# A Kurtosis and Skewness Based Criterion For Model Selection On Gaussian Mixture

Lin Wang and Jinwen Ma
Department of Information Science
School of Mathematical Sciences & LMAM, Peking University
Beijing,100871, China
jwma@math.pku.edu.cn

***Abstract*—The Gaussian mixture model is a powerful statistical tool in data modeling and analysis. Generally, the EM algorithm is utilized to learn the parameters of the Gaussian mixture. However, the EM algorithm is based on the maximum likelihood framework and cannot determine the number of Gaussians for a sample data set. In order to overcome this problem, we propose a new model selection criterion based on the kurtosis and skewness of the estimated Gaussians. Moreover, a new greedy EM algorithm is constructed via the kurtosis and skewness based criterion. The simulation results show that the proposed model selection criterion is efficient and the new greedy EM algorithm is feasible.**

## I. Introduction

The Gaussian mixture model is a very useful probability model for data analysis and information processing [1]. In fact, it has some good properties and can approximate any latent density. The Gaussian mixture model can be also regarded as a model of clustering analysis since a cluster can be defined by a Gaussian distribution [2]. Generally, the gaussian mixture learning is in a unsupervised learning mode. Actually, there have been a variety of learning algorithms for estimating the parameters of the Gaussian mixture with a sample data set, and the most famous one is the Expectation-Maximization (EM) algorithm [3]. But the EM algorithm has certain limitations. Firstly, it may easily be trapped into a local maximum of the likelihood function. Secondly, we must specify the appropriate number of Gaussians in the mixture beforehand. Otherwise, the EM algorithm will lead to a wrong result. Since the number of Gaussians is a kind of scale on Gaussian mixture, the selection of the number of Gaussians in the mixture is referred to as the model selection. In fact, the model selection problem for Gaussian mixture modeling or learning is a very complicated and difficult problem [4].

In a conventional way, we can use the EM algorithm with different numbers of Gaussians for the parameter learning on Gaussian mixture and choose the best based on the model selection criteria such as Akaike's information criterion, the Bayesian inference criterion, the Laplace empirical criterion, and the minimum message length (MML) criterion and so on (see [5] for a review and comparisons of these criteria). Actually, there are over 30 model selection criterions and some of them can be applied to the general finite mixture model [6]-[7]. However, each model selection criterion may has its limitations, and may be good for some sample sets while leads to a wrong result for some other sample sets.

Recently, the Bayesian Ying-Yang (BYY) harmony learning theory and systems [8]-[9] have been developed as a new statistical learning tool, especially for automated model selection on Gaussian mixture. Actually, it has provided a new learning principle, i.e., the BYY harmony learning principle, for the model selection of Gaussian mixture with a set of sample data. Based on it, several BYY harmony learning algorithms have been proposed, such as the BYY gradient, annealing, fixed-point and scale-incremental EM learning algorithms [10]-[13]. The BYY harmony learning algorithms are very efficient for the cases that the overlap of the actual Gaussians in the sample data is less, i.e., these actual Gaussians are separated in a certain degree. However, as the overlap of the actual Gaussians becomes strong, the BYY harmony learning principle and algorithms often leads to a poor result.

In this paper, we try to propose a new model selection criterion for Gaussian mixture based on the kurtosis and skewness of Gaussian distributions. That is, we use the kurtosis and skewness of each estimated Gaussian on the sample data set to measure whether this Gaussian is fitted to the sample data. Actually, Vlassis and Likas already used the kurtosis in the one-dimensional case [14]. Here, we extend this idea from one dimensional variable to the high dimensional vector. Moreover, we also use the information of the skewness. In order to check the new model selection criterion, we construct a greedy EM algorithm [15], [13] which tries to minimize the sum of kurtosis and skewness measures of estimated Gaussians in the mixture. It is demonstrated by the simulation experiments that the proposed model selection criterion is efficient and the new greedy EM algorithm is feasible.

