Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 757240, 9 pages
http://dx.doi.org/10.1155/2013/757240

Research Article

Improved Expectation Maximization Algorithm for Gaussian Mixed Model Using the Kernel Method

Mohd Izhan Mohd Yusoff,[1,2] Ibrahim Mohamed,[1] and Mohd Rizam Abu Bakar[3]

[1] *Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia*
[2] *Telekom Research & Development Sdn Bhd., 63000 Cyberjaya, Selangor, Malaysia*
[3] *Department of Mathematics, Universiti Putra Malaysia, 43400 Serdang, Malaysia*

Correspondence should be addressed to Mohd Izhan Mohd Yusoff; izhan@tmrnd.com.my

Received 25 April 2013; Accepted 17 July 2013

Academic Editor: Kue-Hong Chen

Fraud activities have contributed to heavy losses suffered by telecommunication companies. In this paper, we attempt to use Gaussian mixed model, which is a probabilistic model normally used in speech recognition to identify fraud calls in the telecommunication industry. We look at several issues encountered when calculating the maximum likelihood estimates of the Gaussian mixed model using an Expectation Maximization algorithm. Firstly, we look at a mechanism for the determination of the initial number of Gaussian components and the choice of the initial values of the algorithm using the kernel method. We show via simulation that the technique improves the performance of the algorithm. Secondly, we developed a procedure for determining the order of the Gaussian mixed model using the log-likelihood function and the Akaike information criteria. Finally, for illustration, we apply the improved algorithm to real telecommunication data. The modified method will pave the way to introduce a comprehensive method for detecting fraud calls in future work.

1. Introduction

Every year telecommunication companies register loses amounting to millions of dollars due to fraud activities. Examples of such activities given by [1, 2] include the use of the customer's line in Premium Rate Service without their knowledge or autodialers with no intention to pay for the outgoing calls; PABX for international calls; an unregistered user with an assigned number accessing the network (such activity is called stolen line unknown); and international roaming manipulation. Vendors, seeing the above as an opportunity not to be missed, compete to provide data mining applications that can detect the said activities effectively using various methods such as OLAP, deviation based outlier detection, and hidden Markov model. The focus of this paper is Gaussian mixed model, henceforth, GMM.

A GMM is best known for providing a robust speaker representation for the difficult task of speaker identification on short-time speechspectra [3]. Its function is extended to detect fraud activities on the number (as well as length) of domestic and international calls made on a daily basis during office, evening, and night hours. Reference [4] presented three approaches to fraud detection in communication networks: neural networks with supervised learning, probability density estimation methods, and Bayesian networks. Information describing a subscriber's behavior kept in toll tickets was used. For example, supervised learning used summary statistics over the whole observed time period (especially the number of times fraud activities were recorded in the data). The two latter approaches used a subscriber's daily behavior. To improve the fraud detection system, they recommended the combination of the three presented methods together with the incorporation of rule based systems.

The maximum likelihood estimation for a GMM is generally difficult to obtain directly, but it is made easier with the availability of the Expectation Maximization (EM) algorithm which was first introduced by [5]. Since then, there has been a significant increase in its use especially in finding the maximum likelihood for probabilistic models. For example, [6, 7] developed an online system for detecting fraud calls using a hierarchical switching generative model. The model is trained by using the EM algorithm on an incomplete data set and is further improved by using a gradient-based discriminative method. In this paper, we propose an improved EM algorithm for GMM in detecting fraud calls in telecommunication.

This paper is organized as follows. Section 2 gives an introduction to GMM. We then describe the EM algorithm for a GMM, the kernel method, and eventually the proposed modified EM algorithm for GMM in Section 3. In Section 4, we study the performance of the modified algorithm in estimating the parameters and the effect of overlapping areas of Gaussian components in GMM. In the next section, we propose graphical plots to identify the "best" number of components in a GMM. For illustration, an application of the improvement on a real data set is presented in Section 6.

2. Gaussian Mixed Model

Let $\mathbf{x} \in R^d$ and K be the number of components where each component has its own prior probability a_i and probability density function with mean $\boldsymbol{\mu}_i$ and covariance Σ_i, $i = 1, \ldots, K$. A Gaussian mixed model is then given by

$$\sum_{i=1}^{K} a_i \phi\left(\mathbf{x} \mid \boldsymbol{\mu}_i, \Sigma_i\right) = \sum_{i=1}^{K} a_i \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_i\right|}} \exp\left(\frac{-\left(\mathbf{x}-\boldsymbol{\mu}_i\right)^t \Sigma_i^{-1} \left(\mathbf{x}-\boldsymbol{\mu}_i\right)}{2}\right), \tag{1}$$

where $\sum_{i=1}^{K} a_i = 1$. We next define the likelihood function and the log-likelihood function by $L(\mathbf{X} \mid \theta) = \prod_{j=1}^{n} f(\mathbf{x}_j \mid \theta)$ and $l(\mathbf{X} \mid \theta) = \sum_{j=1}^{n} \log(\sum_{i=1}^{K} a_i \phi(\mathbf{x}_j \mid \boldsymbol{\mu}_i, \Sigma_i))$ where $\mathbf{X} = (\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t)^t$, respectively. The maximum likelihood estimation (m.l.e) method aims at finding $\widehat{\theta}$ that maximizes $l(\mathbf{X} \mid \theta)$; see [8]. The expression $\log(\sum_{i=1}^{K} a_i \phi(\mathbf{x}_j \mid \boldsymbol{\mu}_i, \Sigma_i))$ in $l(\mathbf{X} \mid \theta)$ is difficult to compute. We use the Expectation Maximization (EM) algorithm to overcome this problem.

