

LEARNING GMM USING ELLIPTICALLY CONTOURED DISTRIBUTIONS

Bo Li

(1)School of Automation, Beijing Institute of Technology
(2)Key Laboratory of Complex System Intelligent Control and Decision (Beijing Institute of Technology), Ministry of Education
bli191@sina.com

Wenju Liu

National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy of Science

lwj@nlpr.ia.ac.cn

Lihua Dou

(1)School of Automation, Beijing Institute of Technology
(2)Key Laboratory of Complex System Intelligent Control and Decision (Beijing Institute of Technology), Ministry of Education
doulihua@bit.edu.cn

Abstract-Model order selection and parameter estimation for Gaussian mixture model (GMM) are important issues for clustering analysis and density estimation. Most methods for model selection usually add a penalty term in the objective function that can penalize the models and choose an optimal one from a set of candidate models. This paper presents a simple and novel approach to determine the number of components and simultaneously estimate the parameters for GMM. By introducing the degenerating model, the proposed approach overcomes the drawback of likelihood estimate that is a non-decreasing function and can not be used to select the number of components. The degenerating model is a more general form of mixture component density and it can degenerate into the component density or a crater-like density when its parameter K varies from 1 to a bigger value. The likelihood of the crater-like density evaluated for the training data approximates to zero. This characteristic of the degenerating model forms the foundation of the proposed approach. The experimental results show robust and evident performance improvement of the approach.

1. INTRODUCTION

Mixture models are a type of density model which comprise a number of component densities, usually Gaussian. These component functions are combined to provide a multimodal density. Gaussian mixture model (GMM) is currently among the most statistically mature methods for clustering and density estimation.

Unlike making use of penalty functions, the paper proposed a new approach for estimating the number of components and parameters by constructing degenerating models. The paper is organized as follows: Section 2 reviews some related work and defines some notations. Section 3 introduces the elliptically contoured distributions and presents the degenerating model approach. Section 4 shows the experimental results and section 5 makes some conclusions.

* This work was supported in part by the China National Nature Science Foundation (No. 60675026, No. 60820011, No. 90820303), the 863 China National High Technology Development Projects (No.20060101Z4073, No.2006AA01Z194) and the co-construction fund of Beijing Commission of Education.

2. RELATED WORK

2.1 GMM and EM Algorithm

A finite d -dimensional GMM $p(x)$ parameterized by $\Theta = (\alpha_1, \alpha_2, \dots, \alpha_M, \vartheta_1, \vartheta_2, \dots, \vartheta_M)$ is a linear combination of M Gaussian component densities, as given by the following equation[1]

$$p(x | \Theta) = \sum_{i=1}^M \alpha_i p_i(x | \vartheta_i) \quad (1)$$

where the probability density function of the i -th of the possible Gaussian distributions is denoted by $p_i(x | \vartheta_i)$ and x is a d -dimensional feature vector. The mixture weights α_i are constrained by $\sum_{i=1}^M \alpha_i = 1$ $\alpha_i \geq 0$. The Gaussian component density with mean μ_i and covariance Σ_i for d -dimensional space has the following form:

$$p_i(x) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right\} \quad (2)A$$

n expectation-maximization (EM) algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E step. The parameters found on the M step are then used to begin another E step, and the process is repeated [1][2][3].

2.2 Model Order Selection

The selection of the number of mixture components for GMM is an important issue that achieves widely concern. Most methods for model order selection add a penalty term to the log-likelihood function for the given dataset and then maximize the updated expression. Aiming at deciding the model order, many model selection criteria have been proposed. The criterion of such methods is often written as: $J(M, \hat{\Theta}(M)) = \log(X; \hat{\Theta}(M)) - P(M)$

$\log(X; \hat{\Theta}(M))$ refers to the log-likelihood for the given data. The penalty function $P(M)$ is introduced to balance $\log(X; \hat{\Theta}(M))$ for it is a non-decrease function for the number of components M [1]. These methods can be grouped into the following classes.

Bayesian Approximation Criterion: LEC (Laplace Empirical Criterion) and Bayesian Information Criterion (BIC)

Information Encoding: Minimum Description Length (MDL), Minimum Message Length (MML), Akaike Information Criterion(AIC) and ICOMP(Information Complexity Criterion)

Complete Likelihood Function: Classification Likelihood Criterion, Approximate Weight of Evidence (AWE), Normalized Entropy Criterion (NEC), Integrated Classification Likelihood Criterion (ICL) and ICL-BIC.

Other articles have also discussed the likelihood ratio test based methods.

