Determination of a Sensible Number of Gaussians in a Gaussian Mixture Model

Stuart Heinrich
North Carolina State University

July 28, 2008

Abstract

Expectation Maximization is a well proven technique for converging to the local maximum solution of a maximum likelihood gaussian mixture model from an initial solution. In combination with Monte Carlo searching, it can be used to find a good global solution. However, automatic determination of the most appropriate number of mixtures is not well-defined, and while there have been many proposed solutions, none of them have proven satisfactory. This paper proposes a new method for determining the number of distributions in a gaussian mixture model based on a sequence of good partitionings. It is superior to existing internal criterions, and more robust to noise and highly eccentric distributions than mode-seeking techniques. In addition, it is efficient to compute on high dimensional multivariate models, and easily generalized to work with any other distribution functions which can be sampled – allowing it to be used as a general stopping rule for determining the number of clusters.

1 Introduction

In a gaussian mixture model, the data x_1, ..., x_n are assumed to be samples from a probability distribution $\phi(x)$ with density given by the summation of K multivariate normal distributions [1].

$$\phi(x) = \sum_{k=1}^{K} \alpha_k f(x|\mu_k, \Sigma_k) \tag{1}$$

$$f(x|\mu, \Sigma) = 1/((2\pi)^{D/2}|\Sigma|^{1/2}) \exp(-([x-\mu]^T \Sigma^{-1} [x-\mu])/2) \tag{2}$$

1.1 Expectation Maximization

The likelihood of the observed samples x_1, ..., x_n is given by,

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \alpha_k f(x_i|\mu_k, \Sigma_k) \tag{3}$$

which can be maximized by the Expectation Maximization (EM) algorithm, given K and initial estimates for μ_1, ..., μ_k and Σ_1, ..., Σ_k. Thus, it converges to a local maximum of the likelihood function. This is done through recursive application of the Expectation (E) and Maximization (M) steps. In the expectation step, the probability that the generator y_i of each datum x_i was k (eg, the probability that x_i was generated by distribution k) is updated.

$$p(y_i = k|x_i) \leftarrow \frac{\alpha_k f(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \alpha_j f(x_i|\mu_j, \Sigma_j)} \tag{4}$$

In the maximization step, the distribution weights α_k, mean vectors μ_k, and covariance matrices Σ_k are updated.

$$p_k \leftarrow \frac{1}{n}\sum_{i=1}^{n} p(y_i = k|x_i) \tag{5}$$

$$\mu_k \leftarrow \frac{\sum_{i=1}^{n} p(y_i = k|x_i)x_i}{\sum_{i=1}^{n} p(y_i = k|x_i)} \tag{6}$$

$$\Sigma_k \leftarrow \frac{\sum_{i=1}^{n} p(y_i = k|x_i)[x_i - \mu_k][x_i - \mu_k]^T}{\sum_{i=1}^{n} p(y_i = k|x_i)} \tag{7}$$

Note that the likelihood function is in general an increasing function with the value of K, so it cannot be used as an indicator of an appropriate K value [2].

1.2 Related Methods of Determining K

Determination of an appropriate number of distribution functions is a difficult and unsolved problem [7] [2], which can more generally be stated as the problem of determining the number of clusters in a data set. This is one of the most difficult and persistent problems in clustering [8].

A common technique for determining the number of clusters is by evaluating an internal criterion on a series of partitionings [5]. In order to facilitate this approach, data samples could be assigned into clusters c_k corresponding to their most likely distributions,

$$c_k = \{x_i | \max_j p(y_i = j|x_i) = k\} \tag{8}$$

and any clustering validity index (also known as an internal criterion or stopping rule) can be used as a method for attempting to determine the appropriate K value, by choosing the K that resulted in a partitioning with the maximum value of the validity index.

