ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Sparse mixture of von Mises-Fisher distribution

Florian Barbaro[1] and Fabrice Rossi[2]

1- Université Paris 1 Panthéon-Sorbonne - Laboratoire SAMM EA 4543, France
2- Université Paris Dauphine-PSL - CEREMADE UMR 7534, France

Abstract. Mixtures of von Mises-Fisher distributions can be used to cluster data on the unit hypersphere. This is particularly adapted for high-dimensional directional data such as texts. We propose in this article to estimate a von Mises mixture using a l_1 penalized likelihood. This leads to sparse prototypes that improve both clustering quality and interpretability. We introduce an expectation-maximisation (EM) algorithm for this estimation and show the advantages of the approach on real data benchmark. We propose to explore the trade-off between the sparsity term and the likelihood one with a simple path following algorithm.

1 Introduction

Many classical mixture models are poorly suited to high-dimensional data, including those derived from the vector representation of text. When the data is directional [1], i.e. when it is their correlation rather than their Euclidean distance that matters, Gaussian-type models are even less suitable. For such data, it is natural to carry out a normalisation that places them on the unity sphere. As an alternative to spherical k-means [2], mixtures of von Mises-Fisher (vMF) on this sphere have been shown to provided good clustering results, cf [3, 4, 5].

In this article, following [6], we propose a l_1 penalty for a mixture of von Mises-Fisher to induce sparsity of directional means and thus improve the understanding of classification results for high-dimensional data. Parameters are estimated by an EM algorithm based on the solution from [4]. The penalty parameter β is set automatically using the BIC model selection criterion and leveraging a path following strategy. Sparse prototypes are represented graphically via a simple reordering heuristics inspired by [7, 8] (see Figure 1 for an example): it emphasizes a partial co-clustering structure as well as shared features that are not detectable with co-clustering methods.

Notations: Matrices are denoted with boldface uppercase letters, vectors with boldface lowercase letters. Norm l_1 is noted as $\|.\|_1$ and l_2 as $\|.\|_2$. Data are represented by a matrix $\boldsymbol{X} = (x_{ij})$ of dimension $n \times d$ with $x_{ij} \in \mathbb{R}$ and the i^{th} row of this matrix is represented by a vector $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{id})^T$, where T is the transpose. The partition of all lines I into k clusters is represented by a classification matrix $\boldsymbol{Z}$ with elements z_{ih} in $\{0, 1\}$ satisfying $\sum_{h=1}^{k} z_{ih} = 1$.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

2 Mixture of von Mises-Fisher distribution

First, we recall the mixture model proposed in [4]. The von Mises-Fisher (vMF) distribution in dimension $d \geq 2$ is supported by the $(d-1)$ dimensional unit sphere embedded in $\mathbb{R}^d$ (denoted by $\mathbb{S}^{d-1}$). Its probability density function is given by

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \kappa) = c_d(\kappa) \exp^{\kappa \boldsymbol{\mu}^T \boldsymbol{x}}, \tag{1}$$

where $\|\mu\|_2 = 1$ and $\kappa \geq 0$ are the parameters. The normalization term $c_d(\kappa)$ is

$$c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}, \tag{2}$$

with I_r is the modified Bessel function of the first kind and order r.

A mixture of K von Mises Fisher distributions is specified by K vMF densities $f_h(x|\theta_h)$ with $\theta_h = (\boldsymbol{\mu}_h, \kappa_h)$ for $1 \leq h \leq K$. The mixture density is given by

$$f(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{h=1}^{K} \alpha_h f_h(\boldsymbol{x}|\theta_h), \tag{3}$$

with $\Theta = \{\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K\}$, $\alpha_h \geq 0$ et $\sum_{h=1}^{K} \alpha_h = 1$.

We use a standard hidden variable approach to represent the mixture: an observation $\boldsymbol{x}$ is generated from (3) by first sampling a hidden variable $\boldsymbol{z}$ in $\{0,1\}^K$ with $\sum_h z_h = 1$ and $\mathbb{P}(z_h = 1, \forall l \neq h, z_l = 0|\boldsymbol{\alpha}) = \alpha_h$; and then by sampling $\boldsymbol{x}$ from f_h if $z_h = 1$. For n independently observations $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ we obtain the following complete data log-likelihood

$$\ln L_c(\Theta) = \sum_{i=1}^{n} \sum_{h=1}^{K} z_{ih} \left(\ln \alpha_h + \ln f_h(\boldsymbol{x}_i|\theta_h) \right). \tag{4}$$

Based on this complete data-likelihood, the only complex step of an EM based estimation of the parameters is the computation of κ_h which is detailed in [4].

