MULTIVARIATE GAUSSIAN MIXTURE MODELING WITH UNKNOWN NUMBER OF COMPONENTS

N. Vlassis, A. Likas, B. Kröse

11-2000

Preprint no. 11-00/2000

Department of Computer Science
University of Ioannina
451 10 Ioannina, Greece

Multivariate Gaussian mixture modeling with unknown number of components

Nikos Vlassis Aristidis Likas Ben Kröse*

Abstract

We are dealing with the problem of Gaussian mixture density estimation with an unknown number of mixing components. In previous work we proposed the use of the weighted kurtosis as a test of normality of a component and developed an EM-based dynamic algorithm which, through component splitting, could approximately estimate the number of mixing components. In this paper we treat the multivariate case. We define the multivariate weighted kurtosis of a component and use it as a test of multivariate normality after a theorem of Mardia (1970). For dynamic component allocation we use the VDM theorem of Lindsay (1983) along with partial EM steps, ensuring the monotone increase of the log-likelihood in each step. We demonstrate our method in synthetic and real datasets and compare with the EM solutions with fixed number of components.

Keywords: Gaussian mixture density estimation, number of components, EM algorithm, test of normality, multivariate kurtosis, VDM algorithm.

1 Introduction

Finite mixture distributions [Titterington *et al.*, 1985] provide a simple framework for modeling population heterogeneity and can be particularly useful in cases where a single parametric model is inadequate. The idea is to model the distribution of a random vector by a weighted mixture of component models, each one parametrized on its own set of parameters. Formally, if we denote by $f(\mathbf{x};\boldsymbol{\phi}_j)$ the j-th component model parametrized on $\boldsymbol{\phi}_j$, then the mixture density for the random vector $\mathbf{x}$ assuming k components is

$$p(\mathbf{x}) = \sum_{j=1}^{k} \pi_j f(\mathbf{x};\boldsymbol{\phi}_j) \tag{1}$$

where π_j are the mixing weights satisfying

$$\sum_{j=1}^{k} \pi_j = 1, \quad \pi_j \geq 0. \tag{2}$$

The generality of the mixture models is reflected in the diversity of component models we can use, and recent examples are mixtures of factor analyzers [Ghahramani and Hinton, 1997], the generative topographic mapping [Bishop *et al.*, 1998], and mixtures of principal component analyzers [Tipping and Bishop, 1999].

*N. Vlassis and B. Kröse are with the RWCP, Autonomous Learning Functions SNN, Computer Science Institute, Faculty of Science, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands. Tel: +31-20-5257522, Fax: +31-20-5257490, E-mail: {vlassis,krose}@science.uva.nl .

A. Likas is with the Department of Computer Science, University of Ioannina, 45110 Ioannina, Greece. Tel: +30-651-97310, Fax: +30-651-48131, E-mail: arly@cs.uoi.gr .

The estimation of the parameters of the mixture, i.e., the parameter vector ϕ_j of each component, is often carried out with maximum likelihood and the EM algorithm [Dempster *et al.*, 1977], although other algorithms can also be used [Redner and Walker, 1984]. The increasing preference for using EM can be attributed mainly to its simple implementation together with its good global convergence characteristics and its capability of providing very fast a rough estimation of the model parameters.

One of the most difficult problems in mixture modeling is the specification of the number k of components, i.e., determining the complexity of the model. This *model selection* problem is a fundamental problem arising not only in mixture modeling but in most data modeling and analysis applications such as neural networks (estimating the number of hidden units in feed-forward networks), time series analysis (estimating the number and values of lag parameters), polynomial function approximation (estimating the degree of the fitting polynomial), etc. The difficulty in model selection problems stems from the fact that the estimation of the model order k (in our case the number of components) cannot be done through traditional optimization techniques, since k is a special parameter that assumes discrete values and, in addition, determines the dimensionality of the parameter space. Therefore, it is not easy to directly formulate and apply at each step update equations for k as is the case with the other parameters of the model.

To treat the model selection problem in Gaussian mixture modeling several techniques have been developed. Most of them employ statistical tests for determining the 'correct' value for k and apply these tests at appropriate points of the optimization process. Such known techniques are based on the likelihood ratio test [McLachlan, 1987], tests of homogeneity against mixture alternatives [Zelterman and Chen, 1988], or graphical techniques [Lindsay and Roeder, 1992]. The extension of these methods to multivariate data is however not always easy.

In [Vlassis and Likas, 1999] we proposed the use of kurtosis for deciding on the number of components in one-dimensional mixture problems. In this paper we extend the approach to the multivariate case. The idea is, assuming k components, to apply EM steps until convergence and then add one more component to the mixture placing it near a component that seems to be far from normal. In this context kurtosis is viewed as a measure of non-normality and the insertion of the new component is carried out in the neighborhood of the component whose *weighted* kurtosis is larger than a predefined threshold. This threshold is decided according to the distribution of kurtosis for high-dimensional data [Mardia, 1970].

