GAUSSIAN HIDDEN MARKOV MODELS FOR THE ANALYSIS OF THE DYNAMICS OF SULPHUR DIOXIDE

Roberta Paroli
Istituto di Statistica
Università Cattolica S.C. di Milano
rparoli@mi.unicatt.it

Luigi Spezia
Dipartimento di Ingegneria
Università degli Studi di Bergamo
spezia@unibg.it

Research Group for Statistical Applications
to Environmental Problems

Working Paper n. 9

November 2000

Abstract

Hidden Markov models (HMMs) are sequences of conditionally independent random variables and every observed variable depends only on the contemporary state of an unobserved Markov chain. Here we examine HMMs 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 how the maximum likelihood estimators of the parameters of these models may be suitably obtained performing the EM algorithm. These estimators can be used also when the sequence of observations contains unrecorded data. An application about a time series of hourly mean concentrations of sulphur dioxide with unrecorded data will be shown.

Keywords: discrete time stochastic processes, Markov chains, maximum likelihood estimators, EM algorithm, unrecorded data.

Introduction[1]

Air quality control includes the study of data sets recorded by air pollution testing stations. We consider one of the five stations situated in Bergamo (a town in Northern Italy, with 116.000 inhabitants). It records seven types of pollutants and each hour we have the mean concentration of every pollutant. We are interested in the analysis of the dynamics of hourly mean concentrations of sulphur dioxide (SO_2). SO_2, measured in $\mu g/m^3$, is a gas with a characteristic pungent and choking smell: it is produced by the combustion of substances containing sulphur (coal, fuel oil, diesel oil). The principal sources of SO_2 are the industries that need a lot of energy (refineries, steelworks, thermoelectric power stations), the domestic heating (even if the use of methane drastically reduced the SO_2 emissions), the traffic of the heavy transport (diesel vehicular traffic). SO_2 is responsible, with the nitrogen oxides (NO_X), for the acid rains. It can cause the onset and the worsening of the respiratory tract diseases (persistent cough, bronchitis, sinusitis); very high concentrations can cause tissues destruction, giving rise to emphysema. Figure 1 shows the series of the hourly mean concentrations of SO_2 recorded by the air pollution testing station placed in Via Goisis, Bergamo, from Monday the 16th of November to Sunday the 13th of December, 1998. The series is characterizad by an asymmetric cyclic behaviour, increasing at a faster rate than decreasing. As Chatfield remarks "a non-linear model is much more compelling for describing series with properties such as 'going up faster than coming down' " (Chatfield (1996), p. 195). Figure 2 shows the logarithmic transformation of the series, that has been applied to reduce its variability: the same asymmetric features of the original series may be seen in the plot. Hence we will study the series of the logarithms of the hourly mean concentrations of SO_2 by a non-linear model, which assumes the existence of an unobserved process, modelled by a Markov chain, whose dynamics affects the dynamics of the observed process. Notice that in Figures 1 and 2 some observations are missing, either because the station must be stopped every twenty-four hours for automatic calibration, because of occasional mechanical failure, ordinary maintenance, or because some data are clearly anomalous and so are removed by the technicians.

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

[1]Work partially supported by MURST 1998 Grant "Statistical Methods for the Analysis of the Environment and the Health-Environment Interactions"