A survey of the quality of 30 internal criterions for their use in determining the number of clusters was performed by Milligna & Cooper [10] and later repeated by [13] indicated the best performing internal criterion to be the one used by Calinski & Harabsz [14], which is equivalent to choosing the partitioning that maximizes the F-Statistic

$$F = \frac{\sum_{j=1}^{K} n_j \left(\mu_j - \mu\right)^2}{K-1} \Big/ \frac{\sum_{j=1}^{K}(n_j - 1)\sigma_j^2}{n-K} \tag{9}$$

where n_j is the number of patterns in the jth cluster, σ_j is the standard deviation within the jth cluster, μ_j is the mean of the jth cluster, and μ is the grand mean. A disadvantage of the F-Statistic, as well as nearly all the other internal criterions, is that it is biased towards hyperspherical clusters.

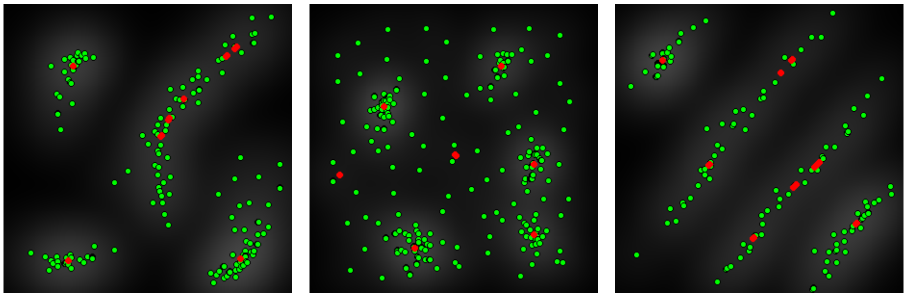
Other researchers have taken a more specific approach to determining the number of clusters. For example, Miloslavsky and van der Laan [7] attempt to minimize the distance between the mixture distribution and a kernel density estimate of the distribution using the Kullback-Leibler divergence function. They only considered univariate normal distributions.

Schlattmann [4] attempted to determine the number of distributions by pruning distributions that had weights less than some threshold $\alpha_k < \alpha_T$, or merging distributions whose means were closer than some threshold. This was combined with statistical bootstrapping, and the mean number of distributions was taken as the final value.

Similarly, Wang et. al [6] described a Merge & Split EM algorithm by which they started with an overestimated number of distributions and recursively merged or split distributions according to certain heuristic conditions. As a splitting function they used the Kullback-Leibler divergence function to measure the agreement between component distribution functions and a local kernel density estimate. As a merging criterion, they used a threshold on the distance between means.

A weakness of the above approaches is a dependence on thresholds and the use of merge criterion that ignores the covariance matrix. Also, none of the above proposed methods showed significantly promising results. A more promising method was proposed by Paclik [11], who used the modes of a kernel density estimate to determine the number of distributions and initialize them.

This appeared to be more successful than other methods discussed so far. Additional testing of this method showed it to work fairly well with automatically determined smoothing kernel bandwidths when there is a low percentage of noise. However, it finds extraneous modes in the presence of noise, and was not found to work well on highly ellipsoidal distributions because the smoothing kernel has a hyperspherical support which tends to break up long clusters unless the density has a consistent gradient towards the mean (Fig. 1).



(a) (b) (c)

Figure 1: Mode detection does not work well for highly ellipsoidal distributions, or in the presence of noise.

2 Automatic Determination of Number of Clusters

The proposed algorithm is based on the observation that, given an optimal partitioning of a data set for a fixed number of clusters, a certain measure of the average divergence of the distribution of patterns in a cluster from the distribution assumed by the cluster model initially increases roughly exponentially as more clusters are used to represent the data set, but only up to the naturally perceived number of clusters in the data set (Fig. 2).