3. Expectation Maximization (EM) Algorithm

3.1. EM for Gaussian Mixed Model. In a general setup of the EM algorithm given in [5], the authors considered an unobservable variable X in sample space $\mathcal{X}$, which is indirectly observed through observed variable Y in sample space $\mathcal{Y}$. Assuming that $f(x \mid \theta)$ is the sampling density depending on the parameter $\theta \in \Omega$, the corresponding family of sampling densities for Y, say $g(y \mid \theta)$, can be derived from

$$g(y \mid \theta) = \int_{\chi(y)} f(x \mid \theta)\, dx, \tag{2}$$

where $\chi(y)$ is a subset of $\mathcal{X}$ under the mapping $x \rightarrow y(x)$ from $\mathcal{X}$ to $\mathcal{Y}$. The main objective of the EM algorithm is to find the value of θ that maximizes (2). Consider the expected value of $\log f(x \mid \theta')$ given y and θ, denoted by $Q(\theta' \mid \theta)$, where

$$Q\left(\theta' \mid \theta\right) = E\left(\log f\left(x \mid \theta'\right) \mid y, \theta\right) \tag{3}$$

with the expectation assumed to exist for all pairs (θ', θ) and $f(x \mid \theta) > 0$ for $\theta \in \Omega$. According to [5], the EM iteration consists of two steps, namely, the E-step and the M-step. At the pth iteration with the estimate of θ denoted by $\theta^{(p)}$, the E-step will give the value of $Q(\theta \mid \theta^{(p)})$ and the M-step will find a new estimate of θ, say $\theta^{(p+1)}$, that maximizes $Q(\theta \mid \theta^{(p)})$. The steps are repeated until convergence is achieved.

For the case of a GMM, we define $Q(\theta' \mid \theta) = E[\log \prod_{i=1}^{n} a'_{y_i} \phi(\mathbf{x}_i \mid \boldsymbol{\mu}'_{y_i}, \Sigma'_{y_i}) \mid \mathbf{X}, \theta]$, where $y_i \in \{1, 2, \ldots, K\}$ and $y_i = k$, if the ith sample is generated by the kth mixture component. It is simplified, by applying, amongst others, the Bayes formula $f(\theta \mid x) \propto f(x \mid \theta) P(\theta)$ where $f(\theta \mid x)$ is the posterior probability, $f(x \mid \theta)$ is the likelihood function, and $P(\theta)$ is the prior probability to the following equations (see [9, 10]):

$$Q\left(\theta' \mid \theta\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} p_{i,k} \log a'_k + \sum_{i=1}^{n} \sum_{k=1}^{K} p_{i,k} \log \phi\left(\mathbf{x}_i \mid \boldsymbol{\mu}'_k, \Sigma'_k\right), \tag{4}$$

where

$$p_{i,k} = \frac{a_k \phi\left(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k\right)}{\sum_l a_l \phi\left(\mathbf{x}_i \mid \boldsymbol{\mu}_l, \Sigma_l\right)}, \tag{5}$$

$$\phi\left(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k\right) = \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_k\right|}} \exp\left(\frac{-\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)^t \Sigma_k^{-1} \left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)}{2}\right). \tag{6}$$

Hence, the EM iteration for a GMM is defined by the following.

E-step: use (5).

M-step: use the formulas

$$a_j = \frac{1}{n} \sum_i p_{ij}, \qquad \boldsymbol{\mu}_j = \frac{\sum_i p_{ij} \mathbf{x}_i}{\sum_i p_{ij}}, \qquad \Sigma_j = \frac{\sum_i p_{ij} \left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)^t}{\sum_i p_{ij}} \tag{7}$$

which are derived from the Lagrange multipliers, $\partial Q / \partial \boldsymbol{\mu}_j = 0$ and $\partial Q / \partial \Sigma_j^{-1} = 0$ (for details, see the appendices).

The above steps (i.e., E-step and M-step) are repeated until convergence is achieved.

3.2. The Kernel Method. The kernel method can be used to find the probability density estimate for univariate data; see, for example, [11]. Let $\alpha < \min(x_i) - 3h$, $\beta > \max(x_i) + 3h$, $M = 2^r$; let h be the bandwidth for some integer r, $\delta = (\beta - \alpha)/M$; and let $t_k = \alpha + k\delta$ be the kth grid point where $k = 0, 1, \ldots, M - 1$. The density estimate at grid point t_k is represented by the following equation:

$$\widehat{f}(t_k) \approx \sum_{l=-M/2}^{M/2} \left(\frac{1}{M} \sum_{k=0}^{M-1} \xi_k \exp\left(\frac{2\pi kl}{M} i \right) \right) \times \exp\left(\left(-\frac{2\pi kl}{M} i \right) - \frac{1}{2} h^2 \left(\frac{2\pi l}{\beta - \alpha} \right)^2 \right), \tag{8}$$

where $i^2 = -1$. For $x \in [t_k, t_{k+1}]$, the density estimate $\widehat{f}(x)$ is defined by $\widehat{f}(x) = (1/nh) \sum_{i=1}^{n} K((x - x_i)/h)$ where $K(t) = (1/\sqrt{2\pi}) \exp(-(1/2)t^2)$. To compute $\widehat{f}(x)$ at a grid of points, a method which makes use of the Fourier transform is employed. Let $\widetilde{f}(s)$ be the Fourier transform of the kernel density estimate $\widehat{f}(x)$. It can be shown that $\widetilde{f}(s) = (2\pi)^{1/2} \widetilde{K}(hs) u(s) = \exp(-(1/2)h^2 s^2) u(s)$, where $\widetilde{K}(s)$ is the Fourier transform of the Gaussian kernel and $u(s) = (2\pi)^{-1/2} n^{-1} \sum_{j=1}^{n} \exp(isx_j)$ is the Fourier transform of the data. Thus, $\widehat{f}(x) = (2\pi)^{-1/2} \int e^{-isx} (2\pi)^{1/2} \widetilde{K}(hs) u(s) ds$ is the convolution of the data with the kernel.

We will use the following algorithm by [11] to discretize the data to very fine grids and to find $\widehat{f}(x)$ by convolving the data with the kernel.

Step A. Discretize the data to find the weight sequence $\{\xi_k\}$ with $M = 2^8$. If $x \in [t_k, t_{k+1}]$, it is split into a weight $(1/n\delta^2)(t_{k+1} - x)$ at t_k and a weight $(1/n\delta^2)(x - t_k)$ at t_{k+1}; these weights are accumulated over all the data points x_i to give a sequence of (ξ_k) weights summing up to $1/\delta$.

