

palgrave  
macmillan



---

Estimating a Markov Transition Matrix from Observational Data

Author(s): Chris Sherlaw-Johnson, Steve Gallivan, Jim Burridge

Source: *The Journal of the Operational Research Society*, Vol. 46, No. 3 (Mar., 1995), pp. 405-410

Published by: Palgrave Macmillan Journals on behalf of the Operational Research Society

Stable URL: <http://www.jstor.org/stable/2584334>

Accessed: 07/03/2010 00:35

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=pal>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).



Palgrave Macmillan Journals and Operational Research Society are collaborating with JSTOR to digitize, preserve and extend access to *The Journal of the Operational Research Society*.

<http://www.jstor.org>



# Estimating a Markov Transition Matrix from Observational Data

CHRIS SHERLAW-JOHNSON<sup>1</sup>, STEVE GALLIVAN<sup>1</sup> and JIM BURRIDGE<sup>2</sup>

<sup>1</sup>Clinical Operational Research Unit and <sup>2</sup>Department of Statistical Science, University College London

Markov chains are frequently used in Operational Research to describe how a system changes over time, its behaviour being governed by its transition matrix. This paper describes a technique for finding a maximum likelihood estimate for such a transition matrix when a system is observed at infrequent time intervals. The technique is called the EM Algorithm which, for this kind of problem, has distinct advantages over other methods of optimization.

*Key words:* Markov chains, maximum likelihood estimation, EM algorithm

## INTRODUCTION

Markov chains play a central role in Operational Research, frequently being used to describe how a system changes over time. If a system can be adequately modelled as a Markov chain then numerous theoretical consequences can be applied to the analysis of the system<sup>1</sup>.

The behaviour of a Markov Chain depends on the values used in the transition matrix, which specifies the probabilities that the system moves from one state to another in unit time. Standard texts assume that the values of such transition matrices are known. However, in most practical studies, this is not the case and the transition matrix needs to be estimated.

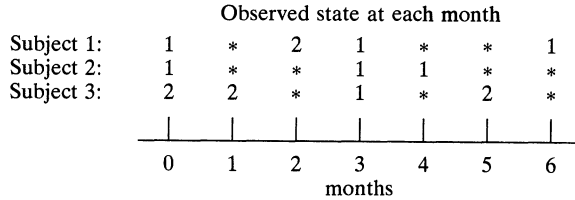
One way of doing such estimation is to use data concerning the observed state of the system at successive time points. If successive observations are all the same interval apart then estimating the transition matrix is straightforward. Unfortunately, however, the practitioner is often faced with problems in which a system has been observed infrequently, where times between successive observations vary. With a large amount of such variation, estimating the transition matrix becomes more complex.

For example, this study arose from work carried out by the authors investigating the progress of coronary artery disease<sup>2</sup>. We were concerned with modelling the progress of patients' symptoms over many years. Symptoms were assessed on a five-point scale by clinicians whenever the patients attended an outpatients' clinic. A preliminary model was based on the assumption that the patient group was homogeneous and that the progression of symptom states formed a Markov chain. However, there was considerable variation in the times between successive visits, ranging from one week to over a year. In this paper we show how to find a maximum likelihood estimate for a Markov transition matrix in such cases, when sequences of observations are irregularly spaced.

## ESTIMATING THE TRANSITION MATRIX FROM IRREGULARLY SPACED SEQUENCES OF OBSERVATIONS

As a simple example, consider the case where there are just two states (1 and 2), and infrequent observations have been made on three subjects over 6 months. The results of the observations are illustrated in Figure 1.

The first subject is initially observed and found to be in state 1. After two months the subject is observed **again** and found to be in state 2. A third observation is made a month later and so on. Such **data** sets are called 'incomplete', as opposed to 'complete' data in which



An asterisk refers to the absence of data at the particular time point.

FIG. 1. Typical incomplete data for three subjects.

the state of a subject is known at all time units between the first and last observations of the subject.

Suppose such a system is modelled as a Markov chain with transition matrix,  $T$ . Let  $O_{ijt}$  denote the number of observed transitions from state  $i$  to state  $j$  occurring over  $t$  time units and  $(T^t)_{ij}$  the  $ij$ th component of the matrix  $T^t$ , (the probability of a subject in state  $i$  being in state  $j$  after  $t$  time units), then the likelihood of the observed data  $Y$  given  $T$  is:

$$g(Y|T) = \prod_i \prod_j \prod_t ((T^t)_{ij})^{O_{ijt}} \tag{1}$$

The log likelihood is given by:

$$\log g(Y|T) = \sum_i \sum_j \sum_t O_{ijt} \log (T^t)_{ij}.$$

This log likelihood function is too complex to be maximized analytically; instead an iterative optimization technique is required. One such technique is the EM Algorithm<sup>3,4</sup>.

