Continuous Observation Hidden Markov Model

Article *in* Revista Kasmera · June 2016

CITATION
1

READS
2,320

1 author:

Academic Network of Loc Nguyen
Loc Nguyen's Academic Network
189 PUBLICATIONS **189** CITATIONS

SEE PROFILE

ISSN: 0075-5222

Continuous Observation Hidden Markov Model

Loc Nguyen
Sunflower Soft Company, An Giang, Vietnam

Abstract

Hidden Markov model (HMM) is a powerful mathematical tool for prediction and recognition but it is not easy to understand deeply its essential disciplines. Previously, I made a full tutorial on HMM in order to support researchers to comprehend HMM. However HMM goes beyond what such tutorial mentioned when observation may be signified by continuous value such as real number and real vector instead of discrete value. Note that state of HMM is always discrete event but continuous observation extends capacity of HMM for solving complex problems. Therefore, I do this research focusing on HMM in case that its observation conforms to a single probabilistic distribution. Moreover, mixture HMM in which observation is characterized by the mixture model of partial probability density functions is also mentioned. Mathematical proofs and practical techniques relevant to continuous observation HMM are main subjects of the research.

Keywords: hidden Markov model, continuous observation, mixture model, evaluation problem, uncovering problem, learning problem

I. Hidden Markov model

The research produces a full tutorial on hidden Markov model (HMM) in case of continuous observations and so it is required to introduce essential concepts and problems of HMM. The main reference of this tutorial is the article "A tutorial on hidden Markov models and selected applications in speech recognition" by author (Rabiner, 1989). Section I – the first section is summary of the tutorial on HMM by author (Nguyen, 2016) whereas sections II and III are main ones of the research. Section IV is the discussion and conclusion. The main problem that needs to be solved is how to learn HMM parameters when discrete observation probability matrix is replaced by continuous density function. In section II, I propose practical technique to calculate essential quantities such as forward variable α_t, backward variable β_t, and joint probabilities ξ_t, γ_t which are necessary to train HMM with regard to continuous observations. Moreover, from expectation maximization (EM) algorithm which was used to learn traditional discrete HMM, I derive the general equation whose solutions are optimal parameters. Such equation specified by formulas II.5 and III.7 is described in sections II, III and discussed more in section IV. My reasoning is based on EM algorithm and Lagrangian function for solving optimization problem.

As a convention, all equations are called formulas and they are entitled so that it is easy for researchers to look up them. Tables, figures, and formulas are numbered according to their sections. For example, formula I.1.1 is the first

formula in sub-section I.1. Most common notations "*exp*" and "*ln*" denote exponential function and natural logarithm function.

There are many real-world phenomena (so-called states) that we would like to model in order to explain our observations. Often, given sequence of observations symbols, there is demand of discovering real states. For example, there are some states of weather: *sunny*, *cloudy*, *rainy* (Fosler-Lussier, 1998, p. 1). Suppose you are in the room and do not know the weather outside but you are notified observations such as wind speed, atmospheric pressure, humidity, and temperature from someone else. Basing on these observations, it is possible for you to forecast the weather by using HMM. Before discussing about HMM, we should glance over the definition of Markov model (MM). First, MM is the statistical model which is used to model the stochastic process. MM is defined as below (Schmolze, 2001):

- Given a finite set of state $S=\{s_1, s_2,\ldots, s_n\}$ whose cardinality is n. Let $\prod$ be the *initial state distribution* where $\pi_i \in \prod$ represents the probability that the stochastic process begins in state s_i. In other words π_i is the initial probability of state s_i, where $\sum_{s_i \in S} \pi_i = 1$.
- The stochastic process which is modeled gets only one state from S at all time points. This stochastic process is defined as a finite vector $X=(x_1, x_2,\ldots, x_T)$ whose element x_t is a state at time point t. The process X is called *state stochastic process* and $x_t \in S$ equals some state $s_i \in S$. Note that X is also called *state sequence*. Time point can be in terms of second, minute, hour, day, month, year, etc. It is easy to infer that the initial probability $\pi_i = P(x_1=s_i)$ where x_1 is the first state of the stochastic process. The state stochastic process X must meet fully the *Markov property*, namely, given previous state x_{t-1} of process X, the conditional probability of current state x_t is only dependent on the previous state x_{t-1}, not relevant to any further past state $(x_{t-2}, x_{t-3},\ldots, x_1)$. In other words, $P(x_t \mid x_{t-1}, x_{t-2}, x_{t-3},\ldots, x_1) = P(x_t \mid x_{t-1})$ with note that $P(.)$ also denotes probability in this research. Such process is called first-order Markov process.
- At each time point, the process changes to the next state based on the *transition probability distribution* a_{ij}, which depends only on the previous state. So a_{ij} is the probability that the stochastic process changes current state s_i to next state s_j. It means that $a_{ij} = P(x_t=s_j \mid x_{t-1}=s_i) = P(x_{t+1}=s_j \mid x_t=s_i)$. The probability of transitioning from any given state to some next state is 1, we have $\forall s_i \in S, \sum_{s_j \in S} a_{ij} = 1$. All transition probabilities a_{ij} (s) constitute the *transition probability matrix* A. Note that A is n by n matrix because there are n distinct states. It is easy to infer that matrix A represents state stochastic process X. It is possible to understand that the initial probability matrix $\prod$ is degradation case of matrix A.

Briefly, MM is the triple $\langle S, A, \prod \rangle$. In typical MM, states are observed directly by users and transition probabilities (A and $\prod$) are unique parameters. Otherwise, hidden Markov model (HMM) is similar to MM except that the underlying states become hidden from observer, they are hidden parameters. HMM adds more output parameters which are called observations. Each state (hidden parameter) has the conditional probability distribution upon such observations. HMM is

responsible for discovering hidden parameters (states) from output parameters (observations), given the stochastic process. The HMM has further properties as below (Schmolze, 2001):

- Suppose there is a finite set of possible observations $\Phi = \{\varphi_1, \varphi_2,\ldots, \varphi_m\}$ whose cardinality is m. There is the second stochastic process which produces *observations* correlating with hidden states. This process is called *observable stochastic process*, which is defined as a finite vector $O = (o_1, o_2,\ldots, o_T)$ whose element o_t is an observation at time point t. Note that $o_t \in \Phi$ equals some φ_k. The process O is often known as *observation sequence*.
- There is a probability distribution of producing a given observation in each state. Let $b_i(k)$ be the probability of observation φ_k when the state stochastic process is in state s_i. It means that $b_i(k) = b_i(o_t=\varphi_k) = P(o_t=\varphi_k \mid x_t=s_i)$. The sum of probabilities of all observations which observed in a certain state is 1, we have $\forall s_i \in S, \sum_{\theta_k \in \Phi} b_i(k) = 1$. All probabilities of observations $b_i(k)$ constitute the *observation probability matrix B*. It is convenient for us to use notation b_{ik} instead of notation $b_i(k)$. Note that B is n by m matrix because there are n distinct states and m distinct observations. While matrix A represents state stochastic process X, matrix B represents observable stochastic process O.

Thus, HMM is the 5-tuple $\Delta = \langle S, \Phi, A, B, \prod \rangle$. Note that components S, Φ, A, B, and $\prod$ are often called parameters of HMM in which A, B, and $\prod$ are essential parameters. Going back weather example, suppose you need to predict how weather tomorrow is: *sunny*, *cloudy* or *rainy* since you know only observations about the humidity: *dry*, *dryish*, *damp*, *soggy*. The HMM is totally determined based on its parameters S, Φ, A, B, and $\prod$ according to weather example. We have $S = \{s_1=sunny, s_2=cloudy, s_3=rainy\}$, $\Phi = \{\varphi_1=dry, \varphi_2=dryish, \varphi_3=damp, \varphi_4=soggy\}$. Transition probability matrix A is shown in table I.1.

		Weather current day (Time point t)		
		sunny	*cloudy*	*rainy*
Weather previous day (Time point t –1)	*sunny*	a_{11}=0.50	a_{12}=0.25	a_{13}=0.25
	cloudy	a_{21}=0.30	a_{22}=0.40	a_{23}=0.30
	rainy	a_{31}=0.25	a_{32}=0.25	a_{33}=0.50

Table I.1. Transition probability matrix A

From table I.1, we have $a_{11}+a_{12}+a_{13}=1$, $a_{21}+a_{22}+a_{23}=1$, $a_{31}+a_{32}+a_{33}=1$. Initial state distribution specified as uniform distribution is shown in table I.2.

sunny	*cloudy*	*rainy*
π_1=0.33	π_2=0.33	π_3=0.33

Table I.2. Uniform initial state distribution $\prod$

From table I.2, we have $\pi_1+\pi_2+\pi_3=1$.

ISSN: 0075-5222

Observation probability matrix B is shown in table I.3.

		Humidity			
		dry	*dryish*	*damp*	*soggy*
Weather	*sunny*	b_{11}=0.60	b_{12}=0.20	b_{13}=0.15	b_{14}=0.05
	cloudy	b_{21}=0.25	b_{22}=0.25	b_{23}=0.25	b_{24}=0.25
	rainy	b_{31}=0.05	b_{32}=0.10	b_{33}=0.35	b_{34}=0.50

Table I.3. Observation probability matrix B

From table I.3, we have $b_{11}+b_{12}+b_{13}+b_{14}=1$, $b_{21}+b_{22}+b_{23}+b_{24}=1$, $b_{31}+b_{32}+b_{33}+b_{34}=1$.

The whole weather HMM is depicted in figure I.1.

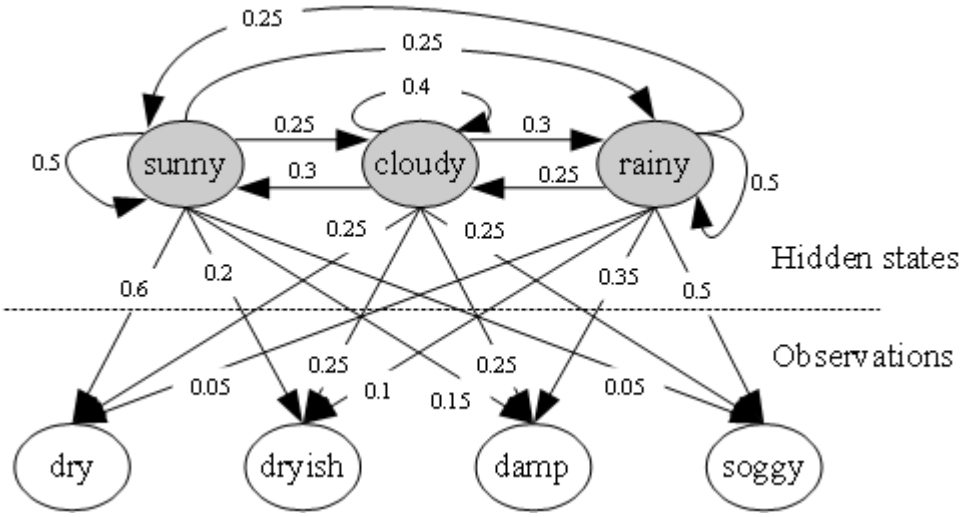

Figure I.1. HMM of weather forecast (hidden states are shaded)

There are three problems of HMM (Schmolze, 2001) (Rabiner, 1989, pp. 262-266):

1. Given HMM Δ and an observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ where $o_t \in \Phi$, how to calculate the probability $P(O|\Delta)$ of this observation sequence. Such probability $P(O|\Delta)$ indicates how much the HMM Δ affects on sequence O. This is *evaluation problem* or *explanation problem*. Note that it is possible to denote $O = \{o_1 \rightarrow o_2 \rightarrow\ldots\rightarrow o_T\}$ and the sequence O is aforementioned observable stochastic process.
2. Given HMM Δ and an observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ where $o_t \in \Phi$, how to find the sequence of states $X = \{x_1, x_2,\ldots, x_T\}$ where $x_t \in S$ so that X is most likely to have produced the observation sequence O. This is *uncovering problem*. Note that the sequence X is aforementioned state stochastic process.
3. Given HMM Δ and an observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ where $o_t \in \Phi$, how to adjust parameters of Δ such as initial state distribution $\prod$, transition probability matrix A, and observation probability matrix B so that the quality of HMM Δ is enhanced. This is *learning problem*.

ISSN: 0075-5222

These problems will be mentioned in sub-sections I.1, I.2, and I.3, in turn.

I.1. HMM evaluation problem

The essence of evaluation problem is to find out the way to compute the probability $P(O|\Delta)$ most effectively given the observation sequence $O = \{o_1, o_2,\ldots, o_T\}$. For example, given HMM Δ whose parameters A, B, and $\prod$ specified in tables I.1, I.2, and I.3, which is designed for weather forecast. Suppose we need to calculate the probability of event that humidity is *soggy* and *dry* in days 1 and 2, respectively. This is evaluation problem with sequence of observations $O = \{o_1=\varphi_4=soggy,\ o_2=\varphi_1=dry,\ o_3=\varphi_2=dryish\}$. There is a complete set of $3^3=27$ mutually exclusive cases of weather states for three days; for example, given a case in which weather states in days 1, 2, and 3 are *sunny*, *sunny*, and *sunny* then, state stochastic process is $X = \{x_1=s_1=sunny,\ x_2=s_1=sunny,\ x_3=s_1=sunny\}$. It is easy to recognize that it is impossible to browse all combinational cases of given observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ as we knew that it is necessary to survey $3^3=27$ mutually exclusive cases of weather states with a tiny number of observations {*soggy*, *dry*, *dryish*}. Exactly, given n states and T observations, it takes extremely expensive cost to survey n^T cases. According to (Rabiner, 1989, pp. 262-263), there is a so-called *forward-backward procedure* to decrease computational cost for determining the probability $P(O|\Delta)$. Let $\alpha_t(i)$ be the joint probability of partial observation sequence $\{o_1, o_2,\ldots, o_t\}$ and state $x_t=s_i$ where $1 \leq t \leq T$, specified by formula I.1.1.

$$\alpha_t(i) = P(o_1, o_2, \ldots, o_t, x_t = s_i|\Delta)$$

Formula I.1.1. Forward variable

The joint probability $\alpha_t(i)$ is also called *forward variable* at time point t and state s_i. Formula I.1.2 specifies recurrence property of forward variable (Rabiner, 1989, p. 262).

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right) b_j(o_{t+1})$$

Formula I.1.2. Recurrence property of forward variable

Where $b_j(o_{t+1})$ is the probability of observation o_{t+1} when the state stochastic process is in state s_j, please see an example of observation probability matrix shown in table I.3. Please pay attention to recurrence property of forward variable specified by formula I.1.2 because this formula is essentially to build up Markov chain.

According to the forward recurrence formula I.1.2, given observation sequence $O = \{o_1, o_2,\ldots, o_T\}$, we have:

$$\alpha_T(i) = P(o_1, o_2, \ldots, o_T, x_T = s_i|\Delta)$$

The probability $P(O|\Delta)$ is sum of $\alpha_T(i)$ over all n possible states of x_T, specified by formula I.1.3.

ISSN: 0075-5222

$$P(O|\Delta) = P(o_1, o_2, \dots, o_T) = \sum_{i=1}^{n} P(o_1, o_2, \dots, o_T, x_T = s_i|\Delta) = \sum_{i=1}^{n} \alpha_T(i)$$

Formula I.1.3. Probability $P(O|\Delta)$ based on forward variable

The forward-backward procedure to calculate the probability $P(O|\Delta)$, based on forward formulas I.1.2 and I.1.3, includes three steps as shown in table I.1.1 (Rabiner, 1989, p. 262).

1. Initialization step: Initializing $\alpha_1(i) = b_i(o_1)\pi_i$ for all $1 \leq i \leq n$
2. Recurrence step: Calculating all $\alpha_{t+1}(j)$ for all $1 \leq j \leq n$ and $1 \leq t \leq T - 1$ according to formula I.1.2.
$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right) b_j(o_{t+1})$$
3. Evaluation step: Calculating the probability $P(O|\Delta) = \sum_{i=1}^{n} \alpha_T(i)$

Table I.1.1. Forward-backward procedure based on forward variable to calculate the probability $P(O|\Delta)$

There is interesting thing that the forward-backward procedure can be implemented based on so-called *backward variable*. Let $\beta_t(i)$ be the backward variable which is conditional probability of partial observation sequence $\{o_t, o_{t+1}, \dots, o_T\}$ given state $x_t=s_i$ where $1 \leq t \leq T$, specified by formula I.1.4.

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, \dots, o_T|x_t = s_i, \Delta)$$

Formula I.1.4. Backward variable

The recurrence property of backward variable specified by formula I.1.5 (Rabiner, 1989, p. 263).

$$\beta_t(i) = \sum_{j=1}^{n} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$

Formula I.1.5. Recurrence property of backward variable

Where $b_j(o_{t+1})$ is the probability of observation o_{t+1} when the state stochastic process is in state s_j, please see an example of observation probability matrix shown in table I.3. The construction of backward recurrence formula I.1.5 is essentially to build up Markov chain.

The probability $P(O|\Delta)$ is sum of product $\pi_i b_i(o_1)\beta_1(i)$ over all n possible states of $x_1=s_i$, specified by formula I.1.6.

$$P(O|\Delta) = \sum_{i=1}^{n} \pi_i b_i(o_1) \beta_1(i)$$

Formula I.1.6. Probability $P(O|\Delta)$ based on backward variable

ISSN: 0075-5222

The forward-backward procedure to calculate the probability $P(O|\Delta)$, based on backward formulas I.1.5 and I.1.6, includes three steps as shown in table I.1.2 (Rabiner, 1989, p. 263).

1. Initialization step: Initializing $\beta_T(i) = 1$ for all $1 \leq i \leq n$
2. Recurrence step: Calculating all $\beta_t(i)$ for all $1 \leq i \leq n$ and $t=T-1$, $t=T-2,\ldots$, $t=1$, according to formula I.1.5.

$$\beta_t(i) = \sum_{j=1}^{n} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$

3. Evaluation step: Calculating the probability $P(O|\Delta)$ according to formula I.1.6, $P(O|\Delta) = \sum_{i=1}^{n} \pi_i b_i(o_1) \beta_1(i)$

Table I.1.2. Forward-backward procedure based on backward variable to calculate the probability $P(O|\Delta)$

Now the uncovering problem is mentioned particularly in successive sub-section I.2.

I.2. HMM uncovering problem

Recall that given HMM Δ and observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ where $o_t \in \Phi$, how to find out a state sequence $X = \{x_1, x_2,\ldots, x_T\}$ where $x_t \in S$ so that X is most likely to have produced the observation sequence O. This is the uncovering problem: which sequence of state transitions is most likely to have led to given observation sequence. In other words, it is required to establish an *optimal criterion* so that the state sequence X leads to maximizing such criterion. The simple criterion is the conditional probability of sequence X with respect to sequence O and model Δ, denoted $P(X|O,\Delta)$. We can apply brute-force strategy: "go through all possible such X and pick the one leading to maximizing the criterion $P(X|O,\Delta)$".

$$X = \underset{X}{\operatorname{argmax}}\big(P(X|O,\Delta)\big)$$

This strategy is impossible if the number of states and observations is huge. Another popular way is to establish a so-called *individually optimal criterion* (Rabiner, 1989, p. 263) which is described right later.

Let $\gamma_t(i)$ be joint probability that the stochastic process is in state s_i at time point t with observation sequence $O = \{o_1, o_2,\ldots, o_T\}$, formula I.2.1 specifies this probability based on forward variable α_t and backward variable β_t.

$$\gamma_t(i) = P(o_1, o_2, \ldots, o_T, x_t = s_i|\Delta) = \alpha_t(i)\beta_t(i)$$

Formula I.2.1. Joint probability of being in state s_i at time point t with observation sequence O

The variable $\gamma_t(i)$ is also called *individually optimal criterion* with note that forward variable α_t and backward variable β_t are calculated according to recurrence formulas I.1.2 and I.1.5, respectively.

ISSN: 0075-5222

Because the probability $P(o_1, o_2, \dots, o_T|\Delta)$ is not relevant to state sequence X, it is possible to remove it from the optimization criterion. Thus, formula I.2.2 specifies how to find out the optimal state x_t of X at time point t.

$$x_t = \underset{i}{\operatorname{argmax}}\, \gamma_t(i) = \underset{i}{\operatorname{argmax}}\, \alpha_t(i)\beta_t(i)$$

Formula I.2.2. Optimal state at time point t

Note that index i is identified with state $s_i \in S$ according to formula I.2.2. The optimal state x_t of X at time point t is the one that maximizes product $\alpha_t(i)\ \beta_t(i)$ over all values s_i. The procedure to find out state sequence $X = \{x_1, x_2,\dots, x_T\}$ based on individually optimal criterion is called *individually optimal procedure* that includes three steps, shown in table I.2.1.

1. Initialization step:
   - Initializing $\alpha_1(i) = b_i(o_1)\pi_i$ for all $1 \le i \le n$
   - Initializing $\beta_T(i) = 1$ for all $1 \le i \le n$
2. Recurrence step:
   - Calculating all $\alpha_{t+1}(i)$ for all $1 \le i \le n$ and $1 \le t \le T-1$ according to formula I.1.2.
   - Calculating all $\beta_t(i)$ for all $1 \le i \le n$ and $t=T-1, t=T-2,\dots, t=1$, according to formula I.1.5.
   - Calculating all $\gamma_t(i)=\alpha_t(i)\beta_t(i)$ for all $1 \le i \le n$ and $1 \le t \le T$ according to formula I.2.1.
   - Determining optimal state x_t of X at time point t is the one that maximizes $\gamma_t(i)$ over all values s_i.
     $$x_t = \underset{i}{\operatorname{argmax}}\, \gamma_t(i)$$
3. Final step: The state sequence $X = \{x_1, x_2,\dots, x_T\}$ is totally determined when its partial states x_t (s) where $1 \le t \le T$ are found in recurrence step.

Table I.2.1. Individually optimal procedure to solve uncovering problem

The individually optimal criterion $\gamma_t(i)$ does not reflect the whole probability of state sequence X given observation sequence O because it focuses only on how to find out each partially optimal state x_t at each time point t. Thus, the individually optimal procedure is heuristic method. Viterbi algorithm (Rabiner, 1989, p. 264) is alternative method that takes interest in the whole state sequence X by using joint probability $P(X,O|\Delta)$ of state sequence and observation sequence as optimal criterion for determining state sequence X. Let $\delta_t(i)$ be the maximum joint probability of observation sequence O and state $x_t=s_i$ over $t-1$ previous states. The quantity $\delta_t(i)$ is called *joint optimal criterion* at time point t, which is specified by formula I.2.3.

$$\delta_t(i) = \max_{x_1,x_2,\dots,x_{t-1}} \left(P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_t = s_i|\Delta)\right)$$

Formula I.2.3. Joint optimal criterion at time point t

ISSN: 0075-5222

The recurrence property of *joint optimal criterion* is specified by formula I.2.4 (Rabiner, 1989, p. 264).

$$\delta_{t+1}(j) = \left(\max_i\left(\delta_t(i)a_{ij}\right)\right) b_j(o_{t+1})$$

Formula I.2.4. Recurrence property of joint optimal criterion

The semantic content of joint optimal criterion δ_t is similar to the forward variable α_t. Given criterion $\delta_{t+1}(j)$, the state $x_{t+1}=s_j$ that maximizes $\delta_{t+1}(j)$ is stored in the backtracking state $q_{t+1}(j)$ that is specified by formula I.2.5.

$$q_{t+1}(j) = \operatorname*{argmax}_i\left(\delta_t(i)a_{ij}\right)$$

Formula I.2.5. Backtracking state

Note that index i is identified with state $s_i \in S$ according to formula I.2.5. The Viterbi algorithm based on joint optimal criterion $\delta_t(i)$ includes three steps described in table I.2.2 (Rabiner, 1989, p. 264).

1. Initialization step:
   - Initializing $\delta_1(i) = b_i(o_1)\pi_i$ for all $1 \le i \le n$
   - Initializing $q_1(i) = 0$ for all $1 \le i \le n$
2. Recurrence step:
   - Calculating all $\delta_{t+1}(j) = \left(\max_i\left(\delta_t(i)a_{ij}\right)\right) b_j(o_{t+1})$ for all $1 \le i, j \le n$ and $1 \le t \le T-1$ according to formula I.2.4.
   - Keeping tracking optimal states $q_{t+1}(j) = \operatorname*{argmax}_i\left(\delta_t(i)a_{ij}\right)$ for all $1 \le j \le n$ and $1 \le t \le T-1$ according to formula I.2.5.
3. State sequence backtracking step: The resulted state sequence $X = \{x_1, x_2,\ldots, x_T\}$ is determined as follows:
   - The last state $x_T = \operatorname*{argmax}_j\left(\delta_T(j)\right)$
   - Previous states are determined by backtracking: $x_t = q_{t+1}(x_{t+1})$ for $t=T-1, t=T-2,\ldots, t=1$.

Table I.2.2. Viterbi algorithm to solve uncovering problem

Now the uncovering problem is described thoroughly in this sub-section I.2. Successive sub-section I.3 will mention the last problem of HMM that is the learning problem.

I.3. HMM learning problem

The learning problem is to adjust parameters such as initial state distribution $\prod$, transition probability matrix A, and observation probability matrix B so that given HMM Δ gets more appropriate to an observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ with note that Δ is represented by these parameters. In other words, the learning

ISSN: 0075-5222

problem is to adjust parameters by maximizing probability of observation sequence *O*, as follows:

$$(A, B, \Pi) = \underset{A,B,\Pi}{\operatorname{argmax}} P(O|\Delta)$$

The Expectation Maximization (EM) algorithm is applied successfully into solving HMM learning problem, which is equivalently well-known Baum-Welch algorithm by authors Leonard E. Baum and Lloyd R. Welch (Rabiner, 1989). The successive sub-section I.3.1 describes shortly EM algorithm before going into Baum-Welch algorithm.

I.3.1. EM algorithm

Expectation Maximization (EM) is effective parameter estimator in case that incomplete data is composed of two parts: observed part and hidden part (missing part). EM is iterative algorithm that improves parameters after iterations until reaching optimal parameters. Each iteration includes two steps: E(xpectation) step and M(aximization) step. In E-step the hidden data is estimated based on observed data and current estimate of parameters; so the lower-bound of likelihood function is computed by the expectation of complete data. In M-step new estimates of parameters are determined by maximizing the lower-bound. Please see document (Sean, 2009) for short tutorial of EM. This sub-section I.3.1 focuses on practice general EM algorithm; the theory of EM algorithm is described comprehensively in article "Maximum Likelihood from Incomplete Data via the EM algorithm" by authors (Dempster, Laird, & Rubin, 1977).

Suppose *O* and *X* are observed data and hidden data, respectively. Note *O* and *X* can be represented in any form such as discrete values, scalar, integer number, real number, vector, list, sequence, sample, and matrix. Let Θ represent parameters of probability distribution. Concretely, Θ includes initial state distribution $\prod$, transition probability matrix *A*, and observation probability matrix *B* inside HMM. In other words, Θ represents HMM Δ itself. EM algorithm aims to estimate Θ by finding out which $\widehat{\Theta}$ maximizes the likelihood function $L(\Theta) = P(O|\Theta)$.

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, L(\Theta) = \underset{\Theta}{\operatorname{argmax}}\, P(O|\Theta) = \underset{\Theta}{\operatorname{argmax}} \sum_{X} P(X|O,\Theta_t) ln\big(P(O,X|\Theta)\big)$$

Where $\widehat{\Theta}$ is the optimal estimate of parameters which is called usually *parameter estimate*. Note that notation "*ln*" denotes natural logarithm function.

The expression $\sum_X P(X|O,\Theta_t) ln\big(P(O,X|\Theta)\big)$ is essentially expectation of $ln\big(P(O,X|\Theta)\big)$ given conditional probability distribution $P(X|O,\Theta_t)$ when $P(X|O,\Theta_t)$ is totally determined. Let $E_{X|O,\Theta_t}\{ln\big(P(O,X|\Theta)\big)\}$ denote this conditional expectation, formula I.3.1.1 specifies EM optimization criterion for determining the parameter estimate, which is the most important aspect of EM algorithm (Sean, 2009, p. 8).

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, E_{X|O,\Theta_t}\{ln\big(P(O,X|\Theta)\big)\}$$

Where,

ISSN: 0075-5222

$$E_{X|O,\Theta_t}\{ln(P(O,X|\Theta))\} = \sum_X P(X|O,\Theta_t) ln(P(O,X|\Theta))$$

Formula I.3.1.1. EM optimization criterion based on conditional expectation

If $P(X|O,\Theta_t)$ is continuous density function, the continuous version of this conditional expectation is:

$$E_{X|O,\Theta_t}\{ln(P(O,X|\Theta))\} = \int_X P(X|O,\Theta_t) ln(P(O,X|\Theta))$$

Finally, the EM algorithm is described in table I.3.1.1.

Starting with initial parameter Θ_0, each iteration in EM algorithm has two steps:

1. *E-step*: computing the conditional expectation $E_{X|O,\Theta_t}\{ln(P(O,X|\Theta))\}$ based on the current parameter Θ_t according to formula I.3.1.1.
2. *M-step*: finding out the estimate $\widehat{\Theta}$ that maximizes such conditional expectation. The next parameter Θ_{t+1} is assigned by the estimate $\widehat{\Theta}$, we have:

   $$\Theta_{t+1} = \widehat{\Theta}$$

   Of course Θ_{t+1} becomes current parameter for next iteration. How to maximize the conditional expectation is optimization problem which is dependent on applications. For example, the popular method to solve optimization problem is Lagrangian duality (Jia, 2013, p. 8).

EM algorithm stops when it meets the terminating condition, for example, the difference of current parameter Θ_t and next parameter Θ_{t+1} is smaller than some pre-defined threshold ε.

$$|\Theta_{t+1} - \Theta_t| < \varepsilon$$

In addition, it is possible to define a custom terminating condition.

Table I.3.1.1. General EM algorithm

In general, it is easy to calculate the EM expectation $E_{X|O,\Theta_t}\{ln(P(O,X|\Theta))\}$ but finding out the estimate $\widehat{\Theta}$ based on maximizing such expectation is complicated optimization problem. It is possible to state that the essence of EM algorithm is to determine the estimate $\widehat{\Theta}$. Now the EM algorithm is introduced to you. How to apply it into solving HMM learning problem is described in successive sub-section I.3.2.

I.3.2. Applying EM algorithm into solving learning problem

Now going back the HMM learning problem, the EM algorithm is applied into solving this problem, which is equivalently well-known Baum-Welch algorithm by authors Leonard E. Baum and Lloyd R. Welch (Rabiner, 1989). The parameter Θ becomes the HMM model $\Delta = (A, B, \prod)$. Recall that the learning problem is to adjust parameters by maximizing probability of observation sequence O, as follows:

$$\widehat{\Delta} = (\hat{A}, \hat{B}, \hat{\Pi}) = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j) = \underset{\Delta}{\operatorname{argmax}} P(O|\Delta)$$

ISSN: 0075-5222

Where $\hat{a}_{ij}$, $\hat{b}_j(k)$, $\hat{\pi}_j$ are parameter estimates and so, the purpose of HMM learning problem is to determine them.

The observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ and state sequence $X = \{x_1, x_2,\ldots, x_T\}$ are observed data and hidden data within context of EM algorithm, respectively. Note O and X is now represented in sequence. According to EM algorithm, the parameter estimate $\hat{\Delta}$ is determined as follows:

$$\hat{\Delta} = \left(\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j\right) = \underset{\Delta}{\operatorname{argmax}}\, E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}$$

Where $\Delta_r = (A_r, B_r, \prod_r)$ is the known parameter at the current iteration. Note that we use notation Δ_r instead of popular notation Δ_t in order to distinguish iteration indices of EM algorithm from time points inside observation sequence O and state sequence X.

It is conventional that $P(x_1|x_0,\Delta) = P(x_1|\Delta)$ where x_0 is pseudo-state, formula I.3.2.1 specifies general EM conditional expectation for HMM:

$$E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\} = \sum_X P(X|O,\Delta_r) ln\left(\prod_{t=1}^{T} P(x_t|x_{t-1},\Delta)P(o_t|x_t,\Delta)\right)$$
$$= \sum_X P(X|O,\Delta_r) \sum_{t=1}^{T} \left(ln(P(x_t|x_{t-1},\Delta)) + ln(P(o_t|x_t,\Delta))\right)$$

Formula I.3.2.1. General EM conditional expectation for HMM

Note that notation "*ln*" denotes natural logarithm function.

Because of the convention $P(x_1|x_0,\Delta) = P(x_1|\Delta)$, matrix $\prod$ is degradation case of matrix A at time point t=1. In other words, the initial probability π_j is equal to the transition probability a_{ij} from pseudo-state x_0 to state x_1=s_j.

$$P(x_1 = s_j|x_0,\Delta) = P(x_1 = s_j|\Delta) = \pi_j$$

Note that n=$|S|$ is the number of possible states and m=$|\Phi|$ is the number of possible observations. Let $I(x_{t-1} = s_i, x_t = s_j)$ and $I(x_t = s_j, o_t = \varphi_k)$ are two index functions so that

$$I(s_i = x_{t-1}, s_j = x_t) = \begin{cases} 1 \text{ if } s_i = x_{t-1} \text{ and } s_j = x_t \\ 0 \text{ otherwise} \end{cases}$$
$$I(x_t = s_j, o_t = \varphi_k) = \begin{cases} 1 \text{ if } x_t = s_j \text{ and } o_t = \varphi_k \\ 0 \text{ otherwise} \end{cases}$$

The EM conditional expectation for HMM is specified by formula I.3.2.2.

$$E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}$$
$$= \sum_X P(X|O,\Delta_r)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} I(x_{t-1} = s_i, x_t = s_j) ln(a_{ij})\right.$$
$$\left. + \sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k) ln\left(b_j(k)\right)\right)$$

Formula I.3.2.2. EM conditional expectation for HMM

ISSN: 0075-5222

Where,

$$I(x_{t-1} = s_i, x_t = s_j) = \begin{cases} 1 \text{ if } x_{t-1} = s_i \text{ and } x_t = s_j \\ 0 \text{ otherwise} \end{cases}$$

$$I(x_t = s_j, o_t = \varphi_k) = \begin{cases} 1 \text{ if } x_t = s_j \text{ and } o_t = \varphi_k \\ 0 \text{ otherwise} \end{cases}$$

$$P(x_1 = s_j | x_0, \Delta) = P(x_1 = s_j | \Delta) = \pi_j$$

Note that the conditional expectation $E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}$ is function of Δ. There are two constraints for HMM as follows:

$$\sum_{j=1}^{n} a_{ij} = 1, \forall i = \overline{1,n}$$

$$\sum_{k=1}^{m} b_j(k) = 1, \forall k = \overline{1,m}$$

Maximizing $E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}$ with subject to these constraints is optimization problem that is solved by Lagrangian duality theorem (Jia, 2013, p. 8). Original optimization problem mentions minimizing target function but it is easy to infer that maximizing target function shares the same methodology. Let $l(\Delta, \lambda, \mu)$ be Lagrangian function constructed from $E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}$ together with these constraints (Ramage, 2007, p. 9), we have formula I.3.2.3 for specifying HMM Lagrangian function as follows:

$$\begin{aligned} l(\Delta, \lambda, \mu) &= l(a_{ij}, b_j(k), \lambda_i, \mu_j) \\ &= E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\} + \sum_{i=1}^{n} \lambda_i \left(1 - \sum_{j=1}^{n} a_{ij}\right) \\ &+ \sum_{j=1}^{n} \mu_j \left(1 - \sum_{k=1}^{m} b_j(k)\right) \end{aligned}$$

Formula I.3.2.3. Lagrangian function for HMM

Where λ is n-component vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ and μ is m-component vector $\mu = (\mu_1, \mu_2, \ldots, \mu_m)$. Factors $\lambda_i \geq 0$ and $\mu_j \geq 0$ are called Lagrange multipliers or Karush-Kuhn-Tucker multipliers (Wikipedia, Karush–Kuhn–Tucker conditions, 2014) or dual variables. The expectation $E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}$ is specified by formula I.3.2.2.

The parameter estimate $\hat{\Delta}$ is extreme point of the Lagrangian function. According to Lagrangian duality theorem (Boyd & Vandenberghe, 2009, p. 216) (Jia, 2013, p. 8), we have:

$$\hat{\Delta} = (\hat{A}, \hat{B}) = \left(\hat{a}_{ij}, \hat{b}_j(k)\right) = \underset{A,B}{\operatorname{argmax}}\, l(\Delta, \lambda, \mu)$$

$$(\hat{\lambda}, \hat{\mu}) = \underset{\lambda,\mu}{\operatorname{argmin}}\, l(\Delta, \lambda, \mu)$$

The parameter estimate $\hat{\Delta} = \left(\hat{a}_{ij}, \hat{b}_j(k)\right)$ is determined by setting partial derivatives of $l(\Delta, \lambda, \mu)$ with respect to a_{ij} and $b_j(k)$ to be zero.

ISSN: 0075-5222

$$\frac{\partial l(\Delta,\lambda,\mu)}{\partial a_{ij}} = 0$$
$$\frac{\partial l(\Delta,\lambda,\mu)}{\partial b_j(k)} = 0$$

By solving these equations, we have formula I.3.2.4 for specifying HMM parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ given current parameter $\Delta = (a_{ij}, b_j(k), \pi_j)$ as follows:

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} P(O, x_{t-1} = s_i, x_t = s_j|\Delta)}{\sum_{t=2}^{T} P(O, x_{t-1} = s_i|\Delta)}$$
$$\hat{b}_j(k) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_k}}^{T} P(O, x_t = s_j|\Delta)}{\sum_{t=1}^{T} P(O, x_t = s_j|\Delta)}$$
$$\hat{\pi}_j = \frac{P(O, x_1 = s_j|\Delta)}{\sum_{i=1}^{n} P(O, x_1 = s_i|\Delta)}$$

Formula I.3.2.4. HMM parameter estimate

The parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ is the ultimate solution of the learning problem. As seen in formula I.3.2.4, it is necessary to calculate probabilities $P(O, x_{t-1}=s_i, x_t=s_j)$ and $P(O, x_{t-1}=s_i)$ when other probabilities $P(O, x_t=s_j)$, $P(O, x_1=s_i)$, and $P(O, x_1=s_j)$ are represented by the joint probability γ_t specified by formula I.2.1.

$$P(O, x_t = s_j|\Delta) = \gamma_t(j) = \alpha_t(j)\beta_t(j)$$
$$P(O, x_1 = s_i|\Delta) = \gamma_1(i) = \alpha_1(i)\beta_1(i)$$
$$P(O, x_1 = s_j|\Delta) = \gamma_1(j) = \alpha_1(j)\beta_1(j)$$

Let $\xi_t(i, j)$ is the joint probability that the stochastic process receives state s_i at time point t–1 and state s_j at time point t given observation sequence O (Rabiner, 1989, p. 264).

$$\xi_t(i,j) = P(O, x_{t-1} = s_i, x_t = s_j|\Delta)$$

Formula I.3.2.5 determines the joint probability $\xi_t(i, j)$ based on forward variable α_t and backward variable β_t.

$$\xi_t(i,j) = \alpha_{t-1}(i)a_{ij}b_j(o_t)\beta_t(j) \text{ where } t \geq 2$$

Formula I.3.2.5. Joint probability $\xi_t(i, j)$

Where forward variable α_t and backward variable β_t are calculated by previous recurrence formulas I.1.2 and I.1.5.

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right) b_j(o_{t+1})$$
$$\beta_t(i) = \sum_{j=1}^{n} a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$$

ISSN: 0075-5222

Recall that $\gamma_t(j)$ is the joint probability that the stochastic process is in state s_j at time point t with observation sequence $O = \{o_1, o_2,\ldots, o_T\}$, specified by previous formula I.2.1.

$$\gamma_t(j) = P(O, x_t = s_j|\Delta) = \alpha_t(j)\beta_t(j)$$

According to total probability rule, it is easy to infer that γ_t is sum of ξ_t over all states with $t \geq 2$, as seen in following formula I.3.2.6.

$$\forall t \geq 2, \gamma_t(j) = \sum_{i=1}^{n} \xi_t(i,j) \text{ and } \gamma_{t-1}(i) = \sum_{j=1}^{n} \xi_t(i,j)$$

Formula I.3.2.6. The γ_t is sum of ξ_t over all states

Deriving from formulas I.3.2.5 and I.3.2.6, we have:

$$P(O, x_{t-1} = s_i, x_t = s_j|\Delta) = \xi_t(i,j)$$

$$P(O, x_{t-1} = s_i|\Delta) = \sum_{j=1}^{n} \xi_t(i,j), \forall t \geq 2$$

$$P(O, x_t = s_j|\Delta) = \gamma_t(j)$$

$$P(O, x_1 = s_j|\Delta) = \gamma_1(j)$$

By extending formula I.3.2.4, we receive formula I.3.2.7 for specifying HMM parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_i(k), \hat{\pi}_i)$ given current parameter $\Delta = (a_{ij}, b_i(k), \pi_i)$ in detailed.

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=2}^{T} \sum_{l=1}^{n} \xi_t(i,l)}$$

$$\hat{b}_j(k) = \frac{\sum_{\substack{t=1 \\ o_t = \varphi_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

$$\hat{\pi}_j = \frac{\gamma_1(j)}{\sum_{i=1}^{n} \gamma_1(i)}$$

Formula I.3.2.7. HMM parameter estimate in detailed

The formula I.3.2.7 and its proof are found in (Ramage, 2007, pp. 9-12). It is easy to infer that the parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ is based on joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ which, in turn, are based on current parameter $\Delta = (a_{ij}, b_j(k), \pi_j)$. The EM conditional expectation $E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}$ is determined by joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$; so, the main task of E-step in EM algorithm is essentially to calculate the joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ according to formulas I.3.2.5 and I.2.1. The EM conditional expectation $E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}$ gets maximal at estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ and so, the main task of M-step in EM algorithm is essentially to calculate $\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j$ according to formula I.3.2.7. The EM algorithm is interpreted in HMM learning problem, as shown in table I.3.2.1.

Starting with initial value for Δ, each iteration in EM algorithm has two steps:

ISSN: 0075-5222

1. *E-step*: Calculating the joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ according to formulas I.3.2.5 and I.2.1 given current parameter $\Delta = (a_{ij}, b_j(k), \pi_j)$.

$$\xi_t(i,j) = \alpha_{t-1}(i)a_{ij}b_j(o_t)\beta_t(j) \text{ where } t \geq 2$$

$$\gamma_t(j) = P(O, x_t = s_j|\Delta) = \alpha_t(j)\beta_t(j)$$

Where forward variable α_t and backward variable β_t are calculated by previous recurrence formulas I.1.2 and I.1.5.

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right) b_j(o_{t+1})$$

$$\beta_t(i) = \sum_{j=1}^{n} a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$$

2. *M-step*: Calculating the estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ based on the joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ determined at E-step, according to formula I.3.2.7.

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=2}^{T}\sum_{l=1}^{n} \xi_t(i,l)}$$

$$\hat{b}_j(k) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

$$\hat{\pi}_j = \frac{\gamma_1(j)}{\sum_{i=1}^{n} \gamma_1(i)}$$

The estimate $\hat{\Delta}$ becomes the current parameter for next iteration.

EM algorithm stops when it meets the terminating condition, for example, the difference of current parameter Δ and next parameter $\hat{\Delta}$ is insignificant. It is possible to define a custom terminating condition.

Table I.3.2.1. EM algorithm for HMM learning problem

The algorithm to solve HMM learning problem shown in table I.3.2.1 is known as Baum-Welch algorithm by authors Leonard E. Baum and Lloyd R. Welch (Rabiner, 1989). Please see document "Hidden Markov Models Fundamentals" by (Ramage, 2007, pp. 8-13) for more details about HMM learning problem. As aforementioned in previous sub-section I.3.1, the essence of EM algorithm applied into HMM learning problem is to determine the estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$.

As seen in table I.3.2.1, it is not difficult to run E-step and M-step of EM algorithm but how to determine the terminating condition is considerable problem. It is better to establish a computational terminating criterion instead of applying the general statement "EM algorithm stops when it meets the terminating condition, for example, the difference of current parameter Δ and next parameter $\hat{\Delta}$ is insignificant". Therefore, author (Nguyen L. , Tutorial on Hidden Markov Model, 2016) proposes the probability $P(O|\Delta)$ as the terminating criterion. Calculating criterion $P(O|\Delta)$ is evaluation problem described in sub-section I.1. Criterion $P(O|\Delta)$ is determined according to forward-backward procedure; please see tables I.1.1 and I.1.2 for more details about forward-backward procedure.

ISSN: 0075-5222

1. Initialization step: Initializing $\alpha_1(i) = b_i(o_1)\pi_i$ for all $1 \leq i \leq n$
2. Recurrence step: Calculating all $\alpha_{t+1}(j)$ for all $1 \leq j \leq n$ and $1 \leq t \leq T-1$ according to formula I.1.2.

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right) b_j(o_{t+1})$$

3. Evaluation step: Calculating the probability $P(O|\Delta) = \sum_{i=1}^{n} \alpha_T(i)$

Concretely, when EM algorithm results out forward variables in E-step, the forward-backward procedure takes advantages of such forward variables so as to determine criterion $P(O|\Delta)$ the at the same time. As a result, the speed of EM algorithm does not decrease. However, there is always a redundant iteration; suppose that the terminating criterion approaches to maximal value at the end of the r^{th} iteration but the EM algorithm only stops at the E-step of the $(r+1)^{th}$ iteration when it really evaluates the terminating criterion. In general, the terminating criterion $P(O|\Delta)$ is calculated based on the current parameter Δ at E-step instead of the estimate $\hat{\Delta}$ at M-step. Table I.3.2.2 (Nguyen, Tutorial on Hidden Markov Model, 2016) shows the proposed implementation of EM algorithm with terminating criterion $P(O|\Delta)$. Pseudo-code like programming language C is used to describe the implementation of EM algorithm. Note, variables are marked as *italic words*, programming language keywords (*while*, *for*, *if*, [], ==, !=, &&, //, etc.) are marked blue and comments are marked gray. For example, notation [] denotes array index operation; concretely, $\alpha[t][i]$ denotes forward variable $\alpha_t(i)$ at time point t with regard to state s_i.

Input:
HMM with current parameter $\Delta = \{a_{ij}, \pi_j, b_{jk}\}$
Observation sequence $O = \{o_1, o_2, \ldots, o_T\}$
Output:
HMM with optimized parameter $\Delta = \{a_{ij}, \pi_j, b_{jk}\}$

Allocating memory for two matrices α and β representing forward variables and backward variables.
previous_criterion = –1
current_criterion = –1
iteration = 0
//Pre-defined number *MAX_ITERATION* is used to prevent from infinite loop.
MAX_ITERATION = 10000
While (*iteration* < *MAX_ITERATION*)

//Calculating forward variables and backward variables
For t = 1 to T
For i = 1 to n
Calculating forward variables $\alpha[t][i]$ and backward variables $\beta[T-t+1][i]$ based on observation sequence O according to formulas I.1.2 and I.1.5.
End for i

ISSN: 0075-5222

```
End for t

//Calculating terminating criterion current_criterion = P(O|Δ)
current_criterion = 0
For i = 1 to n
    current_criterion = current_criterion + α[T][i]
End for i

//Terminating condition
If previous_criterion >= 0 && previous_criterion == current_criterion then
    break //breaking out the loop, the algorithm stops
Else
    previous_criterion = current_criterion
End if

//Updating transition probability matrix
For i = 1 to n
    denominator = 0
    Allocating numerators as a 1-dimension array including n zero elements.
    For t = 2 to T
        For k = 1 to n
            ξ = α[t–1][i] * a_ik * b_k(o_t) * β[t][k]
            numerators[k] = numerators[k] + ξ
            denominator = denominator + ξ
        End for k
    End for t

    If denominator != 0 then
        For j = 1 to n
            a_ij = numerators[j] / denominator
        End for j
    End if

End for i

//Updating initial probability matrix
Allocating g as a 1-dimension array including n elements.
sum = 0
For j = 1 to n
    g[j] = α[1][j] * β[1][j]
    sum = sum + g[j]
End for j

If sum != 0 then
    For j = 1 to n
        π_j = g[j] / sum
```

ISSN: 0075-5222

```
        End for j
    End if

    //Updating observation probability distribution
    For j = 1 to n
        Allocating γ as a 1-dimension array including T elements.
        denominator = 0
        For t = 1 to T
            γ[t] = α[t][j] * β[t][j]
            denominator = denominator + γ[t]
        End for t

        Let m be the columns of observation distribution matrix B.
        For k = 1 to m
            numerator = 0
            For t = 1 to T
                If o_t == k then
                    numerator = numerator + γ[t]
                End if
            End for t

            b_jk = numerator / denominator
        End for k
    End for j

    iteration = iteration + 1
End while
```

Table I.3.2.2. Proposed implementation of EM algorithm for learning HMM with terminating criterion $P(O|\Delta)$

According to table I.3.2.2, the number of iterations is limited by a pre-defined maximum number, which aims to solve a so-called infinite loop optimization. Although it is proved that EM algorithm always converges, maybe there are two different estimates $\hat{\Delta}_1$ and $\hat{\Delta}_2$ at the final convergence. This situation causes EM algorithm to alternate between $\hat{\Delta}_1$ and $\hat{\Delta}_2$ in infinite loop. Therefore, the final estimate $\hat{\Delta}_1$ or $\hat{\Delta}_2$ is totally determined but the EM algorithm does not stop. This is the reason that the number of iterations is limited by a pre-defined maximum number.

Now three main problems of HMM are described; please see an excellent document "A tutorial on hidden Markov models and selected applications in speech recognition" written by author (Rabiner, 1989) for advanced details about HMM. The next section II described a HMM whose observations are continuous.

ISSN: 0075-5222

II. Continuous observation hidden Markov model

Observations of normal HMM mentioned in previous sub-section I are quantified by discrete probability distribution that is concretely observation probability matrix B. In the general situation, observation o_t is continuous variable and matrix B is replaced by probability density function (PDF). Formula II.1 specifies the PDF of continuous observation o_t given state s_j.

$$b_j(o_t) = p_j(o_t|\theta_j)$$

Formula II.1. Probability density function (PDF) of observation

Where the PDF $p_j(o_t|\theta_j)$ belongs to any probability distribution, for example, normal distribution, exponential distribution, etc. The notation θ_j denotes probabilistic parameters, for instance, if $p_j(o_t|\theta_j)$ is normal distribution PDF, θ_j includes mean m_j and variance σ_j^2. The HMM now is specified by parameter $\Delta = (a_{ij}, \theta_j, \pi_j)$, which is called *continuous observation HMM* (Rabiner, 1989, p. 267). The PDF $p_j(o_t|\theta_j)$ is known as *single PDF* because it is atom PDF which is not combined with any other PDF. We will research so-called mixture model PDF that is constituted of many partial PDF (s) later. We still apply EM algorithm known as Baum-Welch algorithm into learning continuous observation HMM. In the field of continuous-speech recognition, authors (Lee, Rabiner, Pieraccini, & Wilpon, 1990) proposed Bayesian adaptive learning for estimating mean and variance of continuous density HMM. Authors (Huo & Lee, 1997) proposed a framework of quasi-Bayes (QB) algorithm based on approximate recursive Bayes estimate for learning HMM parameters with Gaussian mixture model; they described that "The QB algorithm is designed to incrementally update the hyper-parameters of the approximate posterior distribution and the continuous density HMM parameters simultaneously" (Huo & Lee, 1997, p. 161). Authors (Sha & Saul, 2009) and (Cheng, Sha, & Saul, 2009) used the approach of large margin training to learn HMM parameters. Such approach is different from Baum-Welch algorithm when it firstly establishes discriminant functions for correct and incorrect label sequences and then, finds parameters satisfying the margin constraint that separates the discriminant functions as much as possible (Sha & Saul, 2009, pp. 106-108). Authors (Cheng, Sha, & Saul, 2009, p. 4) proposed a fast online algorithm for large margin training, in which "the parameters for discriminant functions are updated according to an online learning rule with given learning rate". Large margin training is very appropriate to speech recognition, which was proposed by authors (Sha & Saul, 2006) in the article "Large Margin Hidden Markov Models for Automatic Speech Recognition". Some other authors used different learning approaches such as conditional maximum likelihood and minimizing classification error, mentioned in (Sha & Saul, 2009, pp. 104-105).

Methods to solve evaluation problem and uncovering problem mentioned previous sub-sections I.1, I.2, and I.3 are kept intact by using the observation PDF specified by formula II.1. For example, forward-backward procedure (based on forward variable, shown in table I.1.1) that solves evaluation problem is based on the recurrence formula I.1.2 as follows:

ISSN: 0075-5222

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right) b_j(o_{t+1})$$

In order to apply forward-backward procedure into continuous observation HMM, it is simple to replace the discrete probability $b_j(o_{t+1})$ by the single PDF specified by formula II.1.

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right) p_j(o_{t+1}|\theta_j)$$

However, there is a change in solution of learning problem. Recall that the essence of EM algorithm applied into HMM learning problem is to determine the estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{\theta}_j, \hat{\pi}_j)$. Formulas for calculating estimates $\hat{a}_{ij}$ and $\hat{\pi}_j$ are kept intact, as aforementioned in formula I.3.2.7.

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=2}^{T} \sum_{l=1}^{n} \xi_t(i,l)}$$

$$\hat{\pi}_j = \frac{\gamma_1(j)}{\sum_{i=1}^{n} \gamma_1(i)}$$

Where joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ are modified based on replacing discrete probability $b_j(o_t)$ by the single PDF $p_j(o_t|\theta_j)$ given current parameter $\Delta = (a_{ij}, b_j(k), \pi_j)$.

$$\xi_t(i,j) = \alpha_{t-1}(i)a_{ij}p_j(o_t|\theta_j)\beta_t(j) \text{ where } t \geq 2$$

$$\gamma_t(j) = P(O, x_t = s_j|\Delta) = \alpha_t(j)\beta_t(j)$$

Where forward variable α_t and backward variable β_t are calculated by recurrence formulas I.1.2 and I.1.5. Recall that $\gamma_t(j)$ is joint probability that the stochastic process is in state s_j at time point t with observation sequence O and $\xi_t(i, j)$ is the joint probability that the stochastic process receives state s_i at time point t–1 and state s_j at time point t given observation sequence O.

Your attention please, quantities $\xi_t(i, j)$, $\gamma_t(j)$, $\alpha_t(i)$, and $\beta_t(j)$ are essentially continuous functions because they are based on PDF $p_j(o_t|\theta_j)$. Their values on a concrete observation o_t are zero because the value of PDF $p_j(o_t|\theta_j)$ given such concrete observation o_t is zero. Therefore, in practice, these quantities are calculated according to integral of PDF $p_j(o_t|\theta_j)$ in ε-vicinity of o_t where ε is very small positive number. The number ε can reflect inherent attribute of observation data with regard to measure bias, for example, if atmosphere humidity at time point t is $o_t = 0.5 \mp 0.01$, the measure bias is 0.01 and so we have ε=0.01. In addition, the number ε can be pre-defined fixedly by arbitrary very small number. For example, given ε=0.01 we have:

$$\int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j(o|\theta_j)\mathrm{d}o = \int_{0.5-0.01}^{0.5+0.01} p_j(o|\theta_j)\mathrm{d}o$$

If all o_t are intervals, for example, $0.1 \leq o_t \leq 0.2, 0.3 \leq o_{t+1} \leq 0.4, \ldots$ then, the integral of PDF $p_j(o_t|\theta_j)$ is calculated directly over such o_t.

$$\int_{o_t} p_j(o|\theta_j)\mathrm{d}o = \int_{0.1}^{0.2} p_j(o|\theta_j)\mathrm{d}o$$

ISSN: 0075-5222

Given the PDF $p_j(o_t|\theta_j)$ conforms normal distribution, it is easy to calculate the probability of o_t in ε-vicinity as the integral of PDF $p_j(o_t|\theta_j)$ in ε-vicinity of o_t as follows:

$$\int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j(o|\theta_j)\mathrm{d}o = ? \text{ if } p_j(o|\theta_j) \text{ is normal PDF}$$

The best way is to standardize the normal PDF $p_j(o_t|\theta_j)$ where $\theta_j = (m_j, \sigma_j^2)$ into cumulative standard normal distribution (Montgomery & Runger, 2003, p. 653). Let Φ be cumulative standard normal distribution (Montgomery & Runger, 2003, p. 653), we have:

$$\int_{-\infty}^{b} p_j(o|\theta_j)\mathrm{d}o = \Phi\left(\frac{b-m_j}{\sqrt{\sigma_j^2}}\right)$$

$$\int_{a}^{b} p_j(o|\theta_j)\mathrm{d}o = \Phi\left(\frac{b-m_j}{\sqrt{\sigma_j^2}}\right) - \Phi\left(\frac{a-m_j}{\sqrt{\sigma_j^2}}\right)$$

The quantities $\frac{b-m_j}{\sqrt{\sigma_j^2}}$ and $\frac{a-m_j}{\sqrt{\sigma_j^2}}$ are standardized values of b and a given PDF $p_j(o_t|\theta_j)$, respectively. The function Φ is always evaluated in popular. For instance, appendix A of the book "Applied Statistics and Probability for Engineers" by authors (Montgomery & Runger, 2003, p. 653) is a good reference for looking up some values of Φ. Please distinguish the function Φ from the set of possible discrete observations $\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_m\}$ aforementioned at the beginning of section I when they share the same notation.

$$\int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j(o|\theta_j)\mathrm{d}o = \Phi\left(\frac{o_t+\varepsilon-m_j}{\sqrt{\sigma_j^2}}\right) - \Phi\left(\frac{o_t-\varepsilon-m_j}{\sqrt{\sigma_j^2}}\right) \text{ if } p_j(o|\theta_j) \text{ is normal PDF}$$

Formula II.2 specifies quantities $\xi_t(i, j)$, $\gamma_t(j)$ according to integral of PDF $p_j(o_t|\theta_j)$.

$$\xi_t(i,j) = P(O, x_{t-1} = s_i, x_t = s_j|\Delta) = \alpha_{t-1}(i)a_{ij}\left(\int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j(o|\theta_j)\mathrm{d}o\right)\beta_t(j) \text{ where } t \geq 2$$

ISSN: 0075-5222

$$\gamma_t(j) = P(O, x_t = s_j|\Delta) = \alpha_t(j)\beta_t(j)$$

Where forward variable α_t and backward variable β_t are calculated based on recurrence formulas I.1.2 and I.1.5.

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right) \int_{o_{t+1}-\varepsilon}^{o_{t+1}+\varepsilon} p_j(o|\theta_j)\mathrm{d}o$$

$$\beta_t(i) = \sum_{j=1}^{n} a_{ij}\left(\int_{o_{t+1}-\varepsilon}^{o_{t+1}+\varepsilon} p_j(o|\theta_j)\mathrm{d}o\right)\beta_{t+1}(j)$$

$$\int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j(o|\theta_j)\mathrm{d}o$$

$$= \Phi\left(\frac{o_t + \varepsilon - m_j}{\sqrt{\sigma_j^2}}\right) - \Phi\left(\frac{o_t - \varepsilon - m_j}{\sqrt{\sigma_j^2}}\right) \text{ if } p_j(o|\theta_j) \text{ is normal PDF}$$

Formula II.2. Joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ based on single PDF

Note that m_j and σ_j^2 are mean and variance of the normal PDF $p_j(o_t|\theta_j)$ and Φ is cumulative standard normal distribution (Montgomery & Runger, 2003, p. 653).

As a convention, quantities $\xi_t(i, j)$, $\gamma_t(j)$, $\alpha_{t+1}(j)$, and $\beta_t(i)$ are still referred as joint probabilities, forward variable, and backward variable. This convention help us to describe traditional HMM and continuous convention HMM in coherent way.

Now it is necessary to determine the estimate $\hat{\theta}_j$. Derived from formula I.3.2.2, the expectation $E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}$ is modified based on replacing discrete probability $b_j(o_t)$ by the continuous PDF $p_j(o_t|\theta_j)$, as seen in following formula II.3 given current parameter Δ_r.

$$E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}$$

$$= \sum_X P(X|O, \Delta_r)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} I(x_{t-1} = s_i, x_t = s_j)ln(a_{ij}) + \sum_{j=1}^{n}\sum_{t=1}^{T} I(x_t = s_j)ln\left(p_j(o_t|\theta_j)\right)\right)$$

Formula II.3. EM conditional expectation for continuous observation HMM with single PDF

ISSN: 0075-5222

Where $I(x_{t-1}=s_i, x_t=s_j)$ and $I(x_t=s_j)$ are index functions so that

$$I(x_{t-1}=s_i, x_t=s_j)=\begin{cases}1 \text{ if } x_{t-1}=s_i \text{ and } x_t=s_j\\ 0 \text{ otherwise}\end{cases}$$

$$I(x_t=s_j)=\begin{cases}1 \text{ if } x_t=s_j\\ 0 \text{ otherwise}\end{cases}$$

Note that notation "*ln*" denotes natural logarithm function. Derived from formula I.3.2.3, the HMM Lagrangian function for continuous observation HMM with single PDF is specified by formula II.4.

$$l(\Delta,\lambda)=l(a_{ij},\theta_j,\lambda_i)=E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}+\sum_{i=1}^{n}\lambda_i\left(1-\sum_{j=1}^{n}a_{ij}\right)$$

Formula II.4. Lagrangian function for continuous observation HMM with single PDF

Where λ is *n*-component vector $\lambda = (\lambda_1, \lambda_2,\ldots, \lambda_n)$. Factors $\lambda_i \geq 0$ are called Lagrange multipliers or Karush-Kuhn-Tucker multipliers (Wikipedia, Karush–Kuhn–Tucker conditions, 2014) or dual variables.

