THE KURTOSIS-EM ALGORITHM FOR GAUSSIAN MIXTURE MODELLING

N. VLASSIS, A. LIKAS

24-98

Preprint no. 24-98/1998

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

The Kurtosis-EM algorithm for Gaussian mixture modelling

Nikos Vlassis*

Department of Computer Science
University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
The Netherlands
E-mail: vlassis@wins.uva.nl

Aristidis Likas

Department of Computer Science
University of Ioannina
45110 Ioannina
Greece
E-mail: arly@cs.uoi.gr

Abstract

We address the problem of probability density function (pdf) estimation using a Gaussian mixture model updated with the EM algorithm. In order to circumvent the major drawback of the approach, i.e., the ignorance of the total number of mixing kernels, we define a new measure for Gaussian mixtures, called total kurtosis, which is based on the weighted sample kurtoses (fourth moments) of the kernels. This measure provides an indication of how well the Gaussian mixture fits the data. Then we propose the algorithm KEM (Kurtosis-EM) which monitors the total kurtosis at each step of the EM algorithm in order to decide dynamically on the correct number of kernels and possibly escape from local maxima. We show the superiority of our technique in approximating unknown densities through a series of examples with several pdf estimation problems.

1 Introduction

The Gaussian mixture model [16] has been proposed long ago as a general model for estimating an unknown probability density function (pdf). The virtues of the model lie mainly with its good approximation properties and the variety of estimation algorithms that exist in the literature [11, 16]. The model assumes that the unknown pdf can be written as a weighted finite sum of Gaussian kernels, with different mixing weights and different parameters, namely, means and covariance matrices. Then, depending on the estimation algorithm, an optimum

*Supported by the Real World Computing Partnership, Autonomous Learning Functions SNN Lab. SN1. This work was partially carried out while the author was with the Dept. of Electrical and Computer Engineering, National Technical University of Athens, Greece.

vector of these parameters is sought that optimizes some criterion. Most often, the estimation of the mixture is done by the maximum likelihood method, aiming at maximizing the likelihood of a set of samples drawn independently from the unknown pdf [12].

One of the first algorithms that were used for Gaussian mixture modelling was the Expectation-Maximization (EM) algorithm, a well-known statistical tool for maximum likelihood problems [8]. The algorithm provides iterative formulae for the estimation of the unknown parameters of the mixture, and can be proven to monotone increase in each step the likelihood of the input samples [11]. However, many researchers have argued about the efficiency of EM for Gaussian mixtures, since it requires an initialization of the parameter vector near the solution and assumes that the total number of mixing kernels is known beforehand.

To overcome the above limitations, we propose in this paper an extension to the original EM algorithm that we call *Kurtosis-EM (KEM)* algorithm. KEM is a dynamic EM algorithm that starts with a small number of kernels K (usually $K = 1$) and performs EM steps in order to maximize the likelihood of the data, while at the same time monitors the value of a new measure of the mixture, called *total kurtosis*, that indicates how well the Gaussian mixture fits the input data. This new measure is computed from the individual *weighted sample kurtoses*, or fourth moments, of the mixing kernels, a quantity defined in analogy to the weighted means and variances of the kernels and was first introduced in [19]. Based on the progressive change of the total kurtosis, KEM performs *kernel splitting* and increases the number K of kernels in the mixture. This splitting aims at making the absolute value of the kurtosis as small as possible.

By performing dynamic kernel allocation, KEM manages to find a good estimation of the number of kernels of the mixture, while it does not require any prior initialization near the solution. As experiments indicate, our approach seems to be superior to the original EM algorithm in approximating an unknown pdf. This fact renders it a good alternative for Gaussian modelling, especially in cases where little or no information about the pdf to be approximated is available in advance.

In the neural networks literature, a feed-forward network that implements a Gaussian mixture is the Probabilistic Neural Network [14]. The network uses one Gaussian kernel for each input sample, while the variance of each kernel is constant and known and the mixing weights are equal to the reciprocal of the total number of inputs. The network can be regarded as a distributed implementation of the Parzen windows method [9]. Some of the limitations of the original network model were relaxed in subsequent works [17, 1, 15, 13, 18], leading

to network models that implement some variations of the EM algorithm. However, in most approaches the number K of kernels of the mixture is considered known in advance, and it turns out that the automatic estimation of K is a difficult problem [16, 12].

Statistical methods or neural network models for estimating the number of kernels of a Gaussian mixture have been proposed in the literature [16, 7, 4, 13, 3]. However, the former usually cannot satisfy the necessary regularity conditions for estimating the asymptotic distributions of the underlying tests, and thus have to resort to costly heuristic techniques, e.g., Monte Carlo bootstrapping, for obtaining a solution.

In Section 2 we revise the use of Gaussian mixtures as models for probability density estimation, and we describe how the EM algorithm can be used for obtaining maximum likelihood solutions. Then, in Section 3 we present our method, the Kurtosis-EM algorithm. We first define the new measure of total kurtosis that is needed by the algorithm and then describe how KEM uses this measure for producing better solutions. We consider the univariate case only. Work is under progress to extend the definition and use of total kurtosis in higher dimensions. Section 4 gives experimental results from the application of the algorithm to pdf estimation problems, while Section 5 summarizes and gives hints for future research.

