ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 99-106

On Convergence Problems of the *EM* Algorithm for Finite Gaussian Mixtures

Cédric Archambeau, John A. Lee, Michel Verleysen*

Université catholique de Louvain – Microelectronics Laboratory
Place du Levant, 3, B-1348 Louvain-la-Neuve, Belgium
Phone: +32 10 47 80 61, Fax: +32 10 47 25 98
E-mail: {archambeau, lee, verleysen}@dice.ucl.ac.be

Abstract. Efficient probability density function estimation is of primary interest in statistics. A popular approach for achieving this is the use of finite Gaussian mixture models. Based on the expectation-maximization algorithm, the maximum likelihood estimates of the model parameters can be iteratively computed in an elegant way. Unfortunately, in some cases the algorithm is not converging properly because of numerical difficulties. They are of two kinds: they can be associated to outliers or to repeated data samples. In this paper, we trace and discuss their origin while providing some theoretical evidence. As a matter of fact, both can be explained by the concept of *isolation*, which is leading to the width of the collapsing mixture component to become zero.

1 Introduction

Probability density function (PDF) estimation is a fundamental concept in statistics. It provides a natural way to investigate the properties of a given data set and perform efficient data mining. Areas of study, which have PDF estimation as a foundation, include machine learning, pattern recognition, neural networks, signal processing, computer vision, feature extraction, and many others.
When we perform density estimation three alternatives can be considered [6] [2]. The first approach, known as parametric density estimation, assumes the data is drawn from a specific density model. The model parameters are then fitted to the data. Unfortunately, an a priori choice of the PDF model is in practice not suited since it might provide a false representation of the true PDF.
An alternative is to build non-parametric PDF estimators, as for example the Parzen windowing PDF estimator [5]. The PDF is estimated by placing a well-defined kernel function on each data point and then determining a common width σ, also denoted as the smoothing parameter. In practice, Gaussian kernels are often used. The estimated PDF is defined as the sum of all the Gaussian kernels, multiplied by a scaling factor.

* Michel Verleysen is a Senior Research Associate of the Belgian F.N.R.S. (National Fund for Scientific Research).

This work was partially supported by the European Commission (grant IST-2000-25145 *OPTIVIP*) and the Belgian F.M.S.R. (project 3.4590.02).

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 99-106

By contrast to the previous methods, such techniques do not assume any functional form of the PDF and allow its shape to be entirely determined from the data.
A third approach consists in using semi-parametric models. As non-parametric techniques, they do not assume the a priori shape of the PDF to estimate. However, unlike the non-parametric methods, the complexity of the model is fixed in advance, in order to avoid a prohibitive increase of the number of parameters with the size of the data set. Finite mixture models are commonly used to serve this purpose. A popular technique for approximating the maximum likelihood estimate (MLE) of the underlying PDF is the expectation-maximization (EM) algorithm.

In this paper, we focus on the convergence problems encountered by EM while training finite Gaussian mixtures. In section 2, we recall the EM algorithm and its relevance for computing the model parameters of Gaussian mixtures. Next, we expose the convergence problems that might occur while using EM and clarify their origin. Finally, in section 4, we illustrate and discuss the numerical difficulties by means of an artificial example.

2 Finite Gaussian Mixtures

Finite mixture distributions [4] can approximate any continuous PDF, provided the model has a sufficient number of components and provided the parameters of the model are chosen correctly. The true PDF is approximated by a linear combination of *M* component densities:

$$p(\mathbf{x}) = \sum_{j=1}^{M} P(j)\, p(\mathbf{x}|j), \; M << N \qquad (1)$$

where *N* is the number of data, $p(\mathbf{x}|j)$ the probability of $\mathbf{x}$ given the component distribution *j* and $P(j)$ are the mixture proportions or priors. The priors are non-negative and must sum to one. In practice, Gaussian kernels are often used for the component densities:

