An introduction to hidden Markov models

J.K. Lindsey
Medical Statistics, De Montfort University, Leicester

1 Introduction

Suppose that a sequence of responses is discrete-valued, often categories that would *appear* to be the observed states of some Markov chain. However, dependence cannot adequately be described by the simple Markov property. In a hidden Markov model, an underlying, *unobserved* sequence of states follows a Markov chain, the hidden state determining the probabilities of the observed states. Such an approach is widely used in speech processing and in biological sequence analysis of nucleic acids in DNA and of amino acids in proteins.

For a binary time series, each event might be generated by one of two Bernoulli distributions. The process switches from the one to the other according to the state of the hidden Markov chain, in this way generating state dependence. Analogous models can be constructed for other discrete distributions, such as the Poisson or binomial distributions.

2 The model

Consider an irreducible homogeneous Markov chain with $M \times M$ transition matrix, $\mathbf{T}$. This gives the probabilities of changing among the hidden states, with marginal stationary distribution, $\boldsymbol{\pi}$. The latter can be calculated from the transition matrix and hence does not introduce any new parameters. Then, the probability of the observed response at time t, $\nu_{mt} = \Pr(y_t|m;\boldsymbol{\kappa}_m)$, will depend on the unobserved state, m, at that time.

The series of responses on a given unit are assumed to be independent, given the hidden state. Thus, there are $M(M-1)$ unknown parameters in the transition matrix as well as M times the length of $\boldsymbol{\kappa}_m$ in the probability distributions. Although the probability of the observed data is complex, it can be written in a recursive form over time:

$$f(\mathbf{y};\boldsymbol{\kappa},\mathbf{T}) = \boldsymbol{\pi}^T \prod_{t=1}^{R} (\mathbf{T}\mathbf{F}_t)\mathbf{J}^T$$

$\mathbf{F}_t$ is an $M \times M$ diagonal matrix containing, on the diagonal, the probabilities, ν_{mt}, of the observed data given the various possible states.

To construct the likelihood function from this, first calculate the marginal probability times the observed probability for each state at time 1, say $a_m = \pi_m \Pr(y_1|m;\boldsymbol{\kappa}_m)$. At the second time point, the first step is to calculate the observed probability for each state multiplied by this quantity and by the transition probabilities in the corresponding column of $\mathbf{T}$. These are summed yielding, say $b_m = \sum_h a_m T_{mh} \Pr(y_2|h;\boldsymbol{\kappa}_h)$. This is the new vector of forward probabilities, but, to prevent underflow, it is divided by its average, yielding a new vector, $\mathbf{a}$. This average is also cumulated as a correction to the likelihood function.

These steps are repeated at each successive time point. Finally, the sum of these a_m at the last time point is the likelihood except that the cumulative correction must be added to it. At each step, the vector, $\mathbf{a}$, divided by its sum gives the (filtered) conditional probabilities of being in the various possible states given the previous observations. This model can be applied in continuous time by using a matrix of transition *intensities*, so that, say, $\mathbf{C}$ has rows summing to zero (instead of one). Then matrix exponentiation is applied to give the transition probabilities $\mathbf{T}_{\Delta t} = \mathrm{e}^{\Delta t \mathbf{C}}$, where Δt is the time interval between observations.

3 Examples

3.1 Locust Behaviour

To investigate the effect of hunger on locomotory behaviour, 24 locusts three days into the fifth larval stage were placed individually in glass observation chambers. Even numbered subjects were not fed for 5.5 hours whereas odd numbered subjects received as much food as they could eat. During subsequent observation, neither food nor water were available. 161 observations, at 30 second intervals, were made on each animal. At each time point, the locust was classified either as locomoting (1) or not (0), the latter including quiescence and grooming.

The question is how locomotory behaviour evolves over time and whether it differs between the two treatment groups. There are 144 locomoting events in the fed group but 973 in the unfed group, each in 1932 observation intervals. As well, there is great individual variability among the locusts.

