STATISTICS IN MEDICINE
Statist. Med. **18**, 3463–3478 (1999)

MONITORING EPIDEMIOLOGIC SURVEILLANCE DATA USING HIDDEN MARKOV MODELS[†]

YANN LE STRAT[1,2]* AND FABRICE CARRAT[1]

[1] *Unité de Recherches 'Epidémiologie et Sciences de l'Information', Institut National de la Santé et de la Recherche Médicale, U444 Institut Fédératif Saint-Antoine de Recherche sur la Santé, 27, rue Chaligny, 75571 Paris, Cedex 12, France*
[2] *Service de Biostatistique et d'Informatique Médicale, Groupe Hospitalier Necker-Enfants malades, 149 rue de Sèvres 75743 Paris, Cedex 15, France*

SUMMARY

The analysis of routinely collected surveillance data is an important challenge in public health practice. We present a method based on a hidden Markov model for monitoring such time series. The model characterizes the sequence of measurements by assuming that its probability density function depends on the state of an underlying Markov chain. The parameter vector includes distribution parameters and transition probabilities between the states. Maximum likelihood estimates are obtained with a modified EM algorithm. Extensions are provided to take into account trend and seasonality in the data. The method is demonstrated on two examples: the first seeks to characterize influenza-like illness incidence rates with a mixture of Gaussian distributions, and the other, poliomyelitis counts with mixture of Poisson distributions. The results justify a wider use of this method for analysing surveillance data. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

The constant increase in the number and complexity of epidemiologic surveillance systems leads to ever more data collection and thus to massive databases that, to be exploited, require appropriate methods for synthesizing the information they contain. These data, collected by time and place, are nearly always presented in the form of epidemiologic indicators (incidence, prevalence) at regular time intervals (weeks, months). They are then analysed with statistical models. Among these, ARIMA (AutoRegressive Integrated Moving Average)[1–4] models are often used, for example, to predict an incidence rate that takes into account the serial correlation of the data. Nevertheless, using these models implies that transformation of the data leads to a stationary time series, for which a single underlying probability distribution is assumed. These two hypotheses are not necessarily true, however; the data may present abrupt and wide changes of magnitude as well as irregular periodicity, in situations such as epidemics, modifications of the case-definition, screening, or vaccination.

* Correspondence to: Yann Le Strat, Service de Biostatistique et d'Informatique Médicale, Hopital Necker-Enfants malades, 149, rue de Sèvres 75743 Paris, Cedex 15, France. E-mail: lestraty@necker.fr

[†] Presented at the International Society for Clinical Biostatistics, Seventeenth International Meeting, Budapest, Hungary, August 1996.

CCC 0277–6715/99/243463–16$17.50

Improved methods for the detection of 'unexpected' phenomena and non-stationary data have been developed. They are based on Bayesian algorithms[5-7] transposed from the domain of biomedicine to that of epidemiologic surveillance. The choice of the initial values of the model's parameters is difficult, however, since it must be based on prior knowledge of the phenomenon studied – knowledge that is not always available.

There are still other methods[8,9] that have been suggested and shown to be effective (for example, in the detection of epidemics),[10] but they have the disadvantage of requiring arbitrary assumptions about the nature of the phenomenon under surveillance.

We propose a method for monitoring surveillance data that is based on a hidden Markov model. The method assumes that the data are generated from a finite mixture of distributions governed by an underlying Markov chain. Separate states can thus be distinguished in the data, without arbitrary choices about either the number of transitions between states, or the instant when these transitions take place. Hidden Markov models have been used in many areas, including automatic speech recognition,[11-13] electrocardiographic signal analysis,[14] epileptic seizure frequency analysis,[15,16] DNA sequence analysis,[17] the modelling of neuron firing[18] and meteorology.[19] To our knowledge, only one study, which used as a mixture of binomial distributions, has applied this type of method to epidemiologic data.[20] We propose applying this model to other distributions (Gaussian, Poisson) particularly appropriate in the analysis of population surveillance data. The model is presented using an example – the characterization of the incidence rates of influenza-like illness.

2. INFLUENZA-LIKE ILLNESS DATA

Influenza-like illness (ILI) data were obtained from the Sentinelles Network,[21] a national computerized surveillance system involving 1 per cent ($n = 500$) of all general practitioners in France. Within the system, an ILI is defined as the combination of a sudden fever of at least 39°C with respiratory signs and myalgia. Weekly ILI incidence rates are computed taking into account the size of the underlying population and the representativeness of the participating physicians. We have considered the data collected between January 1985 and December 1996 (available at www.b3e.jussieu.fr/sentiweb/).[22] Figure 1 represents the time series of ILI weekly incidence rates. It shows a clear mixture of two dynamics: a low-level dynamic with incidence rates that vary according to a seasonal pattern (the non-epidemic dynamic) and a high-level dynamic, in which the incidence rate increases sharply at irregular intervals (the epidemic dynamic).

Of particular interest here is the timing of ILI epidemics. More generally, a basic question that we address involves the definition of an epidemic, usually defined as the occurrence of a number of cases of a disease, in a given period of time and in a given population, that exceeds the 'expected' number.[23] This definition thus assumes a mixture of two (or more) dynamics – one for the 'expected' number of cases, another for the 'excess' cases. Hidden Markov models provide the most natural way of making inferences about such phenomena, by assigning different probability distributions to the two dynamics.

3. HIDDEN MARKOV MODELS

We begin by assuming that the observations, y_t, $t = 1, \ldots, n$ are a realization of the stochastic process $Y = (Y_t; t = 1, \ldots, n)$. The basic idea of hidden Markov models is to associate with each Y_t an unobserved random variable S_t that determines the conditional distribution of Y_t; if $S_t = j$,

