ROBERTA PAROLI (*) – LUIGI SPEZIA (**)

Parameter estimation of Gaussian hidden Markov models when missing observations occur

INTRODUCTION

Hidden Markov models (HMMs) are discrete-time stochastic processes $\{Y_t; X_t\}$ such that $\{Y_t\}$ is an observed sequence of random variables and $\{X_t\}$ is an unobserved Markov chain. $\{Y_t\}$, given $\{X_t\}$, is a sequence of conditionally independent random variables (*conditional independence condition*) with the conditional distribution of Y_t depending on $\{X_t\}$ only through X_t (*contemporary dependence condition*). HMMs are widely used to represent time series characterized by weakly dependence among the observations.

HMMs have been first studied for speech recognition (Juang and Rabiner (1991)) and then applied to time series analysis, assuming as conditional distribution a discrete (Albert (1991); Leroux and Puterman (1992)) or a continuous one (Fredkin and Rice (1992)). A wide survey of HMMs is available in the monographies by Elliot, Aggoun, Moore (1995) and by MacDonald and Zucchini (1997).

In this paper we examine HMMs in which the probability density

(*) Istituto di Statistica, Università Cattolica S.C., Via Necchi 9, 20123 Milano
(**) Dipartimento di Ingegneria, Università degli Studi di Bergamo, via Marconi 5, 24044 Dalmine (BG)

function (*pdf*) of every observation at any time, determined only by the current state of the chain, is Gaussian; so we have those special models $\{Y_t; X_t\}$ said *Gaussian hidden Markov models* (GHMMs).

Our aim is to obtain, through the EM algorithm, the analytic expression of the maximum likelihood estimators of GHMMs, also when the series of observations contains missing data. The basic model used to study univariate stationary non-linear time series will be introduced in Section 1; then, in Section 2, we consider some joint *pdf*s of the process $\{Y_t; X_t\}$ that will be used in Section 3 to obtain, by means of the EM algorithm, the explicit formulae of the estimators of the unknown parameters, which are, at the convergence of the algorithm, the maximum likelihood estimators; we show how these formulae may be applied when missing observations occur. Finally, in Section 4, an environmental application of GHMMs will be shown: we will examine a data set about the duration of geyser eruption, where within the series of 107 observations, we have 11 values that are missing.

1. The basic Gaussian Hidden Markov model

Let $\{X_t\}_{t\in(1,\dots,T)\subset\mathbb{N}}$ be a discrete, homogeneous, aperiodic, irreducible Markov chain on a finite state-space $S_X = \{1, 2, \dots, m\}$. The transition probabilities matrix is $\Gamma = [\gamma_{i,j}]$, with $\gamma_{i,j} = \mathrm{P}(X_t = j \mid X_{t-1} = i)$, for any $i; j \in S_X$, and the initial distribution is $\delta = (\delta_1, \delta_2, \dots, \delta_m)'$, where $\delta_i = \mathrm{P}(X_1 = i)$, for any $i \in S_X$. Since $\{X_t\}$ is a homogeneous, irreducible Markov chain, defined on a finite state-space, it has an initial distribution δ which is stationary, that is $\delta_i = \mathrm{P}(X_t = i)$ for any $t = 1, \dots, T$; hence the equality $\delta' = \delta'\Gamma$ holds. Finally, the hypothesis characterizing HMMs is that the Markov chain $\{X_t\}$ is unobservable: the sequence of states of the Markov chain is *hidden* in the observations. Notice that the states of the chain may have either a convenient interpretation suggested by the nature of the observed phenomenon, or be supposed only for convenience in formulating the model.

Let $\{Y_t\}_{t\in(1,\dots,T)\subset\mathbb{N}}$ be some discrete stochastic process, on a continuous state-space $S_Y \equiv \mathbb{R}$. The process $\{Y_t\}$ must satisfy two conditions: (1) *conditional independence condition* - the random variables $(Y_1, \dots, Y_T)$, given the variables $(X_1, \dots, X_T)$, are conditionally independent; (2) *contemporary dependence condition* - the distribution of

any Y_t, given the variables $(X_1, \ldots, X_T)$, depends only on the contemporary variable X_t. By these two conditions, given a sequence of lenght T of observations $y_1, y_2, \ldots, y_T$ and a sequence of lenght T of states of the unobserved Markov chain $i_1, i_2, \ldots, i_T$, the conditional *pdf* of the observations given the states results

$$f(y_1, y_2, \ldots, y_T \mid i_1, i_2, \ldots, i_T) = \prod_{t=1}^{T} f(y_t \mid i_1, i_2, \ldots, i_T) = \prod_{t=1}^{T} f(y_t \mid i_t),$$

where the generic $f(y \mid i)$ is the *pdf* of the Gaussian random variable Y_t, when $X_t = i$, henceforth denoted Y_t^i, for any $1 \leq t \leq T$:

$$Y_t^i \sim \mathcal{N}\left(\mu_i; \sigma_i^2\right), \text{ for any } i = 1, \ldots, m.$$

The so-defined model $\{Y_t; X_t\}_{t \in (1,\ldots,T) \subset \mathbb{N}}$ is said *Gaussian hidden Markov model* (GHMM) and is characterized by the stationary initial distribution δ, by the transition probabilities matrix Γ and by the state-dependent *pdf*s $f(y \mid i)$.

The GHMM can equivalently be written as a "signal plus noise" model:

$$Y_t^i = \mu_i + E_t^i,$$

where E_t^i denotes the Gaussian random variables E_t, when $X_t = i$, with zero mean and variance σ_i^2 $\left(E_t^i \sim \mathcal{N}\left(0; \sigma_i^2\right)\right)$, for any $i \in S_X$, with the discrete process $\{E_t\}$, given $\{X_t\}$, satisfying the conditional independence and the contemporary dependence conditions.

