EM algorithm for bivariate circular distributions

K. V. Mardia[1], C. C. Taylor[1] and M. Subramaniam* [2]

[1] University of Leeds,
[2] AT&T Labs - Research

1 Introduction

Mardia, Taylor, and Subramaniam (2003) introduced Cosine models ($vMcos$), a class of bivariate circular models and addressed the issue of modeling the data using the Cosine distribution. In studying the properties of the Cosine model and contrasting it with Sine model ($vMsin$), the model parameters were estimated using several bivariate circular data sets. Existence of bimodality in these data sets raised issues about whether these data came from the Cosine or Sine distributions or from mixtures of these distributions. The main goal of this paper is to fit mixtures for both Cosine and Sine model using the EM algorithm. We will derive the mixture-density parameter update equations for a mixture of first von Mises, followed by Cosine and Sine distributions. This paper will also contain a brief discussion of several examples from both simulated and real data to verify the recovery of the mixture parameters using the EM algorithm for the bivariate circular models. The EM discussion is restricted to two component mixture of von Mises, Cosine distribution and Sine distribution to illustrate the performance of EM algorithm for circular data. The theory can be easily extended for mixtures of more than two components.

2 EM Algorithm

The EM algorithm, due to Dempster, Laird, and Rubin (1977) is a method to calculate the maximum likelihood of the parameters of an underlying distribution from the given data set when the data set is incomplete or has missing values. One of the most widely used applications of the EM algorithm is the mixture density parameter estimation. In this, the data is assumed to be drawn from a population that is generated from a mixture of multiple distributions. We will briefly describe the estimation of the parameters associated with the component distributions and the mixing weights using the EM algorithm.

Let us consider the von Mises distribution $vM(x;\mu,\kappa)$ whose density is defined as follows:

$$f(x;\mu,\kappa) = \frac{1}{2\pi\, I_0(\kappa)}\, e^{\kappa\cos(x-\mu)} \tag{1}$$

The mixture density arising from the two von Mises distributions is given by

$$f(x) = \pi f_1(x|\,\mu_1,\kappa_1) + (1-\pi) f_2(x|\,\mu_2,\kappa_2) \tag{2}$$

where $f_1 \sim vM(\mu_1,\kappa_1)$ and $f_2 \sim vM(\mu_2,\kappa_2)$ are von Mises densities.

The log-likelihood is given by

$$L(\pi, \mu_1, \kappa_1, \mu_2, \kappa_2 \mid x) = \sum_{i=1}^{n} \log \Big[\pi f_1(x \mid \mu_1, \kappa_1) + (1-\pi) f_2(x \mid \mu_2, \kappa_2)\Big] \tag{3}$$

Maximizing the likelihood in Equation (3) (which is called the incomplete-data log-likelihood) is an alternative to EM.

To apply the EM algorithm we augment data with missing data. Let Z be a bernoulli random variable with $P(Z=1) = \pi$. Here z is the unobserved missing r.v., x is the observed data and (z, x) is the complete data. Let $\theta = (\mu_1, \kappa_1, \mu_2, \kappa_2)$.

The likelihood function for the complete data is

$$f(\pi, \theta \mid z, x) = \prod_{i=1}^{n} \Big[\pi f_1(x_i \mid \mu_1, \kappa_1)\Big]^{z_i} \Big[(1-\pi)\, f_2(x_i \mid \mu_2, \kappa_2)\Big]^{1-z_i}$$

Then log-likelihood for the complete data after simplification is

$$L(\pi, \theta \mid z, x) = \sum_{i=1}^{n} z_i \log \left(\Big[\frac{\pi}{1-\pi}\Big] \Big[\frac{f_1(x_i \mid \mu_1, \kappa_1)}{f_2(x_i \mid \mu_2, \kappa_2)}\Big] \right) + \log[(1-\pi)\, f_2(x_i \mid \mu_2, \kappa_2)] \tag{4}$$

The expectation step or E-step computes the expected likelihood for the complete data. The expectation is w.r.t the computed conditional distribution of hidden variables. Let ψ be the complete collection of parameters occurring in the mixture, i.e., $\psi = (\pi, \mu_1, \kappa_1, \mu_2, \kappa_2)$.

E-step: We take expectation and get Q function

$$\begin{aligned} Q(\psi \mid \psi^{(0)}, x) &= \mathbb{E}[L(\pi \mid Z, X) \mid x, \psi^{(0)}] \\ &= \sum_{i=1}^{n} \mathbb{E}(Z_i \mid x, \psi^{(0)}) \left(\log \Big[\frac{\pi}{1-\pi}\Big] + \log \Big[\frac{f_1(x_i)}{f_2(x_i)}\Big] \right) + \log[(1-\pi)\, f_2(x_i)] \end{aligned} \tag{5}$$

M-step: Here we maximize the expectation, i.e., the Q-function that we computed in the E-step. The two steps are repeated as necessary.

