

MATH491

Summer Research Project

2005 – 2006

Developing a Hidden Markov Model for Assessing
the Health of Preterm Babies

James Roscoe

Department of Mathematics and Statistics
University of Canterbury

Developing a Hidden Markov Model for assessing the health of preterm babies

Student: James Roscoe (0333864 / 79757400)

Acknowledgements:

Firstly, I would like to thank Dr Dominic Lee, for supervising this project; also, I would like to thank Dr Glynn Russell, for his provision of data and for his medical expertise.

Abstract:

Premature babies, because of their underdeveloped biological systems, often display cardiorespiratory instabilities. Yet, at the same time, many paediatric illnesses also affect cardiorespiratory functions. For a certain baby, it can therefore be difficult to determine the cause of such instabilities, and this ramifies on treatment decisions. We look to develop a Hidden Markov Model for modelling the health of preterm babies, as this is useful for uncovering information on the hidden states of a system – in this case, the health of a premature baby. First, we provide a background for the study of Hidden Markov Models; meanwhile, we develop the variants of Hidden Markov Models that are most desirable for our application, and describe how inference can be made in each case.

Table of Contents:

1)	Introduction	3
2)	The data	5
3)	The Hidden Markov Model	7
4)	Developing an appropriate model	9
-	4.1 Model Topologies	9
-	4.2 Silent States	14
-	4.3 Infinite Hidden Markov Model	16
-	4.4 Bayesian Networks	17
5)	Inference	20
-	5.1 Training	20
-	5.2 Tools for Inference	21
-	5.3 Inference	22
-	5.4 Implementation	23
6)	The Markov-Switching Model	24
7)	Conclusion	28
	References	29

1) Introduction:

The aim of this project is to develop a Hidden Markov Model for the monitoring of the health of premature babies. Because of their underdeveloped biological systems, these babies often undergo cardiorespiratory instabilities, such as lowering of the oxygen level in the blood, or variations in the heart rate and respiratory patterns. Yet, at the same time, various paediatric illnesses may also affect cardiorespiratory functions.

For a particular baby, it can be difficult to discern whether any physiological instabilities are due merely to the baby's being premature, or whether the baby is, in fact, intrinsically ill. This has two ramifications: it impacts upon decisions on whether or not to treat for illness – which is important, because the underdeveloped nature of premature babies means that introducing foreign chemicals, or any other disturbances, should be avoided as much as possible; and, if treatment is given, the efficacy of such treatment is more difficult to gauge. This latter point is accentuated by the vast quantities of data that are obtained from premature babies, with machines able to take readings once every two seconds.

In this project, we look to introduce a Hidden Markov Model for modelling the health of premature babies, because these models have an ability to link observed variables to unobserved states in a system – in this case, the unobserved states relate to the health of the baby. Furthermore, the Hidden Markov Model has a temporal structure that renders it appropriate for coping with the vast quantities of data that can be put out.

In fact, we are unaware of previous attempts to use Hidden Markov Models for our purpose; however, these models have been used in other fields of medical research, like, for example, Scott et al. (2005). In this paper, Scott looks to compare the medications, clozapine and haloperidol, in terms of their ability in treating schizophrenia. These are tested on a designated group of patients, and with their unobserved health states considered as hidden states, clinical change is assessed with regards to the transition probabilities for patients moving between states. Where this differs from our application is that there is no comparative assessment in our application; we are looking to classify babies into health states without modelling effects of treatment.

In this exposition, we begin in Section 2 by describing the sets of data that we use, and then, in Section 3, introduce the Hidden Markov Model, along with its fundamental properties; for a detailed account of Hidden Markov Models, see Cappe et al. (2005). Next, in Section 4, we propose some simple models for tracking the health of premature babies, based upon the Hidden Markov Model presented; and we then present some extensions of Hidden Markov Models, and discuss the applicability of these to our task. In Section 5, we propose how inference should proceed for these models. Finally, in Section 6, we introduce the Markov-Switching Model, and its related inference, along with the motivation of its use for our application. A brief conclusion is offered in Section 7.

2) The data:

In order to make inference about the health of preterm babies, the physiological variables suggested to be most useful include the oxygen concentration in the blood, the pulse rate, and the respiration rate. In intensive care units, measurements for these variables are routinely taken, so there are no problems of having a paucity of data. The readings for both the oxygen concentration and the pulse rate are taken from a pulse oximeter, whilst the respiration rate observations are taken from a separate instrument that measures chest and abdominal inductance or impedance signals. The pulse and respiration rates are measured as a frequency per minute, whereas the oxygen concentration is taken as a straight percentage. The pulse oximeter is insensitive to measurements taken between 0 and 70, and has an internal control which is capable of labelling some readings as potentially unreliable – these labels are included amongst the data, and in Figure 1, are circled in red. The measurements themselves are taken thirty times a minute, at two-second intervals, and comprise two subsets: one pertaining to babies known to be healthy; and another for ill babies. These data have been classified from the posterior information of medical experts.

Also included within the sets of data is information on the activity of the baby. In particular, there is information on the state of sleep of the baby – or the state of wakefulness if the baby is awake: these include details on whether the baby is awake and quiet, awake and crying, in quiet sleep, or active sleep, and so on. There is also information on the mode of respiration, and the means of feeding.

Figure 1(a) shows a portion of the oxygen concentration measurements for one baby, and Figure 1(b) shows a corresponding portion of the pulse rate.

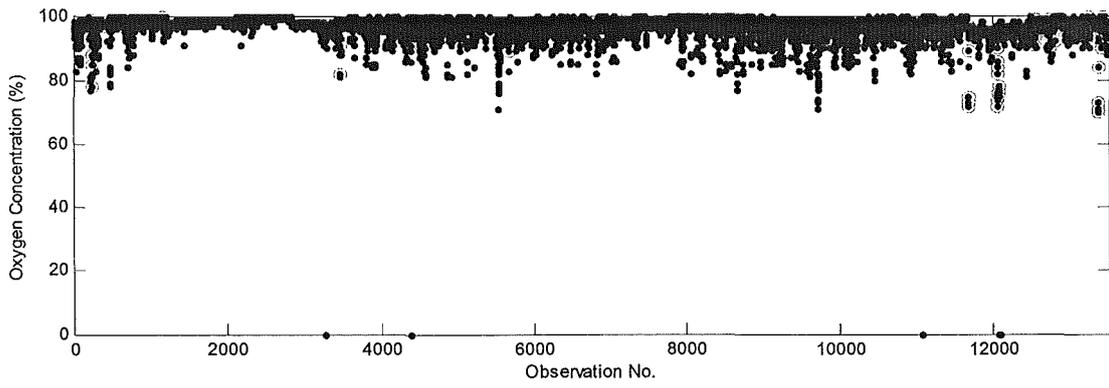


Figure 1(a) – A portion of oxygen concentration readings for a given baby.