Step B. Find the sequence $\{Y_l\}$ defined by $Y_l = M^{-1} \sum_{k=0}^{M-1} \xi_k \exp((2\pi kl/M)i)$, where $-(M/2) \le l \le M/2$. It can be shown that when $\alpha = 0$, $Y_l \approx (2\pi)^{1/2} (\beta - \alpha)^{-1} u(s_l)$, where $s_l = 2\pi l/(\beta - \alpha)$.

Step C. Find the sequence $\{\zeta_l^*\}$, where $\zeta_l^* = \exp(-(1/2)h^2 s_l^2)$ Y_l. Here, $h = 0.9An^{-1/5}$, where $A = \min(\text{sd}, \text{IQR}/1.34)$, sd is the standard deviation, and IQR is the interquartile range. The IQR is chosen here by [11], who claimed that the bandwidth is useful for a wide range of densities.

Step D. Let ζ_k be the inverse discrete Fourier transform of ζ_l^*, that is, $\zeta_k = \sum_{l=-M/2}^{M/2} \zeta_l^* \exp(-(2\pi kl/M)i)$.

It can be shown that when $\alpha = 0$, $\widehat{f}(t_k) \approx \zeta_k$. We then identify x_i where its density estimate, denoted by $\widehat{f}(x_i)$, is greater than those of its nearest neighbors x_{i+1} and x_{i-1}. In other words, $\widehat{f}(x_i) > \widehat{f}(x_{i+1})$ and $\widehat{f}(x_i) > \widehat{f}(x_{i-1})$; refer to Figure 1 where the vertical line that touches t_k and $\widehat{f}(t_k)$ shows the location of the peak. Note that we may obtain more than one maximum points, which means that the data may consist of more than one Gaussian distribution. These results form a very important component of the improved EM algorithm for GMM to be described next.

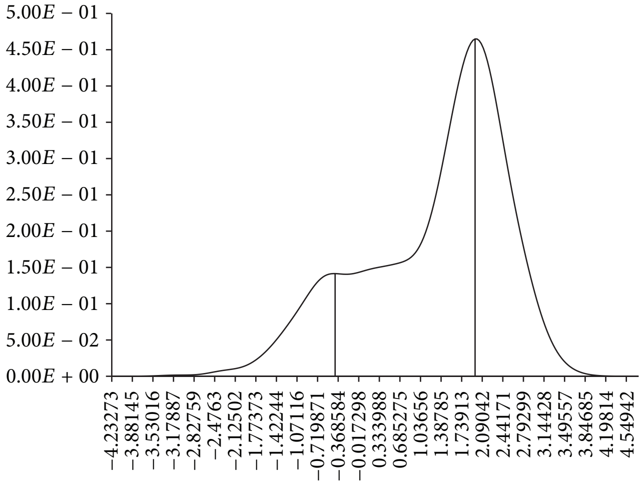


Figure 1: Plot of $\widehat{f}(t_k)$ against t_k.

3.3. Improved EM Algorithm for GMM. A number of authors highlighted the importance of identifying the right number, say k, of components in a GMM and subsequently choosing good initial values for the model parameters μ_i and σ_i^2, $i = 1, 2, \ldots k$, in the EM algorithm. Reference [12] noted the difficulty of using log-likelihood-ratio statistics to test the number of components and subsequently suggested using a nonparametric bootstrapping approach. Similarly, [13] pointed out the same concerns and introduced an algorithm called the stepwise-split-and-merge EM algorithm to solve the said problem. In addition, [14] investigated the possibility of using the minimization of the Kullback-Leiber distance between fitted mixture model and the true density as a method for estimating k where the said distance was estimated using cross validation. Reference [15] viewed the mixture distribution as a contaminated Gaussian density and proposed a recursive algorithm called the Gaussian Mixture Density Decomposition algorithm for identifying each Gaussian component in the mixture. Other works on this topic can also be found, for example, in [16, 17].

In this paper, we propose an improved EM algorithm for GMM which can perform both tasks: identifying the initial number of components and providing automatic initial values for the EM algorithm. The full improved EM algorithm for a GMM is now presented.

Step 1. The kernel method as described in Section 3.2 is used to determine the number, say K_0, of components and also the corresponding means μ_i of each component, where $i = 1, 2, \ldots, K_0$. The initial estimates of the standard deviations σ_{ii} are set to unity while the prior weights a_i are set to be $1/K_0$.

Step 2. The EM algorithm for a GMM as described in Section 3.1 is executed to give the final estimates of parameters μ_i, σ_{ii}, and a_i, $i = 1, 2, \ldots, K_0$. The log-likelihood function

Table 1: List of true values of a's, and μ's, σ's.

Sample name and size (in bracket)	Prior probability	Mean	Variance
Sample 1 Two components	$a_1 = 0.4$ $a_2 = 0.6$	$\mu_1 = 0.0$ $\mu_2 = 2.0$	$\sigma_1^2 = 1.0$ $\sigma_2^2 = 0.25$
Sample 2 Two components	$a_1 = 0.85$ $a_2 = 0.15$	$\mu_1 = 0.0$ $\mu_2 = 2.0$	$\sigma_1^2 = 1.0$ $\sigma_2^2 = 0.25$
Sample 3 Three components	$a_1 = 0.33$ $a_2 = 0.33,$ $a_3 = 0.34$	$\mu_1 = 0.0$ $\mu_2 = -1.0,$ $\mu_3 = 4.0$	$\sigma_1^2 = 1.0$ $\sigma_2^2 = 0.25,$ $\sigma_3^2 = 4.0$

and Akaike information criteria (AIC) are calculated using the said parameters.

Step 3. Step 2 is repeated for other possible number K of components with $\mu_i = 0$, $\sigma_{ii} = 1$ for the other $K - K_0$ components, and $a_i = 1/K$.