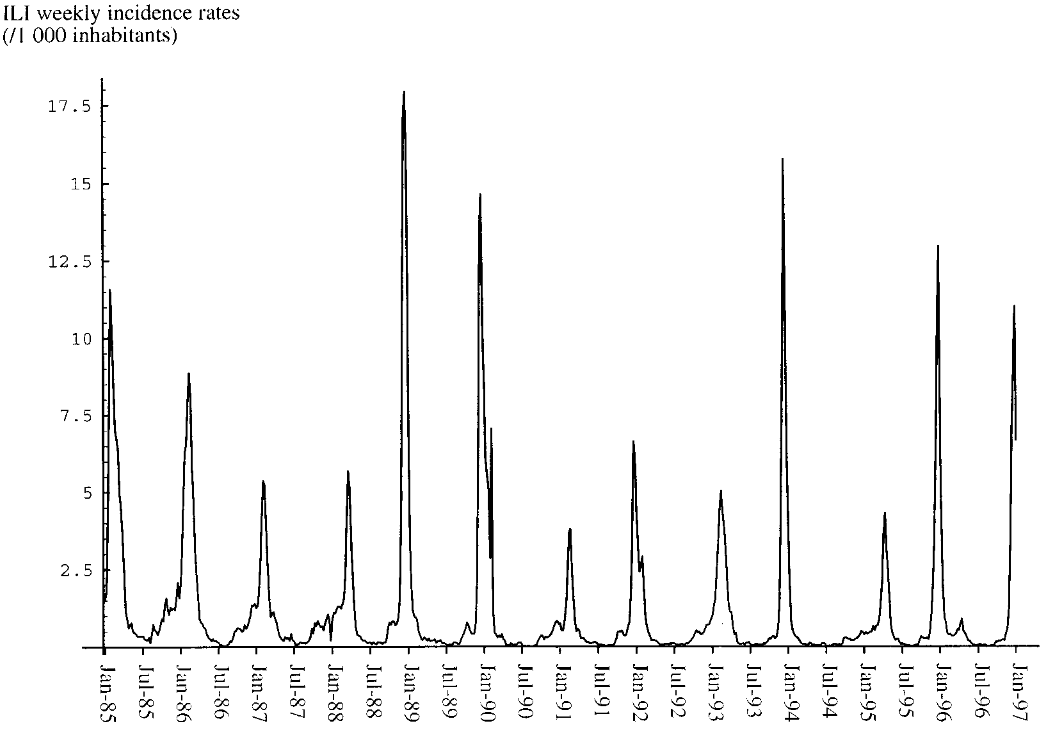

Figure 1. Weekly influenza-like illness (ILI) incidence rates, France, 1985–1996.

then the conditional distribution of Y_t has density $f_{j_t}(y_t; \theta_j)$, where f_{j_t} belongs to a given parameterized family, and θ_j are parameters to be estimated.

We assume that the unobserved sequence $\{S_t\}$ follows an m-state homogeneous Markov chain of order 1 with stationary transition probabilities

$$\alpha_{ij} = P(S_t = j \mid S_{t-1} = i), \qquad i, j = 1, \ldots, m.$$

Note that a stationary Markov chain is homogeneous, whereas the converse is not always true. We further assume that conditionally on the random variables S_t, the variables Y_t are independent. As the sequence of states is unobserved, this Markov-dependent mixture model is called a hidden Markov model.

To illustrate, suppose that there are two underlying distributions ($j = 1, 2$) associated with ILI incidence rates; y_t represents the observed ILI incidence rate at week t and is assumed to be generated from one of the two distributions, and α_{12} represents the probability of switching from distribution 1 to 2 in two successive weeks.

4. PARAMETER ESTIMATION

4.1. Likelihood computation

Let $\psi = (\theta_1, \ldots, \theta_m, \alpha_{11}, \ldots, \alpha_{mm})$ denote the complete parameter vector to be estimated. The likelihood function for ψ, given the data, is defined as the joint density of the random variables

$Y_1, \ldots, Y_n$ at $Y_1 = y_1, \ldots, Y_n = y_n$. We denote the initial probability of $\{S_t\}$ by $\alpha_{s_1} = P(S_1 = s_1)$. Given $S_1, \ldots, S_n$, the conditional density of $Y_1, \ldots, Y_n$ is

$$\prod_{t=1}^{n} f_{j_t}(y_t; \theta_j)$$

and the joint probability of $S_1 = s_1, \ldots, S_n = s_n$ is

$$\alpha_{s_1} \prod_{t=1}^{n} \alpha_{s_{t-1}s_t}.$$

The joint density of $S_1, Y_1, \ldots, S_n, Y_n$ at $Y_1 = y_1, S_1 = s_1, \ldots, Y_n = y_n, S_n = s_n$ is

$$\alpha_{s_1} f_{j_1}(y_1; \theta_j) \prod_{t=2}^{n} \alpha_{s_{t-1}s_t} f_{j_t}(y_t; \theta_j). \tag{1}$$

The likelihood function for ψ can thus be obtained by summing all possible sequences of states:

$$L_n(\psi) = \sum_{s_1=1}^{m} \cdots \sum_{s_n=1}^{m} \alpha_{s_1} f_{j_1}(y_1; \theta_j) \prod_{t=2}^{n} \alpha_{s_{t-1}s_t} f_{j_t}(y_t; \theta_j).$$

When the likelihood function is written in this form, the number of operations required for one evaluation is of the order of $2nm^n$,[11] which seems computationally intractable. It can, however, be rewritten so that one evaluation requires only of the order of nm^2 operations. This evaluation uses Baum's recursive method,[24] which computes the quantities $a_j(t) = f(y_1, \ldots, y_t, S_t = j)$ and evaluates the likelihood as $L_n(\psi) = \sum_{j=1}^{m} a_j(n)$ (see Appendix I).

The likelihood function can be maximized directly by numerical methods[19,20,25-27] or with an iterative scheme such as the EM algorithm.[14-16,28,29] Comparison of the performance of these various maximization methods has been discussed[26,27] and suggests that direct maximization methods are preferable in terms of computing time.

In our work, we have used a version of the EM algorithm for several reasons: this is the most popular and widely used maximization method for HMMs and is well-adapted to the underlying Markov structure. Also its implementation is reasonably easy and usually leads to explicit formulae for parameter estimates at the M-step.

4.2. The EM algorithm

Two recent bibliographic reviews of EM-related papers[30,31] have shown the popularity of this algorithm. Its basic principle is an iterative scheme that first computes the conditional expectations of the unobserved states of the E-step, given the parameters, and then maximizes the likelihood function, given the data and the expected states (called the 'complete-data' likelihood) at the M-step.