In this paper the procedure to estimate the unknown parameters of GHMMs will be studied. The parameters to be estimated are the $m^2 - m$ transition probabilities $\gamma_{i,j}$, for any $i = 1, \ldots, m$; $j = 1, \ldots, m-1$ (the entries of the m^{th} column of Γ are obtained by difference, given that all row of Γ sums equal to one), the m entries of the vector δ, the m parameters μ_i and the m parameters σ_i^2 of the m Gaussian random variables Y_t^i. The initial distribution δ will be estimated by the equality $\delta' = \delta'\Gamma$, after the estimation of the matrix Γ (being δ the stationary distribution). Hence the vector of the $m^2 + m$ unknown parameters is:

$$\phi = \left(\gamma_{1,1}, \ldots, \gamma_{1,m-1}, \ldots, \gamma_{m,1}, \ldots, \gamma_{m,m-1}, \mu_1, \ldots, \mu_m, \sigma_1^2, \ldots, \sigma_m^2\right)',$$

which belongs to the parameter space Φ. The estimator of the vector ϕ will be obtained by the maximum likelihood method, not in the direct analytic way, but in a numerical way by the EM algorithm.

2. Some joint probability density functions of the process

Our initial step is the examination of some joint *pdf*s of the process $\{Y_t; X_t\}$. First we shall obtain the joint *pdf*s of the observed variables $(Y_1, \ldots, Y_T)$, both for a complete sequence of data and for a sequence with missing observations; then we shall obtain the joint *pdf*s of the observed variables and one or two consecutive states of the Markov chain. These *pdf*s will be used in Section 3 to obtain, by means of the EM algorithm, the explicit formulae of the parameters estimators.

2.1. The joint pdf of $(Y_1, ..., Y_T)$

Given a sequence of observations $y_1, y_2, \ldots, y_T$ and a sequence of states of the Markov chain $i_1, i_2, \ldots, i_T$ from a HMM $\{Y_t; X_t\}$, we may obtain the joint *pdf*

$$\begin{aligned} & f(y_1, y_2, \ldots, y_T, i_1, i_2, \ldots, i_T) = \\ & = \delta_{i_1} \gamma_{i_1,i_2} \cdot \ldots \cdot \gamma_{i_{T-1},i_T} f(y_1 \mid i_1) f(y_2 \mid i_2) \cdot \ldots \cdot f(y_T \mid i_T) = \\ & = \delta_{i_1} f(y_1 \mid i_1) \prod_{t=2}^{T} \gamma_{i_{t-1},i_t} f(y_t \mid i_t), \end{aligned} \tag{1}$$

applying the conditional independence, the contemporary dependence and the Markov dependence conditions. Summing over $i_1, i_2, \ldots, i_T$ the equality (1), we have the joint *pdf*

$$f(y_1, y_2, \ldots, y_T) = \sum_{i_1 \in S_X} \sum_{i_2 \in S_X} \cdots \sum_{i_T \in S_X} \delta_{i_1} f(y_1 \mid i_1) \prod_{t=2}^{T} \gamma_{i_{t-1},i_t} f(y_t \mid i_t). \tag{2}$$

Setting $F_t = \text{diag}[f(y_t \mid 1), f(y_t \mid 2), \ldots, f(y_t \mid m)]$, for any $t = 1, \ldots, T$, we obtain

$$f(y_1, y_2, \ldots, y_T) = \delta' F_1 \Gamma F_2 \cdot \ldots \cdot \Gamma F_T 1_{(m)}, \tag{3}$$

where $1_{(m)}$ is the m-dimensional vector of ones. Replacing δ' with $\delta'\Gamma$, given that the initial distribution δ is stationary, and setting $\Gamma F_t = G_t$, we have

$$f(y_1, \ldots, y_T) = \delta' \left(\prod_{t=1}^{T} G_t \right) 1_{(m)}.$$

It is worth to remark that the joint *pdf* $f(y_1, \ldots, y_T)$ may be computed even if some data are missing. If, for example, a subsequence of $w-1$ observations, $y_{v+1}, \ldots, y_{v+w-1}$, is missing within a sequence $y_1, \ldots, y_T$, with $1 < v+1 \leq v+w-1 < T$, the *pdf* (2) becomes

$$\begin{aligned} f(y_1, \ldots, y_v, y_{v+w}, \ldots, y_T) = & \sum_{i_1 \in S_X} \cdots \sum_{i_v \in S_X} \sum_{i_{v+w} \in S_X} \cdots \sum_{i_T \in S_X} \delta_{i_1} \gamma_{i_1, i_2} \cdot \\ & \cdot \ldots \cdot \gamma_{i_{v-1}, i_v} \gamma_{i_v, i_{v+w}}(w) \gamma_{i_{v+w}, i_{v+w+1}} \cdot \ldots \cdot \gamma_{i_{T-1}, i_T} f(y_1 \mid i_1) \cdot \\ & \cdot \ldots \cdot f(y_v \mid i_v) f(y_{v+w} \mid i_{v+w}) \cdot \ldots \cdot f(y_T \mid i_T) = \\ = \delta' F_1 \cdot \ldots \cdot \Gamma F_v \Gamma^w F_{v+w} \cdot \ldots \cdot \Gamma F_T 1_{(m)} = & \\ = \delta' \left(\prod_{t=1}^{v} G_t \right) \Gamma^{w-1} \left(\prod_{t=v+w}^{T} G_t \right) 1_{(m)}, & \end{aligned} \tag{4}$$

where $\gamma_{i,j}(w)$ is the w-step transition probability, $\gamma_{i,j}(w) = \mathrm{P}(X_{v+w} = j \mid X_v = i)$; the w-step transition probabilities matrix is $\Gamma(w) = \left[\gamma_{i,j}(w)\right]$ and, by *Chapman-Kolmogorov equation*, we have $\Gamma(w) = \Gamma^w$.

