# Analysis of stopping criteria for the EM algorithm in the context of patient grouping according to length of stay

**Revlin Abbi**[1]
**Elia El-Darzi**[1]
**Christos Vasilakis**[2]
**Peter Millard**[3]

[1] Harrow School of Computer Science, University of Westminster
[2] Clinical Research Unit, University College London
[3] St. George’s University of London

# Analysis of stopping criteria for the EM algorithm in the context of patient grouping according to length of stay

Revlin Abbi, Elia El-Darzi, Christos Vasilakis, and Peter Millard

***Abstract*— The expectation maximisation (EM) algorithm is an iterative maximum likelihood procedure often used for estimating the parameters of a mixture model. Theoretically, increases in the likelihood function are guaranteed as the algorithm iteratively improves upon previously derived parameter estimates. The algorithm is considered to converge when all parameter estimates become stable and no further improvements can be made to the likelihood value. However, to reduce computational time, it is often common practice for the algorithm to be stopped before complete convergence using heuristic approaches. In this paper, we consider various stopping criteria and evaluate their effect on fitting Gaussian mixture models (GMMs) to patient length of stay (LOS) data. Although the GMM can be successfully fitted to positively skewed data such as LOS, the fitting procedure often requires many iterations of the EM algorithm. To our knowledge, no previous study has evaluated the effect of different stopping criteria on fitting GMMs to skewed distributions. Hence, the aim of this paper is to evaluate the effect of various stopping criteria in order to select and justify their use within a patient spell classification methodology. Results illustrate that criteria based on the difference in the likelihood value and on the GMM parameters may not always be a good indicator for stopping the algorithm. In fact we show that the values of the difference in the variance parameters should be used instead, as these parameters are the last to stabilise. In addition, we also specify threshold values for the other stopping criteria.**

***Index Terms*—GMM fitting, LOS data, patient classification, stopping criteria.**

## I. Introduction

Various techniques have been proposed that use patient length of stay (LOS) data to derive the parameters of patient flow models, which in turn help clinicians and managers to better understand the temporal characteristics of the patients cared for by the health care system [1-5]. An approach complementary to flow modelling and developed by the authors is concerned with deriving the case-mix of patients from LOS observations, and building a LOS patient classification model [6, 7]. In summary, the methodology comprises of several processing steps [6, 7], where the optimal Gaussian mixture model (GMM), based on the minimum description length criterion [8], represents various groups of patients according to their LOS. From the derived GMM, non-overlapping LOS intervals (the classification scheme) are derived and a decision tree is built. In this way, patients are grouped according to their LOS and for each group a profile is derived to help predict a patient's LOS based on various spell characteristics such as age, diagnosis, gender, and others. In this paper we are concerned with determining the parameters of each group using the maximum likelihood approach within a computationally efficient time.

For the single Gaussian function, we are able to obtain a closed form solution to derive the maximum likelihood estimates of the model parameters. In more complex cases however, computational approaches are often used for maximising the likelihood function [9], most commonly the expectation maximization (EM) algorithm [10]. Generally speaking, EM is a computational maximum likelihood procedure used to estimate the parameters of a mixture model. Once initialised, the algorithm readjusts the GMM parameter estimates while guaranteeing increases in the likelihood function. The algorithm is considered to converge when all estimates remain the same and no further improvements can be made.

However, in order to reduce computational time, the EM algorithm is often stopped prior to convergence using alternative stopping criteria. A simplistic approach is to heuristically specify the number of iterations before applying the algorithm [11, 12]. However, this trial and error approach is problem dependant as there is no generally applicable number of iterations before we terminate the algorithm. A better approach is to detect the amount of improvement being made to the likelihood function between successive iterations, and then stop the algorithm when only small improvements are being made, based on a threshold value as a predefined cut-off point [13-16]. Other stopping criteria are based on the

R. Abbi is with the Department of Computer Science, University of Westminster, Harrow, London, HA1 3TP, UK (email: abbiR@westminster.ac.uk)

E. El-Darzi is with the Department of Computer Science, University of Westminster, Harrow, London, HA1 3TP, UK (email: eldarze@westminster.ac.uk)

C. Vasilakis is with the Clinical Research Unit, University College London, London, WC1H 0BT, UK (email: c.vasilakis@ucl.ac.uk)

P. Millard is an Emertus Professor, St. George's University of London, London, UK (email: phmillard@tiscali.co.uk)

978-1-4244-1739-1/08/$25.00 © 2008 IEEE

changes in the GMM parameters (or part of the parameters), between consecutive iterations [17].