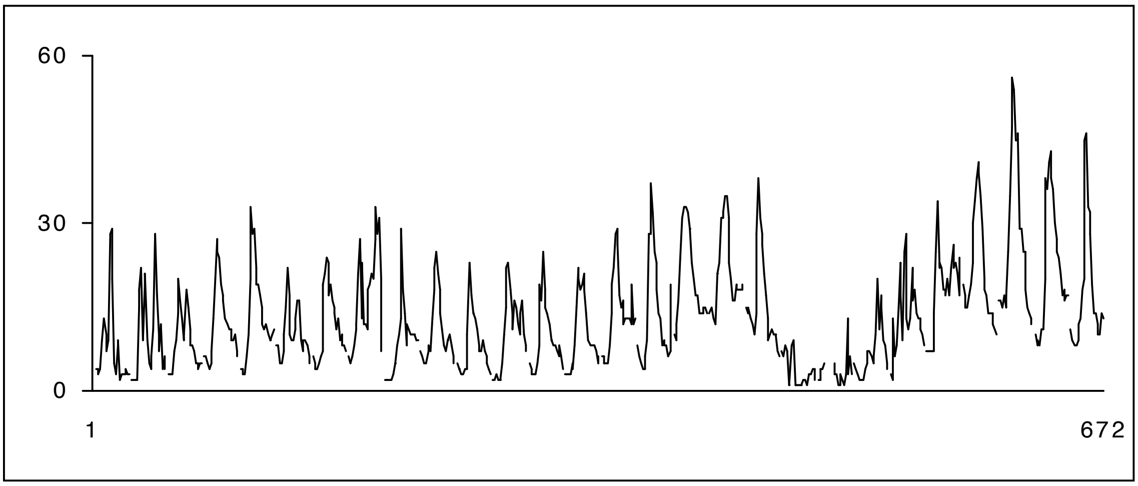


Figure 1: *Series of the hourly mean concentrations of SO2, recorded by the air pollution testing station placed in Via Goisis, Bergamo, from Monday the 16th of November, 1998, 1:00 a.m. (t=1), to Sunday the 13th of December, 1998, 12:00 p.m. (t=672).*

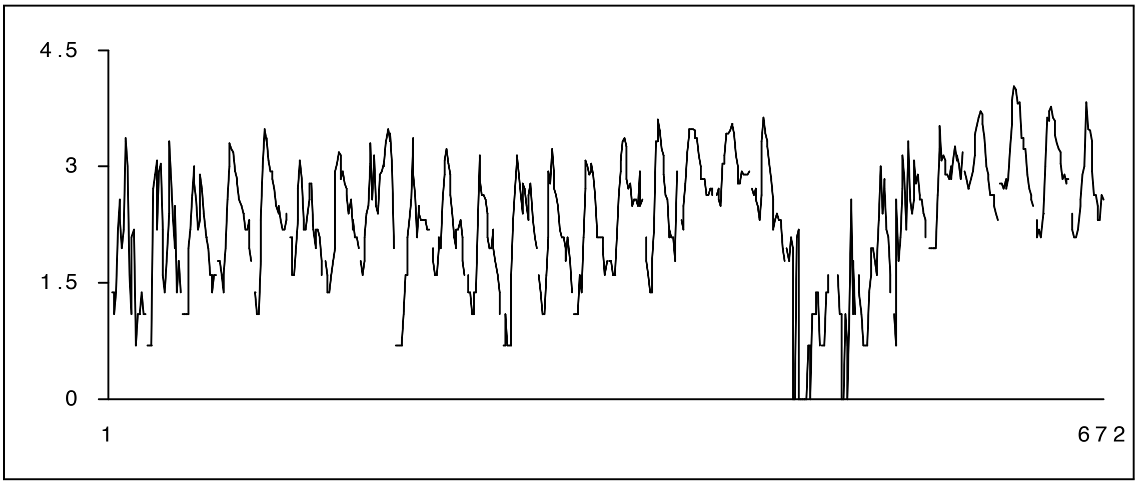
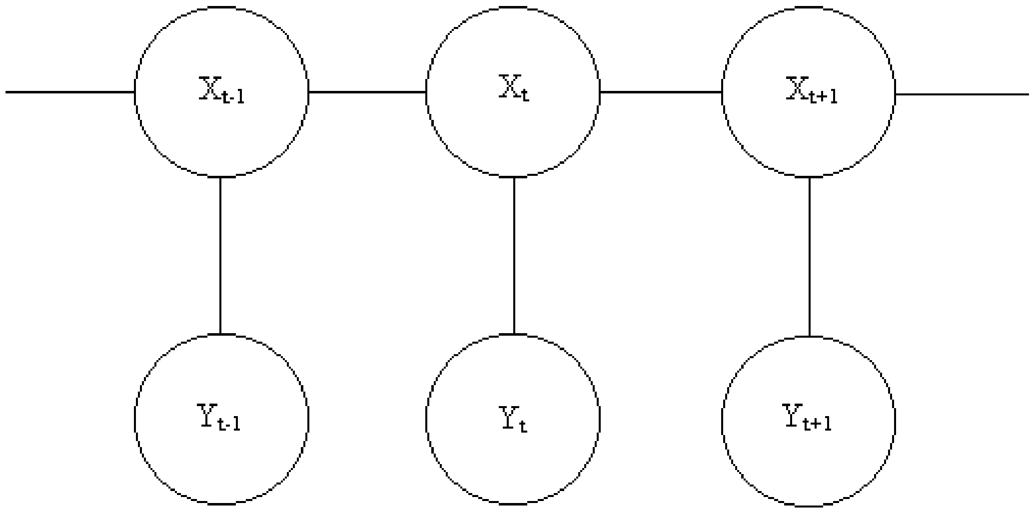


Figure 2: *The logarithmic transformation of the series plotted in Figure 1.*

(*conditional independence condition*) with the conditional distribution of Y_t depending on $\{X_t\}$ only through X_t (*contemporary dependence condition*).

