$. The $(c+1)$st M-step finds $\theta^{(c+1)}$ maximizing $Q(\theta,\theta^{(c)})$.

The characteristics of the EM algorithm are well documented. It leads in general to simple equations, has the nice property of increasing the log-likelihood at each iteration until stationarity, and in many circumstances, it derives sensible parameter estimates and consequently it is a popular tool to obtain maximum likelihood estimation. The EM algorithm can be viewed as a soft algorithm, and the partition can be derived from the parameters by using the MAP.

Note that a hard version CEM [9] can be performed by substituting $Q(\theta,\theta^{(c)})$ by $L_c(\theta)$. The main modifications concern therefore the conditional maximization of complete data log-likelihoods w.r. to $\mathbf{z}$ given θ. in this context, we are not treating the estimation problem but; we are dealing with the problems of the selection of the number of components in a mixture.

2.2.2 The EM steps for a von Mises-Fisher mixture model

The EM algorithm,as explained previously is used to compute the maximum likelihood (ML) estimates of all the parameters through the iterated application of the estimation and maximization of $Q(\theta,\theta^{(c)})$. Starting from

ISBN: 978-988-18210-8-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

an initial situation $\theta^{(0)}$, an iteration ($c > 0$) is defined as follows: After the Estimation step, where the current posterior $s_{ik}^{(c)}$ is computed. The Maximization step compute the ML estimates $\theta^{(c)} = (\mu_k^{(c)}, \pi_k^{(c)}, \xi_k^{(c)})$, as following:

- $\pi_k^{(c)} = \frac{\sum_{i=1}^{n} s_{ik}^{(c)}}{n}$
- $\mu_k^{(c)} = \frac{\sum_{i=1}^{n} s_{ik}^{(c)} x_i}{\left\| \sum_{i=1}^{n} s_{ik}^{(c)} x_i \right\|}$
- $\xi_k^{(c)} = A_d^{-1}(\frac{\left\| \sum_{i=1}^{n} s_{ik}^{(c)} x_i \right\|}{\pi_k^{(c)} \times n})$
  with $A_d(\xi) = \frac{I_{\frac{d}{2}}(\xi)}{I_{\frac{d}{2}-1}(\xi)}$

Then, a partition $z = (z_1, \cdots, z_k)$ of the data can be directly derived from the ML estimates of the mixture parameters by assigning each xi to the component which provided the greatest posterior probability.

3 Number of components selection

Determining the number of components g can be viewed as a model selection problem which can be solved by different criteria: information model selection criteria, or by methods based on confidence interval, and empirical criteria [8]. In the current paper, we focus on the information criteria, for they are the most important and popular practical techniques. This consists of penalizing the model with additional parameters. These criteria split into two terms: one for the model fit, which is data likelihood or complete data likelihood, and one for the model complexity.

3.1 Information criteria

Let L be the $log-$likelihood of observed data, L_c be the complete data $log-$likelihood with the parameter $\hat{\theta}$ obtained by the EM algorithm, υ be the number of free parameters in the mixture model and $E = \sum_{i,k} s_{ik} \log(s_{ik})$ the entropy criterion. The terms L, L_c, υ and E depend on g. In the following, we shall focus on twelve criteria.

- $Bic(g) = -2L(g) + \upsilon \ln n$, proposed by Schwarz [17] and Rissanen [16]
- $Aic(g) = -2L(g) + 2\upsilon$, proposed by Akaike [1]
- $Aic3(g) = -2L(g) + 3\upsilon$, proposed by Bozdogan [7]
- $Aic4(g) = -2L(g) + 4\upsilon$, proposed by Bozdogan [7]
- $Aicc(g) = Aic(g) + \frac{2\upsilon(\upsilon+1)}{n-\upsilon-1}$, proposed by Hurvich and Tsai [12]
- $Aicu(g) = Aicc(g) + n \ln n/(n - \upsilon - 1)$, proposed by McQuarrie, Schwarz and Tsai [14]
- $CAic(g) = -2L(g) + \upsilon(1 + \ln n)$, proposed by Bozdogan [6]
- $Clc(g) = -2L(g) + 2E(g)$, proposed by Biernacki [4]
- $IclBic(g) = Bic(g) + 2E(g)$, proposed by Biernacki, Celeux and Govaert [5]
- $Ll(g) = -L(g) + \frac{\upsilon}{2} \sum_k \ln \frac{n\pi_k}{2} + \frac{g}{2} ln \frac{n}{12} + \frac{g(\upsilon+1)}{2}$, proposed by Figueiredo and Jain (2002) [11]
- $Icl(g) = -2L_c(g) + \upsilon \ln n$, proposed by Biernacki, Celeux and Govaert [5]
- $Awe(g) = -2L_c(g) + 2\upsilon(\frac{3}{2} + \ln n)$, proposed by Banfield and Raftery [3]

3.2 Experimental conditions

In our experiments, we perform a study according to the degree of overlap of clusters and the size of data.