Based on experimental analysis, such approaches often lead to an underestimation of the model parameters. To our knowledge there is no study which evaluates the effect of different stopping criteria on fitting GMMs to skewed distributions such as the LOS distribution. In this paper therefore, various existing stopping criteria based on either the likelihood value or on the parameters of the GMM are evaluated according to the effect on fitting GMMs to patient LOS data. The study uses data from two health administrative datasets. The aim is to understand, define and compare the impact of various stopping criteria in order to justify their use within the patient spell classification methodology. A related issue to the stopping criterion is the initialisation of the model parameters, which also significantly affect computational time. Hence, we introduce an initialisation approach, shown to decrease the number of iterations, based on percentile values derived from the LOS data. This can be used instead of random initialisation.

Results illustrate that criteria based on the difference in the likelihood values and on the GMM parameters may not always be a good indicator for stopping the algorithm. In fact we show that the values of the difference in the variance parameters should be used instead, as these parameters are the last to stabilise.

The rest of the paper is organised as follows. In the next section we briefly describe the approaches used and the criteria evaluated in our study. We also introduce the two healthcare datasets used in this paper. In the results section, we report on the outcomes of the study and assess the impact of applying the various criteria. Lastly, we end the report with some concluding remarks.

## II. METHODS

Using a given set of data $X=\{x_1, x_2,\ldots, x_N,\}$ corresponding to LOS observations, the maximum likelihood approach determines the parameters of a GMM that maximise the likelihood function $L$. The likelihood $L$ of the data $X$ is defined as the product of the probabilities for each data point $x_i$, defined in (1), where $N$ is the number of LOS observations, and $p(x)$ is the probability of a patient staying $x$ days, according to a fitted GMM with $m$ components (2). Within (2), $\omega_j$ is the mixing coefficient for component $j$, representing the percentage of patients belonging to group $j$ and $p(x|j)$ is the conditional probability of the Gaussian component $j$ being distributed according to the mean and variance for group $j$, Equation 3. In addition, the posterior probability of a LOS observation $x_i$ "belonging" to component $j$ is derived using the Bayes rule, Equation 4, where $\omega_j$ can be seen as the prior probability for group $j$.

$$L(X) = \prod_{i=1}^{N} p(x_i) \tag{1}$$

$$p(x) = \sum_{j=1}^{m} \omega_j \, p(x|j) \tag{2}$$

$$p(x|j) = \frac{1}{\sqrt{2\pi\sigma_j{}^2}} exp\left(\frac{(x-\mu_j)^2}{2\sigma_j{}^2}\right) \tag{3}$$

$$p(j|x) = \frac{\omega_j \, p(x|j)}{\sum_{j=1}^{m} \omega_j \, p(x|j)} \tag{4}$$

Based on the LOS observations $X$, the objective of the EM algorithm (see Figure 1) is to provide an iterative computation of the maximum likelihood estimate of the model parameters.

Step 1. Initialise parameters

Step 2. Expectation-step

Compute the posterior probabilities using Bayes Rule (Equation 4) for all LOS data i.e. $i=1,\ldots, N$ and $j=1,\ldots, m$

Step 3. Maximisation-step

$$\omega_j{}^{(t+1)} = \frac{\sum_{i=1}^{N} p(j|x_i)}{N} \tag{5}$$

$$\mu_j{}^{(t+1)} = \frac{\sum_{i=1}^{N} p(j|x_i) x_i}{\sum_{i=1}^{N} p(j|x_i)} \tag{6}$$

$$\sigma_j{}^{2(t+1)} = \frac{p(j|x_i)(x_i-\mu_j{}^{(t+1)})(x_i-\mu_j{}^{(t+1)})}{\sum_{i=1}^{N} p(j|x_i)} \tag{7}$$

Step 4. Repeat step 2 and 3 until convergence

Fig. 1. The EM algorithm.