3. FITTING GMM VIA DEGENERATING MODEL

Traditional multivariate analysis mostly focuses on the multivariate Gaussian distribution. As the further development of traditional multivariate analysis, generalized multivariate analysis developed in recent years extends Gaussian distribution to a more general form. The essence of generalized multivariate analysis is the elliptically contoured distribution [4].

Definition:

A random vector $x = (x_1, x_2, \dots, x_n)^T$ belongs to the elliptically contoured distributions, if its density function has the form

$$f(x) = \frac{c_n}{\sqrt{|\Sigma|}} g_n \left[\frac{1}{2} (x-u)^T \Sigma^{-1} (x-u) \right] \quad (8)$$

$g_n(\bullet)$ is the density generator. The normalizing constant c_n is determined by the following expression

$$c_n = \frac{\Gamma(\frac{n}{2})}{(2\pi)^{\frac{n}{2}}} \left[\int_0^\infty r^{\frac{n}{2}-1} g_n(r) dr \right]^{-1} \quad (9)$$

where $\Gamma(x)$ is the gamma function[4].

3.1 Kotz-Type Distribution

An elliptical vector x belongs to the multivariate Kotz-Type distribution, if its density generator could be written as:

$$g(u) = t^{K-1} \exp(-rt^s) \quad t = (x-u)^T \Sigma^{-1} (x-u) \quad (10)$$

The density function of Kotz-Type distribution can be expressed as:

$$f(x) = \frac{C_n}{\sqrt{|\Sigma|}} [(x-u)^T \Sigma^{-1} (x-u)]^{K-1} \exp\{-r[(x-u)^T \Sigma^{-1} (x-u)]^s\} \quad (11)$$

where the normalizing constant is

$$c_n = s\pi^{-\frac{n}{2}} r^{\frac{2K+n-2}{2s}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{2K+n-2}{2s})} \quad r, s > 0, 2K+n > 2 \quad (12)$$

When $s=1$ alone, Kotz-Type distribution degenerates into the original Kotz-Type distribution. When $K=1, s=1$ and $r=1/2$, it degenerates into the Gaussian distribution. For $K<1$, the density function tends to infinity at the origin, whereas for $K>1$, the density function has a local minimum at the origin and looks like a volcano crater. Inside the crater the density function approximates to zero.

When $K>1$, Kotz-Type distribution has a useful property that with the parameter K becoming bigger, the space of the crater accordingly becomes broader. This property of Kotz type is the basis of the proposed approach for determining the number of components [4].

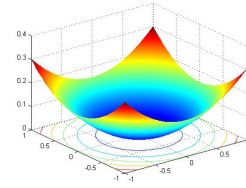


Fig. 1. KOTZ TYPE K=2

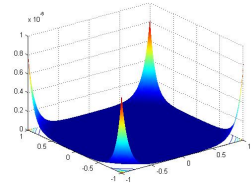


Fig. 2. KOTZ TYPE K=6

3.2 Analysis of the Kotz Type Density

It is known that the likelihood value of the mixture model evaluated for the training data is incapable of estimating the number of components since it is a non-decreasing function. Some mixture components belonging to an over-fitted mixture model maybe nested, that is, given certain training data, the linear combination of two or more mixture components functions as one mixture component. Also in this situation, some mixture densities may have very large value of covariance to maintain the ML value.

Suppose a dataset obeys the Gaussian distribution. A Kotz type density with the constraint $K \geq 1, r = \frac{1}{2}, s = 1$ is

used to fit the dataset. It has been proved that given a dataset from a multivariate Gaussian distribution, the Gaussian based ML estimator (i.e. Kotz type density where $K=1, r=1/2, s=1$) performs better as expected to fit the dataset than the Kotz type ML estimator and vice versa. In other words, a Kotz type density with the constraint $K \geq 1, r = \frac{1}{2}, s = 1$ would degenerate into a Gaussian density if it is used to fit the dataset[4][5][6].

Furthermore consider the problem of model order selection. Suppose an M -th order GMM denoted by (1). Add the Kotz type density constrained by $K \geq 1, r = \frac{1}{2}, s = 1$ in this GMM.

The EM is used to train the new mixture model. Based on the above discussion, it is expected that the Kotz type component density degenerate into a Gaussian component density if the mixture model is under-fitted whereas expand

its crater to cover the training feature vectors if the mixture model is over-fitted. The probabilities and likelihood value of the training feature vectors evaluated for the crater-like density approximates to zero.