Step 4. The log-likelihood function and AIC values for $K = 1, 2, \ldots, 10$ are plotted. The final number of components K_f is chosen when adding extra components in the model does not significantly increase or decrease the values of the log-likelihood function and the AIC, respectively.

4. Simulation

We use simulation to investigate the performance of the proposed modified algorithm.

4.1. Simulation Scheme. Simulation data were generated using the Box and Muller Transformation [18] as defined by (9) below:

$$\begin{aligned} z_j &= \mu + \left(-2\sigma^2 \log u_j\right)^{1/2} \cos 2\pi u_{j+1}, \\ z_{j+1} &= \mu + \left(-2\sigma^2 \log u_j\right)^{1/2} \sin 2\pi u_{j+1}, \end{aligned} \tag{9}$$

where $u_j, u_{j+1} \sim U(0,1)$. For the case of two components, we start by generating a random number $u_1 \sim U(0,1)$. If $0 < u_1 < a_1$, we generate two random numbers $u_2 \sim U(0,1)$ and $u_3 \sim U(0,1)$ and calculate $z_2 + z_3$ using the first and second equations of (9) with $\mu^* = \mu_1/2$ and $\sigma^* = \sigma_1/\sqrt{2}$. Otherwise, we use $\mu^* = \mu_2/2$, and $\sigma^* = \sigma_2/\sqrt{2}$. The process continues until the required sample size is obtained. The scheme is easily extended to any number of components. For further details, refer to [19].

4.2. Study of Performance Based on the Log-Likelihood Function. We first look at the performance of the standard method, called Method 1, followed by that of the modified method, called Method 2. For Method 1, in place of Step 1 of the modified method, we assign values zero and unity, respectively, to the means and variances of all components. We compare the performance by looking at the log-likelihood function via simulation study.

Following [20], we consider two cases with two and one case with three components with the true values of the parameters given in Table 1. For each case, we generate 100 samples of size 1000 where the chosen sample size reflects the large size of data sets found in the telecommunication industry, the focus of our interest. We then apply Method 1 and Method 2 on the simulated data. For each case, for better quality viewing, we plot only 50 values of the log-likelihood function for both methods on the same plot, as given in Figure 2. It can be seen that, for Samples 1 and 3, the proposed Method 2 clearly outperforms the standard Method 1 with the values of the log-likelihood function corresponding to Method 2 being always larger than those of Method 1. However, we see that some values overlap for Sample 2, though the proposed Method 2 still generally performs better. In this case, the prior probabilities a_i are distinctly different from the chosen values of a_i in Sample 1 while other true values remain the same which leads to different percentages of overlapping of the Gaussian components in the GMM. Hence, we will investigate the performance of the improved EM algorithm in estimating the parameters of the GMM by taking into account the effect of different percentages of overlapping between the components observed in the data.

4.3. The Effects of Different Overlapping Percentages on Performance. The main objective here is to investigate the performance of the modified EM algorithm for different overlapping percentages of the components in the GMM. For simplicity, we restrict our attention to two components so that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6) = (a_1, a_2, \mu_1, \mu_2, \sigma_{11}, \sigma_{22})$ are to be estimated. Data is simulated using the simulation scheme described in Section 4.1.

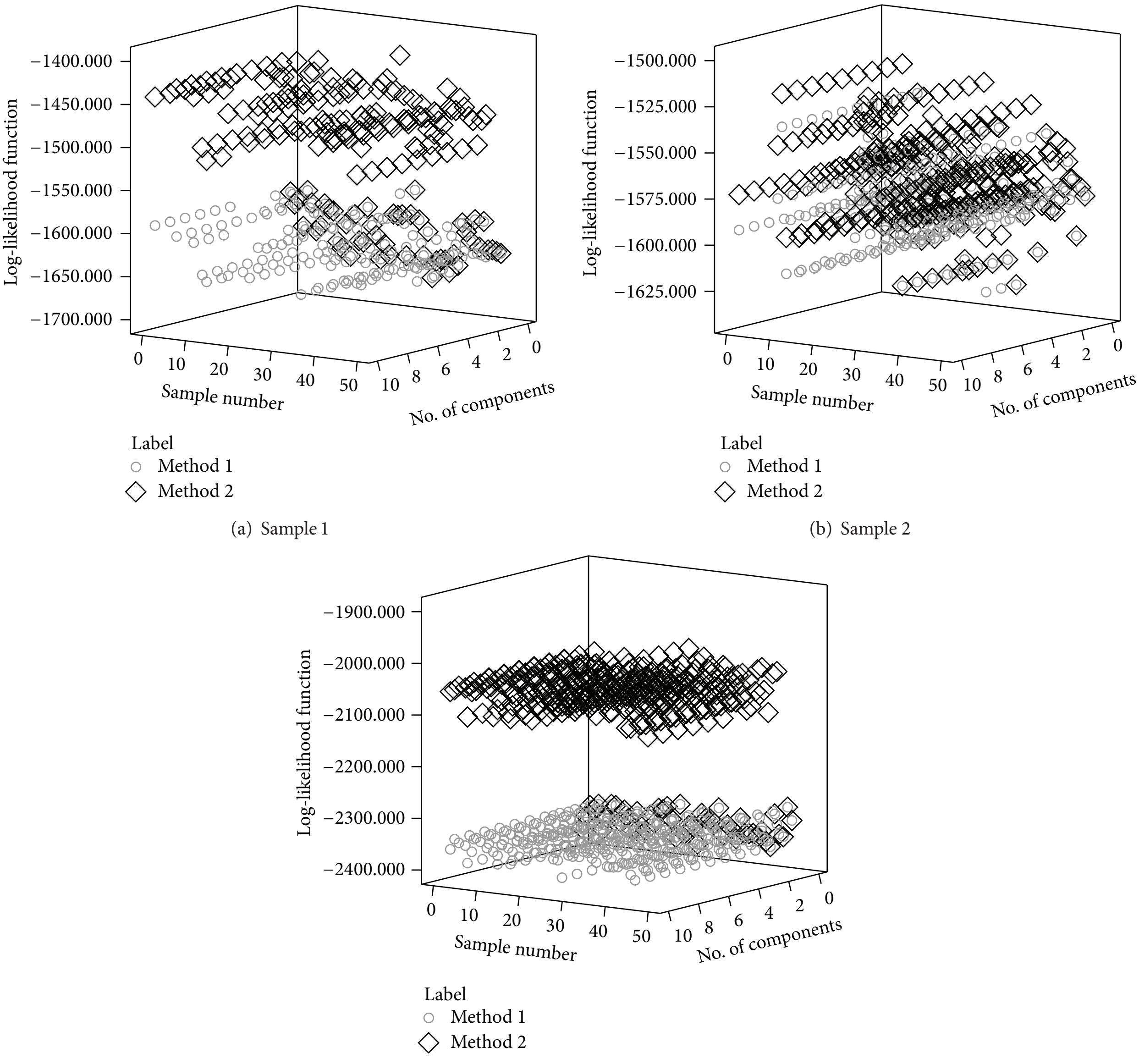
After performing Steps 1 and 2, we find $D_i = \theta_i - \widehat{\theta}_i$, where θ_i is the true value of the ith parameter and $\widehat{\theta}_i$ is the EM estimate of the parameter, $i = 1, 2, \ldots, n$. The sample mean and standard deviation of D_i are computed using the formulas $\overline{D} = (1/n)\sum_{i=1}^{n} D_i$ and $S_D = \sqrt{1/(n-1)\sum_{i=1}^{n}(D_i - \overline{D})^2}$. The estimates are considered good if $\overline{D}$ is close to zero, indicating small biases observed in the simulation results, and S_D is also close to zero, indicating that the parameter estimates are concentrated around their respective true values.