For all experiments, we initialised the EM algorithm (step 1 above) based on the solution derived from the *k*-means clustering algorithm [18], where the mean, variance and the mixing coefficients are derived from each of the estimated clusters. Percentile values derived from the LOS data were used as mean parameter inputs for the *k*-means algorithm. We set the mean of each of the $m$ groups equal to a percentile value (denoted as *pc)* of the LOS data defined according to (8). Based on experimental analysis, this form of initialisation was shown to decrease the number of iterations needed, due to the fact that the initialisation parameters are selected according to the nature of the GMM for modelling positively skewed distributions such as LOS. From the *k*-means cluster model, mean estimates were derived from cluster centres, as defined in (9), where $x_j$ is a LOS observation belonging to cluster $j$ and $N_j$ is the number of LOS observations belonging to cluster $j$. The variance $\sigma_j^2$ for component $j$ is defined in (10). The mixing coefficient $\omega_j$ for Gaussian component $j$ is valued according to the number of LOS values that belong to the cluster $j$, defined in (11).

$$c_j = pc\left(\frac{100}{m+1} * j\right) \tag{8}$$

$$\mu_j = \frac{\sum X_j}{N_j} \tag{9}$$

$$\sigma_j^2 = \frac{\sum(\mu_j - x_j)^2}{N_j} \tag{10}$$

$$\omega_j = \frac{N_j}{N} \qquad (11)$$

Although the objective of the maximum likelihood approach is to increase the likelihood function, in practice we maximise the logarithm of the likelihood (log-likelihood). This is equivalent to maximising the likelihood directly, as the logarithm is a monotonically increasing function of its argument. It is convenient to use the log-likelihood ($LL$ in equation 12) function because the product of a large number of small probabilities (extremely small floating-point values) can easily underflow the numerical precision of the computer [9].

$$LL = \sum_{i=1}^{N} \ln p(x_i) \qquad (12)$$

Whilst fitting the GMMs to the LOS data, if at any time the variance of component $j$ reduces below the threshold of $10^{-3}$, (i.e. step 3) then some perturbation (values of $10^{-7}$) was added to the data belonging to component $j$ [19]. This ensures that the variance does not converge to zero and thus the singularity problem can be avoided [9]. The divisor of zero (i.e. the variance in the Gaussian function) results in the probability of component $j$ to reach infinity. The value of $10^{-7}$ was chosen to ensure that very little impact is imposed on the derived model parameters, whilst enabling the EM algorithm to update the parameters of the other $m$-1 components.

We fitted various GMMs, from three to eight components, to LOS data and evaluated various stopping criteria for different threshold values. The stopping criteria were based on either the changes in the log-likelihood value, defined in (13), or changes in the model parameters $\mu_j$, $\sigma_j^2$, $\omega_j$, defined in equations 14, 15, and 16, respectively, where $t$ is the iteration number.

$$(ln(LL)^t - ln(LL)^{t+1}) < 10^{-\delta} \qquad (13)$$

$$\left[\left(\,|\mu_j^t - \mu_j^{t+1}|\right) < 10^{-\delta}\right];\ \forall\, j, j = 1, \dots, m \qquad (14)$$

$$\left[\left(\,|(\sigma^2)_j^t - (\sigma^2)_j^{t+1}|\right) < 10^{-\delta}\right];\ \forall\, j, j = 1, \dots, m \qquad (15)$$

$$\left[\left(|\omega_j^t - \omega_j^{t+1}|\right) < 10^{-\delta}\right];\ \forall\, j, j = 1, \dots m \qquad (16)$$

Based on the literature and our experience with EM we considered various threshold values for $\delta$ ($10^{-2}$, $10^{-4}$, $10^{-6}$, and $10^{-8}$) at which to stop the algorithm. For all stopping criteria considered, the number of iterations and the derived GMM parameter estimates were recorded. This information was then used to determine the effect on model parameters and the running time of using a given criterion as opposed to complete convergence. We also assessed how effective each criterion was in terms of the number of iterations, compared with complete convergence.

For notational purposes we use $\nabla_\delta^z(\omega_j,\mu_j,\sigma^2{}_j)$ to indicate the absolute difference in the GMM parameters for the $j^{th}$ component between using a stopping criterion $z$ (as defined in equations 13-16) and complete convergence.