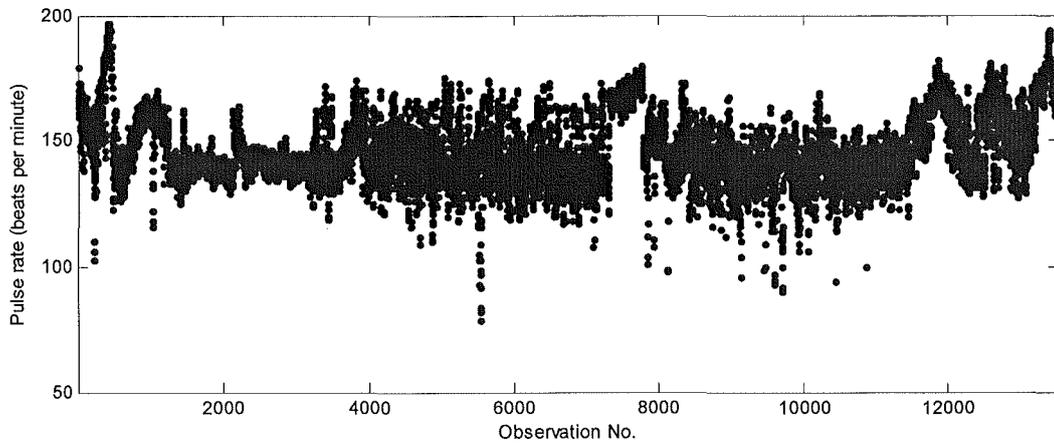


Figure 1(b) – Corresponding pulse rate readings for the same baby.

3) The Hidden Markov Model:

A Hidden Markov Model (often abbreviated to the acronymic "HMM") comprises two sets: a set of *observed states*, \mathbf{X} , and a set of *hidden states*, Y , which contains precisely K elements. A time series is observed, whose observations x_1, \dots, x_T are each a member of \mathbf{X} . The probability distribution of any one such observation, x_t , depends only on the *hidden state* at time t , $y_t \in Y$, where the hidden state is not observed. This means that, given y_t , x_t is conditionally independent of the other observations. Meanwhile, the hidden states, y_1, \dots, y_T are assumed to follow a Markov chain, with a stationary transition matrix $\mathbf{Q}(i, j) = P(y_t = j | y_{t-1} = i)$. In other words, \mathbf{Q} is a $K \times K$ matrix which governs the probabilities of transition between all the pairs of members in Y . This is depicted in Figure 2.

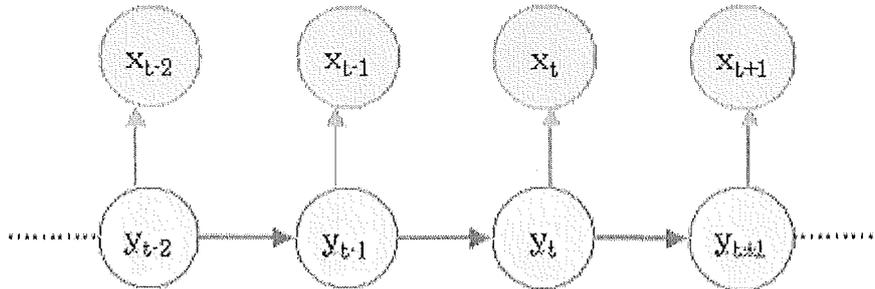


Figure 2 – An illustration of the Hidden Markov Model.

For the remainder of this text, we shall assume that \mathbf{X} is, like Y , a finite set. This assumption can be relaxed for continuous observations.

Now, note that a requisite condition of the hidden states following a Markov chain, is that the current hidden state y_t is conditionally independent of all hidden states prior to y_{t-1} given the value of y_{t-1} . This means that we can rewrite the joint probability distribution of the observed sequence $x = x_1, \dots, x_T$ and the hidden sequence (or the *path*) $y = y_1, \dots, y_T$:

$$P(x, y) = \dot{P}(y_1)P(x_T | y_T) \prod_{i=1}^{T-1} P(x_i | y_i) \mathbf{Q}(y_i, y_{i+1}) \quad (3.1)$$

If this is maximised with respect to y , then we get the *most probable path* through the model. This can be found using the *Viterbi* algorithm, and this algorithm is a common means of inference for Hidden Markov Models (see, for example, Ewens and Grant (2001), or Durbin et al. (1998), for more details). However, given that the time series for our physiological variables are very long, such a joint probability will be very small, and there are likely to be other combinations of paths and observations whose joint probabilities lie very close to the probability of the most probable path. Also, in our application, what we are really interested in is the health state of a baby at one particular time t . With these ideas in mind, it becomes apparent that what we really seek is $P(x, y_t)$, rather than $P(x, y)$.

To this end, it can be shown that

$$P(x, y_t = k) = f_k(t) b_k(t), \quad (3.2)$$

$$\text{where } f_k(t) = P(x_1, \dots, x_t, y_t = k), \quad (3.3)$$

$$\text{and } b_k(t) = P(x_{t+1}, \dots, x_T | y_t = k). \quad (3.4)$$

The $f_k(t)$ and $b_k(t)$ values are found by the forward-backward algorithm, and inference for the Hidden Markov Model proceeds from there.

4) Developing an appropriate model:

4.1) Model Topologies

Now, say that we assume that our physiological observations are based on a Hidden Markov Model with K hidden states, and that all transition probabilities are allowed to be nonzero. Then, we must model K transitions emanating from each hidden state, and, since there are K hidden states, there are a total of K^2 transition probabilities to be estimated. This number is likely to be too large for inference to be practicable. Therefore, we need to set some of the probabilities to zero, because not only will the HMM algorithms run more smoothly, this is often a more apt reflection of the model interpretation – if, in reality, some of the transitions are impossible, then it is appropriate to disallow such transitions in the model, by enforcing the probabilities of such transitions to be zero.