The allocation of the new component is carried out according to the *vertex direction method* (VDM) [Lindsay, 1983] and a theorem that specifies the conditions that hold at the global maximum of the log-likelihood function. We propose the use of partial EM steps to search for the maxima of the new log-likelihood after component allocation, starting from solutions along the first PCA direction of the high-kurtosis component.

In the following we first outline the EM algorithm on Gaussian mixtures and then describe the idea of using the weighted multivariate kurtosis as part of a statistical test for deciding when a component exhibits non-normality. Next we propose a method for adding a new component to the mixture near the component with high kurtosis and according to the VDM theorem. The particular form of the mixture density after component allocation allows the use of partial EM steps which we describe next. Finally we demonstrate the use of the algorithm in synthetic and real datasets.

2 The EM algorithm for k-component Gaussian mixtures

We first describe briefly the EM algorithm for Gaussian mixtures with a fixed number of components k. Details can be found in several textbooks, e.g., [McLachlan and Krishnan, 1997]. A brief description is also given in [Vlassis and Likas, 1999].

A multivariate Gaussian mixture is defined as the weighted sum (1) with $f(\mathbf{x};\phi_j)$ the d-dimensional Gaussian density

$$f(\mathbf{x};\phi_j) = (2\pi)^{-d/2}|\mathbf{S}_j|^{-1/2}\exp\left[-0.5(\mathbf{x}-\mathbf{m}_j)^T\mathbf{S}_j^{-1}(\mathbf{x}-\mathbf{m}_j)\right] \quad (3)$$

parametrized on the mean $\mathbf{m}_j$ and the covariance matrix $\mathbf{S}_j$, collectively denoted by the parameter vector $\boldsymbol{\phi}_j$. We assume a training set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of independent and identically distributed points sampled from (1) and the task is to estimate the parameters of the mixture that maximize the log-likelihood

$$\mathcal{L} = n^{-1} \sum_{i=1}^{n} \log p(\mathbf{x}_i). \tag{4}$$

One of the algorithms often used for Gaussian mixture modeling is the Expectation-Maximization (EM) algorithm, a well-known statistical tool for maximum likelihood problems involving hidden or unobserved variables [Dempster *et al.*, 1977]. In the case of mixtures the unobserved variable can be regarded as the component each input point has been sampled from, and the EM update equations are [Redner and Walker, 1984]

$$\pi_j := \frac{1}{n} \sum_{i=1}^{n} P(j|\mathbf{x}_i), \tag{5}$$

$$\mathbf{m}_j := \frac{\sum_{i=1}^{n} P(j|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} P(j|\mathbf{x}_i)}, \tag{6}$$

$$\mathbf{S}_j := \frac{\sum_{i=1}^{n} P(j|\mathbf{x}_i)(\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T}{\sum_{i=1}^{n} P(j|\mathbf{x}_i)}, \tag{7}$$

where $P(j|\mathbf{x}_i)$ is the posterior probability that the point $\mathbf{x}_i$ was sampled from the component j and is given by the Bayes rule

$$P(j|\mathbf{x}_i) = \frac{\pi_j f(\mathbf{x}_i; \boldsymbol{\phi}_j)}{p(\mathbf{x}_i)} \tag{8}$$

where π_j and $\boldsymbol{\phi}_j$ the parameters of the mixture from the previous EM step. These estimates can be regarded as *weighted* sample averages, with the weights being the posterior probabilities $P(j|\mathbf{x}_i)$. It can be shown that in each EM step the log-likelihood cannot decrease [Ripley, 1996]. Details about the convergence properties of EM can be found, e.g., in [Redner and Walker, 1984].

2.1 EM, maximum likelihood, and model complexity

When trying to estimate the density of a given set of unlabeled data using a Gaussian mixture model with k components we must first impose appropriate bounds on the singular values of the covariance matrices of the components so that the log-likelihood of the model is bounded above. Then, let $\hat{\mathcal{L}}_k$ be the maximum of the log-likelihood of a k-component mixture model (for a given dataset) and let $\mathcal{L}_k$ be the log-likelihood of the solution reached after training the same model using the EM algorithm. Although EM ensures monotone increase of the log-likelihood in each step, it is a local optimization algorithm and may get trapped in local maxima in the parameter space, i.e., $\mathcal{L}_k < \hat{\mathcal{L}}_k$, often due to improper parameter initialization [Redner and Walker, 1984].

When the number of components is unknown, additional problems arise since we must also deal with the model complexity issue, i.e., how many components to use in order to maximize the log-likelihood of a given dataset. Let $\hat{\mathcal{L}}$ denote the *global* maximum of the log-likelihood for a dataset generated from a mixture of m components. Then, if the number of components is also treated as a model parameter, the maximum likelihood solutions using a k-component model with $k < m$ will be suboptimal in the sense that $\hat{\mathcal{L}}_k < \hat{\mathcal{L}}$ [Lindsay, 1983].

Therefore, an attempt to reach the global maximum of the log-likelihood for a given dataset in the space of all possible mixture models may fail either because of local maxima of EM or because of limited model complexity. To solve the problem one can think of the following approaches: i) fit a mixture model with a large number of components (e.g., in the order of the sample size) and at convergence remove those components with near zero mixing weights, or ii) fit several mixture models with increasing complexity and stop when the difference between the solutions $\mathcal{L}_{k+1}$ and $\mathcal{L}_k$ is not significant. The first approach suffers from the problem of overfitting (i.e., choosing too many components), while the second is computationally expensive.