The EM Algorithm has two stages, the E-step and the M-step, the essences of which are as follows. A parameter, in this case the transition matrix, is assigned an initial value. Using this estimate for the parameter, the E-step reconstructs a ‘complete’ set of data,  $X$  from the ‘incomplete’, observed data  $Y$ . The M-step then uses the reconstructed, complete data to obtain a new estimate of the parameter. With this new estimate the E-step is then repeated, followed by the M-step again to obtain a third estimate. Iteration continues until successive estimates of the parameter appear to converge. The method is particularly suited to the maximum likelihood problem being considered since both steps involve easy computations. We now describe the algorithm in detail.

First note that, given the observed data  $Y$  and a transition matrix  $T$ , there are several possible sets of ‘complete’ data  $X$  each with an associated probability of occurring. The likelihood of such a set  $X$  given the transition matrix  $T$  is given by:

$$f(X|T) = \prod_i \prod_j (T_{ij})^{N_{ij}}$$

where  $N_{ij}$  is the number of transitions from state  $i$  to state  $j$  in one time unit occurring within the ‘complete’ data,  $X$ , so that  $X = \{N_{ij}\}$ .

Define  $Q(T|T')$  as the expected value, given  $T'$  and the observed data  $Y$ , of the log likelihood of the ‘complete’ data with transition matrix  $T$ , i.e.:

$$Q(T|T') = E[\log f(X|T)|Y, T'] \tag{2}$$

The EM Algorithm operates as follows.

Step 1. Find a suitable starting matrix  $T^{(0)}$ .

Step 2. E-STEP (expectation): for  $p \geq 0$  compute  $Q(T|T^{(p)})$ .

Step 3. M-STEP (maximization): choose  $T^{(p+1)}$  as a global maximum for  $Q(T|T^{(p)})$ .

Step 4. Repeat the E and M-steps until the sequence  $\{T^{(p)}\}$  converges; its limit will be a maximum likelihood estimate for  $T$  given the observed data  $Y$ .

We now derive an expression for  $Q(\mathbf{T}|\mathbf{T}^{(p)})$  in the Markov case, from which it is straightforward to derive its global maximum.

If  $S_{ij}(\mathbf{T})$  is defined as the expected number of transitions from state  $i$  to state  $j$  occurring within the ‘complete’ data, given the observed data  $Y$ , i.e.:

$$S_{ij}(\mathbf{T}) = E[N_{ij}|Y, \mathbf{T}] \tag{3}$$

then, from (2):

$$\begin{aligned} Q(\mathbf{T}|\mathbf{T}^{(p)}) &= E[\log f(X|\mathbf{T})|Y, \mathbf{T}^{(p)}] \\ &= E\left[\sum_i \sum_j N_{ij} \log(T_{ij})|Y, \mathbf{T}^{(p)}\right] \\ &= \sum_i \sum_j \log(T_{ij}) E[N_{ij}|Y, \mathbf{T}^{(p)}] \\ &= \sum_i \sum_j S_{ij}(\mathbf{T}^{(p)}) \log(T_{ij}). \end{aligned} \tag{4}$$

This can be shown to have a unique global maximum with the matrix  $\mathbf{T}$  having elements:

$$T_{ij} = \frac{S_{ij}(\mathbf{T}^{(p)})}{\sum_k S_{ik}(\mathbf{T}^{(p)})} \left( \sum_k S_{ik} \neq 0 \right). \tag{5}$$

We now show how to calculate  $S_{ij}(\mathbf{T})$ .

Suppose the system is observed to be in state  $m$  at time  $t_0$  and, when next observed,  $t$  time units later, is in state  $n$ . The probability that a transition between states  $i$  and  $j$  occurs at time  $t_0 + k$ , where  $0 \leq k \leq t - 1$  is given by:

$$P_{ijk,mnt} = \frac{(T^k)_{mi} T_{ij} (T^{t-k-1})_{jn}}{(T^t)_{mn}} \tag{6}$$

with the convention that  $T^0 = I$ , the identity matrix.

It follows that the expected number of transitions between states  $i$  and  $j$  occurring within the observed time interval, given a single transition from  $m$  to  $n$  over  $t$  time units, is given by:

$$\sum_{k=0}^{t-1} P_{ijk,mnt}. \tag{7}$$

Therefore the expected number for all such time intervals is given by:

$$O_{mnt} \sum_{k=1}^{t-1} P_{ijk,mnt}.$$

Hence, summing over all the observed data:

$$S_{ij}(\mathbf{T}) = \sum_m \sum_n \sum_t O_{mnt} \sum_{k=0}^{t-1} P_{ijk,mnt}. \tag{8}$$

So, in practice, the algorithm can be implemented as follows.