To this end, note that the hidden state Markov chain can be represented in the usual fashion for Markov chains. In particular, we can draw a graph whose vertices represent the members of \mathcal{Y} , with a directed edge linking one vertex to another precisely when there is a strictly positive possibility of transition, in the direction indicated, between these two states. Such a graphical representation is sometimes, in the Hidden Markov Model literature, referred to as the *topology* of the HMM – the term "topology" used to indicate that it is the relationships between the variables which are of importance. Then, once we have specified numerical quantities for the allowed transitions, and have modelled the emission distribution pertaining to each hidden state, we have fully specified the Hidden Markov Model.

Our task is, therefore, to find an appropriate topology, so that we can use Hidden Markov Modelling to make inference about the health state of a new baby, given the observed time series for its physiological variables. Whilst analytical methods have been developed as aids in selecting correct model topologies, in our problem, the hidden states have a direct scientific interpretation, so it is sensible to incorporate expert knowledge into the structure of our design – note that there is always a bias-variance tradeoff in such model selection problems, anyhow.

As a starting point, consider a model in which there are two hidden states: one corresponding to health (denoted by 1); and one corresponding to illness (denoted by

2). In this case, $Y = \{1,2\}$. There is an emission distribution taking the hidden state 1 to the observations; likewise, there is another emission distribution for hidden state 2. This HMM is depicted in Figure 3:

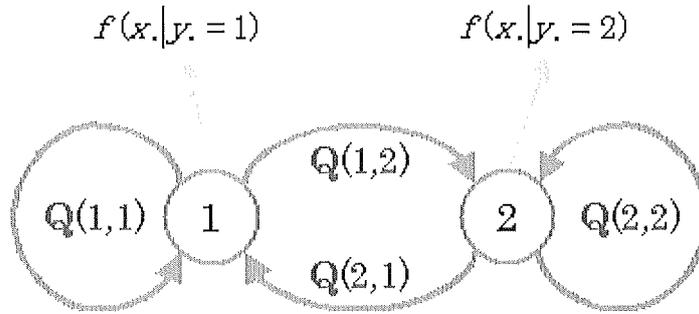


Figure 3 – Illustration of model with two hidden states.

Because our data are labelled, there is no information from which we can base estimates of the hidden state transition probabilities. Therefore, we must use expert medical knowledge to prescribe sensible values for these probabilities. Once this is achieved, the model gets its weight from the estimation of the emission distributions.

Now, a critical property of Hidden Markov Models is that the observations are held to be conditionally independent given the path. This implies that two observations at times t_1 and t_2 should be independently distributed whenever y_{t_1} and y_{t_2} are known. Again, our data are labelled, and so, for this model, the observations for each time series should be identically and independently distributed. This is clearly an unrealistic assumption, because there are patterns and jumps that, throughout the time series, are immediately detectable to the eye. Therefore, having two such hidden states is not sufficient to encapsulate the generation process of these time series. A more complex topology is required, in order to model the activities of the baby – over and above the health of the baby – and to account for the effects of these activities on the three chosen physiological variables.

Notwithstanding, it should be pointed out that, simply because the conditional independence assumption has been violated, this does not necessarily mean that the above model is worthless. If enough data are accumulated for both ill and healthy babies, then the emission distributions, averaged over the other influences on the

physiological variables, may still be sufficiently distinguishable in order for inferences to be made about a new baby's state of health.

Nevertheless, it is clear that the model can be vastly improved upon, with the hidden states further "partitioned" in order to account for the activities of the baby. In particular, we consider that there are two types of influence on the three physiological variables, which are not directly related to the health of the baby: firstly, there is the state of consciousness, or *consciousness state*, of the baby; and secondly, there is the state of intervention, or *intervention state*. The former differs from the latter in that the state of consciousness is a natural phenomenon, which is impervious to human intervention, whereas the intervention state is not directly attributable to the natural response of the baby, but reflects the responses which the caregivers deem most appropriate for ensuring the comfort and longevity of the baby.

Hence, if C_t is the consciousness state at time t , and I_t the intervention state, then we shall denote $A_t = (C_t, I_t)$ as the *activity state* of the baby at time t . Let it be remarked that rather than being merely a means of reacquiring conditional independence given the hidden states, including the activity state as part of the hidden state topology is likely to incorporate information that is fruitful for discrimination, because the transition probabilities between the activity states are affected by the state of health of the baby, and so these ought to be modelled.

With this in mind, the following tree diagrams (Figure 4) summarise the considered main influences on our three observation variables - these were based on expert advice, and also on classifications in both Sahni et al. (1999) and Galland et al. (2000):

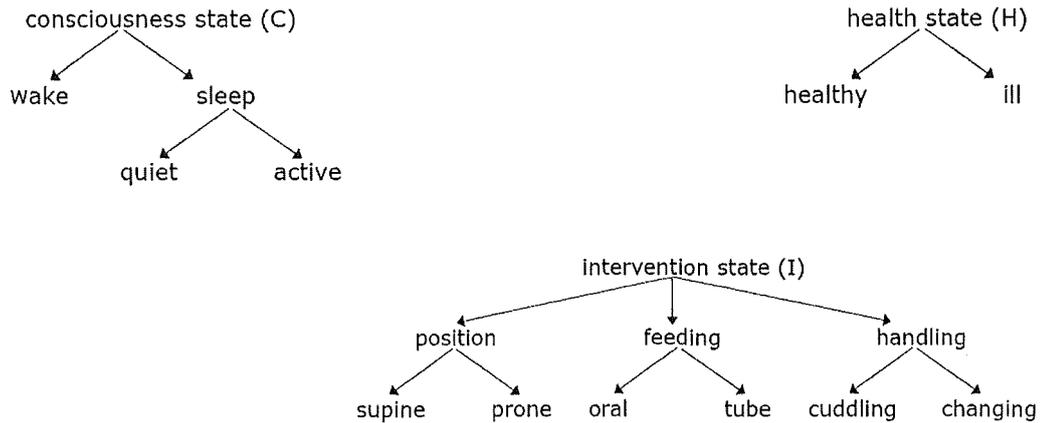


Figure 4 – Health and activity state classifications.

We shall now expound two possible means of incorporating the information on the hidden states.

The first model has a sort of hierarchical structure. Indeed, there are similarities to the hierarchical model which was developed by Skounakis et al. (2003) to model syntax, but whereas they embed Hidden Markov Models in the "second tier" of hierarchy, our secondary Markov chains are fully observed.