The parameter estimate $\hat{\theta}_j$ which is extreme point of the Lagrangian function $l(\Delta, \lambda)$ is determined by setting partial derivatives of $l(\Delta, \lambda)$ with respect to θ_j to be zero. The partial derivative of $l(\Delta, \lambda)$ with respect to θ_j is:

$$\frac{\partial l(\Delta,\lambda)}{\partial\theta_j}=\frac{\partial E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}}{\partial\theta_j}+\frac{\partial}{\partial\theta_j}\left(\sum_{i=1}^{n}\lambda_i\left(1-\sum_{j=1}^{n}a_{ij}\right)\right)$$

$$=\frac{\partial E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}}{\partial\theta_j}$$

$$=\frac{\partial}{\partial\theta_j}\left(\sum_{X}P(X|O,\Delta_r)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}I(x_{t-1}=s_i,x_t=s_j)ln(a_{ij})+\sum_{j=1}^{n}\sum_{t=1}^{T}I(x_t=s_j)ln\left(p_j(o_t|\theta_j)\right)\right)\right)$$

$$=\sum_{X}P(X|O,\Delta_r)\sum_{j=1}^{n}\sum_{t=1}^{T}I(x_t=s_j)\frac{\partial}{\partial\theta_j}\left(ln\left(p_j(o_t|\theta_j)\right)\right)$$

$$=\sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I(x_t=s_j)\frac{\partial ln\left(p_j(o_t|\theta_j)\right)}{\partial\theta_j}$$

$$=\sum_{t=1}^{T}\sum_{X}I(x_t=s_j)P(X|O,\Delta_r)\frac{\partial ln\left(p_j(o_t|\theta_j)\right)}{\partial\theta_j}$$

$$=\sum_{t=1}^{T}\sum_{X}I(x_t=s_j)P(x_1,\ldots,x_t,\ldots,x_T|O,\Delta_r)\frac{\partial ln\left(p_j(o_t|\theta_j)\right)}{\partial\theta_j}$$

ISSN: 0075-5222

$$= \sum_{t=1}^{T} P(x_t = s_j | O, \Delta_r) \frac{\partial ln\left(p_j(o_t|\theta_j)\right)}{\partial \theta_j}$$

(Due to total probability rule)

$$= \sum_{t=1}^{T} \frac{P(O, x_t = s_j | \Delta_r)}{P(O|\Delta_r)} \frac{\partial ln\left(p_j(o_t|\theta_j)\right)}{\partial \theta_j}$$

(Due to multiplication rule)

$$= \frac{1}{P(O|\Delta_r)} \sum_{t=1}^{T} P(O, x_t = s_j | \Delta_r) \frac{\partial ln\left(p_j(o_t|\theta_j)\right)}{\partial \theta_j}$$

$$= \frac{1}{P(O|\Delta_r)} \sum_{t=1}^{T} \gamma_t(j) \frac{\partial ln\left(p_j(o_t|\theta_j)\right)}{\partial \theta_j}$$

Setting the partial derivative $\frac{\partial l(\Delta,\lambda)}{\partial \theta_j}$ to be zero, we get the equation whose solution is estimate $\hat{\theta}_j$, specified by formula II.5.

$$\frac{1}{P(O|\Delta_r)} \sum_{t=1}^{T} \gamma_t(j) \frac{\partial ln\left(p_j(o_t|\theta_j)\right)}{\partial \theta_j} = 0 \Leftrightarrow \sum_{t=1}^{T} \gamma_t(j) \frac{\partial ln\left(p_j(o_t|\theta_j)\right)}{\partial \theta_j} = 0$$

Formula II.5. Equation of single PDF parameter

Note that notation "*ln*" denotes natural logarithm function.

It is possible to solve the above equation (formula II.5) by Newton-Raphson method (Burden & Faires, 2011, pp. 67-69) – a numeric analysis method but it is easier and simpler to find out more precise solution if the PDF $p_j(o_t|\theta_j)$ belongs to well-known distributions: normal distribution (Montgomery & Runger, 2003, pp. 109-110), exponential distribution (Montgomery & Runger, 2003, pp. 122-123), etc. According to author (Couvreur, 1996, p. 32), the estimate $\hat{\theta}_j$ is determined by the more general formula as follows:

$$\hat{\theta}_j = \underset{\theta_j}{\operatorname{argmax}} \sum_{t=1}^{T} \gamma_t(j) ln\left(p_j(o_t|\theta_j)\right)$$

The easy way to find out $\hat{\theta}_j$ is to solve formula II.5 by taking advantages of derivatives.

Suppose $p_j(o_t|\theta_j)$ is normal PDF whose parameter is $\theta_j = (m_j, \sigma_j^2)$ where m_j and σ_j^2 are mean and variance, respectively.

$$p_j(o_t|\theta_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(-\frac{1}{2}\frac{(o_t - m_j)^2}{\sigma_j^2}\right)$$

Note that notation "*exp*" denotes exponential function. The equation specified by formula II.5 is re-written with regard to parameter m_j as follows:

ISSN: 0075-5222

$$\sum_{t=1}^{T}\gamma_t(j)\frac{\partial ln\left(\frac{1}{\sqrt{2\pi\sigma_j^2}}exp\left(-\frac{1}{2}\frac{(o_t-m_j)^2}{\sigma_j^2}\right)\right)}{\partial m_j}=0$$

$$\Rightarrow\sum_{t=1}^{T}\gamma_t(j)\frac{\partial\left(-\frac{1}{2}\left(ln(2\pi)+ln(\sigma_j^2)+\frac{(o_t-m_j)^2}{\sigma_j^2}\right)\right)}{\partial m_j}=0$$

$$\Rightarrow-\frac{1}{2}\sum_{t=1}^{T}\gamma_t(j)\frac{(o_t-m_j)}{\sigma_j^2}=0\Rightarrow-\frac{1}{2\sigma_j^2}\sum_{t=1}^{T}\gamma_t(j)(o_t-m_j)=0$$

$$\Rightarrow\sum_{t=1}^{T}\gamma_t(j)(o_t-m_j)=0\Rightarrow\sum_{t=1}^{T}\gamma_t(j)o_t-m_j\sum_{t=1}^{T}\gamma_t(j)=0$$

$$\Rightarrow m_j=\frac{\sum_{t=1}^{T}\gamma_t(j)o_t}{\sum_{t=1}^{T}\gamma_t(j)}$$

Therefore, the estimate $\hat{m}_j$ is

$$\hat{m}_j=\frac{\sum_{t=1}^{T}\gamma_t(j)o_t}{\sum_{t=1}^{T}\gamma_t(j)}$$

The equation specified by formula II.5 is re-written with regard to parameter σ_j^2 as follows:

$$\sum_{t=1}^{T}\gamma_t(j)\frac{\partial ln\left(\frac{1}{\sqrt{2\pi\sigma_j^2}}exp\left(-\frac{1}{2}\frac{(o_t-m_j)^2}{\sigma_j^2}\right)\right)}{\partial\sigma_j^2}=0$$

$$\Rightarrow\sum_{t=1}^{T}\gamma_t(j)\frac{\partial\left(-\frac{1}{2}\left(ln(2\pi)+ln(\sigma_j^2)+\frac{(o_t-m_j)^2}{\sigma_j^2}\right)\right)}{\partial\sigma_j^2}=0$$

$$\Rightarrow-\frac{1}{2}\sum_{t=1}^{T}\gamma_t(j)\left(\frac{1}{\sigma_j^2}-\frac{(o_t-m_j)^2}{(\sigma_j^2)^2}\right)=0$$

$$\Rightarrow-\frac{1}{2\sigma_j^2}\sum_{t=1}^{T}\gamma_t(j)\left(1-\frac{(o_t-m_j)^2}{\sigma_j^2}\right)=0$$

$$\Rightarrow\sum_{t=1}^{T}\gamma_t(j)\left(1-\frac{(o_t-m_j)^2}{\sigma_j^2}\right)=0$$

ISSN: 0075-5222

$$\Rightarrow \sum_{t=1}^{T} \gamma_t(j) - \frac{1}{\sigma_j^2} \sum_{t=1}^{T} \gamma_t(j)\left(o_t - m_j\right)^2 = 0$$

$$\Rightarrow \sigma_j^2 = \frac{\sum_{t=1}^{T} \gamma_t(j)\left(o_t - m_j\right)^2}{\sum_{t=1}^{T} \gamma_t(j)}$$

It implies that given the estimate $\hat{m}_j$, the estimate $\hat{\sigma}_j^2$ is:

$$\hat{\sigma}_j^2 = \frac{\sum_{t=1}^{T} \gamma_t(j)\left(o_t - \hat{m}_j\right)^2}{\sum_{t=1}^{T} \gamma_t(j)}$$

In general, the normal parameter estimate $\hat{\theta}_j$ is:

$$\hat{\theta}_j = \left(\hat{m}_j = \frac{\sum_{t=1}^{T} \gamma_t(j) o_t}{\sum_{t=1}^{T} \gamma_t(j)}, \hat{\sigma}_j^2 = \frac{\sum_{t=1}^{T} \gamma_t(j)\left(o_t - \hat{m}_j\right)^2}{\sum_{t=1}^{T} \gamma_t(j)}\right)$$

Suppose $p_j(o_t|\theta_j)$ is exponential PDF whose parameter is $\theta_j = \kappa_j$ as follows:

$$p_j\left(o_t \middle| \theta_j\right) = \kappa_j e^{-\kappa_j o_t}$$

Note that notations "*exp*" and "$e^{(.)}$" denote exponential function. The equation specified by formula II.5 is re-written with regard to parameter κ_j as follows:

$$\sum_{t=1}^{T} \gamma_t(j) \frac{\partial ln\left(\kappa_j e^{-\kappa_j o_t}\right)}{\partial \kappa_j} = 0$$

$$\Rightarrow \sum_{t=1}^{T} \gamma_t(j) \frac{\partial \left(ln\left(\kappa_j\right) - \kappa_j o_t\right)}{\partial \kappa_j} = 0$$

$$\Rightarrow \sum_{t=1}^{T} \gamma_t(j) \left(\frac{1}{\kappa_j} - o_t\right) = 0$$

$$\Rightarrow \frac{1}{\kappa_j} \sum_{t=1}^{T} \gamma_t(j) - \sum_{t=1}^{T} \gamma_t(j) o_t = 0$$

$$\Rightarrow \kappa_j = \frac{\sum_{t=1}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j) o_t}$$

Therefore, the exponential parameter estimate $\hat{\theta}_j$ is

$$\hat{\theta}_j = \hat{\kappa}_j = \frac{\sum_{t=1}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j) o_t}$$

Shortly, the continuous observation HMM parameter estimate $\hat{\Delta} = \left(\hat{a}_{ij}, \hat{\theta}_j, \hat{\pi}_i\right)$ with single PDF given current parameter $\Delta = (a_{ij}, b_i(k), \pi_i)$ is specified by formula II.6.

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=2}^{T} \sum_{l=1}^{n} \xi_t(i,l)}$$

$$\hat{\pi}_j = \frac{\gamma_1(j)}{\sum_{i=1}^{n} \gamma_1(i)}$$

ISSN: 0075-5222

$$\hat{\theta}_j \text{ is the solution of } \sum_{t=1}^{T} \gamma_t(j) \frac{\partial ln\left(p_j(o_t|\theta_j)\right)}{\partial \theta_j} = 0$$

With normal distribution:

$$\hat{\theta}_j = \left(\hat{m}_j = \frac{\sum_{t=1}^{T} \gamma_t(j) o_t}{\sum_{t=1}^{T} \gamma_t(j)}, \hat{\sigma}_j^2 = \frac{\sum_{t=1}^{T} \gamma_t(j)\left(o_t - \hat{m}_j\right)^2}{\sum_{t=1}^{T} \gamma_t(j)} \right)$$

With exponential distribution:

$$\hat{\theta}_j = \hat{\kappa}_j = \frac{\sum_{t=1}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j) o_t}$$

Formula II.6. Continuous observation HMM parameter estimate with single PDF

Where joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ based on single PDF $p_j(o_t|\theta_j)$ is specified by formula II.2.

The EM algorithm applied into learning continuous observation HMM parameter with single PDF is described in table II.1.

Starting with initial value for Δ, each iteration in EM algorithm has two steps:

1. *E-step*: Calculating the joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ according to formula II.2.

$$\xi_t(i,j) = \alpha_{t-1}(i) a_{ij} \left(\int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j(o|\theta_j) \mathrm{d}o \right) \beta_t(j) \text{ where } t \geq 2$$

$$\gamma_t(j) = \alpha_t(j) \beta_t(j)$$

2. *M-step*: Calculating the estimate $\hat{\Delta} = \left(\hat{a}_{ij}, \hat{\theta}_j, \hat{\pi}_j\right)$ based on the joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ determined at E-step, according to formula II.6.The estimate $\hat{\Delta}$ becomes the current parameter for next iteration.

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=2}^{T} \sum_{l=1}^{n} \xi_t(i,l)}$$

$$\hat{\pi}_j = \frac{\gamma_1(j)}{\sum_{i=1}^{n} \gamma_1(i)}$$

$$\hat{\theta}_j \text{ is the solution of } \sum_{t=1}^{T} \gamma_t(j) \frac{\partial ln\left(p_j(o_t|\theta_j)\right)}{\partial \theta_j} = 0$$

With normal distribution:

$$\hat{\theta}_j = \left(\hat{m}_j = \frac{\sum_{t=1}^{T} \gamma_t(j) o_t}{\sum_{t=1}^{T} \gamma_t(j)}, \hat{\sigma}_j^2 = \frac{\sum_{t=1}^{T} \gamma_t(j)\left(o_t - \hat{m}_j\right)^2}{\sum_{t=1}^{T} \gamma_t(j)} \right)$$

With exponential distribution:

$$\hat{\theta}_j = \hat{\kappa}_j = \frac{\sum_{t=1}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j) o_t}$$

EM algorithm stops when it meets the terminating condition, for example, the

ISSN: 0075-5222

difference of current parameter Δ and next parameter $\widehat{\Delta}$ is insignificant. It is possible to define a custom terminating condition. The terminating criterion $P(O|\Delta)$ described in table I.3.2.2 is a suggestion.

Table II.1. EM algorithm applied into learning continuous observation HMM parameter with single PDF

Going back the weather example, there are some states of weather: *sunny*, *cloudy*, and *rainy*. Suppose you are in the room and do not know the weather outside but you are notified air humidity measures as observations from someone else. You can forecast weather based on humidity. However, humidity is not still categorized into discrete values such as *dry*, *dryish*, *damp*, and *soggy*. The humidity is now continuous real number, which is used to illustrate continuous observation HMM. It is required to discuss humidity a little bit.

Absolute humidity of atmosphere is measured as amount of water vapor (kilogram) in 1 cubic meter (m^3) volume of air (Gallová & Kučerka).

$$h = \frac{m_w}{V}$$

Where m_w is the amount of water vapor and V is the volume of air. The SI unit (NIST, 2008) of absolute humidity is kg/m^3.

The amount of water vapor in the air conforms to environment conditions such as temperature and pressure. Given environment conditions, there is a saturation point at which absolute humidity becomes maximal, denoted h_{max}. *Relative humidity* is ratio of the absolute humidity h to its maximal value h_{max} (Gallová & Kučerka).

$$rh = \frac{h}{h_{max}}$$

The relative humidity rh is always less than or equal to 1. Relative humidity rh is near to 0 then, the air is dry. Relative humidity rh is near to 1 then, the air is soggy. It is comfortable for human if relative humidity is between 0.5 and 0.7. Relative humidity is used in our weather example instead of absolute humidity. Suppose continuous observation sequence is

$$O = \{o_1=0.88, o_2=0.13, o_3=0.38\}$$

These observations are relative humidity measures. The bias for all measures is $\varepsilon=0.01$, for example, the first observation $o_1=0.88$ ranges in interval [0.88–0.01, 0.88+0.01]. Given weather HMM Δ whose parameters A and $\prod$ specified in tables I.1 and I.2 is shown below:

		Weather current day (Time point t)		
		sunny	*cloudy*	*rainy*
Weather previous day (Time point t –1)	*sunny*	$a_{11}=0.50$	$a_{12}=0.25$	$a_{13}=0.25$
	cloudy	$a_{21}=0.30$	$a_{22}=0.40$	$a_{23}=0.30$
	rainy	$a_{31}=0.25$	$a_{32}=0.25$	$a_{33}=0.50$

sunny	*cloudy*	*rainy*
$\pi_1=0.33$	$\pi_2=0.33$	$\pi_3=0.33$

ISSN: 0075-5222

The observation probability distribution *B* now includes three normal PDF (s): $p_1(o_t|\theta_1)$, $p_2(o_t|\theta_2)$, and $p_3(o_t|\theta_3)$ corresponding to three states: s_1=*sunny*, s_2=*cloudy*, and s_3=*rainy*. Following is the specification of these normal PDF (s).

$$p_1(o_t|\theta_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left(-\frac{1}{2}\frac{(o_t - m_1)^2}{\sigma_1^2}\right)$$
$$p_2(o_t|\theta_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\left(-\frac{1}{2}\frac{(o_t - m_2)^2}{\sigma_2^2}\right)$$
$$p_3(o_t|\theta_3) = \frac{1}{\sqrt{2\pi\sigma_3^2}} exp\left(-\frac{1}{2}\frac{(o_t - m_3)^2}{\sigma_3^2}\right)$$

Where $\theta_1 = (m_1, \sigma_1^2)$, $\theta_2 = (m_2, \sigma_2^2)$, and $\theta_3 = (m_3, \sigma_3^2)$ are means and variances of $p_1(o_t|\theta_1)$, $p_2(o_t|\theta_2)$, and $p_3(o_t|\theta_3)$.

As a convention, observation PDF (s) such as $p_1(o_t|\theta_1)$, $p_2(o_t|\theta_2)$, and $p_3(o_t|\theta_3)$ are represented by theirs means and variances (m_1, σ_1^2), (m_2, σ_2^2), and (m_3, σ_3^2). These means and variances are also called *observation probability parameters* that substitute for discrete matrix *B*. Table II.2 shows observation probability parameters for our weather example.

$p_1(o_t\|\theta_1)$	m_1 = 0.87	σ_1^2 = 0.9
$p_2(o_t\|\theta_2)$	m_2 = 0.14	σ_1^2 = 0.9
$p_3(o_t\|\theta_3)$	m_3 = 0.39	σ_1^2 = 0.9

Table II.2. Observation probability parameters for weather example

Obviously, we have:

$$p_1(o_t|\theta_1) = \frac{1}{\sqrt{2\pi 0.9}} exp\left(-\frac{1}{2}\frac{(o_t - 0.87)^2}{0.9}\right)$$
$$p_2(o_t|\theta_2) = \frac{1}{\sqrt{2\pi 0.9}} exp\left(-\frac{1}{2}\frac{(o_t - 0.14)^2}{0.9}\right)$$
$$p_3(o_t|\theta_3) = \frac{1}{\sqrt{2\pi 0.9}} exp\left(-\frac{1}{2}\frac{(o_t - 0.39)^2}{0.9}\right)$$

EM algorithm described in table II.1 is applied into calculating the parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{\theta}_j, \hat{\pi}_j)$ given continuous observation sequence O = {o_1=0.88, o_2=0.13, o_3=0.38} and continuous normal PDF (s) whose means and variances shown in table II.2. For convenience, all floating-point values are rounded off until ten decimal numbers.

As a convention, let

$$b_j(o_t) = \int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j(o|\theta_j)\mathrm{d}o = \Phi\left(\frac{o_t + \varepsilon - m_j}{\sqrt{\sigma_j^2}}\right) - \Phi\left(\frac{o_t - \varepsilon - m_j}{\sqrt{\sigma_j^2}}\right)$$

At the first iteration (r=1) we have:

ISSN: 0075-5222

$$b_1(o_1) = b_1(0.88) = \int_{0.88-0.01}^{0.88+0.01} p_1(o|\theta_1)\mathrm{d}o$$
$$= \Phi\left(\frac{0.88 + 0.01 - 0.87}{\sqrt{0.9}}\right) - \Phi\left(\frac{0.88 - 0.01 - 0.87}{\sqrt{0.9}}\right)$$
$$= 0.0084098112$$
$$b_1(o_2) = b_1(0.13) = \int_{0.13-0.01}^{0.13+0.01} p_1(o|\theta_1)\mathrm{d}o$$
$$= \Phi\left(\frac{0.13 + 0.01 - 0.87}{\sqrt{0.9}}\right) - \Phi\left(\frac{0.13 - 0.01 - 0.87}{\sqrt{0.9}}\right)$$
$$= 0.0062043131$$
$$b_1(o_3) = b_1(0.38) = \int_{0.38-0.01}^{0.38+0.01} p_1(o|\theta_1)\mathrm{d}o$$
$$= \Phi\left(\frac{0.38 + 0.01 - 0.87}{\sqrt{0.9}}\right) - \Phi\left(\frac{0.38 - 0.01 - 0.87}{\sqrt{0.9}}\right)$$
$$= 0.0073600784$$
$$b_2(o_1) = b_2(0.88) = \int_{0.88-0.01}^{0.88+0.01} p_2(o|\theta_2)\mathrm{d}o$$
$$= \Phi\left(\frac{0.88 + 0.01 - 0.14}{\sqrt{0.9}}\right) - \Phi\left(\frac{0.88 - 0.01 - 0.14}{\sqrt{0.9}}\right)$$
$$= 0.0062043061$$
$$b_2(o_2) = b_2(0.13) = \int_{0.13-0.01}^{0.13+0.01} p_2(o|\theta_2)\mathrm{d}o$$
$$= \Phi\left(\frac{0.13 + 0.01 - 0.14}{\sqrt{0.9}}\right) - \Phi\left(\frac{0.13 - 0.01 - 0.14}{\sqrt{0.9}}\right)$$
$$= 0.0084098205$$
$$b_2(o_3) = b_2(0.38) = \int_{0.38-0.01}^{0.38+0.01} p_2(o|\theta_2)\mathrm{d}o$$
$$= \Phi\left(\frac{0.38 + 0.01 - 0.14}{\sqrt{0.9}}\right) - \Phi\left(\frac{0.38 - 0.01 - 0.14}{\sqrt{0.9}}\right)$$
$$= 0.0081454189$$
$$b_3(o_1) = b_3(0.88) = \int_{0.88-0.01}^{0.88+0.01} p_3(o|\theta_3)\mathrm{d}o$$
$$= \Phi\left(\frac{0.88 + 0.01 - 0.39}{\sqrt{0.9}}\right) - \Phi\left(\frac{0.88 - 0.01 - 0.39}{\sqrt{0.9}}\right)$$
$$= 0.0073600784$$

ISSN: 0075-5222

$$b_3(o_2) = b_3(0.13) = \int_{0.13-0.01}^{0.13+0.01} p_3(o|\theta_3)\mathrm{d}o$$
$$= \Phi\left(\frac{0.13 + 0.01 - 0.39}{\sqrt{0.9}}\right) - \Phi\left(\frac{0.13 - 0.01 - 0.39}{\sqrt{0.9}}\right)$$
$$= 0.0081003029$$
$$b_3(o_3) = b_3(0.38) = \int_{0.38-0.01}^{0.38+0.01} p_3(o|\theta_3)\mathrm{d}o$$
$$= \Phi\left(\frac{0.38 + 0.01 - 0.39}{\sqrt{0.9}}\right) - \Phi\left(\frac{0.38 - 0.01 - 0.39}{\sqrt{0.9}}\right)$$
$$= 0.0084098112$$

$$\alpha_1(1) = b_1(o_1)\pi_1 = 0.0027752379$$
$$\alpha_1(2) = b_2(o_1)\pi_2 = 0.0020474212$$
$$\alpha_1(3) = b_3(o_1)\pi_3 = 0.0024288259$$
$$\alpha_2(1) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i1}\right) b_1(o_2) = 0.0000161874$$
$$\alpha_2(2) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i2}\right) b_2(o_2) = 0.0000178287$$
$$\alpha_2(3) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i3}\right) b_3(o_2) = 0.0000204326$$
$$\alpha_3(1) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i1}\right) b_1(o_3) = 0.0000001365$$
$$\alpha_3(2) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i2}\right) b_2(o_3) = 0.0000001327$$
$$\alpha_3(3) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i3}\right) b_3(o_3) = 0.0000001649$$
$$\beta_3(1) = \beta_3(2) = \beta_3(3) = 1$$
$$\beta_2(1) = \sum_{j=1}^{n} a_{1j}b_j(o_3)\beta_3(j) = 0.0078188467$$
$$\beta_2(2) = \sum_{j=1}^{n} a_{2j}b_j(o_3)\beta_3(j) = 0.0079891346$$
$$\beta_2(3) = \sum_{j=1}^{n} a_{3j}b_j(o_3)\beta_3(j) = 0.0080812799$$

ISSN: 0075-5222

$$\beta_1(1) = \sum_{j=1}^{n} a_{1j} b_j(o_2) \beta_2(j) = 0.0000574173$$
$$\beta_1(2) = \sum_{j=1}^{n} a_{2j} b_j(o_2) \beta_2(j) = 0.0000610663$$
$$\beta_1(3) = \sum_{j=1}^{n} a_{3j} b_j(o_2) \beta_2(j) = 0.0000616548$$

Within the E-step of the first iteration (r=1), the terminating criterion $P(O|\Delta)$ is calculated according to forward-backward procedure (see table I.1.1) as follows:

$$P(O|\Delta) = \alpha_3(1) + \alpha_3(2) + \alpha_3(3) = 0.0000004341$$

Within the E-step of the first iteration (r=1), the joint probabilities $\xi_t(i,j)$ and $\gamma_t(j)$ are calculated based on formula II.2 as follows:

$$\xi_2(1,1) = \alpha_1(1) a_{11} b_1(o_2) \beta_2(1) = 0.0000000673$$
$$\xi_2(1,2) = \alpha_1(1) a_{12} b_2(o_2) \beta_2(2) = 0.0000000466$$
$$\xi_2(1,3) = \alpha_1(1) a_{13} b_3(o_2) \beta_2(3) = 0.0000000454$$
$$\xi_2(2,1) = \alpha_1(2) a_{21} b_1(o_2) \beta_2(1) = 0.0000000298$$
$$\xi_2(2,2) = \alpha_1(2) a_{22} b_2(o_2) \beta_2(2) = 0.0000000550$$
$$\xi_2(2,3) = \alpha_1(2) a_{23} b_3(o_2) \beta_2(3) = 0.0000000402$$
$$\xi_2(3,1) = \alpha_1(3) a_{31} b_1(o_2) \beta_2(1) = 0.0000000295$$
$$\xi_2(3,2) = \alpha_1(3) a_{32} b_2(o_2) \beta_2(2) = 0.0000000408$$
$$\xi_2(3,3) = \alpha_1(3) a_{33} b_3(o_2) \beta_2(3) = 0.0000000795$$
$$\xi_3(1,1) = \alpha_2(1) a_{11} b_1(o_3) \beta_3(1) = 0.0000000596$$
$$\xi_3(1,2) = \alpha_2(1) a_{12} b_2(o_3) \beta_3(2) = 0.0000000330$$
$$\xi_3(1,3) = \alpha_2(1) a_{13} b_3(o_3) \beta_3(3) = 0.0000000340$$
$$\xi_3(2,1) = \alpha_2(2) a_{21} b_1(o_3) \beta_3(1) = 0.0000000394$$
$$\xi_3(2,2) = \alpha_2(2) a_{22} b_2(o_3) \beta_3(2) = 0.0000000581$$
$$\xi_3(2,3) = \alpha_2(2) a_{23} b_3(o_3) \beta_3(3) = 0.0000000450$$
$$\xi_3(3,1) = \alpha_2(3) a_{31} b_1(o_3) \beta_3(1) = 0.0000000376$$
$$\xi_3(3,2) = \alpha_2(3) a_{32} b_2(o_3) \beta_3(2) = 0.0000000416$$
$$\xi_3(3,3) = \alpha_2(3) a_{33} b_3(o_3) \beta_3(3) = 0.0000000859$$

$$\gamma_1(1) = \alpha_1(1) \beta_1(1) = 0.0000001593$$
$$\gamma_1(2) = \alpha_1(2) \beta_1(2) = 0.0000001250$$
$$\gamma_1(3) = \alpha_1(3) \beta_1(3) = 0.0000001497$$
$$\gamma_2(1) = \alpha_2(1) \beta_2(1) = 0.0000001266$$
$$\gamma_2(2) = \alpha_2(2) \beta_2(2) = 0.0000001424$$
$$\gamma_2(3) = \alpha_2(3) \beta_2(3) = 0.0000001651$$
$$\gamma_3(1) = \alpha_3(1) \beta_3(1) = 0.0000001365$$
$$\gamma_3(2) = \alpha_3(2) \beta_3(2) = 0.0000001327$$
$$\gamma_3(3) = \alpha_3(3) \beta_3(3) = 0.0000001649$$

Within the M-step of the first iteration (r=1), the estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ is calculated based on joint probabilities $\xi_t(i,j)$ and $\gamma_t(j)$ determined at E-step.

ISSN: 0075-5222

$$\hat{a}_{11} = \frac{\sum_{t=2}^{3} \xi_t(1,1)}{\sum_{t=2}^{3} \sum_{l=1}^{3} \xi_t(1,l)} = 0.443786$$
$$\hat{a}_{12} = \frac{\sum_{t=2}^{3} \xi_t(1,2)}{\sum_{t=2}^{3} \sum_{l=1}^{3} \xi_t(1,l)} = 0.278330$$
$$\hat{a}_{13} = \frac{\sum_{t=2}^{3} \xi_t(1,3)}{\sum_{t=2}^{3} \sum_{l=1}^{3} \xi_t(1,l)} = 0.277883$$
$$\hat{a}_{21} = \frac{\sum_{t=2}^{3} \xi_t(2,1)}{\sum_{t=2}^{3} \sum_{l=1}^{3} \xi_t(2,l)} = 0.258587$$
$$\hat{a}_{22} = \frac{\sum_{t=2}^{3} \xi_t(2,2)}{\sum_{t=2}^{3} \sum_{l=1}^{3} \xi_t(2,l)} = 0.422909$$
$$\hat{a}_{23} = \frac{\sum_{t=2}^{3} \xi_t(2,3)}{\sum_{t=2}^{3} \sum_{l=1}^{3} \xi_t(2,l)} = 0.318504$$
$$\hat{a}_{31} = \frac{\sum_{t=2}^{3} \xi_t(3,1)}{\sum_{t=2}^{3} \sum_{l=1}^{3} \xi_t(3,l)} = 0.212952$$
$$\hat{a}_{32} = \frac{\sum_{t=2}^{3} \xi_t(3,2)}{\sum_{t=2}^{3} \sum_{l=1}^{3} \xi_t(3,l)} = 0.261709$$
$$\hat{a}_{33} = \frac{\sum_{t=2}^{3} \xi_t(3,3)}{\sum_{t=2}^{3} \sum_{l=1}^{3} \xi_t(3,l)} = 0.525339$$

$$\hat{m}_1 = \frac{\sum_{t=1}^{3} \gamma_t(1) o_t}{\sum_{t=1}^{3} \gamma_t(1)} = 0.493699$$
$$\hat{\sigma}_1^2 = \frac{\sum_{t=1}^{3} \gamma_t(1)(o_t - \hat{m}_1)^2}{\sum_{t=1}^{3} \gamma_t(1)} = 0.100098$$
$$\hat{m}_2 = \frac{\sum_{t=1}^{3} \gamma_t(2) o_t}{\sum_{t=1}^{3} \gamma_t(2)} = 0.447242$$
$$\hat{\sigma}_2^2 = \frac{\sum_{t=1}^{3} \gamma_t(2)(o_t - \hat{m}_2)^2}{\sum_{t=1}^{3} \gamma_t(2)} = 0.095846$$
$$\hat{m}_3 = \frac{\sum_{t=1}^{3} \gamma_t(3) o_t}{\sum_{t=1}^{3} \gamma_t(3)} = 0.450017$$
$$\hat{\sigma}_3^2 = \frac{\sum_{t=1}^{3} \gamma_t(3)(o_t - \hat{m}_3)^2}{\sum_{t=1}^{3} \gamma_t(3)} = 0.094633$$

$$\hat{\pi}_1 = \frac{\gamma_1(1)}{\gamma_1(1) + \gamma_1(2) + \gamma_1(3)} = 0.367053$$
$$\hat{\pi}_2 = \frac{\gamma_1(2)}{\gamma_1(1) + \gamma_1(2) + \gamma_1(3)} = 0.288002$$
$$\hat{\pi}_3 = \frac{\gamma_1(3)}{\gamma_1(1) + \gamma_1(2) + \gamma_1(3)} = 0.344945$$

ISSN: 0075-5222

At the second iteration (r=2), the current parameter $\Delta = (a_{ij}, \theta_j, \pi_j)$ is received values from the previous estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{\theta}_j, \hat{\pi}_j)$, as seen in table II.3.

