EM Approach for Fitting Gaussian Mixtures to Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper we apply the EM algorithm to fitting a Gaussian mixture model to data and investigate different aspects of the algorithm.

1 Basic EM on some sample data sets

In this section we apply the EM algorithm to identify the mixture components in three different data sets with different sizes. We implement the EM with a random initialization; to avoid underflow we computed the log likelihoods.

Using different number of mixture components for the three different data sets, we obtain Table 1. Table shows the log likelihood for each number of mixture components ($k = 1, \dots, 5$). For this table we used the convergence criterion in which we terminate the algorithm whenever the difference between two consecutive log likelihoods is less than 10^{-5}. It is clear from Table 1 that increasing the number of components generally increases the log likelihood. We also observe that for large data sets we have much larger log likelihood values; this can be used as a justification of using an average log likelihood (which we use in the following sections).

	Data set 1		Data set 2		Data set 3	
K	LL (s)	LL (l)	LL (s)	LL (l)	LL (s)	LL (l)
1	-103.548	-510.245	-109.179	-535.404	-168.361	-853.712
2	-87.158	-479.133	-104.004	-366.528	-159.344	-838.811
3	-84.901	-475.087	-90.751	-362.540	-153.115	-833.157
4	-72.016	-469.831	-63.712	-359.657	-149.341	-809.200
5	-65.278	-462.274	-56.400	-351.959	-144.129	-805.090

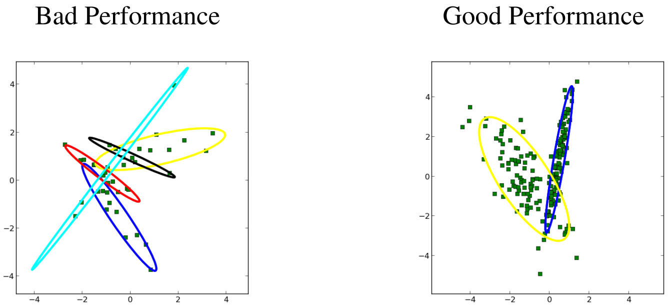
Table 1: The log likelihoods for large (l) and small (s) data sets for different number of mixture components (k).

We also tried different convergence criterion and different initializations. We observed that using a large convergence criterion can sometimes result in early termination of the algorithm and as a result lower log likelihood value. For instance, using convergence criterion 0.1 instead of 10^{-5} will result in log likelihood of -81.6286 as opposed to -72.016 that we get from convergence criterion of 10^{-5}. Using a initial value of mean vector $(0, 0)$ and identity covariance matrix, the results does not change significantly.

Having higher values for the log likelihood does not necessarily mean we have a better clusters. This point is clear in Figure 1.

2 Variations of EM on sample data sets

In this section we apply two variations to the EM algorithm and analyze the results. The first variation is using a diagonal covariance matrix. The difference with the original EM is basically



Bad Performance

Good Performance

Figure 1: The relatively good and bad performance of the EM algorithm. These results are for K=5 for the first small data (bad result) and K=2 for second large data set (good result).

in the model class that we are using. Now the responsibilities are computed with a similar formula with the difference that now the multivariate density is the multiplication of 2 independent normal densities (as the covariance between the two dimensions is zero now).

Now, for re-estimating the parameters; the same identity as the original EM can be used for the mean vectors and the mixing coefficients. However, for the covariance matrix we can have a simpler formula which is as follows:

$$\sigma_{k,1}^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_{n,1} - \mu_{k,1}^{new})^2$$

$$\sigma_{k,2}^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_{n,2} - \mu_{k,2}^{new})^2$$

where $\sigma_{k,1}^{new}$ is the new estimated **variance** of the points in the first dimension for the kth cluster (similarly we have that for the second dimension), $x_{n,i}$ is the ith dimension of the data, and $\mu_{k,i}$ is the mean of the ith dimension of data (estimated).

The results for using the diagonal covariance matrices are presented in Table 2. It is clear that we have generally lower log likelihood values; that is, due to the fact that diagonal covariance matrices are not suitable for the clusters in the three data sets. Figure 2 shows bad and good performances of the algorithm on two of the sample data sets. From these figure we can observe that having diagonal matrices does not help in these data sets.

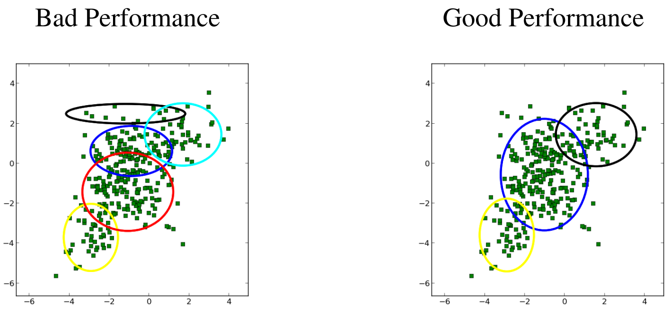
	Data set 1		Data set 2		Data set 3	
K	LL (s)	LL (l)	LL (s)	LL (l)	LL (s)	LL (l)
1	-105.047	-515.691	-109.636	-535.856	-179.235	-904.937
2	-96.989	-489.331	-102.711	-472.020	-170.144	-857.276
3	-93.020	-485.909	-79.244	-445.467	-156.543	-824.770
4	-86.661	-474.472	-73.336	-424.604	-154.548	-822.557
5	-66.278	-469.446	-67.529	-403.832	-148.747	-817.315

Table 2: The log likelihoods for large (l) and small (s) data sets for different number of mixture components (k) in an EM algorithm with diagonal covariance.