The study used LOS data from two health administrative datasets. The first dataset, referred to here as the Surgical dataset, consists of 7723 records detailing the spells of patients undergoing surgery in a tertiary hospital in Australia between 4 February 1997 and 30 June 1998 [20]. In this dataset the mean LOS was 5.8 days with a standard deviation of 9.2 days, and a range of between one and 228 days. The second dataset is referred to here as the Stroke dataset, and originates from the English Hospital Episode Statistics (HES) database. It concerns all finished consultant episodes (FCE) of stroke patients, aged 65 and over discharged from all English hospitals between April 1st 1994 and March 31st 1995 [21]. The Stroke data consists of 103,881 LOS observations, with a mean of 14.0 days, a standard deviation of 52.3 days, and a range of between one and 4,907 days.

## III. RESULTS

Various GMMs consisting of three to eight components were fitted to both the Surgical and Stroke LOS data. The singularity problem was encountered whilst fitting the GMMs with six or more components to the Surgical data and five or more components to the Stroke data. In this case, the first component variance converged towards zero, with a mean LOS of one day. In addition, we observed that models with more components take longer to converge, Table 1.

Table 1. Number of iterations needed for convergence of the GMM with $m$ components

| # of components ($m$) | # iterations for Surgical data ($t$) | # iterations for Stroke data ($t$) |
|---|---|---|
| 2 | 214 | 85 |
| 3 | 204 | 266 |
| 4 | **589** | 406 |
| 5 | 1,094 | 533 |
| 6 | 2,452 | **2,389** |

For all the fitted GMMs in the two datasets, the behaviour of the criteria, in terms of the number of iterations needed to converge, was consistent. Therefore we selected the models with four components in the Surgical dataset and six components in the Stroke dataset. These groupings were optimal and shown to be representative of the LOS data [6, 22, 23]. They are used below for illustrative purposes. The optimal parameter estimates of the GMMs, based on complete convergence, were derived in 589 iterations for the Surgical data (four components), and 2,389 iterations for the Stroke data (six components), Table 1. The number of iterations for each stopping criterion is described in Table 2, where the parameter estimates for the complete convergence are described in Table 3 (Surgical data) and Table 4 (Stroke data).

The number of iterations for each stopping criterion differed extensively and the percentage of iterations compared with complete convergence points to a saving of computational time. Higher precision and more accurate parameter estimates were derived using smaller threshold values, obviously at the expense of computing time, Table 2.

Table 2. Number of iterations for each stopping criterion for the Surgical and Stroke dataset

| Criteria Type | Threshold Value ($\delta$) | No. of iterations Surgical Data | % of iterations | No. of iterations Stroke Data | % of iterations |
|---|---|---|---|---|---|
| $\omega$ | -2 | 10 | 1.7% | 7 | 0.3% |
| $\mu$ | -2 | 93 | 15.8% | 350 | 14.6% |
| $\sigma^2$ | -2 | 168 | 28.5% | 988 | 41.2% |
| $LL$ | -2 | 68 | 11.5% | 165 | 6.9% |
| $\omega$ | -4 | 80 | 13.6% | 214 | 8.9% |
| $\mu$ | -4 | 203 | 34.5% | 784 | 32.7% |
| $\sigma^2$ | -4 | 279 | 47.4% | 1,431 | 59.7% |
| $LL$ | -4 | 121 | 20.5% | 457 | 19.1% |
| $\omega$ | -6 | 188 | 31.9% | 629 | 26.2% |
| $\mu$ | -6 | 314 | 53.3% | 1,228 | 51.1% |
| $\sigma^2$ | -6 | 390 | 66.2% | 1,846 | 76.9% |
| $LL$ | -6 | 176 | 29.9% | 645 | 26.9% |
| $\omega$ | -8 | 299 | 50.7% | 1,073 | 44.7% |
| $\mu$ | -8 | 425 | 72.2% | 1,668 | 69.6% |
| $\sigma^2$ | -8 | 501 | 85.1% | 2,006 | 83.6% |
| $LL$ | -8 | 227 | 38.5% | 767 | 31.9% |
| Complete Convergence | | 589 | 100% | 2,398 | 100% |