In the sequel, we review the Gaussian mixture model and the EM algorithm in Section 2. In section 3 we introduce the new model selection criterion and propose the new greedy EM algorithm. Section 4 presents the data analysis on the new criterion and the experimental results of the proposed greedy EM algorithm. Finally, we conclude briefly in Section 5.

## II. Gaussian Mixture and EM algorithm

### A. $a$ $a$ , ,

The probability density function of the Gaussian mixture $p(x)$ is given as a finite weighted sum of Gaussian distribu-

tions. That is,

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k), \tag{1}$$

where the $k$-th component $N(x|\mu_k, \Sigma_k)$ is a $d$-dimensional Gaussian probability density function parameterized by the mean $\mu_k$ and the covariance matrix $\Sigma_k$, $\pi = \{\pi_k\}$ are the mixing proportions (or weights) that satisfy:

$$\sum_{k=1}^{K} \pi_k = 1, \pi_k \geq 0. \tag{2}$$

Given $X = \{x_1, \ldots, x_n\}$ as $n$ independent and identically distributed samples from the above Gaussian mixture model, the Gaussian mixture modeling on the sample data of $X$ is to determine the right number of Gaussians in the mixture and to estimate the true parameters $\theta^* = (\pi_1^*, \mu_1^*, \Sigma_1^*, \ldots, \pi_K^*, \mu_K^*, \Sigma_K^*)$ that maximizes the log likelihood function defined as follows:

$$L(\theta^*) = \prod_{i=1}^{n} p(x_i). \tag{3}$$

Although the number $K$ of Gaussians in the mixture is difficult to estimate with $X$, it is quite easy to get the ML estimate of the parameter of the Gaussian mixture when $K$ is right. The commonly used method is just the EM algorithm being introduced in the next subsection.

### B. E A

The EM algorithm is used in statistics for finding the maximum likelihood estimates of the parameters in a probabilistic model where the model depends on some unobserved latent variables. It has two steps: The E step computes an expectation of the log likelihood with respect to current estimate of the distribution for the latent variables, while the M step computes the parameters which maximize the expected log likelihood on the E step. These parameters are then used to determine the distribution of the latent variables in the next E step.

For the Gaussian mixture model, given specific $K$, the latent variables are the component indexes that the input samples statistically belong to. So, we can use the EM algorithm to estimate the other parameters. Here, we just give the parameter iterative formulas:

$$P(k|x_i) = \frac{\pi_k^{(t)} N(x_i|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} N(x_i|\mu_k^{(t)}, \Sigma_k^{(t)})} \tag{4}$$

$$\pi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} P(k|x_i) \tag{5}$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} P(k|x_i) x_i}{\sum_{i=1}^{n} P(k|x_i)} \tag{6}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n} P(k|x_i)(x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{n} P(k|x_i)} \tag{7}$$

It can be shown that the log likelihood in each EM step cannot be decreased. So it must converge to a local maximum. Unfortunately, it cannot make sure to find the global solution and may just get one local one. Moreover, in practice the number of Gaussians should be determined from the input samples, so the formulas above do not always lead to a good result. In order to solve the model selection and parameter learning problems, we will introduce a new model selection criterion based on kurtosis and skewness for multidimensional case in the next section.