We also implemented the k-means for initializing the clusters. The resuls are presented in Table 3. We use the general covariance matrices for obtaining these results. It is clear from this table that we can have generally higher log likelihood compared to the case where we use diagonal covariance matrices. However, we can get comparable results with the random initialization.

3 Model Selection

To perform the model selection; a naive approach could be ranking the models based on their average log likelihood. In the previous section we presented the results for 5 different number of components and two different covariance structures (diagonal and general). The results of the ranking based on


Bad Performance

Good Performance

Figure 2: The relatively good and bad performance of the EM algorithm using diagonal covariance matrices. These results are for K=3 for the third large data (relatively good result) and K=5 for the third large data set (bad result).

	Data set 1		Data set 2		Data set 3	
K	LL (s)	LL (l)	LL (s)	LL (l)	LL (s)	LL (l)
1	-103.547	-510.245	-109.179	-535.404	-168.361	-853.712
2	-87.158	-479.133	-102.501	-366.528	-161.241	-838.811
3	-82.376	-473.704	-67.376	-362.130	-154.186	-827.920
4	-77.624	-465.883	-62.317	-354.027	-149.855	-807.532
5	-68.656	-465.639	-56.688	-351.959	-140.078	-803.945

Table 3: The log likelihoods for large (l) and small (s) data sets for different number of mixture components (k) in an EM algorithm with K-Means initialization.

the log likelihood (consequently the average log likelihoods) are presented in Tables 1 and 2. It is clear that using the average log likelihood will result in using the more complex model (higher k value). It is worth mentioning that we also tried initializing from K-Means with 5 different seeds and taking the average of the log likelihoods; the results of the ranking did not change.

A better approach is using cross-validation. In this section we use $K = 5$ and $K = 10$ fold cross-validation to choose the number of clusters. The results for $K = 5$ with a general covariance matrix is presented in Table 4. The top ranked k is consistent for the first two pair of small and large data sets; however, for the last data set we have inconsistent results. This issue may be solved by trying a different initialization.

	Data set 1		Data set 2		Data set 3	
K	LL (s)	LL (l)	LL (s)	LL (l)	LL (s)	LL (l)
1	-2.773	-2.568	-2.829	-2.691	**-2.905**	-2.865
2	**-2.563**	**-2.436**	-2.993	-1.881	-2.933	-2.833
3	-4.074	-2.446	**-2.413**	**-1.928**	-2.931	**-2.791**
4	-4.389	-2.497	-2.547	-1.946	-3.213	-2.802
5	-4.637	-2.559	-3.402	-2.004	-3.713	-2.799

Table 4: The average log likelihoods for large (l) and small (s) data sets for different number of mixture components (k) in an EM algorithm with general covariance matrix using 5-fold cross validation. Best K is in bold.

For a 10-fold cross validation we provided the results in Table 5. The top ranked k is consistent with that of 5-fold cross validation for the first and third data sets and also the second large data set. We can see from these two experiments that for larger data sets we tend to find more clusters.

We also repeated the cross validation experiments with a diagonal covariance matrix. The results of the 5-fold cross validation with a diagonal covariance matrix is provided in Table 6. Comparing the results with the 5-fold with a general diagonal matrix we see that in data set 2 we can have a higher number of components. This can be related to the fewer number of parameters in the model; that

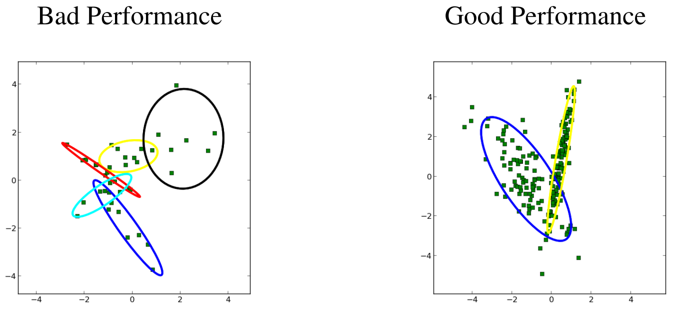


Figure 3: The relatively good and bad performance of the EM algorithm using K-Means initialization. These results are for K=5 for the first small data (bad result) and K=2 for second large data set (good result).