Table 3. The optimal parameter estimates of the four-component GMM fitted to Surgical LOS data

| | Comp 1 | Comp 2 | Comp 3 | Comp 4 |
|---|---|---|---|---|
| $\omega$ | 0.382941891 | 0.394322113 | 0.188141358 | 0.034594636 |
| $\mu$ | 2.237109797 | 5.361430352 | 13.393280227 | 39.231905692 |
| $\sigma^2$ | 0.300740522 | 4.683453924 | 36.517687080 | 676.04638607 |

### *A. Stopping Criteria*

The mixing coefficient stopping criteria often stops the EM algorithm first, followed by the mean, and then the variance, Table 2. This results in fewer updates to the mean and variance parameters. In the specific case where the threshold value is set as *δ=2*, there is a large impact on the derived model parameters because of the fact that the EM algorithm is stopped after only a few iterations. Therefore we could conclude, that the threshold value of *δ=2* for the mixing coefficient criterion is not appropriate. Table 5 shows the 'absolute-difference' between the parameter estimates for *δ=2* and the optimal ones. In this case, there is a difference in the mixing coefficient of 0.156 for the first component, 0.043 for the second component, 0.085 for the third component, etc. For patient LOS groupings, such differences have a large impact on the interpretation of the LOS of patients. Based on the derived parameters, 15.6% of the population have been over estimated for group one, 4.3% for group two, 8.5% for group three, etc.

Smaller threshold values for the mixing coefficient i.e. where *δ=4* and *δ=6*, arrive at more accurate estimates in terms of the mean parameters of the shorter stay components. In the case where *δ=4*, the number of iterations is reduced by a minimum of 85%, compared with complete convergence, Table 2. However the mean of the sixth component is underestimated by 4.4 days for the Stroke data, i.e. $\nabla^{\omega}_{\delta=4}(\omega_6, \mu_6, {\sigma^2}_6)$ = (0.000050618535, 4.433971700618, 3681.300412265990). Although the longer stay groups are affected, in the case of the Surgical data, the difference is quite small i.e. 0.4 days for the mean of the last group, and 9 days for the variance. The absolute differences $\nabla^{\omega}_{\delta=4}(\omega_4, \mu_4, {\sigma^2}_4)$ = (0.000865944783, 0.393699359448, 9.358568474631).

With regards to the criterion based on the mean where *δ=2*, the variance of the longer stay components for both datasets tend to be underestimated. For the Stroke dataset, the mean is under estimated by one day, and the variance is underestimated by 742, $\nabla^{\mu}_{\delta=2}(\omega_6, \mu_6, {\sigma^2}_6)$ = (0.000010410438, 0.903770345244, 742.328224505996). To overcome this, at the expense of more iterations, a threshold value of *δ=4* may be used, $\nabla^{\mu}_{\delta=4}(\omega_6, \mu_6, {\sigma^2}_6)$ = (0.000000110011, 0.009528670325, 7.930092132010), which results in 7.9 overestimation of the variance. However, this tends to provide a balance between computational time whilst still being able to derive reasonably accurate parameter estimates for $\omega$ and $\mu$. However for the Surgical data, the difference is quite small, i.e. the difference of 0.05 in the variance for the last component. Smaller threshold values i.e. *δ=6* and *δ=8* may be used if a greater precision is of significance. For instance, *δ=6* provides accurate parameter estimates for the Stroke data, up to one decimal place, $\nabla^{\mu}_{\delta=6}(\omega_6, \mu_6, {\sigma^2}_6)$ = (0.000000001100, 0.000095259677, 0.079277967976).

Criteria based on the variance parameters are the most computationally expensive stopping criteria but provide accurate parameter estimates. In such cases, values such as *δ=2* produce reasonably accurate parameter estimates, similar to the *δ=4* or *δ=6* where the mean stopping criteria is used, $\nabla^{\sigma}_{\delta=2}(\omega_6, \mu_6, {\sigma^2}_6)$ = (0.000000013255, 0.001148110676, 8.504560780013). Threshold values of *δ=4* for the variance, tend to give very close estimates to the maximum likelihood estimates, $\nabla^{\sigma}_{\delta=4}(\omega_6, \mu_6, {\sigma^2}_6)$ = (0.000000000134, 0.000011575548, 0.009633265028).