$a_{11} = \hat{a}_{11} = 0.443786$	$a_{12} = \hat{a}_{12} = 0.278330$	$a_{13} = \hat{a}_{13} = 0.277883$
$a_{21} = \hat{a}_{21} = 0.258587$	$a_{22} = \hat{a}_{22} = 0.422909$	$a_{23} = \hat{a}_{23} = 0.318504$
$a_{31} = \hat{a}_{31} = 0.212952$	$a_{32} = \hat{a}_{32} = 0.261709$	$a_{33} = \hat{a}_{33} = 0.525339$
$m_1 = \hat{m}_1 = 0.493699$	$\sigma_1^2 = \hat{\sigma}_1^2 = 0.100098$	
$m_2 = \hat{m}_2 = 0.447242$	$\sigma_2^2 = \hat{\sigma}_2^2 = 0.095846$	
$m_3 = \hat{m}_3 = 0.450017$	$\sigma_3^2 = \hat{\sigma}_3^2 = 0.094633$	
$\pi_1 = \hat{\pi}_1 = 0.367053$	$\pi_2 = \hat{\pi}_2 = 0.288002$	$\pi_3 = \hat{\pi}_3 = 0.344945$
Terminating criterion $P(O\|\Delta) = 0.0000004341$		

Table II.3. Continuous observation HMM parameters resulted from the first iteration of EM algorithm

We have:

$$b_1(o_1) = b_1(0.88) = \int_{0.88-0.01}^{0.88+0.01} p_1(o|\theta_1)\mathrm{d}o$$
$$= \Phi\left(\frac{0.88 + 0.01 - 0.493699}{\sqrt{0.100098}}\right) - \Phi\left(\frac{0.88 - 0.01 - 0.493699}{\sqrt{0.100098}}\right)$$
$$= 0.0119683407$$

$$b_1(o_2) = b_1(0.13) = \int_{0.13-0.01}^{0.13+0.01} p_1(o|\theta_1)\mathrm{d}o$$
$$= \Phi\left(\frac{0.13 + 0.01 - 0.493699}{\sqrt{0.100098}}\right) - \Phi\left(\frac{0.13 - 0.01 - 0.493699}{\sqrt{0.100098}}\right)$$
$$= 0.0130255492$$

$$b_1(o_3) = b_1(0.38) = \int_{0.38-0.01}^{0.38+0.01} p_1(o|\theta_1)\mathrm{d}o$$
$$= \Phi\left(\frac{0.38 + 0.01 - 0.493699}{\sqrt{0.100098}}\right) - \Phi\left(\frac{0.38 - 0.01 - 0.493699}{\sqrt{0.100098}}\right)$$
$$= 0.0236385148$$

$$b_2(o_1) = b_2(0.88) = \int_{0.88-0.01}^{0.88+0.01} p_2(o|\theta_2)\mathrm{d}o$$
$$= \Phi\left(\frac{0.88 + 0.01 - 0.447242}{\sqrt{0.095846}}\right) - \Phi\left(\frac{0.88 - 0.01 - 0.447242}{\sqrt{0.095846}}\right)$$
$$= 0.0097034648$$

ISSN: 0075-5222

$$b_2(o_2) = b_2(0.13) = \int_{0.13-0.01}^{0.13+0.01} p_2(o|\theta_2)\mathrm{d}o$$
$$= \Phi\left(\frac{0.13 + 0.01 - 0.447242}{\sqrt{0.095846}}\right) - \Phi\left(\frac{0.13 - 0.01 - 0.447242}{\sqrt{0.095846}}\right)$$
$$= 0.0152455261$$
$$b_2(o_3) = b_2(0.38) = \int_{0.38-0.01}^{0.38+0.01} p_2(o|\theta_2)\mathrm{d}o$$
$$= \Phi\left(\frac{0.38 + 0.01 - 0.447242}{\sqrt{0.095846}}\right) - \Phi\left(\frac{0.38 - 0.01 - 0.447242}{\sqrt{0.095846}}\right)$$
$$= 0.0251673553$$
$$b_3(o_1) = b_3(0.88) = \int_{0.88-0.01}^{0.88+0.01} p_3(o|\theta_3)\mathrm{d}o$$
$$= \Phi\left(\frac{0.88 + 0.01 - 0.450017}{\sqrt{0.094633}}\right) - \Phi\left(\frac{0.88 - 0.01 - 0.450017}{\sqrt{0.094633}}\right)$$
$$= 0.0097667109$$
$$b_3(o_2) = b_3(0.13) = \int_{0.13-0.01}^{0.13+0.01} p_3(o|\theta_3)\mathrm{d}o$$
$$= \Phi\left(\frac{0.13 + 0.01 - 0.450017}{\sqrt{0.094633}}\right) - \Phi\left(\frac{0.13 - 0.01 - 0.450017}{\sqrt{0.094633}}\right)$$
$$= 0.0150984107$$
$$b_3(o_3) = b_3(0.38) = \int_{0.38-0.01}^{0.38+0.01} p_3(o|\theta_3)\mathrm{d}o$$
$$= \Phi\left(\frac{0.38 + 0.01 - 0.450017}{\sqrt{0.094633}}\right) - \Phi\left(\frac{0.38 - 0.01 - 0.450017}{\sqrt{0.094633}}\right)$$
$$= 0.0252694804$$

$$\alpha_1(1) = b_1(o_1)\pi_1 = 0.0043930188$$
$$\alpha_1(2) = b_2(o_1)\pi_2 = 0.0027946138$$
$$\alpha_1(3) = b_3(o_1)\pi_3 = 0.0033689782$$
$$\alpha_2(1) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i1}\right) b_1(o_2) = 0.0000441519$$
$$\alpha_2(2) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i2}\right) b_2(o_2) = 0.0000501009$$
$$\alpha_2(3) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i3}\right) b_3(o_2) = 0.0000585924$$

ISSN: 0075-5222

$$\alpha_3(1) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i1}\right) b_1(o_3) = 0.0000010644$$
$$\alpha_3(2) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i2}\right) b_2(o_3) = 0.0000012284$$
$$\alpha_3(3) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i3}\right) b_3(o_3) = 0.0000014911$$
$$\beta_3(1) = \beta_3(2) = \beta_3(3) = 1$$
$$\beta_2(1) = \sum_{j=1}^{n} a_{1j}b_j(o_3)\beta_3(j) = 0.0245172558$$
$$\beta_2(2) = \sum_{j=1}^{n} a_{2j}b_j(o_3)\beta_3(j) = 0.0248045456$$
$$\beta_2(3) = \sum_{j=1}^{n} a_{3j}b_j(o_3)\beta_3(j) = 0.0248954357$$
$$\beta_1(1) = \sum_{j=1}^{n} a_{1j}b_j(o_2)\beta_2(j) = 0.0003514276$$
$$\beta_1(2) = \sum_{j=1}^{n} a_{2j}b_j(o_2)\beta_2(j) = 0.0003622263$$
$$\beta_1(3) = \sum_{j=1}^{n} a_{3j}b_j(o_2)\beta_2(j) = 0.0003644390$$

Within the E-step of the second iteration (r=2), the terminating criterion $P(O|\Delta)$ is calculated according to forward-backward procedure (see table I.1.1) as follows:

$$P(O|\Delta) = \alpha_3(1) + \alpha_3(2) + \alpha_3(3) = 0.0000037839$$

Within the E-step of the second iteration (r=2), the joint probabilities $\xi_t(i,j)$ and $\gamma_t(j)$ are calculated based on formula II.2 as follows:

$$\xi_2(1,1) = \alpha_1(1)a_{11}b_1(o_2)\beta_2(1) = 0.0000006226$$
$$\xi_2(1,2) = \alpha_1(1)a_{12}b_2(o_2)\beta_2(2) = 0.0000004624$$
$$\xi_2(1,3) = \alpha_1(1)a_{13}b_3(o_2)\beta_2(3) = 0.0000004589$$
$$\xi_2(2,1) = \alpha_1(2)a_{21}b_1(o_2)\beta_2(1) = 0.0000002308$$
$$\xi_2(2,2) = \alpha_1(2)a_{22}b_2(o_2)\beta_2(2) = 0.0000004469$$
$$\xi_2(2,3) = \alpha_1(2)a_{23}b_3(o_2)\beta_2(3) = 0.0000003346$$
$$\xi_2(3,1) = \alpha_1(3)a_{31}b_1(o_2)\beta_2(1) = 0.0000002291$$
$$\xi_2(3,2) = \alpha_1(3)a_{32}b_2(o_2)\beta_2(2) = 0.0000003334$$
$$\xi_2(3,3) = \alpha_1(3)a_{33}b_3(o_2)\beta_2(3) = 0.0000006653$$
$$\xi_3(1,1) = \alpha_2(1)a_{11}b_1(o_3)\beta_3(1) = 0.0000004632$$
$$\xi_3(1,2) = \alpha_2(1)a_{12}b_2(o_3)\beta_3(2) = 0.0000003093$$
$$\xi_3(1,3) = \alpha_2(1)a_{13}b_3(o_3)\beta_3(3) = 0.0000003100$$
$$\xi_3(2,1) = \alpha_2(2)a_{21}b_1(o_3)\beta_3(1) = 0.0000003062$$

ISSN: 0075-5222

$\xi_3(2,2) = \alpha_2(2)a_{22}b_2(o_3)\beta_3(2) = 0.0000005332$
$\xi_3(2,3) = \alpha_2(2)a_{23}b_3(o_3)\beta_3(3) = 0.0000004032$
$\xi_3(3,1) = \alpha_2(3)a_{31}b_1(o_3)\beta_3(1) = 0.0000002949$
$\xi_3(3,2) = \alpha_2(3)a_{32}b_2(o_3)\beta_3(2) = 0.0000003859$
$\xi_3(3,3) = \alpha_2(3)a_{33}b_3(o_3)\beta_3(3) = 0.0000007778$

$\gamma_1(1) = \alpha_1(1)\beta_1(1) = 0.0000015438$
$\gamma_1(2) = \alpha_1(2)\beta_1(2) = 0.0000010123$
$\gamma_1(3) = \alpha_1(3)\beta_1(3) = 0.0000012278$
$\gamma_2(1) = \alpha_2(1)\beta_2(1) = 0.0000010825$
$\gamma_2(2) = \alpha_2(2)\beta_2(2) = 0.0000012427$
$\gamma_2(3) = \alpha_2(3)\beta_2(3) = 0.0000014587$
$\gamma_3(1) = \alpha_3(1)\beta_3(1) = 0.0000010644$
$\gamma_3(2) = \alpha_3(2)\beta_3(2) = 0.0000012284$
$\gamma_3(3) = \alpha_3(3)\beta_3(3) = 0.0000014911$

Within the M-step of the second iteration (r=2), the estimate $\hat{\Delta}= \left(\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j\right)$ is calculated based on joint probabilities $\xi_t(i,j)$ and $\gamma_t(j)$ determined at E-step.

$$\hat{a}_{11} = \frac{\sum_{t=2}^{3}\xi_t(1,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(1,l)} = 0.413419$$
$$\hat{a}_{12} = \frac{\sum_{t=2}^{3}\xi_t(1,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(1,l)} = 0.293817$$
$$\hat{a}_{13} = \frac{\sum_{t=2}^{3}\xi_t(1,3)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(1,l)} = 0.292764$$
$$\hat{a}_{21} = \frac{\sum_{t=2}^{3}\xi_t(2,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(2,l)} = 0.238147$$
$$\hat{a}_{22} = \frac{\sum_{t=2}^{3}\xi_t(2,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(2,l)} = 0.434668$$
$$\hat{a}_{23} = \frac{\sum_{t=2}^{3}\xi_t(2,3)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(2,l)} = 0.327184$$
$$\hat{a}_{31} = \frac{\sum_{t=2}^{3}\xi_t(3,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(3,l)} = 0.195073$$
$$\hat{a}_{32} = \frac{\sum_{t=2}^{3}\xi_t(3,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(3,l)} = 0.267764$$
$$\hat{a}_{33} = \frac{\sum_{t=2}^{3}\xi_t(3,3)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(3,l)} = 0.537163$$

$$\hat{m}_1 = \frac{\sum_{t=1}^{3}\gamma_t(1)o_t}{\sum_{t=1}^{3}\gamma_t(1)} = 0.515827$$
$$\hat{\sigma}_1^2 = \frac{\sum_{t=1}^{3}\gamma_t(1)(o_t - \hat{m}_1)^2}{\sum_{t=1}^{3}\gamma_t(1)} = 0.104459$$

ISSN: 0075-5222

$$\hat{m}_2 = \frac{\sum_{t=1}^{3}\gamma_t(2)o_t}{\sum_{t=1}^{3}\gamma_t(2)} = 0.436110$$
$$\hat{\sigma}_2^2 = \frac{\sum_{t=1}^{3}\gamma_t(2)(o_t-\hat{m}_2)^2}{\sum_{t=1}^{3}\gamma_t(2)} = 0.091798$$
$$\hat{m}_3 = \frac{\sum_{t=1}^{3}\gamma_t(3)o_t}{\sum_{t=1}^{3}\gamma_t(3)} = 0.439658$$
$$\hat{\sigma}_3^2 = \frac{\sum_{t=1}^{3}\gamma_t(3)(o_t-\hat{m}_3)^2}{\sum_{t=1}^{3}\gamma_t(3)} = 0.091739$$

$$\hat{\pi}_1 = \frac{\gamma_1(1)}{\gamma_1(1)+\gamma_1(2)+\gamma_1(3)} = 0.407999$$
$$\hat{\pi}_2 = \frac{\gamma_1(2)}{\gamma_1(1)+\gamma_1(2)+\gamma_1(3)} = 0.267524$$
$$\hat{\pi}_3 = \frac{\gamma_1(3)}{\gamma_1(1)+\gamma_1(2)+\gamma_1(3)} = 0.324477$$

Table II.4 summarizes HMM parameters resulted from the first iteration and the second iteration of EM algorithm.

Iteration	HMM parameters		
1st	$\hat{a}_{11} = 0.443786$	$\hat{a}_{12} = 0.278330$	$\hat{a}_{13} = 0.277883$
	$\hat{a}_{21} = 0.258587$	$\hat{a}_{22} = 0.422909$	$\hat{a}_{23} = 0.318504$
	$\hat{a}_{31} = 0.212952$	$\hat{a}_{32} = 0.261709$	$\hat{a}_{33} = 0.525339$
	$\hat{m}_1 = 0.493699$	$\hat{\sigma}_1^2 = 0.100098$	
	$\hat{m}_2 = 0.447242$	$\hat{\sigma}_2^2 = 0.095846$	
	$\hat{m}_3 = 0.450017$	$\hat{\sigma}_3^2 = 0.094633$	
	$\hat{\pi}_1 = 0.367053$	$\hat{\pi}_2 = 0.288002$	$\hat{\pi}_3 = 0.344945$
	Terminating criterion $P(O\|\Delta) = 0.0000004341$		
2nd	$\hat{a}_{11} = 0.413419$	$\hat{a}_{12} = 0.293817$	$\hat{a}_{13} = 0.292764$
	$\hat{a}_{21} = 0.238147$	$\hat{a}_{22} = 0.434668$	$\hat{a}_{23} = 0.327184$
	$\hat{a}_{31} = 0.195073$	$\hat{a}_{32} = 0.267764$	$\hat{a}_{33} = 0.537163$
	$\hat{m}_1 = 0.515827$	$\hat{\sigma}_1^2 = 0.104459$	
	$\hat{m}_2 = 0.436110$	$\hat{\sigma}_2^2 = 0.091798$	
	$\hat{m}_3 = 0.439658$	$\hat{\sigma}_3^2 = 0.091739$	
	$\hat{\pi}_1 = 0.407999$	$\hat{\pi}_2 = 0.267524$	$\hat{\pi}_3 = 0.324477$

ISSN: 0075-5222

	Terminating criterion $P(O\|\Delta) = 0.0000037839$

Table II.4. Continuous observation HMM parameters resulted from the first iteration and the second iteration of EM algorithm

As seen in table II.4, the EM algorithm does not converge yet when it produces two different terminating criteria at the first iteration and the second iteration. It is necessary to run more iterations so as to gain the most optimal estimate. Within this example, the EM algorithm converges absolutely after 14 iterations when the criterion $P(O|\Delta)$ approaches to the same value 1 at the 13rd and 14th iterations. Table II.5 shows HMM parameter estimates along with terminating criterion $P(O|\Delta)$ at the 1st, 2nd, 13rd, and 14th iterations of EM algorithm.

Iteration	HMM parameters		
1st	$\hat{a}_{11} = 0.443786$	$\hat{a}_{12} = 0.278330$	$\hat{a}_{13} = 0.277883$
	$\hat{a}_{21} = 0.258587$	$\hat{a}_{22} = 0.422909$	$\hat{a}_{23} = 0.318504$
	$\hat{a}_{31} = 0.212952$	$\hat{a}_{32} = 0.261709$	$\hat{a}_{33} = 0.525339$
	$\hat{m}_1 = 0.493699$	$\hat{\sigma}_1^2 = 0.100098$	
	$\hat{m}_2 = 0.447242$	$\hat{\sigma}_2^2 = 0.095846$	
	$\hat{m}_3 = 0.450017$	$\hat{\sigma}_3^2 = 0.094633$	
	$\hat{\pi}_1 = 0.367053$	$\hat{\pi}_2 = 0.288002$	$\hat{\pi}_3 = 0.344945$
	Terminating criterion $P(O\|\Delta) = 0.0000004341$		
2nd	$\hat{a}_{11} = 0.413419$	$\hat{a}_{12} = 0.293817$	$\hat{a}_{13} = 0.292764$
	$\hat{a}_{21} = 0.238147$	$\hat{a}_{22} = 0.434668$	$\hat{a}_{23} = 0.327184$
	$\hat{a}_{31} = 0.195073$	$\hat{a}_{32} = 0.267764$	$\hat{a}_{33} = 0.537163$
	$\hat{m}_1 = 0.515827$	$\hat{\sigma}_1^2 = 0.104459$	
	$\hat{m}_2 = 0.436110$	$\hat{\sigma}_2^2 = 0.091798$	
	$\hat{m}_3 = 0.439658$	$\hat{\sigma}_3^2 = 0.091739$	
	$\hat{\pi}_1 = 0.407999$	$\hat{\pi}_2 = 0.267524$	$\hat{\pi}_3 = 0.324477$
	Terminating criterion $P(O\|\Delta) = 0.0000037839$		
13rd	$\hat{a}_{11} = 0$	$\hat{a}_{12} = 1$	$\hat{a}_{13} = 0$
	$\hat{a}_{21} = 0$	$\hat{a}_{22} = 0$	$\hat{a}_{23} = 1$
	$\hat{a}_{31} = 0$	$\hat{a}_{32} = 0$	$\hat{a}_{33} = 1$

ISSN: 0075-5222

	$\hat{m}_1 = 0.88$	$\hat{\sigma}_1^2 = 4.2e-09$	
	$\hat{m}_2 = 0.13$	$\hat{\sigma}_2^2 = 1.4e-14$	
	$\hat{m}_3 = 0.38$	$\hat{\sigma}_3^2 = 1.0e-21$	
	$\hat{\pi}_1 = 1$	$\hat{\pi}_2 = 0$	$\hat{\pi}_3 = 0$
	Terminating criterion $P(O\|\Delta) = 1$		
	$\hat{a}_{11} = 0$	$\hat{a}_{12} = 1$	$\hat{a}_{13} = 0$
	$\hat{a}_{21} = 0$	$\hat{a}_{22} = 0$	$\hat{a}_{23} = 1$
	$\hat{a}_{31} = 0$	$\hat{a}_{32} = 0$	$\hat{a}_{33} = 1$
	$\hat{m}_1 = 0.88$	$\hat{\sigma}_1^2 = 4.2e-09$	
14^{th}	$\hat{m}_2 = 0.13$	$\hat{\sigma}_2^2 = 1.4e-14$	
	$\hat{m}_3 = 0.38$	$\hat{\sigma}_3^2 = 1.0e-21$	
	$\hat{\pi}_1 = 1$	$\hat{\pi}_2 = 0$	$\hat{\pi}_3 = 0$
	Terminating criterion $P(O\|\Delta) = 1$		

Table II.5. Continuous observation HMM parameters along with terminating criteria after 14 iterations of EM algorithm

Note that the format like "4.2*e* – 09" indicates scientific notation for real number, namely, 4.2*e* – 09 = 4.2*10^{-9}.

As a result, the learned parameters *A*, *B*, and $\prod$ are shown in table II.6:

		Weather current day (Time point *t*)		
		sunny	*cloudy*	*rainy*
Weather previous day (Time point *t* –1)	*sunny*	a_{11}=0	a_{12}=1	a_{13}=0
	cloudy	a_{21}=0	a_{22}=0	a_{23}=1
	rainy	a_{31}=0	a_{32}=0	a_{33}=1

sunny	*cloudy*	*rainy*
π_1=1	π_2=0	π_3=0

		Humidity		
Weather	*sunny*	$p_1(o_t\|\theta_1)$	m_1 = 0.88	σ_1^2 = 4.2*e* – 09
	cloudy	$p_2(o_t\|\theta_2)$	m_2 = 0.13	σ_1^2 = 1.4*e* –14
	rainy	$p_3(o_t\|\theta_3)$	m_3 = 0.38	σ_1^2 = 1.0*e* –21

ISSN: 0075-5222

Table II.6. Continuous observation HMM parameters of weather example learned from EM algorithm

Such learned parameters are more appropriate to the continuous observation sequence O = {o_1=0.88, o_2=0.13, o_3=0.38} than the original ones shown in tables I.1, I.2, and II.2. In this section, observation conforms to a single PDF. Next section mentions mixture HMM in which observation is characterized by the mixture model of partial PDF (s).

III. Mixture hidden Markov model

Suppose observation o_t is result of experiment that is done at laboratory with K methods or equipments. Each observation is product of K methods or equipments. Given aforementioned weather example, humidity is measured at time point t by K=2 equipments. Equipment 1 and equipment 2 produces two humidity measures a_1=0.63 and a_2=0.88, respectively at time point t. Hence, the observation o_t receives a random value among a_1=0.63 and a_2=0.88. If there is 60% that equipment 1 is selected then, we have o_t=a_1=0.63 with confidence 60%. The percentage 60% is called normalized weight of equipment 1 and so, it is easy to infer that normalized weight of equipment 2 is 40%.

Shortly, observation o_t is created from mixture of K methods or equipments. If the observation o_t is characterized by a PDF as aforementioned in formula II.1, this PDF is constituted of K partial PDF (s) and each partial PDF corresponds to one method or equipment. Such PDF is called *mixture model PDF*, specified by formula III.1 (Rabiner, 1989, p. 267).

$$b_j(o_t) = p_j(o_t|\theta_j) = \sum_{k=1}^{K} c_j^{(k)} p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)$$

Where $c_j^{(k)}$ are non-negative *normalized weights*, $0 \leq c_j^{(k)} \leq 1$ such that

$$\sum_{k=1}^{K} c_j^{(k)} = 1$$

Formula III.1. Mixture model probability density function (PDF) of observation

Each *partial PDF* $p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)$ belongs to any probability distribution, for example, normal distribution, exponential distribution, etc. It is not specified that K partial PDF (s) constructing the formula III.1 must have the same type of distribution but they often share the same distribution type in practice. For example, humidity is measured by K=2 equipments where the first equipment produces values conforming normal distribution with mean 0 and variance 1 while the second equipment produces values conforming normal distribution with mean 0.5 and variance 2. The case that two equipments measuring the same metric like humidity produce values in accordance with two different distributions (for example, normal distribution and exponential distribution) is very rare or

ISSN: 0075-5222

impossible. The notation $\theta_j^{(k)}$ denotes *partial probabilistic parameters*, for instance, if $p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)$ is normal distribution PDF, $\theta_j^{(k)}$ includes mean $m_j^{(k)}$ and variance ${\sigma^2}_j^{(k)}$. The HMM now is specified by parameter $\Delta=\left(a_{ij}, c_j^{(k)}, \theta_j^{(k)}, \pi_j\right)$, which is called *mixture continuous observation HMM*. The main subject of learning problem is to calculate partial estimates $\hat{c}_j^{(k)}$ and $\hat{\theta}_j^{(k)}$ when estimates $\hat{a}_{ij}$ and $\hat{\pi}_j$ are kept intact, as aforementioned in formula I.3.2.7.

$$\hat{a}_{ij}=\frac{\sum_{t=2}^{T}\xi_t(i,j)}{\sum_{t=2}^{T}\sum_{l=1}^{n}\xi_t(i,l)}$$

$$\hat{\pi}_j=\frac{\gamma_1(j)}{\sum_{i=1}^{n}\gamma_1(i)}$$

The joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ according to formulas II.2.

$$\xi_t(i,j)=\alpha_{t-1}(i)a_{ij}\left(\int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j\left(o\middle|\theta_j\right)\mathrm{d}o\right)\beta_t(j) \text{ where } t\geq 2$$

$$\gamma_t(j)=P\left(O,x_t=s_j\middle|\Delta\right)=\alpha_t(j)\beta_t(j)$$

Let $\gamma_t(j, k)$ be joint probability that the stochastic process counting for k^{th} partial PDF is in state s_i at time point t with observation sequence O. Let $\xi_t(i, j, k)$ be the joint probability that the stochastic process counting for k^{th} receives state s_i at time point t–1 and state s_j at time point t given observation sequence O. Formula III.2 specifies partial $\gamma_t(j, k)$ and $\xi_t(i, j, k)$.

$$\begin{aligned}\xi_t(i,j,k)&=P^{(k)}\left(O,x_{t-1}=s_i,x_t=s_j\middle|\Delta\right)\\&=\alpha_{t-1}(i)a_{ij}\left(\int_{o_t-\varepsilon}^{o_t+\varepsilon} c_j^{(k)}p_j^{(k)}\left(o\middle|\theta_j^{(k)}\right)\mathrm{d}o\right)\beta_t(j) \text{ where } t\geq 2\end{aligned}$$

$$\gamma_t(j,k)=P^{(k)}\left(O,x_t=s_j\middle|\Delta\right)=\alpha_t(j,k)\beta_t(j,k)$$

Where partial forward variable $\alpha_{t+1}(j, k)$ and partial backward variable $\beta_t(i, k)$ count for k^{th} partial PDF.

$$\alpha_{t+1}(j,k)=\left(\sum_{i=1}^{n}\alpha_t(i,k)a_{ij}\right)\int_{o_{t+1}-\varepsilon}^{o_{t+1}+\varepsilon} c_j^{(k)}p_j^{(k)}\left(o\middle|\theta_j^{(k)}\right)\mathrm{d}o$$

$$\beta_t(i,k)=\sum_{j=1}^{n}a_{ij}\left(\int_{o_{t+1}-\varepsilon}^{o_{t+1}+\varepsilon} c_j^{(k)}p_j^{(k)}\left(o\middle|\theta_j^{(k)}\right)\mathrm{d}o\right)\beta_{t+1}(j,k)$$

ISSN: 0075-5222

$$\int_{o_t-\varepsilon}^{o_t+\varepsilon} c_j^{(k)} p_j^{(k)}\left(o\middle|\theta_j^{(k)}\right)\mathrm{d}o$$

$$= c_j^{(k)}\Phi\left(\frac{o_t+\varepsilon-m_j^{(k)}}{\sqrt{{\sigma^2}_j^{(k)}}}\right)$$

$$- c_j^{(k)}\Phi\left(\frac{o_t-\varepsilon-m_j^{(k)}}{\sqrt{{\sigma^2}_j^{(k)}}}\right) \text{ if } p_j^{(k)}\left(o\middle|\theta_j^{(k)}\right) \text{ is normal PDF}$$

Formula III.2. Partial joint probabilities $\xi_t(i, j, k)$ and $\gamma_t(j, k)$ based on mixture model PDF

Note that $m_j^{(k)}$ and ${\sigma^2}_j^{(k)}$ are mean and variance of the normal PDF $p_j^{(k)}\left(o\middle|\theta_j^{(k)}\right)$ and Φ is cumulative standard normal distribution (Montgomery & Runger, 2003, p. 653).

After comparing formula II.2 with formula III.2, it is easy to draw the relationship between quantities $\xi_t(i, j)$, $\gamma_t(j)$ and partial quantities $\xi_t(i, j, k)$, $\gamma_t(j, k)$, as seen in formula III.3.

$$\xi_t(i,j,k) = \frac{c_j^{(k)} p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)}{\sum_{l=1}^{K} c_j^{(l)} p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)}\xi_t(i,j) \text{ where } t \geq 2$$

$$\gamma_t(j,k) = \frac{c_j^{(k)} p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)}{\sum_{l=1}^{K} c_j^{(l)} p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)}\gamma_t(j)$$

Formula III.3. Relationship between quantities $\xi_t(i, j)$, $\gamma_t(j)$ and partial quantities $\xi_t(i, j, k)$, $\gamma_t(j, k)$

The ratio $\frac{c_j^{(k)} p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)}{\sum_{l=1}^{K} c_j^{(l)} p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)}$ reflects how much the partial PDF $p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)$ contributes to the mixture model PDF $p_j\left(o_t\middle|\theta_j\right)$. Following is the proof of formula III.3.