	Indep	Markov	HMM
	Common intercept		
Null	2324.5	1869.0	1599.1
Trend	2263.2	1849.2	1591.6
	Treatment intercepts		
Null	2322.9	1869.2	1596.6
Same trend	2261.4	1849.1	1582.7
Diff trend	2248.2	1844.4	1583.2
	Individual intercepts		
Null	1667.6	1577.6	1494.0
Same trend	1575.6	1521.6	1492.1
Diff trend	1566.5	1513.5	1488.3

The trends, common to the two states, are estimated to be 0.0140 in the fed group and 0.0048 in the unfed group. Locomotory behaviour increases faster in the former group, perhaps because it stays rather stable throughout the observation period for the unfed group. For the fed group, the probability of locomotion is small in both states, whereas, for the unfed group, there is a clear distinction between the two states, one of them indicating higher locomotory behaviour. Indeed, for the fed group, a simple Markov chain fits better: the AIC is 386.7 with 14 parameters as compared to 389.4 with 27 parameters for the hidden Markov model.

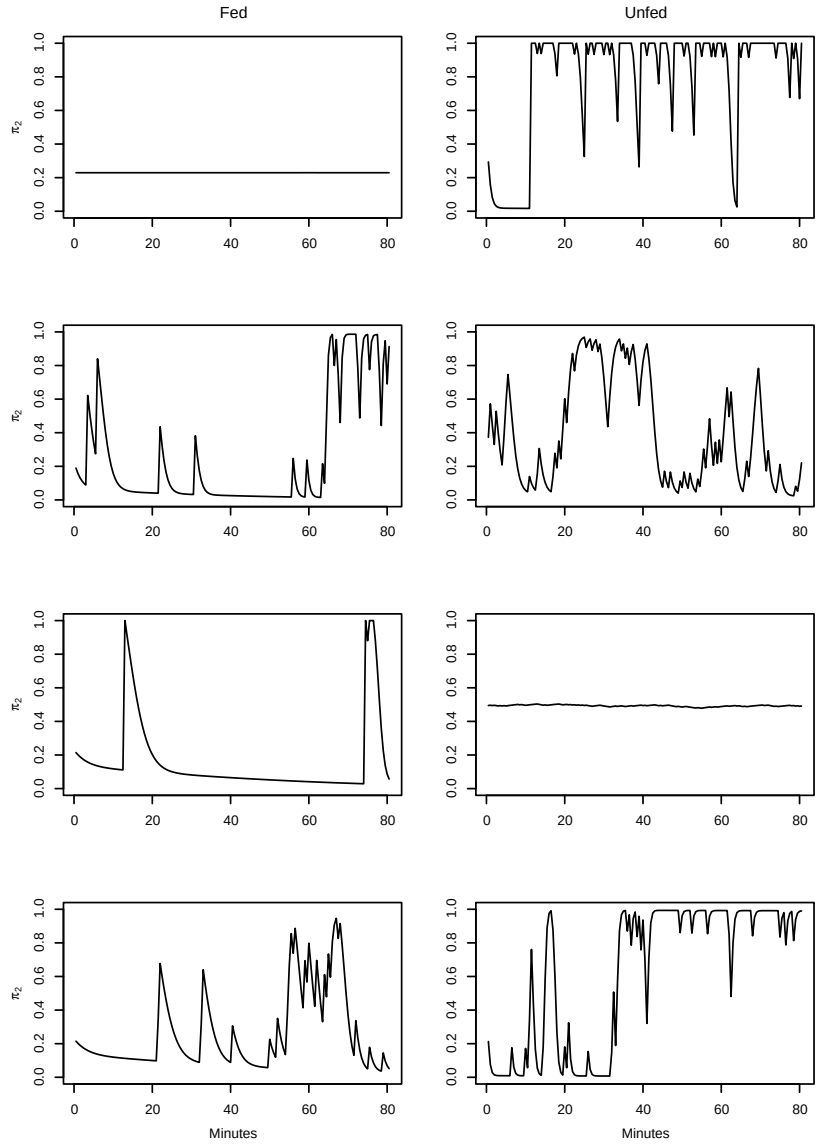
The hidden transition matrices are estimated respectively to be