## III. New Model Selection Criterion and Learning Algorithm

### A. a Ba C

As well known, for a Gaussian distribution, its kurtosis and skewness are both zero. if $Y = \{y_1, \ldots, y_m\}$ are independent and identically distributed samples from a Gaussian distribution with the parameters $\mu$ and $\sigma^2$, the sample kurtosis is $\frac{1}{n}\sum_{i=1}^{m}(\frac{y_i-\mu}{\sigma})^4 - 3$, and the sample skewness $\frac{1}{n}\sum_{i=1}^{m}(\frac{y_i-\mu}{\sigma})^3$. According to the law of large numbers, the sample kurtosis and sample skewness are both approximatively equal to zero. That is, the sample kurtosis and skewness can indicate how well a Gaussian distribution fit the sample data. For a true Gaussian mixture model, if every component is a Gaussian distribution, we can compute their sample kurtosis and skewness using the samples belonged to this component. So, the kurtosis and skewness offer us information to decide whether we have gotten the right number of components and the estimation of other parameters. In [14], a model selection criterion called the total kurtosis was proposed to find the right number of Gaussians in the mixture and according to that criterion, a dynamic learning algorithm was also constructed. According to the sample kurtosis of each component, the algorithm chooses one component with the largest kurtosis to be split and initialize the parameters of the split Gaussians and make the algorithm to more likely get the global solution. But it can just be used in univariate case and do not consider about the skewness information. We try to find a way to use the two kinds of information in the multidimensional case.

A straightforward idea is that we can find a projection direction and project the samples to this direction to get one dimension data. If the component is Gaussian distribution, the data also come from a Gaussian distribution. Then, we can compute the kurtosis and skewness and check whether they are zero or nearby. For convenience, we can make singular value decomposition for the covariance, and choose the first $F$ eigenvectors to be the projection directions. The section of $F$ is proper to reflect the information of the high dimensional samples.

We consider the following three ways to define the kurtosis and skewness of the projected data. We can firstly classify the

samples to $K$ Gaussians according to the maximum posterior $P(k|x_i)$ principle, that is,

$$j = \arg\max_k p(k|x_i), x_i \in j. \tag{8}$$

Let $X^k = \{x_1^k, \ldots, x_m^k\}$ be the samples belonged to Gaussian $k$, $v_1^k, \ldots, v_F^k$ are the first $F$ eigenvectors of the covariance $\Sigma_k$. We can define the kurtosis of the Gaussian $k$ in the direction $v_f^k$, named $ku_f^k(1)$, as follows:

$$ku_f^k(1) = \frac{1}{m}\sum_{i=1}^{m}(\frac{y_{fi}^k - m_f^k(1)}{\sigma_f^k(1)})^4 - 3, \tag{9}$$

where $y_{fi}^k$ are the projected data of $x_i^k$ in the direction of $v_f^k$, $m_f^k(1)$ and $\sigma_f^k(1)$ are the mean and standard deviation of data $Y_f^k = \{y_{f1}^k, \ldots, y_{fm}^k\}$. We can use the original $\mu_k$ and $\Sigma_k$ to produce another mean and standard deviation:

$$m_f^k(2) = \mu_k^T * v_f^k, \sigma_f^k(2) = \sqrt{v_f^{kT} * \Sigma_k * v_f^k}. \tag{10}$$

This leads to the second definition:

$$ku_f^k(2) = \frac{1}{m}\sum_{i=1}^{m}(\frac{y_{fi}^k - m_f^k(2)}{\sigma_f^k(2)})^4 - 3 \tag{11}$$

The third definition is quite different from the previous ones. We do not need to classify the samples in advance. Instead, we define a weighted kurtosis according to the posterior $P(k|x_i)$ as follows:

$$ku_f^k(3) = \frac{\sum_{i=1}^{n}(\frac{y_{fi}^k - m_f^k(2)}{\sigma_f^k(2)})^4 P(k|x_i)}{\sum_{i=1}^{n} P(k|x_i)} - 3 \tag{12}$$