2 Gaussian mixture modelling with the EM algorithm

2.1 Gaussian mixtures

We say that a random variable x has a finite mixture distribution when the probability density function $p(x)$ of x can be written as a finite weighted sum of known densities, or kernels, $f(x)$. In cases where $f(x)$ is the Gaussian pdf we say that x follows a *Gaussian mixture*.

For the univariate case and for a number K of Gaussian kernels, the unknown mixture is written

$$p(x) = \pi_1 f_1(x; \mu_1, \sigma_1) + \cdots + \pi_K f_K(x; \mu_K, \sigma_K), \tag{1}$$

where $f_j(x; \mu_j, \sigma_j)$ stands for the univariate Gaussian $\mathcal{N}(\mu_j, \sigma_j)$

$$f_j(x; \mu_j, \sigma_j) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left[\frac{-(x-\mu_j)^2}{2\sigma_j^2}\right], \tag{2}$$

parameterized on the mean μ_j and variance σ_j^2. In order for $p(x)$ to be a probability density function with integral 1 over the input space, the additional constraint on the weights π_j of the mixture must hold

$$\sum_{j=1}^{K} \pi_j = 1, \quad \pi_j \geq 0. \tag{3}$$

The Gaussian mixture model is general and under regular conditions it may approximate every continuous function having a finite number of discontinuities (universal approximation theorem [5]). Also, compared to other pdf models, the Gaussian mixture model exhibits a number of attractive properties:

- It yields continuous and differentiatable functions.
- It assumes no discretization of the input space.
- It does not require the whole training set beforehand.
- Under appropriate conditions, it may also approximate nonstationary distributions.

For the estimation problem, we assume a training set X of n independent and identically distributed samples of the random variable x, taking values from an input space, e.g., in the univariate case the real line $\mathbb{R}$, and write

$$X = (x_1, \ldots, x_n), \quad x_i \in \mathbb{R}. \tag{4}$$

Training aims at finding the number of kernels K and the optimum vector θ^* of the $3K$ parameters of the mixture

$$\theta^* = (\pi_1^*, \mu_1^*, \sigma_1^*, \ldots, \pi_K^*, \mu_K^*, \sigma_K^*), \tag{5}$$

which leads to an approximation of the unknown pdf $p(x)$.

For the estimation of the optimum vector θ^*, we try to maximize the likelihood $p(X) = p(x_1) \cdots p(x_n)$ of the training set with respect to the unknown parameter vector θ. The point θ^* of the parameter space that maximizes the above function is called the *maximum likelihood estimate* and constitutes a solution to the problem:

$$\theta^* = \arg\max_{\theta} L(\theta), \quad L(\theta) = \prod_{x_i \in X} p(x_i). \tag{6}$$

Although efficient methods exist for the estimation of the $3K$ parameters of the mixture from a set of samples of x, the automatic estimation of K remains a difficult problem [12].

2.2 Maximum likelihood solutions with EM

The EM algorithm [2, 11, 8] is a powerful statistical tool for finding maximum likelihood solutions to problems involving observed and hidden variables. The algorithm applies in cases where we ask for maximum likelihood estimates for some observed variables, or observations,

but we do not know the exact form of their probability density function. Instead, we know the joint density of these variables and the hidden variables.

At each EM step the algorithm computes the quantity

$$Q(\theta|\theta^{(t)}) = E_Y[\log p(X, Y|\theta) \mid X, \theta^{(t)}] \tag{7}$$

which is a function of the parameter vector θ and which is obtained by averaging the logarithm of the joint density of X and Y, conditioned on θ, over the hidden variables Y, given the observations X and the current estimate of the parameter vector $\theta^{(t)}$ (E step). Then, a better estimate $\theta^{(t+1)}$ of the parameter vector is computed as the value that maximizes the quantity Q (M step). Alternating these two steps, the EM algorithm can be shown [11] to monotone increase the likelihood of the observations X, thus yielding an optimum θ^* in the maximum likelihood sense.

The problem of estimating an unknown Gaussian mixture by maximizing the likelihood of the parameter vector θ can be regarded as a problem with hidden variables and thus solved by using the EM algorithm. In this case, the hidden variables are the kernels the input samples statistically belong to, while each EM step provides an improved estimate of the parameters π_j, μ_j, and σ_j of each kernel j, $j = 1, \ldots, K$. These iterative formulae can be shown [15, 12] to be

$$\pi_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} P(j|x_i), \tag{8}$$

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} P(j|x_i)\ x_i}{\sum_{i=1}^{n} P(j|x_i)}, \tag{9}$$

$$\sigma_j^{2\,(t+1)} = \frac{\sum_{i=1}^{n} P(j|x_i)\ (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^{n} P(j|x_i)}, \tag{10}$$

$$P(j|x_i) = \frac{\pi_j^{(t)} f_j(x_i|\mu_j^{(t)}, \sigma_j^{(t)})}{p(x_i|\theta^{(t)})}. \tag{11}$$

where the index $t+1$ denotes the next (better) approximation of a parameter based on its previous value at t. It is easy to see that the priors π_j satisfy the condition $\sum_{j=1}^{K} \pi_j = 1$ after applying the above formulae for all kernels.