A useful way to show the dependence structure of HMMs is the following simple graph in which the edges between the vertices representing the Markov chain $\{X_t\}$ indicate the Markov dependence condition; the edges between the vertices representing the observed process $\{Y_t\}$ and the vertices of the contemporary states of the chain indicate the contemporary dependence condition; the lack of edges between the vertices representing the observed process $\{Y_t\}$ indicates the conditional independence condition:



In this paper we examine HMMs in which the probability density 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).

The aim of this paper is to show how the maximum likelihood estimators of GHMMs may be suitably obtained using the EM algorithm. 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 maximum likelihood estimators of the unknown parameters of GHMMs; finally, in Section 4, an application of GHMMs will be shown: we will examine the data set about the hourly mean concentrations of SO_2, where within the series of 672 observations, we have 33 means that have not been recorded or validated.

1 The basic Gaussian Hidden Markov Model

Let $\{X_t\}_{t\in(1,\ldots,T)\subset\mathbb{N}}$ be a discrete, homogeneous, aperiodic, irreducible Markov chain on a finite state-space $S_X = \{1, 2, \ldots, m\}$. The transition probability from state i, at time $t-1$, to state j, at time t, is denoted by $\gamma_{i,j}$, for any state i, j and for any time t: $\gamma_{i,j} = \mathrm{P}(X_t = j \mid X_{t-1} = i) = \mathrm{P}(X_2 = j \mid X_1 = i)$. The transition probabilities $(m \times m)$ matrix is $\Gamma = [\gamma_{i,j}]$, with $\sum_{j\in S_X} \gamma_{i,j} = 1$, for any $i \in S_X$. The initial distribution is the vector $\delta = (\delta_1, \delta_2, \ldots, \delta_m)'$, where $\delta_i = \mathrm{P}(X_1 = i)$, for any $i = 1, 2, \ldots, m$, with $\sum_{i\in S_X} \delta_i = 1$. 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, for any time t, $\delta_i = \mathrm{P}(X_t = i)$, for any state $i = 1, 2, \ldots, m$. Since δ is a stationary distribution, the equality $\delta' = \delta'\Gamma$ holds: δ is the left eigenvector of the matrix Γ, associated with the eigenvalue one, which always exists, because Γ is a stochastic matrix. Finally, the hypothesis characterizing HMMs is that the Markov chain $\{X_t\}$ is unobservable.

Let $\{Y_t\}_{t\in(1,\ldots,T)\subset\mathbb{N}}$ 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 $(Y_1, \ldots, Y_T)$, given the variables $(X_1, \ldots, X_T)$, 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 called *gaussian hidden Markov model* (GHM M) 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 model is called *Markov* because the unobserved sequence of states is the realization of a Markov chain; the model is called *hidden* because the sequence of realizations of the stochastic process $\{Y_t\}$ is observed, but not the sequence of states of the Markov chain, which is hidden in the observations; the model is called *gaussian* because $f(y \mid i)$ is a gaussian *pdf*.

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.

Given that $\{X_t\}$ is a homogeneous, irreducible Markov chain, defined on a finite state-space, it is a strongly stationary process and also the observed process $\{Y_t\}$ is strongly stationary; hence Y_t, for any t, has the same marginal distribution, obtained by applying the definition of conditional probability:

$$f(y_t) = \sum_{i\in S_X} f(y_t \mid i)\mathrm{P}(X_t = i) = \sum_{i\in S_X} \delta_i f(y_t \mid i),$$

which is a mixture of gaussian *pdf*s.