We determine the area of overlapping between the two components for each model by using the misclassification concept given in [21], the details of which are provided in Appendix B. The formula to estimate the overlapping areas depends on the mean and standard deviation of the components. The effects of prior probabilities should not affect the estimates greatly as their sum equals unity.

We consider three cases for different combinations of parameter $\boldsymbol{\theta}$ which give different percentages of overlapping of the GMM components. The results are tabulated in Tables 2–4. Table 2 deals with case 1, where the true values of $\mu_1 = 0$, $\mu_2 = 3.0$, and $\sqrt{\sigma_{11}} = \sqrt{\sigma_{22}} = 0.316$ are fixed but the true values of a_1 and a_2 are varied. In all cases, the percentage of overlapping is 0% as the separation of the means is rather large with small values of dispersion. We can see that the values of the mean are close to zero with the small standard errors less than unity for all parameters considered. On the other hand, Table 3 gives the results for case 2 where $\mu_1 = 0$,

Table 2: Simulation results for the case $\mu_1 = 0$, $\mu_2 = 3.0$, and $\sqrt{\sigma_1^2} = \sqrt{\sigma_2^2} = 0.316$.

Prior prob.		Bias, D_i											
		a_1		a_2		μ_1		μ_2		σ_1^2		σ_2^2	
a_1	a_2	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D
0.1	0.9	−0.001	0.010	0.001	0.010	0.007	0.033	−0.003	0.013	0.000	0.014	0.001	0.004
0.2	0.8	0.002	0.013	−0.002	0.013	0.002	0.022	−0.003	0.011	0.000	0.011	0.002	0.007
0.3	0.7	−0.002	0.014	0.002	0.014	0.004	0.017	−0.002	0.010	0.002	0.009	0.001	0.005
0.4	0.6	0.003	0.020	−0.003	0.020	−0.004	0.019	−0.006	0.012	0.000	0.009	0.001	0.005



(a) Sample 1

(b) Sample 2

(c) Sample 3

Figure 2: Plots of values of log-likelihood function.

$\mu_2 = 1.0$, $\sqrt{\sigma_{22}} = 0.707$, and $\sqrt{\sigma_{11}} = 0.447$ are fixed but a_1 and a_2 are varied to give 25% of overlapping. The bias is still considered small but generally larger than that for case 1. In addition, the values are also more dispersed here. Finally Table 4 shows the results of case 3 where $\mu_1 = 0$, $\mu_2 = 0.25$, $\sqrt{\sigma_{11}} = 0.577$, and $\sqrt{\sigma_{22}} = 1.414$ are fixed with 45% of overlapping. As expected, the results deteriorate when the percentage of overlapping increases. We conclude that the modified EM algorithm for GMM performs well when the percentages of overlapping are small, but its performance is

Table 3: Simulation results for the case $\mu_1 = 0$, $\mu_2 = 1.0$, $\sqrt{\sigma_2^2} = 0.707$, and $\sqrt{\sigma_1^2} = 0.447$.

Prior prob.						Bias, D_i							
		a_1		a_2		μ_1		μ_2		σ_1^2		σ_2^2	
a_1	a_2	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D
0.1	0.9	−0.112	0.148	0.112	0.148	−0.130	0.203	−0.067	0.139	−0.030	0.103	0.034	0.068
0.2	0.8	−0.014	0.073	0.014	0.073	−0.006	0.087	0.006	0.091	0.022	0.056	0.005	0.057
0.3	0.7	0.020	0.087	−0.020	0.087	0.031	0.069	0.023	0.104	0.044	0.048	−0.006	0.075
0.4	0.6	0.067	0.075	−0.067	0.075	−0.112	0.444	0.145	0.232	−0.002	0.085	−0.021	0.099

Table 4: Simulation results for the case $\mu_1 = 0$, $\mu_2 = 0.25$, $\sqrt{\sigma_1^2} = 0.577$, and $\sqrt{\sigma_2^2} = 1.414$.

Prior prob.						Bias, D_i							
		a_1		a_2		μ_1		μ_2		σ_1^2		σ_2^2	
a_1	a_2	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D	$\overline{D}$	S_D
0.1	0.9	0.089	0.007	−0.089	0.007	−0.817	3.770	0.045	0.048	−0.341	0.823	0.305	0.095
0.2	0.8	0.157	0.108	−0.157	0.108	0.291	3.521	0.075	0.073	−0.169	0.374	0.413	0.256
0.3	0.7	0.237	0.100	−0.237	0.100	−0.205	2.924	0.105	0.121	−0.278	0.389	0.578	0.278
0.4	0.6	0.245	0.187	−0.245	0.187	0.602	2.520	0.092	0.108	−0.008	0.258	0.508	0.435

affected when the percentages increase. More comprehensive simulation results can be obtained from the authors upon request.