E-Step. For each  $i$  and  $j$ , compute  $S_{ij}(\mathbf{T}^{(p)})$  using (6) and (8).

M-Step. Compute  $\mathbf{T}^{(p+1)}$  as the matrix whose elements are defined by (5).

It is shown in the Appendix, with reference to the paper by Dempster *et al.*<sup>3</sup>, that the sequence of successive log likelihoods,  $\{\log g(Y|\mathbf{T}^{(p)})\}$ , converges and that if  $\{\mathbf{T}^{(p)}\}$  also converges then its limit is either a local maximum or saddle point for the likelihood function. Conditions under which the sequence  $\{\mathbf{T}^{(p)}\}$  does converge are also discussed in the Appendix.

## PRACTICAL CONSIDERATIONS

There are a number of practicalities which should be taken into account when applying the EM Algorithm in the way described in the previous section.

The starting matrix,  $T^{(0)}$ , can be chosen arbitrarily. However, if one of its elements is zero, ( $T_{kl}^{(0)}$ , say), and also  $O_{kl1} = 0$ , then, from above,  $S_{kl}(T^{(0)}) = 0$  and  $T_{kl}^{(p)} = 0$  for all  $p > 0$ . Hence, the corresponding elements will be zero in all the generated matrices. If this constraint is to be avoided then the starting matrix should contain no zero elements.

All observations of the same subject are assumed to be whole numbers of time units apart. In practice, to meet this assumption, time intervals may have to be rounded to the nearest whole number of time units. What effect this rounding has on the final result is unclear.

The larger the number of time units between successive observations, the higher the powers of the transition matrix that need to be computed during the E-step. This leads to slower computation time. If there is a relatively small number of observed transitions a large interval apart, then the benefit of including them in the analysis may not be enough to justify the extra computation time. In practice, it is preferable to set a maximum time interval above which observed transitions are ignored. If this maximum time interval is still very large then it may be sensible to increase the duration regarded as unit time.

In computation, convergence is assumed to occur when the maximum difference between the terms of successive matrices is less than a pre-specified value,  $\varepsilon$ , ( $\varepsilon = 10^{-4}$ , say) i.e. given  $p$ :

$$|T_{ij}^{(p+1)} - T_{ij}^{(p)}| < \varepsilon \quad \text{for all } i, j$$

Even when the algorithm can be shown to converge, (see the Appendix), it is difficult to find the limit because, given  $i$  and  $j$ , the sequence  $\{|T_{ij}^{(p+1)} - T_{ij}^{(p)}|\}$  is not monotonic. It is only possible to conclude apparent convergence.

If the algorithm does converge then its limit is either a local maximum or a saddle point of the likelihood function. However, it is possible that more than one of these stationary values exist and, as with all such algorithms, there is no guarantee that the limit will be the global maximum. Therefore, the algorithm should be repeated with a variety of starting matrices to see whether higher maxima can be found.

In deriving the expression for the likelihood function,  $g(Y|T)$ , and the equation for computing  $S_{ij}(T)$ , (equation (8)), we have assumed that the exact time a subject enters each state is unknown. In general, this may not be the case. There may be an absorbing state of 'dysfunction' or 'death', say, for which the times of entry are known with some precision. In this situation, the likelihood function and the equation for computing  $S_{ij}(T)$  would be slightly different but the remaining analysis would remain the same.

For example, suppose a subject was observed to be in state  $m$  and is no longer observed until it dies  $t$  time units later. At time  $t - 1$  the subject would be still alive, so, given this information, the log likelihood of a transition matrix,  $T$ , is given by:

$$\log \sum_{k \neq D} (T^{t-1})_{ik} T_{kD}$$

where  $D$  is the death state.

The full log likelihood of observed data  $Y$  is then given by:

$$\log g(Y|T) = \sum_i \sum_{j \neq D} \sum_t N_{ijt} \log (T^t)_{ij} + \sum_i \sum_t N_{iDt} \log \sum_{k \neq D} (T^{t-1})_{ik} T_{kD}.$$

## CONCLUSIONS

This paper is concerned with systems that are assumed to be Markov chains. By observing how the system changes over time it is possible to estimate an appropriate transition matrix.