3.3 Non-linear Programming Based EM

Before further discussion, the non-linear programming based EM is introduced to train the mixture model with different types of mixture components. To simplify the problem, consider a mixture model containing M Gaussian components denoted by (1) and one Kotz type component. The new mixture model has the form

$$p(x | \Theta) = \sum_{i=1}^M \alpha_i p_i(x | \vartheta_i) + p_{kotz}(x | \vartheta_{kotz}) \quad (13)$$

p_{kotz} refers to the Kotz type component density where $K \geq 1, r = \frac{1}{2}, s = 1$ and the parameters are denoted by ϑ_{kotz} .

The Q function of the new mixture model is written as:

$$\begin{aligned} & \sum_{j=1}^N \sum_{i=1}^{M+1} \log(\alpha_i) P(i | x_j, \Theta) + \sum_{j=1}^N \sum_{i=1}^M \log(p_i(x_j | \vartheta_i)) P(i | x_j, \Theta) \\ & + \sum_{j=1}^N \log(p_{kotz}(x_j | \vartheta_{kotz})) P(kotz | x_j, \Theta) \\ & = I_a + I_b + I_c \quad (14) \end{aligned}$$

N denotes the number of training feature vectors.

Similar to (4), I_a, I_b, I_c can be maximized respectively.

The estimation for the updated parameters of GMM is the same as (6) and (7). It is observed that $\max(I_c) = \min(-I_c)$ and this is a multivariable optimization problem where the independent variable y could be rewritten as: $y = (u_1, u_2 \dots u_n, \sigma_1, \sigma_2 \dots \sigma_{n*n}, \tau_1, \tau_2 \dots \tau_h)'$

$u_1, u_2 \dots u_n, \sigma_1, \sigma_2 \dots \sigma_{n*n}$ denotes the elements of the mean vector and covariance matrix. $\tau_1, \tau_2 \dots \tau_h$ denote the other parameters of the component density p_{kotz} . This is

a non-linear programming optimization problem [8]. Sequential Quadratic Programming (SQP) can be applied to solve the problem for it is the most efficient and accurate method developed in recent years for medium-scale nonlinear constrained optimization problem. More introduction of SQP can be found in [7][8]. If the original mixture is not GMM, the estimation for the mixture parameters can also applied non-linear programming approach.

3.4 Definition of the Crater

If the criterion that the Kotz type degenerates into the Gaussian distribution or not is directly used to determine if the GMM is over-fitted, mistakes may occur. On occasion, the likelihood value of an under-fitted GMM mixed with a Kotz type component may continue to increase and the estimate of K is still greater than 1. This is due to the fact that a Kotz type component may better fit the data clusters hidden in the training feature vectors than one or several

Gaussian components from an under-fitted GMM. Similarly an over-fitted GMM mixed with the Kotz type component where $K > I$ may also produce a higher likelihood due to the peculiar density function of Kotz type. Hence $K > I$ can not be employed for the stopping criterion to choose the number of components. The solution to this difficulty depends on the definition of the crater formed by the Kotz type density where $K > I$.

The definition includes the following three points:

(1) The aim of defining the crater space is to make the Kotz type density degenerate into Gaussian density if GMM is under-fitted and expand its crater to cover all the training feature vectors if GMM is over-fitted. On the basis of the analysis, the estimate of K of the added Kotz type component is limited to $1 \cup [H_1, H_2]$. The value of the lower bound H_1 must make the space of the crater become broader enough to cover all the training feature vectors.

(2) The GMMs with different number of components are well trained in advance and their ML values converge to global maximum. Since the estimate of K has been limited to $1 \cup [H_1, H_2]$, the Kotz type component either degenerates into a Gaussian component or produces a broader crater. If a GMM is with the optimal number of components but has not been well trained, the Kotz type component is also possible to degenerate into a Gaussian component. The well trained GMMs can be obtained by a good initialization or adding a random disturbance to the updated parameters at each EM iteration.

(3) The initial value of K is bigger enough to make the space of the crater cover the feature space of the training data. If its initial value is equal to 1, the Kotz type may also degenerate into a Gaussian component with very large value of covariance if GMM is over-fitted.

3.5 Degenerating Model

The Kotz Type density where $K \geq 1, r = \frac{1}{2}, s = 1$ is

named as *degenerating model* for GMM. The essence for constructing the degenerating model is to make it degenerate into the mixture component density when $K=I$. Hence the degenerating model of GMM is a Kotz-Type distribution where $K \geq 1, r = \frac{1}{2}, s = 1$.

Complete Algorithm

- a) Suppose a finite mixture model with M mixture components as shown in (1)
- b) The mixture model comprised M mixture components is well trained.
- c) Train the new mixture composed of the M -th order mixture model and the degenerating model, store the operation result of parameter K
- d) If the value of stored K changes, stop;
 - Else if $K > I, M=M-I$ and turn to c);
 - Else if $K=I, M=M+I$ and turn to c).