The difference between formulae (3) and (4) lies in replacing the matrix F_t, for any $t = v+1, \ldots, v+w-1$, with the identity matrix.

2.2. The joint pdf of the observations and one state of the Markov chain

Now we want to obtain the joint *pdf*s of the observations $y_1, \ldots, y_T$ and the state i at time t of the Markov chain, i.e. $f(y_1, \ldots, y_T, X_t = i)$, for any $t = 1, \ldots, T$. Notice the two different notations for the joint *pdf*s of the observations and the states of the chain: if we have a complete sequence of states, we use $f(y_1, \ldots, y_T, i_1, \ldots, i_T)$, while, when we want to highlight one or two states, i or j, we use $f(y_1, \ldots, y_T, X_t = i, X_{t+1} = j)$ or $f(y_1, \ldots, y_T, X_1 = i_1, \ldots, X_t = i, X_{t+1} = j, \ldots, X_T = i_T)$. In the former case, time t appears only in the subscript of the states, while, in the latter, time t appears in the subscript of the variables X_t which have a generic realization i, j or i_t.

The joint *pdf* $f(y_1, \ldots, y_T, X_t = i)$ can be written as

$$\sum_{i_1 \in S_X} \cdots \sum_{i_{t-1} \in S_X} \sum_{i_{t+1} \in S_X} \cdots \sum_{i_T \in S_X} f(y_1, \ldots, y_T, X_1 = i_1, \ldots, X_t = i, \ldots, X_T = i_T).$$

Applying, as in Subsection 2.1, the conditional independence, the contemporary dependence and the Markov dependence conditions, in the following three situations we have:

(a) $t = 1$:

$$f(y_1, \ldots, y_T, X_1 = i) = \delta_i f(y_1 \mid i)\Gamma_{i\bullet} F_2 \Gamma F_3 \cdot \ldots \cdot \Gamma F_T 1_{(m)};$$

(b) $1 < t < T$:

$$f(y_1, \ldots, y_T, X_t = i) = \delta' F_1 \Gamma F_2 \cdot \ldots \cdot \Gamma F_{t-1} \Gamma_{\bullet i} f(y_t \mid i) \Gamma_{i\bullet} F_{t+1} \cdot \ldots \cdot \Gamma F_T 1_{(m)};$$

(c) $t = T$:

$$f(y_1, \ldots, y_T, X_T = i) = \delta' F_1 \Gamma F_2 \cdot \ldots \cdot \Gamma F_{T-1} \Gamma_{\bullet i} f(y_T \mid i),$$

where $\Gamma_{i\bullet}$ denotes the i^{th} row of Γ and $\Gamma_{\bullet i}$ the i^{th} column of Γ.

2.3. The joint pdf of the observations and two consecutive states of the Markov chain

Finally we want to obtain the joint *pdf*s of the observations $y_1, \ldots, y_T$ and two consecutive states i, j at times $t, t+1$ of the Markov chain, i.e. $f(y_1, \ldots, y_T, X_t = i, X_{t+1} = j)$, for any $t = 1, \ldots, T-1$. The joint *pdf* $f(y_1, \ldots, y_T, X_t = i, X_{t+1} = j)$ can be written as

$$\sum_{i_1 \in S_X} \cdots \sum_{i_{t-1} \in S_X} \sum_{i_{t+2} \in S_X} \cdots \sum_{i_T \in S_X} f(y_1, \ldots, y_T, X_1 = i_1, \ldots, X_t = i, X_{t+1} = j, \ldots, X_T = i_T).$$

Applying the conditional independence, the contemporary dependence and the Markov dependence conditions, in the following three situations we have:

(a) $t = 1$:

$$f(y_1, \ldots, y_T, X_1 = i, X_2 = j) = \\ = \delta_i f(y_1 \mid i) \gamma_{i,j} f(y_2 \mid j) \Gamma_{j\bullet} F_3 \Gamma F_4 \cdot \ldots \cdot \Gamma F_T 1_{(m)};$$

(b) $1 < t < T - 1$:

$$f(y_1, \ldots, y_T, X_t = i, X_{t+1} = j) =$$

$$= \delta' F_1 \Gamma F_2 \cdot \ldots \cdot \Gamma F_{t-1} \Gamma_{\bullet i} f(y_t \mid i) \gamma_{i,j} f(y_{t+1} \mid j) \Gamma_{j\bullet} F_{t+2} \cdot \ldots \cdot \Gamma F_T 1_{(m)};$$

(c) $t = T - 1$:

$$f(y_1, \ldots, y_T, X_{T-1} = i, X_T = j) = \\ = \delta' F_1 \Gamma F_2 \cdot \ldots \cdot \Gamma F_{T-2} \Gamma_{\bullet i} f(y_{T-1} \mid i) \gamma_{i,j} f(y_T \mid j).$$

3. Parameter estimation of GHMMs

In the context of HMMs, parameter estimation is based on maximum likelihood technique performed via the EM algorithm (for details on the EM algorithm, see McLachlan and Krishnan (1997)). The EM algorithm is an iterative procedure with two steps at each iteration: the first step, E step, provides the computation of an *Expectation*; the second step, M step, provides a *Maximization*.