$$\mathbf{T} = \begin{pmatrix} 0.978 & 0.022 \\ 0.073 & 0.927 \end{pmatrix}$$

and

$$\mathbf{T} = \begin{pmatrix} 0.975 & 0.025 \\ 0.025 & 0.975 \end{pmatrix}$$

in the fed and unfed groups with respective stationary distributions (0.77, 0.23) and $(0.51, 0.49)$. The large diagonal values indicate that the locusts tend to remain a relatively long time in the same state, a spell, inducing a dependence among consecutive observations. The fed group stays about three-quarters of the time in the first state (but recall that two states are not required) whereas the unfed group spends about half the time in each state.



3.2 DNA Sequence Analysis

The double-stranded helical form of DNA is well known. Each strand consists of a linear sequence of the four nucleic acid bases, adenine (`A`), cytosine (`C`), guanine (`G`), and thymine (`T`). Opposite strands contain complementary pairs: `A` with `T` and `C` with `G` so that only one of the strands need by studied. In a gene, consecutive, non-overlapping triplets of bases code corresponding sequences consisting of the twenty different amino acids that make up a protein.

Because there are 64 possible combinations of the bases, the code is redundant, particularly in the third base, with several triplets often coding the same amino acid.

Most bases in a DNA sequence do not code for proteins. Only selective sections of the strands are actually active. In addition, the bases coding a given protein are not necessarily all consecutive but may be split into several sections. These are called the *exons* of the gene whereas the non-coding sections in between are called *introns*. Because the set of exons define a protein, they are subject to natural selection; one may expect the bases in the introns to be more random. A mutation in an exon sequence will often result in a code for a non-viable or inappropriate protein, whereas a mutation in an intron does not have this harmful effect.

DNA sequences coding similar proteins must be similar. This will be true of two proteins in the same organism but also of those in two closely related organisms. On the other hand, the non-coding sequences may differ widely. In order to compare such sequences, the DNA must be aligned. To do this optimally, gaps may have to be left in some of the sequences, where breaks may have occurred during mutations in the evolutionary process.