$$p(\mathbf{x}|j) = \frac{1}{\left(2\pi\sigma_j^{\,2}\right)^{\frac{d}{2}}} \exp\left(-\frac{\left\|\mathbf{x}-\mathbf{c}_j\right\|^2}{2\sigma_j^{\,2}}\right), \qquad (2)$$

where $\mathbf{c}_j$ and σ_j are the centres and widths of the kernels respectively and *d* the dimension of the data space.
By applying the EM algorithm [3], the MLEs of the model parameters $P(j)$, $\mathbf{c}_j$ and σ_j can be computed iteratively while avoiding the complexities of a non-linear optimisation scheme. Let us define the likelihood function:

$$\mathcal{L} = \prod_{n=1}^{N} p(\mathbf{x}_n). \qquad (3)$$

Maximizing the likelihood function is then equivalent to finding the most probable PDF estimate provided the data set $\{\mathbf{x}_n\}_{n=1}^{N}$. In order to compute the MLE of the likelihood function the EM operates in two stages. First, in the *E-step*, the expected value of some "unobserved" data is computed, using the current parameter estimates

and the observed data. Here the "unobserved" data are the data labels of the samples. They correspond to the identification number of the different mixture components and specify by which one the data were generated. Subsequently, during the *M-step*, the expected values computed in the E-step are used to compute the MLE and the model parameters are updated. Each iteration step t can be summarized as follows [2] [1] [4]:

E-step:

$$P^{(t)}(j|\mathbf{x}_n) = \frac{p^{(t)}(\mathbf{x}_n|j) \cdot P^{(t)}(j)}{p^{(t)}(\mathbf{x}_n)}, \tag{4}$$

where $p^{(t)}(\mathbf{x}_n)$ and $p^{(t)}(\mathbf{x}_n|j)$ are computed according to equation (1) and (2) respectively.

M-step:

$$\mathbf{c}_j^{(t+1)} = \frac{\sum_{n=1}^{N} P^{(t)}(j|\mathbf{x}_n) \cdot \mathbf{x}_n}{\sum_{n=1}^{N} P^{(t)}(j|\mathbf{x}_n)}, \tag{5}$$

$$\sigma_j^{2\,(t+1)} = \frac{\sum_{n=1}^{N} P^{(t)}(j|\mathbf{x}_n) \cdot \left\| \mathbf{x}_n - \mathbf{c}_j^{(t+1)} \right\|^2}{d \sum_{n=1}^{N} P^{(t)}(j|\mathbf{x}_n)}, \tag{6}$$

$$P^{(t+1)}(j) = \frac{1}{N} \cdot \sum_{n=1}^{N} P^{(t)}(j|\mathbf{x}_n). \tag{7}$$

Note that in equation (4) $P^{(t)}(j|\mathbf{x}_n)$ corresponds to the posterior probability that $\mathbf{x}_n$ is generated by component j provided that the data point $\mathbf{x}_n$ is known.

3 Convergence Problems of the *EM* Algorithm

The convergence properties of the EM algorithm have been discussed in [8]. In theory, the EM is guaranteed to converge and it provides a MLE of the model parameters at a relatively fast convergence rate. However, in practice, the algorithm frequently fails due to numerical difficulties, especially when the available data is sparsely distributed in the input space. We have found that, in a number of cases, the variance of a kernel approaches zero, causing the EM to collapse (see section 4 for examples). This phenomenon appears in two situations:

1. Outliers occur in the database.
2. Data repetitions occur among the data samples.

The former was already partially traced by Yang and Chen [7]. Both numerical difficulties can be explained by the concept of *isolation* as will be described next.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 99-106

3.1 Effect of Outliers

First, let us consider we are nearby a local maximum of the likelihood function, where component j_0 is close to an outlier $\mathbf{x}_{out}$. Because of the exponentially decreasing nature of Gaussian functions, the latter is more likely to be generated only by j_0, the other components being far from the isolated sample, whereas the other data points are unlikely to be generated by it:

$$\forall n \neq out : P^{(t)}(j_0 | \mathbf{x}_n) << P^{(t)}(j_0 | \mathbf{x}_{out}) \approx 1. \qquad (8)$$