In this paper the procedure to estimate the unknown parameters of GHMMs $\{Y_t; X_t\}$ 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 Γ is a stochastic matrix and therefore each row sum equals 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 unrecorded data; 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 $(Y_1, \ldots, Y_T)$

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*

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

Setting $F_t = \mathrm{diag}(f(y_t \mid 1), f(y_t \mid 2), \ldots, f(y_t \mid m))$, 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

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

We can observe that the joint *pdf* $f(y_1, \ldots, y_T)$ may be computed even if some data are not available. If, for example, a subsequence of $w-1$ observations, $y_{v+1}, \ldots, y_{v+w-1}$, is not available within a sequence $y_1, \ldots, y_T$, with $1 < v+1 \leq v+w-1 < T$, the *pdf* (2) becomes

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

where $\gamma_{i,j}(w)$ is the w-step transition probability, $\gamma_{i,j}(w) = \mathrm{P}(X_{v+w} = j \mid X_v = i) = \mathrm{P}(X_{1+w} = j \mid X_1 = i)$; the w-step transition probabilities matrix is $\Gamma(w) = [\gamma_{i,j}(w)]$ 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

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

We separately analyze the following three situations: $t = 1$, $1 < t < T$, $t = T$; henceforth we shall denote the i^{th} row of Γ with $\Gamma_{i\bullet}$ and the i^{th} column of Γ with $\Gamma_{\bullet i}$.

(a) $t = 1$:

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

In fact: summing $f(y_1, \ldots, y_T, X_1 = i, X_2 = i_2, \ldots, X_T = i_T)$ over $i_2, \ldots, i_T$ and applying the conditional independence, the contemporary dependence and the Markov dependence conditions, we obtain

$$\begin{aligned}
&f(y_1, \ldots, y_T, X_1 = i) = \\
&= \sum_{i_2 \in S_X} \cdots \sum_{i_T \in S_X} \delta_i \gamma_{i,i_2} \cdot \ldots \cdot \gamma_{i_{T-1},i_T} f(y_1 \mid i) f(y_2 \mid i_2) \cdot \ldots \cdot f(y_T \mid i_T) = \\
&= \delta_i f(y_1 \mid i) \Gamma_{i\bullet} F_2 \Gamma F_3 \cdot \ldots \cdot \Gamma F_T 1_{(m)} \;\square
\end{aligned}$$

(b) $1 < t < T$:

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

In fact: summing $f(y_1,\ldots,y_T,X_1=i_1,\ldots,X_{t-1}=i_{t-1},X_t=i,X_{t+1}=i_{t+1},\ldots,X_T=i_T)$ over $i_1,\ldots,i_{t-1},i_{t+1},\ldots,i_T$ and applying the conditional independence, the contemporary dependence and the Markov dependence conditions, we obtain

$f(y_1,\ldots,y_T,X_t=i) =$

$= \sum_{i_1 \in S_X} \cdots \sum_{i_{t-1} \in S_X} \sum_{i_{t+1} \in S_X} \cdots \sum_{i_T \in S_X} \delta_{i_1} \gamma_{i_1,i_2} \cdot \ldots \cdot \gamma_{i_{t-1},i} \gamma_{i,i_{t+1}} \cdot \ldots \cdot \gamma_{i_{T-1},i_T} \cdot$
$\cdot f(y_1 \mid i_1) \cdot \ldots \cdot f(y_{t-1} \mid i_{t-1}) f(y_t \mid i) f(y_{t+1} \mid i_{t+1}) \cdot \ldots \cdot f(y_T \mid i_T) =$

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

(c) $t = T$:

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

In fact: summing $f(y_1,\ldots,y_T,X_1=i_1,\ldots,X_{T-1}=i_{T-1},X_T=i)$ over $i_1,\ldots,i_{T-1}$ and applying the conditional independence, the contemporary dependence and the Markov dependence conditions, we obtain

$f(y_1,\ldots,y_T,X_T=i) =$

$= \sum_{i_1 \in S_X} \cdots \sum_{i_{T-1} \in S_X} \delta_{i_1} \gamma_{i_1,i_2} \cdot \ldots \cdot \gamma_{i_{T-1},i} f(y_1 \mid i_1) \cdot \ldots \cdot f(y_{T-1} \mid i_{T-1}) f(y_T \mid i) =$
$= \delta' F_1 \Gamma F_2 \cdot \ldots \cdot \Gamma F_{T-1} \Gamma_{\bullet i} f(y_T \mid i)$ □

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 the 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

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

We separately analyze the following three situations: $t=1$, $1<t<T-1$, $t=T-1$.