Consider four such aligned sequences, for coding two closely-related proteins (α_1- and θ_1-globin) in two primates (the orang-utan, *Pongo pygmeus*, and the olive baboon, *Papio anubis*).

```
CCAATGAGCG CCGCCCGGCC GGGCGTGCCC CTGCGCCCCA AGCATAAA++
CCAATGAGCA CCGCCCTGCC GGGCGTGCCC CCGCGCCCGG AGCATAAA++
CCAATTTTTG TGTTTTTAGT AGAGACTAAA AACCATATGG TGAACACCTA
CCAATTTTTG TGTTTTTAGT AGAGACTAAA AACTATATGG TGAACACCTA

++++++++++ ++++++++++ ++++++++++ ++++++++++ ++++++++++
++++++++++ ++++++++++ ++++++++++ ++++++++++ ++++++++++
AGACG+GGGG GCCTTGGATC CAGGGCAATT CAGAGGGCCC CCGGTCGGAG
AGACGCGGGG GCCTTGGATC CAGGGCGATT CAGAGTTCCC CCGGTCGGAG

++++++++++ ++++++++++ ++++++++++ ++++++++++ ++++++++++
++++++++++ ++++++++++ ++++++++++ ++++++++++ ++++++++++
CTGTCGGAGA TGGAGCGCGC GCGCTCCCGG GATCCCGGAC GAGGCCCTGC
CAGTCGGAGA TGGAGGCCGC GCGGTCCCGG GATCCGGGAC CAGGCCCTGC

++++++++++ ++++++++++ ++++++++++ ++++++++++ ++++++++++
++++++++++ ++++++++++ ++++++++++ ++++++++++ ++++++++++
GCCCCAGGGC GGCGAGGCTG CAGCGCGGCG CCCCCTGGAG GCCGCGGGAC
ACCCCAGGGT GGCGAGGCTG CAGCGCGGCG CCCCCTGGTG GCCGCGGGAC

CCCTGGCGCG CTCGCGGCCC CGCACTCTTC TGGTCCCCAC AGACTCAGAA
CCCTGGCGCG CTCGCGGCCC GGCACTCTTC TGGTCCCCAC AGACTCAGAA
CCCTAGCCGG TCCGCGCAGG CGCGGCGGGG ACGCAGGGCG CGGCGGGTTC
CCCTGGCCGG TCCGCGCAGG CGCAGCGCGG GCGCAGGGCG CGGCGGGTTC
```

```
AGAACCCACC ATG GTG CTG TCT CCT GCC GAC AAG ACC AAC GTC
AGAACCCACC ATG GTG CTG TCT CCT GAC GAC AAG AAA CAC GTC
CAGCGCGGGG ATG GCG CTG TCC GCG GAG GAC CGG GCG CTG GTG
CAGCGCGGGA ATG GCG CTG TCC GCG GAG GAC CGG GCG CTG GTG

AAG ACC GCC TGG GGG AAG GTC GGC GCG CAC GCC GGC GAC
AAG GCC GCC TGG GGT AAG GTC GGC GAG CAC GCT GGC GAG
CGC GCC CTG TGG AAG AAG CTG GGC AGC AAC GTC GGC GTC
CGC GCC CTG TGG AAG AAA CTG GGA AGC AAT GTT GGC GTC

TAT GGT GCG GAG GCC CTG GAG AG GTGAGGCTCC CTCCCCTGCT
TAT GGT GCG GAG GCC CTG GAG AG GTGAGGCTCC CTCCCCTGCT
TAC ACG ACA GAG GCC CTG GAG AG GTGCGGC+++ +GAGGCTGGG
TAT GCT ACT GAG GCC CTG GAG AG GTGCGGC+++ +GAGGCTGGG
```

The total length of the three exons is 429 bases for each of the four genes, whereas the total length of the two introns is 260 bases. The relative frequencies are

	A	C	G	T
Exon	0.17	0.36	0.29	0.17
Intron	0.10	0.40	0.36	0.14

Let us fit hidden Markov models with various numbers of states, separately to the complete sequences of exons and of introns. For the moment, for the exons, these models will not take into account the triplet structure of the coding.

Model		Exons	Introns
Independence		2290.2	1109.0
Markov chain		2236.4	1058.6
Hidden Markov chain			
States	2	2287.2	1087.0
	3	2268.1	1085.4
	4	2250.5	1075.0
	5	2227.5	1075.7
	6	2207.3	

The hidden Markov models provide an improvement over the ordinary Markov chain only for the coding sequences. However, the intron base sequences are not completely random because the Markov chain fits better than the independence model. The six-state hidden Markov model contains 48 parameters but, for the exons, it can be greatly simplified by setting a number of the probabilities to zero.

These models do not take into account the fact that the coding sections of a DNA sequence define the amino acids in the corresponding protein by triplets of

nucleic acids. Let us then modify the hidden Markov model by allowing state-dependent probabilities of the four nucleic acid bases to be different for each of the three positions in a triplet. Now, only a four-state model is required, with a much improved AIC of 2032.4, containing 24 parameters after setting various probabilities to zero. There is little indication of difference between the two globin types (AIC of 2031.0 with 48 parameters) and none for difference among all four sequences (AIC of 2067.9).