It is useful here to make a qualitative analysis of the above formulae. In each EM step we use the posterior probability $P(j|x_i)$ that a sample x_i belongs statistically to kernel j when the prior probability of the latter is π_j, and which is computed by applying the continuous version of the Bayes' theorem. This quantity $P(j|x_i)$ is computed for every kernel j from

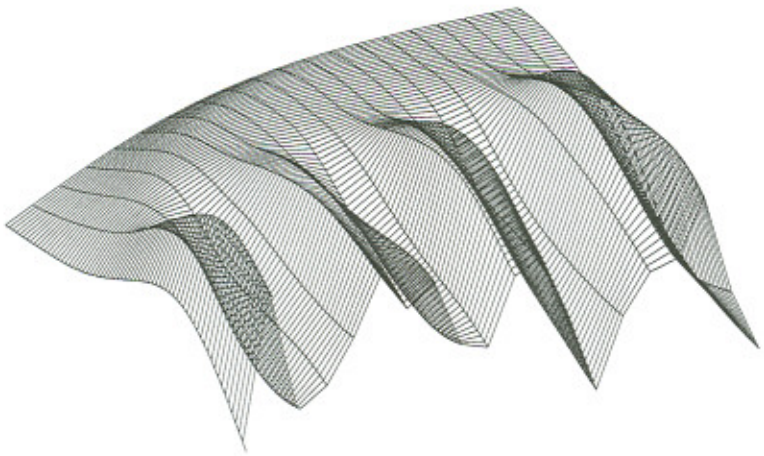

Figure 1: Likelihood function and local maxima.

the previous estimates of the parameters π_j, μ_j, and σ_j for this kernel and input x_i. This quantity is summed over all input samples x_i for the estimation of the new priors π_j (8), is summed weighted by the inputs x_i for the estimation of the new means (9), and is summed weighted by the square distances of the input samples x_i to μ_j for the estimation of the new variances σ_j^2 (10). In analogy to the formulae of the sample moments in statistics [20], the estimated parameters in each EM step of the EM algorithm can be regarded as *weighted sample moments* of the inputs X, with the weights being the posterior probabilities $P(j|x_i)$.

2.3 Convergence

In problems of Gaussian mixtures, due to the nonlinearities of the underlying densities with respect to the parameters μ and σ, the likelihood function exhibits almost surely local maxima or saddle points. In Fig. 1 it is shown a typical likelihood function for a problem of a Gaussian mixture, where a number of local maxima are evident. It is reasonable, therefore, to ask from a maximum likelihood method to escape from such local maxima and converge to the global maximum.

Although the EM algorithm monotone increases in each EM step the likelihood of the observations, it cannot ensure convergence of the parameter vector θ^* to the global maximum of the parameter space [11]. The algorithm may easily get stuck to a local maximum or saddle point. Moreover, so far we have assumed a known number of mixing kernels and thus have obtained the iterative solutions (8)–(10). In practice, though, this is hardly true: K is usually unknown and has also to be estimated from the input samples. These two constraints hamper

severely the efficiency of the EM algorithm.

Unfortunately, for the estimation of the number K of mixing kernels we cannot use the maximum likelihood method. Maximizing the likelihood with respect to the number of kernels implies choosing one kernel for each input sample of X, with the kernel mean coinciding with the sample. Apparently, such a solution would lead to a large number of kernels, equal to the cardinality of X, giving rise to overfitting problems.

Concluding, it appears that the EM algorithm for Gaussian mixtures suffers from the problems of the local maxima and the unknown number of kernels. In order to overcome these two problems on-line, i.e., while the algorithm evolves, we need a measure of the quality of the approximation at any instant, as an indicator how well the model fits the data. A bad fitting would necessitate a change to the parameter vector, or even an increase of the dimensionality of the parameter space. In the following we touch these issues.

3 The Kurtosis-EM algorithm for Gaussian mixtures

3.1 The total kurtosis measure

Having assumed that the random variable x follows a Gaussian mixture actually implies that the input samples x_i originate from K Gaussian kernels. Moreover, using the posterior probability (11) that a sample x_i statistically belongs to a kernel j, each input sample can be regarded as originating from one of the kernels j with probability $P(j|x_i)$.

The basic idea behind pdf estimation using Gaussian mixtures is that the distribution of samples in a local region can be approximated by a Gaussian kernel. Nevertheless, if the number of kernels is not adequate for a given pdf, this assumption may not be valid, and the approximation results may be poor. Thus, although the parameters of the kernels, i.e., means and variances, are correctly estimated from the weighted sample moments of the input samples (9)–(10), it is probable that the fit is not adequate. In Fig. 2 we show such a case: the input samples (vertical bars) follow a bimodal distribution while EM tries to fit them to a single Gaussian kernel. In this case, there is no way to use the first two moments (mean and variance) of the kernel to reveal this hidden multimodality.

To overcome the above inefficiencies, it is reasonable to resort to higher moments in order to decide whether kernel j fits adequately the samples lying in its vicinity. One indicator of this type is the *fourth sample moment*, called *kurtosis* (κ) that is defined for a kernel with

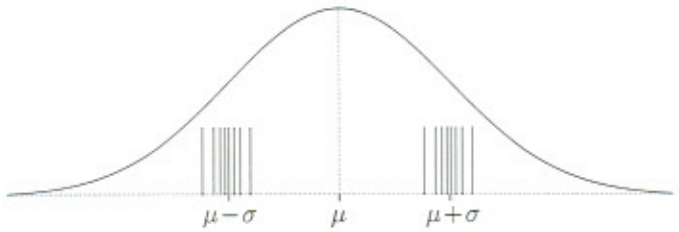

Figure 2: Wrong fitting: the EM algorithm tries to fit the input samples (vertical bars) to a Gaussian kernel, while the samples actually follow a bimodal distribution.

mean μ and variance σ as

$$\kappa = \frac{1}{n}\sum_{i=1}^{n}[(x_i - \mu)/\sigma]^4 - 3 \tag{12}$$

where the term 3 makes the quantity zero if the data follow the Gaussian distribution [10].