1. The concept of cluster separation is difficult to visualize easily for our model, but the degree of overlap can be measured by the true error rate approximated by comparing the partitions simulated with those we obtained by applying a classification step. From our numerical experiments, we present only 3 situations corresponding to 3 levels of overlap degrees: clusters well separated ($\approx 5\%$), moderately separated ($\approx 15\%$) and poorly separated ($\approx 23\%$).
2. We selected several sizes of data 600×3, 1800×3, 6000×3, 6000×50 and 6000×50 data arising from 3-components mixture model corresponding to the three degrees of overlap.

To evaluate the EM algorithm and the previous criteria, many applications on simulated data was realized. For each θ leading the degree of overlap, we generated 20 samples. For each sample and to avoid local optima in the generated estimation process, the $EM(g)$ algorithm $g = 2, \ldots, 5$ regarding the the general model $[\pi_k, \xi_k]$, is repeated 20 times starting from the best partition obtained by the spherical $k$means [2] which is a CEM applied with the model $[\pi, \xi]$. From the best solution,

1. we compute the percent of documents misclassified by comparing the true partition and the obtained partition with the same number of clusters,
2. we compute all criteria previously cited in function of different values of g,
3. we count the number of times on 20 that each criterion detects the original number of clusters *fit*, overestimates it *over-fit* or underestimates it *under-fit*. In table 1 are reported all results obtained by all criteria.

ISBN: 978-988-18210-8-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

From these experiments, the main points arising are the following.

- The EM algorithm gives good results by comparing the true partition and the obtained one by EM(3).
- When the clusters are well or moderately separated Aic3, Aic4, Aicu and Bic are the more efficient for the studied sizes.
- When the clusters are poorly separated, the quality of these criteria increases with the size of the data n and when $n >> d$.
- Moreover note that Aic3 and Aicu outperform Bic when the number of columns increases and remain interesting in the most situations. In fact, Bic seems very sensitive to the dimension, it underestimates the number of clusters.

In these first experiments, we can consider that Aic3 and Aicu are the best criteria. Note that Aic3 is also interesting for the Bernoulli mixture model for the binary data [15]. Nevertheless, we have noted that their performances decrease when we are in the high dimension. Then we illustrate the behavior of all criteria by using a well known set of data known as Classic3 as a real data application. This is a set of documents from three well separated sources. Classic3 contains 3893 documents (vectors) with a total of 4303 features (words). The data matrix consists of 1400 Cranfield documents from aeronautical system papers, 1033 from Medline documents obtained from medical journals, and 1460 Cisi documents obtained from information retrieval papers. Each vector was normalized in order to be used as a unit vector. In order to select a number of clusters in $g = 2, \ldots, 5$, we have computed the same criteria as previously, we applied the $EM(g)$ algorithm regarding the general model $[\pi_k, \xi_k]$ and we obtained the following results:

- Bic, Caic, Icl-Bic, Icl overestimate the number of clusters and give 4 clusters.
- Aic, Aic3, Aic4, Aicc, Clc overestimate the number of clusters and give 5 clusters.
- Aicu, Ll, Awe underestimate the number of clusters and give 2 clusters.

4 Conclusion

Setting the clustering of directional data in the mixture approach context, we have performed some experiments in order to evaluate the EM algorithm and to assess the number of clusters. Different information criteria have been tested on different sizes of data according different degree of overlap. We have observed that some of them such as Aic3, Aic, Aicu and Bic are interesting. Moreover we have noted that their performance increases on the size of data and Aic3 and Aicu appear as the best.

In future work, it will be interesting 1) to take into account the high dimension in these criteria and 2) to tackle simultaneously the problem of assessing of the number of clusters combined to the choice of the parsimonious models $[\pi_k, \xi]$, $[\pi_k, \xi]$ and $[\pi, \xi]$.

References

[1] Akaike, H., "Information theory and an extension of maximum likelihood principle," *Second International Symposium on Information Theory*, Akademia Kiado, 267-281, 1973.

[2] Banerjee, A., Dhillon, I. S., Ghosh J., Sra S., "Clustering on the Unit Hypersphere Using Von Mises-Fisher Distributions," *Journal of Machine Learning Research*, 6:1345-1382, 2005.

[3] J. D. Banfield and A. E. Raftery., "Model-based gaussian and non-gaussian clustering," *Biometrics*, 49:803821, 1993.

[4] Biernacki, C., "Choix de modèles en Classification," *PhD Thesis*, Compiègne University of Technology, 1997.

[5] Biernacki, C., Celeux, G. and Govaert, G., "Assessing a Mixture model for Clustering with the integrated Completed Likelihood," *IEEE Transactions on Pattern analysis and Machine Intelligence 22 (7)*, pp. 719-725, 2000.

[6] Bozdogan, H., "Model Selection and Akaikes's Information Criterion (AIC): The General Theory and its Analytical Extensions", *Psycometrika 52 (3)*, pp. 345-370, 1987.

[7] Bozdogan, H.,"Mixture-Model Cluster Analysis Using Model Selection Criteria and a New Information Measure of Complexity" *Proceedings of the first US/Japan conference on the Frontiers of Statistical Modeling: An Informational Approach*, 1 ed. 3 vols. Vol. 1., Dordrecht, Kluwer Academic Publishers, 1994.