Table 4. The optimal parameter estimates of the six-component GMM fitted to Stroke LOS data

| | Comp 1 | Comp 2 | Comp 3 | Comp 4 | Comp 5 | Comp 6 |
|---|---|---|---|---|---|---|
| $\omega$ | 0.1230542583 | 0.3033356781 | 0.3512135607 | 0.1619508790 | 0.0564964670 | 0.0039491567 |
| $\mu$ | 1.0000000000 | 4.8108552766 | 11.6246572037 | 25.0823316875 | 59.9468473021 | 488.4917157336 |
| $\sigma^2$ | 0.0000000000 | 4.6403628210 | 19.4153227730 | 94.2261690700 | 806.6491375100 | 402707.1197309430 |

Table 5. Absolute differences of parameter estimates regarding the six-component GMM for Stroke data, derived between using the mixing coefficient stopping criterion where *δ=2* compared with complete convergence

| Parameter | Comp 1 | Comp 2 | Comp 3 | Comp 4 | Comp 5 | Comp 6 |
|---|---|---|---|---|---|---|
| $\omega$ | 0.15616191524 | 0.04335332928 | 0.08498798740 | 0.06585144482 | 0.04631357792 | 0.00236223437 |
| $\mu$ | 1.24457437460 | 3.41286440122 | 5.98235460972 | 18.15016890400 | 62.46374976959 | 442.09746720871 |
| $\sigma^2$ | 1.79884499700 | 4.09575529100 | 25.68679251300 | 289.38932050300 | 2301.44403437100 | 267378.44082966200 |

Stopping criteria based on the log-likelihood value are more expensive than the mixing coefficient, but less expensive compared with the mean criteria, except for where *δ*=8 in both datasets. When using the log-likelihood criteria, *δ*=2 and *δ*=6 can often result in early stoppage, hence underestimation of the parameter values, for the Stroke data $\nabla^{LL}_{\delta=6}(\omega_6,\ \mu_6,\ \sigma^2{}_6)$ = (0.000000465857, 0.040353855745, 33.583330418973). Results have shown that it may not be appropriate to use the log-likelihood stopping criteria, unless the value of *δ*=8 or greater is adopted, $\nabla^{LL}_{\delta=8}(\omega_6,\ \mu_6,\ \sigma^2{}_6)$ = (0.000000131235, 0.011367070791, 9.460063125007). Even in this case, the variance of the last component has been underestimated by 9.4 for the stroke data.

## IV. DISCUSSION

In this study, the LOS distribution of two patient populations was modelled using the GMM while the EM algorithm was used to estimate the parameters of the fitted model. Conventionally, iterative maximum likelihood algorithms such as EM are considered to have converged when the parameters ($\mu_j$, $\sigma_j^2$, $\omega_j$) being estimated become stable and do not change for two consecutive iterations. However, this results in a large number of iterations, where each iteration of the algorithm has little improvement on the estimated parameters. An effective criterion for stopping the EM before convergence may result in both good parameter estimates (i.e. very close to the optimal values) and a reduction in computational time. The objective of this research was to consider various stopping criteria and to evaluate their effect on the fitting of GMMs to patient LOS data.

For any particular application, it is important to understand the behaviour of the EM algorithm for estimating the parameters of the GMM in order to be able to develop robust methodologies that are computationally efficient. The criteria described in this report are either based on the change in the likelihood value or on the change between parameter estimates of the GMM. A threshold value of $10^{-\delta}$ for various values of *δ* was adopted (values 2, 4, 6, and 8). For each criterion, smaller values for *δ* result in early termination of the algorithm at some cost to the model parameter estimates. Larger values of *δ* obviously result in more iterations of the EM algorithm but with better estimates of model parameters.

A related issue to the stopping criterion is initialisation, which significantly affects computational time. For example, if we initialised the algorithm with parameters that are close to the maximum likelihood solution, fewer iterations would be needed compared with model parameters that are far off. To avoid random initialisation, in this paper we introduced an initialisation approach based on percentile values derived from the LOS data. This has shown to be very effective.