In analogy with the definitions of weighted sample moments (9)–(10), we define the *weighted kurtosis* of a Gaussian kernel j as the quantity

$$\kappa_j = \frac{\sum_{i=1}^{n} P(j|x_i)\left(\frac{x_i-\mu_j}{\sigma_j}\right)^4}{\sum_{i=1}^{n} P(j|x_i)} - 3. \tag{13}$$

Although difficult to prove analytically, experiments showed that in the case where the data originate from a Gaussian mixture, then the weighted kurtosis of each kernel of the mixture is approximately zero. Intuitively, we expect this to be true for a mixture of well-separated kernels where the posteriors are almost one for samples belonging to the correct kernel and almost zero for distant samples. In this case, (13) reduces to the original definition of kurtosis (12) for all kernels. On the other hand, if for some kernel j the distribution of the samples in its vicinity is non-Gaussian, the associated weighted kurtosis κ_j deviates from zero to a positive or negative number.

To test how large this deviation is for the whole mixture, a new measure is needed that weighs the deviation k_j of each kernel according to its importance π_j for the whole mixture. In this sense, we define a new quantity called *total kurtosis* as

$$K_T = \sum_{j=1}^{K} \pi_j |\kappa_j|, \tag{14}$$

which is a weighted sum over the individual kurtoses of the kernels of the mixture. The absolute values are needed to compensate for the individual kurtoses taking positive or negative values.

The total kurtosis K_T can be regarded as a measure of how well a Gaussian mixture fits the data, since a low value (near zero) indicates that each individual kernel fits naturally the samples in its vicinity, therefore the mixture constitutes a good approximation to the unknown pdf that generated the samples. On the other hand, a large value of the total kurtosis means that there are kernels that do not fit adequately their corresponding samples, or their parameters (mean and variance) are not properly adjusted.

The measure of kurtosis is important since the value of the likelihood alone does not provide much information regarding the effectiveness of the fit. For example, by using the EM algorithm we arrive at a solution of maximum likelihood for a specific number of kernels, but we cannot be sure whether the solution constitutes an acceptable approximation to the the unknown pdf; the two densities may differ significantly based on other distance measures [6]. On the other hand, we know that a lower bound for the total kurtosis is the zero value. Therefore, we can expect that the lower the total kurtosis value of the obtained solution is, the better is the approximation of the unknown pdf.

As an example, consider the approximation of two unknown pdfs using a Gaussian mixture with $K = 10$ kernels. The first pdf was a Gaussian mixture with four components, while the second was a uniform pdf. The maximization of the likelihood provided solutions with likelihood values -12202 and -11732 respectively. These values contain no information of how well the obtained solutions approximate the corresponding pdfs. As expected, the first solution accurately approximated the known Gaussian mixture, while the second solution was only a coarse approximation to the uniform pdf. The total kurtosis of the solutions was 0.03 and 0.31 respectively for the two cases. This means that the total kurtosis value revealed that the first approximation was accurate, while the second approximation was coarse. In general, although it is not easy to prove it theoretically, solutions of lower total kurtosis, provide better fit to the samples compared to solutions with higher total kurtosis. In the following section we propose a technique based on the EM algorithm that automatically increases the number of kernels based on the value of the total kurtosis of the mixture. We call this dynamic technique the *Kurtosis-EM (KEM)* algorithm.

3.2 The Kurtosis-EM algorithm

Based on the definition of the total kurtosis presented above, we have developed a new technique for pdf estimation based on Gaussian mixtures that exploits the EM algorithm for parameter estimation and automatically adjusts the number of kernels K using criteria based

on the total kurtosis of the mixture. The proposed algorithm is based on the idea that we try to maximize the likelihood by performing EM steps that in general lead to a decrease of the total kurtosis value.

More specifically, we start with a small number K of kernels (usually K is selected from one to three) and perform EM steps using the K kernels. These EM steps adjust the parameters of the kernels so that the likelihood is increasing and the total kurtosis is decreasing. This procedure is continued until either a local maximum of the likelihood is encountered or the total kurtosis reaches a minimum value and starts increasing. We distinguish between these two cases. In the first case a local maximum of the likelihood is also a local minimum of the total kurtosis, since no further update of the parameters is possible. If this happens, we check the value of the total kurtosis and, if it is sufficient low, we accept the solution, otherwise we consider that the solution is inadequate. Then, using the current local maximum parameters of the kernels, we create a new initial point for the EM algorithm by splitting one of the kernels in two as it will be described later in this section.

In the second case where the total kurtosis starts increasing without the likelihood having reached yet a local maximum, we consider that the EM algorithm has made its best in trying to fit each Gaussian to the given samples. Consequently, more kernels are needed to approximate the samples better. Therefore, if an EM step leads to an increase in the value of the total kurtosis, this is considered as the event that triggers the split of a kernel in two kernels in order to provide the capability for better approximation of the unknown pdf by the Gaussian mixture. After splitting, the $K + 1$ kernels continue to be updated at each step using the EM algorithm. In general, the splitting of the kernel leads to a decrease in the value of total kurtosis. Nevertheless, in some cases it is possible that, due to improper initialization of the two new kernels, the value of the kurtosis temporarily increases for some steps, until the kernels move to the right positions and the kurtosis starts decreasing again. For this reason, for a number of EM steps after a split, no further splits are permitted, since we allow the kernels to move to their appropriate positions even if this temporarily leads to an increase of the total kurtosis (of course, since EM steps are performed, the value of the likelihood always increases at each step).