The equivalent ordinary Markov chain, with a different marginal probability distribution at each of the three positions but the same transition matrix, has an AIC of 2091.9 with 21 parameters. The hidden transition matrix is estimated to be

$$\begin{pmatrix} 0.41 & 0 & 0 & 0.59 \\ 0.26 & 0.74 & 0 & 0 \\ 0.12 & 0.62 & 0.26 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

with stationary distribution (0.23, 0.45, 0.18, 0.14).

For the first position, the probabilities of the four bases, when in a given state, are

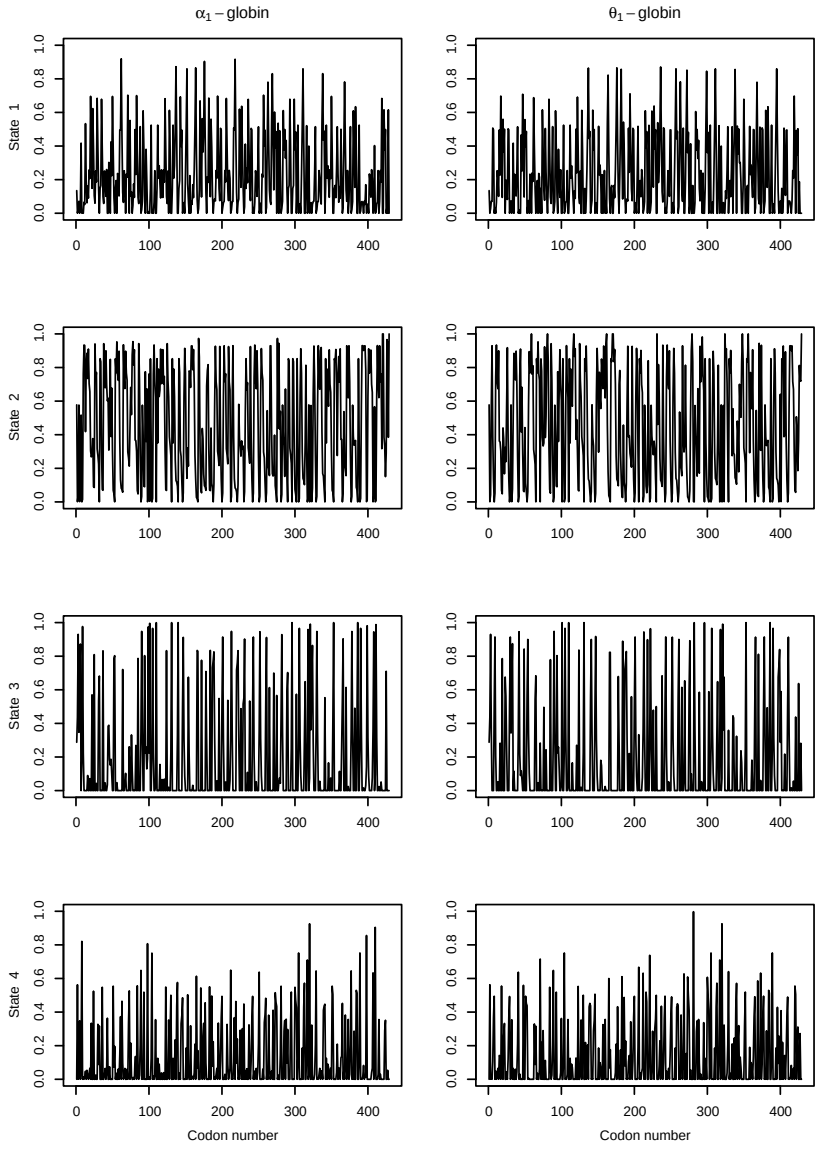
State	A	C	G	T
1	0.11	0.67	0.22	0
2	0.25	0.20	0.39	0.17
3	0.30	0	0.58	0.12
4	0	0.51	0.13	0.35