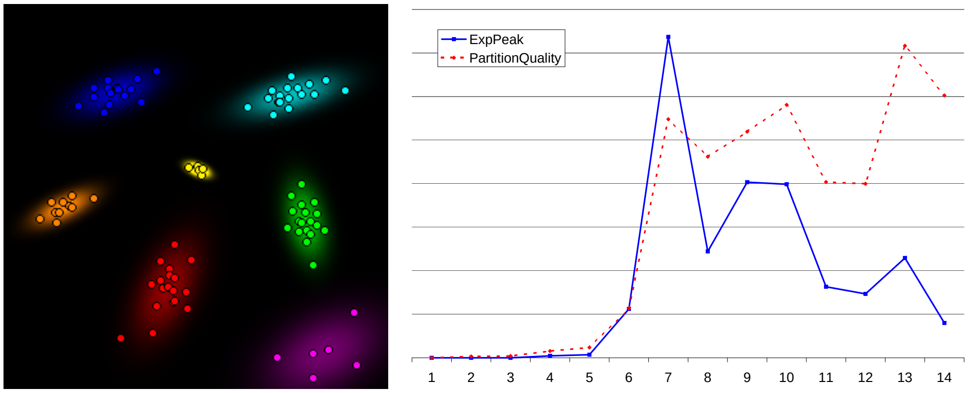
Therefore, an algorithm for determining the natural number of clusters was accomplished by creating partitionings P_k for $k = 1, 2, ..., \mathcal{K}_{max}$, evaluating those partitionings based on a heuristic measure of the divergence, and then using a robust method for detecting the last point of exponential increase in the heuristic as a function of k.

2.1 Heuristic Divergence Measure

The only method found to exhibit the exponential property described in the previous section was based on generating Monte Carlo samples according to the distribution and taking the average of the exponentiated distances to the nearest pattern in the cluster to the sample. Thus, partition quality was estimated by the following heuristic,

$$Q(k) = \frac{1}{ks} \sum_{j=1}^{k} \sum_{i=1}^{s} \min_{n \in C_j} ||x_i - n||^{-4} \tag{10}$$

where x_i are multivariate samples generated according to the distribution function of the cluster C_j. This heuristic serves as an estimate of the divergence of the data from the proposed distribution, by penalizing distributions for not containing patterns in areas predicted to have high probability.



(a) (b)

Figure 2: Example showing how the divergence measure of partition quality increases exponentially up until the perceived number of clusters. The method of exponential peak detection correctly identifies $K = 7$.

Several other measures of divergence were also evaluated including the Kullback-Leibler divergence with a kernel density estimate of the cluster distribution, but none performed as well as the aforementioned one for this purpose.

In general, it was found that $Q(k)$ tends to 0 as k diverges from the natural number of clusters in either direction, but the peak frequently occurred beyond the perceived number of clusters, making the detection of the last exponential point necessary.

2.2 Sampling the Multivariate Normal Distribution

The proposed divergence test requires drawing samples from a multivariate normal distribution, which is characterized by a mean vector μ and positive-definite covariance matrix Σ. This was accomplished by a transformation with the Cholesky decomposition A of the covariance matrix,

$$AA^T = \Sigma \tag{11}$$

Then, given a vector of normally distributed variables $Z = [z_1, z_2, ..., z_d]^T$, a single d-dimensional multivariate sample X is generated by

$$X = \mu + AZ \tag{12}$$

2.3 Robust Exponential Detection

A robust method for determining the last value K of $Q(k)$ belonging to an exponential curve starting with $Q(1)$ was developed,

$$K = \operatorname{argmax} k \left\{ \left(ae^{bk} - Q(k)\right)^2 \frac{Q(k)}{Q(k-1)} \right\} \tag{13}$$

where a and b are the parameters of an exponential regression of $Q(1), ..., Q(k)$ (the exponential regression was computed by a linear regression on the log-transformed data). Essentially, the error of the last residual is multiplied by the last ratio of increase. Because the regression requires at least 3 samples, it only works for $k > 2$.

Finally, K is increased while $Q(K + 1) > Q(K)$. This was found to improve the number of clusters found in some cases, and did not worsen the result in any of the tests.

3 Results

The proposed method for detecting the number of distributions was tested on a suite of hand-made 2D data sets that were designed to provide a range of easy and hard cases, some of which clearly do not fit the Gaussian distribution model (Fig. 3). The algorithm's performance in these cases serves as an indicator of its robustness. The true number of distributions is considered subjective.