Let $y = (y_1, \ldots, y_T)'$ be the vector of the observed data, that is the sequence of the realizations of the stochastic process $\{Y_t\}$; the vector y is incomplete because the sequence $(i_1, \ldots, i_T)'$ of the states of the chain $\{X_t\}$ is unobserved (or *hidden*). Moreover let $L_T^c(\phi)$ be the likelihood function of complete data and $L_T(\phi)$ be that of observed data:

$$L_T^c(\phi) = f(y_1, \ldots, y_T, i_1, \ldots, i_T) = \delta_{i_1} f(y_1 \mid i_1) \prod_{t=2}^{T} \gamma_{i_{t-1}, i_t} f(y_t \mid i_t);$$

$$L_T(\phi) = f(y_1, \ldots, y_T) = \sum_{i_1 \in S_X} \sum_{i_2 \in S_X} \cdots \sum_{i_T \in S_X} \delta_{i_1} f(y_1 \mid i_1) \prod_{t=2}^{T} \gamma_{i_{t-1} i_t} f(y_t \mid i_t).$$

The EM algorithm finds the value of ϕ that maximizes the log-likelihood of incomplete data, $\ln L_T(\phi)$, that is the maximum likelihood estimator based on the observations. The iterative scheme is the following. Let $\phi^{(0)}$ be some starting value of ϕ; at the first iteration, the E step requires the computation of the conditional expectation of the complete data log-likelihood, given the observed data, in $\phi = \phi^{(0)}$: $Q\left(\phi; \phi^{(0)}\right) = \mathrm{E}_{\phi^{(0)}}\left(\ln L_T^c(\phi) \mid y\right)$; the M step provides the search for that special value $\phi^{(1)}$ which maximize $Q\left(\phi; \phi^{(0)}\right)$, for any $\phi \in \Phi$. In general, at the $(k+1)^{th}$ iteration, the E and M steps are so defined:

E STEP - given $\phi^{(k)}$, compute $Q\left(\phi; \phi^{(k)}\right) = \mathrm{E}_{\phi^{(k)}}\left(\ln L_T^c(\phi) \mid y\right)$;

M STEP - search for that $\phi^{(k+1)}$ which maximize $Q\left(\phi; \phi^{(k)}\right)$, for any $\phi \in \Phi$.

The E and M steps must be repeated in an alternating way until we have the convergence of the sequence of the log-likelihood values $\left\{\ln L_T\left(\phi^{(k)}\right)\right\}$. The main property of the EM algorithm is the monotonicity of the log-likelihood function for incomplete data, with respect to the iterations of the algorithm (Dempster, Laird, Rubin

(1977), Theorem 1). If the algorithm converges at the $(k+1)^{th}$ iteration, $\left(\phi^{(k+1)}; \ln L_T\left(\phi^{(k+1)}\right)\right)$ is a stationary point of $\ln L_T(\phi)$. Assuming the standard hypotheses (Bickel, Ritov, Rydén (1998), Example 1) on μ_i and σ_i, i.e. $\mu_i \in [-1/\varepsilon, 1/\varepsilon]$ and $\sigma_i \in [\varepsilon, 1/\varepsilon]$, for any $i \in S_X$ and for some small $\varepsilon > 0$, the four regularity conditions on the convergence of the EM algorithm to a stationary point (Wu (1983), conditions (5), (6), (7), p. 96; (10), p. 98) are satisfied. Nevertheless, if the likelihood surface is multimodal, the convergence of the EM algorithm to the global maximum depends on the starting value $\phi^{(0)}$. To avoid the convergence to a stationary point which is not a global maximum, the best strategy is to start the algorithm from several different, possibly random, points in Φ and to compare the stationary points obtained at each run.

In case of GHMMs the regularity conditions let down by Leroux (1992) and Bickel, Ritov, Rydén (1998) are satisfied; hence the maximum likelihood estimator of ϕ is strongly consistent and asymptotically normal.

In practice the M step can be performed in two way: direcly, using a standard numerical optimization procedure (like Newton-Raphson method) or, when possible, analitically, obtainig the closed-form solutions for the parameter estimators. The explicit formulae of the parameter estimators of special HMMs, computed via the EM algorithm by means of the *backward and forward quantities* (Baum *et al.*, (1970)), are well-known in literature (Leroux and Puterman (1992); MacDonald and Zucchini (1997)), but they can not be adopted if we have missing observations. Moreover backward and forward quantities often give rise to problems of underflow, as T increases, and the scaling techniques available in literature do not allow to remove this problem. Hence we propose new versions of the analytic expression of the function $Q\left(\phi; \phi^{(0)}\right)$ at the E step and of the explicit formulae of the estimators of ϕ, which can be used also in the case of missing data.

Now the two steps of the EM algorithm at the $(k+1)^{th}$ iteration are analyzed in detail, remembering that at the k^{th} iteration the vector of estimates $\phi^{(k)}$ has been obtained:

$$\phi^{(k)} = \left(\gamma_{1,1}^{(k)}, \ldots, \gamma_{1,m-1}^{(k)}, \ldots, \gamma_{m,1}^{(k)}, \ldots, \gamma_{m,m-1}^{(k)}, \mu_1^{(k)}, \ldots, \mu_m^{(k)}, \sigma_1^{2^{(k)}}, \ldots, \sigma_m^{2^{(k)}}\right)'.$$

Henceforth the superscript (k) will denote a quantity obtained at the k^{th} iteration as a function of the vector $\phi^{(k)}$; notice that $\delta^{(k)}$ is

the left eigenvector of the matrix $\Gamma^{(k)} = \left[\gamma_{i,j}^{(k)}\right]$, associated with the eigenvalue one, such that $\delta'^{(k)} = \delta'^{(k)}\Gamma^{(k)}$.