The unobserved states are represented by indicator random variables that we define as follows:

$$v_{jk}(t) = 1 \text{ if } s_{t-1} = j \quad \text{and} \quad s_t = k$$

and

$$u_j(t) = 1 \quad \text{if} \quad s_t = j.$$

The complete-data log-likelihood can be expressed directly from (1) as

$$\log(\alpha_{s_1}) + \sum_{j=1}^{m} \sum_{k=1}^{m} \log \alpha_{jk} \sum_{t=2}^{n} v_{jk}(t) + \sum_{j=1}^{m} \sum_{t=1}^{n} u_j(t) \log f_{j_t}(y_t; \theta_j). \tag{2}$$

Let ω be one iteration of the EM algorithm. At the E-step, $\hat{v}_{jk}^{(\omega)}(t)$ and $\hat{u}_j^{(\omega)}(t)$, which are the conditional expectations of $v_{jk}(t)$ and $u_j(t)$, are computed with the 'forward–backward' algorithm (see Appendix I).[24] Then, at the M-step, they are applied to (2) to obtain new estimates of $\hat{\alpha}_{jk}^{(\omega+1)}$ and $\hat{\theta}_j^{(\omega+1)}$. The number of computations in any E-step is of the order of $2nm^2$ calculations.[11]

The EM algorithm has some well-known limitations, however.[30] One criticism is that its convergence can be slow. Another weakness concerns its strong dependence on the starting position. A third limitation is that it may converge to a saddle point rather than a maximum of the likelihood function.

Stochastic versions of the EM algorithm have been developed to overcome these limitations. These include the stochastic EM algorithm (SEM),[32,33] the stochastic approximation EM algorithm (SAEM)[34] and the Monte Carlo EM algorithm (MCEM).[35,36] The first two algorithms incorporate a stochastic step (S-step) between the E-step and the M-step by introducing random perturbations for the estimation of $\hat{u}_j^{(\omega)}(t)$. These perturbations are introduced mainly to extricate the parameter sequence from any potentially unstable stationary point of the likelihood function. The random perturbations in SEM are too large when the number of observations is small[34] even though SEM overcomes most of the limitations of EM. The SAEM algorithm has a median position between EM and SEM. By contrast to SEM, the random perturbations that SAEM introduces decrease to 0 as the number of iterations increases. This implies that it will perform best for small sample sizes. These considerations led us to choose the SAEM algorithm as an alternative to the EM algorithm.

4.3. The SAEM algorithm

The SAEM algorithm introduces an S-step just after the E-step and before the M-step. During the S-step new values of $\hat{u}_j^{(\omega)}(t)$, which we note as $\hat{u}_j^{(\omega)\text{new}}(t)$ are computed by the following method.

First, for each $t = 1, \dots, n$, random perturbations are generated by a time-dependent random variable $c_j^{(\omega)}(t)$ that follows a multinomial distribution of one draw of m categories with probabilities $p_j = \hat{u}_j^{(\omega)}(t)$, $j = 1, \dots, m$. For example, in the case of a two-state hidden Markov model, n uniform $(0, 1)$ random numbers $U(t)$ are generated. Then $c_1^{(\omega)}(t) = 1$ and $c_2^{(\omega)}(t) = 0$ if $0 \leqslant U(t) \leqslant \hat{u}_1^{(\omega)}(t)$. Otherwise $c_1^{(\omega)}(t) = 0$ and $c_2^{(\omega)}(t) = 1$.

Second, to avoid numerical singularities, the means $1/n \sum_{t=1}^{n} c_j^{(\omega)}(t)$ are calculated for each $j = 1, \dots, m$. If each mean is smaller than a given threshold, which we note $c(n)$, then we return to the previous step and draw new values of $c_j^{(\omega)}(t)$. The threshold $c(n)$ is chosen so that $0 < c(n) < 1$ and $\lim_{n \to \infty} c(n) = 0$; a possible choice is $c(n) = 1/n$.

Third, for each t, $\hat{u}_j^{(\omega)\text{new}}(t)$ is calculated by

$$\hat{u}_j^{(\omega)\text{new}}(t) = (1 - \gamma(\omega))\, \hat{u}_j^{(\omega)}(t) + \gamma(\omega)\, c_j^{(\omega)}(t).$$

The function $\gamma(\omega)$ is chosen so that it starts with $\gamma(0) = 1$, stays near 1 during the first iterations and decreases slowly towards 0 when ω increases to an arbitrarily-chosen maximum number of iterations I. A possible choice is $\gamma(\omega) = \cos(\pi\omega/2I)$. To our knowledge, no rule is indicated for choosing I. A value of $I = 500$ was sufficient to allow convergence.

5. SELECTING AND COMPARING HIDDEN MARKOV MODELS

Likelihood-based criteria, such as Akaike or Bayesian information criteria (BIC), can be used to select the number of states of the hidden Markov model and to compare non-nested models. We used the Bayesian information criterion[37] because it does not favour the more complicated model asymptotically.[38] We note the maximized log-likelihood as $l_{\max}$ and the number of free parameters as k. Then

$$\mathrm{BIC} = l_{\max} - \frac{\log(n)k}{2}.$$

The models with the highest BIC values were selected.

6. RECONSTRUCTING THE MOST LIKELY SEQUENCE OF STATES

Reconstructing the most likely sequence of states associated with the time series under study is useful in analysing epidemiologic surveillance data. This sequence cannot, however, simply be deduced by calculating $\max_j(\hat{u}_j(t))$, which only determines the most likely state at time t, without regard to the probability of any particular sequence of states. The Viterbi algorithm,[12,39] which is a recursive procedure, can be used to uncover the most likely state sequence (see Appendix II).

7. MODEL EXTENSIONS

Epidemiologic surveillance data often reveal trend or seasonality, as Figure 1 shows for the ILI data. These features can be introduced in this model by defining the distribution parameters as linear or periodic functions of time. Let β_j represent a linear trend and δ_j, ε_j be the parameters associated with an r-period seasonality. The mean of a mixture of Gaussian distributions can then be parameterized as

$$\mu_j(t) = \gamma_j + \beta_j t + \delta_j \cos\left(\frac{2\pi t}{r}\right) + \varepsilon_j \sin\left(\frac{2\pi t}{r}\right).$$

For a mixture of Poisson distributions, the parameter can be expressed by

$$\log(\lambda_j(t)) = \gamma_j + \beta_j t + \delta_j \cos\left(\frac{2\pi t}{r}\right) + \varepsilon_j \sin\left(\frac{2\pi t}{r}\right).$$

The logarithm function is used to ensure positive values of $\lambda_j(t)$.

Introducing these additional terms, however, can lead to difficulties with likelihood maximization. With a Gaussian distribution, the parameters are obtained by simple linear algebra. For Poisson distributions, though, the M-step does not always have a closed-form solution, and estimates cannot be obtained by an explicit expression. To avoid this problem, we used the algorithm of Rai and Matthews,[40] which consists of an unchanged E-step and has only one cycle in the iterative solution of the likelihood maximization. Other standard maximization algorithms can also be considered at the M-step. Examples include the simplex or Newton–Raphson algorithms.