Given $t \geq 2$, we have:

$$\xi_t(i,j,k) = P^{(k)}\left(O, x_{t-1} = s_i, x_t = s_j\middle|\Delta\right)$$

$$= \alpha_{t-1}(i) a_{ij}\left(c_j^{(k)} p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\beta_t(j)$$

(The integral $\int_{o_t-\varepsilon}^{o_t+\varepsilon} c_j^{(k)} p_j^{(k)}\left(o\middle|\theta_j^{(k)}\right)\mathrm{d}o$ shown in formula III.2 is turned back the original PDF $c_j^{(k)} p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)$)

ISSN: 0075-5222

$$= \alpha_{t-1}(i)a_{ij}\left(\frac{c_j^{(k)}p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)}{\sum_{l=1}^{K}c_j^{(l)}p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)}\sum_{l=1}^{K}c_j^{(l)}p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)\right)\beta_t(j)$$

$$= \frac{c_j^{(k)}p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)}{\sum_{l=1}^{K}c_j^{(l)}p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)}\alpha_{t-1}(i)a_{ij}\left(\sum_{l=1}^{K}c_j^{(l)}p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)\right)\beta_t(j)$$

$$= \frac{c_j^{(k)}p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)}{\sum_{l=1}^{K}c_j^{(l)}p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)}\alpha_{t-1}(i)a_{ij}p_j\left(o_t\middle|\theta_j\right)\beta_t(j)$$

(due to $p_j\left(o_t\middle|\theta_j\right) = \sum_{l=1}^{K}c_j^{(l)}p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)$ according to formula III.1)

$$= \frac{c_j^{(k)}p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)}{\sum_{l=1}^{K}c_j^{(l)}p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)}\xi_t(i,j)$$

We also have:

$$\gamma_t(j,k) = P^{(k)}\left(O, x_t = s_j\middle|\Delta\right) = \sum_{i=1}^{n}\xi_t(i,j,k)$$

(due to $\gamma_t(j,k) = \sum_{i=1}^{n}\xi_t(i,j,k)$ according to formula I.3.2.6)

$$= \sum_{i=1}^{n}\frac{c_j^{(k)}p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)}{\sum_{l=1}^{K}c_j^{(l)}p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)}\xi_t(i,j)$$

$$= \frac{c_j^{(k)}p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)}{\sum_{l=1}^{K}c_j^{(l)}p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)}\sum_{i=1}^{n}\xi_t(i,j)$$

$$= \frac{c_j^{(k)}p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)}{\sum_{l=1}^{K}c_j^{(l)}p_j^{(l)}\left(o_t\middle|\theta_j^{(l)}\right)}\gamma_t(j)$$

(due to $\gamma_t(j) = \sum_{i=1}^{n}\xi_t(i,j)$ according to formula I.3.2.6)

Because estimates $\hat{a}_{ij}$ and $\hat{\pi}_j$ are kept intact, it is necessary to calculate partial estimates $\hat{c}_j^{(k)}$ and $\hat{\theta}_j^{(k)}$. Formula III.4 specifies the EM expectation $E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}$ for continuous observation HMM given mixture model PDF.

$$E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}$$

$$= \sum_{X}P(X|O,\Delta_r)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}I(x_{t-1}=s_i, x_t=s_j)ln(a_{ij})\right.$$

$$\left.+\sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{T}I(x_t=s_j,k)ln\left(c_j^{(k)}p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)$$

ISSN: 0075-5222

Formula III.4. EM conditional expectation for continuous observation HMM given mixture model PDF

Where $I(x_t = s_j, k)$ is index function such that,

$$I(x_t = s_j, k) = \begin{cases} 1 \text{ if } x_t = s_j \text{ counting for } k^{th} \text{ partial PDF} \\ 0 \text{ otherwise} \end{cases}$$

Note that notation "*ln*" denotes natural logarithm function.

Derived from formula I.3.2.3, the HMM Lagrangian function for continuous observation HMM given mixture model PDF and constraint $\sum_{k=1}^{K} c_j^{(k)} = 1$ is specified by formula III.5.

$$\begin{aligned} l(\Delta, \lambda, \mu) &= l\left(a_{ij}, c_j^{(k)}, \theta_j^{(k)}, \lambda_i, \mu_j\right) \\ &= E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\} + \sum_{i=1}^{n} \lambda_i \left(1 - \sum_{j=1}^{n} a_{ij}\right) \\ &+ \sum_{j=1}^{n} \mu_j \left(1 - \sum_{k=1}^{K} c_j^{(k)}\right) \end{aligned}$$

Formula III.5. Lagrangian function for continuous observation HMM with single PDF

Where λ is *n*-component vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ and μ is *n*-component vector $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$. Factors $\lambda_i \geq 0$ and $\mu_j \geq 0$ are called Lagrange multipliers or Karush-Kuhn-Tucker multipliers (Wikipedia, Karush–Kuhn–Tucker conditions, 2014) or dual variables.

The estimate $\hat{c}_j^{(k)}$ which is extreme point of the Lagrangian function $l(\Delta, \lambda, \mu)$ is determined by setting partial derivatives of $l(\Delta, \lambda, \mu)$ with respect to $c_j^{(k)}$ to be zero. The partial derivative of $l(\Delta, \lambda)$ with respect to $c_j^{(k)}$ is:

$$\begin{aligned} \frac{\partial l(\Delta,\lambda)}{\partial c_j^{(k)}} &= \frac{\partial E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}}{\partial c_j^{(k)}} + \frac{\partial}{\partial c_j^{(k)}}\left(\sum_{i=1}^{n} \lambda_i \left(1 - \sum_{j=1}^{n} a_{ij}\right)\right) \\ &+ \frac{\partial}{\partial c_j^{(k)}}\left(\sum_{j=1}^{n} \mu_j \left(1 - \sum_{k=1}^{K} c_j^{(k)}\right)\right) \\ &= \frac{\partial E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}}{\partial c_j^{(k)}} - \mu_j \\ &= \frac{\partial}{\partial c_j^{(k)}}\left(\sum_{X} P(X|O,\Delta_r)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} I(x_{t-1} = s_i, x_t = s_j) ln(a_{ij})\right.\right. \\ &\left.\left. + \sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{T} I(x_t = s_j, k) ln\left(c_j^{(k)} p_j^{(k)}\left(o_t \middle| \theta_j^{(k)}\right)\right)\right)\right) - \mu_j \end{aligned}$$

ISSN: 0075-5222

$$= \left(\sum_{X} P(X|O,\Delta_r) \sum_{j=1}^{n} \sum_{k=1}^{K} \sum_{t=1}^{T} I(x_t = s_j, k) \frac{\partial}{\partial c_j^{(k)}} \left(ln\left(c_j^{(k)} p_j^{(k)}\left(o_t \middle| \theta_j^{(k)}\right)\right)\right)\right) - \mu_j$$

$$= \left(\sum_{X} P(X|O,\Delta_r) \sum_{j=1}^{n} \sum_{k=1}^{K} \sum_{t=1}^{T} I(x_t = s_j, k) \frac{\partial \left(ln\left(c_j^{(k)}\right) + ln\left(p_j^{(k)}\left(o_t \middle| \theta_j^{(k)}\right)\right)\right)}{\partial c_j^{(k)}}\right) - \mu_j$$

$$= \sum_{X} P(X|O,\Delta_r) \sum_{k=1}^{K} \sum_{t=1}^{T} I(x_t = s_j, k) \frac{1}{c_j^{(k)}} - \mu_j$$

$$= \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{X} I(x_t = s_j, k) P(X|O,\Delta_r) \frac{1}{c_j^{(k)}} - \mu_j$$

$$= \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{X} I(x_t = s_j, k) P(x_1, \dots, x_t, \dots, x_T|O,\Delta_r) \frac{1}{c_j^{(k)}} - \mu_j$$

$$= \sum_{k=1}^{K} \sum_{t=1}^{T} I(k) P\left(x_t = s_j \middle| O,\Delta_r\right) \frac{1}{c_j^{(k)}} - \mu_j$$

(Due to total probability rule)

Where $I(x_t = s_j, k)$ is index function such that,

$$I(x_t = s_j, k) = \begin{cases} 1 \text{ if } x_t = s_j \text{ counting for } k^{th} \text{ partial PDF} \\ 0 \text{ otherwise} \end{cases}$$

$$I(k) = \begin{cases} 1 \text{ if counting for } k^{th} \text{ partial PDF} \\ 0 \text{ otherwise} \end{cases}$$

It implies

$$\frac{\partial l(\Delta,\lambda,\mu)}{\partial c_j^{(k)}} = \frac{1}{c_j^{(k)}} \sum_{k=1}^{K} \sum_{t=1}^{T} I(k) P\left(x_t = s_j \middle| O,\Delta_r\right) - \mu_j$$

$$= \frac{1}{c_j^{(k)}} \sum_{k=1}^{K} \sum_{t=1}^{T} I(k) \frac{P\left(O, x_t = s_j \middle| \Delta_r\right)}{P(O|\Delta_r)} - \mu_j$$

(Due to multiplication rule)

$$= \frac{1}{c_j^{(k)} P(O|\Delta_r)} \sum_{k=1}^{K} \sum_{t=1}^{T} I(k) P\left(O, x_t = s_j \middle| \Delta_r\right) - \mu_j$$

$$= \frac{1}{c_j^{(k)}} \sum_{t=1}^{T} P^{(k)}\left(O, x_t = s_j \middle| \Delta_r\right) - \mu_j$$

(Where $P^{(k)}\left(O, x_t = s_j \middle| \Delta_r\right)$ is the joint probability of state x_t=s_j and observation sequence O counting for k^{th} PDF)

ISSN: 0075-5222

$$= \frac{1}{c_j^{(k)} P(O|\Delta_r)} \sum_{t=1}^{T} \gamma_t(j,k) - \mu_j$$

(Due to $\gamma_t(j,k) = P^{(k)}(O, x_t = s_j|\Delta_r)$ according to formula II.8)

Setting the partial derivative $\frac{\partial l(\Delta,\lambda)}{\partial c_j^{(k)}}$ to be zero, we get the equation whose solution is estimate $\hat{c}_j^{(k)}$, as follows:

$$\frac{1}{c_j^{(k)} P(O|\Delta_r)} \sum_{t=1}^{T} \gamma_t(j,k) - \mu_j = 0$$

$$\Rightarrow \sum_{t=1}^{T} \gamma_t(j,k) - \mu_j c_j^{(k)} P(O|\Delta_r) = 0$$

Summing the expression $\sum_{t=1}^{T} \gamma_t(j,k) - \mu_j c_j^{(k)} P(O|\Delta_r)$ over K mixture components, we have:

$$\sum_{k=1}^{K} \sum_{t=1}^{T} \gamma_t(j,k) - \sum_{k=1}^{K} \mu_j c_j^{(k)} P(O|\Delta_r) = 0$$

$$\Rightarrow \sum_{k=1}^{K} \sum_{t=1}^{T} \gamma_t(j,k) - \mu_j P(O|\Delta_r) \sum_{k=1}^{K} c_j^{(k)} = 0$$

$$\Rightarrow \sum_{k=1}^{K} \sum_{t=1}^{T} \gamma_t(j,k) - \mu_j P(O|\Delta_r) = 0$$

(Due to constraint $\sum_{k=1}^{K} c_j^{(k)} = 1$)

$$\Rightarrow \mu_j P(O|\Delta_r) = \sum_{k=1}^{K} \sum_{t=1}^{T} \gamma_t(j,k)$$

Substituting $\mu_j P(O|\Delta_r) = \sum_{k=1}^{K} \sum_{t=1}^{T} \gamma_t(j,k)$ into equation $\sum_{t=1}^{T} \gamma_t(j,k) - \mu_j c_j^{(k)} P(O|\Delta_r) = 0$, we have:

$$\sum_{t=1}^{T} \gamma_t(j,k) - c_j^{(k)} \sum_{l=1}^{K} \sum_{t=1}^{T} \gamma_t(j,l) = 0$$

It implies that the weight estimate $\hat{c}_j^{(k)}$ is specified by formula III.6:

$$\hat{c}_j^{(k)} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{l=1}^{K} \sum_{t=1}^{T} \gamma_t(j,l)}$$

Formula III.6. Weight estimate of partial PDF

The partial joint probabilities $\gamma_t(j, k)$ is determined by formula III.2.

The parameter estimate $\hat{\theta}_j^{(k)}$ which is extreme point of the Lagrangian function $l(\Delta, \lambda, \mu)$ is determined by setting partial derivatives of $l(\Delta, \lambda, \mu)$ with respect to $\theta_j^{(k)}$ to be zero. The partial derivative of $l(\Delta, \lambda, \mu)$ with respect to $\theta_j^{(k)}$ is:

ISSN: 0075-5222

$$\frac{\partial l(\Delta,\lambda,\mu)}{\partial \theta_j^{(k)}} = \frac{\partial E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}}{\partial \theta_j^{(k)}} + \frac{\partial}{\partial \theta_j^{(k)}}\left(\sum_{i=1}^{n}\lambda_i\left(1-\sum_{j=1}^{n}a_{ij}\right)\right) + \frac{\partial}{\partial \theta_j^{(k)}}\left(\sum_{j=1}^{n}\mu_j\left(1-\sum_{k=1}^{K}c_j^{(k)}\right)\right)$$

$$= \frac{\partial E_{X|O,\Delta_r}\{ln(P(O,X|\Delta))\}}{\partial \theta_j^{(k)}}$$

$$= \frac{\partial}{\partial \theta_j^{(k)}}\left(\sum_{X}P(X|O,\Delta_r)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}I(x_{t-1}=s_i,x_t=s_j)ln(a_{ij}) + \sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{T}I(x_t=s_j,k)ln\left(c_j^{(k)}p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)\right)$$

$$= \sum_{X}P(X|O,\Delta_r)\sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{T}I(x_t=s_j,k)\frac{\partial}{\partial \theta_j^{(k)}}\left(ln\left(c_j^{(k)}p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)$$

$$= \sum_{X}P(X|O,\Delta_r)\sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{T}I(x_t=s_j,k)\frac{\partial\left(ln\left(c_j^{(k)}\right)+ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)}{\partial \theta_j^{(k)}}$$

$$= \sum_{X}P(X|O,\Delta_r)\sum_{k=1}^{K}\sum_{t=1}^{T}I(x_t=s_j,k)\frac{\partial\left(ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)}{\partial \theta_j^{(k)}}$$

$$= \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{X}I(x_t=s_j,k)P(X|O,\Delta_r)\frac{\partial\left(ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)}{\partial \theta_j^{(k)}}$$

$$= \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{X}I(x_t=s_j,k)P(x_1,\dots,x_t,\dots,x_T|O,\Delta_r)\frac{\partial\left(ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)}{\partial \theta_j^{(k)}}$$

$$= \sum_{k=1}^{K}\sum_{t=1}^{T}I(k)P(x_t=s_j|O,\Delta_r)\frac{\partial\left(ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)}{\partial \theta_j^{(k)}}$$

(Due to total probability rule)

Where $I(x_t=s_j,k)$ is index function such that,

$$I(x_t=s_j,k)=\begin{cases}1 \text{ if } x_t=s_j \text{ counting for } k^{th} \text{ partial PDF}\\ 0 \text{ otherwise}\end{cases}$$

$$I(k)=\begin{cases}1 \text{ if counting for } k^{th} \text{ partial PDF}\\ 0 \text{ otherwise}\end{cases}$$

It implies

ISSN: 0075-5222

$$\frac{\partial l(\Delta,\lambda,\mu)}{\partial \theta_j^{(k)}} = \sum_{k=1}^{K}\sum_{t=1}^{T} I(k)\frac{P(O,x_t=s_j|\Delta_r)}{P(O|\Delta_r)}\frac{\partial\left(ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)}{\partial\theta_j^{(k)}}$$

(Due to multiplication rule)

$$= \frac{1}{P(O|\Delta_r)}\sum_{k=1}^{K}\sum_{t=1}^{T} I(k)P(O,x_t=s_j|\Delta_r)\frac{\partial\left(ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)}{\partial\theta_j^{(k)}}$$

$$= \frac{1}{P(O|\Delta_r)}\sum_{t=1}^{T} P^{(k)}(O,x_t=s_j|\Delta_r)\frac{\partial\left(ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)}{\partial\theta_j^{(k)}}$$

(Where $P^{(k)}(O,x_t=s_j|\Delta_r)$ is the joint probability of state x_t=s_j and observation sequence O counting for k^{th} PDF)

$$= \frac{1}{P(O|\Delta_r)}\sum_{t=1}^{T} \gamma_t(j,k)\frac{\partial\left(ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)\right)}{\partial\theta_j^{(k)}}$$

(Due to $\gamma_t(j,k) = P^{(k)}(O,x_t=s_j|\Delta_r)$ according to formula III.2)

Setting the partial derivative $\frac{\partial l(\Delta,\lambda)}{\partial\theta_j^{(k)}}$ to be zero, we get the equation whose solution is estimate $\hat{\theta}_j^{(k)}$, specified by formula III.7.

$$\frac{1}{P(O|\Delta_r)}\sum_{t=1}^{T}\gamma_t(j,k)\frac{\partial ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)}{\partial\theta_j^{(k)}} = 0 \Leftrightarrow \sum_{t=1}^{T}\gamma_t(j)\frac{\partial ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)}{\partial\theta_j^{(k)}} = 0$$

Formula III.7. Equation of partial PDF parameter

Note that notation "*ln*" denotes natural logarithm function.

The equation specified by formula is similar to the one specified by formula II.5 except that the single PDF $p_j(o_t|\theta_j)$ is replaced by partial PDF $p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)$. It is possible to solve the above equation (formula III.7) by Newton-Raphson method (Burden & Faires, 2011, pp. 67-69) – a numeric analysis method but it is easier and simpler to find out more precise solution if the PDF $p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)$ belongs to well-known distributions: normal distribution (Montgomery & Runger, 2003, pp. 109-110), exponential distribution (Montgomery & Runger, 2003, pp. 122-123), etc. Please see formula II.5 for more details of such specific solutions.

Therefore, without of replication, if $p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)$ is normal PDF, its parameter estimate $\hat{\theta}_j^{(k)} = \left(\hat{m}_j^{(k)}, {\sigma^2}_j^{(k)}\right)$ where $\hat{m}_j$ is mean estimate and $\hat{\sigma}_j^2$ is variance estimate is:

ISSN: 0075-5222

$$\hat{\theta}_j^{(k)} = \left(\hat{m}_j^{(k)} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) o_t}{\sum_{t=1}^{T} \gamma_t(j,k)}, {\sigma^2}_j^{(k)} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \left(o_t - \hat{m}_j^{(k)} \right)^2}{\sum_{t=1}^{T} \gamma_t(j,k)} \right)$$

If $p_j^{(k)}\left(o_t \middle| \theta_j^{(k)}\right)$ is exponential PDF, its parameter estimate $\hat{\theta}_j = \hat{\kappa}_j$ is:

$$\hat{\theta}_j^{(k)} = \hat{\kappa}_j^{(k)} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \gamma_t(j,k) o_t}$$

Shortly, the mixture continuous observation HMM estimate $\hat{\Delta} = \left(\hat{a}_{ij}, \hat{c}_j^{(k)}, \hat{\theta}_j^{(k)}, \hat{\pi}_j\right)$ with mixture PDF given current parameter $\Delta = \left(a_{ij}, c_j^{(k)}, \theta_j^{(k)}, \pi_j\right)$ is specified by formula III.8.

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=2}^{T} \sum_{j=1}^{n} \xi_t(i,j)}$$

$$\hat{\pi}_j = \frac{\gamma_1(j)}{\sum_{i=1}^{n} \gamma_1(i)}$$

$$\hat{c}_j^{(k)} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{l=1}^{K} \sum_{t=1}^{T} \gamma_t(j,l)}$$

$$\hat{\theta}_j \text{ is the solution of } \sum_{t=1}^{T} \gamma_t(j) \frac{\partial ln\left(p_j^{(k)}\left(o_t \middle| \theta_j^{(k)}\right)\right)}{\partial \theta_j^{(k)}} = 0$$

With normal distribution:

$$\hat{\theta}_j^{(k)} = \left(\hat{m}_j^{(k)} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) o_t}{\sum_{t=1}^{T} \gamma_t(j,k)}, {\sigma^2}_j^{(k)} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \left(o_t - \hat{m}_j^{(k)} \right)^2}{\sum_{t=1}^{T} \gamma_t(j,k)} \right)$$

With exponential distribution:

$$\hat{\theta}_j^{(k)} = \hat{\kappa}_j^{(k)} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \gamma_t(j,k) o_t}$$

Formula III.8. Continuous observation HMM parameter estimate with mixture PDF

The joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ follow formulas II.2.

$$\xi_t(i,j) = \alpha_{t-1}(i) a_{ij} \left(\int_{o_t - \varepsilon}^{o_t + \varepsilon} p_j\left(o \middle| \theta_j\right) \mathrm{d}o \right) \beta_t(j) \text{ where } t \geq 2$$

$$\gamma_t(j) = P\left(O, x_t = s_j \middle| \Delta\right) = \alpha_t(j) \beta_t(j)$$

The partial joint probabilities $\gamma_t(j, k)$ is determined by formula III.2.

$$\gamma_t(j,k) = P^{(k)}\left(O, x_t = s_j \middle| \Delta\right) = \alpha_t(j,k) \beta_t(j,k)$$

The EM algorithm applied into learning continuous observation HMM parameter with mixture PDF is described in table III.1.

ISSN: 0075-5222

Starting with initial value for Δ, each iteration in EM algorithm has two steps:

1. *E-step*: Calculating the joint probabilities $\xi_t(i, j)$, $\gamma_t(j)$, and $\gamma_t(j, k)$ according to formulas II.2 and III.2.

$$\xi_t(i,j) = \alpha_{t-1}(i)a_{ij}\left(\int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j(o|\theta_j)\mathrm{d}o\right)\beta_t(j) \text{ where } t \geq 2$$

$$\gamma_t(j) = \alpha_t(j)\beta_t(j)$$

$$\gamma_t(j,k) = \alpha_t(j,k)\beta_t(j,k)$$

2. *M-step*: Calculating the estimate $\hat{\Delta}= \left(\hat{a}_{ij}, \hat{c}_j^{(k)}, \hat{\theta}_j^{(k)}, \hat{\pi}_j\right)$ based on the joint probabilities $\xi_t(i, j)$, $\gamma_t(j)$, and $\gamma_t(j, k)$ determined at E-step, according to formula III.8. The estimate $\hat{\Delta}$ becomes the current parameter for next iteration.

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T}\xi_t(i,j)}{\sum_{t=2}^{T}\sum_{j=1}^{n}\xi_t(i,j)}$$

$$\hat{\pi}_j = \frac{\gamma_1(j)}{\sum_{i=1}^{n}\gamma_1(i)}$$

$$\hat{c}_j^{(k)} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)}{\sum_{l=1}^{K}\sum_{t=1}^{T}\gamma_t(j,l)}$$

$$\hat{\theta}_j \text{ is the solution of } \sum_{t=1}^{T}\gamma_t(j)\frac{\partial ln\left(p_j^{(k)}\left(o_t\middle|\theta_j^{(k)}\right)\right)}{\partial\theta_j^{(k)}} = 0$$

With normal distribution:

$$\hat{\theta}_j^{(k)} = \left(\hat{m}_j^{(k)} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)o_t}{\sum_{t=1}^{T}\gamma_t(j,k)}, {\sigma^2}_j^{(k)} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)\left(o_t-\hat{m}_j^{(k)}\right)^2}{\sum_{t=1}^{T}\gamma_t(j,k)}\right)$$

With exponential distribution:

$$\hat{\theta}_j^{(k)} = \hat{\kappa}_j^{(k)} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)}{\sum_{t=1}^{T}\gamma_t(j,k)o_t}$$

EM algorithm stops when it meets the terminating condition, for example, the difference of current parameter Δ and next parameter $\hat{\Delta}$ is insignificant. It is possible to define a custom terminating condition. The terminating criterion $P(O|\Delta)$ described in table I.3.2.2 is a suggestion.

Table III.1. EM algorithm applied into learning continuous observation HMM parameter with mixture PDF

Going back given HMM Δ whose parameters *A* and ∏ are specified in tables I.1 and I.2, suppose continuous observation sequence is *O* = {o_1=0.88, o_2=0.13, o_3=0.38, the EM algorithm described in table III.1 is applied into calculating the parameter estimate $\hat{\Delta}= \left(\hat{a}_{ij}, \hat{\theta}_j, \hat{\pi}_j\right)$.

	Weather current day (Time point *t*)		
	sunny	*cloudy*	*rainy*

ISSN: 0075-5222

Weather previous day (Time point t –1)	*sunny*	a_{11}=0.50	a_{12}=0.25	a_{13}=0.25
	cloudy	a_{21}=0.30	a_{22}=0.40	a_{23}=0.30
	rainy	a_{31}=0.25	a_{32}=0.25	a_{33}=0.50

sunny	*cloudy*	*rainy*
π_1=0.33	π_2=0.33	π_3=0.33

Recall that observations are relative humidity measures. The bias for all measures is ε=0.01, for example, the first observation o_1=0.88 ranges in interval [0.88–0.01, 0.88+0.01]. Humidity is measured by K=2 equipments. The normalized weights of equipments 1 and 2 are 60% and 40%, respectively. Shortly, observation o_t is created from mixture of K=2 equipments. The observation probability distribution B is now includes three normal mixture PDF (s): $p_1(o_t|\theta_1)$, $p_2(o_t|\theta_2)$, and $p_3(o_t|\theta_3)$ corresponding to three states: s_1=*sunny*, s_2=*cloudy*, and s_3=*rainy*. Following is specification of these normal PDF (s).

$$p_1(o_t|\theta_1) = \sum_{k=1}^{2} c_1^{(k)} p_1^{(k)}\left(o_t\middle|\theta_1^{(k)}\right)$$

$$p_2(o_t|\theta_2) = \sum_{k=1}^{2} c_2^{(k)} p_2^{(k)}\left(o_t\middle|\theta_2^{(k)}\right)$$

$$p_3(o_t|\theta_3) = \sum_{k=1}^{2} c_3^{(k)} p_3^{(k)}\left(o_t\middle|\theta_3^{(k)}\right)$$

Note that $c_1^{(k)}$, $c_2^{(k)}$, and $c_3^{(k)}$ are normalized weights; $p_1^{(k)}\left(o\middle|\theta_1^{(k)}\right)$, $p_2^{(k)}\left(o\middle|\theta_2^{(k)}\right)$, and $p_3^{(k)}\left(o\middle|\theta_3^{(k)}\right)$ are normal partial PDF (s).

$$p_1^{(k)}\left(o_t\middle|\theta_1^{(k)}\right) = \frac{1}{\sqrt{2\pi\sigma^2{}_1^{(k)}}} exp\left(-\frac{1}{2}\frac{\left(o_t - m_1^{(k)}\right)^2}{\sigma^2{}_1^{(k)}}\right)$$

$$p_2^{(k)}\left(o_t\middle|\theta_2^{(k)}\right) = \frac{1}{\sqrt{2\pi\sigma^2{}_2^{(k)}}} exp\left(-\frac{1}{2}\frac{\left(o_t - m_2^{(k)}\right)^2}{\sigma^2{}_2^{(k)}}\right)$$

$$p_3^{(k)}\left(o_t\middle|\theta_3^{(k)}\right) = \frac{1}{\sqrt{2\pi\sigma^2{}_3^{(k)}}} exp\left(-\frac{1}{2}\frac{\left(o_t - m_3^{(k)}\right)^2}{\sigma^2{}_3^{(k)}}\right)$$

Where $\theta_1^{(k)} = \left(m_1^{(k)}, \sigma^2{}_1^{(k)}\right)$, $\theta_2^{(k)} = \left(m_2^{(k)}, \sigma^2{}_2^{(k)}\right)$, and $\theta_3^{(k)} = \left(m_3^{(k)}, \sigma^2{}_3^{(k)}\right)$ are means and variances of $p_1^{(k)}\left(o_t\middle|\theta_1^{(k)}\right)$, $p_2^{(k)}\left(o_t\middle|\theta_2^{(k)}\right)$, and $p_3^{(k)}\left(o_t\middle|\theta_3^{(k)}\right)$.

As a convention, normal mixture PDF (s) such as $p_1(o_t|\theta_1)$, $p_2(o_t|\theta_2)$, and $p_3(o_t|\theta_3)$ are represented by theirs weights, partial means and partial variances.

$$\theta_1 = \left(c_1^{(1)}, m_1^{(1)}, \sigma^2{}_1^{(1)}, c_1^{(2)}, m_1^{(2)}, \sigma^2{}_1^{(2)}\right)$$

$$\theta_2 = \left(c_2^{(1)}, m_2^{(1)}, \sigma^2{}_2^{(1)}, c_2^{(2)}, m_2^{(2)}, \sigma^2{}_2^{(2)}\right)$$

ISSN: 0075-5222

$$\theta_3 = \left(c_3^{(1)}, m_3^{(1)}, {\sigma^2}_3^{(1)}, c_3^{(2)}, m_3^{(2)}, {\sigma^2}_3^{(2)}\right)$$

These weights, means and variances are also called observation probability parameters that substitute for discrete matrix *B*. Table III.2 shows observation probability parameters for our weather example.

$p_1(o_t\|\theta_1)$	$c_1^{(1)} = 0.6$	$m_1^{(1)} = 0.87$	${\sigma^2}_1^{(1)} = 1$	$c_1^{(2)} = 0.4$	$m_1^{(2)} = 0.15$	${\sigma^2}_1^{(2)} = 1$
$p_2(o_t\|\theta_2)$	$c_2^{(1)} = 0.6$	$m_2^{(1)} = 0.39$	${\sigma^2}_2^{(1)} = 1$	$c_2^{(2)} = 0.4$	$m_2^{(2)} = 0.89$	${\sigma^2}_2^{(2)} = 1$
$p_3(o_t\|\theta_3)$	$c_3^{(1)} = 0.6$	$m_3^{(1)} = 0.14$	${\sigma^2}_3^{(1)} = 1$	$c_3^{(2)} = 0.4$	$m_3^{(2)} = 0.37$	${\sigma^2}_3^{(2)} = 1$

Table III.2. Observation probability parameters for weather example in case of mixture model

Obviously, we have:

$$p_1(o_t|\theta_1) = \frac{0.6}{\sqrt{2\pi 1}} exp\left(-\frac{1}{2}\frac{(o_t - 0.87)^2}{1}\right) + \frac{0.4}{\sqrt{2\pi 1}} exp\left(-\frac{1}{2}\frac{(o_t - 0.15)^2}{1}\right)$$

$$p_2(o_t|\theta_2) = \frac{0.6}{\sqrt{2\pi 1}} exp\left(-\frac{1}{2}\frac{(o_t - 0.39)^2}{1}\right) + \frac{0.4}{\sqrt{2\pi 1}} exp\left(-\frac{1}{2}\frac{(o_t - 0.89)^2}{1}\right)$$

$$p_3(o_t|\theta_3) = \frac{0.6}{\sqrt{2\pi 1}} exp\left(-\frac{1}{2}\frac{(o_t - 0.14)^2}{1}\right) + \frac{0.4}{\sqrt{2\pi 1}} exp\left(-\frac{1}{2}\frac{(o_t - 0.37)^2}{1}\right)$$

EM algorithm described in table III.1 is applied into calculating the parameter estimate $\hat{\Delta} = \left(\hat{a}_{ij}, \hat{\theta}_j, \hat{\pi}_j\right)$ given continuous observation sequence O = {o_1=0.88, o_2=0.13, o_3=0.38} and mixture normal PDF (s) whose means and variances shown in table III.2. For convenience, all floating-point values are rounded off until ten decimal numbers.