Once we have decided when to perform kernel splitting by monitoring the value of the total kurtosis after each EM step, it remains to specify which kernel will be selected for splitting.

A deviation of the total kurtosis K_T from zero implies that the weighted kurtosis κ of one or more kernels deviates also from the zero value. Therefore, a reasonable selection criterion is

to split the kernel j that contributes most significantly to the high value of the total kurtosis, i.e., we select the kernel with the highest value of $\pi_j|\kappa_j|$. The two new kernels that are created have means equal to $\mu_j + \sigma_j$ and $\mu_j - \sigma_j$ respectively, and variances both equal to σ_j. Their priors are also set equal to $\pi_j/2$.

Finally, we must also specify termination criteria for the proposed method. Since the EM algorithm converges for fixed number of kernels, we must specify criteria for disabling kernel splitting in future steps. Consequently, the EM algorithm will converge to a maximum value of the likelihood. A criterion of this kind is a measure of the effectiveness of the split: we keep the value of the total kurtosis at the time of a split and the corresponding value at the time of the next split. If the difference is very small, we consider that splitting is no longer effective and from that time on, we keep the number of kernels fixed and perform EM steps until reaching a local maximum of the likelihood.

The algorithm just described, that dynamically adds kernels based on the value of the total kurtosis, is called *Kurtosis-EM*, or simply *KEM*. The attractive feature of KEM compared to EM is that KEM requires no initial knowledge about the number K of the kernels of the mixture. KEM starts using a small number of kernels, and adds kernels in the mixture dynamically, while the algorithm evolves. Moreover, it requires no initialization at a point θ which is already near the optimum solution, as is the case with EM, while, by dynamically increasing the number of kernels, the algorithm is capable of potentially escaping from local maxima of the likelihood function, and thus leading to a better approximation of the unknown pdf.

The KEM algorithm is summarized below. In the following description ϵ_1, ϵ_2 are user defined variables, *limit* denotes the number of steps after the last splitting during which a new splitting is not allowed, *nosteps* denotes the number of steps after the last splitting and K_{split} denotes the total kurtosis value at the time of the last split. Moreover if the variable *enableSplitting* is set equal to 1 then no further kernel splitting is allowed.

1. Initialization: Set the initial number K of kernels and initialize the parameters of the kernels (means and variances).

2. Compute the initial value of the total kurtosis K_T^{old} and the initial value of the likelihood L^{old}.

3. Set $nosteps := 0$, $K_{split} := K_T^{old}$, $enableSplitting := 1$.

   (a) Perform an EM step. Set $nosteps := nosteps + 1$.

(b) Compute the new value of the total kurtosis K_T^{new} and the new value of the likelihood L^{new}.

(c) Check for convergence: if $|L^{new} - L^{old}| < \epsilon_1$ goto step (f).

(d) If $(K_T^{new} > K_T^{old})$ and $(nosteps > limit)$ and $(enableSplitting = 1)$ then

- Perform kernel splitting.
- $K := K + 1$, $nosteps := 0$.
- $K_T^{old} := K_T^{new}$, $L^{old} := L^{new}$
- Compute K_T^{new}, L^{new}
- If $|K_{split} - K_T^{new}| < \epsilon_2$ then $enableSplitting := 0$.
- Set $K_{split} := K_T^{new}$

(e) Goto step (a)

(f) If (K_T^{new} is not small enough) and $(enableSplitting = 1)$ then

- Perform kernel splitting.
- $K := K + 1$, $nosteps := 0$.
- $K_T^{old} := K_T^{new}$, $L^{old} := L^{new}$
- Compute K_T^{new}, L^{new}
- If $|K_{split} - K_T^{new}| < \epsilon_2$ then $enableSplitting := 0$.
- Set $K_{split} := K_T^{new}$
- Goto step (a)

4. end

4 Examples

To assess the effectiveness of our approach we have conducted experiments with data drawn independently from known distributions, which in turn we tried to approximate using the KEM algorithm and the conventional EM algorithm. After training, we tested the accuracy of the obtained approximations with respect to the true pdf's.

In every problem considered, we have created a data set of $n = 5000$ points drawn independently from the corresponding pdf to be approximated. In all experiments the EM algorithm started with the means of the K kernels being uniformly distributed within the range of the data, while the deviance σ of each kernel was set equal to 0.5. On the other