8. ILLUSTRATION

8.1. Influenza-like illness data

Hidden Markov models were applied to ILI data, by using a mixture of Gaussian distributions. Several models were tested with or without trend terms. Depending on the BIC values, models were improved by the inclusion of trend terms. Linear trend and periodic parameters ($r = 52$) were included as described above.

Table I shows the transition probability matrix, the stationary probabilities, the maximized log-likelihood value, and the BIC corresponding to each hidden Markov model tested (six containing from one to six states). The maximized log-likelihood of the two-state model is substantially better than that of the one-state model; BIC is highest for the five-state model. Because what we know, graphically, epidemiologically and biologically, about ILI incidence rates suggests a mixture of two dynamics (non-epidemic, epidemic), we focus on the two-state model. State 1 is associated with the highest mean and has a stationary probability of 26 per cent. The probability that any given week is in state 1 ($\hat{u}_1(t)$) is shown in Figure 2. This probability is always close to 0 or to 1, which indicates that the model clearly differentiates between the two states.

Figure 3 shows the most likely state sequence obtained with the two-state model. Two distinct states are clearly visible, one epidemic (state 1), and the other not (state 2). Moreover, the results obtained are consistent (92·3 per cent concordance) with the current method used by the Sentinelles system to determine the existence of an ILI epidemic.[10] Results disagreed between the two methods only in one direction; 48 weeks were classified as epidemic by the hidden Markov model and non-epidemic by the other method.

8.2. Poliomyelitis data

Epidemiologic data are often presented as counts, a form that suggests an underlying Poisson distribution. To illustrate the use of hidden Markov models with data containing a mixture of Poisson distributions, we apply this method to the time series of monthly poliomyelitis cases in the United States between January 1970 and December 1983 (available at http://www.maths.monash.edu.au/~hyndman/tseries/health/polio.dat).

We tested several models and observed the best results from the non-seasonal two-state hidden Markov model, which yielded the highest BIC value ($-270{\cdot}28$, compared with $-302{\cdot}58$ for a one-state model, $-277{\cdot}04$ for a three-state model, and $-274{\cdot}38$ for a seasonal two-state model). The initial state was state 1.

The transition probability matrix was

$$\begin{pmatrix} 0{\cdot}932 & 0{\cdot}068 \\ 0{\cdot}331 & 0{\cdot}670 \end{pmatrix}.$$

Its corresponding stationary probabilities (0·840, 0·160) and parameter values were

$$\lambda_1 = 0{\cdot}791 \text{ and } \lambda_2 = 4{\cdot}180.$$

Figure 4 shows the probability of state 2 in any given month, and Figure 5 the optimal state sequence obtained with the Viterbi algorithm.

Table I. Comparison of seasonal Gaussian hidden Markov models fitted to weekly ILI incidence rates

m	γ_j	β_j	δ_j	ε_j	Δ_j†	Transition probability matrices						Stationary probabilities	Log-likelihood	k‡	BIC
1	1·716	– 0·0015	1·641	0·550	7·721	1·000						1·000	– 1360·05	5	– 1376·14
2	*1·597	0·0009	3·349	0·517	29·295	0·919	0·081					0·260	– 286·60	12	– 325·23
	0·551	– 0·0006	0·286	0·053	0·146	0·028	0·972					0·740			
3	*2·927	– 0·0007	3·282	– 0·005	37·264	0·889	0·000	0·111				0·202	– 41·45	21	– 109·05
	0·302	– 0·0002	0·176	0·032	0·040	0·000	0·946	0·054				0·494			
	0·775	– 0·0007	0·384	0·184	0·527	0·074	0·088	0·839				0·303			
4	9·216	– 0·0052	0·354	– 4·687	47·946	0·874	0·000	0·111	0·015			0·155	39·67	32	– 63·33
	0·299	– 0·0002	0·178	0·032	0·042	0·000	0·935	0·000	0·065			0·491			
	*0·930	– 0·0001	0·652	0·187	1·565	0·165	0·000	0·634	0·201			0·126			
	0·653	– 0·0005	0·332	0·145	0·334	0·000	0·141	0·122	0·738			0·228			
5	14·640	– 0·0077	– 1·986	– 8·465	68·712	0·790	0·018	0·192	0·000	0·000		0·099	84·45	45	– 60·40
	0·654	– 0·0005	0·332	0·145	0·349	0·000	0·738	0·000	0·121	0·140		0·229			
	4·150	– 0·0017	0·269	– 1·412	11·009	0·405	0·000	0·274	0·322	0·000		0·055			
	*0·931	– 0·0001	0·656	0·185	1·616	0·000	0·205	0·165	0·630	0·000		0·127			
	0·299	– 0·0002	0·178	0·032	0·042	0·000	0·065	0·000	0·000	0·935		0·491			
6	14·498	– 0·0077	– 1·844	– 8·384	68·002	0·793	0·017	0·190	0·000	0·000	0·000	0·100	129·32	60	– 63·81
	0·862	– 0·0011	0·522	0·134	0·525	0·000	0·741	0·000	0·110	0·100	0·049	0·191			
	4·095	– 0·0019	0·434	– 1·396	10·882	0·407	0·000	0·252	0·340	0·000	0·000	0·054			
	1·167	– 0·0005	0·773	0·136	1·787	0·000	0·263	0·268	0·425	0·044	0·000	0·081			
	0·312	– 0·0002	0·171	0·102	0·031	0·000	0·033	0·000	0·000	0·922	0·045	0·351			
	*0·552	– 0·0008	0·335	– 0·045	0·226	0·000	0·065	0·000	0·033	0·022	0·880	0·223			

m = number of states·

* the initial state of the Markov chain that maximized the likelihood

† $\Delta_j = \frac{\sum_{t=1}^{n} \hat{u}_j(t)\, y_t^2}{\sum_{t=1}^{n} \hat{u}_j(t)}$ · The variance $\sigma_j^2(t)$ is estimated by $\hat{\sigma}_j^2(t) \; = \Delta_j - \hat{\mu}_j^2(t)$ (see Appendix I)·

‡ k is the number of free parameters, equal to $5m$ parameters $(\gamma_j, \beta_j, \delta_j, \varepsilon_j, \Delta_j) + m(m-1)$ free transition probabilities