Therefore, the only data point contributing significantly to the computation of the centre of j_0 by equation (5) is the isolated sample:

$$\mathbf{c}_{j_0} \rightarrow \mathbf{x}_{out}, \qquad (9)$$

where "→" stands for "tends to".
Next, equation (6) can be rewritten in expanded form according to (2) and (4):

$$\sigma_{j_0}{}^{2(t+1)} = \frac{1}{d\sum_{n=1}^{N} P^{(t)}(j_0 | \mathbf{x}_n)} \sum_{n=1}^{N} \frac{P^{(t)}(j_0) \frac{1}{\left(2\pi\sigma_{j_0}{}^{2(t)}\right)^{\frac{d}{2}}} \exp\left(-\frac{y_n{}^{(t)}}{2\sigma_{j_0}{}^{2(t)}}\right) y_n{}^{(t+1)}}{p^{(t)}(\mathbf{x}_n)}, \qquad (10)$$

where:

$$y_n{}^{(t)} \equiv \left\| \mathbf{x}_n - \mathbf{c}_{j_0}{}^{(t)} \right\|^2. \qquad (11)$$

Suppose $\mathbf{x}_n$ is far away from $\mathbf{c}_{j0}{}^{(t)}$. Its contribution to the variance in (10) is then:

$$\lim_{y_n{}^{(t)} \rightarrow +\infty} \alpha \exp\left(-\frac{y_n{}^{(t)}}{\beta}\right) \frac{y_n{}^{(t+1)}}{\beta} =$$

$$\lim_{y_n{}^{(t)} \rightarrow +\infty} \alpha \exp\left(-\frac{y_n{}^{(t)}}{\beta}\right) \frac{y_n{}^{(t)}}{\beta} + \lim_{y_n{}^{(t)} \rightarrow +\infty} \alpha \exp\left(-\frac{y_n{}^{(t)}}{\beta}\right) \frac{\Delta y_n{}^{(t)}}{\beta} = 0, \qquad (12)$$

where α and β are positive scalars and $y_n{}^{(t+1)}$ can be rewritten as:

$$y_n{}^{(t+1)} = y_n{}^{(t)} + \Delta y_n{}^{(t)}. \qquad (13)$$

The first term of equation (12) is clearly zero in the limit. Seeing that we are close to the local maximum of the likelihood function, $\Delta y_n{}^{(t)} << y_n{}^{(t)}$ and therefore the second term also tends to be zero. Thus, when the centre of a kernel j_0 approaches an isolated data point $\mathbf{x}_{out}$, only the outlier contributes to the width (the other terms being zero), so that at the equilibrium the width will be approximately:

$$\sigma_{j_0}{}^2 \propto \left\| \mathbf{x}_{out} - \mathbf{c}_{j_0} \right\|^2 \rightarrow 0 \text{ as } \mathbf{c}_{j_0} \rightarrow \mathbf{x}_{out}. \qquad (14)$$

The isolation of a data point combined to the local character of Gaussian components makes the EM possibly collapse. Actually, because of their exponential shape it is more likely that an outlier is generated by a highly improbable isolated component than that it was generated by a component consistent with the database.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 99-106

3.2 Effect of Repetitions

The physical isolation associated to an outlier can be extended to the concept of *relative isolation* of repeated data points. Consider $\mathbf{x}_{rep}$ is repeated K times and it is not an outlier. Suppose we are close to a local maximum of the likelihood function where component j_0 is close to $\mathbf{x}_{rep}$. We can decompose equation (5) as follows:

$$\mathbf{c}_{j_0}^{(t+1)} = \frac{1}{\sum_{n=1}^{N} P^{(t)}(j_0|\mathbf{x}_n)} \left(K \cdot P^{(t)}(j_0|\mathbf{x}_{rep})\mathbf{x}_{rep} + \sum_{n\neq rep}^{N-K} P^{(t)}(j_0|\mathbf{x}_n)\mathbf{x}_n \right) \qquad (15)$$