(a) $t = 1$:

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

In fact: summing $f(y_1,\ldots,y_T,X_1=i,X_2=j,X_3=i_3,\ldots,X_T=i_T)$ over $i_3,\ldots,i_T$ and applying the conditional independence, the contemporary dependence and the Markov dependence conditions, we obtain

$f(y_1,\ldots,y_T,X_1=i,X_2=j) =$

$$= \sum_{i_3 \in S_X} \cdots \sum_{i_T \in S_X} \delta_i \gamma_{i,j} \gamma_{j,i_3} \cdot \ldots \cdot \gamma_{i_{T-1},i_T} f(y_1 \mid i) f(y_2 \mid j) f(y_3 \mid i_3) \cdot \ldots \cdot f(y_T \mid i_T) =$$

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

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

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

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

In fact: summing $f(y_1, \ldots, y_T, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = i, X_{t+1} = j, X_{t+2} = i_{t+2}, \ldots, X_T = i_T)$ over $i_1, \ldots, i_{t-1}, i_{t+2}, \ldots, i_T$ and applying the conditional independence, the contemporary dependence and the Markov dependence conditions, we obtain

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

$$= \sum_{i_1 \in S_X} \cdots \sum_{i_{t-1} \in S_X} \sum_{i_{t+2} \in S_X} \cdots \sum_{i_T \in S_X} \delta_{i_1} \gamma_{i_1,i_2} \cdot \ldots \cdot \gamma_{i_{t-1},i} \gamma_{i,j} \gamma_{j,i_{t+2}} \cdot \ldots \cdot \gamma_{i_{T-1},i_T} \cdot$$
$$\cdot f(y_1 \mid i_1) \cdot \ldots \cdot f(y_{t-1} \mid i_{t-1}) f(y_t \mid i) f(y_{t+1} \mid j) f(y_{t+2} \mid i_{t+2}) \cdot \ldots \cdot f(y_T \mid i_T) =$$

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

(c) $t = T-1$:

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

In fact: summing $f(y_1, \ldots, y_T, X_1 = i_1, \ldots, X_{T-2} = i_{T-2}, X_{T-1} = i, X_T = j)$ over $i_1, \ldots, i_{T-2}$ and applying the conditional independence, the contemporary dependence and the Markov dependence conditions, we obtain

$$f(y_1, \ldots, y_T, X_{T-1} = i, X_T = j) =$$

$$= \sum_{i_1 \in S_X} \cdots \sum_{i_{T-2} \in S_X} \delta_{i_1} \gamma_{i_1,i_2} \cdot \ldots \cdot \gamma_{i_{T-2},i} \gamma_{i,j} \cdot$$
$$\cdot f(y_1 \mid i_1) \cdot \ldots \cdot f(y_{T-2} \mid i_{T-2}) f(y_{T-1} \mid i) f(y_T \mid j) =$$

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

3 Parameters estimation of GHMMs

To obtain the estimators of unknown parameters by the maximum likelihood method, the likelihood function, or equivalently the log-likelihood function, must be maximized with respect to the parameters. Hence we must determine the roots of the likelihood system. Because in practice this system cannot be constructed in an analytic closed form and directly solved, given that the data are incomplete, we perform an iterative numerical method of maximization: the EM algorithm.

The name EM was given by Dempster, Laird, Rubin (1977) because the algorithm is based on 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* (for details on the EM algorithm, see McLachlan and Krishnan (1997)).

Let $y = (y_1, \dots, y_T)'$ 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 of the states of the chain $\{X_t\}$ is missing (or *hidden*). Let $x = (i_1, \dots, i_T)'$ be the vector of the unobserved states of the chain $\{X_t\}$; hence $z = (i_1, y_1, \dots, i_T, y_T)'$ is the vector of the complete data. Moreover let $L_T^c(\phi)$ be the likelihood function of complete data and $L_T(\phi)$ be that of observed data:

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

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 $Q\left(\phi; \phi^{(0)}\right)$, that is the special $\phi^{(1)}$, such that

$$Q\left(\phi^{(1)}; \phi^{(0)}\right) \geq Q\left(\phi; \phi^{(0)}\right),$$

for any $\phi \in \Phi$. At the second iteration, the E and M steps must be repeated, replacing $\phi^{(0)}$ with $\phi^{(1)}$. 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 $Q\left(\phi; \phi^{(k)}\right)$, i.e. such that

$$Q\left(\phi^{(k+1)}; \phi^{(k)}\right) \geq Q\left(\phi; \phi^{(k)}\right),$$

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 $\left\{\ln L_T\left(\phi^{(k)}\right)\right\}$, i.e. until the difference

$$\ln L_T\left(\phi^{(k+1)}\right) - \ln L_T\left(\phi^{(k)}\right)$$

is not greater than a sufficiently small arbitrary value. The entries of the vector $\phi^{(k+1)}$ are then the maximum likelihood estimators.