When the observations are irregularly spaced, a maximum likelihood estimate can be found

by the EM Algorithm. Given a matrix to start the algorithm, the elements of the next matrix are proportional to the expected frequency of moving from one state to the other. The matrix after that is generated from the former in the same way, and so on until the sequence of matrices appears to converge.

The EM Algorithm has become a popular technique in applied statistics over the past 10 years<sup>5,6,7</sup>. Compared with other optimization algorithms, which could be used in its place, it is readily applicable to maximum likelihood estimation and, unlike methods which involve the computation of derivatives, is easy to implement on a computer. If the maximum time interval and the number of states are fairly small then each iteration takes very little computation time; however, a large number of iterations may be required before convergence.

### APPENDIX

In this appendix we demonstrate that if the algorithm, described in this paper, generates a sequence of transition matrices  $\{T^{(p)}\}$  then the sequence of likelihood functions  $\{L(T^{(p)})\}$  converges. We also discuss the convergence of  $\{T^{(p)}\}$  and show that, in cases when it does converge, its limit is either a local maximum or a saddle point for the likelihood function. The theorems are derived from those stated for the general case by Dempster *et al*<sup>3</sup>. The reader is referred to this paper or the book by Little and Rubin<sup>4</sup> for proofs.

Put  $L(T) = \log g(Y|T)$ , with  $g(Y|T)$  as defined in the paper.

#### Theorem 1

If  $\{T^{(p)}\}$  is a sequence generated by an EM algorithm then  $L(T^{(p+1)}) \geq L(T^{(p)})$  for all  $p > 0$ .

The sequence  $\{L(T^{(p)})\}$  is bounded above by 0 and hence as a corollary to Theorem 1 it must converge.

#### Theorem 2

Suppose  $\{T^{(p)}\}$  is a sequence generated by an EM algorithm such that:

- (i) the sequence  $\{L(T^{(p)})\}$  is bounded, and
- (ii) there exists  $\lambda > 0$  such that

$$\frac{S_{ij}(T^{(p)})}{(T_{ij}^{(p)})^2} > \lambda \quad \text{and} \quad \frac{S_{ij}(T^{(p)})}{(T_{ij}^{(p+1)})^2} > \lambda$$

for all  $i, j$  and  $p$  where  $T_{ij}^{(p)}$  and  $T_{ij}^{(p+1)} \neq 0$ .

The sequence  $\{T^{(p)}\}$  then converges to some transition matrix  $T^*$  which is a local maximum or saddle point for  $L(T)$ .

Condition (ii) is sufficient, but not necessary, for convergence. If it is not satisfied the algorithm may still converge.

One situation in which condition (ii) holds is clear from (8), since  $S_{ij}(T^{(p)}) \geq O_{ij}$  for all  $p$ . Hence, the condition is satisfied if  $O_{ij} > 0$  for all  $i$  and  $j$ , i.e. if the observations over one time unit include at least one transition between each pair of states.

For a more general situation, define  $M_{ij}$  as the transition from state  $i$  to state  $j$  over one time unit, and  $Z = \{M_{ij} | M_{ij} \text{ is observed in the data}\}$ . It can easily be verified that condition (ii) will then be satisfied if:

- (a) at least one transition is observed from each state, and
- (b) if  $O_{ij} = 0$  then there exists states  $m$  and  $n$  and  $t > 1$  with  $O_{mnt} > 0$  such that the system can change from state  $m$  to  $n$  with a combination of  $t - 1$  elements of  $Z$  and the unobserved transition  $M_{ij}$ .

Even more general situations under which condition (ii) holds may exist but are yet to be established by the authors.

## REFERENCES

1. D. R. COX and H. D. MILLER (1965, reprinted 1978) *The Theory of Stochastic Processes*. Chapman and Hall, London.
2. J. MITCHARD, S. GALLIVAN, D. L. H. PATTERSON, T. TREASURE and R. R. P. JACKSON (1990) Modelling the progress of coronary artery disease. Clinical Operational Research Unit: Research Paper 167.
3. A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* **39**, 1–38.
4. R. J. A. LITTLE and D. B. RUBIN (1987) *Statistical Analysis with Missing Data*. Wiley, New York.
5. B. W. SILVERMAN, M. C. JONES, J. D. WILSON and D. W. NYCHKA (1990) A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission topography. *J. Roy. Statist. Soc. B* **52**, 271–324.
6. R. D. BOCK and M. AITKIN (1981) Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* **46**, 443–459.
7. C. FUCHS and J. B. GREENHOUSE (1988) The EM algorithm for maximum likelihood estimation in the mover–stayer model. *Biometrics* **44**, 605–613.

*Received August 1993; accepted August 1994 after one revision*