The above discussion suggests that a sensible policy to reach (or get close to) the global maximum $\hat{\mathcal{L}}$ of the log-likelihood would be to allow a controlled flexibility in the complexity of the model while the EM algorithm evolves, i.e., permitting the number of components to change dynamically. In particular, it would be desirable to have an algorithm which, starting from a single component, adds components dynamically to the mixture to reach a solution with log-likelihood close to $\hat{\mathcal{L}}$. In order for such an approach to be applicable two main issues must be addressed:

- A statistical test is needed to decide whether to add a new component after convergence of the EM algorithm with k components.
- A methodology is required concerning the initial placement of the new component (with the other k components, e.g., remaining unchanged), ensuring that the log-likelihood of the resulting $(k+1)$-component mixture does not decrease.

These issues are discussed in the sections that follow.

3 The multivariate kurtosis and a test of normality

In earlier work we have proposed the use of kurtosis for testing whether a component fitted with EM is actually Gaussian or not, according to the distribution of the points in its vicinity. The approach was applicable to one-dimensional mixtures only and consisted of dynamic insertion of components, along with EM steps, until a global complexity measure was satisfied [Vlassis and Likas, 1999].

We now generalize this approach to multivariate Gaussian mixtures. For a random vector $\mathbf{x} \in \mathbb{R}^d$ with mean $\mathbf{m}$ and covariance matrix $\mathbf{S}$ its multivariate kurtosis is defined as [Mardia, 1970]

$$\beta = E[\{(\mathbf{x}-\mathbf{m})^T\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m})\}^2] \tag{9}$$

i.e., the average fourth power of the Mahalanobis distances from $\mathbf{m}$ with covariance $\mathbf{S}$. It turns out that if $\mathbf{x}$ follows d-variate Gaussian then its kurtosis equals $\beta = d(d+2)$, e.g., for $d=1$ we have $\beta = 3$. Moreover, to test the null hypothesis of normality against a multimodality alternative (assuming zero skewness) we have the following result.

Theorem 1 (Mardia 1970, Eq. 4.2) *Under the null hypothesis of normality the test statistic*

$$B = \frac{\beta - d(d+2)}{\sqrt{8d(d+2)/n}} \tag{10}$$

is distributed as $N(0,1)$ for large n.

The simple form of the asymptotic distribution of B allows the formulation of a test of normality by simply checking if $|B|$ is larger than a threshold corresponding to the right boundary of the critical region of the test [Papoulis, 1991, ch. 9]. Such a test could be, e.g., $|B| > 1.5$, since $N(0,1)$ is negligible outside the interval $[-3,3]$.

3.1 The weighted multivariate kurtosis

To apply the above test of normality to a component j of the mixture we rewrite kurtosis (9) as

$$\beta_j = \int_{-\infty}^{\infty} \{(\mathbf{x}-\mathbf{m}_j)^T\mathbf{S}_j^{-1}(\mathbf{x}-\mathbf{m}_j)\}^2 f(\mathbf{x};\boldsymbol{\phi}_j)d\mathbf{x} \tag{11}$$

where $f(\mathbf{x};\boldsymbol{\phi}_j)$ the Gaussian component j under the null hypothesis. We can use Bayes rule (8) to express $f(\mathbf{x};\boldsymbol{\phi}_j)$ in terms of the mixture $p(\mathbf{x})$ to get

$$\beta_j = \frac{1}{\pi_j}\int_{-\infty}^{\infty} P(j|\mathbf{x})\{(\mathbf{x}-\mathbf{m}_j)^T\mathbf{S}_j^{-1}(\mathbf{x}-\mathbf{m}_j)\}^2 p(\mathbf{x})d\mathbf{x} \tag{12}$$

We want to test the null hypothesis that the points $\mathbf{x}_i$ in the vicinity of component j follow Gaussian density against the alternative hypothesis of multimodality. Since all other components

are unaffected by the test, it is reasonable to assume that under the null hypothesis the $\mathbf{x}$ points are sampled from $p(\mathbf{x})$ and thus approximate the integral in (12) with the sum

$$\tilde{\beta}_j = \frac{1}{n\pi_j} \sum_{i=1}^{n} P(j|\mathbf{x}_i)\{(\mathbf{x}_i - \mathbf{m}_j)^T \mathbf{S}_j^{-1} (\mathbf{x}_i - \mathbf{m}_j)\}^2. \tag{13}$$

In the above formula we can substitute the mixing weights π_j from the last EM step (5) to arrive at the definition of the weighted kurtosis

$$\tilde{\beta}_j = \frac{\sum_{i=1}^{n} P(j|\mathbf{x}_i)\{(\mathbf{x}_i - \mathbf{m}_j)^T \mathbf{S}_j^{-1} (\mathbf{x}_i - \mathbf{m}_j)\}^2}{\sum_{i=1}^{n} P(j|\mathbf{x}_i)} \tag{14}$$

and this has to be approximately $d(d+2)$ under the null hypothesis of normality of $f(\mathbf{x}; \boldsymbol{\phi}_j)$. Thus, if n is large enough we can substitute $\tilde{\beta}_j$ from above into (10) and test the hypothesis of normality of $f(\mathbf{x}; \boldsymbol{\phi}_j)$ by checking if the resulting statistic B_j is within its critical region.