### *A. Contrast to the literature*

Manual approaches that require the analyst to specify the number of iterations prior to parameter estimation is often used in many studies. For example, Gilland et al [11] specify 25 and 50 iterations, and Permuter et al [12] specify 30 iterations as stopping criteria. However, our study put forward more appropriate stopping criteria that are able to stop the EM algorithm according to its progress. Chandramouli and Srikantam [17] use the change in the mixing coefficients as an indictor to stop the EM algorithm considering values *δ>0*. However, our results showed that the mean and variance estimates may be underestimated when stopping the EM algorithm early, especially if *δ*<=4. However, for the LOS distribution, this criterion can underestimate GMM parameters.

In addition, it may not always be appropriate to use the likelihood function as a stopping criterion. Carson and Greenspan [13] suggested that the EM algorithm be stopped when the increase in the log-likelihood is less than 1%. They also specified that if this criterion is not met, a stop can be made at the tenth iteration. In our application, such criteria would lead to an underestimation of the parameters. Similarly, Roch et al [14] proposed to stop the EM algorithm when the increase in the log-likelihood is less than 2%. We found that this approach would result in underestimating the model parameters. As mentioned above, when using the likelihood criteria, *δ* should ideally be set to 8. As such, the work of Xing et al [15], which suggested that the EM algorithm be stopped when the increase in the log likelihood is less than $10^{-6}$ is adequate enough to reduce computation time and derive good parameter estimates. Zhang et al [16] considered $10^{-\delta}$, for various different values of *δ*, which is very important because the appropriate criterion can be problem dependant as shown by the difference between the two datasets in this paper.

### *B. Summary of Results*

In summary, we found that the components converge in ascending order of their mean value. Thus if we stop early, we are more likely to affect the longer stay group parameter estimates than those of the shorter stay groups. This could be because the variability in longer stay groups is larger than the variability in shorter stay groups.

Our findings also showed that the mixing coefficient of the GMM converges first, then the mean, and lastly the variance parameters. The mixing coefficient parameters stabilise quicker than mean and variance parameters, because the proportion of LOS observations that belong to each component of the GMM are determined fairly early whilst still computing the mean and variance parameters. As such, when the variance for each component becomes stable the mean and mixing coefficient parameters have already converged, hence the maximum likelihood estimates of the GMM parameters. The possible consequence of stopping early, is that the mean and variance parameters of the medium and longer stay groups will be affected rather than the estimates of the shorter stay groups.

The recommendations of this paper for stopping criteria are set out below.

- For the mixing coefficient stopping criteria only $\delta$=8 should be used, which reduces the number of iterations by about 49%. Any value below $\delta$=8 will result in the variance parameters being underestimated, especially for the longer stay groups.
- Criteria based on the mean are more reliable than the mixing coefficient. As such, $\delta$=6 may be used, which reduces the number of iterations by about 46%. On the other hand, $\delta$=4 was shown to derive reasonably accurate parameter estimates but at the expense of the last component.
- Criteria based on the variance are the most reliable. As such, $\delta$=2 and $\delta$=4 may be used although $\delta$=6 will produce the most accurate parameter estimates. The latter reduced the number of iterations by about 33% compared with complete convergence, but is a lot more expensive than $\delta$=2 or $\delta$=4.
- By using criteria based on the likelihood value, we take into account the relative changes of all GMM parameters. The experiments showed that a value no larger than $10^{-8}$ would be appropriate, because a small change in the likelihood value is not a good indicator of the parameter set convergence. This criterion resulted in approximately 61% fewer iterations than complete convergence (Table 2).

## REFERENCES

[1] Millard P.H, "Geriatric medicine: a new method of measuring bed usage and a theory for planning," *St. George's Hospital Medical School*: University of London, 1988.

[2] Harrison G.W and Millard P.H, "Balancing acute and long term care: the mathematics of throughput in departments of geriatric medicine," *Methods of Information in Medicine,* vol. 30, pp. 221-228, August 1991.

[3] Millard P.H, "Flow rate modelling: a method of comparing performance in departments of geriatric medicine," *St George's Hospital Medical School*: University of London, 1992.