The importance of the sample $x_i$ in the formula is decided by the posterior $P(k|x_i)$ which means the probability of the event that $x_i$ belongs to Gaussian $k$. No matter what the definition is, we compute the sum of the kurtosis of Gaussians by

$$ku_k = \sum_{f=1}^{F} abs(ku_f^k) \tag{13}$$

The absolute values are needed to compensate for the individual kurtosis taking positive or negative values. To test how well the whole mixture model fit the samples we should compute the weighted average of the individual kurtosis of the components of the mixtures. The total kurtosis is

$$K_T = \sum_{k=1}^{K} \pi_k ku_k \tag{14}$$

In the same way, we can define the total skewness as

$$S_T = \sum_{k=1}^{K} \pi_k sk_k. \tag{15}$$

Finally, we define the total sum of the kurtosis and skewness as

$$Sum_T = K_T + S_T. \tag{16}$$

The total kurtosis and total skewness is a useful information to check the fitness of the Gaussian mixture model. It can be regarded as a measure on how well a Gaussian mixture fits the sample data. Since a low value indicates every component of the mixtures fit the samples in its vicinity, therefore the mixture is a good approximation of the latent distribution. If the value is large, it shows that the samples belong to some components is not likely come from a Gaussian so we should change the number of components or adjust the parameters.

We use both the kurtosis and skewness because they introduce more information than either of them, especially after projection. One dimensional data may have low kurtosis or low skewness, but they are likely to have large value of total kurtosis and total skewness. As for the three definitions of $ku_f^k$, each of them will be evaluated by the experiments. Indeed, they all behave well and everyone may be better than the other ones in some cases.

### B. E A

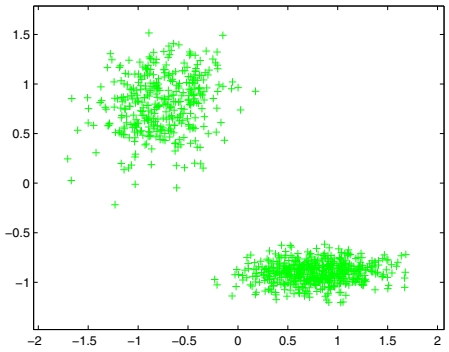
Based on the total kurtosis and skewness, we can construct a new greedy EM algorithm which can solve the two problems of the EM algorithm we have mentioned above. It utilizes the EM algorithm to estimate the parameters and split the component with the largest value of kurtosis and skewness step by step and use the minimum total kurtosis and skewness criterion to find the right number of Gaussians in the sample data. That is, the lowest value of the total kurtosis and skewness corresponds to the right number of Gaussians.

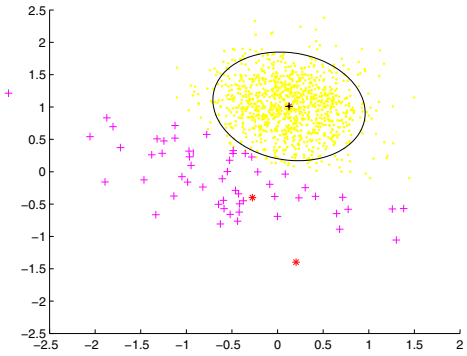
Specifically, we starts with a small number of Gaussians such as $k = 1$, and then perform the EM algorithm until convergence, and compute the total kurtosis and total skewness measure. If the value of the measure is too large being compared with a previously given threshold value, we perform the component splitting. We choose the component which has the maximum value of $\pi_k(ku_k + sk_k)$. The two new components have means $\mu_k + v_1^f, \mu_k - \nu_1^f$ and the same covariance $\Sigma_k$, their weights are set to $\pi_k/2$ so that Eq.(2) still holds. Furthermore, we perform the EM algorithm again, until the measure is small enough or perform the procedure until the number is big enough and choose the one which has the least value of total kurtosis and total skewness as the tight number of components. As we initialize the parameters when a new component is added, we may escape from the local maximum of the log likelihood function. Thus, the proposed greedy EM algorithm can behave better than the standard EM algorithm.

Particularly, we give the new greedy EM algorithm as follows:

Step 1. Initialization: Set the initial number $K = 1$ and initialize the parameters of the components randomly. Set the threshold value $\varepsilon > 0$.

Step 2. Perform the EM algorithm until convergence.


1.5
1
0.5
0
−0.5
−1
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2


2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2