	Data set 1		Data set 2		Data set 3	
K	LL (s)	LL (l)	LL (s)	LL (l)	LL (s)	LL (l)
1	-2.977	-2.575	-2.865	-2.895	**-2.900**	-2.866
2	**-2.817**	**-2.471**	**-2.705**	-1.970	-2.926	-2.836
3	-3.541	-2.487	-3.046	**-1.966**	-3.005	**-2.785**
4	-3.614	-2.540	-3.192	-1.991	-3.099	-2.882
5	-3.828	-2.651	-3.219	-2.010	-3.256	-2.919

Table 5: The average log likelihoods for large (l) and small (s) data sets for different number of mixture components (k) in an EM algorithm with general covariance matrix using 10-fold cross validation. Best K is in bold.

is, in the absence of the correlation terms in the covariance matrix the model is now trying to fit the data with higher number of clusters.

	Data set 1		Data set 2		Data set 3	
K	LL (s)	LL (l)	LL (s)	LL (l)	LL (s)	LL (l)
1	-2.698	-2.593	-2.828	-2.693	**-3.055**	-3.029
2	**-2.681**	**-2.493**	-2.914	-2.600	-3.072	-2.889
3	-3.106	-2.497	-2.883	-2.348	-3.114	**-2.788**
4	-3.770	-2.506	**-2.793**	**-2.263**	-3.099	-2.798
5	-3.578	-2.745	-3.071	-2.276	-3.265	-2.800

Table 6: The average log likelihoods for large (l) and small (s) data sets for different number of mixture components (k) in an EM algorithm with diagonal covariance matrix using 5-fold cross validation. Best K is in bold.

We also tried the 10-fold cross validation for the model with diagonal covariance matrices. The general result that the model tends to have higher number of clusters seems to be valid in this case (similar to the 5-fold case).

4 Prediction for 3 different data sets

We use the cross validation, explained in the previous section, to choose the number of clusters and estimate the mixture parameters. We apply the 5-fold and 10-fold cross validations with diagonal and general matrices. The results are presented in Tables 7 and 8. As in the previous sections, we can see the tendency to having more clusters in the model with diagonal covariance matrix.

By choosing the model with the highest average likelihood and also by some exploratory analysis, we predict a model with a general covariance matrix and 2 mixture components for the first data set; a model with a general covariance matrix and 1 mixture component for the second data set, and, a model with a diagonal covariance matrix and 1 mixture component for the third data set. The plot for the predicted model is presented in Figure 4. We should mention as the average log likelihoods are not significantly different; the results may not be completely reliable; thus, further investigation is necessary.

	Mystery 1		Mystery 2		Mystery 3	
K	5-Fold	10-Fold	5-Fold	10-Fold	5-Fold	10-Fold
1	-3.289	-3.291	**-2.238**	**-2.244**	**-2.433**	**-2.417**
2	**-3.280**	**-3.279**	-2.359	-2.316	-2.533	-2.510
3	-3.357	-3.372	-2.949	-2.605	-2.809	-2.560
4	-3.362	-3.349	-3.356	-2.906	-3.210	-2.685
5	-3.503	-3.456	-3.439	-3.384	-3.281	-3.285

Table 7: The average log likelihoods for three mystery data sets for different number of mixture components (k) in an EM algorithm with general covariance matrix using 5-fold and 10-fold cross validation. Best K is in bold.

	Mystery 1		Mystery 2		Mystery 3	
K	5-Fold	10-Fold	5-Fold	10-Fold	5-Fold	10-Fold
1	-3.306	-3.317	**-2.688**	-2.650	**-2.417**	**-2.413**
2	-3.280	**-3.279**	-2.749	-2.574	-2.489	-2.477
3	**-3.278**	-3.334	-3.134	**-2.563**	-2.574	-2.548
4	-3.287	-3.280	-3.424	-3.009	-2.842	-2.618
5	-3.384	-3.305	-3.441	-3.303	-3.102	-2.763

Table 8: The average log likelihoods for three mystery data sets for different number of mixture components (k) in an EM algorithm with diagonal covariance matrix using 5-fold and 10-fold cross validation. Best K is in bold.

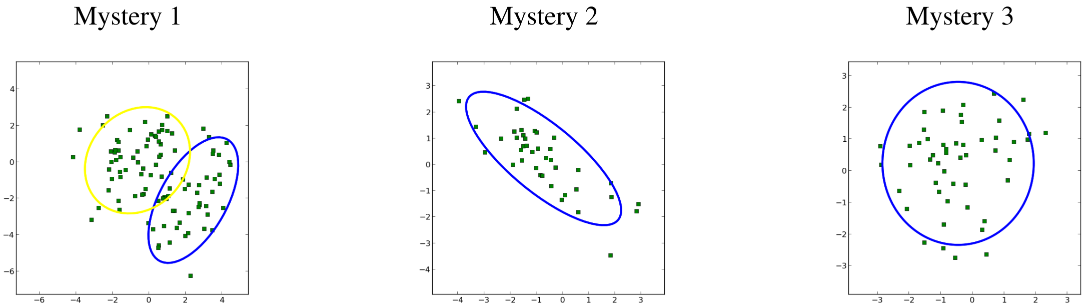


Figure 4: Predicted model for three mystery data sets based on the cross validation results.