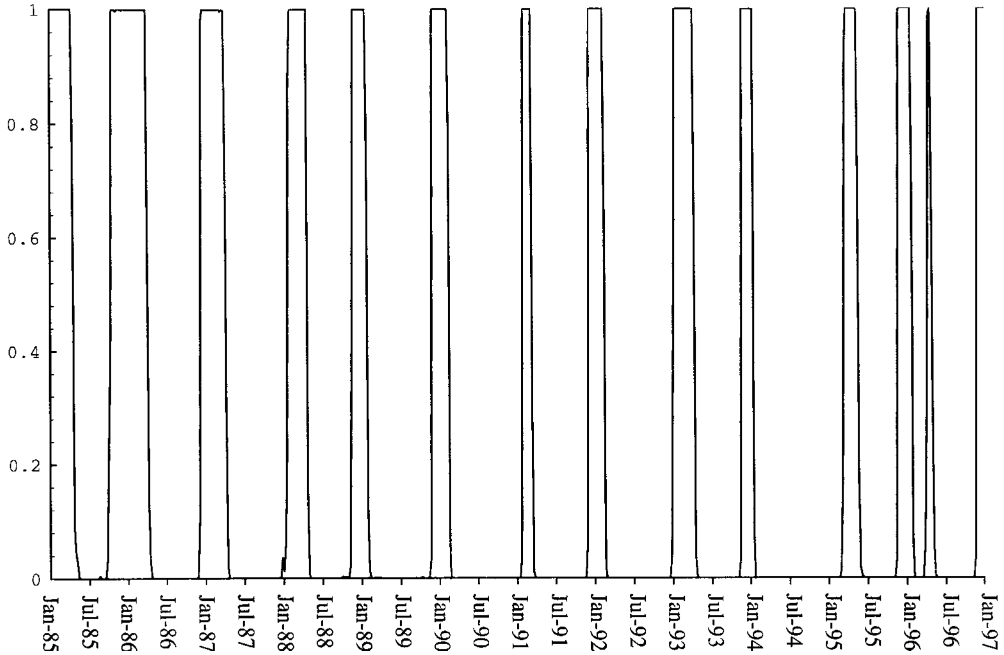

Figure 2. Estimated probabilities of state 1 over time for ILI

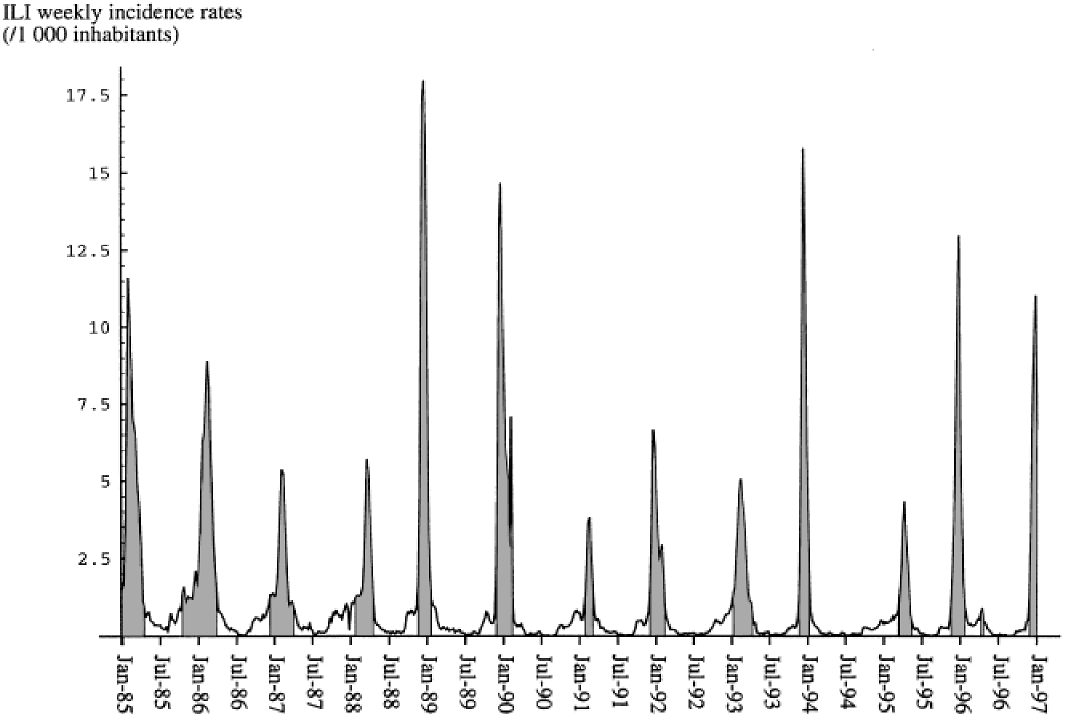

Figure 3. The state sequence of weeekly ILI incidence rates obtained by applying the Viterbi algorithm