Near the local maximum, it is more probable that $\mathbf{x}_{rep}$ was generated by component j_0 than $\mathbf{x}_{n\neq rep}$ was. Therefore:

$$\forall n \neq rep : P^{(t)}(j_0|\mathbf{x}_n) \leq P^{(t)}(j_0|\mathbf{x}_{rep}). \qquad (16)$$

In addition, the contribution of $\mathbf{x}_{rep}$ is multiplied by a positive factor K. Thus, looking at equation (5), the position of the centre of j_0 is meanly determined by the repeated sample:

$$\mathbf{c}_{j_0} \rightarrow \mathbf{x}_{rep}. \qquad (17)$$

Similarly, equation (6) can be reformulated as:

$$\sigma_{j_0}^{2\,(t+1)} = \frac{1}{d\sum_{n=1}^{N} P^{(t)}(j_0|\mathbf{x}_n)} \left(K \cdot \frac{P^{(t)}(j_0|\mathbf{x}_{rep}) \left\| \mathbf{x}_{rep} - \mathbf{c}_{j_0}^{(t+1)} \right\|^2}{p^{(t)}(\mathbf{x}_{rep})} + \sum_{n\neq rep}^{N-K} \frac{P^{(t)}(j_0|\mathbf{x}_n) \left\| \mathbf{x}_n - \mathbf{c}_{j_0}^{(t+1)} \right\|^2}{p^{(t)}(\mathbf{x}_n)} \right) \qquad (18)$$

If we are close to the local maximum, the first term in this sum is approximately zero. According to (16) the second term is small and it diminishes when the number of repetitions $K > 1$. As a result, the width of component j_0 is small with respect to the other mixture components. Next, suppose the width shrinks a little at iteration t. Clearly, because of the exponentially decreasing shape of the component, a width reduction will lessen the influence of the neighbouring data points on the computation of the corresponding mixture parameters. At the next computation of the E-step, the posterior probability that $\mathbf{x}_{n\neq rep}$ is generated by component j_0 will be reduced:

$$\forall n \neq rep : P^{(t)}(j_0|\mathbf{x}_n) \leq P^{(t+1)}(j_0|\mathbf{x}_n). \qquad (19)$$

Thus, at iteration t+1, the width of the collapsing component will decrease again, reducing even more the contribution of the neighbouring samples to the width computation. By snowball-effect the resulting width will be approximately:

$$\sigma_{j_0}^{2} \propto K \cdot \left\| \mathbf{x}_{rep} - \mathbf{c}_{j_0} \right\|^2 + \varepsilon_K \approx 0 \text{ as } \mathbf{c}_{j_0} \rightarrow \mathbf{x}_{rep}, \qquad (20)$$

where ε_K is the residual contribution of the neighbouring samples. As in the previous section, the numerical difficulty is due to the local character of the mixture components. However, by contrast to the outlier case where the algorithm converged to a stable local maximum of the likelihood function, in this situation it is not the case as illustrated in the next section.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 99-106

4 Experimental Results

As shown in figure 1, the EM algorithm might fail computing a consistent MLE of the parameters of a finite Gaussian mixture model when we are facing a data set containing an outlier (left column) or a repeated sample (right column). We have considered an artificial example containing a data set of 300 samples. The data samples $\{\mathbf{x}_n\}_{n=1}^{N=300}$ are realizations of a random variable χ_n, of which the PDF is the sum of two 2D uniform distributions. Five components were used in order to estimate the PDF of χ_n. The position of the centres (marked by a cross) and their respective widths are drawn for successive iteration steps, starting from initialisation till the collapsing of one component of the mixture. The repeated sample was chosen randomly among the data samples and repeated $K = 0.05 \cdot N$ times.