In summary, the EM algorithm starts by assigning initial values to all parameters to be estimated. It then iteratively alternates between two steps, the E-step and M-step. The E-step computes the expected likelihood for the complete data, given the current settings of parameters and the observed incomplete data. The M-step re-estimates all the parameters by maximizing the Q-function. If the convergence criteria is not met, then the parameters $\psi = (\pi, \mu_1, \kappa_1, \mu_2, \kappa_2)$ are updated. We can repeat E-step followed by the M-step until the likelihood converges. Each iteration is guaranteed to increase the log-likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function. A more comprehensive review of EM can be found in McLachlan and Krishnan (1997) and Titterington *et al.*, (1985).

3 Mixtures for Cosine Models

The **Cosine distribution** is $vMcos(\theta, \phi; \kappa_1, \kappa_2, \kappa_3, \mu, \nu)$ has density

$$f_{cos}(\theta, \phi) = \{c(\kappa_1, \kappa_2, \kappa_3)\}^{-1} \exp\{\kappa_1 \cos(\theta - \mu) + \kappa_2 \cos(\phi - \nu) - \kappa_3 \cos(\theta - \mu - \phi + \nu)\} \tag{1}$$

for $-\pi < \theta,\ \phi \leq \pi,\ -\pi < \mu,\ \nu \leq \pi,\ \kappa_1 \geq 0, \kappa_2 \geq 0$ and $\kappa_1 \geq \kappa_3 \geq 0, \kappa_2 \geq \kappa_3 \geq 0$. The constant term $\{c(\kappa_1, \kappa_2, \kappa_3)\}^{-1}$ is the appropriate normalizing constant for the density.

The **Sine distribution** is $vMsin(\theta, \phi; \kappa_1, \kappa_2, \lambda, \mu, \nu)$ has density

$$f_{sin}(\theta, \phi) = C \exp\{\kappa_1 \cos(\theta - \mu) + \kappa_2 \cos(\phi - \nu) + \lambda \sin(\theta - \mu) \sin(\phi - \nu)\} \quad (2)$$

μ and ν are *mean directions,* κ_1 and κ_2, are *concentration parameters* for θ and ϕ, respectively. The parameter λ describes the correlation and C is the normalizing constant. Further the parameters are constrained by the following conditions

$$-\pi < \theta, \phi \leq \pi, -\pi < \mu, \nu \leq \pi, \kappa_1, \kappa_2 \geq 0 \text{ and } -\infty < \lambda < \infty$$

Consider a mixture data (x, y) from two Cosine distributions

$$f_{cos}(x, y) = \pi f_{cos_1}(x, y) + (1 - \pi) f_{cos_2}(x, y) \quad (3)$$

where f_{cos_i} denotes a Cosine density with parameters $\kappa_{i,j}$ $j = 1, 2, 3,$ $\mu_i,$ $\nu_i,$ $i = 1, 2.$ The same definitions apply to a mixture data (x, y) from two Sine distributions.

4 Implementation Details

The EM algorithm was implemented in R. As closed form iterative equations for the EM were not computed, both the E and M steps were estimated numerically. The EM required the optimization step for every iteration. In the bivariate case, there are a total of eleven parameters to estimate, we use the real advantage of EM by making the optimization program (optim in R) call M-step twice, each time solving five parameters. Thus it reduces to "two" five parameter problem, instead of solving one eleven parameter problem. The algorithm implemented in R converged to the solutions within 100 iterations for all the data sets. The start values were chosen so that the means for the component densities were well separated. Random start values were chosen for concentration parameters. Several start values were used and the final solution was based on minimum negative log-likelihood criteria. To estimate the standard error for the parameters, the incomplete-data log-likelihood in Equation (3) was used. The hessian matrix associated with the solution was computed. The solutions were all local optimum as the associated hessian matrix were all negative definite.