[4] Harrison G.W, "Compartmental models of hospital patient occupancy patterns. In: Millard P.H, McClean S.I (eds). *Modelling hospital resource use: a different approach to the planning and control of health care systems.,*" *Royal Society of Medicine: London,* pp. 53-61, 1994.

[5] Faddy M.J and McClean S.I, "Analysing data on length of stay of hospital patients using phase-type distributions," *Applied Stochastic Models in Business and Industry,* vol. 15, pp. 311-317, 1999.

[6] Abbi R, El-Darzi E, Vasilakis C, and Millard P.H, "Intelligent methods for modelling patient flow," *The International Conference on Industrial Engineering and Systems Management (IESM 2007)*, Beijing CHINA., 2007.

[7] Abbi R, El-Darzi E, Vasilakis C, and Millard P, "Length of stay based grouping and classification methodology for modelling patient flow," *Journal of Operations and Logistics (Accepted),* 2008.

[8] Rissanen J, "Modelling by the shortest data description," *Automatica,* vol. 14, pp. 465-471, 1978.

[9] Bishop C.M, *Pattern Recognition and Machine Learning*. New York, USA: Springer, 2006.

[10] Dempster A.P, Laird N.M, and Rubin D.B, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 39, pp. 1-38, 1977.

[11] Gilland D.R, Tsui B.M.W, Metz C.E, Jaszczak R.J, and Perry J.R, "An Evaluation of Maximum Likelihood-Expectation Maximization Reconstruction for SPECT by ROC Analysis," *The Journal of Nuclear Medicine* vol. 33, pp. 451-457., 1992.

[12] Permuter H, Francos J, and Jermyn I.H, "A study of Gaussian mixture models of colour and texture features for image classification and segmentation," *Pattern Recognition,* vol. 36, pp. 695-706, 2006.

[13] Carson C and Greenspan H, "Blobworld: Image Segmentation using Expectation-Maximization and its Application to Image Querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, pp. 1026-1038, 2002.

[14] Roch M.A, Soldevilla M.S, and Burtenshaw J.C, "Gaussian mixture model classification of odontocetes in the Southern California Bright and the Gulf of California," *Journal of Acoustical Society of America,* vol. 121, 2007.

[15] Xing Y, Yu T, Wu Y.N, Roy M, Kim J, and Lee C, "An expectation-maximisation algorithm for probabilistic reconstuctions of full-length isoforms from splice graphs," *Nucleic Acids Research,* vol. 34, pp. 3150-3160, 2006.

[16] Zhang Y, Xu W, and Callan J, "Exact Maximum Likelihood Estimation for Word Mixtures," *Text Learning Workshop in International Conference on Machine Learning (ICML 2002)*, 2002.

[17] Chandramouli R and Srikantam V.K, "On Mixture Density and Maximum Likelihood Power estimation via Expectation-Maximization," *Proceedings of the 2000 conference on Asia South Pacific design automation*, Yokohama, Japan, 2000, pp. 423-428.

[18] MacQueen J. B, "Some Methods for classification and Analysis of Multivariate Observations," *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1967, pp. 281-297.

[19] Lin X and Zhu Y, "Degenerate Expectation-Maximisation Algorithm for local Dimension Reduction," in *Classification, Clustering, and Data Mining Applications*, Banks D, House L, McMorris F.R, Arabie P, and Gaul W, Eds. Chicago: Springer 2004.

[20] Millard P.H, Mackay M, Vasilakis C, and Christodoulou G, "Measuring and modelling surgical bed usage," *Annals of the Royal College of surgeons of England,* vol. 82, pp. 75-82, March 2000.

[21] Marshall A.H, Vasilakis C, and El-Darzi E, "Length of Stay-Based Patient Flow Models: Recent Developments and Future Directions," *Health Care Management Science,* vol. 8, pp. 213-320, August 2005.

[22] Abbi R, El-Darzi E, Vasilakis C, and Millard P, "Length of stay based grouping and classification methodology for modelling patient flow," *Journal of Operations and Logistics (Submitted),* 2007.

[23] Abbi R, El-Darzi E, Vasilakis C, and Millard P.H, "A Gaussian mixture model approach to grouping patients according to their hospital length of stay," *submitted to 21st IEEE International Symposium on Computer-based Medical Systems* Finland, 2008.