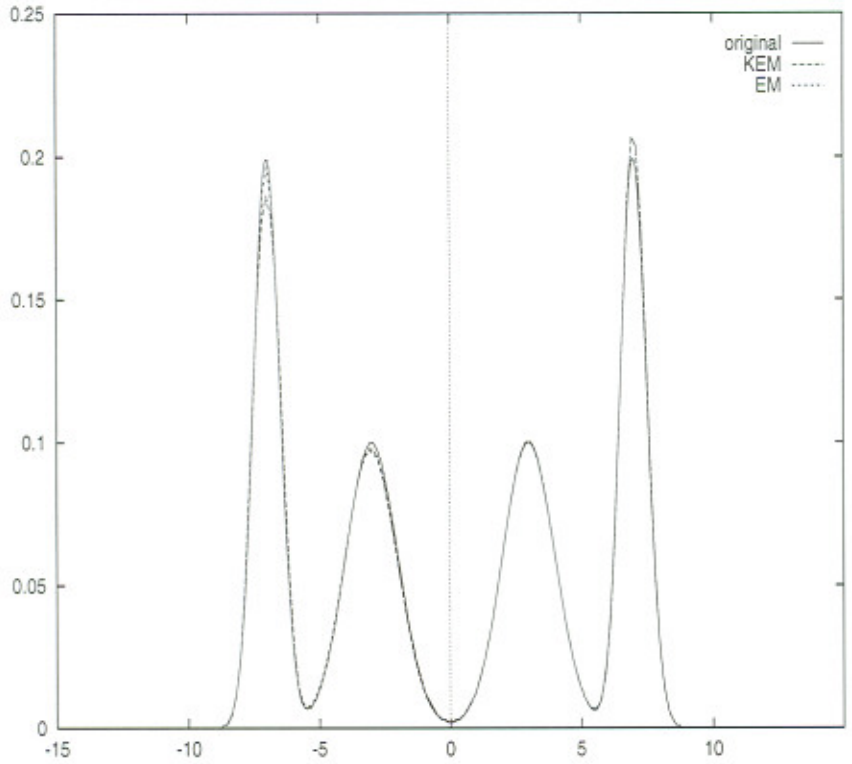

Figure 3: The approximation of a Gaussian mixture pdf with four kernels.

hand, the KEM algorithm always started with $K = 1$ kernel with mean in the center of the data range and σ also equal to 0.5.

We have considered four one-dimensional problems: i) a Gaussian mixture pdf with four kernels, ii) a a Gaussian mixture pdf with five kernels iii) a pdf with three uniform kernels and iv) a pdf with two Gaussian and two uniform kernels.

In all experiments, since the original pdf $g(x)$ is known, we could compute the theoretically optimal log-likelihood $\tilde{L}$ for the specific set of samples x_i $(i = 1, \ldots, n)$ drawn from the respective pdf:

$$\tilde{L} = \sum_{i=1}^{n} \log g(x_i) \tag{15}$$

4.1 Example 1

In this experiment we have generated samples using the following Gaussian mixture pdf:

$$g(x) = 0.25\mathcal{N}(-7, 0.5) + 0.25\mathcal{N}(-3, 1) + 0.25\mathcal{N}(3, 1) + 0.25\mathcal{N}(7, 0.5)$$

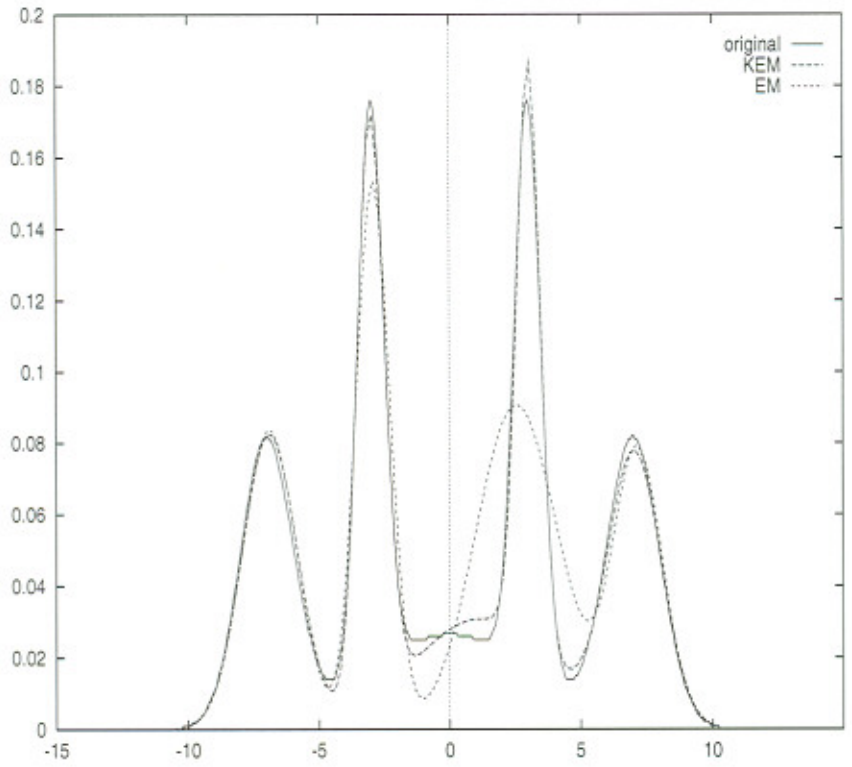

Figure 4: The approximation of a Gaussian mixture pdf with five kernels.

where $\mathcal{N}(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ. The value of the theoretical likelihood was $\tilde{L} = -12201$.

Fig. 3 displays the original pdf $g(x)$ as well as the obtained solutions using the KEM and the EM algorithms. The EM algorithm was applied on $K = 4$ kernels and provided accurate solution with $L = -12201.4$ and total kurtosis $K_T = 0.03$. The KEM algorithm was able to identify the correct number of kernels and provided an accurate solution with $K = 4$ kernels having $L = -12201.9$ and $K_T = 0.033$.