3 Proposed model

3.1 Penalized likelihood

Following [6], we add a l_1 norm penalisation to the log-likelihood in order to favor sparsity in the parameters. More precisely, we seek to estimate Θ by maximizing the log-likelihood penalized

$$\ln L_p(\Theta) = \ln L(\Theta) - \beta \sum_{h=1}^{K} \|\boldsymbol{\mu}_h\|_1, \tag{5}$$

where β regulates the trade-off between likelihood and sparsity. Similarly, the penalized complete data log-likelihood $\ln L_{c,p}(\Theta)$ is obtained by subtracting the same penalty term from $\ln L_c(\Theta)$.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

3.2 EM algorithm

To derive the EM algorithm, let us first denote $\Theta^{(m)}$ the estimate of the parameters at iteration m of the algorithm. We denote $\tau_{ih}^{(m)} = \mathbb{P}(z_{ih} = 1|\boldsymbol{x}_i, \Theta^{(m)})$ the probability that $\boldsymbol{x}_i$ was generated by component h of the mixture. Then the E phase consist in computing

$$\begin{aligned} Q_P(\Theta|\Theta^{(m)}) &:= \mathbb{E}_{\boldsymbol{Z}\sim\mathbb{P}(\boldsymbol{Z}|\boldsymbol{X},\Theta^{(m)})} \ln L_{c,p}(\Theta), \\ &= \sum_{i=1}^{n}\sum_{h=1}^{K} \tau_{ih}^{(m)} \left(\ln \alpha_h + \ln f_h(x_i|\theta_h)\right) - \beta \sum_{h=1}^{K} \|\boldsymbol{\mu}_h\|_1 . \end{aligned} \tag{6}$$

M step maximizes $Q_P(\Theta|\Theta^{(m)})$ with respect to Θ and under the constraints $\sum_{h=1}^{K} \alpha_h = 1$, $\|\boldsymbol{\mu}_h\|_2 = 1$ and $\kappa_h \geq 0$ for $1 \leq h \leq K$. To respect these constraints on α_h and $\boldsymbol{\mu}_h$, we introduce $K+1$ Lagrange multipliers, respectively ζ et $(\lambda_h)_{1\leq h\leq K}$ leading to

$$\mathcal{L}(\Theta, \zeta, \boldsymbol{\lambda}|\Theta^{(m)}) = Q_P(\Theta|\Theta^{(m)}) + \zeta(\sum_{h=1}^{K} \alpha_h - 1) + \sum_{h=1}^{K} \lambda_h(1 - \boldsymbol{\mu}_h^T \boldsymbol{\mu}_h). \tag{7}$$

Compared to the derivations of [4], the main difference comes from the calculation of the subgradient $\mathcal{L}$ with respect to μ_{hj}. We have indeed

$$\partial_{\mu_{hj}} \mathcal{L}(\Theta, \zeta, \boldsymbol{\lambda}|\Theta^{(m)}) = \kappa_h \left(\sum_{i=1}^{n} \tau_{ih}^{(m)} x_{ij} \right) - 2\lambda_h \mu_{hj} - \beta \partial_{\mu_{hj}} |\mu_{hj}|. \tag{8}$$

The first-order optimality condition is $0 \in \partial_{\mu_{hj}} \mathcal{L}(\Theta, \zeta, \boldsymbol{\lambda}|\Theta^{(m)})$. By writing $r_{hj}^{(m)} = \sum_{i=1}^{n} \tau_{ih}^{(m)} x_{ij}$, we get :

$$\partial_{\mu_{hj}} \mathcal{L}(\Theta, \zeta, \boldsymbol{\lambda}|\Theta^{(m)}) = \begin{cases} \kappa_h r_{hj}^{(m)} - 2\lambda_h \mu_{hj} + \beta & \text{if } \mu_{hj} < 0 \\ \kappa_h r_{hj}^{(m)} - \epsilon\beta, \epsilon \in [-1; 1] & \text{if } \mu_{hj} = 0 \\ \kappa_h r_{hj}^{(m)} - 2\lambda_h \mu_{hj} - \beta & \text{if } \mu_{hj} > 0 \end{cases} \tag{9}$$