As a convention, let

$$b_j(o_t, k) = \int_{o_t-\varepsilon}^{o_t+\varepsilon} c_j^{(k)} p_j^{(k)}\left(o\middle|\theta_j^{(k)}\right)\mathrm{d}o = c_j^{(k)} \int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j^{(k)}\left(o\middle|\theta_j^{(k)}\right)\mathrm{d}o$$

$$= c_j^{(k)}\Phi\left(\frac{o_t + \varepsilon - m_j^{(k)}}{\sqrt{{\sigma^2}_j^{(k)}}}\right) - c_j^{(k)}\Phi\left(\frac{o_t - \varepsilon - m_j^{(k)}}{\sqrt{{\sigma^2}_j^{(k)}}}\right)$$

$$b_j(o_t) = \int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j(o|\theta_j)\mathrm{d}o = \sum_{k=1}^{2} \int_{o_t-\varepsilon}^{o_t+\varepsilon} c_j^{(k)} p_j^{(k)}\left(o\middle|\theta_j^{(k)}\right)\mathrm{d}o$$

$$= \sum_{k=1}^{2} c_j^{(k)} \int_{o_t-\varepsilon}^{o_t+\varepsilon} p_j^{(k)}\left(o\middle|\theta_j^{(k)}\right)\mathrm{d}o = b_j(o_t, 1) + b_j(o_t, 2)$$

At the first iteration (r=1) we have:

ISSN: 0075-5222

$$b_1(o_1,1) = b_1(0.88,1)$$
$$= c_1^{(1)}\Phi\left(\frac{0.88+0.01-m_1^{(1)}}{\sqrt{\sigma^2{}_1^{(1)}}}\right) - c_1^{(1)}\Phi\left(\frac{0.88-0.01-m_1^{(1)}}{\sqrt{\sigma^2{}_1^{(1)}}}\right)$$
$$= 0.0047869841$$
$$b_1(o_1,2) = b_1(0.88,2)$$
$$= c_1^{(2)}\Phi\left(\frac{0.88+0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right) - c_1^{(2)}\Phi\left(\frac{0.88-0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right)$$
$$= 0.0024449981$$
$$b_1(o_1) = b_1(0.88) = b_1(0.88,1) + b_1(0.88,2) = 0.0072319824$$
$$b_1(o_2,1) = b_1(0.13,1)$$
$$= c_1^{(1)}\Phi\left(\frac{0.13+0.01-m_1^{(1)}}{\sqrt{\sigma^2{}_1^{(1)}}}\right) - c_1^{(1)}\Phi\left(\frac{0.13-0.01-m_1^{(1)}}{\sqrt{\sigma^2{}_1^{(1)}}}\right)$$
$$= 0.0036406445$$
$$b_1(o_2,2) = b_1(0.13,2)$$
$$= c_1^{(2)}\Phi\left(\frac{0.13+0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right) - c_1^{(2)}\Phi\left(\frac{0.13-0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right)$$
$$= 0.0031908476$$
$$b_1(o_2) = b_1(0.13) = b_1(0.13,1) + b_1(0.13,2) = 0.0068314923$$
$$b_1(o_3,1) = b_1(0.38,1)$$
$$= c_1^{(1)}\Phi\left(\frac{0.38+0.01-m_1^{(1)}}{\sqrt{\sigma^2{}_1^{(1)}}}\right) - c_1^{(1)}\Phi\left(\frac{0.38-0.01-m_1^{(1)}}{\sqrt{\sigma^2{}_1^{(1)}}}\right)$$
$$= 0.0042456910$$
$$b_1(o_3,2) = b_1(0.38,2)$$
$$= c_1^{(2)}\Phi\left(\frac{0.38+0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right) - c_1^{(2)}\Phi\left(\frac{0.38-0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right)$$
$$= 0.0031081766$$
$$b_1(o_3) = b_1(0.38) = b_1(0.38,1) + b_1(0.38,2) = 0.0073538674$$

ISSN: 0075-5222

$$b_2(o_1,1) = b_2(0.88,1)$$
$$= c_2^{(1)}\Phi\left(\frac{0.88+0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right) - c_2^{(1)}\Phi\left(\frac{0.88-0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right)$$
$$= 0.0035380756$$
$$b_2(o_1,2) = b_2(0.88,2)$$
$$= c_2^{(2)}\Phi\left(\frac{0.88+0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right) - c_2^{(2)}\Phi\left(\frac{0.88-0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right)$$
$$= 0.0039891531$$
$$b_2(o_1) = b_2(0.88) = b_2(0.88,1) + b_2(0.88,2) = 0.0075272284$$
$$b_2(o_2,1) = b_2(0.13,1)$$
$$= c_2^{(1)}\Phi\left(\frac{0.13+0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right) - c_2^{(1)}\Phi\left(\frac{0.13-0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right)$$
$$= 0.0038567744$$
$$b_2(o_2,2) = b_2(0.13,2)$$
$$= c_2^{(2)}\Phi\left(\frac{0.13+0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right) - c_2^{(2)}\Phi\left(\frac{0.13-0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right)$$
$$= 0.0029887035$$
$$b_2(o_2) = b_2(0.13) = b_2(0.13,1) + b_2(0.13,2) = 0.0068454780$$
$$b_2(o_3,1) = b_2(0.38,1)$$
$$= c_2^{(1)}\Phi\left(\frac{0.38+0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right) - c_2^{(1)}\Phi\left(\frac{0.38-0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right)$$
$$= 0.0039891531$$
$$b_2(o_3,2) = b_2(0.38,2)$$
$$= c_2^{(2)}\Phi\left(\frac{0.38+0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right) - c_2^{(2)}\Phi\left(\frac{0.38-0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right)$$
$$= 0.0035028723$$
$$b_2(o_3) = b_2(0.38) = b_2(0.38,1) + b_2(0.38,2) = 0.0074920254$$

ISSN: 0075-5222

$$b_3(o_1,1) = b_3(0.88,1)$$
$$= c_3^{(1)}\Phi\left(\frac{0.88+0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right) - c_3^{(1)}\Phi\left(\frac{0.88-0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right)$$
$$= 0.0024270937$$
$$b_3(o_1,2) = b_3(0.88,2)$$
$$= c_3^{(2)}\Phi\left(\frac{0.88+0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right) - c_3^{(2)}\Phi\left(\frac{0.88-0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right)$$
$$= 0.0042034471$$
$$b_3(o_1) = b_3(0.88) = b_3(0.88,1) + b_3(0.88,2) = 0.0066305408$$
$$b_3(o_2,1) = b_3(0.13,1)$$
$$= c_3^{(1)}\Phi\left(\frac{0.13+0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right) - c_3^{(1)}\Phi\left(\frac{0.13-0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right)$$
$$= 0.0031913263$$
$$b_3(o_2,2) = b_3(0.13,2)$$
$$= c_3^{(2)}\Phi\left(\frac{0.13+0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right) - c_3^{(2)}\Phi\left(\frac{0.13-0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right)$$
$$= 0.0046513272$$
$$b_3(o_2) = b_3(0.13) = b_3(0.13,1) + b_3(0.13,2) = 0.0078426534$$
$$b_3(o_3,1) = b_3(0.38,1)$$
$$= c_3^{(1)}\Phi\left(\frac{0.38+0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right) - c_3^{(1)}\Phi\left(\frac{0.38-0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right)$$
$$= 0.0031008814$$
$$b_3(o_3,2) = b_3(0.38,2)$$
$$= c_3^{(2)}\Phi\left(\frac{0.38+0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right) - c_3^{(2)}\Phi\left(\frac{0.38-0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right)$$
$$= 0.0047869841$$
$$b_3(o_3) = b_3(0.38) = b_3(0.38,1) + b_3(0.38,2) = 0.0078878654$$

$$\alpha_1(1,1) = b_1(o_1,1)\pi_1 = 0.0015797048$$
$$\alpha_1(1,2) = b_1(o_1,2)\pi_1 = 0.0008068494$$
$$\alpha_1(1) = b_1(o_1)\pi_1 = 0.0023865544$$
$$\alpha_1(2,1) = b_2(o_1,1)\pi_2 = 0.0011675650$$

ISSN: 0075-5222

$$\alpha_1(2,2) = b_2(o_1,2)\pi_2 = 0.0013164206$$
$$\alpha_1(2) = b_2(o_1)\pi_2 = 0.0024839854$$
$$\alpha_1(3,1) = b_3(o_1,1)\pi_3 = 0.0008009410$$
$$\alpha_1(3,2) = b_3(o_1,2)\pi_3 = 0.0013871376$$
$$\alpha_1(3) = b_3(o_1)\pi_3 = 0.0021880786$$
$$\alpha_2(1,1) = \left(\sum_{i=1}^{3} \alpha_1(i,1)a_{i1}\right) b_1(o_2,1) = 0.0000048798$$
$$\alpha_2(1,2) = \left(\sum_{i=1}^{3} \alpha_1(i,2)a_{i1}\right) b_1(o_2,2) = 0.0000036540$$
$$\alpha_2(1) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i1}\right) b_1(o_2) = 0.0000169796$$
$$\alpha_2(2,1) = \left(\sum_{i=1}^{3} \alpha_1(i,1)a_{i2}\right) b_2(o_2,1) = 0.0000040966$$
$$\alpha_2(2,2) = \left(\sum_{i=1}^{3} \alpha_1(i,2)a_{i2}\right) b_2(o_2,2) = 0.0000032131$$
$$\alpha_2(2) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i2}\right) b_2(o_2) = 0.0000146305$$
$$\alpha_2(3,1) = \left(\sum_{i=1}^{3} \alpha_1(i,1)a_{i3}\right) b_3(o_2,1) = 0.0000036562$$
$$\alpha_2(3,2) = \left(\sum_{i=1}^{3} \alpha_1(i,2)a_{i3}\right) b_3(o_2,2) = 0.0000060012$$
$$\alpha_2(3) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i3}\right) b_3(o_2) = 0.0000191037$$
$$\alpha_3(1,1) = \left(\sum_{i=1}^{3} \alpha_2(i,1)a_{i1}\right) b_1(o_3,1) = 0.0000000195$$
$$\alpha_3(1,2) = \left(\sum_{i=1}^{3} \alpha_2(i,2)a_{i1}\right) b_1(o_3,2) = 0.0000000133$$
$$\alpha_3(1) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i1}\right) b_1(o_3) = 0.0000001298$$
$$\alpha_3(2,1) = \left(\sum_{i=1}^{3} \alpha_2(i,1)a_{i2}\right) b_2(o_3,1) = 0.0000000150$$
$$\alpha_3(2,2) = \left(\sum_{i=1}^{3} \alpha_2(i,2)a_{i2}\right) b_2(o_3,2) = 0.0000000130$$

ISSN: 0075-5222

$$\alpha_3(2) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i2}\right) b_2(o_3) = 0.0000001114$$

$$\alpha_3(3,1) = \left(\sum_{i=1}^{3} \alpha_2(i,1)a_{i3}\right) b_3(o_3,1) = 0.0000000133$$

$$\alpha_3(3,2) = \left(\sum_{i=1}^{3} \alpha_2(i,2)a_{i3}\right) b_3(o_3,2) = 0.0000000234$$

$$\alpha_3(3) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i3}\right) b_3(o_3) = 0.0000001434$$

$$\beta_3(1,1) = \beta_3(2,1) = \beta_3(3,1) = 1$$

$$\beta_3(1,2) = \beta_3(2,2) = \beta_3(3,2) = 1$$

$$\beta_3(1) = \beta_3(2) = \beta_3(3) = 1$$

$$\beta_2(1,1) = \sum_{j=1}^{n} a_{1j}b_j(o_3,1)\beta_3(j,1) = 0.0038953541$$

$$\beta_2(1,2) = \sum_{j=1}^{n} a_{1j}b_j(o_3,2)\beta_3(j,2) = 0.0036265524$$

$$\beta_2(1) = \sum_{j=1}^{n} a_{1j}b_j(o_3)\beta_3(j) = 0.0075219064$$

$$\beta_2(2,1) = \sum_{j=1}^{n} a_{2j}b_j(o_3,1)\beta_3(j,1) = 0.0037996331$$

$$\beta_2(2,2) = \sum_{j=1}^{n} a_{2j}b_j(o_3,2)\beta_3(j,2) = 0.0037696972$$

$$\beta_2(2) = \sum_{j=1}^{n} a_{2j}b_j(o_3)\beta_3(j) = 0.0075693303$$

$$\beta_2(3,1) = \sum_{j=1}^{n} a_{3j}b_j(o_3,1)\beta_3(j,1) = 0.0036091517$$

$$\beta_2(3,2) = \sum_{j=1}^{n} a_{3j}b_j(o_3,2)\beta_3(j,2) = 0.0040462543$$

$$\beta_2(3) = \sum_{j=1}^{n} a_{3j}b_j(o_3)\beta_3(j) = 0.0076554059$$

$$\beta_1(1,1) = \sum_{j=1}^{n} a_{1j}b_j(o_2,1)\beta_2(j,1) = 0.0000136339$$

ISSN: 0075-5222

$$\beta_1(1,2) = \sum_{j=1}^{n} a_{1j} b_j(o_2, 2)\beta_2(j, 2) = 0.0000133076$$

$$\beta_1(1) = \sum_{j=1}^{n} a_{1j} b_j(o_2)\beta_2(j) = 0.0000536565$$

$$\beta_1(2,1) = \sum_{j=1}^{n} a_{2j} b_j(o_2, 1)\beta_2(j, 1) = 0.0000135716$$

$$\beta_1(2,2) = \sum_{j=1}^{n} a_{2j} b_j(o_2, 2)\beta_2(j, 2) = 0.0000136243$$

$$\beta_1(2) = \sum_{j=1}^{n} a_{2j} b_j(o_2)\beta_2(j) = 0.0000541536$$

$$\beta_1(3,1) = \sum_{j=1}^{n} a_{3j} b_j(o_2, 1)\beta_2(j, 1) = 0.0000129680$$

$$\beta_1(3,2) = \sum_{j=1}^{n} a_{3j} b_j(o_2, 2)\beta_2(j, 2) = 0.0000151198$$

$$\beta_1(3) = \sum_{j=1}^{n} a_{3j} b_j(o_2)\beta_2(j) = 0.0000558197$$

Within the E-step of the first iteration (r=1), the terminating criterion $P(O|\Delta)$ is calculated according to forward-backward procedure (see table I.1.1) as follows:

$$P(O|\Delta) = \alpha_3(1) + \alpha_3(2) + \alpha_3(3) = 0.0000003847$$

Within the E-step of the first iteration (r=1), the joint probabilities $\xi_t(i,j)$, $\gamma_t(j)$, and $\gamma_t(j,k)$ are calculated based on formulas II.2 and III.2 as follows:

$$\xi_2(1,1) = \alpha_1(1)a_{11}b_1(o_2)\beta_2(1) = 0.0000000613$$
$$\xi_2(1,2) = \alpha_1(1)a_{12}b_2(o_2)\beta_2(2) = 0.0000000309$$
$$\xi_2(1,3) = \alpha_1(1)a_{13}b_3(o_2)\beta_2(3) = 0.0000000358$$
$$\xi_2(2,1) = \alpha_1(2)a_{21}b_1(o_2)\beta_2(1) = 0.0000000383$$
$$\xi_2(2,2) = \alpha_1(2)a_{22}b_2(o_2)\beta_2(2) = 0.0000000515$$
$$\xi_2(2,3) = \alpha_1(2)a_{23}b_3(o_2)\beta_2(3) = 0.0000000447$$
$$\xi_2(3,1) = \alpha_1(3)a_{31}b_1(o_2)\beta_2(1) = 0.0000000281$$
$$\xi_2(3,2) = \alpha_1(3)a_{32}b_2(o_2)\beta_2(2) = 0.0000000283$$
$$\xi_2(3,3) = \alpha_1(3)a_{33}b_3(o_2)\beta_2(3) = 0.0000000657$$
$$\xi_3(1,1) = \alpha_2(1)a_{11}b_1(o_3)\beta_3(1) = 0.0000000624$$
$$\xi_3(1,2) = \alpha_2(1)a_{12}b_2(o_3)\beta_3(2) = 0.0000000318$$
$$\xi_3(1,3) = \alpha_2(1)a_{13}b_3(o_3)\beta_3(3) = 0.0000000335$$
$$\xi_3(2,1) = \alpha_2(2)a_{21}b_1(o_3)\beta_3(1) = 0.0000000323$$
$$\xi_3(2,2) = \alpha_2(2)a_{22}b_2(o_3)\beta_3(2) = 0.0000000438$$
$$\xi_3(2,3) = \alpha_2(2)a_{23}b_3(o_3)\beta_3(3) = 0.0000000346$$
$$\xi_3(3,1) = \alpha_2(3)a_{31}b_1(o_3)\beta_3(1) = 0.0000000351$$
$$\xi_3(3,2) = \alpha_2(3)a_{32}b_2(o_3)\beta_3(2) = 0.0000000358$$

ISSN: 0075-5222

$\xi_3(3,3) = \alpha_2(3)a_{33}b_3(o_3)\beta_3(3) = 0.0000000753$

$\gamma_1(1,1) = \alpha_1(1,1)\beta_1(1,1) = 0.0000000215$
$\gamma_1(1,2) = \alpha_1(1,2)\beta_1(1,2) = 0.0000000107$
$\gamma_1(1) = \alpha_1(1)\beta_1(1) = 0.0000001281$
$\gamma_1(2,1) = \alpha_1(2,1)\beta_1(2,1) = 0.0000000158$
$\gamma_1(2,2) = \alpha_1(2,2)\beta_1(2,2) = 0.0000000179$
$\gamma_1(2) = \alpha_1(2)\beta_1(2) = 0.0000001345$
$\gamma_1(3,1) = \alpha_1(3,1)\beta_1(3,1) = 0.0000000104$
$\gamma_1(3,2) = \alpha_1(3,2)\beta_1(3,2) = 0.0000000210$
$\gamma_1(3) = \alpha_1(3)\beta_1(3) = 0.0000001221$
$\gamma_2(1,1) = \alpha_2(1,1)\beta_2(1,1) = 0.0000000190$
$\gamma_2(1,2) = \alpha_2(1,2)\beta_2(1,2) = 0.0000000133$
$\gamma_2(1) = \alpha_2(1)\beta_2(1) = 0.0000001277$
$\gamma_2(2,1) = \alpha_2(2,1)\beta_2(2,1) = 0.0000000156$
$\gamma_2(2,2) = \alpha_2(2,2)\beta_2(2,2) = 0.0000000121$
$\gamma_2(2) = \alpha_2(2)\beta_2(2) = 0.0000001107$
$\gamma_2(3,1) = \alpha_2(3,1)\beta_2(3,1) = 0.0000000132$
$\gamma_2(3,2) = \alpha_2(3,2)\beta_2(3,2) = 0.0000000243$
$\gamma_2(3) = \alpha_2(3)\beta_2(3) = 0.0000001462$
$\gamma_3(1,1) = \alpha_3(1,1)\beta_3(1,1) = 0.0000000195$
$\gamma_3(1,2) = \alpha_3(1,2)\beta_3(1,2) = 0.0000000133$
$\gamma_3(1) = \alpha_3(1)\beta_3(1) = 0.0000001298$
$\gamma_3(2,1) = \alpha_3(2,1)\beta_3(2,1) = 0.0000000150$
$\gamma_3(2,2) = \alpha_3(2,2)\beta_3(2,2) = 0.0000000130$
$\gamma_3(2) = \alpha_3(2)\beta_3(2) = 0.0000001114$
$\gamma_3(3,1) = \alpha_3(3,1)\beta_3(3,1) = 0.0000000133$
$\gamma_3(3,2) = \alpha_3(3,2)\beta_3(3,2) = 0.0000000234$
$\gamma_3(3) = \alpha_3(3)\beta_3(3) = 0.0000001434$

Within the M-step of the first iteration (r=1), the estimate $\hat{\Delta}= \left(\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j\right)$ is calculated based on joint probabilities $\xi_t(i,j)$ and $\gamma_t(j)$ determined at E-step.

$$\hat{a}_{11} = \frac{\sum_{t=2}^{3}\xi_t(1,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(1,l)} = 0.483829$$
$$\hat{a}_{12} = \frac{\sum_{t=2}^{3}\xi_t(1,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(1,l)} = 0.245210$$
$$\hat{a}_{13} = \frac{\sum_{t=2}^{3}\xi_t(1,3)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(1,l)} = 0.270961$$
$$\hat{a}_{21} = \frac{\sum_{t=2}^{3}\xi_t(2,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(2,l)} = 0.287734$$
$$\hat{a}_{22} = \frac{\sum_{t=2}^{3}\xi_t(2,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(2,l)} = 0.388684$$
$$\hat{a}_{23} = \frac{\sum_{t=2}^{3}\xi_t(2,3)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(2,l)} = 0.323582$$

ISSN: 0075-5222

$$\hat{a}_{31} = \frac{\sum_{t=2}^{3} \xi_t(3,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(3,l)} = 0.235597$$
$$\hat{a}_{32} = \frac{\sum_{t=2}^{3} \xi_t(3,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(3,l)} = 0.238932$$
$$\hat{a}_{33} = \frac{\sum_{t=2}^{3} \xi_t(3,3)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(3,l)} = 0.525471$$

$$\hat{c}_1^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(1,1)}{\sum_{l=1}^{K}\sum_{t=1}^{T} \gamma_t(1,l)} = 0.614137$$
$$\hat{m}_1^{(1)} = \frac{\sum_{t=1}^{3} \gamma_t(1,1) o_t}{\sum_{t=1}^{3} \gamma_t(1,1)} = 0.484616$$
$$\hat{\sigma^2}_1^{(1)} = \frac{\sum_{t=1}^{3} \gamma_t(1,1)\left(o_t - \hat{m}_1^{(1)}\right)^2}{\sum_{t=1}^{3} \gamma_t(1,1)} = 0.100338$$
$$\hat{c}_1^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(1,2)}{\sum_{l=1}^{K}\sum_{t=1}^{T} \gamma_t(1,l)} = 0.385863$$
$$\hat{m}_1^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(1,2) o_t}{\sum_{t=1}^{T} \gamma_t(1,2)} = 0.442128$$
$$\sigma^2{}_1^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(1,2)\left(o_t - \hat{m}_1^{(2)}\right)^2}{\sum_{t=1}^{T} \gamma_t(1,2)} = 0.093132$$
$$\hat{c}_2^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(2,1)}{\sum_{l=1}^{K}\sum_{t=1}^{T} \gamma_t(2,l)} = 0.582310$$
$$\hat{m}_2^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(2,1) o_t}{\sum_{t=1}^{T} \gamma_t(2,1)} = 0.524775$$
$$\sigma^2{}_2^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(2,1)\left(o_t - \hat{m}_2^{(1)}\right)^2}{\sum_{t=1}^{T} \gamma_t(2,1)} = 0.109779$$
$$\hat{c}_2^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(2,2)}{\sum_{l=1}^{K}\sum_{t=1}^{T} \gamma_t(2,l)} = 0.417690$$
$$\hat{m}_2^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(2,2) o_t}{\sum_{t=1}^{T} \gamma_t(2,2)} = 0.566734$$
$$\sigma^2{}_2^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(2,2)\left(o_t - \hat{m}_2^{(2)}\right)^2}{\sum_{t=1}^{T} \gamma_t(2,2)} = 0.108245$$
$$\hat{c}_3^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(3,1)}{\sum_{l=1}^{K}\sum_{t=1}^{T} \gamma_t(3,l)} = 0.465338$$
$$\hat{m}_3^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(3,1) o_t}{\sum_{t=1}^{T} \gamma_t(3,1)} = 0.438194$$
$$\sigma^2{}_3^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(3,1)\left(o_t - \hat{m}_3^{(1)}\right)^2}{\sum_{t=1}^{T} \gamma_t(3,1)} = 0.099469$$

ISSN: 0075-5222

$$\hat{c}_3^{(2)} = \frac{\sum_{t=1}^{T}\gamma_t(3,2)}{\sum_{l=1}^{K}\sum_{t=1}^{T}\gamma_t(3,l)} = 0.534662$$
$$\hat{m}_3^{(2)} = \frac{\sum_{t=1}^{T}\gamma_t(3,2)o_t}{\sum_{t=1}^{T}\gamma_t(3,2)} = 0.448495$$
$$\sigma^{2}{}_3^{(2)} = \frac{\sum_{t=1}^{T}\gamma_t(3,2)\left(o_t - \hat{m}_3^{(2)}\right)^2}{\sum_{t=1}^{T}\gamma_t(3,2)} = 0.101523$$

$$\hat{\pi}_1 = \frac{\gamma_1(1)}{\gamma_1(1)+\gamma_1(2)+\gamma_1(3)} = 0.332860$$
$$\hat{\pi}_2 = \frac{\gamma_1(2)}{\gamma_1(1)+\gamma_1(2)+\gamma_1(3)} = 0.349659$$
$$\hat{\pi}_3 = \frac{\gamma_1(3)}{\gamma_1(1)+\gamma_1(2)+\gamma_1(3)} = 0.317481$$

At the second iteration (r=2), the current parameter $\Delta = (a_{ij}, \theta_j, \pi_j)$ is received values from the previous estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{\theta}_j, \hat{\pi}_j)$, as seen in table III.3.

$a_{11} = \hat{a}_{11} = 0.483829$	$a_{12} = \hat{a}_{12} = 0.245210$	$a_{13} = \hat{a}_{13} = 0.270961$
$a_{21} = \hat{a}_{21} = 0.287734$	$a_{22} = \hat{a}_{22} = 0.388684$	$a_{23} = \hat{a}_{23} = 0.323582$
$a_{31} = \hat{a}_{31} = 0.235597$	$a_{32} = \hat{a}_{32} = 0.238932$	$a_{33} = \hat{a}_{33} = 0.525471$
$c_1^{(1)} = \hat{c}_1^{(1)} = 0.614137$	$m_1^{(1)} = \hat{m}_1^{(1)} = 0.484616$	$\sigma^{2}{}_1^{(1)} = \hat{\sigma}^{2}{}_1^{(1)} = 0.100338$
$c_1^{(2)} = \hat{c}_1^{(2)} = 0.385863$	$m_1^{(2)} = \hat{m}_1^{(2)} = 0.442128$	$\sigma^{2}{}_1^{(2)} = \hat{\sigma}^{2}{}_1^{(2)} = 0.093132$
$c_2^{(1)} = \hat{c}_2^{(1)} = 0.582310$	$m_2^{(1)} = \hat{m}_2^{(1)} = 0.524775$	$\sigma^{2}{}_2^{(1)} = \hat{\sigma}^{2}{}_2^{(1)} = 0.109779$
$c_2^{(2)} = \hat{c}_2^{(2)} = 0.417690$	$m_2^{(2)} = \hat{m}_2^{(2)} = 0.566734$	$\sigma^{2}{}_2^{(2)} = \hat{\sigma}^{2}{}_2^{(2)} = 0.108245$
$c_3^{(1)} = \hat{c}_3^{(1)} = 0.465338$	$m_3^{(1)} = \hat{m}_3^{(1)} = 0.438194$	$\sigma^{2}{}_3^{(1)} = \hat{\sigma}^{2}{}_3^{(1)} = 0.099469$
$c_3^{(2)} = \hat{c}_3^{(2)} = 0.534662$	$m_3^{(2)} = \hat{m}_3^{(2)} = 0.448495$	$\sigma^{2}{}_3^{(2)} = \hat{\sigma}^{2}{}_3^{(2)} = 0.101523$
$\pi_1 = \hat{\pi}_1 = 0.332860$	$\pi_2 = \hat{\pi}_2 = 0.349659$	$\pi_3 = \hat{\pi}_3 = 0.317481$
Terminating criterion $P(O\|\Delta) = 0.0000003847$		

Table III.3. Mixture HMM parameters resulted from the first iteration of EM algorithm

We have:

$$b_1(o_1, 1) = b_1(0.88,1)$$
$$= c_1^{(1)}\Phi\left(\frac{0.88 + 0.01 - m_1^{(1)}}{\sqrt{\sigma^{2}{}_1^{(1)}}}\right) - c_1^{(1)}\Phi\left(\frac{0.88 - 0.01 - m_1^{(1)}}{\sqrt{\sigma^{2}{}_1^{(1)}}}\right)$$
$$= 0.0070989374$$

ISSN: 0075-5222

$$b_1(o_1,2) = b_1(0.88,2)$$
$$= c_1^{(2)}\Phi\left(\frac{0.88+0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right) - c_1^{(2)}\Phi\left(\frac{0.88-0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right)$$
$$= 0.0036046472$$
$$b_1(o_1) = b_1(0.88) = b_1(0.88,1) + b_1(0.88,2) = 0.0107035842$$
$$b_1(o_2,1) = b_1(0.13,1)$$
$$= c_1^{(1)}\Phi\left(\frac{0.13+0.01-m_1^{(1)}}{\sqrt{\sigma^2{}_1^{(1)}}}\right) - c_1^{(1)}\Phi\left(\frac{0.13-0.01-m_1^{(1)}}{\sqrt{\sigma^2{}_1^{(1)}}}\right)$$
$$= 0.0082668914$$
$$b_1(o_2,2) = b_1(0.13,2)$$
$$= c_1^{(2)}\Phi\left(\frac{0.13+0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right) - c_1^{(2)}\Phi\left(\frac{0.13-0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right)$$
$$= 0.0059796064$$
$$b_1(o_2) = b_1(0.13) = b_1(0.13,1) + b_1(0.13,2) = 0.0142464973$$
$$b_1(o_3,1) = b_1(0.38,1)$$
$$= c_1^{(1)}\Phi\left(\frac{0.38+0.01-m_1^{(1)}}{\sqrt{\sigma^2{}_1^{(1)}}}\right) - c_1^{(1)}\Phi\left(\frac{0.38-0.01-m_1^{(1)}}{\sqrt{\sigma^2{}_1^{(1)}}}\right)$$
$$= 0.0146460999$$
$$b_1(o_3,2) = b_1(0.38,2)$$
$$= c_1^{(2)}\Phi\left(\frac{0.38+0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right) - c_1^{(2)}\Phi\left(\frac{0.38-0.01-m_1^{(2)}}{\sqrt{\sigma^2{}_1^{(2)}}}\right)$$
$$= 0.0098798331$$
$$b_1(o_3) = b_1(0.38) = b_1(0.38,1) + b_1(0.38,2) = 0.0245259330$$
$$b_2(o_1,1) = b_2(0.88,1)$$
$$= c_2^{(1)}\Phi\left(\frac{0.88+0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right) - c_2^{(1)}\Phi\left(\frac{0.88-0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right)$$
$$= 0.0078930566$$