4.2 Example 2

In this experiment we have generated samples using the following pdf:

$$g(x) = 0.2\mathcal{N}(-7, 1) + 0.2\mathcal{N}(-3, 0.5) + 0.2\mathcal{N}(0, 3) + 0.2\mathcal{N}(3, 0.5) + 0.2\mathcal{N}(7, 1)$$

The value of the theoretical likelihood was $\tilde{L} = -13382.6$.

Fig. 4 displays the original pdf $g(x)$ as well as the obtained solutions using the KEM and the EM algorithms. The EM algorithm was applied on $K = 5$ kernels and was stuck

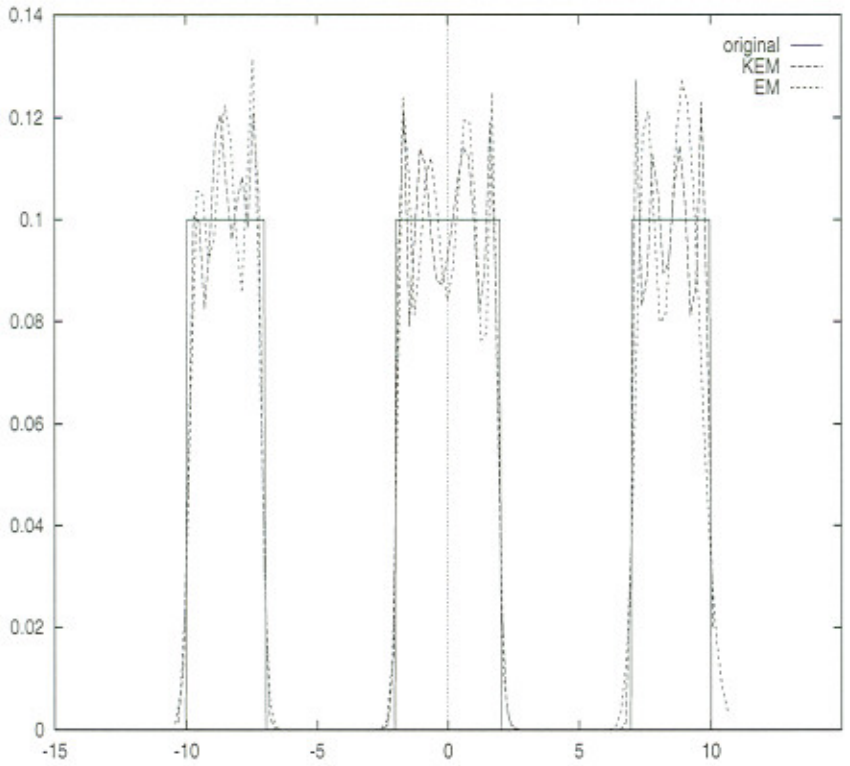

Figure 5: The approximation of a uniform pdf with three kernels.

in a local maximum with $L = -13718$ and total kurtosis $K_T = 0.35$. On the contrary, the KEM algorithm provided a much better solution with $K = 6$ kernels having $L = -13385$ and $K_T = 0.062$.

4.3 Example 3

In this experiment we have generated samples using a mixture of uniform kernels:

$$g(x) = \frac{1}{3}U(-10, -7) + \frac{1}{3}U(-2, 2) + \frac{1}{3}U(7, 10)$$

where $U(a, b)$ is the pdf of the uniform distribution in $[a, b]$. The value of the theoretical likelihood was $\tilde{L} = -11513.4$.

It is known that the Gaussian mixture method encounters difficulties in approximating uniform pdfs. This fact was confirmed by our experiments (Fig. 5), however the KEM algorithm was still able to give better solutions than EM. KEM yielded a solution with $K = 13$ kernels having $L = -11673$ and total kurtosis $K_T = 0.22$. EM, tested on $K = 13$ kernels,

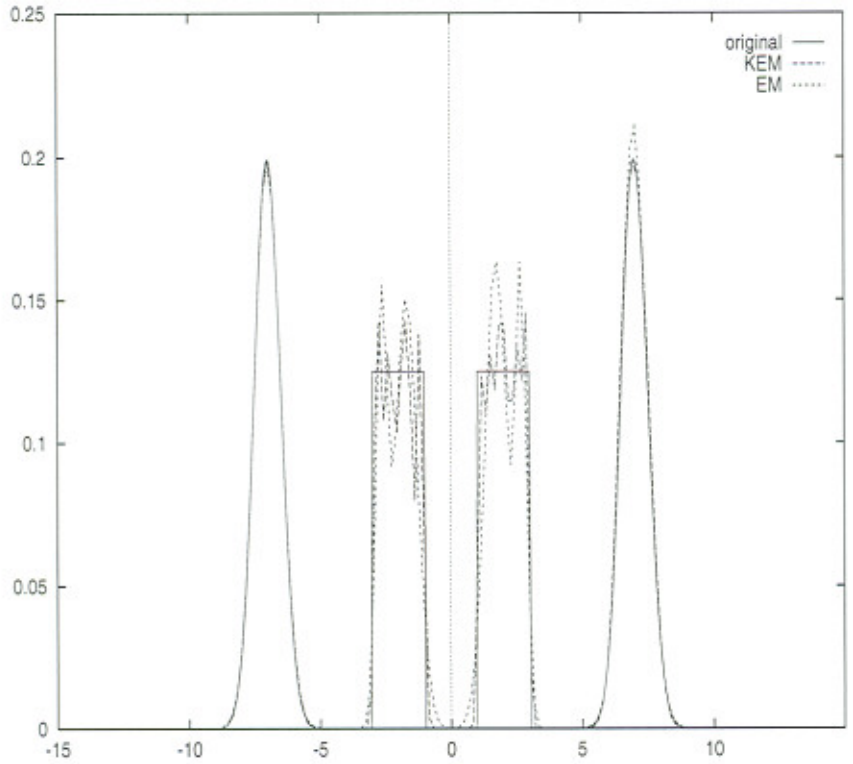

Figure 6: The approximation of a pdf with two Gaussian and two uniform kernels.