5. Determination of the Final Number of Components in the GMM

In the last two steps of the modified algorithm, we intend to confirm that the choice of the initial number K_0 of components in the GMM using kernel method is final. This can be done by considering extra components in the model. For that, as stated in Section 3.3, we repeat Step 2 for other possible number K of components, by setting $\mu_i = 0$, $\sigma_{ii} = 1$ for the other $K - K_0$ components, and $a_i = 1/K$. The final number of components K_f is determined when adding extra component neither increases the log-likelihood nor decreases the AIC values significantly. The changes can easily be seen on a line plot of the values.

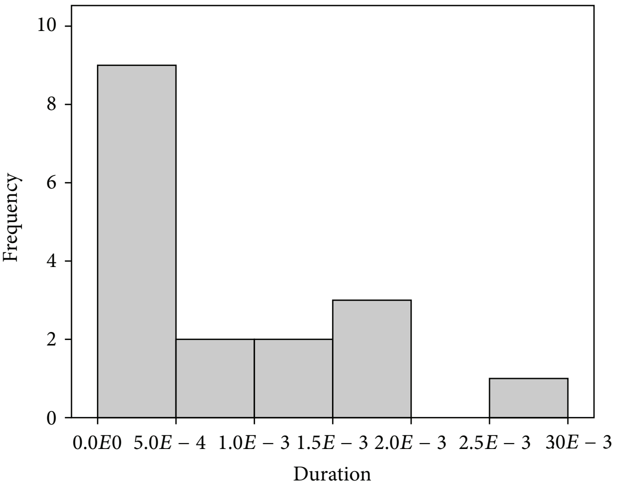
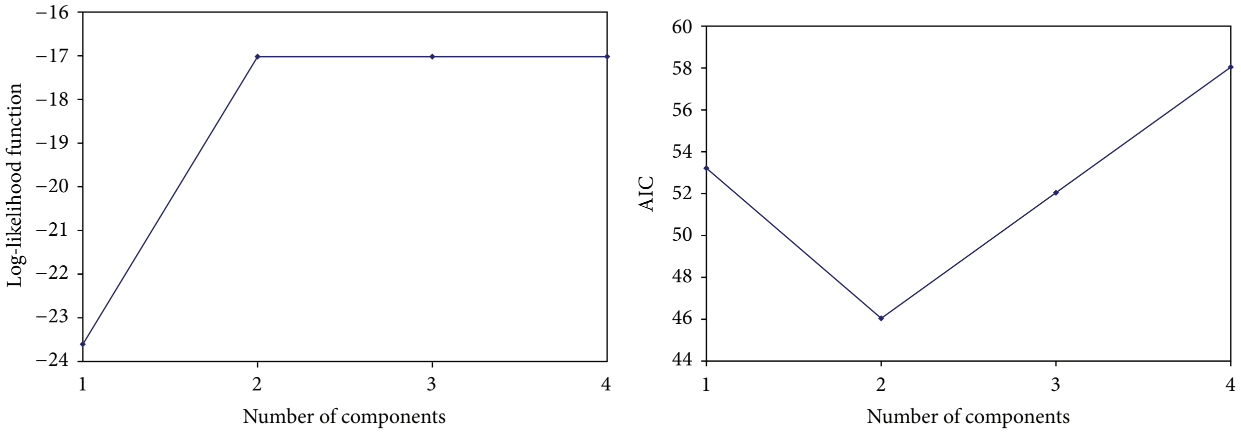


Figure 3: Duration (in day format) is displayed in the histogram.

6. Real Example: Phone Call Data

The call detail record, which was supplied by Telekom Malaysia Berhad (henceforth, TM), consists of calls made by customers that fell victim to fraud activities. Table 5 shows the format of the call detail record for each TM's customer. We performed several steps on the original data in order to have the data in a desired format, that is, group the real data according to service number, find the country that matches with the country code as well as dialed digits, and sort the real data according to seize time. The column entitled "Seize time" gives the time when the call was made; the fourth column details the duration of the calls in the following format: hour (hh), minute (mm), and second (ss); and the fifth column is the result of converting the information in the fourth column into day format.

We consider real data consisting of the converted duration of each call made by Customer A (referring to the fifth column of Table 5), whose identity is not revealed to ensure confidentiality, on March 31, 2011. Seventeen (17) calls were made and the data are displayed in Figure 3. Step 1 of the improved EM algorithm for GMM identifies two initial components. The plots of the log-likelihood function and AIC in Figures 4(a) and 4(b) are the results from performing Steps 2, 3, and 4 of the improved EM algorithm for GMM, which reveal that the EM algorithm fails to achieve convergence when the number of components equals to five or above. It can also be seen that a GMM with 2 components is identified as the "best" model, since the inclusion of more components not only fails to increase the value of the log-likelihood but also fails to decrease the values of the AIC. The final EM estimates for the two-component GMM are $\widehat{a}_1 = 0.64$, $\widehat{a}_2 = 0.36$, $\widehat{\mu}_1 = -0.66$, $\widehat{\mu}_2 = 1.17$, $\widehat{\sigma}_{11} = 0.07$, and $\widehat{\sigma}_{22} = 0.35$, and they represent the behavior of calls made by Customer A on March 31, 2011. In a future paper, we will show



(a) Log-likelihood

(b) AIC

FIGURE 4: Plots of log-likelihood and AIC values.

TABLE 5: An extract from Customer A' call detail record.