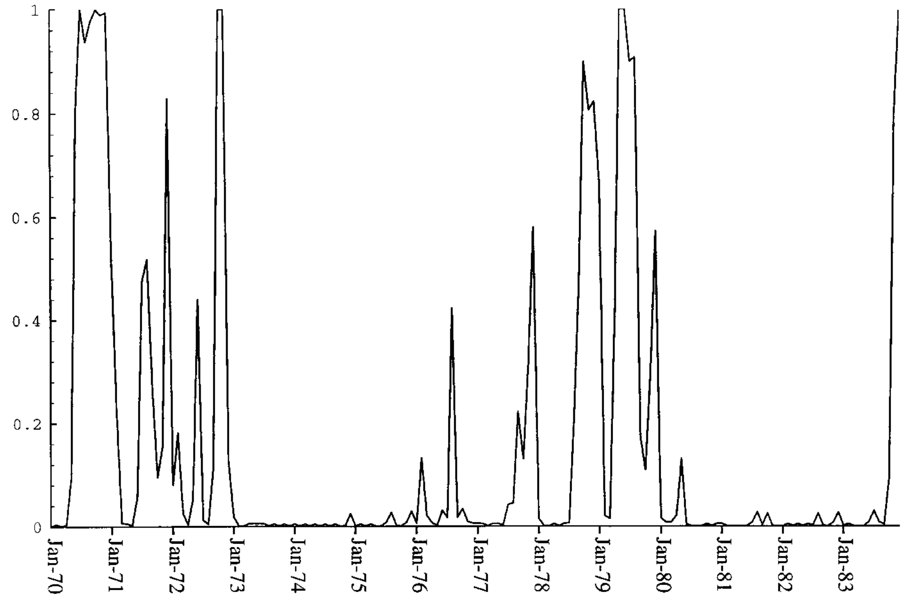

Figure 4. Estimated probabilities of state 2 over time for poliomyelitis

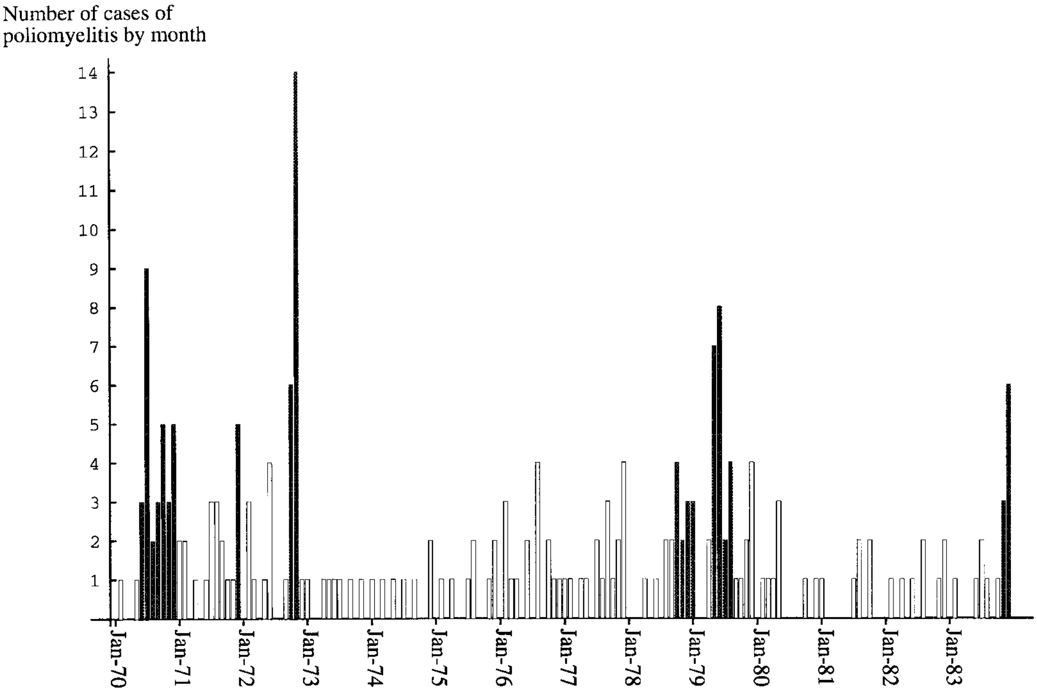

Figure 5. The state sequence of monthly poliomyelitis counts, U.S.A. 1970–1983, obtained by applying the Viterbi algorithm

Four periods were associated with state 2: one between June and December 1970; a second between October and November 1972; a third between September 1978 and August 1979, another in November and December 1983.

The first three periods coincided with the three poliomyelitis outbreaks reported in United States over this period, in Texas (1970), Connecticut (1972), and in Pennsylvania, Wisconsin, Iowa and Missouri (1979).[41] The final period was not classified as epidemic following the CDC's classification.[42] It corresponded to six cases which occurred in oral poliomyelitis vaccine recipients and four subsequent cases among contacts of these subjects.

9. DISCUSSION

We have shown here that hidden Markov models are useful for analysing epidemiologic surveillance data. When a two-state hidden Markov model was applied to the analysis of ILI incidence rates, it clearly differentiated between epidemic and non-epidemic rates. We used this model as an illustration because it fulfilled a standard epidemiologic objective, that is, to determine the timing of epidemic periods. The best results for the likelihood criterion, however, came from a five-state model. There are several explanations for this discrepancy. First, the choice of Gaussian distributions may be inappropriate in this case, since it implies that negative values of ILI rates could theoretically be observed. Other choices for the distributions (for example, Gamma, Weibull) would have been more appropriate but would have led to computational difficulties during the maximization steps. Second, the validity of the hypothesis that ILI incidence rates are independent conditional on the state, is questionable. In particular, during epidemic periods, incidence rates increase successively for a period and then decrease, again over consecutive time intervals. It would have been possible to deal with this problem by introducing autoregressive terms in the model or by explicitly modelling the state duration.[11] Again, for simplicity, we did not use these two methods in this paper. None the less, these limitations should not obscure the model's substantial advantages. In particular, it does not require any arbitrary choice about the timing of epidemic periods. Furthermore, in the two-state model, a threshold can be computed directly from the non-epidemic state (by estimating $\mu(t)$ and $\sigma^2(t)$) and used as an early warning system, exactly as in ARIMA models.[4]

Applied to poliomyelitis data, the hidden Markov model clearly identified, without any arbitrary hypothesis, all three outbreaks reported over the data period. The use of this method to analyse such count data was straightforward. We did not calculate the standard errors of parameters, which would be useful, for example, to determine confidence intervals. The parametric bootstrap method[26] can be used for this purpose, but will require extensive computation.

Another attractive aspect of hidden Markov models, not shown in this paper, lies in the possibility of introducing covariates into the model, for example, by defining the distribution parameters as a function of the covariates. This is particularly important with epidemiologic monitoring data, to correlate disease and exposure rates, or to evaluate the effect of an intervention on the observed disease rate (for example, modification of the case definition, or a public health intervention).[20]

Finally, we want to stress that hidden Markov models provide a very flexible tool for analysing time series of discrete values. They should rapidly join the array of methods used for monitoring epidemiologic surveillance data.

All computations were performed on a UNIX station using computer programs written in C. The source code is available on request.

APPENDIX I: THE EM ALGORITHM

The estimation of parameters by the EM algorithm starts by initializing the parameters

$$(\alpha_1^{(0)}, \dots, \alpha_m^{(0)}), (\theta_1^{(0)}, \dots, \theta_m^{(0)})$$

as well as the transition probability matrix

$$[\alpha_{jk}^{(0)}]_{1 \leqslant j,k \leqslant m} = \begin{bmatrix} \alpha_{11}^{(0)} & \cdots & \alpha_{1m}^{(0)} \\ \vdots & \ddots & \vdots \\ \alpha_{m1}^{(0)} & \cdots & \alpha_{mm}^{(0)} \end{bmatrix}.$$

At each iteration ω, $\omega = 1, \dots, I$, the E-step is followed by the M-step.

The E-step

The forward–backward algorithm[24] was used to replace $v_{jk}(t)$ and $u_j(t)$ by their conditional expectations $\hat{v}_{jk}(t)$ and $\hat{u}_j(t)$, given the following observations and current parameter estimates:

$$\hat{v}_{jk}(t) = P(S_{t-1} = j, S_t = k | y_1, \dots, y_n) \quad \text{and} \quad \hat{u}_j(t) = P(S_t = j | y_1, \dots, y_n).$$

The values of $\hat{v}_{jk}(t)$ and $\hat{u}_j(t)$ were obtained by computing

$$a_j(t) = f(y_1, \dots, y_t, S_t = j) \quad \text{and} \quad b_j(t) = f(y_{t+1}, \dots, y_n | S_t = j)$$

with the recursive 'forward–backward' formulae.