At the E step of the $(k+1)^{th}$ iteration, the analytic closed-form of the function $Q\left(\phi; \phi^{(k)}\right)$ is

$$
\begin{aligned}
Q\left(\phi; \phi^{(k)}\right) &= \mathrm{E}_{\phi^{(k)}}\left(\ln L_T^c(\phi) \mid y\right) = \\
&= \sum_{i \in S_X} \frac{f^{(k)}(y_1, \ldots, y_T, X_1 = i)}{f^{(k)}(y_1, \ldots, y_T)} \ln \delta_i + \\
&\quad + \sum_{i \in S_X} \sum_{j \in S_X} \frac{\sum_{t=1}^{T-1} f^{(k)}(y_1, \ldots, y_T, X_t = i, X_{t+1} = j)}{f^{(k)}(y_1, \ldots, y_T)} \ln \gamma_{i,j} + \\
&\quad + \sum_{i \in S_X} \frac{\sum_{t=1}^{T} f^{(k)}(y_1, \ldots, y_T, X_t = i)}{f^{(k)}(y_1, \ldots, y_T)} \ln f(y_t \mid i),
\end{aligned}
$$

where

$$
f(y_t \mid i) = \frac{1}{\sqrt{2\pi}\ \sigma_i} \exp\left[-\frac{1}{2}\left(\frac{y_t - \mu_i}{\sigma_i}\right)^2\right],
$$

for any $i = 1, \ldots, m$.

At the M step of the $(k+1)^{th}$ iteration, to obtain $\phi^{(k+1)}$, the function $Q\left(\phi; \phi^{(k)}\right)$ must be maximized with respect to the $m^2 - m$ parameters $\gamma_{i,j}$, for any $i = 1, \ldots, m; j = 1, \ldots, m-1$, the m parameters μ_i and the m parameters σ_i^2, for any $i \in S_X$. The analytic expression of $Q\left(\phi; \phi^{(k)}\right)$ is the sum of three terms: the first two are functions only of the parameters of the Markov chain, while the third is a function only of the parameters of the Gaussian *pdf*s. This separation of parameters makes the global maximization of $Q\left(\phi; \phi^{(k)}\right)$ into a simple closed-form, that gives the explicit solutions of the maximzation problem. We shall not maximize $Q\left(\phi; \phi^{(k)}\right)$ with respect to the m parameters δ_i, for any $i \in S_X$, because, as we said previously, the initial distribution δ will be estimated by the equality $\delta' = \delta'\Gamma$, after the estimation of the matrix Γ. But, by the stationarity assumption, δ contains informations about the transition probabilities matrix Γ, since $\delta_j = \sum_{i \in S_X} \delta_i \gamma_{i,j}$, for any $j \in S_X$. However, for large T, the effect of δ is negligible; so the first term of the function $Q(\phi; \phi^{(k)})$ can be

ignored searching for the maximum likelihood estimator of $\gamma_{i,j}$, for any i, j (Basawa and Prakasa Rao (1980), pp. 53-54).

Directly solving the M step, it is possible to obtain the explicit formulae of the estimators of $\gamma_{i,j}, \mu_i, \sigma_i^2$:

$$\gamma_{i,j}^{(k+1)} = \frac{\sum_{t=1}^{T-1} f^{(k)}(y_1, \dots, y_T, X_t = i, X_{t+1} = j)}{\sum_{t=1}^{T-1} f^{(k)}(y_1, \dots, y_T, X_t = i)} = \frac{1'_{(T-1)} A_{i,j}^{(k)}}{1'_{(T-1)} B_i^{(k)}}, \quad (5)$$

$$\mu_i^{(k+1)} = \frac{\sum_{t=1}^{T} f^{(k)}(y_1, \dots, y_T, X_t = i) y_t}{\sum_{t=1}^{T} f^{(k)}(y_1, \dots, y_T, X_t = i)} = \frac{1'_{(T)} \left(C_i^{(k)} \odot y \right)}{1'_{(T)} C_i^{(k)}}, \quad (6)$$

$$\sigma_i^{2(k+1)} = \frac{\sum_{t=1}^{T} f^{(k)}(y_1, \dots, y_T, X_t = i) \left(y_t - \mu_i^{(k+1)} \right)^2}{\sum_{t=1}^{T} f^{(k)}(y_1, \dots, y_T, X_t = i)} = \quad (7)$$

$$= \frac{1'_{(T)} \left(C_i^{(k)} \odot \left(y - \mu_i^{(k+1)} 1_{(T)} \right)^2 \right)}{1'_{(T)} C_i^{(k)}},$$

for any state $i = 1, \dots, m$ and $j = 1, \dots, m-1$ of the Markov chain $\{X_t\}$, where

$$A_{i,j}^{(k)} = \begin{bmatrix} \delta_i^{(k)} f^{(k)}(y_1 \mid i) \gamma_{i,j}^{(k)} f^{(k)}(y_2 \mid j) \Gamma_{j\bullet}^{(k)} F_3^{(k)} \Gamma^{(k)} F_4^{(k)} \cdot \dots \cdot \Gamma^{(k)} F_T^{(k)} 1_{(m)} \\ \vdots \\ \delta'^{(k)} F_1^{(k)} \Gamma^{(k)} F_2^{(k)} \cdot \dots \cdot \Gamma^{(k)} F_{t-1}^{(k)} \Gamma_{\bullet i}^{(k)} f^{(k)}(y_t \mid i) \gamma_{i,j}^{(k)} \cdot \\ \cdot f^{(k)}(y_{t+1} \mid j) \Gamma_{j\bullet}^{(k)} F_{t+2}^{(k)} \cdot \dots \cdot \Gamma^{(k)} F_T^{(k)} 1_{(m)} \\ \vdots \\ \delta'^{(k)} F_1^{(k)} \Gamma^{(k)} F_2^{(k)} \cdot \dots \cdot \Gamma^{(k)} F_{T-2}^{(k)} \Gamma_{\bullet i}^{(k)} f^{(k)}(y_{T-1} \mid i) \gamma_{i,j}^{(k)} f^{(k)}(y_T \mid j) \end{bmatrix},$$