In this case, the activity state A is brought to the same conceptual level as the time series observations. Once the hidden state at time t is chosen, then both the activity state and the observation at time t are emitted. In the meantime, the activity states are allowed to follow their own observed Markov chain whenever the underlying hidden state remains the same. This topology is more appropriate if we wish to self-specify a minimal number of probabilities, since these are contained only in the first tier of the hierarchy.

In Figure 5, we illustrate the case where $(H_t, A_t) \in \{1, 2\} \times \{1, 2, 3\}$.

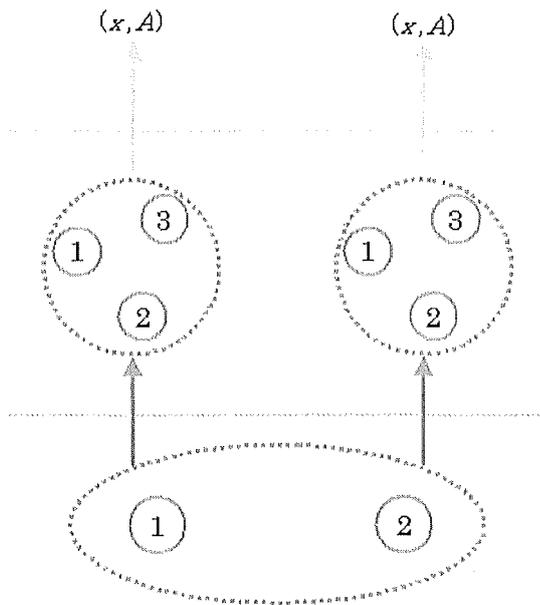


Figure 5 – Illustration of HMM for two health states and three activity states.

By contrast, in the second model, we enforce the information about the activity state to be contained within each hidden state. Indeed, we let each hidden state contain information on the health state, the consciousness state, and the intervention state; in other words, each hidden state is an ordered triple whose components represent the three types of substate, where the last two relate to the activity state. Here, at most one component of the hidden state is allowed to change in a given transition; otherwise the possible number of transitions would be too large to practicably estimate. Allowing only one component to make a transition is by no means unreasonable, because the probability of two components changing simultaneously is exceptionally small, and even if it is perceived that two components change at once, medical knowledge can be used to specify which ought to have preceded which, thereby avoiding ambiguity.

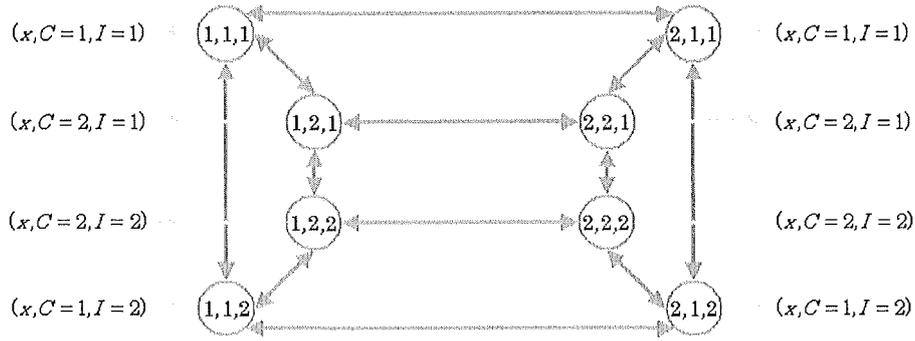


Figure 6 – Illustration of HMM with two each of health states, consciousness states, and intervention states.

From Figure 6, we see that this model displays more symmetry, and has the potential to encapsulate information on more complex relationships than does the previous model.

4.2) Silent States:

The use of silent states can often reduce the number of transition probabilities to be estimated. This is important, as the maximum likelihood estimation procedures, which are commonly used for Hidden Markov Model inference, are susceptible to both over-fitting, and the prevalence of local maxima in the likelihood function. Reducing the number of transition probabilities to be estimated helps to curtail both of these problems.

A *silent state* is a state which is unable to make a transition to itself, and which makes a null emission. Consider two subsets of Y : Y_1 and Y_2 . To illustrate how silent states can be used to reduce the size of the model, say that Y_1 and Y_2 are two disjoint sets of hidden states. Unless some of the probabilities are set to zero, in order to model Markov chain transitions between Y_1 and Y_2 , we have to estimate the transition probability for each ordered pair from both $Y_1 \times Y_2$ and $Y_2 \times Y_1$. However, if we add a silent state, and force any transitions from Y_1 to Y_2 to visit this silent state, then the number of transitions is likely to be reduced – if $|Y_1| = K_1$, and $|Y_2| = K_2$, then the number of transitions alters from $K_1 K_2$ to $K_1 + K_2$ – and when this is repeated for transitions from Y_2 to Y_1 , the reduction in the number of parameter estimates is doubled.

Now, say that, given the observations, we wish to model a silent state between x_j and x_{j+1} . We can achieve this by nestling a null emission between the observations x_j and x_{j+1} , and posit that this null observation can only be emitted by the aforementioned silent state; in other words, we force the model to visit the silent state between time j and time $j + 1$.

For example, consider Figure 7. In this example, $Y_1 = \{A, B, C\}$, and $Y_2 = \{X, Y, Z\}$. Without the use of silent states, we have to specify eighteen transition probabilities; if a silent state is forced whenever the model makes a transition from Y_1 to Y_2 , this number reduces to twelve.

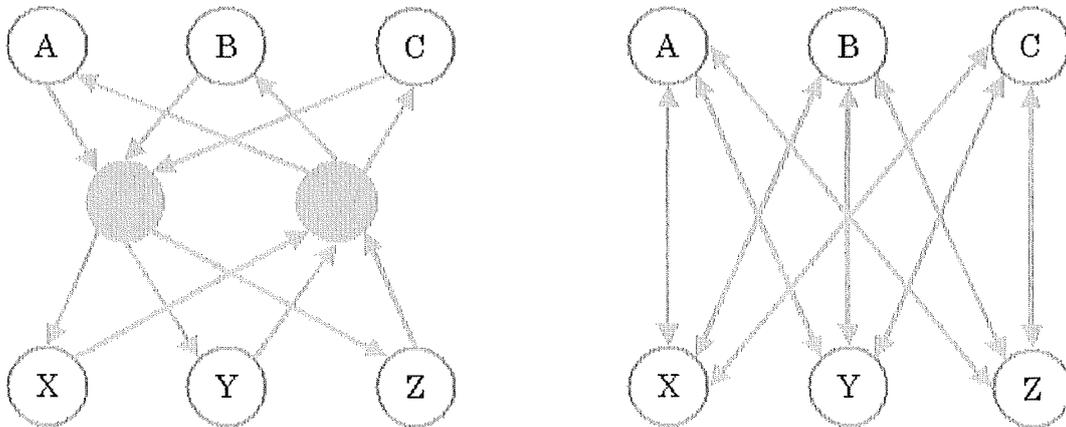


Figure 7 – Illustration of the use of silent states for transitions between $\{A, B, C\}$ and $\{X, Y, Z\}$.