4.1 Results

This section contains results from executing the EM algorithm and estimating parameters of the mixture distributions. A combination of real data and simulated data, i.e., data sampled from simulated mixtures of von Mises and Cosine distributions were used. First, a mixture data is simulated by drawing from two von Mises distributions, 200 random variates from $vM(\pi/3, 1.5)$ and 800 variates from $vM(4\pi/3, 8.5)$. Also a mixture data is simulated from two Cosine distributions. Gibbs sampling was used to simulate the individual component densities, 200 random variates drawn from $vMcos(3.5, 3.3, 1, 0, 0)$ and 800 random variates drawn from $vMcos(2.2, 1.7, 0.7, \pi/3, \pi/2)$. As Tables 2 and 3 indicate, for both the von Mises and Cosine distribution mixtures, the parameters of the component densities were recovered very well with small variance. The variance for the concentration parameters was generally higher than means. This is true for estimation of scale parameters in general.

We also tested the algorithm on conformational angles from protein, Malate dehydrogenase. The initial sample size of the protein data set was 1422. As they are angles that are sequenced in time, it has high autocorrelations. After some filtering of data (e.g., removal of 360° angles),

Case	μ_1	κ_1	μ_2	κ_2	π
True	1.047	1.5	4.189	8.5	0.2
Estimated	1.045	1.644	4.207	8.694	0.20
Std. Error	0.07	0.28	0.01	0.56	0.02

Table 2: Parameter Estimation for vM mixtures on a circle

Cluster 1	κ_1	κ_2	κ_3	μ	ν	π
True	3.5	3.3	1	0	0	0.20
Estimated	3.39	3.47	1.087	0.009	-0.0352	0.195
Std. Error	0.252	0.293	0.171	0.035	0.032	0.035
Cluster 2	κ_1	κ_2	κ_3	μ	ν	π
True	2.2	1.7	0.7	1.047	1.57	0.80
Estimated	2.40	1.82	0.62	0.989	1.597	0.804
Std. Error	0.383	0.473	0.326	0.120	0.169	0.035

Table 3: Estimated Parameter Values for the Simulated Cosine Mixture

we considered every third observation in the data set to remove the serial correlation. The sample size was reduced to 350. The circular histogram in Figure 1 shows the first angle θ to be bimodal. The Cosine density surface plots for (a) single component density, i.e., assuming the data doesn't arise from a mixture distribution (b) component density 1 and (c) component density 2 are shown in Figure 2 for the protein example. The surface plots for the single component case reveals several modes. The surface plots for the clusters show that they are well separated. The results suggest that the data comes from a mixture of Cosine distributions.

5 Conclusions

This paper explored the applicability of EM algorithm to Cosine and Sine models. First, this was established for the univariate von Mises distribution by simulating a mixture data. This was extended for the Cosine distribution. The m.l.e. equations for the mixture of von Mises, Cosine and Sine distributions were derived. An EM algorithm to estimate MLE parameters was also developed. The algorithm was verified by running them on both simulated and real data. We observe that for data sampled from mixture of von Mises distribution, we get fairly good estimates from the EM algorithm. For the Cosine and Sine distributions, although good results were observed, there were some numerical difficulties in convergence in the M-step. It could be

Data Type	κ_1	κ_2	κ_3	μ	ν	π
Single component	1.54	3.374	1.574	7.144	4.764	1
Cluster 1	5.772	5.031	5.791	1.394	0.025	0.262
Std. Error	0.654	0.592	0.657	0.064	0.078	0.027
Cluster 2	6.994	11.735	7.344	0.734	-1.808	0.738
Std. Error	0.543	0.851	0.569	0.044	0.029	0.027

Table 4: Estimated Parameter Values for the Cosine Density Mixture for Protein Data

Data Type	κ_1	κ_2	λ	μ	ν	π
Single component	0.3	2.83	-1.5	0.8	4.76	1
Cluster 1	15.113	14.440	-9.387	-0.545	-1.257	0.544
Std. Error	1.573	1.547	1.407	0.024	0.024	0.027
Cluster 2	2.002	0.144	-4.320	1.338	-0.235	0.456
Std. Error	0.322	0.218	0.340	0.064	0.068	0.027

Table 5: Estimated Parameter Values for the Sine Density Mixture for Protein Data