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 proved $\ln L_T(\phi)$ is not decreased after any EM iteration, i.e.

$$\ln L_T\left(\phi^{(k+1)}\right) \geq \ln L_T\left(\phi^{(k)}\right)$$

for any $k = 0, 1, 2, \ldots$ (Dempster, Laird, Rubin (1977), Theorem 1).

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.

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 $\delta'^{(k)} = \delta'^{(k)}\Gamma^{(k)}$.

Proposition 1 - *The function $Q\left(\phi; \phi^{(k)}\right)$ obtained at the E step of the $(k+1)^{th}$ iteration of the EM algorithm is*

$$\begin{aligned} Q\left(\phi;\phi^{(k)}\right) &= \mathrm{E}_{\phi^{(k)}}\left(\ln L_T^c(\phi) \mid y\right) = \\ &= \sum_{i\in S_X} \frac{f^{(k)}(y_1,\ldots,y_T,X_1=i)}{f^{(k)}(y_1,\ldots,y_T)} \ln \delta_i + \\ &+ \sum_{i\in S_X}\sum_{j\in S_X} \frac{\sum_{t=1}^{T-1} f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)}{f^{(k)}(y_1,\ldots,y_T)} \ln \gamma_{i,j} + \\ &+ \sum_{i\in S_X} \frac{\sum_{t=1}^{T} f^{(k)}(y_1,\ldots,y_T,X_t=i)}{f^{(k)}(y_1,\ldots,y_T)} \ln f(y_t \mid i), \end{aligned} \tag{5}$$

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 $i = 1, \ldots, m$.

Proof - See Appendix □

This 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. 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 $i \in S_X$. 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).

Now the explicit formulae of the estimators of $\gamma_{i,j}, \mu_i, \sigma_i^2$ are introduced.

Proposition 2 - *Given a generic HMM $\{Y_t; X_t\}$, the expression of the unknown transition probability $\gamma_{i,j}$ of the Markov chain $\{X_t\}$, obtained at the $(k+1)^{th}$ iteration of the EM algorithm, is, for large T,*

$$\gamma_{i,j}^{(k+1)} = \frac{\sum_{t=1}^{T-1} f^{(k)}(y_1, \dots, y_T, X_t = i, X_{t+1} = j)}{\sum_{t=1}^{T-1} f^{(k)}(y_1, \dots, y_T, X_t = i)} = \frac{1'_{(T-1)} A_{i,j}^{(k)}}{1'_{(T-1)} B_i^{(k)}}, \tag{6}$$

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

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

and

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

Proof - See Appendix □

Notice that the explicit expression of the estimator $\gamma_{i,j}^{(k+1)}$ holds for the generic HMM, not only for the gaussian case.

Proposition 3 - *Given a GHMM $\{Y_t; X_t\}$, the expressions of the unknown parameters μ_i and σ_i^2 of the gaussian random variable $Y_{t(i)}$, for any t, obtained at the $(k+1)^{th}$ iteration of the EM algorithm, are*

$$\mu_i^{(k+1)} = \frac{\sum_{t=1}^{T} f^{(k)}(y_1, \dots, y_T, X_t = i)\ y_t}{\sum_{t=1}^{T} f^{(k)}(y_1, \dots, y_T, X_t = i)} = \frac{1'_{(T)} \left(C_i^{(k)} \odot y \right)}{1'_{(T)} C_i^{(k)}} \tag{7}$$

and

$$\sigma_i^{2(k+1)} = \frac{\sum_{t=1}^{T} f^{(k)}(y_1, \dots, y_T, X_t = i)\ \left(y_t - \mu_i^{(k+1)}\right)^2}{\sum_{t=1}^{T} f^{(k)}(y_1, \dots, y_T, X_t = i)} = \frac{1'_{(T)} \left(C_i^{(k)} \odot \left(y - \mu_i^{(k+1)} 1_{(T)} \right)^2 \right)}{1'_{(T)} C_i^{(k)}}, \tag{8}$$