In our case, the appropriateness of adding silent states transcends the reduction in the number of transition probabilities. Consider the case whereby the consciousness states are modelled, while the health state remains the same. Rather than modelling all of the transitions that traverse between states of wakefulness and states of sleep, we can instead enforce a silent state, so that whenever a transition is made between wakefulness and sleep, the Markov chain must enter this silent state. This seems particularly apt for our task, as the state of wakefulness should have little or no bearing on the following state of sleep, and vice versa. Therefore, it is sensible to estimate the respective probabilities of falling asleep from each state of wakefulness, and *then* estimate the probabilities, once asleep, of entering each state of sleep. So, not only does this reduce the number of transitions, but it also has a

corresponding scientific interpretation. Similar methodology can, of course, be used elsewhere in the model.

We now look at some other extensions of the Hidden Markov Model, and examine their appropriateness for our task.

4.3) **Infinite Hidden Markov Model:**

For many problems in Hidden Markov Modelling, the assumption of a finite number of states does not fairly reflect reality; however, trying to use a Hidden Markov Model for infinitely many states is difficult, because such a model would have infinitely many parameters contained within both its transition matrix and its emission distributions. Notwithstanding this, Beal et al. (2003) look to build a theory of Infinite Hidden Markov Models, in which the transition parameters are integrated out, to leave just three hyperparameters, which completely determine the characteristics of the model: the first relates to the propensity of the hidden states to make transitions to themselves; the second gives the ability of the model to enter the so-called "oracle", and thirdly, there is another parameter which determines to what extent, once the oracle has been entered, the model will visit entirely new states.

The model uses an observed transition matrix \mathbf{N} , which keeps count of the number of transitions between each pair of states. Then, given that the model is in state i , the next transition is modelled as a Dirichlet process, for which all but one of the parameters are determined from the i -th row of \mathbf{N} : this ensures that popular states continue to be visited more frequently. The remaining hyperparameter gives the probability that the model instead enters the oracle: a similar Dirichlet process, using a count vector \mathbf{n} , is used for the oracle – otherwise, when a new state is entered, there would be no history with which the model could be guided back to existing states – whereupon the remaining hyperparameter gives the probability that the model can enter an entirely new state. Note that this is indeed the only mechanism for model entering new states. Finally, a third, auxiliary parameter is incorporated to model self-transitions, which have a special importance in many applications.

Once the path of the Infinite Hidden Markov Model is determined, a similar technique is used for the emission mechanism, and this is again needed, for it is infeasible to estimate infinitely many emission distributions. Hence, having finally

yielded an observed sequence, procedures are proposed by Beal et al. (2003) for estimating the path, and also the three hyperparameters.

They contend that, although the model has been purged of all but three hyperparameters, these parameters allow a rich variety of models, whose infinity of states gives them ascendancy over the traditional finite-state Hidden Markov Models. However, no natural applications are proposed as yet, and indeed, the ideas in this paper are not directly applicable to our problem, as we wish to model a finite number of carefully chosen hidden states. Furthermore, the Infinite Hidden Markov Model disregards many of the complex relations in the model by virtue of integrating out the majority of parameters. However, some of these ideas in this model could, in principle, be used to reduce the dimensionality of the model, should this prove to be too cumbersome.

4.4) Bayesian Networks:

The Bayesian network framework involves a wider range of models, into which the Hidden Markov Model fits. Bayesian networks are a means of representing dependencies amongst a set of random variables.

More specifically, say we have N random variables W_1, \dots, W_N , with a joint probability distribution $P(W_1, \dots, W_N)$. Elementary probability theory can be used to factorise this distribution into the product of conditional probabilities; such a factorisation may not, necessarily, be unique.

A *Bayesian network* represents such interdependencies as a graph: the nodes correspond to the random variables W_1, \dots, W_N , and a directed edge is drawn from the node W_I to the node W_J if W_J is conditioned upon W_I in some factor of $P(W_1, \dots, W_N)$. Since the factorisation of $P(W_1, \dots, W_N)$ is not necessarily unique, neither is any given Bayesian network representation of this distribution.

When a Bayesian network models a time series, it is known as a *dynamic* Bayesian network. In this case, the directed edges flow forwards in time, since the probability distribution of a time series random variable should only depend on past events, not future events. Both ordinary Markov chains, and Hidden Markov Models, are types of dynamic Bayesian networks. For the latter, this can be seen by inspecting

equation (3.1) above. For more details about Bayesian Networks, and their relationship to Hidden Markov Models, see Ghahramani (2001).

Estimation for a complete set of data:

Say that we are given a set of independent, identically distributed observations X^1, \dots, X^R . The maximum likelihood estimates of the model parameters θ are found by maximising the log-likelihood

$$L(\theta) = \sum_{c=1}^R \log P(X^c | \theta).$$

With the observations including all variables in the network, this further factorises as

$$\log P(X^c | \theta) = \sum_{d \in \{W_1, \dots, W_N\}} \log P(X_d^c | X_{P(d)}^c, \theta)$$

where $P(d)$ denotes the *parents* of d , the parents of a random variable W being the variables whose nodes have a directed edges linking them to W .

In our example, where the variables are discrete-valued, parameter estimation results in keeping a normalised table of counts. For example, say that we want to estimate the conditional probabilities for W_j given its parents. Then we count the number of times the model makes a transition to W_j from each parent, and divide these values by the number of times spent in the node of each parent.

Estimation without a complete set of data:

In the case where the set of data is incomplete, maximum likelihood estimation can still be implemented; this time using the *EM (expectation-maximisation)* algorithm, developed by Dempster et al. (1977). The EM algorithm can be thought of as a series of optimisation problems, with each one providing an improved estimate of the model parameter vector θ , using information on what the

model is likely to have done given the previous estimate for θ . The details of this algorithm are, in general, quite involved; for more details, see Tanner (1996).