The k-partitions were generated by running the EM algorithm initialized with the lowest SSE solution from a Monte Carlo k-means clustering, with $\mathcal{K}_{max} = 15$. The results are compared vs the F-statistic and mode finding on a kernel density estimate. A Gaussian support kernel was used with bandwidth equal to half the "rule of thumb" [12] bandwidth,

$$\hat{h}_{rot} = 1.06\sigma n^{-1/5} \tag{14}$$

4 Discussion

The F-statistic usually works well for hyperspherical clusters in the presence of noise, but fails when ellipsoidal clusters are introduced. The number of modes is a robust indicator that tends to over estimate the number of clusters, and is sensitive to outlying patterns. It tends to produce multiple modes nearby which could be merged using a distance threshold. It handles ellipsoidal clusters better than the F-statistic, but only when there is a distinct gradient towards the mean. The results of the newly proposed statistic appear subjectively superior.

References

[1] A.P.Dempster, N.M.Laird, D.B.Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, Vol. 39, pp.1-38, 1977.

[2] G. Celeux, G. Soromenho, "An Entropy Criterion for Assessing the number of clusters in a mixture model," *Journal of Classification*, Vol. 13, No. 2, pp.195-212, 1996.

[3] X. Zhuang, Y. Huang, K. Palaniappan, Y. Zhao, "Gaussian Mixture Density Modeling, Decomposition, and Applications," *IEEE Transactions on Image Processing*, Vol. 5, no. 9, 1996.

[4] P. Schlattmann, "Estimating the number of components in a finite mixture model: the special case of homogeneity," *Computational Statistics & Data Analysis*, Vol. 41, pp.441-451, 2003.

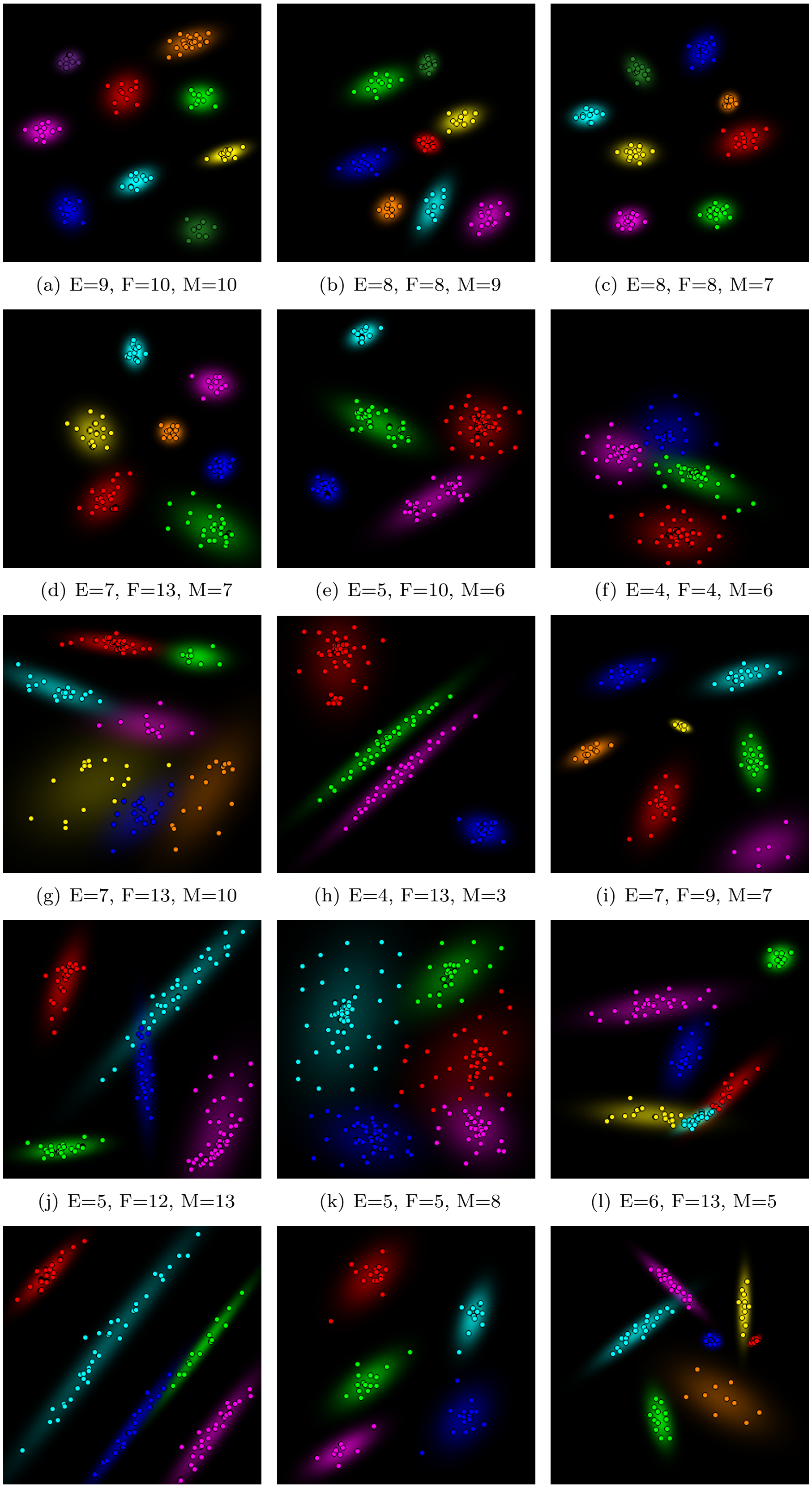
[5] A.K. Jain, M.N. Murty, P.J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–322, September 1999.