Service number	Dialed digits	Seize time	Duration (hhmmss)	Converted duration
Xxx	yyy	8:41:37	000339	$2.53E-03$
Xxx	yyy	9:27:03	000035	$4.05E-04$
Xxx	yyy	9:43:46	000048	$5.56E-04$
Xxx	yyy	9:50:21	000031	$3.59E-04$
Xxx	yyy	10:54:30	000138	$1.13E-03$

how the above information produced from the improved EM algorithm for GMM can be used in the process of detecting fraud activities in the telecommunication industry.

7. Conclusion

In this paper, we proposed a modified EM algorithm which can numerically identify the number of components of a GMM and estimate the parameters of the model using the kernel method. We showed via simulation that the performance of the algorithm is generally good but, as expected, is affected by increasing percentages of overlapping of the Gaussian components. We then used the line plots of the log-likelihood and AIC values to identify the final number of GMM components. They could clearly be determined via the concave-like shape of the AIC plot which indicates that the AIC decreases to a minimum value and then increases as the number of components increases. Finally, the modified EM algorithm for GMM was tested on real telecommunication data. The results serve as testimony to the effectiveness of the improved EM algorithm for GMM and should be useful when considering the problem of fraud calls faced by the telecommunication companies.

Appendices

A. Derivation of the First, Second, and Third Equations of (7)

(A.1) Using Lagrange multipliers defined by max/min $F(x,y,z)$ subject to $\Phi(x,y,z) = 0$, $G(x,y,z) = F(x,y,z) + \lambda\Phi(x,y,z)$, $\partial G/\partial x = 0$, $\partial G/\partial y = 0$, $\partial G/\partial z = 0$ [22] on $\max\sum_i \sum_j p_{ij}\log(a_j)$ subject to $\sum_j a_j = 1$ (or $(\sum_j a_j - 1) = 0$), we get the first equation of (7).

(A.2) From $(\partial/\partial\boldsymbol{\mu}_j)((1/2)\sum_i\sum_j p_{ij}(\mathbf{x}_i - \boldsymbol{\mu}_j)^t\Sigma_j^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_j)) = 0$, we get the second equation of (7) by using the following matrix properties: $\partial\mathbf{x}^t\mathbf{A}\mathbf{y}/\partial\mathbf{x} = \mathbf{A}\mathbf{y}$, $\partial\mathbf{a}^t\mathbf{x}/\partial\mathbf{x} = \mathbf{a}$.

(A.3) The first and second expressions of

$$\frac{\partial}{\partial\Sigma_j^{-1}}\left(\frac{1}{2}\sum_i\sum_j p_{ij}\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)^t \sum_j^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)\right) + \frac{\partial}{\partial\Sigma_j^{-1}}\left(\frac{1}{2}\sum_i\sum_j p_{ij}\log\left|\sum_j^{-1}\right|\right) = 0 \tag{A.1}$$

use the following matrix properties: $\partial\,\mathrm{tr}(\mathbf{x}\mathbf{y})/\partial\mathbf{x} = \mathbf{y} + \mathbf{y}^t - \mathrm{Diag}(\mathbf{y})$ and $\sum\mathbf{x}_i^t\mathbf{A}\mathbf{x}_i = \mathrm{tr}(\mathbf{A}\sum\mathbf{x}_i\mathbf{x}_i^t)$ to get the third equation of (7) [8].

B. The Value of Intersections

For the case when $\mu_1 \neq \mu_2$ and $\sigma_1 \neq \sigma_2$, $f_1(x) = (1/\sigma_1\sqrt{2\pi})\,e^{-(1/2)((x-\mu_1)/\sigma_1)^2}$ and $f_2(y) = 1/\sigma_2\sqrt{2\pi}e^{-(1/2)((y-\mu_2)/\sigma_2)^2}$ are obtained from $x_{11} = (-b + \sqrt{b^2 - 4ac})/2a$, and $x_{12} = (-b - \sqrt{b^2 - 4ac})/2a$ where $a = (\sigma_2^2 - \sigma_1^2)$, $b = 2(\sigma_1^2\mu_2 - \sigma_2^2\mu_1)$ and $c = (\sigma_2^2\mu_1^2 - \sigma_1^2\mu_2^2) - 2\sigma_1^2\sigma_2^2\log(\sigma_2/\sigma_1)$.

Firstly, using the above formula as well as $P((x-\mu)/\sigma) = \int_{-\infty}^{(x-\mu)/\sigma}(1/\sqrt{2\pi})e^{-(1/2)t^2}\,dt$, we find the area between x_{11} and

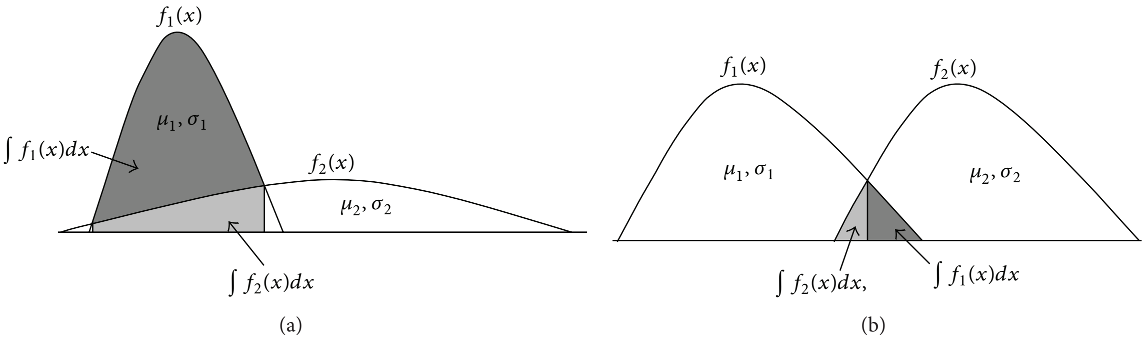


Figure 5: $\int f_1(x)dx$ and $\int f_2(x)dx$ are used to find the shaded areas.