5) Inference:

Once we have chosen a model topology, then we use the labelled data to create an HMM, from which inferences can be made about the state of health of a new baby.

5.1) Training:

We have two sets of labelled data here: one corresponding to babies designated as being healthy; and one for ill babies. As mentioned earlier, not only is the state of health known for each set, but also the activity state – indeed, knowledge of the activity state is always possible. For each set, we can estimate the transition probabilities given its health state; and we can also estimate the emission distribution given the hidden state. Because, for each labelled set, the path is known, both of these tasks are easy to implement.

If, however, the path is unknown, there are still algorithms that can be used to find the maximum likelihood estimates. One such algorithm is the *Baum-Welch algorithm*, a special case of the more general EM algorithm alluded to above, developed by Baum et al. (1970). The Baum-Welch algorithm is an iterative procedure in which each iteration produces a new estimate for θ . Specifically, each iteration consists of an "E" step, and an "M" step: the E step calculates the expected value $E[P(y|x)]$; the M step maximises θ , using these expectations as the weighted probabilities for each sequence. It can be shown that each iteration of this algorithm is guaranteed to increase the model likelihood.

In our application, however, in order to estimate the transition probabilities, it can be shown that, when the path is known, the method of maximum likelihood estimation instead yields the estimates

$$\hat{Q}(i, j) = \frac{\text{number from state } i \text{ to state } j}{\text{number from state } i \text{ to any state.}}$$

Similarly, when trying to estimate the emission distributions, if we treat our data as discrete observations, then this amounts to specifying a matrix, whose (i,j) -th

entry indicates the proportion of times that x_j has been visited when the model is in hidden state y_i . To avoid the problems which are incurred when trying to estimate a transition which has not, in fact, occurred, we can include *pseudocounts*, which are small positive counts imposed for each transition, and these avoid problems of dividing by zero, or other very small numbers. In fact, pseudocounts can also be used to incorporate prior beliefs about the transition probabilities – the size of the pseudocounts should increase with the strength of the beliefs.

5.2) Tools for Inference:

Now say that we are given an observed sequence of length T , from a new baby for whom we are unable to directly classify the health state.

For inferring the health state, the main tool that we can use is known as the *forward-backward* algorithm. This is based upon the equations (3.2) – (3.4) above. The forward-backward algorithm comprises the *forward* algorithm, followed by the *backward* algorithm.

The forward algorithm iteratively calculates $f_k(t) = P(x_1, \dots, x_t, y_t = k)$, for all k . Calculations are based on the recursive formula

$$f_l(t+1) = P(x_{t+1} | y_{t+1} = l) \sum_k f_k(t) P(y_{t+1} = l | y_t = k). \quad (5.1)$$

Similarly, the backward algorithm iteratively calculates $b_k(t) = P(x_{t+1}, \dots, x_T | y_t = k)$, for all k . This, in turn, is based on the recursive formula

$$b_k(t) = \sum_l P(x_{t+1} | y_{t+1} = l) b_l(t+1) P(y_{t+1} = l | y_t = k). \quad (5.2)$$

Initialising the algorithm, we set $f_k(1) = P(x_1 | y_1 = k) P(y_1 = k)$, and $b_l(T) = 1$.

In this context, the f_k values are known as *forward variables*, and the b_k values are known as the *backward variables*. As we can see from these formulae, this is where information on the topology of the model, as well as on the numerical quantities for the transition parameters, filters through into the inference – the estimated $\hat{Q}(i, j)$ values serve as proxies for the unobserved $P(y_{i+1} = l | y_i = k)$ probabilities.

So now, the algorithm has finally allowed us to infer the hidden state probabilities. To be more explicit, recall (3.1) whence it follows that inference can be based upon the equation

$$P(y_i = k | x) = \frac{P(x, y_i = k)}{P(x)} = \frac{P(x, y_i = k)}{\sum_k f_k(T)}.$$

Using the algorithm thus, however, is likely to cause computational problems, due to underflow errors resulting from the very small probability values. In the forward-backward algorithm, this problem can be bypassed, at each step, by rescaling the forward and backward variables. Specifically, each forward variable is replaced by the quantity formed from dividing itself by the sum of the other forward variables at the t -th step; likewise for the backward variables. Initialisation remains as before, with the exception that $b_1(T)$ is first set to $1/|Y|$.

5.3) Inference:

For the new baby as described, the ultimate objective is to infer the state of health of the infant as time progresses. Therefore, we must sum up the posterior probabilities for each activity state in which the baby is healthy; its probabilistic complement is thence the posterior probability of the baby's being ill.

In other words, we wish to calculate

$$\sum_{k \in Y_0} P(y_i = k | x),$$

where Y_0 is the subset of hidden states whereby the baby is healthy.

The forward-backward algorithm yields the $P(y_i = k | x)$ values; the aforementioned maximum likelihood estimates give the parameter values for their computation.

5.4) Implementation:

The intention was to decimate the data, making them more computationally tractable, and also uncorrelated – being purged of their correlation was thought to be a useful proxy for conditional independence. Then the emission distributions were to be estimated.

Unfortunately, when the autocorrelation functions were plotted (see Figure 8), the autocorrelations of the time series were found to be significant and prolonged.

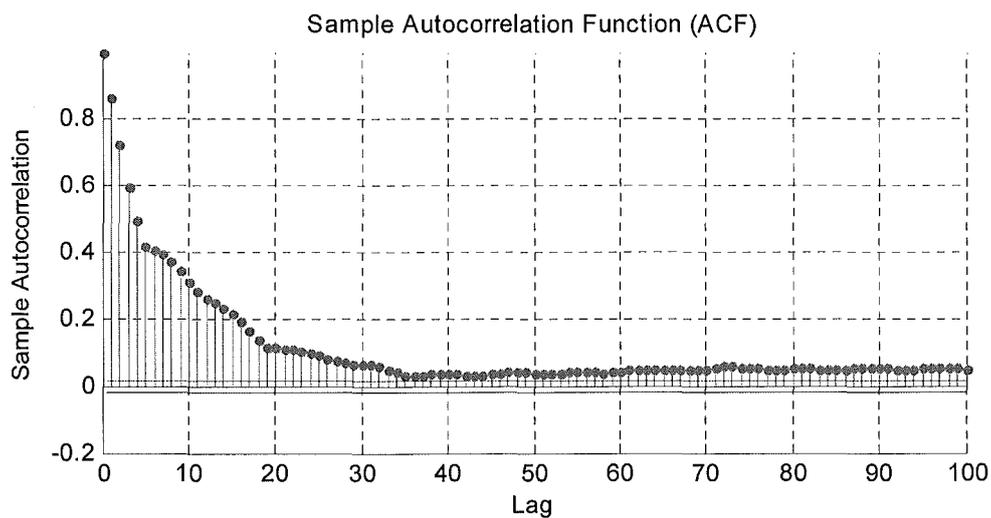


Figure 8 – Autocorrelation function for the oxygen concentration data ; confidence bands are in blue.