for any state i of the Markov chain $\{X_t\}$, where

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

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

Proof - See Appendix □

In the case of convergence of the algorithm at the $(k+1)^{th}$ iteration, $\left(\phi^{(k+1)}; \ln L_T\left(\phi^{(k+1)}\right)\right)$ is a stationary point of $\ln L_T(\phi)$, and therefore $\phi^{(k+1)} = \left(\gamma_{1,1}^{(k+1)}, \ldots, \gamma_{1,m-1}^{(k+1)}, \ldots, \gamma_{m,1}^{(k+1)}, \ldots, \gamma_{m,m-1}^{(k+1)}, \mu_1^{(k+1)}, \ldots, \mu_m^{(k+1)}, \sigma_1^{2(k+1)}, \ldots, \sigma_m^{2(k+1)}\right)'$ is the maximum likelihood estimator of the unknown parameter ϕ. This fundamental result holds because 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.

Propositions 2 and 3 give the explicit formulae of the parameters estimators when the sequence of the observed data $y_1, \ldots, y_T$ is complete, i.e. no missing observations are in the sequence. The sequence of hourly mean concentrations of SO_2 we are studying, however, contains missing observations; so expressions (6), (7), (8) 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 new versions of the explicit formulae of the estimators. Besides, for the estimators in (7) and (8), 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

$$I_t = \begin{cases} 1 & \text{if } y_t \text{ has been recorded at time } t \\ 0 & \text{if } y_t \text{ has been unrecorded at time } t \end{cases}.$$

4 Application to air pollution data

The foregoing iterative procedure for the identification of the parameters of GHMMs 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 repeates 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 two maximum penalized likelihood methods, that is 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. Depending on the value of $a_{m,T}$, we have two special maximum penalized likelihood criteria. 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 (i.e., $m^2 + m$), we have the *Akaike Information Criterion* (AIC); if $a_{m,T} = (\ln T)d_m/2$, we have the *Bayesian Information Criterion* (BIC).

As m increases, also the risk that the Hessian matrix is not invertible increases; hence the optimal model we will consider is that maximizing the AIC and the BIC under the constraint that the standard errors of the estimates exist.

Now we examine in detail the sequence of the time series about the hourly mean concentrations of SO_2 in the 46th, 47th, 48th, 49th weeks of 1998, recorded by the air pollution testing station placed in Via Goisis, Bergamo. As Figure 1 and 2 (in Introduction) show, some data have not been recorded or validated, but, as we saw in Subsection 2.1, it is not difficult to estimate the parameters of the model because the likelihood function may be obtained even if some data are not available. In the series of hourly mean concentrations $y_1, \ldots, y_{672}$, the values $y_{2+24\tau}$ $(\tau = 0, \ldots, 27)$, y_{488}, y_{489}, y_{490}, y_{491}, y_{492} have not been recorded; so we have to consider the w-step transition probabilities $\gamma_{i_{1+24\tau}, i_{3+24\tau}}(2)$, for any τ and $\gamma_{i_{487}, i_{493}}(6)$. Hence the likelihood function of the observed data, according to (4), is

$$\begin{aligned} L_{672}(\phi) \;=\; & \delta' F_1 \prod_{\tau=0}^{19} \left[\Gamma \left(\prod_{t=3+24\tau}^{25+24\tau} G_t \right) \right] \Gamma \left(\prod_{t=483}^{487} G_t \right) \Gamma^5 \left(\prod_{t=493}^{505} G_t \right) \cdot \\ & \cdot \prod_{\tau=0}^{5} \left[\Gamma \left(\prod_{t=507+24\tau}^{529+24\tau} G_t \right) \right] \Gamma \left(\prod_{t=651}^{672} G_t \right) 1_{(m)}. \end{aligned}$$

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 (6), (7), (8), the matrices $F_{2+24\tau}^{(k)}$ $(\tau = 0, \ldots, 27)$, $F_{488}^{(k)}$, $F_{489}^{(k)}$, $F_{490}^{(k)}$, $F_{491}^{(k)}$, $F_{492}^{(k)}$ with the identity matrix.