x_{12} (and convert it into percentage) for each component; refer to Figure 5(a). Secondly, we find the minimum between the areas of the two components. This value represents the percentage of overlapping between two components (which is an approximation). For the case when $\mu_1 \neq \mu_2$, $\sigma_1 = \sigma_2$, $f_1(x) = (1/\sigma_1\sqrt{2\pi})e^{-(1/2)((x-\mu_1)/\sigma_1)^2}$, and $f_2(y) = (1/\sigma_2\sqrt{2\pi})e^{-(1/2)((y-\mu_2)/\sigma_2)^2}$, let $d = (\mu_1 + 2\sigma_1) - (\mu_2 - 2\sigma_2)$. The value of the intersection, say x_1, is obtained from the following formula (which is an approximation):

$$x_1 = \begin{cases} 0, & d < 0, \\ (\mu_1 + 2\sigma_1), & d = 0, \\ (\mu_1 + 2\sigma_1) - \dfrac{d}{2}, & d > 0. \end{cases} \qquad \text{(B.1)}$$

Taking similar steps, the area for the component on the left hand side of Figure 5(b) is obtained from $1 - P((x_1 - \mu_1)/\sigma_1) = 1 - \int_{-\infty}^{(x_1-\mu_1)/\sigma_1}(1/\sqrt{2\pi})e^{-(1/2)t^2}dt$ and that of the component on the right hand side of Figure 5(b) from $P((x_1 - \mu_2)/\sigma_2) = \int_{-\infty}^{(x_1-\mu_2)/\sigma_2}(1/\sqrt{2\pi})e^{-(1/2)t^2}dt$. We convert them into percentages before adding them up to represent the percentage of overlapping between two components (which is an approximation).

Conflict of Interests

Mohd Izhan Mohd Yusoff, the main author of the paper worked in Telekom Research & Development Sdn Bhd. (a subsidiary of Telekom Malaysia Berhad) for almost 20 years, the same organization that is sponsoring his Ph.D. study at the Institute of Mathematical Sciences, University of Malaya. The title of his study was proposed by two supervisors, who are the coauthors of the paper, namely, Ibrahim Mohamed and Mohd Rizam Abu Bakar. Due to the affiliation of the first author to Telekom Malaysia Berhad, the authors are able to obtain the relevant data sets to illustrate the development of the theory considered in this paper. In other words, Telekom Malaysia Berhad, as mentioned in the paper, is merely supplying the customer's call detail record and does not provide any financial contribution towards the project. The data analysis and the paper preparation have been carried out independently. In conclusion, there is no potential conflict of interests in the study.

Acknowledgments

The authors would like to express their thanks to Telekom Malaysia Berhad for their cooperation in this study. This research is financially supported by the Fundamental Research Grant Scheme (no. FP012-2013A).

References

[1] R. Jacobs, "Telecommunications Fraud: the single biggest cause of revenue loss for telecommunication providers," White Paper, Dimension Data, 2002.

[2] S. McClelland, "Plumbing network leaks: telecom fraud, perhaps the ultimate in virtual white collar crime, is alive, well, and on the increase," in *Commentary—Industry Overview*, Telecommunications International, 2003.

[3] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[4] M. Tanigushi, M. Haft, J. Hollmen, and V. Tresp, "Fraud detection in communications networks using neural and probabilistic methods," in *Proceeding of the IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 2, pp. 1241–1244, 1998.

[5] 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*, vol. 39, no. 1, pp. 1–38, 1977.

[6] J. Hollmen and V. Tresp, "Call based fraud detection in Mobile communication networks using a hierarchical regime-switching model," in *Proceedings of the Advances in Neural Information Processing Systems II Conference (NIPS 'II)*, M. Kearns, S. Solla, and D. Cohn, Eds., pp. 889–895, MIT Press, 1998.

[7] J. Hollmen, V. Tresp, and O. Simula, "A learning vector quantization algorithm for probabilistic models," in *Proceedings of the European Signal Processing Conference (EUSIPCO '00)*, vol. 2, pp. 721–724, 2000.

[8] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, London, UK, 1979.

[9] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models," Tech. Rep. ICSI-TR-97-021, University of Berkeley, 1998.

[10] R. S. Tsay, *Analysis of Financial Time Series: Financial Econometrics*, John Wiley & Sons, 2005.

[11] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, UK, 1986.

[12] P. Schlattmann, “Estimating the number of components in a finite mixture model: the special case of homogeneity,” *Computational Statistics and Data Analysis*, vol. 41, no. 3-4, pp. 441–451, 2003.

[13] H. X. Wang, B. Luo, Q. B. Zhang, and S. Wei, “Estimation for the number of components in a mixture model using stepwise split-and-merge EM algorithm,” *Pattern Recognition Letters*, vol. 25, no. 16, pp. 1799–1809, 2004.

[14] M. Miloslavsky and M. J. van der Laan, “Fitting of Mixtures with unspecified number of components cross validation distance estimate,” *Computational Statistics and Data Analysis*, vol. 41, no. 3-4, pp. 413–428, 2003.

[15] X. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao, “Gaussian mixture density modeling, decomposition, and applications,” *IEEE Transactions on Image Processing*, vol. 5, no. 9, pp. 1293–1302, 1996.

[16] Y. Lee, K. Y. Lee, and J. Lee, “The estimating optimal number of Gaussian Mixtures based on incremental k-means for Speaker Identification,” *International Journal of Information Technology*, vol. 12, no. 7, pp. 13–21, 2006.

[17] G. Celeux and G. Soromenho, “An entropy criterion for assessing the number of clusters in a mixture model,” *Journal of Classification*, vol. 13, no. 2, pp. 195–212, 1996.

[18] G. E. P. Box and M. E. Muller, “A note on the generating of random normal deviates,” *Annals of Mathematical Statistics*, vol. 29, pp. 610–611, 1958.

[19] G. S. Fishman, *Discrete-Event Simulation: Modeling, Programming, and Analysis*, Springer, 2001.

[20] B. S. Everitt and D. J. Hand, *Finite Mixture Distributions*, Chapman & Hall, London, UK, 1981.

[21] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, 1998.

[22] M. R. Spiegel, *Theory and Problems of Advanced Calculus: SI*, Shaum’s Outline Series, McGraw-Hill, 1974.