The log-likelihood function $\ln(\mathcal{L})$ is represented in figure 2 in function of the number of iterations. In the presence of an outlier (a) the algorithm converges to a stable local maximum, but gets trapped in a lower maximum than the optimal achievable MLE that can be attained in the regular case (d), i.e. when the database does not include any outlier or repeated sample. In fact, the resulting model corresponds to the MLE of a finite mixture of the 4 non-collapsing Gaussian components.

By contrast in the presence of a repeated sample, the EM seems first to converge to a poor local maximum involving a mixture of five components. Subsequently, when the collapsing of the closest component of the repeated sample has started, the likelihood increases rapidly. This is due to the significant contribution of the repeated sample to the likelihood, the estimated probability of the sample being close to unity. A similar increase is observed when the outlier is repeated K times as shown in figure 2 (c). Finally, remark that a sparsely distributed data set favours the birth of the collapsing process in the case of repetitions.

5 Conclusions

In this paper, we have clarified a number of numerical difficulties that might occur while computing the MLE of a Gaussian mixture model by the EM algorithm. Experimentally, it has been observed that when the data set includes outliers and/or repeated samples, the algorithm may fail. In section 3, we have suggested that, because of the local character of Gaussian kernels, some data points can be considered as isolated, leading the EM to collapse. Indeed, we have demonstrated that once a component of the mixture is nearby an isolated point, the corresponding centre will be strongly attracted by it. Meanwhile, because of the local character of the kernels, the neighbouring samples are not contributing to the width computation any more, leading, in the limit, to a zero value. As a consequence, the resultant component collapses and the iterative scheme crashes.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 99-106

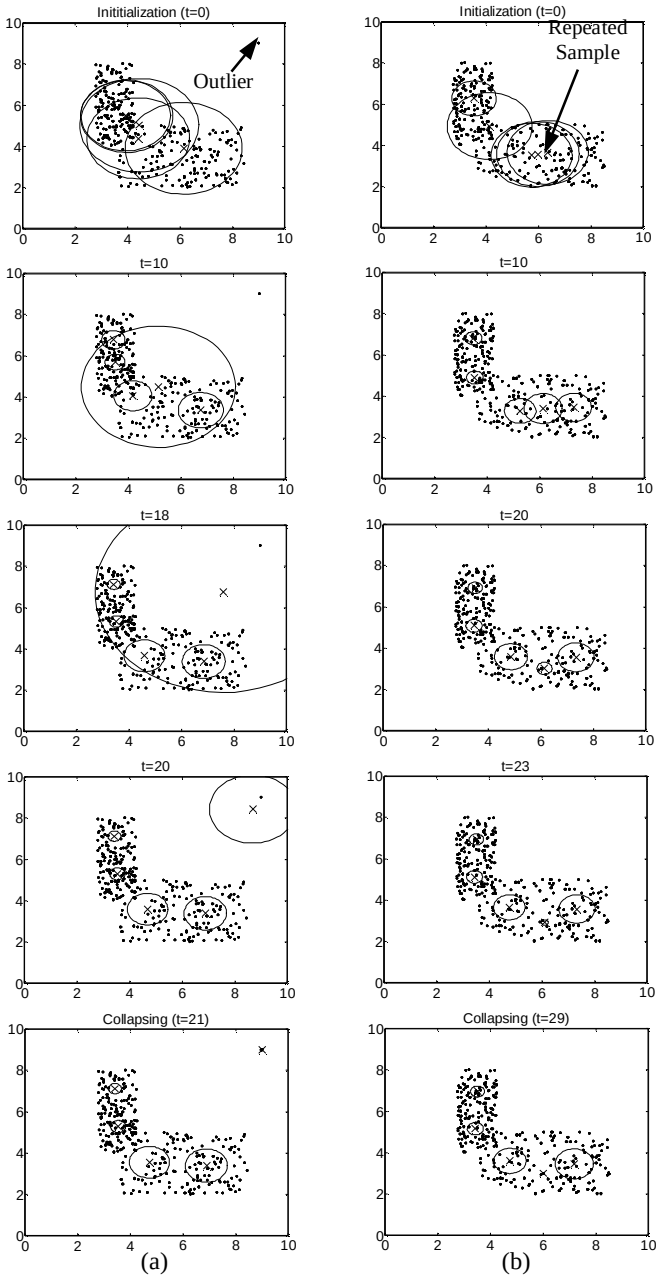

Figure 1: Computation of the MLE of a finite Gaussian mixture by the EM in function of the number of iterations in the presence of (a) an outlier in the database or (b) a repeated data sample. In both cases the algorithm is converging towards a local maximum non-representative of the data set and finally collapses.