For the second position, they are

State	A	C	G	T
1	0.87	0.13	0	0
2	0.19	0.57	0.24	0
3	0	0	0.19	0.81
4	0	0	0	1

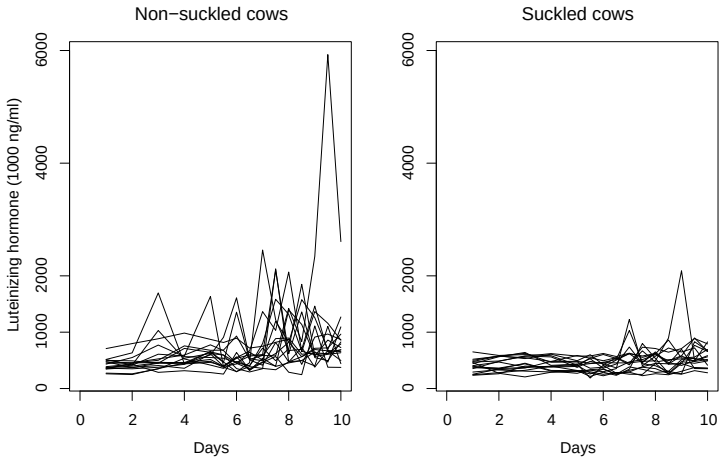
and, for the third position,

State	A	C	G	T
1	0	0.59	0.39	0.02
2	0.10	0.64	0.11	0.14
3	0	0	1	0
4	0	0.45	0.51	0.05



3.3 Luteinizing Hormone Levels

Thirty-two cows divided into two groups, suckled and non-suckled, were followed for ten days post-partum. Their levels of luteinizing hormone (ng/ml×1000) were measured 15 times at unequally-spaced intervals. Concentrations of luteinizing hormone are influenced by semi-periodic pulsing of the glands that produce the hormone.



Variability increases as time goes on with non-suckled cows show more variation than the suckled ones.

		HMM	
Distribution	Indep	Same	Different
Log normal	3270.5	3198.3	3174.1
Gamma	3314.6	3225.9	3195.0
Weibull	3387.1	3262.3	3261.9
Inverse Gauss	3264.8	3187.0	3176.7
Burr	3250.6	3160.4	3150.3

A different transition matrix is clearly necessary for suckled and for non-suckled cows.

The Burr distribution

$$f(y;\mu,\kappa,\nu) = \frac{\nu\kappa\left(\frac{y}{\mu}\right)^{\kappa-1}}{\mu^{\kappa}\left[1+\left(\frac{y}{\mu}\right)^{\kappa}\right]^{\nu+1}}$$

fits considerably better than the others.

For the non-suckled cows, the location parameters in the two hidden states are estimated to be $\hat{\mu} = 346.8$ and 553.4, whereas the other parameters are $\hat{\kappa} = 9.39$ and $\hat{\nu} = 0.30$ for both states. For the suckled cows, the corresponding values are $\hat{\mu} = 285.0$ and 475.1, and $\hat{\kappa} = 10.4$ and $\hat{\nu} = 0.50$.

The intensity transition matrices, and the corresponding (one-day) probability transition matrices, the latter obtained by matrix exponentiation, are respectively