4 Dynamic component allocation

Having specified a test of normality for checking whether a component of the mixture fits adequately the data points in its vicinity, we can develop an algorithm which dynamically inserts new components in the mixture until the hypothesis of normality holds for all components. Our algorithm is based on the following result.

Theorem 2 (Lindsay 1983, Th. 4.1, 5.3) *Let $p_k(\mathbf{x})$ be a k-component mixture and $f_{k+1}(\mathbf{x}; \boldsymbol{\phi})$ a new component model outside the mixture with parameter vector $\boldsymbol{\phi}$. If*

$$D(\boldsymbol{\phi}) \equiv n^{-1} \sum_{i=1}^{n} \left\{ \frac{f_{k+1}(\mathbf{x}_i; \boldsymbol{\phi})}{p_k(\mathbf{x}_i)} - 1 \right\} \tag{15}$$

is a function of the parameter vector $\boldsymbol{\phi}$ of the new component then:

1. *At the global maximum of the log-likelihood holds*

$$\sup_{\boldsymbol{\phi}} D(\boldsymbol{\phi}) = 0. \tag{16}$$

2. *If for some $\boldsymbol{\phi}^*$ holds $D(\boldsymbol{\phi}^*) > 0$, then the log-likelihood cannot decrease if we add the component $f_{k+1}(\mathbf{x}; \boldsymbol{\phi}^*)$ to the mixture, with weight $a \in (0,1)$ so that the new mixture is*

$$p_{k+1}(\mathbf{x}) = a f_{k+1}(\mathbf{x}; \boldsymbol{\phi}^*) + (1-a) p_k(\mathbf{x}). \tag{17}$$

The first part of the theorem specifies the conditions that hold at the global maximum of the log-likelihood, namely, that the function $D(\boldsymbol{\phi})$ is everywhere in the parameter space negative or at most zero. The second part of the theorem is more useful in practice: it states that adding a new component to a mixture in the form (17) will always lead to an increase of the log-likelihood, unless we have already reached the global maximum $\hat{\mathcal{L}}$.

Proving the first part of the theorem for the mixture model (17) provides some insight to the problem. The difference between the log-likelihood after and before insertion of a component with parameter vector $\boldsymbol{\phi}$ is

$$\begin{aligned}
\Delta\mathcal{L} &= n^{-1} \sum_{i=1}^{n} \log p_{k+1}(\mathbf{x}_i) - n^{-1} \sum_{i=1}^{n} \log p_k(\mathbf{x}_i) \\
&= n^{-1} \sum_{i=1}^{n} \log \left\{ \frac{a f_{k+1}(\mathbf{x}_i; \boldsymbol{\phi}) + (1-a) p_k(\mathbf{x}_i)}{p_k(\mathbf{x}_i)} \right\} \\
&= n^{-1} \sum_{i=1}^{n} \log \left\{ a \left(\frac{f_{k+1}(\mathbf{x}_i; \boldsymbol{\phi})}{p_k(\mathbf{x}_i)} - 1 \right) + 1 \right\}.
\end{aligned} \tag{18}$$

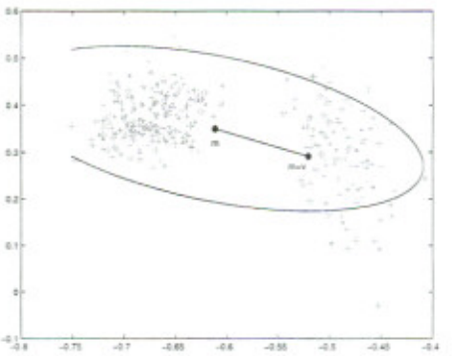

Figure 1: The multimodality of a component appears along its first PCA direction.

If we use the inequality $\log x \leq x - 1$ we get

$$\Delta\mathcal{L} \leq aD(\boldsymbol{\phi}) \tag{19}$$

which, since a is positive, states that the log-likelihood cannot increase (i.e., we reached the global maximum) if $D(\boldsymbol{\phi}) \leq 0$ for all $\boldsymbol{\phi}$ in the parameter space (i.e., $\sup_{\boldsymbol{\phi}} D(\boldsymbol{\phi}) = 0$).

Intuitively, the above theorem says that in order to increase the log-likelihood of our data set we must place a new component so that it fits as many input data as possible (numerator in (15)), which are at the same time inadequately fitted by the existing k-component mixture (denominator in (15)).