Some algebraic manipulations yield

$$\mu_{hj}^{(m)} = \text{sign}(r_{hj}^{(m)}) \max \left(\frac{\kappa_h |r_{hj}^{(m)}| - \beta}{\sqrt{\sum_{j=1}^{d} (\kappa_h |r_{hj}^{(m)}| - \beta)^2}}, 0 \right), \tag{10}$$

with $\lambda_h = \frac{1}{2}\sqrt{\sum_{j=1}^{d} (\kappa_h |r_{hj}^{(m)}| - \beta)^2}$. Note that the addition of the penalty introduces a coupling between κ_h and μ_h that does not exist in its absence (we see that if we fix $\beta = 0$, κ_h is no longer involved in the definition of $\boldsymbol{\mu}_h$). One must therefore solve

$$\frac{c_d'(\kappa_h)}{c_d(\kappa_h)} = -\frac{\boldsymbol{\mu}_h \sum_{i=1}^{n} \tau_{ih}^{(m)} \boldsymbol{x}_i}{\sum_{i=1}^{n} \tau_{ih}^{(m)}}. \tag{11}$$

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

We use the approximation proposed in [4]. If we introduce $\bar{r}_h^{(m)} = \frac{\boldsymbol{\mu}_h \sum_{i=1}^n \tau_{ih}^{(m)} \boldsymbol{x}_i}{\sum_{i=1}^n \tau_{ih}^{(m)}}$, κ_h is given by

$$\kappa_h = \frac{\bar{r}_h^{(m)} d - (\bar{r}_h^{(m)})^3}{1 - (\bar{r}_h^{(m)})^2}. \tag{12}$$

We propose to estimate from $\boldsymbol{\mu}_h$ from $\kappa_h^{(m)}$, then to update κ_h. We could iterate those updates to reach a fixed point, but experiments showed that this was not necessary.

3.3 Path following approach

Rather than testing for different values of β on a grid, we propose to use a path following approach. We initialise the path by estimating the parameter for the non sparse model with $\beta_0 = 0$. Equation (10) shows that provided that β is small enough, no component will be set to zero. This enables us to determine the smallest value of β that creates some sparsity, i.e. $\beta_1 = \min_{h,j,\kappa_h r_{hj} \neq 0} |\kappa_h r_{hj}|$. Starting from the solution with $\beta = 0$, we set β to this value and run the EM algorithm to convergence.

This process is repeated by adding to the sum of the β obtained in previous iterations the minimal increment that increases the parsimony of the $\boldsymbol{\mu}_h$. We obtain β_N using

$$\beta_N = \beta_{N-1} + \min_{h,j,|\kappa_h r_{hj} - \beta_{N-1}| > 0} |\kappa_h r_{hj} - \beta_{N-1}|. \tag{13}$$

To avoid taking too many steps on this path, we set values smaller than the chosen numerical precision threshold to zero after updating β.

3.4 Model selection

We propose to select the model retained for a dataset by using the BIC criterion. Only non-zero parameters for μ_{hj} are considered as effective parameters (according to the consistency results obtained in [9]). So we have

$$BIC = -2 \times L(\hat{\Theta}) + C \times \log(n), \tag{14}$$

with C, number of parameters, equal to $C = (K - 1 + K) + \sum_h \sum_j \mathbb{I}_{\mu_{hj} \neq 0}$.

4 Experimental results

We compare our approach to alternative solutions on the popular dataset called CSTR [10][1]. It consists in $n = 475$ abstracts of technical reports (TRs) published in the Department of Computer Science at the University of Rochester between 1991 and 2002. Each abstract is given by a $d = 1000$ vector. The abstract are classified into four research areas: Natural Language Processing(NLP),

[1] Available here: https://github.com/dbmovMFs/DirecCoclus/tree/master/Data.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Robotics/Vision, Systems, and Theory. We use the proposed model with values of K ranging from 2 to 5. For each mixture, we use the path following technique coupled with the BIC criterion to select an optimal value of β.