[6] H. X. Wang, B. Luo, Q. B. Zhang, S. Wei, "Estimation for the number of components in a mixture model using stepwise split-and-merge EM algorithm," *Pattern Recognition Letters*, Vol. 25, pp.1799-1809, 2004.

[7] M. Miloslavsky, M. J. van der Laan, "Fitting of mixture with unspecified number of components using cross validation distance estimate," *Computational Statistics & Data Analysis*, Vol. 41, pp.413-428, 2003.

[8] R. T. Ng, J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," *Proc. 20th Int. Conf. on Very Large Data Bases*, pp. 144-155. Santiago, Chile, 1994.

[9] G.W. Milligan, "A Monte Carlo study of thirty internal criterion measures for cluster analysis," *Psychometrika*, vol. 46, pp. 187–199, 1981.



(a) E=9, F=10, M=10

(b) E=8, F=8, M=9

(c) E=8, F=8, M=7

(d) E=7, F=13, M=7

(e) E=5, F=10, M=6

(f) E=4, F=4, M=6

(g) E=7, F=13, M=10

(h) E=4, F=13, M=3

(i) E=7, F=9, M=7

(j) E=5, F=12, M=13

(k) E=5, F=5, M=8

(l) E=6, F=13, M=5

(m) E=5, F=13, M=20

(n) E=5, F=5, M=5

(o) E=7, F=10, M=9

Figure 3: E = number of clusters found using exponential method, F = by F-statistic, M = number of modes in kernel density estimate.

[10] G.W. Milligan, M.C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, pp. 159–179, 1985.

[11] P. Paclik, J. Novovicova, "Number of Components and Initialization in Gaussian Mixture Model for Pattern Recognition"

[12] B. Turlach, "Bandwidth Selection in Kernel Density Estimation: A Review," *CORE and Institut de Statistique*

[13] Y. Shim, J. Chung, Choi, "A Comparison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm," *Computational Intelligence for Modelling, Control and Automation*, Vol. 1, pp.28-30, 2005.

[14] R. B. Calinski, L. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, Vol. 3, pp.1-27, 1974.