This dependence must be accounted for in the model. To this end, we shall now introduce the autoregressive Hidden Markov Model, a special case of the so-called Markov-Switching Model.

6) The Markov-Switching Model:

A *Markov-Switching Model* is an extension of the Hidden-Markov Model whereby the observations are not required to be conditionally independent given the hidden states. In the meantime, however, the dynamics of the hidden state Markov chain remain intact. This notion is depicted in Figure 9, where we see that the underlying Markov chain runs just as before, except that there is an extra layer of dependency amongst the observations.

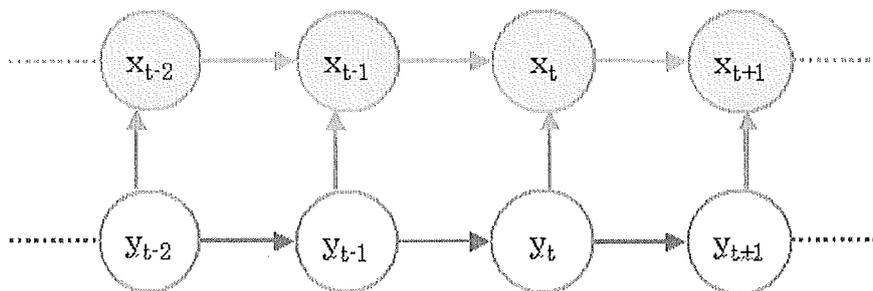


Figure 9 – Conditional structure of an autoregressive HMM with one time-dependency in the past.

Although we have depicted dependency only on the previous observation and the current state, in more generality, this dependency can be extended to variables further back in time as well. In fact, one way to think of this is to consider the model as a *partially observed* Markov chain; that is, a model whereby, at each time step, part of the underlying state is observed, while the other part is hidden. If the dependencies run μ steps into the past, then such a structure can be transformed into the depicted model, by mapping each possible group of μ adjacent states to a single state.

Many of the ideas behind Markov-Switching Models are first attributable to Hamilton (1989). In the 1989 paper, Hamilton successfully used a Markov-Switching Model in order to model the U.S.A. economy: there were two hidden states, with one corresponding to expansion, and the other to recession. The scalar-valued time series for GNP (gross national product) was assumed to follow a fourth-order autoregressive process, whose mean was itself a function of the hidden state. If we liken the recessions and expansions, respectively, to relapses and improvements in health, then we can see a correspondence between Hamilton's application and ours. In fact, Markov-Switching Models have proved very popular in econometrics and finance (a

partial review, as well as some various applications within these fields are given in Hamilton and Raj (2002)); however, other fields of research have been less forthcoming in making use of their attributes.

We now introduce the foundations of the Markov-Switching Model as expounded by Hamilton. The Markov-Switching Model is essentially made of two components.

Firstly, the model specifies that the probability distribution for one observation x_t depends only on the m most recent values of both the observations and the path. In other words,

$$P(x_t | y_t, y_{t-1}, \dots, y_1, x_{t-1}, x_{t-2}, \dots, x_1; \theta) = P(x_t | y_t, y_{t-1}, \dots, y_{t-m}, x_{t-1}, x_{t-2}, \dots, x_{t-m}; \theta), \quad (6.1)$$

where θ is the vector of parameters, which includes the vector of transition probabilities, as well as the autoregressive parameters. Clearly, when $m = 1$, the above pictorial representation holds.

Secondly, the transitions between the hidden states are governed by a Markov chain, again with a transition matrix

$$Q(i, j) = P(y_t = j | y_{t-1} = i). \quad (6.2)$$

As a particular case, if $P(x_t | y_t, y_{t-1}, \dots, y_1, x_{t-1}, x_{t-2}, \dots, x_1; \theta)$ specifies an autoregressive process for x_t , then this model is said to be an *autoregressive Hidden Markov Model*. With the probability distribution for x_t itself conditioned upon the path y , we allow the parameters of this expression, including the autoregressive parameters, to depend on the current and previous hidden states. This is appropriate for our application, where we postulate that the observations are essentially based on a hidden health state, except that this dependency is muddled by an autoregressive layer amidst the observations.

Hamilton (1990) proposed an EM algorithm for making inferences from these models. If the two equations are specified according to equations (6.1) and (6.2)

above, then the algorithm can be described thus: first, a guess θ_0 is made – it can be arbitrarily – for the parameter vector θ ; then, each step consists of first calculating the *smoothed probabilities* $P(y_t, \dots, y_{t-m}; \theta_t)$; after which the three equations (6.3) – (6.5) are solved for θ , to give θ_{t+1} . This process is continued, until some criterion for convergence is satisfied ; perhaps based upon the size of the difference between θ_{t+1} and θ_t .

$$\hat{Q}(i, j)_{t+1} = \frac{\sum_{c=m+1}^T P(y_c = j, y_{c-1} = i | x; \theta_t)}{\sum_{c=m+1}^T P(y_{c-1} = i | x; \theta_t)}, \quad i, j = 1, \dots, K, \quad (6.3)$$

$$\sum_{c=m+1}^T \sum_{y_c=1}^K \dots \sum_{y_{c-m}=1}^K \frac{\partial \log P(x_c | y_c, \dots, y_{c-m}, x_{c-1}, \dots, x_{c-m}; \theta)}{\partial \theta} \Big|_{\theta=\theta_{t+1}} P(y_c, \dots, y_{c-m} | x; \theta_t) = \mathbf{0}, \quad (6.4)$$

$$\rho_{i_m, i_{m-1}, \dots, i_1}^{t+1} = P(y_m = i_m, y_{m-1} = i_{m-1}, \dots, y_1 = i_1 | x; \theta_t), \quad i_1, \dots, i_m = 1, \dots, K. \quad (6.5)$$

where $\hat{Q}(i, j)$ and ρ are elements of θ , with

$$\rho_{y_m, \dots, y_1} = P(y_m, \dots, y_1 | x_m, x_{m-1}, \dots, x_1),$$

modelled as a separate distribution.