Our model is compared to a standard non sparse mixture of vMF distribution with $K = 4$ as well as to co-clustering variant dbmovMF [7]. We use also as reference Gaussian mixture models adapted to high-dimensional data, namely HDDC [11] and Fisher EM [12]. We use the Adjusted Rand Index (ARI) to compare the results. All reference models are used with $K = 4$ clusters, while this value is automatically selected based on the BIC for our model. Results are reported in table 1. They show that our approach is able to retrieve the structure of the CSTR in an unsupervised and fully automated way by improving both over the ARI and the sparsity. Those results have been confirmed on other data sets.

Models	ARI	Sparsity
K = 4 beta = 0	0.56	0
K = 2 beta = 87.17	0.42	0.34
K = 3 beta = 77.59	0.53	0.47
K = 4 beta = 65.37	**0.67**	**0.58**
K = 5 beta = 45.19	0.507	0.55
dbmovMF K = 4	0.55	0
HDDC K=4	0.04	0.22
Fisher EM K=4	0.50	0.42

Table 1: Results CSTR

Figure 1 shows the sparsity of $\boldsymbol{\mu}$ for CSTR dataset where black blocks mean important features in the cluster representation (non zero values). Moreover, $\boldsymbol{\mu}$ was reordered as in [7, 8] to reveal its block structure. In contrast to the two articles cited, which require a variable to belong to only one class, our approach highlights variables of common importance for each class and those that are more discriminating for one or more classes. All this makes it easy to see the similarities but also the discriminating elements of each class.

5 Conclusion

In this article, we seek to penalize the likelihood of a mixture of von Mises-Fisher distributions to increase the parsimony of directional means. We show that the penalization by means of the l_1 norm allows us to reach our end using an adaptation of the EM algorithm as well as the combination between a path following approach and the BIC criterion to automatically select an optimal penalization parameter. Experimental results show that we outperformed reference method on a real world data set.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

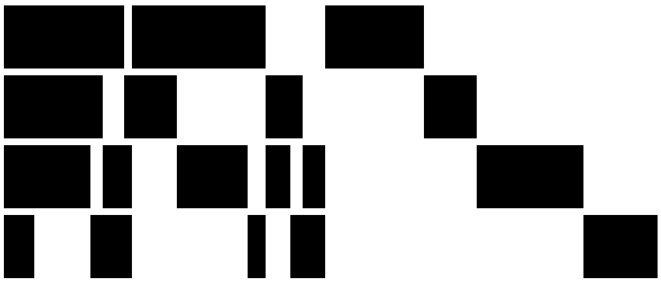

Figure 1: Sparsity of $\boldsymbol{\mu}$ for K = 4 beta = 65.37 on CSTR dataset. From common (left) to more discriminative features (right).

References

[1] K.V. Mardia and P.E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, 2009.

[2] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Mach. Learn.*, 42(1–2):143–175, January 2001.

[3] Shi Zhong and Joydeep Ghosh. Generative model-based document clustering: A comparative study. *Knowl. Inf. Syst.*, 8(3):374–384, September 2005.

[4] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *J. Mach. Learn. Res.*, 6:1345–1382, December 2005.

[5] Siddharth Gopal and Yiming Yang. Von Mises-Fisher Clustering Models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 154–162, Bejing, China, 22–24 Jun 2014. PMLR.

[6] Wei Pan and Xiaotong Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(41):1145–1164, 2007.

[7] Aghiles Salah, Nicoleta Rogovschi, and Mohamed Nadif. Model-based co-clustering for high dimensional sparse data. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th Int. Conf. on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 866–874, Cadiz, Spain, 09–11 May 2016. PMLR.

[8] Aghiles Salah and Mohamed Nadif. Model-based von Mises-Fisher co-clustering with a conscience. In *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM'17)*, pages 246–254, Houston, TX, United States, 2017. SIAM.

[9] Hui Zou, Trevor Hastie, and Robert Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5), Oct 2007.

[10] Tao Li. A general model for clustering binary data. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, page 188–197, New York, NY, USA, 2005. Association for Computing Machinery.

[11] Charles Bouveyron, Stéphane Girard, and Cordelia Schmid. High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52(1):502–519, 2007.

[12] Charles Bouveyron and Camille Brunet. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, 22(1):301–324, January 2012.