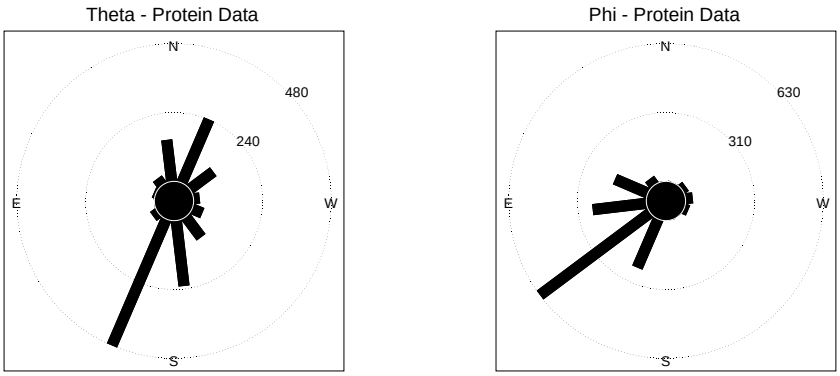

Figure 1: Circular Histograms for Protein Data

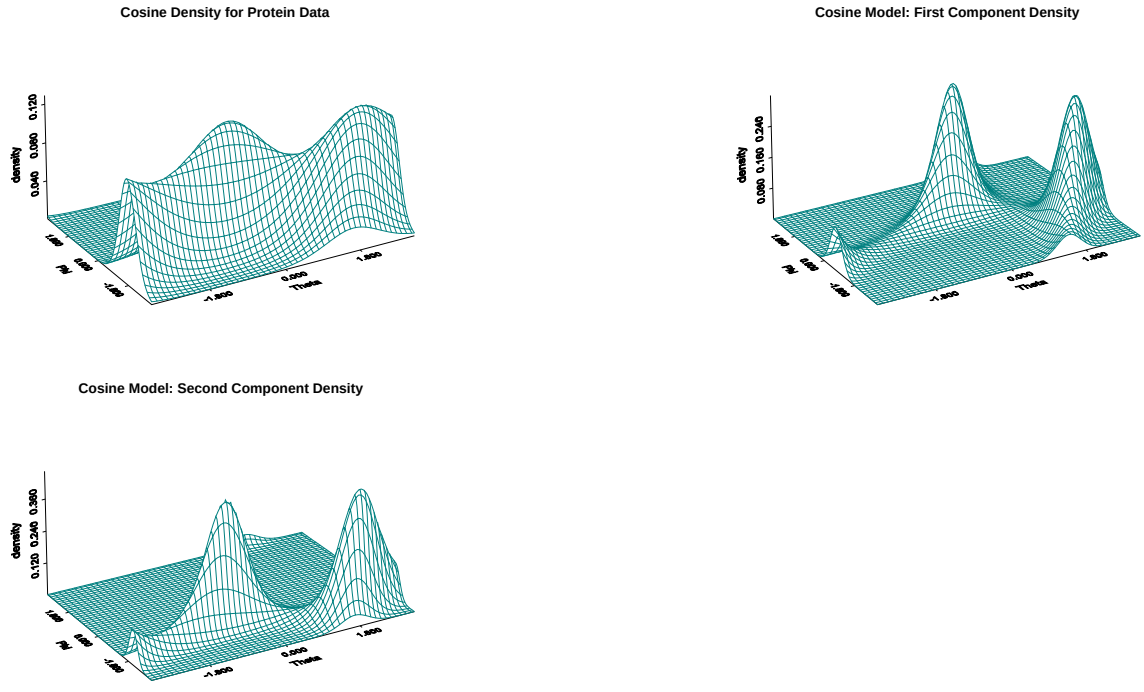

Figure 2: Surface plots for mixture distributions and the two clusters for Cosine Density

due to the normalizing constant, which is estimated by solving an integral. It easily converges to local optimal solutions. Higher component mixtures might pose convergence problems. The concentration parameters have higher uncertainty than the means. The challenge was higher with real data to get a good set of clusters that are well separated. Another issue is the start values, which is very critical. Further work is needed to determine better start values for the algorithm.

References

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. Wiley.

Mardia, K. V. (1975) Statistics of directional data (with Discussion) *J. R. Statist. Soc.* **B 37**, 349-393.

Mardia, K. V. and Glasbey, C. A. (2001). A penalized likelihood approach to image warping. *J. R. Statist. Soc.* **B 63**,

Mardia, K. V, Taylor, C. C. and Subramaniam, M. (2003). Application of Circular Distributions to Conformational Angles in Protein. *Proceedings of the 22nd LASR Workshop, edited by K.V. Mardia, R.G. Aykroyd & M.J. Langdon, 149-152. Leeds University Press, Leeds.*

Mardia, K. V. and Westhead, D. R. (2002). New major challenges in bioinformatics. *Proceedings of the 21st LASR Workshop, edited by K.V. Mardia, R.G. Aykroyd & P. McDonnell, 9-15. Leeds University Press, Leeds.*

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum-likelihood from Incomplete data via EM algorithm. *J. Royal Statist. Soc. Ser.* ***39****, 1-38.*

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* Wiley.

Titterington, D. M. Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley

Singh, H, Hnizdo, V. and Demchuk, E. (2002) Probablistic model for two dependent circular variables. *Biometrika* **89**, 719–723.