$$B_i^{(k)} = \begin{bmatrix} \delta_i^{(k)} f^{(k)}(y_1 \mid i)\Gamma_{i\bullet}^{(k)} F_2^{(k)}\Gamma^{(k)} F_3^{(k)} \cdot \ldots \cdot \Gamma^{(k)} F_T^{(k)} 1_{(m)} \\ \vdots \\ \delta'^{(k)} F_1^{(k)}\Gamma^{(k)} F_2^{(k)} \cdot \ldots \cdot \Gamma^{(k)} F_{t-1}^{(k)}\Gamma_{\bullet i}^{(k)} f^{(k)}(y_t \mid i)\cdot \\ \cdot\Gamma_{i\bullet}^{(k)} F_{t+1}^{(k)} \cdot \ldots \cdot \Gamma^{(k)} F_T^{(k)} 1_{(m)} \\ \vdots \\ \delta'^{(k)} F_1^{(k)}\Gamma^{(k)} F_2^{(k)} \cdot \ldots \cdot \Gamma^{(k)} F_{T-2}^{(k)}\Gamma_{\bullet i}^{(k)} f^{(k)}(y_{T-1} \mid i)\Gamma_{i\bullet}^{(k)} F_T^{(k)} 1_{(m)} \end{bmatrix},$$

$$C_i^{(k)} = \begin{bmatrix} B_i^{(k)} \\ c_i^{(k)} \end{bmatrix};$$

$$c_i^{(k)} = \delta'^{(k)} F_1^{(k)}\Gamma^{(k)} F_2^{(k)} \cdot \ldots \cdot \Gamma^{(k)} F_{T-1}^{(k)}\Gamma_{\bullet i}^{(k)} f^{(k)}(y_T \mid i)$$

and the symbol $\odot$ denotes the Hadamard product.

These explicit formulae of the parameters estimators are obtained when no missing observations are in the sequence of the observed data $y_1, \ldots, y_T$. If the series contains missing observations, expressions (5), (6), (7) must be modified. As related in Subsection 2.1, we must use the w-step transition probabilities matrices and change the structure of the various joint *pdf*s, obtaining the modified versions of the explicit formulae of the estimators. Besides, for the estimators in (6) and (7), we must multiply every y_t and every $\left(y_t - \mu_i^{(k+1)}\right)^2$ by the indicator function I_t, so defined

$$I_t = \begin{cases} 1 & \text{if } y_t \text{ is available} \\ 0 & \text{if } y_t \text{ is missing} \end{cases} \qquad \text{for any } t.$$

4. Application to Old Faithful Geyser data

Geysers are volcanic phenomena whose activity is based on intermittent eruptions of hot and mineralized water. Thus geysers are intermittent thermal springs, from which the water gushes forth by violent jets alternate to break periods.

We consider the series of the duration of the eruptions, in minutes, of the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA, recorded from August 1st until August 8th, 1978 (Silverman (1986)). The series presents no missing observation, but we artificially put them in: it is not a strangeness that there are some missing observations in a geyser eruptions series, e.g. Azzalini and Bowman

(1990) (we could not analyse that series because the number of missing observations is too high). We chose three isolated missing observations, at time $t = 20, 52, 78$ and eight missing observations gathered in two separated blocks of three (at time $t = 34, 35, 36$) and five (from time $t = 90$ to time $t = 94$) values.

The series is plotted in Figure 1, in which the presence of two hidden states, representing two different levels of the volcanic activity, is evident. The presence of two hidden states, noticing a two-peaks distribution, is confirmed by the continuous approximation of the histogram plotted in Figure 2.

The iterative procedure for the identification of the parameters of GHMMs, introduced in Section 3, has been implemented in a GAUSS code.

As we have already observed, the choice of the starting values is a matter of primary importance to identify the global maximum, given that the log-likelihood surface for HMMs is often irregular and characterized by many local maxima. Therefore the code repeats more than once the iterative procedure, starting from several different points, randomly chosen in the parameter space Φ and we compare the stationary points obtained at each run, choosing that with the largest log-likelihood value. Furthermore δ has been assumed known and fixed for any iteration of the EM algorithm, given that the initial distribution is non-informative about the transition probabilities.

The variance-covariance matrix of the parameters estimates are obtained from the inverse of a numerical approximation of the Hessian matrix with reverse sign.

To estimate the dimension m of the state-space of the Markov chain, according to Leroux and Puterman (1992), we use the *Akaike Information Criterion* (AIC) and the *Bayesian Information Criterion* (BIC): we search for that special value m^* which maximizes the difference $\ln L_T^{(m)}(\phi) - a_{m,T}$, where $\ln L_T^{(m)}(\phi)$ is the log-likelihood function maximized over a HMM with an m-state Markov chain, while $a_{m,T}$ is the penalty term depending on the number of states m and the length T of the observed sequence. If $a_{m,T} = d_m$, where d_m is the dimension of the model, that is the number of the parameters estimated with the EM algorithm $\left(\text{i.e. } m^2 + m\right)$, we have the AIC; if $a_{m,T} = (\ln T)d_m/2$, we have the BIC.