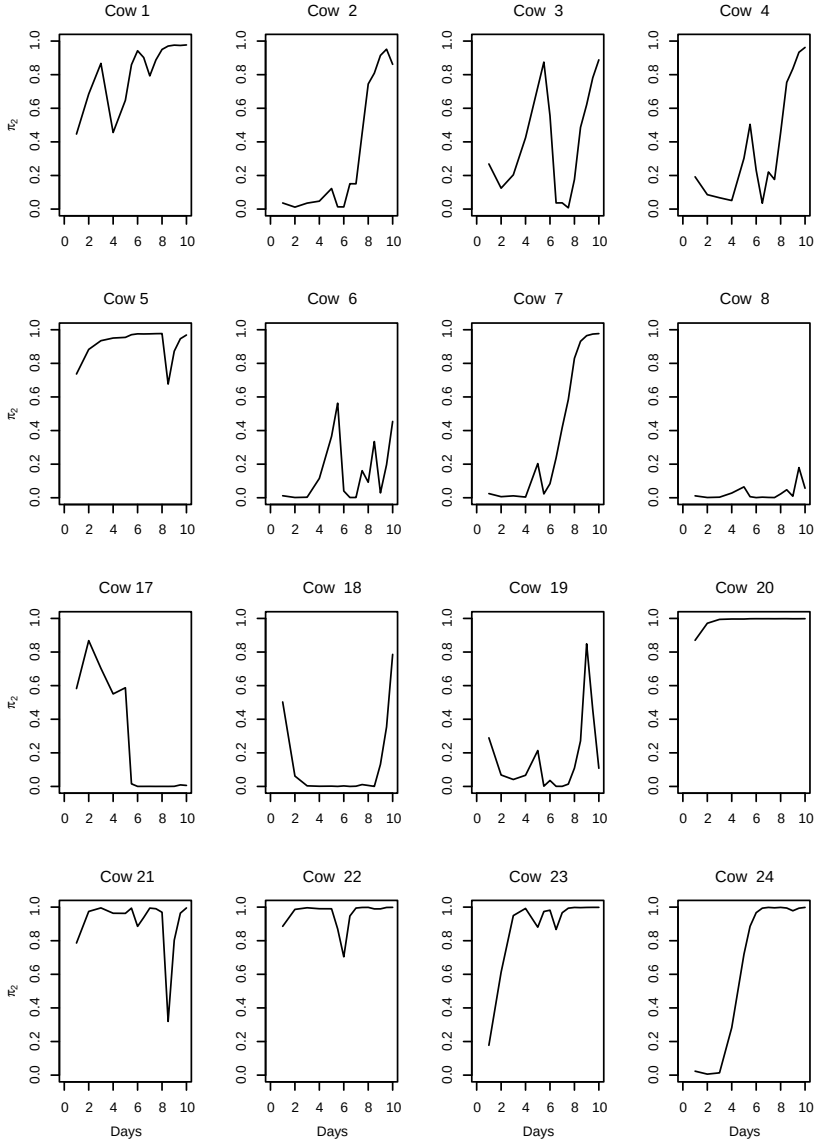
$$\mathbf{T} = \begin{pmatrix} -0.102 & 0.102 \\ 0.119 & -0.119 \end{pmatrix} \text{ and } \begin{pmatrix} 0.908 & 0.092 \\ 0.107 & 0.893 \end{pmatrix}$$

for the suckled cows and

$$\mathbf{T} = \begin{pmatrix} -0.0556 & 0.0556 \\ 0.0392 & -0.0392 \end{pmatrix} \text{ and } \begin{pmatrix} 0.947 & 0.053 \\ 0.037 & 0.963 \end{pmatrix}$$

for the non-suckled ones.

The stationary distributions are, respectively, (0.54, 0.46). and (0.41, 0.59).



Some of the cows, such as number 8 (and 10), have such low levels of hormone that they probably stay primarily in the lower state most of the time, at least as compared to the other cows. Others, such as 1, 5, 20, 22 seem to stay in the high state. However, most cows switch between the two states during the period of observation, with increasing probability of being in the high state as time passes. However, the two states have somewhat different levels in the two groups.

4 Selected References

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids.* Cambridge: Cambridge University Press.

Elliot, R.J., Aggoun, L., and Moore, J.B. (1995) *Hidden Markov Models.* Berlin: Springer-Verlag.

Hamilton, J.D. (1990) Analysis of time series subject to changes in regime. *Journal of Econometrics* **45**, 39–70.

Juang, B.H. and Rabiner, L.R. (1991) Hidden Markov models for speech recognition. *Technometrics* **33**, 251–272.

MacDonald, I.L. and Zucchini, W. (1997) *Hidden Markov and other Models for Discrete-valued Time Series.* London: Chapman & Hall.

Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* **77**, 257–286.