Based on the above theorem, an incremental algorithm called the *vertex direction method* (VDM) has been proposed in [Lindsay, 1983] for searching for the global maximum likelihood solutions. The idea is to find in each step the parameter vector $\boldsymbol{\phi}^*$ that maximizes $D(\boldsymbol{\phi})$ and then to estimate a in (17) that maximizes $\Delta\mathcal{L}$. The VDM algorithm can be shown to converge to the global maximum of the log-likelihood.

In practice, however, finding the maximum of the function $D(\boldsymbol{\phi})$ over the whole parameter space can be computationally very expensive, and gradient-based methods like quasi-Newton may get trapped in local maxima. Therefore we need a heuristic method to quickly search for maxima of $D(\boldsymbol{\phi})$, or equivalently, maxima of $\Delta\mathcal{L}$. Since $\mathcal{L}_k$ is constant, this maximization corresponds, in turn, to maximization of the log-likelihood $\mathcal{L}_{k+1} = n^{-1} \sum_i \log p_{k+1}(\mathbf{x}_i)$ of the $(k+1)$-component mixture (17).

A heuristic mechanism to search for this optimum is through the kurtosis test: we expect that a potential (local) maximum of $\mathcal{L}_{k+1}$ is near this component with large kurtosis. We notice that a common source of multimodality of a component appears along its direction of largest variance, i.e., along its first PCA direction (Fig. 1). This implies that a potential position for a new component added to the mixture is along (or near) this direction which is computed by an eigenvalue analysis of the covariance matrix of this component [Press *et al.*, 1992].

Moreover, the new mixture $p_{k+1}(\mathbf{x})$ in (17) can be regarded as a two-component mixture, the first component being the new added component $f_{k+1}(\mathbf{x}; \boldsymbol{\phi}^*)$ and the second component being the old mixture $p_k(\mathbf{x})$. Thus, *partial* EM steps can be used to find the parameters a and $\boldsymbol{\phi}^*$ which maximize $\mathcal{L}_{k+1}$ while leaving the parameters of $p_k(\mathbf{x})$ unchanged. This means applying (5), (6), and (7) only on the mixing weight a, the mean $\mathbf{m}^*$, and the covariance matrix $\mathbf{S}^*$ of the newly inserted component, i.e.,

$$a := \frac{1}{n} \sum_{i=1}^{n} P(k+1|\mathbf{x}_i), \tag{20}$$

$$\mathbf{m}^* := \frac{\sum_{i=1}^{n} P(k+1|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} P(k+1|\mathbf{x}_i)}, \tag{21}$$

$$\mathbf{S}^* := \frac{\sum_{i=1}^{n} P(k+1|\mathbf{x}_i)(\mathbf{x}_i - \mathbf{m}^*)(\mathbf{x}_i - \mathbf{m}^*)^T}{\sum_{i=1}^{n} P(k+1|\mathbf{x}_i)}, \tag{22}$$

with

$$P(k+1|\mathbf{x}_i) = \frac{a f_{k+1}(\mathbf{x}_i; \mathbf{m}^*, \mathbf{S}^*)}{a f_{k+1}(\mathbf{x}_i; \mathbf{m}^*, \mathbf{S}^*) + (1-a) p_k(\mathbf{x}_i)}. \tag{23}$$

Since only the parameters of the new component are updated, partial EM steps constitute a simple and fast iterative procedure for sensible component placement, obviating the need for expensive gradient-based non-linear optimization of $D(\phi)$.

4.1 The proposed algorithm

Based on the previous ideas we have developed an EM-based algorithm which dynamically adds components to the mixture until the log-likelihood gets no significant improvement and all the components are close to normal according to the kurtosis test.

The algorithm starts with one component. At some point suppose that k components have already been dynamically added to the mixture. We apply EM steps using the k-component mixture until the log-likelihood gets not significant improvement. At convergence we apply the kurtosis test (10) to all components and identify the component which maximally deviates from normality. If this deviation is smaller than a threshold then we terminate the algorithm with k components.

If this deviation is large enough we try to allocate a new component along (or near) the direction of the principal eigenvector of the covariance matrix of this component. We initialize the center $\mathbf{m}^*$, the covariance $\mathbf{S}^*$, and the mixing weight a of the candidate new component (as shown below), and then apply partial EM steps until convergence. We allocate the new component if the resulting log-likelihood is larger than the old one, i.e., if $\Delta\mathcal{L} > 0$, otherwise we terminate the algorithm. The whole algorithm can be summarized as follows:

1. Initialize using one component.
2. Repeat EM steps until $|\mathcal{L}^t \;/\; \mathcal{L}^{t-1} - 1| <$ CONV_THRES (e.g., 1e-6).
3. Compute the weighted kurtosis (14) and test statistic B (10) of all components. The number n in (10) for component j is approximated by $n\pi_j$.
4. Find the component c having largest $|B|$ and $n\pi_j >$ SIZE_THRES (e.g., 30).
5. If $|B_c| <$ KURT_THRES (e.g., 1.5) the algorithm terminates.
6. Initialize the center of the new component as

$$\mathbf{m}^* = \mathbf{m}_c \pm \sqrt{\lambda}(\mathbf{v}_c + 0.1\mathbf{w}), \tag{24}$$

   where λ the largest eigenvalue of $\mathbf{S}_c$, $\mathbf{v}_c$ the respective eigenvector, and $\mathbf{w}$ random perturbation vector sampled from a d-variate Gaussian.
7. Initialize the covariance matrix of the new component as $\mathbf{S}^* = 0.25\lambda\mathbf{I}_d$, where $\mathbf{I}_d$ the d-dimensional identity matrix.
8. Initialize its mixing weight as $a = 0.5$.
9. For each initial value of $\mathbf{m}^*$ from (24), apply partial EM steps (20)–(22) until convergence, and keep the solution with largest log-likelihood.
10. If $\Delta\mathcal{L} \leq 0$ the algorithm terminates, otherwise go to 2.