Specifically, for $j = 1, \dots, m$

$$a_j^{(\omega)}(1) = \alpha_j^{(0)} f_{j_1}(y_1; \hat{\theta}_j^{(\omega)})$$

and

$$a_j^{(\omega)}(t) = \sum_{k=1}^{m} a_k^{(\omega)}(t-1) \hat{\alpha}_{kj}^{(\omega)} f_{j_t}(y_t; \hat{\theta}_j^{(\omega)}) \qquad t = 2, \dots, n.$$

Then

$$b_j^{(\omega)}(n) = 1$$

and

$$b_j^{(\omega)}(t) = \sum_{k=1}^{m} \hat{\alpha}_{jk}^{(\omega)} f_{k_{t+1}}(y_{t+1}; \hat{\theta}_k^{(\omega)}) b_k^{(\omega)}(t+1) \qquad t = n-1, \dots, 1.$$

The likelihood is calculated by

$$L_n(\psi) = \sum_{j=1}^{m} a_j^{(\omega)}(t) b_j^{(\omega)}(t) \qquad t = 1, \dots, n$$

and particularly by

$$L_n(\psi) = \sum_{j=1}^{m} a_j^{(\omega)}(n).$$

Next, for $t = 1, \ldots, n$ the conditional expectations $\hat{v}_{jk}^{(\omega)}(t)$ and $\hat{u}_j^{(\omega)}(t)$ were computed by

$$\hat{u}_j^{(\omega)}(t) = \frac{a_j^{(\omega)}(t)\, b_j^{(\omega)}(t)}{\sum_{l=1}^{m} a_l^{(\omega)}(n)}$$

and

$$\hat{v}_{jk}^{(\omega)}(t) = \frac{\hat{\alpha}_{jk}^{(\omega)} f_{k_t}(y_t; \hat{\theta}_k^{(\omega)})\, a_j^{(\omega)}(t-1)\, b_k^{(\omega)}(t)}{\sum_{l=1}^{m} a_l^{(\omega)}(n)}.$$

We used Leroux and Puterman's scaling method[29] to avoid the numerical instability of the computation of $a_j(t)$ and $b_j(t)$. Instead, they were divided by 10^p, with p such that $10^{-p} \sum_{j=1}^{m} a_j(t)$ and $10^{-p} \sum_{j=1}^{m} b_j(t)$ lie between 0·1 and 1.

The M-step

This maximizes the two parts of the complete-data log-likelihood separately. Maximization of the first part yields the values of the transition probabilities:

$$\hat{\alpha}_{jk}^{(\omega+1)} = \frac{\sum_{t=2}^{n} \hat{v}_{jk}^{(\omega)}(t)}{\sum_{t=2}^{n} \sum_{l=1}^{m} \hat{v}_{jl}^{(\omega)}(t)}.$$

The stationary probabilities α_j are given by

$$\hat{\alpha}_j^{(\omega+1)} = \frac{\sum_{t=1}^{n} \hat{u}_j^{(\omega)}(t)}{\sum_{k=1}^{m} \sum_{t=1}^{n} \hat{u}_k^{(\omega)}(t)}.$$

We estimate θ_j by maximizing the second part of the complete-data log-likelihood:

$$\sum_{t=1}^{n} \hat{u}_j^{(\omega)}(t) \log f_{j_t}(y_t; \theta_j).$$

Using the Gaussian case with mean μ_j and variance σ_j^2 implies that

$$\hat{\mu}_j^{(\omega+1)} = \frac{\sum_{t=1}^{n} \hat{\mu}_j^{(\omega)}(t)\, y_t}{\sum_{t=1}^{n} \hat{\mu}_j^{(\omega)}(t)}.$$

and that

$$\hat{\sigma}_j^{2\,(\omega+1)} = \Delta_j^{(\omega)} - \hat{\mu}_j^{(\omega+1)^2} \quad \text{with} \quad \Delta_j^{(\omega)} = \frac{\sum_{t=1}^{n} \hat{\mu}_j^{(\omega)}(t)\, y_t^2}{\sum_{t=1}^{n} \hat{\mu}_j^{(\omega)}(t)}.$$

In the Poisson case with parameter λ_j

$$\hat{\lambda}_j^{(\omega+1)} = \frac{\sum_{t=1}^{n} \hat{\mu}_j^{(\omega)}(t)\, y_t}{\sum_{t=1}^{n} \hat{\mu}_j^{(\omega)}(t)}.$$

APPENDIX II

The Viterbi algorithm allows the best state-sequence to be deduced by computing

$$\xi_{tk} = \max_{s_1, \ldots, s_{t-1}} f(S_1 = s_1, \ldots, S_t = k, y_1, \ldots, y_t), \quad \text{for } t = 2, \ldots, n.$$

That is, ξ_{tk} is the highest density along a single path at time t.

The recursive computation of ξ_{tk} is straightforward:

$$\xi_{tk} = \max_j (\xi_{(t-1)j}\, \hat{\alpha}_{jk}) f_{k_t}(y_t; \hat{\theta}_k) \quad \text{for } t = 2, \ldots, n \text{ and } k = 1, \ldots, m$$

with

$$\xi_{1k} = \hat{\alpha}_k f_{k_1}(y_1; \hat{\theta}_k) \quad \text{for } k = 1, \ldots, m.$$

Then, the index sequence of states $\hat{j}_t$ is calculated starting from $\hat{j}_n = \arg\max_j (\xi_{nj})$, and recursively for $t = n-1, \ldots, 1$ by