4. EXPERIMENTAL RESULTS

Experimental Setup

The constraint of K is set as $1 \cup [10, 100]$. The upper bound of the mean of degenerating model is twice as much as the maximal absolute value of the corresponding elements of training feature vectors. The lower bound of the mean is equal to the negative upper bound. The mean and covariance of a randomly selected Gaussian component of the well trained GMM is used for the initial mean of the degenerating model. The initial mixture weight of the degenerating model is set as 0.001 and the mixture weights of trained GMM are normalized as $0.999 * \alpha_i$.

The three Gaussian Dataset

The first experiment considers the three Gaussian components with 900 samples.

$$\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3} \quad u_1 = [0, 2]^T \quad u_2 = [0, 0]^T \quad u_3 = [0, -2]^T$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 0.2 \end{bmatrix}$$

The initial number of Gaussian components is set as 1. When the number of Gaussian components is smaller than three, the proposed degenerating model approach needs 3~5 EM iterations to make the degenerating model degenerate into the Gaussian component. When the number of Gaussian components is more than three, the mixture weight of the degenerating model approximates to zero and the value of K varies from 12~30. The algorithm runs 100 times and every time selects the correct number three.

Comparison with MDL

At last consider the situation where the Gaussian components almost overlap and evaluate the influence of the degree of overlap to the performance of the degenerating model approach and MDL. Suppose a GMM with $u_1 = [50, 50]^T$, $u_2 = [50 + \tau, 0]^T$, $\Sigma_1 = \Sigma_2 = E$

Figure 3 shows the percentage of correct selection over 100 times by MDL and the proposed approach respectively (1000 data samples). Owing to the different stopping criterion, it can be seen from figure 3 that the proposed degenerating model approach performs better when the degree of overlap τ varies. The MDL selects the number of components according to the ML values of the GMMs penalized and the optimal model chose from a set of candidate models that has the largest ML value. If the mean points of the two Gaussian components are very close, the likelihood values of them are also very close. If the separation is too small, the penalty function penalizes a bigger value in the likelihood. So in most cases MDL chooses one component as the optimal number. The proposed approach however, uses the criterion $K > 1$ as the stopping criterion. If the initial number of Gaussian components is greater than two, the degenerating model

contributes a likelihood approximating zero. If the GMM is under fitted and can produce a larger likelihood value, the degenerating model would degenerate the Gaussian component.

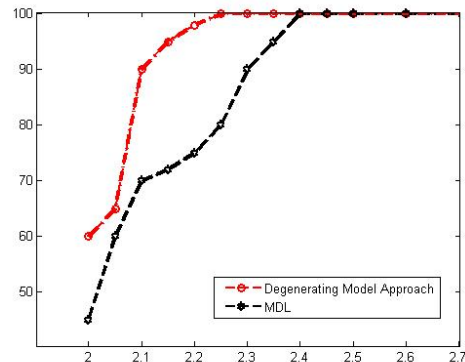


FIG 3 PERFORMANCE COMPARISON

5. CONCLUSIONS

This paper proposed a new approach for estimating the number components and parameters for finite mixture models. Unlike using penalty functions, the approach introduces the construction of degenerating model, which produces a crater-like density. Inside the crater the density approximately equals zero. It overcomes the drawback of the likelihood value of the mixture model that is a non-decreasing function and can be used for model selection.

6. REFERENCE

- [1]. Geoffrey McLachlan, David Peel "Finite Mixture Models" Wiley press, September 2000
- [2]. Dempster. N. Laird, and, D. Rubin, "maximum likelihood from incomplete data via the EM algorithm," J. Royal stat.soc, vol.39, pp: 1-38, 1977.
- [3]. Jeff A.Bilmes, "A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gauss mixture Model and Hidden Markov Models," April, 1998
- [4]. Fang K.T. and Zhang Y Generalized Multivariate Analysis. Springer-Verlag and Science Press, Berlin and Beijing,(1990).
- [5]. Amal Helu, Dayanand N. Naik, Estimation of interclass correlation via a Kotz-type distribution Computational Statistics & Data Analysis, Volume 51, Issue 3 (December 2006)Pages 1523-1534
- [6]. Nadarajah, 2003. The Kotz-type distribution with applications. Statistics. v37. 341-358.
- [7]. Han, S.P, "A Globally Convergent Method for Nonlinear Programming," Journal of Optimization Theory and Applications, Vol. 22, 1977.
- [8]. M.S. Bazaraa, C.M. Shetty, Nonlinear Programming Theory and Algorithms, Wiley, New York, 1979.