gave a worse solution compared to KEM: the likelihood value was $L = -11825$ and the value of the total kurtosis was $K_T = 0.36$.

4.4 Example 4

Finally we have conducted experiments with the following pdf consisting of two Gaussian and two uniform kernels:

$$g(x) = 0.25\mathcal{N}(-7, 0.5) + 0.25U(-3, -1) + 0.25U(1, 3) + 0.25\mathcal{N}(7, 0.5)$$

The value of the theoretical likelihood was $\tilde{L} = -10492$.

The obtained solutions are shown in Fig. 6. The KEM algorithm provided a solution with $K = 12$ kernels having $L = -10577$ and total kurtosis $K_T = 0.15$. The EM algorithm was also tested with $K = 12$ kernels, but again the obtained solution was worse compared to KEM: the likelihood value was $L = -10730$ and the value of the total kurtosis was $K_T = 0.31$.

As a conclusion, we can state that the dynamic addition of new kernels which is guided by monitoring the value of the total kurtosis makes the KEM algorithm an effective approach

that yields considerable improvement over the classical EM algorithm. Moreover, as shown in examples 1 and 2, KEM algorithm has the ability to approximately identify the number of Gaussian kernels of an unknown Gaussian mixture pdf, which is of major importance in many applications.

5 Conclusions

An approach has been presented for Gaussian mixture modelling which dynamically adjusts the number of mixing kernels. For this reason we have defined the total kurtosis measure as an indication of how well a Gaussian mixture fits the data. The proposed KEM algorithm constitutes an adaptation of the well-known EM algorithm that monitors the value of the total kurtosis and increases the number of kernels in the case where this measure starts increasing. The increase in the number of kernels is performed through splitting of the kernel that contributes more significantly to the value of the total kurtosis. In this sense the KEM algorithm proceeds by performing both likelihood maximization and kurtosis minimization. The increase in the number of kernels stops when no further progress in the minimization of the kurtosis seems possible. Experimental results on several test problems indicate that the KEM algorithm constitutes a promising dynamic alternative to the EM approach. In this work we have examined the univariate case. Current work focuses on a multi-dimensional definition of the weighted kurtosis (and the total kurtosis) and the application of the KEM algorithm to pdf estimation problems of higher dimensionality. Moreover, we aim at testing the effectiveness of the KEM algorithm on several applications where the EM approach has already been employed, e.g., classification, robotics etc.

References

[1] W.-S. Chou and Y.-C. Chen. A new fast algorithm for effective training of neural classifiers. *Pattern Recognition*, 25(4):423–429, 1992.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[3] O. Fambon and C. Jutten. Pruning kernel density estimators. In M. Verleysen, editor, *ESANN95-European Symposium on Artificial Neural Networks*, Brussels, Belgium, Apr. 1995. D facto publications.

[4] W. D. Furman and B. G. Lindsay. Testing for the number of components in a mixture of normal distributions using moment estimators. *Computational Statistics & Data Analysis*, 17:473–492, 1994.

[5] F. Girosi and T. Poggio. Networks and the best approximation property. *Biological Cybernetics*, 63:169–176, 1990.

[6] S. Ingrassia. A comparison between the simulated annealing and the EM algorithms in normal mixture decompositions. *Statistics and Computing*, 2:203–211, 1992.

[7] 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.

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

[9] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

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

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

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

[13] S. Shimoji. *Self-Organizing Neural Networks Based on Gaussian Mixture Model for PDF Estimation and Pattern Classification.* PhD thesis, University of Southern California, 1994.

[14] D. F. Specht. Probabilistic neural networks. *Neural Networks*, 3:109–118, 1990.

[15] R. L. Streit and T. E. Luginbuhl. Maximum likelihood training of probabilistic neural networks. *IEEE Trans. on Neural Networks*, 5(5):764–783, 1994.

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

[17] H. G. C. Tråvén. A neural network approach to statistical pattern classification by "semiparametric" estimation of probability density functions. *IEEE Trans. on Neural Networks*, 2(3):366–377, May 1991.

[18] N. A. Vlassis, A. Dimopoulos, and G. Papakonstantinou. The probabilistic growing cell structures algorithm. In *Proc. ICANN'97, 7th Int. Conf. on Artificial Neural Networks*, pages 649–654, Lausanne, Switzerland, Oct. 1997.

[19] N. A. Vlassis, G. Papakonstantinou, and P. Tsanakas. Mixture density estimation based on maximum likelihood and test statistics. *Neural Processing Letters*, 9(1), Feb. 1999, to appear.

[20] R. von Mises. *Mathematical Theory of Probability and Statistics*. Academic Press, New York, 1964.