In the series of the duration of geyser eruptions, $y_1, \ldots, y_{107}$, the values y_{20}, $y_{34} - y_{36}$, y_{52}, y_{78}, $y_{90} - y_{94}$ have been dropped out of the

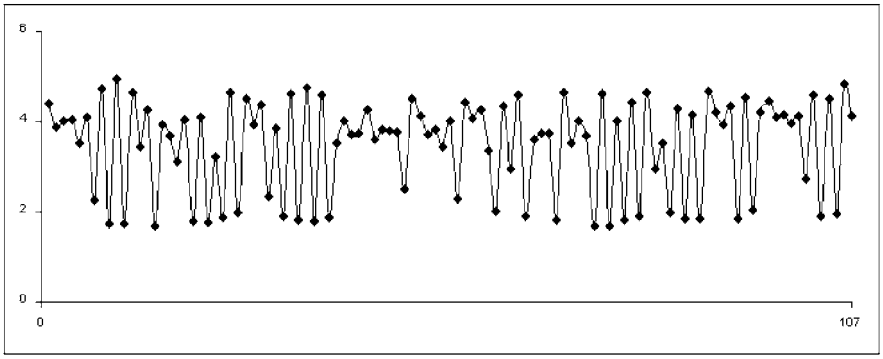

Fig. 1. Series of the duration of the eruptions of the Old Faithful Geyser, from 1.8 to 8.8 1978

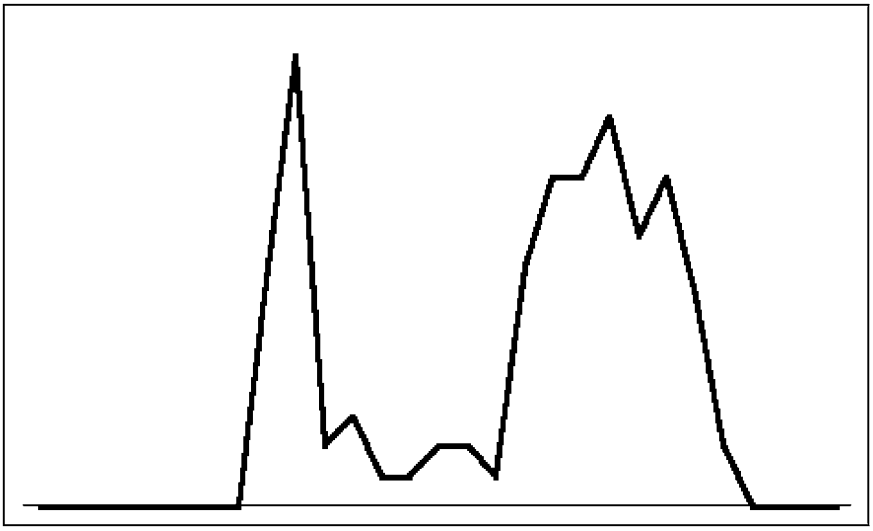

Fig. 2. Continuous approximation of the histogram of the data plotted in Fig. 1

sample; so, as we saw in Subsection 2.1, we have to consider the w-step transition probabilities $\gamma_{i_{19},i_{21}}(2)$, $\gamma_{i_{33},i_{37}}(4)$, $\gamma_{i_{51},i_{53}}(2)$, $\gamma_{i_{77},i_{79}}(2)$, $\gamma_{i_{89},i_{95}}(6)$. According to (4), the likelihood function of the observed

data, is

$$L_{107}(\phi) = \delta' \left(\prod_{t=1}^{19} G_t \right) \Gamma \left(\prod_{t=21}^{33} G_t \right) \Gamma^3 \left(\prod_{t=37}^{51} G_t \right) \cdot$$

$$\cdot \Gamma \left(\prod_{t=53}^{77} G_t \right) \Gamma \left(\prod_{t=79}^{89} G_t \right) \Gamma^5 \left(\prod_{t=95}^{107} G_t \right) 1_{(m)}.$$

In the same way, the w-step transition probabilities will be adopted to obtain the explicit formulae of the estimators $\gamma_{i,j}^{(k+1)}$, $\mu_i^{(k+1)}$, $\sigma_i^{2(k+1)}$ replacing in vectors $A_{i,j}^{(k)}$, $B_i^{(k)}$, $C_i^{(k)}$, in expressions (5), (6), (7), the matrices $F_{20}^{(k)}$, $F_{34}^{(k)} - F_{36}^{(k)}$, $F_{52}^{(k)}$, $F_{78}^{(k)}$, $F_{90}^{(k)} - F_{94}^{(k)}$ with the identity matrix and, for any other t, $F_t^{(k)} = \text{diag}\left[f^{(k)}(y_t \mid 1),\ f^{(k)}(y_t \mid 2), \ldots,\ f^{(k)}(y_t \mid m) \right]$.

Performing the EM algorithm for a range of values of the number of states of the Markov chain ($m = 1, \ldots, 4$), we obtain the following maximized values of the log-likelihood and the corresponding AIC and BIC:

m	log-likelihood	AIC	BIC
1	−142.62739	−144.62739	−147.3002
2	−125.91443	**−131.91443**	**−139.9329**
3	−125.91442	−137.91442	−153.9514
4	−125.91443	−145.91443	−172.6427

Considering both the AIC and the BIC as model selection criteria, we choose a 2-states Markov chain. The sequence of the log-likelihood $\left\{ \ln L_{107}\left(\phi^{(k)}\right) \right\}$ converges at the 41^{th} iteration to $\ln L_{107}\left(\phi^{(41)}\right) = -125.91443$; the estimates of the parameters (standard errors in brackets) of the 2 Gaussian *pdf*s are

i	1	2
$\mu_i^{(41)}$	1.6263	3.4786
	(0.0000)	(0.0000)
$\sigma_i^{2(41)}$	0.0470	0.5546
	(0.0000)	(0.0000)

The estimate of the transition probabilities matrix of the Markov chain (standard errors in brackets) is

$$\Gamma^{(41)} = \begin{bmatrix} 0 & 1 \\ (0.0077) & (0.0077) \\ 0.2318 & 0.7682 \\ (0.1038) & (0.1038) \end{bmatrix}$$

from which we have the estimate of the stationary initial distribution

$$\delta^{(41)} = (0.1882; 0.8118)' .$$

From the diagonal entries of the transition probabilities matrix, it is also possible to compute the time spent in state i of the Markov chain upon each return to it, which has a geometric distribution with mean $1/(1-\gamma_{i,i})$; hence the mean number of consecutive eruptions occuring in state i is

i	1	2
eruptions	1	4.3144

Observing Figure 1, it is possible to see that no eruption in the low level is followed by another eruption in the same level; for this reason we have $\gamma_{1,1} = 0$ and so 1 is the mean number of consecutive eruptions in state 1.