Just as before, in our application, use of the EM algorithm is unnecessary, because, for the labelled data, the path is known. We can use the same maximum likelihood techniques for estimating the transition probabilities, and use established time series methods for specifying the structure of the autoregression.

Once this has been performed, we must now outline how inference can be made from this model. Indeed, the basic concepts of the forward-backward algorithm carry through to this case. More explicitly, the backward variables $b_k(t)$ are redefined as $b_k(t) = P(x_{t+1}, \dots, x_T | x_t, y_t = k)$, while the definition of the forward variables remains intact. The two recursive formulae (5.1) and (5.2) become:

$$f_l(t+1) = P(x_{t+1}|x_t, y_{t+1} = l) \sum_k f_k(t) P(y_{t+1} = l | y_t = k); \quad (6.6)$$

$$b_k(t) = \sum_l P(x_{t+1}|x_t, y_{t+1} = l) b_l(t+1) P(y_{t+1} = l | y_t = k). \quad (6.7)$$

Here, we show the derivation of equations (6.6) and (6.7):

$$\begin{aligned} P(x, y_t = k) &= P(x_1, \dots, x_t, x_{t+1}, \dots, x_T, y_t = k) \\ &= P(x_1, \dots, x_t, y_t = k) P(x_{t+1}, \dots, x_T | x_1, \dots, x_t, y_t = k) \\ &= P(x_1, \dots, x_t, y_t = k) P(x_{t+1}, \dots, x_T | x_t, y_t = k) \\ &= f_k(t) b_k(t). \end{aligned}$$

$$\begin{aligned} f_l(t+1) &= P(x_1, \dots, x_{t+1}, y_{t+1} = l) \\ &= \sum_k P(x_1, \dots, x_{t+1}, y_t = k, y_{t+1} = l) \\ &= \sum_k P(x_{t+1} | x_t, y_{t+1} = l) P(x_1, \dots, x_t, y_t = k, y_{t+1} = l) \\ &= P(x_{t+1} | x_t, y_{t+1} = l) \sum_k P(x_1, \dots, x_t, y_t = k) P(y_{t+1} = l | y_t = k) \\ &= P(x_{t+1} | x_t, y_{t+1} = l) \sum_k f_k(t) P(y_{t+1} = l | y_t = k). \end{aligned}$$

$$\begin{aligned} b_k(t) &= P(x_{t+1}, \dots, x_T | x_t, y_t = k) \\ &= \sum_l P(x_{t+1}, \dots, x_T, y_{t+1} = l | x_t, y_t = k) \\ &= \sum_l P(x_{t+2}, \dots, x_T | x_{t+1}, y_{t+1} = l) P(x_{t+1}, y_{t+1} = l | x_t, y_t = k) \\ &= \sum_l b_l(t+1) P(x_{t+1} | x_t, y_{t+1} = l) P(y_{t+1} = l | y_t = k). \end{aligned}$$

7) **Conclusion:**

Thus it can be seen that, although we are unaware of previous attempts of using Hidden Markov Models for modelling premature baby health, the use of these, or their variants, show promise in being able to recover the states of a biological system that are not immediately discernable. This would impact upon decisions regarding treatment, and the consequences of this would be significant for the care of premature babies.

Some models have been proposed for capturing the pertinent information, and methods of inference have been described for each case. Unfortunately, in the time available, we were unable to test these methods and discriminate amongst the models, but it is hoped that these ideas can later be utilised, and ultimately be programmed to function as an aid to a doctor, or any other caregiver, who is required to make a treatment decision based on the perceived state of health of a premature baby.

References:

1. L. E. Baum, T. Petrie, G. Soules, and N. Weiss (1970). "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics*, 41: 164-171.
2. M. J. Beal, Z. Ghahramani, and C. E. Rasmussen (2003). "The Infinite Hidden Markov Model." *Advances in Neural Information Processing Systems* Volume 14: 577-584, MIT Press.
3. O. Cappe, E. Moulines, and T. Ryden (2005). "Inference in Hidden Markov Models," *Springer*.
4. P. Dempster, N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society* 39: 1-38.
5. R. Durbin, S. Eddy, A. Krogh, G. Mitchison (1998). "Biological sequence analysis: Probabilistic models of proteins and nucleic acids," *Cambridge University Press*.
6. W. J. Ewens, and G. R. Grant (2001). "Statistical Methods in Bioinformatics," *Springer*.
7. B. C. Galland, R. M. Hayman, B. J. Taylor, D. P. G. Bolton, R. M. Sayers, and S. M. Williams (2000). "Factors Affecting Heart Rate Variability and Heart Rate Responses to Tilting in Infants Aged 1 and 3 Months," *Pediatric Research*, 48(3): 360-369.
8. Z. Ghahramani (2001). "An Introduction to Hidden Markov Models and Bayesian Networks," *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1): 9-42.
9. J. D. Hamilton (1989). "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, 57(2): 357-384
10. J. D. Hamilton (1990). "Analysis of time series subject to changes in regime," *Journal of Econometrics*, 45: 39-70.
11. J. D. Hamilton, and B. Raj (Eds.) (2002). "Advances in Markov-Switching Models," *Physica-Verlag*.
12. R. Sahni, K. Schulze, S. Kashyap, K. Ohira-Kist, M. M. Myers, and W. P. Fifer (1999). "Body position, sleep states, and cardiorespiratory activity in developing low birth weight infants," *Early Human Development*, 54: 197-206.
13. M. Skounakis, M. Craven, and S. Ray (2003). "Hierarchical Hidden Markov Models for Information Extraction," *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexico. Morgan Kaufmann.
14. S. L. Scott (2002). "Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century," *Journal of the American Statistical Association*, 97: 337-351.
15. S. L. Scott, G. M. James, and C. A. Sugar (2005). "Hidden Markov Models for Longitudinal Comparisons," *Journal of the American Statistical Association*, 100: 359-369.
16. M. A. Tanner (1996). "Tools for Statistical inference, (third edition)," *Springer*.