ISSN: 0075-5222

$$b_2(o_1,2) = b_2(0.88,2)$$
$$= c_2^{(2)}\Phi\left(\frac{0.88+0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right) - c_2^{(2)}\Phi\left(\frac{0.88-0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right)$$
$$= 0.0064374940$$
$$b_2(o_1) = b_2(0.88) = b_2(0.88,1) + b_2(0.88,2) = 0.0143305510$$
$$b_2(o_2,1) = b_2(0.13,1)$$
$$= c_2^{(1)}\Phi\left(\frac{0.13+0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right) - c_2^{(1)}\Phi\left(\frac{0.13-0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right)$$
$$= 0.0068958909$$
$$b_2(o_2,2) = b_2(0.13,2)$$
$$= c_2^{(2)}\Phi\left(\frac{0.13+0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right) - c_2^{(2)}\Phi\left(\frac{0.13-0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right)$$
$$= 0.0041976823$$
$$b_2(o_2) = b_2(0.13) = b_2(0.13,1) + b_2(0.13,2) = 0.0110935736$$
$$b_2(o_3,1) = b_2(0.38,1)$$
$$= c_2^{(1)}\Phi\left(\frac{0.38+0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right) - c_2^{(1)}\Phi\left(\frac{0.38-0.01-m_2^{(1)}}{\sqrt{\sigma^2{}_2^{(1)}}}\right)$$
$$= 0.0127444649$$
$$b_2(o_3,2) = b_2(0.38,2)$$
$$= c_2^{(2)}\Phi\left(\frac{0.38+0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right) - c_2^{(2)}\Phi\left(\frac{0.38-0.01-m_2^{(2)}}{\sqrt{\sigma^2{}_2^{(2)}}}\right)$$
$$= 0.0086217178$$
$$b_2(o_3) = b_2(0.38) = b_2(0.38,1) + b_2(0.38,2) = 0.0213661827$$
$$b_3(o_1,1) = b_3(0.88,1)$$
$$= c_3^{(1)}\Phi\left(\frac{0.88+0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right) - c_3^{(1)}\Phi\left(\frac{0.88-0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right)$$
$$= 0.0044138478$$

ISSN: 0075-5222

$$b_3(o_1,2) = b_3(0.88,2)$$
$$= c_3^{(2)}\Phi\left(\frac{0.88+0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right) - c_3^{(2)}\Phi\left(\frac{0.88-0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right)$$
$$= 0.0053523076$$
$$b_3(o_1) = b_3(0.88) = b_3(0.88,1) + b_3(0.88,2) = 0.0097661559$$
$$b_3(o_2,1) = b_3(0.13,1)$$
$$= c_3^{(1)}\Phi\left(\frac{0.13+0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right) - c_3^{(1)}\Phi\left(\frac{0.13-0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right)$$
$$= 0.0073030768$$
$$b_3(o_2,2) = b_3(0.13,2)$$
$$= c_3^{(2)}\Phi\left(\frac{0.13+0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right) - c_3^{(2)}\Phi\left(\frac{0.13-0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right)$$
$$= 0.0081239808$$
$$b_3(o_2) = b_3(0.13) = b_3(0.13,1) + b_3(0.13,2) = 0.0154270576$$
$$b_3(o_3,1) = b_3(0.38,1)$$
$$= c_3^{(1)}\Phi\left(\frac{0.38+0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right) - c_3^{(1)}\Phi\left(\frac{0.38-0.01-m_3^{(1)}}{\sqrt{\sigma^2{}_3^{(1)}}}\right)$$
$$= 0.0115717780$$
$$b_3(o_3,2) = b_3(0.38,2)$$
$$= c_3^{(2)}\Phi\left(\frac{0.38+0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right) - c_3^{(2)}\Phi\left(\frac{0.38-0.01-m_3^{(2)}}{\sqrt{\sigma^2{}_3^{(2)}}}\right)$$
$$= 0.0130808186$$
$$b_3(o_3) = b_3(0.38) = b_3(0.38,1) + b_3(0.38,2) = 0.0246525966$$

$$\alpha_1(1,1) = b_1(o_1,1)\pi_1 = 0.0023629516$$
$$\alpha_1(1,2) = b_1(o_1,2)\pi_1 = 0.0011998425$$
$$\alpha_1(1) = b_1(o_1)\pi_1 = 0.0035627941$$
$$\alpha_1(2,1) = b_2(o_1,1)\pi_2 = 0.0027598760$$
$$\alpha_1(2,2) = b_2(o_1,2)\pi_2 = 0.0022509256$$
$$\alpha_1(2) = b_2(o_1)\pi_2 = 0.0050108018$$
$$\alpha_1(3,1) = b_3(o_1,1)\pi_3 = 0.0014013147$$
$$\alpha_1(3,2) = b_3(o_1,2)\pi_3 = 0.0016992582$$
$$\alpha_1(3) = b_3(o_1)\pi_3 = 0.0031005731$$

ISSN: 0075-5222

$$\alpha_2(1,1) = \left(\sum_{i=1}^{3} \alpha_1(i,1)a_{i1}\right) b_1(o_2,1) = 0.0000187454$$

$$\alpha_2(1,2) = \left(\sum_{i=1}^{3} \alpha_1(i,2)a_{i1}\right) b_1(o_2,2) = 0.0000097380$$

$$\alpha_2(1) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i1}\right) b_1(o_2) = 0.0000555050$$

$$\alpha_2(2,1) = \left(\sum_{i=1}^{3} \alpha_1(i,1)a_{i2}\right) b_2(o_2,1) = 0.0000137018$$

$$\alpha_2(2,2) = \left(\sum_{i=1}^{3} \alpha_1(i,2)a_{i2}\right) b_2(o_2,2) = 0.0000066118$$

$$\alpha_2(2) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i2}\right) b_2(o_2) = 0.0000395162$$

$$\alpha_2(3,1) = \left(\sum_{i=1}^{3} \alpha_1(i,1)a_{i3}\right) b_3(o_2,1) = 0.0000165755$$

$$\alpha_2(3,2) = \left(\sum_{i=1}^{3} \alpha_1(i,2)a_{i3}\right) b_3(o_2,2) = 0.0000158124$$

$$\alpha_2(3) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i3}\right) b_3(o_2) = 0.0000650412$$

$$\alpha_3(1,1) = \left(\sum_{i=1}^{3} \alpha_2(i,1)a_{i1}\right) b_1(o_3,1) = 0.0000002478$$

$$\alpha_3(1,2) = \left(\sum_{i=1}^{3} \alpha_2(i,2)a_{i1}\right) b_1(o_3,2) = 0.0000001022$$

$$\alpha_3(1) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i1}\right) b_1(o_3) = 0.0000013133$$

$$\alpha_3(2,1) = \left(\sum_{i=1}^{3} \alpha_2(i,1)a_{i2}\right) b_2(o_3,1) = 0.0000001769$$

$$\alpha_3(2,2) = \left(\sum_{i=1}^{3} \alpha_2(i,2)a_{i2}\right) b_2(o_3,2) = 0.0000000753$$

$$\alpha_3(2) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i2}\right) b_2(o_3) = 0.0000009510$$

ISSN: 0075-5222

$$\alpha_3(3,1) = \left(\sum_{i=1}^{3} \alpha_2(i,1)a_{i3}\right) b_3(o_3,1) = 0.0000002109$$

$$\alpha_3(3,2) = \left(\sum_{i=1}^{3} \alpha_2(i,2)a_{i3}\right) b_3(o_3,2) = 0.0000001712$$

$$\alpha_3(3) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i3}\right) b_3(o_3) = 0.0000015286$$

$$\beta_3(1,1) = \beta_3(2,1) = \beta_3(3,1) = 1$$

$$\beta_3(1,2) = \beta_3(2,2) = \beta_3(3,2) = 1$$

$$\beta_3(1) = \beta_3(2) = \beta_3(3) = 1$$

$$\beta_2(1,1) = \sum_{j=1}^{n} a_{1j}b_j(o_3,1)\beta_3(j,1) = 0.0133467782$$

$$\beta_2(1,2) = \sum_{j=1}^{n} a_{1j}b_j(o_3,2)\beta_3(j,2) = 0.0104386732$$

$$\beta_2(1) = \sum_{j=1}^{n} a_{1j}b_j(o_3)\beta_3(j) = 0.0237854524$$

$$\beta_2(2,1) = \sum_{j=1}^{n} a_{2j}b_j(o_3,1)\beta_3(j,1) = 0.0129121714$$

$$\beta_2(2,2) = \sum_{j=1}^{n} a_{2j}b_j(o_3,2)\beta_3(j,2) = 0.0104266040$$

$$\beta_2(2) = \sum_{j=1}^{n} a_{2j}b_j(o_3)\beta_3(j) = 0.0233387756$$

$$\beta_2(3,1) = \sum_{j=1}^{n} a_{3j}b_j(o_3,1)\beta_3(j,1) = 0.0125762706$$

$$\beta_2(3,2) = \sum_{j=1}^{n} a_{3j}b_j(o_3,2)\beta_3(j,2) = 0.0112612559$$

$$\beta_2(3) = \sum_{j=1}^{n} a_{3j}b_j(o_3)\beta_3(j) = 0.0238375263$$

$$\beta_1(1,1) = \sum_{j=1}^{n} a_{1j}b_j(o_2,1)\beta_2(j,1) = 0.0001001042$$

$$\beta_1(1,2) = \sum_{j=1}^{n} a_{1j}b_j(o_2,2)\beta_2(j,2) = 0.0000657217$$

ISSN: 0075-5222

$$\beta_1(1) = \sum_{j=1}^{n} a_{1j} b_j(o_2) \beta_2(j) = 0.0003270814$$
$$\beta_1(2,1) = \sum_{j=1}^{n} a_{2j} b_j(o_2, 1) \beta_2(j, 1) = 0.0000960759$$
$$\beta_1(2,2) = \sum_{j=1}^{n} a_{2j} b_j(o_2, 2) \beta_2(j, 2) = 0.0000645752$$
$$\beta_1(2) = \sum_{j=1}^{n} a_{2j} b_j(o_2) \beta_2(j) = 0.0003171307$$
$$\beta_1(3,1) = \sum_{j=1}^{n} a_{3j} b_j(o_2, 1) \beta_2(j, 1) = 0.0000955318$$
$$\beta_1(3,2) = \sum_{j=1}^{n} a_{3j} b_j(o_2, 2) \beta_2(j, 2) = 0.0000732366$$
$$\beta_1(3) = \sum_{j=1}^{n} a_{3j} b_j(o_2) \beta_2(j) = 0.0003349345$$

Within the E-step of the second iteration (r=2), the terminating criterion $P(O|\Delta)$ is calculated according to forward-backward procedure (see table I.1.1) as follows:

$$P(O|\Delta) = \alpha_3(1) + \alpha_3(2) + \alpha_3(3) = 0.0000037929$$

Within the E-step of the second iteration (r=2), the joint probabilities $\xi_t(i,j)$, $\gamma_t(j)$, and $\gamma_t(j,k)$ are calculated based on formulas II.2 and III.2 as follows:

$$\xi_2(1,1) = \alpha_1(1) a_{11} b_1(o_2) \beta_2(1) = 0.0000005841$$
$$\xi_2(1,2) = \alpha_1(1) a_{12} b_2(o_2) \beta_2(2) = 0.0000002262$$
$$\xi_2(1,3) = \alpha_1(1) a_{13} b_3(o_2) \beta_2(3) = 0.0000003550$$
$$\xi_2(2,1) = \alpha_1(2) a_{21} b_1(o_2) \beta_2(1) = 0.0000004886$$
$$\xi_2(2,2) = \alpha_1(2) a_{22} b_2(o_2) \beta_2(2) = 0.0000005043$$
$$\xi_2(2,3) = \alpha_1(2) a_{23} b_3(o_2) \beta_2(3) = 0.0000005963$$
$$\xi_2(3,1) = \alpha_1(3) a_{31} b_1(o_2) \beta_2(1) = 0.0000002475$$
$$\xi_2(3,2) = \alpha_1(3) a_{32} b_2(o_2) \beta_2(2) = 0.0000001918$$
$$\xi_2(3,3) = \alpha_1(3) a_{33} b_3(o_2) \beta_2(3) = 0.0000005991$$
$$\xi_3(1,1) = \alpha_2(1) a_{11} b_1(o_3) \beta_3(1) = 0.0000006586$$
$$\xi_3(1,2) = \alpha_2(1) a_{12} b_2(o_3) \beta_3(2) = 0.0000002908$$
$$\xi_3(1,3) = \alpha_2(1) a_{13} b_3(o_3) \beta_3(3) = 0.0000003708$$
$$\xi_3(2,1) = \alpha_2(2) a_{21} b_1(o_3) \beta_3(1) = 0.0000002789$$
$$\xi_3(2,2) = \alpha_2(2) a_{22} b_2(o_3) \beta_3(2) = 0.0000003282$$
$$\xi_3(2,3) = \alpha_2(2) a_{23} b_3(o_3) \beta_3(3) = 0.0000003152$$
$$\xi_3(3,1) = \alpha_2(3) a_{31} b_1(o_3) \beta_3(1) = 0.0000003758$$
$$\xi_3(3,2) = \alpha_2(3) a_{32} b_2(o_3) \beta_3(2) = 0.0000003320$$
$$\xi_3(3,3) = \alpha_2(3) a_{33} b_3(o_3) \beta_3(3) = 0.0000008426$$

$$\gamma_1(1,1) = \alpha_1(1,1) \beta_1(1,1) = 0.0000002365$$

ISSN: 0075-5222

$\gamma_1(1,2) = \alpha_1(1,2)\beta_1(1,2) = 0.0000000789$
$\gamma_1(1) = \alpha_1(1)\beta_1(1) = 0.0000011653$
$\gamma_1(2,1) = \alpha_1(2,1)\beta_1(2,1) = 0.0000002652$
$\gamma_1(2,2) = \alpha_1(2,2)\beta_1(2,2) = 0.0000001454$
$\gamma_1(2) = \alpha_1(2)\beta_1(2) = 0.0000015891$
$\gamma_1(3,1) = \alpha_1(3,1)\beta_1(3,1) = 0.0000001339$
$\gamma_1(3,2) = \alpha_1(3,2)\beta_1(3,2) = 0.0000001244$
$\gamma_1(3) = \alpha_1(3)\beta_1(3) = 0.0000010385$
$\gamma_2(1,1) = \alpha_2(1,1)\beta_2(1,1) = 0.0000002502$
$\gamma_2(1,2) = \alpha_2(1,2)\beta_2(1,2) = 0.0000001017$
$\gamma_2(1) = \alpha_2(1)\beta_2(1) = 0.0000013202$
$\gamma_2(2,1) = \alpha_2(2,1)\beta_2(2,1) = 0.0000001769$
$\gamma_2(2,2) = \alpha_2(2,2)\beta_2(2,2) = 0.0000000689$
$\gamma_2(2) = \alpha_2(2)\beta_2(2) = 0.0000009223$
$\gamma_2(3,1) = \alpha_2(3,1)\beta_2(3,1) = 0.0000002085$
$\gamma_2(3,2) = \alpha_2(3,2)\beta_2(3,2) = 0.0000001781$
$\gamma_2(3) = \alpha_2(3)\beta_2(3) = 0.0000015504$
$\gamma_3(1,1) = \alpha_3(1,1)\beta_3(1,1) = 0.0000002478$
$\gamma_3(1,2) = \alpha_3(1,2)\beta_3(1,2) = 0.0000001022$
$\gamma_3(1) = \alpha_3(1)\beta_3(1) = 0.0000013133$
$\gamma_3(2,1) = \alpha_3(2,1)\beta_3(2,1) = 0.0000001769$
$\gamma_3(2,2) = \alpha_3(2,2)\beta_3(2,2) = 0.0000000753$
$\gamma_3(2) = \alpha_3(2)\beta_3(2) = 0.0000009510$
$\gamma_3(3,1) = \alpha_3(3,1)\beta_3(3,1) = 0.0000002109$
$\gamma_3(3,2) = \alpha_3(3,2)\beta_3(3,2) = 0.0000001712$
$\gamma_3(3) = \alpha_3(3)\beta_3(3) = 0.0000015286$

Within the M-step of the second iteration (r=2), the estimate $\hat{\Delta}= \left(\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j\right)$ is calculated based on joint probabilities $\xi_t(i,j)$ and $\gamma_t(j)$ determined at E-step.

$$\hat{a}_{11} = \frac{\sum_{t=2}^{3}\xi_t(1,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(1,l)} = 0.499998$$
$$\hat{a}_{12} = \frac{\sum_{t=2}^{3}\xi_t(1,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(1,l)} = 0.208001$$
$$\hat{a}_{13} = \frac{\sum_{t=2}^{3}\xi_t(1,3)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(1,l)} = 0.292001$$
$$\hat{a}_{21} = \frac{\sum_{t=2}^{3}\xi_t(2,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(2,l)} = 0.305584$$
$$\hat{a}_{22} = \frac{\sum_{t=2}^{3}\xi_t(2,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(2,l)} = 0.331468$$
$$\hat{a}_{23} = \frac{\sum_{t=2}^{3}\xi_t(2,3)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(2,l)} = 0.362948$$
$$\hat{a}_{31} = \frac{\sum_{t=2}^{3}\xi_t(3,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3}\xi_t(3,l)} = 0.240779$$

ISSN: 0075-5222

$$\hat{a}_{32} = \frac{\sum_{t=2}^{3} \xi_t(3,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(3,l)} = 0.202342$$
$$\hat{a}_{33} = \frac{\sum_{t=2}^{3} \xi_t(3,3)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(3,l)} = 0.556879$$

$$\hat{c}_1^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(1,1)}{\sum_{l=1}^{K}\sum_{t=1}^{T} \gamma_t(1,l)} = 0.717523$$
$$\hat{m}_1^{(1)} = \frac{\sum_{t=1}^{3} \gamma_t(1,1) o_t}{\sum_{t=1}^{3} \gamma_t(1,1)} = 0.438209$$
$$\hat{\sigma^2}_1^{(1)} = \frac{\sum_{t=1}^{3} \gamma_t(1,1)\left(o_t - \hat{m}_1^{(1)}\right)^2}{\sum_{t=1}^{3} \gamma_t(1,1)} = 0.091906$$
$$\hat{c}_1^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(1,2)}{\sum_{l=1}^{K}\sum_{t=1}^{T} \gamma_t(1,l)} = 0.282477$$
$$\hat{m}_1^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(1,2) o_t}{\sum_{t=1}^{T} \gamma_t(1,2)} = 0.414942$$
$$\sigma^2{}_1^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(1,2)\left(o_t - \hat{m}_1^{(2)}\right)^2}{\sum_{t=1}^{T} \gamma_t(1,2)} = 0.085791$$
$$\hat{c}_2^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(2,1)}{\sum_{l=1}^{K}\sum_{t=1}^{T} \gamma_t(2,l)} = 0.728952$$
$$\hat{m}_2^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(2,1) o_t}{\sum_{t=1}^{T} \gamma_t(2,1)} = 0.592928$$
$$\sigma^2{}_2^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(2,1)\left(o_t - \hat{m}_2^{(1)}\right)^2}{\sum_{t=1}^{T} \gamma_t(2,1)} = 0.104641$$
$$\hat{c}_2^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(2,2)}{\sum_{l=1}^{K}\sum_{t=1}^{T} \gamma_t(2,l)} = 0.271048$$
$$\hat{m}_2^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(2,2) o_t}{\sum_{t=1}^{T} \gamma_t(2,2)} = 0.623464$$
$$\sigma^2{}_2^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(2,2)\left(o_t - \hat{m}_2^{(2)}\right)^2}{\sum_{t=1}^{T} \gamma_t(2,2)} = 0.100478$$
$$\hat{c}_3^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(3,1)}{\sum_{l=1}^{K}\sum_{t=1}^{T} \gamma_t(3,l)} = 0.675739$$
$$\hat{m}_3^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(3,1) o_t}{\sum_{t=1}^{T} \gamma_t(3,1)} = 0.370764$$
$$\sigma^2{}_3^{(1)} = \frac{\sum_{t=1}^{T} \gamma_t(3,1)\left(o_t - \hat{m}_3^{(1)}\right)^2}{\sum_{t=1}^{T} \gamma_t(3,1)} = 0.071249$$
$$\hat{c}_3^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(3,2)}{\sum_{l=1}^{K}\sum_{t=1}^{T} \gamma_t(3,l)} = 0.324261$$

ISSN: 0075-5222

$$\hat{m}_3^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(3,2) o_t}{\sum_{t=1}^{T} \gamma_t(3,2)} = 0.399837$$

$$\sigma^{2}{}_3^{(2)} = \frac{\sum_{t=1}^{T} \gamma_t(3,2)\left(o_t - \hat{m}_3^{(2)}\right)^2}{\sum_{t=1}^{T} \gamma_t(3,2)} = 0.082774$$

$$\hat{\pi}_1 = \frac{\gamma_1(1)}{\gamma_1(1) + \gamma_1(2) + \gamma_1(3)} = 0.307239$$

$$\hat{\pi}_2 = \frac{\gamma_1(2)}{\gamma_1(1) + \gamma_1(2) + \gamma_1(3)} = 0.418962$$

$$\hat{\pi}_3 = \frac{\gamma_1(3)}{\gamma_1(1) + \gamma_1(2) + \gamma_1(3)} = 0.273799$$

Table III.4 summarizes mixture HMM parameters resulted from the first iteration and the second iteration of EM algorithm.

Iteration	HMM parameters		
1st	$\hat{a}_{11} = 0.483829$	$\hat{a}_{12} = 0.245210$	$\hat{a}_{13} = 0.270961$
	$\hat{a}_{21} = 0.287734$	$\hat{a}_{22} = 0.388684$	$\hat{a}_{23} = 0.323582$
	$\hat{a}_{31} = 0.235597$	$\hat{a}_{32} = 0.238932$	$\hat{a}_{33} = 0.525471$
	$\hat{c}_1^{(1)} = 0.614137$	$\hat{m}_1^{(1)} = 0.484616$	$\hat{\sigma}^{2}{}_1^{(1)} = 0.100338$
	$\hat{c}_1^{(2)} = 0.385863$	$\hat{m}_1^{(2)} = 0.442128$	$\hat{\sigma}^{2}{}_1^{(2)} = 0.093132$
	$\hat{c}_2^{(1)} = 0.582310$	$\hat{m}_2^{(1)} = 0.524775$	$\hat{\sigma}^{2}{}_2^{(1)} = 0.109779$
	$\hat{c}_2^{(2)} = 0.417690$	$\hat{m}_2^{(2)} = 0.566734$	$\hat{\sigma}^{2}{}_2^{(2)} = 0.108245$
	$\hat{c}_3^{(1)} = 0.465338$	$\hat{m}_3^{(1)} = 0.438194$	$\hat{\sigma}^{2}{}_3^{(1)} = 0.099469$
	$\hat{c}_3^{(2)} = 0.534662$	$\hat{m}_3^{(2)} = 0.448495$	$\hat{\sigma}^{2}{}_3^{(2)} = 0.101523$
	$\hat{\pi}_1 = 0.332860$	$\hat{\pi}_2 = 0.349659$	$\hat{\pi}_3 = 0.317481$
	Terminating criterion $P(O\|\Delta) = 0.0000003847$		
2nd	$\hat{a}_{11} = 0.499998$	$\hat{a}_{12} = 0.208001$	$\hat{a}_{13} = 0.292001$
	$\hat{a}_{21} = 0.305584$	$\hat{a}_{22} = 0.331468$	$\hat{a}_{23} = 0.362948$
	$\hat{a}_{31} = 0.240779$	$\hat{a}_{32} = 0.202342$	$\hat{a}_{33} = 0.556879$
	$\hat{c}_1^{(1)} = 0.717523$	$\hat{m}_1^{(1)} = 0.438209$	$\hat{\sigma}^{2}{}_1^{(1)} = 0.091906$
	$\hat{c}_1^{(2)} = 0.282477$	$\hat{m}_1^{(2)} = 0.414942$	$\hat{\sigma}^{2}{}_1^{(2)} = 0.085791$
	$\hat{c}_2^{(1)} = 0.728952$	$\hat{m}_2^{(1)} = 0.592928$	$\hat{\sigma}^{2}{}_2^{(1)} = 0.104641$
	$\hat{c}_2^{(2)} = 0.271048$	$\hat{m}_2^{(2)} = 0.623464$	$\hat{\sigma}^{2}{}_2^{(2)} = 0.100478$
	$\hat{c}_3^{(1)} = 0.675739$	$\hat{m}_3^{(1)} = 0.370764$	$\hat{\sigma}^{2}{}_3^{(1)} = 0.071249$

ISSN: 0075-5222

	$\hat{c}_3^{(2)} = 0.324261$	$\hat{m}_3^{(2)} = 0.399837$	$\hat{\sigma}^2{}_3^{(2)} = 0.082774$
	$\hat{\pi}_1 = 0.307239$	$\hat{\pi}_2 = 0.418962$	$\hat{\pi}_3 = 0.273799$
	Terminating criterion $P(O\|\Delta) = 0.0000037929$		

Table III.4. Mixture HMM parameters resulted from the first iteration and the second iteration of EM algorithm

As seen in table III.4, the EM algorithm does not converge yet when it produces two different terminating criteria at the first iteration and the second iteration. It is necessary to run more iterations so as to gain the most optimal estimate. Within this example, the EM algorithm converges absolutely after 11 iterations when the criterion $P(O|\Delta)$ approaches to value 1 at the 10th and 11st iterations. Table III.5 shows mixture HMM parameter estimates along with terminating criterion $P(O|\Delta)$ at the 1st, 2nd, 10th, and 11st iterations of EM algorithm.

Iteration	HMM parameters		
1st	$\hat{a}_{11} = 0.483829$	$\hat{a}_{12} = 0.245210$	$\hat{a}_{13} = 0.270961$
	$\hat{a}_{21} = 0.287734$	$\hat{a}_{22} = 0.388684$	$\hat{a}_{23} = 0.323582$
	$\hat{a}_{31} = 0.235597$	$\hat{a}_{32} = 0.238932$	$\hat{a}_{33} = 0.525471$
	$\hat{c}_1^{(1)} = 0.614137$	$\hat{m}_1^{(1)} = 0.484616$	$\hat{\sigma}^2{}_1^{(1)} = 0.100338$
	$\hat{c}_1^{(2)} = 0.385863$	$\hat{m}_1^{(2)} = 0.442128$	$\hat{\sigma}^2{}_1^{(2)} = 0.093132$
	$\hat{c}_2^{(1)} = 0.582310$	$\hat{m}_2^{(1)} = 0.524775$	$\hat{\sigma}^2{}_2^{(1)} = 0.109779$
	$\hat{c}_2^{(2)} = 0.417690$	$\hat{m}_2^{(2)} = 0.566734$	$\hat{\sigma}^2{}_2^{(2)} = 0.108245$
	$\hat{c}_3^{(1)} = 0.465338$	$\hat{m}_3^{(1)} = 0.438194$	$\hat{\sigma}^2{}_3^{(1)} = 0.099469$
	$\hat{c}_3^{(2)} = 0.534662$	$\hat{m}_3^{(2)} = 0.448495$	$\hat{\sigma}^2{}_3^{(2)} = 0.101523$
	$\hat{\pi}_1 = 0.332860$	$\hat{\pi}_2 = 0.349659$	$\hat{\pi}_3 = 0.317481$
	Terminating criterion $P(O\|\Delta) = 0.0000003847$		
2nd	$\hat{a}_{11} = 0.499998$	$\hat{a}_{12} = 0.208001$	$\hat{a}_{13} = 0.292001$
	$\hat{a}_{21} = 0.305584$	$\hat{a}_{22} = 0.331468$	$\hat{a}_{23} = 0.362948$
	$\hat{a}_{31} = 0.240779$	$\hat{a}_{32} = 0.202342$	$\hat{a}_{33} = 0.556879$
	$\hat{c}_1^{(1)} = 0.717523$	$\hat{m}_1^{(1)} = 0.438209$	$\hat{\sigma}^2{}_1^{(1)} = 0.091906$
	$\hat{c}_1^{(2)} = 0.282477$	$\hat{m}_1^{(2)} = 0.414942$	$\hat{\sigma}^2{}_1^{(2)} = 0.085791$
	$\hat{c}_2^{(1)} = 0.728952$	$\hat{m}_2^{(1)} = 0.592928$	$\hat{\sigma}^2{}_2^{(1)} = 0.104641$
	$\hat{c}_2^{(2)} = 0.271048$	$\hat{m}_2^{(2)} = 0.623464$	$\hat{\sigma}^2{}_2^{(2)} = 0.100478$

ISSN: 0075-5222

	$\hat{c}_3^{(1)} = 0.675739$	$\hat{m}_3^{(1)} = 0.370764$	$\hat{\sigma}^2{}_3^{(1)} = 0.071249$
	$\hat{c}_3^{(2)} = 0.324261$	$\hat{m}_3^{(2)} = 0.399837$	$\hat{\sigma}^2{}_3^{(2)} = 0.082774$
	$\hat{\pi}_1 = 0.307239$	$\hat{\pi}_2 = 0.418962$	$\hat{\pi}_3 = 0.273799$
	Terminating criterion $P(O\|\Delta) = 0.0000037929$		
10^{th}	$\hat{a}_{11} = 0.998665$	$\hat{a}_{12} = 0$	$\hat{a}_{13} = 0.001335$
	$\hat{a}_{21} = 0$	$\hat{a}_{22} = 0$	$\hat{a}_{23} = 1$
	$\hat{a}_{31} = 1$	$\hat{a}_{32} = 0$	$\hat{a}_{33} = 0$
	$\hat{c}_1^{(1)} = 1$	$\hat{m}_1^{(1)} = 0.38$	$\hat{\sigma}^2{}_1^{(1)} = 4.1e-16$
	$\hat{c}_1^{(2)} = 0$	$\hat{m}_1^{(2)} = 0.255083$	$\hat{\sigma}^2{}_1^{(2)} = 0.015675$
	$\hat{c}_2^{(1)} = 1$	$\hat{m}_2^{(1)} = 0.88$	$\hat{\sigma}^2{}_2^{(1)} = 3.3e-33$
	$\hat{c}_2^{(2)} = 0$	$\hat{m}_2^{(2)} = 0.88$	$\hat{\sigma}^2{}_2^{(2)} = 9.0e-10$
	$\hat{c}_3^{(1)} = 1$	$\hat{m}_3^{(1)} = 0.130001$	$\hat{\sigma}^2{}_3^{(1)} = 3.2e-07$
	$\hat{c}_3^{(2)} = 0$	$\hat{m}_3^{(2)} = 0.463332$	$\hat{\sigma}^2{}_3^{(2)} = 0.097222$
	$\hat{\pi}_1 = 0$	$\hat{\pi}_2 = 1$	$\hat{\pi}_3 = 0$
	Terminating criterion $P(O\|\Delta) = 1$		
11^{st}	$\hat{a}_{11} = 0.998665$	$\hat{a}_{12} = 0$	$\hat{a}_{13} = 0.001335$
	$\hat{a}_{21} = 0$	$\hat{a}_{22} = 0$	$\hat{a}_{23} = 1$
	$\hat{a}_{31} = 1$	$\hat{a}_{32} = 0$	$\hat{a}_{33} = 0$
	$\hat{c}_1^{(1)} = 1$	$\hat{m}_1^{(1)} = 0.38$	$\hat{\sigma}^2{}_1^{(1)} = 4.1e-16$
	$\hat{c}_1^{(2)} = 0$	$\hat{m}_1^{(2)} = 0.255083$	$\hat{\sigma}^2{}_1^{(2)} = 0.015675$
	$\hat{c}_2^{(1)} = 1$	$\hat{m}_2^{(1)} = 0.88$	$\hat{\sigma}^2{}_2^{(1)} = 3.3e-33$
	$\hat{c}_2^{(2)} = 0$	$\hat{m}_2^{(2)} = 0.88$	$\hat{\sigma}^2{}_2^{(2)} = 9.0e-10$
	$\hat{c}_3^{(1)} = 1$	$\hat{m}_3^{(1)} = 0.130001$	$\hat{\sigma}^2{}_3^{(1)} = 3.2e-07$
	$\hat{c}_3^{(2)} = 0$	$\hat{m}_3^{(2)} = 0.463332$	$\hat{\sigma}^2{}_3^{(2)} = 0.097222$
	$\hat{\pi}_1 = 0$	$\hat{\pi}_2 = 1$	$\hat{\pi}_3 = 0$
	Terminating criterion $P(O\|\Delta) = 1$		

Table III.5. Mixture HMM parameters along with terminating criteria after 14 iterations of EM algorithm

ISSN: 0075-5222

Note that the format like "4.1e – 16" indicates scientific notation for real number, namely, 4.1e – 16 = 4.1*10^{-16}.