[8] Bubna K., Stewart, C.V., "Model Selection Techniques and Merging Rules for Range Data Segmentation Algorithms," *Computer Vision and Image Understanding*, 80: 215-245, 2000.

[9] Celeux, G., Govaert, G., "A classification *EM* Algorithm for clustering and two stochastic versions," *Computational statistics & Data analysis*, 14:315-332, 1992.

[10] Dempster, A.P., Laird, N.M., Rubin, D., "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J of the Royal Stat Soc*, B 39: 1-38, 1977

ISBN: 978-988-18210-8-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[11] Figueiredo, M.A.T., and Jain, A.K., "Unsupervised Learning of Finite Mixture Models," *IEEE Transactions on pattern analysis and Machine Intelligence 24 (3)*, pp. 1-16, 2002.

[12] Hurvich, C.M., and Tsai, C.-L., "Regression and Time Series Model Selection in Small Samples," *Biometrika 76 (2)*, pp. 297-307, 1989.

[13] McLachlan, G.J., Peel, D., *Finite mixture models*, *Wiley*, New York, 2000.

[14] McQuarrie, A., Shumway, R. and Tsai, C.-L., "The model selection criterion AICu," *Statistics & Probability Letters 34*, pp. 285-292, 1997.

[15] Nadif, M., Govaert, G., "Clustering for binary data and mixture models: Choice of the model," *Applied Stochastic Models and Data Analysis*, 13: 269-278, 1998.

[16] Rissanen, J., "Modelling by shortest data description," *Automatica*, 14:465-471, 1978.

[17] Schwarz, G. "Estimating the Dimension of a Model," *The Annals of Statistics 6 (2)*, pp. 461-464, 1978.

ISBN: 978-988-18210-8-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table 1: Evaluation of EM and all information criteria for the model $[\pi_k, \xi_k]$. For each criterion, the numbers of times on 20 indicate that a criterion detects or not the good number of clusters.

size	*degree*	$EM(3)$	*fit*	*Bic*	*Aic*	*Aic3*	*Aic4*	*Aicc*	*Aicu*	*CAic*	*Clc*	$Icl - Bic$	*Ll*	*Icl*	*Awe*
600×3	4.88%	5.17%	under-fit	0	0	0	0	0	0	0	0	0	0	0	0
			fit	20	15	19	20	15	19	20	15	20	20	20	20
			over-fit	0	5	1	0	5	1	0	5	0	0	0	0
1800×3	5.16%	4.83%	under-fit	0	0	0	0	0	0	0	0	0	0	0	0
			fit	20	18	19	20	18	19	20	20	20	20	20	20
			over-fit	0	2	1	0	2	1	0	0	0	0	0	0
3000×50	4.74%	6.85%	under-fit	0	0	0	0	0	0	0	0	1	7	0	4
			fit	20	1	20	20	1	20	20	6	19	13	20	16
			over-fit	0	19	0	0	19	0	0	14	0	0	0	0
600×3	14.63%	16.33%	under-fit	0	0	0	0	0	0	0	9	16	7	7	16
			fit	20	17	20	20	17	20	20	9	4	13	13	4
			over-fit	0	3	0	0	3	0	0	2	0	0	0	0
1800×3	15.10%	15.83%	under-fit	0	0	0	0	0	0	0	14	18	0	2	7
			fit	20	19	20	20	19	20	20	6	2	20	18	13
			over-fit	0	1	0	0	1	0	0	0	0	0	0	0
3000×50	13.68%	14.10%	under-fit	0	0	0	0	0	0	0	0	3	0	0	20
			fit	20	10	20	20	10	20	20	18	17	20	20	0
			over-fit	0	10	0	0	10	0	0	2	0	0	0	0
600×3	24.96%	29.17%	under-fit	20	15	17	20	15	18	20	20	20	20	20	20
			fit	0	3	3	0	3	2	0	0	0	0	0	0
			over-fit	0	2	0	0	2	0	0	0	0	0	0	0
1800×3	25.19%	35.94%	under-fit	20	12	17	19	12	17	20	20	20	20	20	20
			fit	0	8	3	1	8	3	0	0	0	0	0	0
			over-fit	0	0	0	0	0	0	0	0	0	0	0	0
6000×3	27.49%	30.95%	under-fit	0	0	0	0	0	0	0	20	20	8	20	20
			fit	20	20	20	20	20	20	20	0	0	12	0	0
			over-fit	0	0	0	0	0	0	0	0	0	0	0	0
3000×50	24.75%	32.26%	under-fit	18	0	0	0	0	0	20	20	20	20	20	20
			fit	2	8	20	20	8	20	0	0	0	0	0	0
			over-fit	0	12	0	0	12	0	0	0	0	0	0	0
6000×50	25.61%	39.17%	under-fit	20	0	1	16	0	1	20	20	20	20	20	20
			fit	0	11	19	4	11	19	0	0	0	0	0	0
			over-fit	0	9	0	0	9	0	0	0	0	0	0	0

ISBN: 978-988-18210-8-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)