Since EM increases the log-likelihood in each step and our partial EM solutions are accepted only if $\Delta\mathcal{L} > 0$, the log-likelihood is monotone increasing in the course of the algorithm.

5 Experiments

In order to assess the performance of the proposed method we have conducted a series of experiments considering artificial (synthetic) as well as real datasets. Our interest is to evaluate the proposed algorithm in terms of its ability to determine the number of mixture components, and also to compare its performance with the EM algorithm with fixed number of components. The latter comparison was conducted because we noticed that the solutions obtained with the proposed algorithm (with k final components) were on average better than those obtained with the standard EM algorithm with k components. This is due to the fact that the original EM algorithm suffers from the problem of poor initialization, while our approach gradually increases the number of components and places them on appropriate positions. In all experiments with our algorithm (except for the phoneme dataset) the value of the kurtosis threshold was set equal to 1.5.

First, in order to illustrate the operation of the proposed algorithm we used a synthetic two-dimensional dataset with 500 points arranged in six clusters. Fig. 2 displays the evolution of the mixture model that starts with one component and through EM and dynamic component allocation provides a mixture model with six components each one placed at the corresponding cluster center. The contour lines around each cluster correspond to points in the input space with equal mixture density and give an impression of the covariance structure of each component.

In a second experiment we have evaluated our technique on a five-dimensional artificial dataset with 4000 points produced by a mixture of five Gaussians. The theoretical log-likelihood of the dataset was -5.918. The application of the proposed algorithm to this dataset led to the almost accurate recovery of the original mixture: $k = 5$ components were created with centers, covariance matrices, and mixing weights almost identical to the original ones. This is clearly indicated from the value of the log-likelihood of the obtained solution which was -5.902. We have also applied the fixed EM algorithm with $k = 5$ components. More specifically, a series of 15 experiments were conducted using the fixed EM algorithm with random component initializations. In five of the experiments the original mixture was approximately obtained (with log-likelihood -5.902). In one experiment a solution with three 'active' components was obtained, in the sense that the other two components had very small mixing weights and did not contribute essentially to the mixture ('dead' components). In the remaining nine experiments solutions with four 'active' components were obtained (one 'dead' component). The average log-likelihood value in the 15 experiments was -6.25.

We have also tested our method on two well-studied classification datasets (of two classes each): the *synthetic* dataset [Ripley, 1996] and the *phoneme* dataset [ELENA, 1995]. We used the class-independent mixture modeling approach where the training data of each class are separately used (using the proposed algorithm) to develop an independent model of each class density $p(\mathbf{x}|C_i)$, $i = 1, 2$. To classify a new point $\mathbf{x}$, the corresponding densities $p(\mathbf{x}|C_i)$ are computed and the class with maximum value is selected (according to the Bayes rule for classification and assuming equal class priors).

The synthetic dataset is a two-class two-dimensional dataset where each class has a bimodal distribution. Each class contains 125 training points and 500 test points. The application of our technique on the training data of each class led to mixtures with the correct number of components ($k = 2$) for each class. Fig. 3 displays the training data for each class and the solutions obtained with our technique. The application of the Bayes rule to the test dataset gave a very satisfactory classification error of 9%, whereas according to [Ripley, 1996] the best possible error is about 8%.

The phoneme dataset contains vowels coming from 1809 isolated syllables. Each vowel is characterized by five features and is classified either as Nasal (class 1) or as Oral (class 2). The total number of data points is 5404 and was divided in a training set with 2500 points and a test with 2904 points. The application of the proposed algorithm to the training data of class 1 resulted in a mixture of $k_1 = 10$ components and corresponding log-likelihood $\mathcal{L}_1 = -3.48$. Similarly, the application of the algorithm to the training data of class 2 gave a mixture with $k_2 = 3$ components and log-likelihood $\mathcal{L}_2 = -4.85$. The classification performance of the two mixtures on the test set was 84.1%. It must be noted that the value of the kurtosis threshold used in these experiments