5. Conclusions

In this paper special *hidden Markov models* (HMMs) $\{Y_t; X_t\}$ used to study univariate non-linear time series have been introduced. They are called *Gaussian hidden Markov models* (GHMMs), because every observed variable Y_t, given a special state i of the Markov chain at time t, is a Gaussian random variable with unknown parameters μ_i and σ_i^2. The attention has been focused on the estimation of the parameters $\delta_i, \gamma_{i,j}, \mu_i, \sigma_i^2$, for any state i, j of the Markov chain state-space, when missing observations occur. The estimators of $\gamma_{i,j}, \mu_i, \sigma_i^2$ have been obtained by the maximum likelihood method performing the EM algorithm, while the estimators of δ_i have been obtained by means of the equality $\delta' = \delta'\Gamma$. We could solve the M step exactly, without using a numerical maximization algorithm, such as the Newton-Raphson

method and obtain the explicit formulae of the estimators of the parameters. Hence the procedure is more stable and converges faster in the neighborhood of the maximum. Furthermore an application of GHMMs to geyser data have been shown and the estimates of the parameters of the model computed. In this application, the dimension m of the Markov chain state-space has been estimated by two maximum penalized likelihood methods, the *Akaike Information Criterion* (AIC) and the *Bayes Information Criterion* (BIC).

ACKNOWLEDGMENTS

The authors would like to thanks the referee for his very constructive comments. This research has been supported by the Italian Ministry of University and Scientific Research (MURST) 2000 Grant "Statistics in Environmental Risk Evaluation".

REFERENCES

ALBERT, P.S. (1991) A Two-State Markov Mixture Model for a Time Series of Epileptic Seizure Counts, *Biometrics*, 47, 1371-1381.

AZZALINI, A. and BOWMAN, A.W. (1990) A Look at Some Data on the Old Faithful Geyser, *Applied Statistics*, 39, 357-365.

BASAWA, I.V. and PRAKASA RAO, B.L.S. (1980) *Statistical Inference for Stochastic Processes*, Academic Press, London.

BAUM, L.E., PETRIE, T., SOULES, G., and WEISS, N. (1970) A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains, *The Annals of Mathematical Statistics*, 41, 164-171.

BICKEL, P.J., RITOV, Y., and RYDÉN, T. (1998) Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models, *The Annals of Statistics*, 26, 1614-1635.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with Discussion), *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

ELLIOTT, R.J., AGGOUN, L., and MOORE, J.B. (1995) *Hidden Markov Models: Estimation and Control*, Springer, New York.

FREDKIN, D.R. and RICE, J.A. (1992) Maximum likelihood estimation and identification directly from single-channel recordings, *Proceedings of the Royal Society of London, Series B*, 249, 125-132.

JUANG, B.H. and RABINER, L.R. (1991) Hidden Markov Models for Speech Recognition, *Technometrics*, 33, 251-272.

LEROUX, B.G. (1992) Maximum-likelihood estimation for hidden Markov models, *Stochastic Processes and their Applications*, 40, 127-143.

LEROUX, B.G. and PUTERMAN, M.L. (1992) Maximum-Penalized-Likelihood Estimation for Independent and Markov-Dependent Mixture Models, *Biometrics*, 48, 545-558.

MACDONALD, I.L. and ZUCCHINI, W. (1997) *Hidden Markov and Other Models for Discrete-valued Time Series*, Chapman & Hall, London.

MCLACHLAN, G.J. and KRISHNAN, T. (1997) *The EM algorithm and extensions*, John Wiley & Sons, New York.

SILVERMAN, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

WU, C.F.J. (1983) On the Convergence Properties of the EM Algorithm, *The Annals of Statistics*, 11, 95-103.

Parameter estimation of Gaussian hidden Markov models when missing observations occur

SUMMARY

We examine hidden Markov models in which the probability density function of every observed variable, given a state of the Markov chain, is Gaussian. The aim of this paper is to show a methodology to obtain the maximum likelihood estimators of the parameters of this class of models that can be computed also when the time series contains missing observations. Our methodology, based on the EM algorithm, is explained analysing a time series about the duration of geyser eruptions.

Stima dei parametri di un modello markoviano latente gaussiano con osservazioni mancanti

RIASSUNTO

In questo lavoro si considerano quei particolari modelli markoviani latenti in cui la funzione di densità di probabilità di ogni variabile osservata, dato lo stato della catena di Markov, è gaussiana. Si vuole qui mostrare una metodologia per ottenere gli stimatori di massima verosimiglianza dei parametri di questa classe di modelli che sia valida anche nel caso in cui la successione di osservazioni sia incompleta. Questa metodologia, che si basa sull'algoritmo EM, è illustrata attraverso l'analisi di una serie storica relativa alla durata delle eruzioni di un geyser.

KEY WORDS

Discrete-time stochastic processes; Markov chains; Maximum likelihood estimators; EM algorithm; Geyser data.

[Manuscript received March 2002; final version received June 2002.]