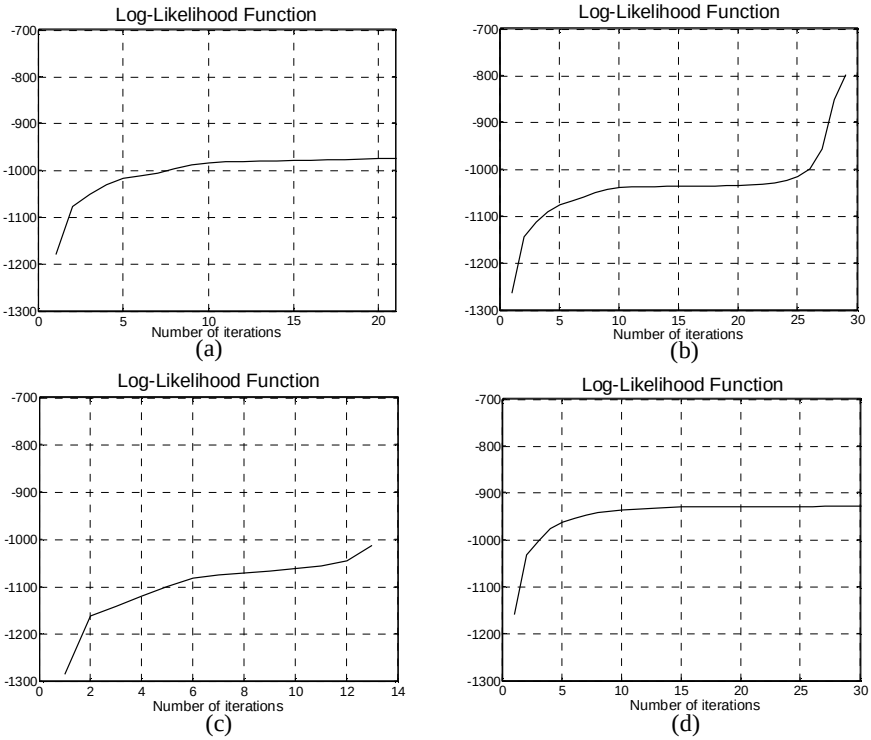

Figure 2: Log-likelihood function of the MLE of a finite Gaussian mixture model computed by EM in the presence of (a) an outlier in the database, (b) a repeated data sample or (c) both, and (d) for a regular data set.

References

[1] J. Bilmes: A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report of the International Computer Science Institute, Berkeley, CA (1998).

[2] C.M. Bishop: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995).

[3] A.P. Dempster, N.M. Laird, D.B. Rubin: Maximum Likelihood from Incomplete Data via the *EM* algorithm. J. Roy. Stat. Soc. (B), 39, 1–38 (1977).

[4] G. McLachlan and D. Peel: Finite Mixture Models. Wiley, New York (2000).

[5] E. Parzen: On Estimation of a Probability Density Function and Mode. Annals of Math. Statistics, 33, 1065–1076 (1962).

[6] B.W. Silverman: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986).

[7] Z.R. Yang, S. Chen: Robust Maximum Likelihood Training of Heteroscedastic Probabilistic Neural Networks. Neural Networks, 11, 739–747 (1998).

[8] L. Xu, M.I. Jordan: On Convergence Properties of the EM Algorithm for Gaussian Mixtures. Neural Computation, 8(1), 129–151 (1996).