As a result, the learned parameters A, B, and $\prod$ are shown in table III.6:

		Weather current day (Time point t)		
		sunny	*cloudy*	*rainy*
Weather previous day (Time point t –1)	*sunny*	a_{11}=0.998665	a_{12}=0	a_{13}=0.001335
	cloudy	a_{21}=0	a_{22}=0	a_{23}=1
	rainy	a_{31}=1	a_{32}=0	a_{33}=0

sunny	*cloudy*	*rainy*
π_1=0	π_2=1	π_3=0

			Humidity		
Weather	*sunny*	$p_1(o_t\|\theta_1)$	$\hat{c}_1^{(1)} = 1$	$\hat{m}_1^{(1)} = 0.38$	$\hat{\sigma}^2{}_1^{(1)} = 4.1e-16$
			$\hat{c}_1^{(2)} = 0$	$\hat{m}_1^{(2)} = 0.255083$	$\hat{\sigma}^2{}_1^{(2)} = 0.015675$
	cloudy	$p_2(o_t\|\theta_2)$	$\hat{c}_2^{(1)} = 1$	$\hat{m}_2^{(1)} = 0.88$	$\hat{\sigma}^2{}_2^{(1)} = 3.3e-33$
			$\hat{c}_2^{(2)} = 0$	$\hat{m}_2^{(2)} = 0.88$	$\hat{\sigma}^2{}_2^{(2)} = 9.0e-10$
	rainy	$p_3(o_t\|\theta_3)$	$\hat{c}_3^{(1)} = 1$	$\hat{m}_3^{(1)} = 0.130001$	$\hat{\sigma}^2{}_3^{(1)} = 3.2e-07$
			$\hat{c}_3^{(2)} = 0$	$\hat{m}_3^{(2)} = 0.463332$	$\hat{\sigma}^2{}_3^{(2)} = 0.097222$

Table III.6. Mixture HMM parameters of weather example learned from EM algorithm

Such learned parameters are more appropriate to the continuous observation sequence O = {o_1=0.88, o_2=0.13, o_3=0.38} than the original ones shown in tables I.1, I.2, and III.2. This section ends up full description of continuous observation HMM. Next section is the discussion and conclusion.

IV. Discussion and Conclusion

My main contribution is to propose the practical technique to calculate basic quantities such as forward variable α_t, backward variable β_t, and joint probabilities ξ_t, γ_t based on integral of observation PDF within a very small velocity of observation value. These quantities are necessary to evaluation problem, uncovering problem, and learning problem of HMM. However, the most hazardous issue of continuous observation HMM is to determine parameter estimate $\hat{\theta}$ of the observation PDF $p(o_t|\theta)$. In previous sections, it is easy to solve the equation specified by formulas II.5 and III.7 to find out $\hat{\theta}$ if the PDF $p(o_t|\theta)$ belongs to normal/exponential distributions when the natural logarithm function eliminates "exponent form" of these distributions, which in turn leads to take derivatives easier. Otherwise, if such PDF belongs to complicated distributions, a

new hazardous problem is issued when equation II.5 is relevant to complex derivatives as follows:

$$\sum_{t=1}^{T} \gamma_t \frac{\partial ln\big(p(o_t|\theta)\big)}{\partial \theta} = 0$$

The suggestion of using Newton-Raphson method (Burden & Faires, 2011, pp. 67-69) to solve such equation is not optimal solution. Your attention please, we should concern the "exponent form" of normal/exponential distributions. The recognition of "exponent form" is very important. In fact, probabilistic distributions such as normal and exponent belong to exponential family of distribution. We should take advantages of exponential family instead of finding the most general method to solve formula II.5. Therefore, this section focuses on exponential family of observation PDF. Later on, we will know that exponential family ranges over most of common distributions. In other words, solving successfully formula II.5 with regard to exponential family is close to finding the general method for determining parameter estimate $\hat{\theta}$.

According to (Wikipedia, Exponential family, 2016), exponential family refers to a set of probabilistic distributions whose PDF (s) have the same exponential form as follows:

$$f(x|\theta) = h(x)exp\left(\eta(\theta) \cdot T(x) - A\big(\eta(\theta)\big)\right)$$

Where,

- The $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ is vector of n parameters θ_i (s). Note that a scalar known as a number is considered as 1-element vector.
- The $h(x)$ is a function of x, which is called base measure.
- The $\eta(\theta)$ is function of θ, which is a vector. If $\eta(\theta) = \theta$, the PDF is called canonical. The function $\eta(\theta)$ is known as natural parameter η whose dimension may be larger than θ.
- The $T(x)$ is function of x, which is sufficient statistic.
- The notation $\eta \cdot T(x)$ denotes scalar product of $\eta(\theta)$ and $T(x)$. Note that the scalar product of two scalars is multiplication of such two scalars.
- The $A(\theta)$ is log-partition function which is used to normalize PDF. It will be mentioned later on.

All functions $\eta(\theta)$, $T(x)$, and $A(\eta(\theta))$ are known. Formula IV.1 specifies the exponential family PDF with note that function $\eta(\theta)$ is considered as natural parameter η.

$$f(x|\eta) = h(x)exp\big(\eta \cdot T(x) - A(\eta)\big)$$

Where η is natural parameter which is function of θ and so $A(\eta)$ is always specified by function of θ.

Formula IV.1. Exponential family PDF

According to (Wikipedia, Exponential family, 2016), most of common distributions belong to exponential family, for instance: normal, log-normal, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, categorical, Poisson, geometric, inverse Gaussian, von Mises, von Mises-Fisher, Wishart, Inverse

ISSN: 0075-5222

Wishart, binomial with fixed number of trials, multinomial with fixed number of trials, negative binomial with fixed number of failures.

The log-partition function $A(\eta)$ is determined based on a property of PDF that integral of any PDF from negative infinity to positive infinity is 1.

$$\int_x f(x|\theta)dx = \int_x h(x)exp\big(\eta \cdot T(x) - A(\eta)\big)dx$$
$$= exp\big(-A(\eta)\big)\int_x h(x)exp\big(\eta \cdot T(x)\big)dx = 1$$

It implies $A(\eta)$ is determined by formula IV.2 (Jebara, 2015, p. 7).

$$A(\eta) = ln\left(\int_x h(x)exp\big(\eta \cdot T(x)\big)dx\right)$$

Formula IV.2. Log-partition function $A(\eta)$

According to author (Jebara, 2015, p. 7), the first derivative of $A(\eta)$ with regard to η is:

$$\frac{dA(\eta)}{d\eta} = \frac{1}{\int_x h(x)exp\big(\eta \cdot T(x)\big)dx}\frac{d\big(\int_x h(x)exp\big(\eta \cdot T(x)\big)dx\big)}{d\eta}$$
$$= \frac{\int_x h(x)exp\big(\eta \cdot T(x)\big)T(x)dx}{\int_x h(x)exp\big(\eta \cdot T(x)\big)dx}$$
$$= \frac{exp\big(-A(\eta)\big)\int_x h(x)exp\big(\eta \cdot T(x)\big)T(x)dx}{exp\big(-A(\eta)\big)\int_x h(x)exp\big(\eta \cdot T(x)\big)dx}$$
$$= \frac{\int_x h(x)exp\big(\eta \cdot T(x) - A(\eta)\big)T(x)dx}{\int_x h(x)exp\big(\eta \cdot T(x) - A(\eta)\big)dx} = \frac{\int_x f(x)T(x)dx}{\int_x f(x)dx}$$
$$= E(T(x)|\eta)$$

In general, the first derivative of $A(\eta)$ is the expectation of $T(x)$ given exponential family PDF according to formula IV.3 as follows:

$$A'(\eta) = E(T(x)|\eta) = E(T(x)|\theta)$$

Formula IV.3. First derivative of log-partition function $A(\eta)$ with regard to η

Note that the exponential family PDF is specified by formula IV.1 and it is possible to determine the derivative $A'(\eta)$ if $A(\eta)$ is known; for example, function $A(\eta)$ of normal PDF with mean μ and variance σ^2 is $\frac{\mu^2}{2\sigma^2} + ln\sigma$. It is not always to calculate $A'(\eta)$ based on the expectation $E(T(x)|\eta)$.

Let $LnL(\eta)$ be the log-likelihood of exponential family PDF which is natural logarithm of exponential family PDF as seen in formula IV.4.

$$LnL(\eta) = ln\big(f(x|\eta)\big) = ln\left(h(x)exp\big(\eta \cdot T(x) - A(\eta)\big)\right)$$
$$= ln\big(h(x)\big) + \eta \cdot T(x) - A(\eta)$$

Formula IV.4. Log-likelihood function of exponential family PDF

ISSN: 0075-5222

Note that $LnL(\eta)$ is function of parameters η and it is also a vector. The first derivative of $LnL(\eta)$ is specified by formula IV.5.

$$DLnL(\eta) = \frac{dLnL(\eta)}{d\eta} = T(x) - A'(\eta) = T(x) - E(T(x)|\eta)$$

Formula IV.5. Derivative of log-likelihood function of exponential family PDF

The formula II.5 becomes formula IV.6 as follows:

$$\sum_{t=1}^{T} \gamma_t DLnL(\eta|o_t) = 0$$

Formula IV.6. Equation of exponential family PDF parameters

Where o_t is continuous observation, T is the number of observations and γ_t is the joint probability specified by formula I.2.1. As seen in section II, formulas II.5, III.7 and IV.6 are derived from EM algorithm and so they are essentially maximum likelihood equations with additional information of joint probabilities γ_t (s). Please refer to article "Maximum Likelihood from Incomplete Data via the EM Algorithm" by author (Dempster, Laird, & Rubin, 1977, p. 5) for more details about maximum likelihood method and EM algorithm.

For example, we have normal PDF with mean μ and variance σ^2. Its specification includes:

$$h(x) = \frac{1}{\sqrt{2\pi}}$$
$$T(x) = (x, x^2)$$
$$\theta = (\mu, \sigma^2)$$
$$\eta = \left(\eta_1 = \frac{\mu}{\sigma^2}, \eta_2 = -\frac{1}{2\sigma^2}\right)$$
$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2} ln\left|\frac{1}{2\eta_2}\right| = \frac{\mu^2}{2\sigma^2} + ln\sigma$$

According to formula IV.1, the normal PDF is:

$$f(x|\eta) = h(x)exp\big(\eta \cdot T(x) - A(\eta)\big)$$
$$= \frac{1}{\sqrt{2\pi}} exp\left(\left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right) \cdot (x, x^2) - \left(\frac{\mu^2}{2\sigma^2} + ln\sigma\right)\right)$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

We have:

$$E(x^2) = \mu^2 + E((x-\mu)^2) = \mu^2 + \sigma^2$$

The derivative of $A(\eta)$ of normal PDF is a vector as follows:

$$A'(\eta) = E(T(x)|\eta) = \big(E(x), E(x^2)\big) = (\mu, \mu^2 + \sigma^2)$$

We have the same result if taking directly derivative on $A(\eta)$:

ISSN: 0075-5222

$$\frac{dA(\eta)}{d\eta_1} = \frac{d\left(-\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}ln\left|\frac{1}{2\eta_2}\right|\right)}{d\eta_1} = -\frac{\eta_1}{2\eta_2} = \mu$$

$$\frac{dA(\eta)}{d\eta_2} = \frac{d\left(-\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}ln\left|\frac{1}{2\eta_2}\right|\right)}{d\eta_2} = \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} = \mu^2 + \sigma^2$$

The derivative of log-likelihood of normal PDF is:

$$DLnL(\eta) = T(x) - E(T(x)|\eta) = (x, x^2) - (\mu, \mu^2 + \sigma^2)$$
$$= (x - \mu, x^2 - \mu^2 - \sigma^2)$$

Now we solve formula IV.6 given normal observation PDF $p(o_t|\theta)$ with mean μ and variance σ^2. The natural parameter of such PDF is:

$$\eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$$

We have:

$$\sum_{t=1}^{T} \gamma_t DLnL(\eta|o_t) = (0,0) \Leftrightarrow \begin{cases} \sum_{t=1}^{T} \gamma_t(o_t - \mu) = 0 \\ \sum_{t=1}^{T} \gamma_t(o_t^2 - \hat{\mu}^2 - \hat{\sigma}^2) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \hat{\mu} = \frac{\sum_{t=1}^{T} \gamma_t o_t}{\sum_{t=1}^{T} \gamma_t} \\ \hat{\sigma}^2 = \frac{\sum_{t=1}^{T} \gamma_t(o_t^2 - \hat{\mu}^2)}{\sum_{t=1}^{T} \gamma_t} \end{cases}$$

The estimate $\hat{\mu}$ is the same to one in formula II.6 but the estimate $\hat{\sigma}^2$ is different. The previous estimate $\hat{\sigma}^2$ specified by formula II.6 is:

$$\hat{\sigma}^2 - \text{previous} = \frac{\sum_{t=1}^{T} \gamma_t(o_t - \hat{\mu})^2}{\sum_{t=1}^{T} \gamma_t}$$

However, our method is still correct because the expectation of $\hat{\sigma}^2$ is kept intact. In fact, we have:

$$E(\hat{\sigma}^2 - \text{previous}) = E\left(\frac{\sum_{t=1}^{T} \gamma_t(o_t - \hat{\mu})^2}{\sum_{t=1}^{T} \gamma_t}\right) = \frac{\sum_{t=1}^{T} \gamma_t E(o_t - \hat{\mu})^2}{\sum_{t=1}^{T} \gamma_t}$$
$$= \frac{\sum_{t=1}^{T} \gamma_t(E(o_t^2) - 2\hat{\mu}E(o_t) + \hat{\mu}^2)}{\sum_{t=1}^{T} \gamma_t} = \frac{\sum_{t=1}^{T} \gamma_t(E(o_t^2) - 2\hat{\mu}\hat{\mu} + \hat{\mu}^2)}{\sum_{t=1}^{T} \gamma_t}$$
$$= \frac{\sum_{t=1}^{T} \gamma_t(E(o_t^2) - \hat{\mu}^2)}{\sum_{t=1}^{T} \gamma_t} = \frac{\sum_{t=1}^{T} \gamma_t E(o_t^2 - \hat{\mu}^2)}{\sum_{t=1}^{T} \gamma_t}$$
$$= E\left(\frac{\sum_{t=1}^{T} \gamma_t(o_t^2 - \hat{\mu}^2)}{\sum_{t=1}^{T} \gamma_t}\right) = E(\hat{\sigma}^2)$$

In general, it is easy for us to solve estimation equation IV.6 for finding parameter estimates or to know that the equation IV.6 is impossible if the continuous PDF belongs to exponential family because the derivative of log-likelihood specified by formula IV.5 always exists and it is totally possible to calculate the expectation of function $T(x)$.

ISSN: 0075-5222

Now there is a question "how to estimate parameters of a PDF which does not conform to exponential family, for instance, Student's t-distribution, F-distribution, Cauchy distribution (Wikipedia, Exponential family, 2016). In that case, we will solve formula II.5 with subject to each concrete distribution. The research gives an example of estimating t-distribution parameters. Similarly, our main point is to calculate the derivative $DLnL(\theta)$ of log-likelihood function of t-distribution. Formula IV.7 specifies the PDF of t-distribution (Wikipedia, Student's t-distribution, 2016).

$$f(x|v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}$$

Where v is the number of degrees of freedom and Γ is gamma function.

Formula IV.7. Student's t-distribution

Following are gamma function $\Gamma(x)$ and digamma function $\psi(x)$ which is derivative of natural logarithm of gamma function. Trigamma function $\psi_1(x)$ is derivative of digamma function $\psi(x)$. Formula IV.8 specifies $\Gamma(x)$, $\psi(x)$, and $\psi_1(x)$ (Nguyen, Beta Likelihood Estimation and Its Application to Specify Prior Probabilities in Bayesian Network, 2016).

$$\Gamma(x) = \int_0^{+\infty} t^{x-1}exp(-t)dt$$

$$\psi(x) = d\left(ln\left(\Gamma(x)\right)\right) = \frac{\Gamma'(x)}{\Gamma(x)}$$

$$\psi_1(x) = \psi'(x)$$

Formula IV.8. Functions gamma, digamma, trigamma

The t-distribution replaces normal distribution for testing on means of normal population in case of small size samples. It is important distribution to hypothesis testing. When v is parameter of t-distribution, we need to estimate it according to formula II.5. The log-likelihood of t-distribution PDF is:

$$LnL(v) = ln\left(\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}\right)$$

$$= ln\left(\Gamma\left(\frac{v+1}{2}\right)\right) - \frac{1}{2}ln\pi - \frac{1}{2}lnv - ln\left(\Gamma\left(\frac{v}{2}\right)\right)$$

$$-\frac{v+1}{2}ln\left(1+\frac{x^2}{v}\right)$$

The derivative $DLnL(v)$ of log-likelihood function is:

$$DLnL(v) = \frac{x^2-1}{2(x^2+v)} + \frac{1}{2}lnv - \frac{1}{2}ln(x^2+v) + \psi\left(\frac{v+1}{2}\right) - \psi\left(\frac{v}{2}\right)$$

Given the t-distribution PDF of observation, the formula II.5 to find out the parameter estimate $\hat{v}$ becomes:

ISSN: 0075-5222

$$\sum_{t=1}^{T} \gamma_t DLnL(v|o_t) = 0$$

$$\Rightarrow \sum_{t=1}^{T} \gamma_t \left(\frac{o_t^2 - 1}{2(o_t^2 + v)} + \frac{1}{2} lnv - \frac{1}{2} ln(o_t^2 + v) - \psi\left(\frac{v+1}{2}\right) - \psi\left(\frac{v}{2}\right) \right) = 0$$

$$\Rightarrow \left(lnv + 2\psi\left(\frac{v+1}{2}\right) - 2\psi\left(\frac{v}{2}\right) \right) \sum_{t=1}^{T} \gamma_t + \sum_{t=1}^{T} \gamma_t \left(\frac{o_t^2 - 1}{o_t^2 + v} - ln(o_t^2 + v) \right) = 0$$

Shortly, the estimate $\hat{v}$ is solution of equation specified by formula IV.9.

$$g(v) = 0 \text{ where } g(v) = \left(lnv + 2\psi\left(\frac{v+1}{2}\right) - 2\psi\left(\frac{v}{2}\right) \right) \sum_{t=1}^{T} \gamma_t + \sum_{t=1}^{T} \gamma_t \left(\frac{o_t^2 - 1}{o_t^2 + v} - ln(o_t^2 + v) \right)$$

Formula IV.9. Equation of t-distribution PDF parameter

Where o_t is continuous observation, T is the number of observations and γ_t is the joint probability specified by formula I.2.1.

The formula IV.9 is much more complicated than formula IV.6. As aforementioned suggestion, the Newton-Raphson method (Burden & Faires, 2011, pp. 67-69) is used to find solution of given equation $g(v)=0$ along with the tangent of $g(v)$. It starts with an arbitrary value of v_0 as a solution candidate. Suppose the current value is v_k, the next value v_{k+1} is calculated based on following formula:

$$v_{k+1} = v_k - \frac{g(v_k)}{g'(v_k)}$$

The value v_k is solution of $g(v)=0$ if $g(v_k)=0$ which means that $v_{k+1}=v_k$. The most important aspect of Newton-Raphson method is that derivative of $g(v)$ must be determinate within domain of $g(v)$. Fortunately, the derivative $g'(v)$ is always determinate for all $v > 0$.

$$g'(v) = \left(\frac{1}{v} + \psi_1\left(\frac{v+1}{2}\right) - \psi_1\left(\frac{v}{2}\right) \right) \sum_{t=1}^{T} \gamma_t - \sum_{t=1}^{T} \gamma_t \frac{2o_t^2 - 1 + v}{(o_t^2 + v)^2}$$

Suppose there is only one observation $o_1=1$ and the joint probability γ_1 is 1, we have:

$$g(v) = ln\left(\frac{v}{1+v}\right) + 2\psi\left(\frac{v+1}{2}\right) - 2\psi\left(\frac{v}{2}\right)$$

$$g'(v) = \frac{1}{v(1+v)} + \psi_1\left(\frac{v+1}{2}\right) - \psi_1\left(\frac{v}{2}\right)$$

Figure IV.1 shows function $g(v)$ with $o_1=1$ and $\gamma_1=1$.

ISSN: 0075-5222

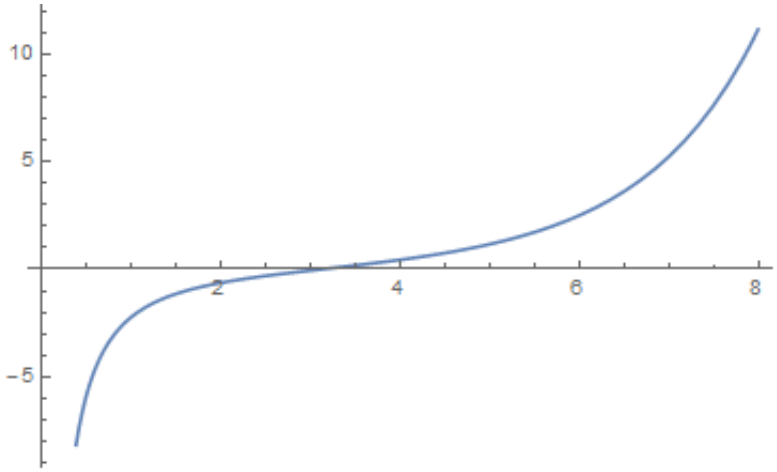

Figure IV.1. Graph of t-distribution parameter function

By applying the Newton-Raphson method, we receive the solution $\hat{v} \approx 3.12$ of $g(v)=0$ given starting value 3, after 6 times to run.

In conclusion, it is optimal if observation PDF belongs to exponential family where the derivative $DLnL(\eta)$ of log-likelihood specified by formula IV.5 is simple, which makes solving estimation equation easier. Otherwise, Newton-Raphson method is a work-around because it can find out solution if the target function is increasing (decreasing) within the interval containing solution. We cannot solve effectively an arbitrary equation. In future, I try my best to prove whether the solution of general estimation equation specified by formula II.5 exists. When asserting the existence of solution, we can apply artificial intelligence approaches such as neural network and genetic algorithm into solving formula II.5 instead of using only Newton-Raphson method.

List of Formulas

Formula I.1.1. Forward variable
Formula I.1.2. Recurrence property of forward variable
Formula I.1.3. Probability $P(O|\Delta)$ based on forward variable
Formula I.1.4. Backward variable
Formula I.1.5. Recurrence property of backward variable
Formula I.1.6. Probability $P(O|\Delta)$ based on backward variable
Formula I.2.1. Joint probability of being in state s_i at time point t with observation sequence O
Formula I.2.2. Optimal state at time point t
Formula I.2.3. Joint optimal criterion at time point t
Formula I.2.4. Recurrence property of joint optimal criterion
Formula I.2.5. Backtracking state
Formula I.3.1.1. EM optimization criterion based on conditional expectation
Formula I.3.2.1. General EM conditional expectation for HMM
Formula I.3.2.2. EM conditional expectation for HMM
Formula I.3.2.3. Lagrangian function for HMM

ISSN: 0075-5222

Formula I.3.2.4. HMM parameter estimate
Formula I.3.2.5. Joint probability $\xi_t(i, j)$
Formula I.3.2.6. The γ_t is sum of ξ_t over all states
Formula I.3.2.7. HMM parameter estimate in detailed
Formula II.1. Probability density function (PDF) of observation
Formula II.2. Joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ based on single PDF
Formula II.3. EM conditional expectation for continuous observation HMM with single PDF
Formula II.4. Lagrangian function for continuous observation HMM with single PDF
Formula II.5. Equation of single PDF parameter
Formula II.6. Continuous observation HMM parameter estimate with single PDF
Formula III.1. Mixture model probability density function (PDF) of observation
Formula III.2. Partial joint probabilities $\xi_t(i, j, k)$ and $\gamma_t(j, k)$ based on mixture model PDF
Formula III.3. Relationship between quantities $\xi_t(i, j)$, $\gamma_t(j)$ and partial quantities $\xi_t(i, j, k)$, $\gamma_t(j, k)$
Formula III.4. EM conditional expectation for continuous observation HMM given mixture model PDF
Formula III.5. Lagrangian function for continuous observation HMM with single PDF
Formula III.6. Weight estimate of partial PDF
Formula III.7. Equation of partial PDF parameter
Formula III.8. Continuous observation HMM parameter estimate with mixture PDF
Formula IV.1. Exponential family PDF
Formula IV.2. Log-partition function $A(\eta)$
Formula IV.3. First derivative of log-partition function $A(\eta)$ with regard to η
Formula IV.4. Log-likelihood function of exponential family PDF
Formula IV.5. Derivative of log-likelihood function of exponential family PDF
Formula IV.6. Equation of exponential family PDF parameters
Formula IV.7. Student's t-distribution
Formula IV.8. Functions gamma, digamma, trigamma
Formula IV.9. Equation of t-distribution PDF parameter

References

Boyd, S., & Vandenberghe, L. (2009). *Convex Optimization.* New York, NY, USA: Cambridge University Press.

Burden, R. L., & Faires, D. J. (2011). *Numerical Analysis* (9th Edition ed.). (M. Julet, Ed.) Brooks/Cole Cengage Learning.

Cheng, C.-C., Sha, F., & Saul, L. K. (2009). A Fast Online Algorithm for Large Margin Training of Continuous Density Hidden Markov Models. *Proceedings of the Tenth Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, (pp. 668-671).

ISSN: 0075-5222

Brighton. Retrieved May 28, 2016, from http://cseweb.ucsd.edu/~saul/papers/is09_hmm.pdf

Couvreur, C. (1996). *Hidden Markov Models and Their Mixtures.* Université Catholique de Louvain, Department of Methematics. Louvain-la-Neuve: ResearchGate. Retrieved May 7, 2016, from https://www.researchgate.net/publication/2677995_Hidden_Markov_Models_and_Their_Mixtures

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. (M. Stone, Ed.) *Journal of the Royal Statistical Society, Series B (Methodological), 39*(1), 1-38.

Fosler-Lussier, E. (1998). *Markov Models and Hidden Markov Models: A Brief Tutorial.* Technical Report TR-98-041, International Computer Science Institute, USA.

Gallová, J., & Kučerka, N. (n.d.). *Measurement of air humidity.* Technical Report, Comenius University in Bratislava, Faculty of Pharmacy.

Huo, Q., & Lee, C.-H. (1997, March). On-Line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate. *IEEE Transactions on Speech and Audio Processing, 5*(2), 161-172. doi:10.1109/89.554778

Jebara, T. (2015). *The Exponential Family of Distributions.* Columbia University, Computer Science Department. New York: Columbia Machine Learning Lab. Retrieved April 27, 2016, from http://www.cs.columbia.edu/~jebara/4771/tutorials/lecture12.pdf

Jia, Y.-B. (2013). *Lagrange Multipliers.* Lecture notes on course "Problem Solving Techniques for Applied Computer Science", Iowa State University of Science and Technology, USA.

Lee, C. H., Rabiner, L. R., Pieraccini, R., & Wilpon, J. G. (1990, April). Acoustic modeling for large vocabulary speech recognition. *Computer Speech & Language, 4*(2), 127-165. doi:10.1016/0885-2308(90)90002-N

Montgomery, D. C., & Runger, G. C. (2003). *Applied Statistics and Probability for Engineers* (3rd Edition ed.). New York, NY, USA: John Wiley & Sons, Inc.

Nguyen, L. (2016, April 27). Beta Likelihood Estimation and Its Application to Specify Prior Probabilities in Bayesian Network. (T.-X. He, P. Bracken, J. C. Valverde, & J. Li, Eds.) *British Journal of Mathematics, 16*(3). doi:10.9734/BJMCS/2016/25731

Nguyen, L. (2016). Tutorial on Hidden Markov Model. (L. Nguyen, & M. A. MELLAL, Eds.) *Special Issue "Some Novel Algorithms for Global Optimization and Relevant Subjects", Applied and Computational Mathematics (ACM).*

NIST. (2008, March). International System of Units (SI). (330 - 2008 Edition). (B. N. Taylor, & A. Thompson, Eds.) Gaithersburg, Maryland, USA: National Institute of Standards and Technology.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. (J. H. Trussell, Ed.) *Proceedings of the IEEE, 77*(2), 257-286.

ISSN: 0075-5222

Ramage, D. (2007). *Hidden Markov Models Fundamentals*. Lecture Notes, Stanford University, USA.

Schmolze, J. G. (2001). *An Introduction to Hidden Markov Models*. Lecture Notes on course "COMP 232-Knowledge Based Systems", Department of Computer Science, Tufts University.

Sean, B. (2009). *The Expectation Maximization Algorithm - A short tutorial*. University of Notre Dame, Indiana, Department of Electrical Engineering. Sean Borman's Homepage.

Sha, F., & Saul, L. K. (2006). Large Margin Hidden Markov Models for Automatic Speech Recognition. (B. Schölkopf, J. C. Platt, & T. Hoffman, Eds.) *Advances in Neural Information Processing Systems (NIPS 2006), 19*, 1249-1256. Retrieved May 29, 2016, from http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2006_143.pdf

Sha, F., & Saul, L. K. (2009). Large margin training of continuous-density hidden Markov models. In J. Keshet, S. Bengio, J. Keshet, & S. Bengio (Eds.), *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods* (pp. 101-114). Chichester, UK: Wiley & Sons. doi:10.1002/9780470742044.ch7

Wikipedia. (2014, August 4). *Karush–Kuhn–Tucker conditions*. (Wikimedia Foundation) Retrieved November 16, 2014, from Wikipedia website: http://en.wikipedia.org/wiki/Karush–Kuhn–Tucker_conditions

Wikipedia. (2016, March September). *Exponential family*. (Wikimedia Foundation) Retrieved 2015, from Wikipedia website: https://en.wikipedia.org/wiki/Exponential_family

Wikipedia. (2016, April 12). *Student's t-distribution*. (Wikimedia Foundation) Retrieved April 24, 2016, from Wikipedia website: https://en.wikipedia.org/wiki/Student%27s_t-distribution