$$\hat{j}_t = \arg \max_j (\xi_{tj}\, \hat{\alpha}_{j\hat{j}_{t+1}}).$$

REFERENCES

1. Box, G. E. P. and Jenkins, G. M. *Time Series Analysis. Forecasting and Control*, Holden–Day, San Fransisco, 1970.
2. Helfenstein, U. 'The use of transfer function models, intervention analysis and related time series methods in epidemiology', *International Journal of Epidemiology*, **20**, 808–815 (1991).
3. Martinez-Schnell, B. and Zaidi, A. 'Time series analysis of injuries', *Statistics in Medicine*, **8**, 1497–1508 (1989).
4. Watier, L., Richardson, S. and Hubert, B. 'A time series construction of an alert threshold with application to S.bovismorficans in France', *Statistics in Medicine*, **10**, 1493–1509 (1991).
5. Smith, A. F. M. and West, M. 'Monitoring renal transplants: an application of the multiprocess Kalman filter', *Biometrics*, **39**, 867–878 (1983).
6. Gordon, K. and Smith, A. F. M. 'Modeling and monitoring biomedical time series', *Journal of the American Statistical Association*, **85**, 328–337 (1990).
7. Stroup, D. F. and Thacker, S. B. 'A Bayesian approach to the detection of aberrations in public health-surveillance data', *Epidemiology*, **4**, 435–443 (1993).
8. Stroup, D. F., Williamson, G. D. and Herndon, J. L. 'Detection of aberrations in the occurrence of notifiable diseases surveillance data', *Statistics in Medicine*, **8**, 323–329 (1989).
9. Serfling, R. E. 'Methods of current statistical analysis of excess pneumonia-influenza deaths', *Public Health Reports*, **78**, 494–506 (1963).
10. Costagliola, D., Flahault, A., Galinec, D., Garnerin, P., Menares, J. and Valleron, A-J. 'A routine tool for detection and assessment of epidemics of influenza-like syndromes in France', *American Journal of Public Health*, **81**, 97–99 (1991).

11. Rabiner, L. R. 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proceedings of the IEEE*, **77,** 257–286 (1989).
12. Juang, B. H. and Rabiner, L. R. 'Hidden Markov models for speech recognition', *Technometrics*, **33**, 251–272 (1991).
13. Kundu, A. and He, Y. 'On optimal order in modeling sequence of letters in words of common language as a Markov chain', *Pattern Recognition*, **24**, 603–608 (1991).
14. Coast, D. A., Cano, G. G. and Briller, S. A. 'Use of Hidden Markov Models for electrocardiographic signal analysis', *Journal of Electrocardiology*, **23**, 184–191 (1990).
15. Albert, P. S. 'A two-state Markov mixture model for a time series of epileptic seizure counts', *Biometrics*, **47**, 1371–1381 (1991).
16. Le, N. D., Leroux, B. G. and Puterman, M. L. 'Exact Likelihood evaluation in a Markov mixture model for time series of seizure counts', *Biometrics*, **48**, 317–323 (1992).
17. Churchill, G. A. 'Stochastic models for heterogeneous DNA sequences', *Bulletin of Mathematical Biology*, **51**, 79–94 (1989).
18. Camproux, A-C., Saunier, F., Chouvet, G., Thalabard, J-C. and Thomas, G. 'A hidden model approach to neuron firing pattern analysis', *Biophysical Journal*, **71**, 2404–2412 (1996).
19. Zucchini, W. and Guttorp, P. 'A hidden Markov model for space-time precipitation', *Water Resources Research*, **27**, 1917–1923 (1991).
20. MacDonald, I. L. and Lerer, L. B. 'A time series analysis of trends in firearm-related homicide and suicide', *International Journal of Epidemiology*, **23**, 66–72 (1994).
21. Valleron, A-J. and Garnerin, P. 'Computer networking as a tool for public health surveillance: the French experiment', *Morbidity and Mortality Weekly Report*, **41**, 606–614 (1992).
22. Boussard, E., Flahault, A., Vibert, J. F. and Valleron, A-J. 'Sentiweb: French communicable diseases surveillance on the World-Wide Web', *British Medical Journal*, **313**, 1381–1384 (1996).
23. *Control of Communicable Diseases* 16th edn, Abram S. Benenson, 1995, p. 535.
24. Baum, L. E., Petrie, T., Soules, G. and Weiss, N. 'A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains', *Annals of Mathematical Statistics*, **41**, 164–171 (1970).
25. Levinson, L. E., Rabiner, L. R. and Sondhi, M. M. 'An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition', *Bell System Technical Journal*, **62**, 1035–1074 (1983).
26. MacDonald, I. L. and Zucchini, W. *Hidden Markov and Other Models for Discrete-valued Time Series*, Chapman & Hall, London, 1997.
27. Fredkin, D.R. and Rice, J. A. 'Maximum likelihood estimation and identification directly from single-channel recordings', *Proceedings of the Royal Society of London*, *Series B*, **249**, 125–132 (1992).
28. Dempster, A. P., Laird, N. M. and Rubin, D. B. 'Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)', *Journal of the Royal Statistical Society*, *Series B*, **39**, 1–38 (1977).
29. Leroux, B. G. and Puterman, M. L. 'Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models', *Biometrics*, **48**, 545–558 (1992).
30. Meng, X. L. and Van Dyk, D. 'The EM algorithm – an old folk-song sung to a fast tune', *Journal of the Royal Statistical Society*, *Series B*, **59**, 511–567 (1997).
31. 'The impact of the EM algorithm on medical statistics', *Statistical Methods in Medical Research*, **6**, 000–000 (1997).
32. Celeux, G. and Diebolt, J. 'The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem', *Computational Statistics Quarterly*, **2**, 73–82 (1985).
33. Celeux, G. and Diebolt, J. 'A probabilistic teacher algorithm for iterative maximum likelihood estimation', *in* Bock, H. H. (ed.), *Classification and Related Methods of Data Analysis*, North-Holland, Amsterdam, 1988, pp. 617–623.
34. Celeux, G. and Diebolt, J. 'A stochastic approximation type EM algorithm for the mixture problem', *Stochastics and Stochastics Reports*, **41**, 119–134 (1992).
35. Wei, G. C. G. and Tanner, M. A. 'A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms', *Journal of the American Statistical Association*, **85**, 699–704 (1990).

36. Tanner, M. A. *Tools for Statistical Inference*, Lectures Notes in Statistics 67, Springer-Verlag, New York, 1991.
37. Schwartz, G. 'Estimating the dimension of a model', *Annals of Statistics*, **6**, 461–464 (1978).
38. Katz, R. W. 'On some criteria for estimating the order of a Markov chain', *Technometrics*, **23**, 243–249 (1981).
39. Forney, G. D. 'The Viterbi algorithm', *Proceedings of the IEEE*, **61**, 268–278 (1973).
40. Rai, S. N. and Matthews, D. E. 'Improving the EM algorithm', *Biometrics*, **49**, 587–591 (1993).
41. Moore, M., Katona, P., Kaplan, J. E Schonberger, L. B. and Hatch, M. H. 'Poliomyelitis in the United States, 1969–1981', *Journal of Infectious Diseases*, **4**, 558–563 (1982).
42. Anonymous. 'Paralytic poliomyelitis – United States, 1982 and 1983', *Morbidity Mortality Weekly Report*, **33**, 635–638 (1984).