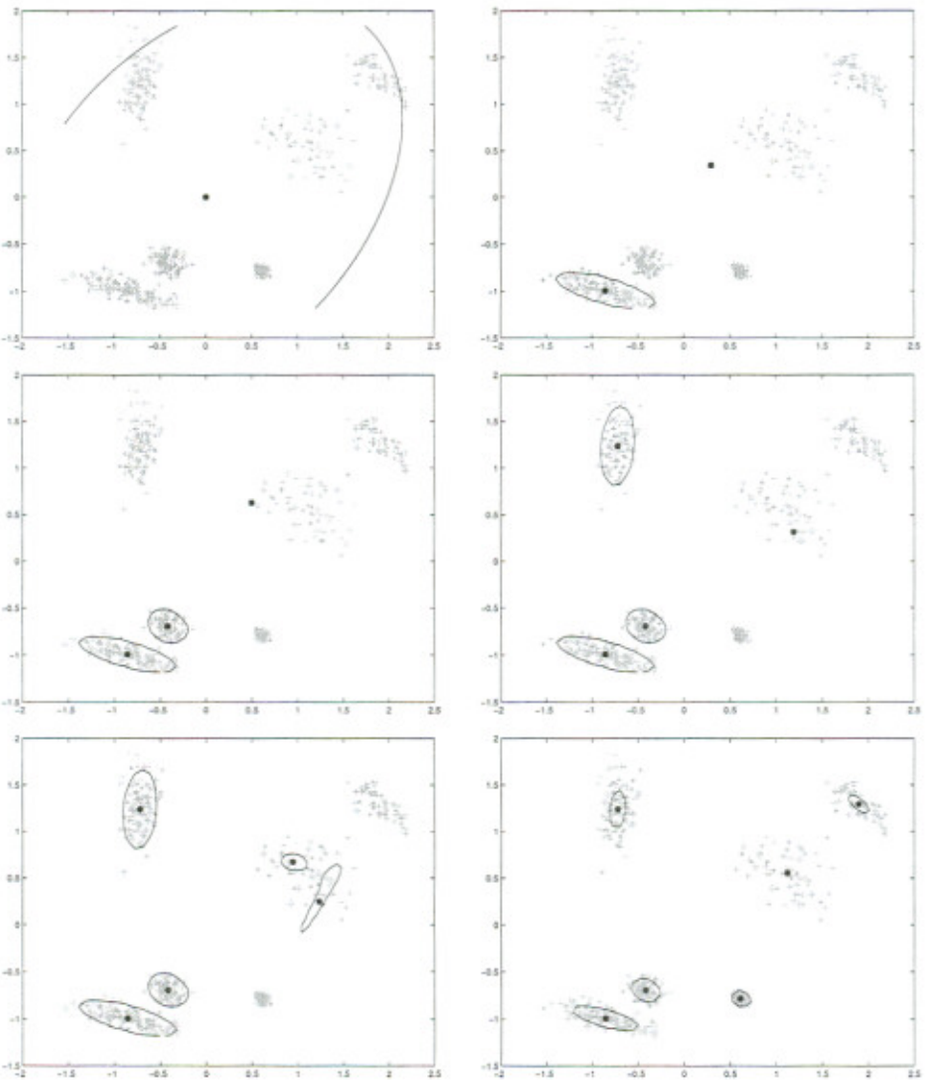

Figure 2: Mixture model evolution using our algorithm for an artificial two-dimensional dataset with six clusters. Each figure displays the data points, the center of each component, and the contours of the mixture density.

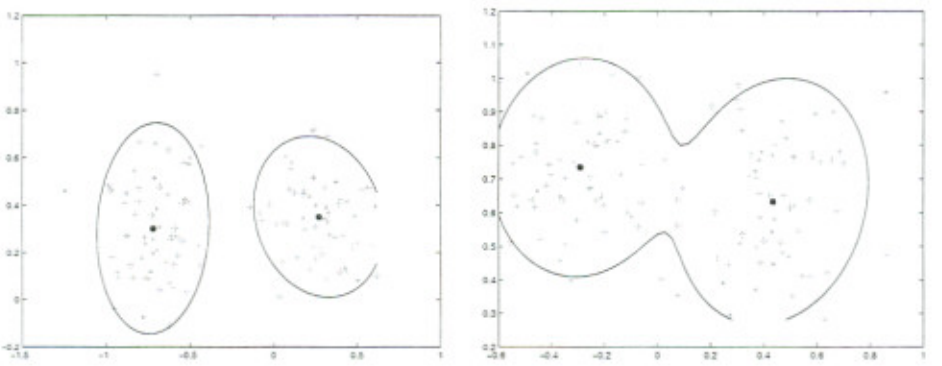

Figure 3: Results from the application of our algorithm to the training data of the two classes of the synthetic dataset. Each figure displays the data points, the center of each component and the contours of the density function.

was set equal to 3 to avoid the creation of excessive number of components.

In addition, we have conducted 15 experiments with the phoneme dataset by applying the EM algorithm with fixed number of components (and random initialization) on the training dataset of each class using the same number of components as that obtained using our technique, i.e., $k_1 = 10$ and $k_2 = 3$. The average log-likelihood values for the two training datasets were $\mathcal{L}_1 = -3.98$ and $\mathcal{L}_2 = -5.09$, while the average classification performance on the test set was 82.7%. From these experimental results it is clear that our algorithm is capable of discovering the correct number k of mixture components and in addition it provides better solutions compared to the EM algorithm with k components.

Another issue that must be mentioned is related with the robustness of the proposed algorithm when outliers are present in the dataset. Three cases can be distinguished: i) the outliers do not have any effect on the normal operation of the algorithm, ii) during EM or component allocation a component is placed on the location of an outlier, and iii) the kurtosis test fails due to the existence of outliers. In the third case there is no action that can be taken to treat the problem, apart from data preprocessing that identifies and removes outliers. The second case introduces no problem since the 'outlier' components have very small π_j values compared to the rest of the components and therefore can be easily identified and removed from the mixture.

A last issue concerns the determination of the kurtosis threshold which is in general user specified. In the case where a validation dataset is available, it is possible to adjust the threshold value by applying the algorithm with several threshold values and measuring the performance (for example the log-likelihood or the generalization error) of the resulting mixtures using the validation set.

6 Conclusions and discussion

We have presented a dynamic approach to multivariate Gaussian mixture modeling that treats effectively the model selection problem and in many cases provides solutions that are very close to optimal. Moreover we have shown experimentally that the solutions obtained using our technique (that allocates for example k components) are on average of better quality compared to those obtained using the EM algorithm with a fixed number of components k.

The effectiveness of the method can be attributed to two factors. The first is the multivariate kurtosis test of normality used to decide whether an optimal number of components have been allocated or additional component allocations are needed to improve the accuracy of density estimation. The second factor is related to the mechanism that allocates the new component to the mixture so that the log-likelihood of the new mixture always increases. This mechanism uses

principal component analysis and partial EM steps to adjust the parameters of the new component and has proven to be very effective.

A similar approach to Gaussian mixture modeling which imposes a certain degree of control over the complexity of the model has been recently proposed in [Ueda *et al.*, 2000]. The idea is to apply EM until convergence and then search for components that can be split or merged according to specific criteria. In this approach the number of components remains unchanged, while a form of backtracking is required to ensure that the log-likelihood increases in each split-merge step. The split test statistic employed in this approach is the Kullback divergence between a component density and the empirical density (e.g., through kernel smoothing) in the vicinity of the component. However, the asymptotic distribution of this statistic is not known and the critical region of the test must be set heuristically.

Future research will focus mainly on ways to further improve the effectiveness of the proposed method by trying several modifications to the method. For example, investigating alternative forms of the function $p_{k+1}(\mathbf{x})$ in (17) that specifies the way in which the new $(k+1)$-component mixture is related to the old k-component one. In addition, it may be possible to develop partial EM update schemes which adjust not only the parameters of the newly added component but also adjust the parameters of the large-kurtosis component.

Another direction of research is to abandon the kurtosis criterion as a criterion for deciding when and where to add a new component and proceed in a global search approach. For example, after EM convergence of a k-component mixture we could perform a search to identify whether there exist regions in the input space to place a new component and whether this initial placement causes a considerable increase in the log-likelihood of the new $(k+1)$-component mixture. If such a region is found, the new component is added, otherwise the algorithm terminates. However care must be taken in such an approach to efficiently deal with computational complexity issues.

References

[Bishop *et al.*, 1998] C. M. Bishop, M. Svensén, and C. K. I Williams. GTM: The generative topographic mapping. *Neural Computation*, 10:215–234, 1998.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977.

[ELENA, 1995] *ESPRIT Basic Research Project ELENA (no. 6891)*, 1995. Datasets available at `ftp://ftp.dice.ucl.ac.be/pub/neural-nets/ELENA/databases`.

[Ghahramani and Hinton, 1997] Z. Ghahramani and G.E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical report, University of Toronto, 1997. CRG-TR-96-1.

[Lindsay and Roeder, 1992] B. G. Lindsay and K. Roeder. Residual diagnostics for mixture models. *J. Amer. Statist. Assoc.*, 87(419):785–795, 1992.

[Lindsay, 1983] B. G. Lindsay. The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, 11(1):86–94, 1983.

[Mardia, 1970] K. V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

[McLachlan and Krishnan, 1997] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

[McLachlan, 1987] G. J. McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 36:318–324, 1987.

[Papoulis, 1991] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 3rd edition, 1991.

[Press *et al.*, 1992] W. H. Press, S. A. Teukolsky, B. P. Flannery, and W. T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 2nd edition, 1992.

[Redner and Walker, 1984] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, April 1984.

[Ripley, 1996] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, U.K., 1996.

[Tipping and Bishop, 1999] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[Titterington *et al.*, 1985] D. M. Titterington, A. F. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1985.

[Ueda *et al.*, 2000] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. SMEM algorithm for mixture models. *Neural Computation*, 2000. (to appear).

[Vlassis and Likas, 1999] N. Vlassis and A. Likas. A kurtosis-based dynamic approach to Gaussian mixture modeling. *IEEE Trans. on Systems, Man, and Cybernetics, Part A*, 29(4):393–399, July 1999.

[Zelterman and Chen, 1988] D. Zelterman and C.F. Chen. Homogeneity tests against central-mixture alternatives. *J. Amer. Statist. Assoc.*, 83(401):179–182, 1988.