Performing the EM algorithm we obtain the following values of log-likelihood, as a function of the number of states m, and the corresponding values of AIC and BIC:

m	log-likelihood	AIC	BIC
1	−777.417	−779.417	−783.927
2	−591.713	**−597.713**	**−611.243**
3	−598.808	−610.808	−637.869
4	−580.997	−600.997	−646.099

Considering both the AIC and the BIC as model selection criteria, we choose a two-states Markov chain. The sequence $\left\{\ln L_{672}\left(\phi^{(k+1)}\right)\right\}$ converges at the 20^{th} iteration to $\ln L_{672}\left(\phi^{(20)}\right) = -591.713$, starting from $\ln L_{672}\left(\phi^{(0)}\right) = -1172.107$. The estimates of the parameters (standard errors in brackets) of the two gaussian *pdf*s are

i	1	2
$\mu_i^{(20)}$	1.0676	2.5660
	(0.2001)	(0.0890)
$\sigma_i^{2(20)}$	0.2666	0.2760
	(0.0406)	(0.0514)

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

$$\Gamma^{(20)} = \begin{bmatrix} 0.9039 & 0.0961 \\ (0.0315) & (0.0315) \\ 0.0294 & 0.9706 \\ (0.1095) & (0.1095) \end{bmatrix}$$

from which we have the estimate of the stationary initial distribution

$$\delta^{(20)} = (0.2344; 0.7656)'.$$

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 expected time spent in state i, is:

i	1	2
hours	10.4102	34.0053

Conclusions and extensions

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. 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 $\delta' = \delta'\Gamma$. The use of explicit formulae of the maximum likelihood estimators of the parameters simplifies the optimization problem, because it allows us to solve the M step exactly, without using a numerical maximization algorithm, such as the Newton-Raphson method. Hence the procedure is more stable and converges faster in the neighborhood of the maximum. Furthermore an application of GHMMs to air pollution data have been shown and the estimates of the parameters of the model, togheter with their standard errors, 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).

We are currently studying how to reinforce the dependence among the observed variables by adding an autoregressive feature: we replace the conditional independence condition of HMMs with an order-p dependence condition. Hence the present observation depends also on the p past observations, the autoregressive coefficients depend on the current state of the Markov chain and, as in the HMMs case, every observation depends on the contemporary state of the chain. These *hidden Markov autoregressive models* are often used by econometricians (Hamilton, 1994, Chapter 22 and the references therein) and they are also said *Markov-switching autoregressive models*.

Appendix - Prooves of Propositions 1, 2, 3

Proof of Proposition 1 - The E step of the EM algorithm, at the $(k+1)^{th}$ iteration, is so defined:

$$Q(\phi;\phi^{(k)}) = \mathrm{E}_{\phi^{(k)}}\left\{\ln L_T^c(\phi) \mid y\right\} =$$
$$= \sum_{i_1} \dots \sum_{i_T} \left\{\left[\ln f\left(y_1, \dots, y_T, i_1, \dots, i_T\right)\right] \cdot \right. \tag{9}$$
$$\left. \cdot \mathrm{P}\left(X_1 = i_1, \dots, X_T = i_T \mid y_1, \dots, y_T; \phi^{(k)}\right)\right\}.$$

Consider only the log-likelihood of (9):

$$\ln f(y_1, \dots, y_T, i_1, \dots, i_T) =$$
$$= \ln[\delta_{i_1} f(y_1 \mid i_1) \prod_{t=2}^{T} \gamma_{i_{t-1},i_t} f(y_t \mid i_t)] = \tag{10}$$
$$= \ln \delta_{i_1} + \sum_{t=1}^{T-1} \ln \gamma_{i_t,i_{t+1}} + \sum_{t=1}^{T} \ln f(y_t \mid i_t).$$

Replacing the right-hand side of (10) in (9), we have

$$Q(\phi;\phi^{(k)}) = \sum_{i_1} \dots \sum_{i_T} \frac{f^{(k)}(y_1,\dots,y_T,i_1,\dots,i_T)}{f^{(k)}(y_1,\dots,y_T)} \left[\ln \delta_{i_1} + \sum_{t=1}^{T-1} \ln \gamma_{i_t,i_{t+1}} + \sum_{t=1}^{T} \ln f(y_t \mid i_t)\right] =$$

$$= \sum_{i_1} \dots \sum_{i_T} \frac{f^{(k)}(y_1,\dots,y_T,i_1,\dots,i_T)}{f^{(k)}(y_1,\dots,y_T)} \ln \delta_{i_1} +$$
$$+ \sum_{i_1} \dots \sum_{i_T} \frac{f^{(k)}(y_1,\dots,y_T,i_1,\dots,i_T)}{f^{(k)}(y_1,\dots,y_T)} \ln \gamma_{i_1,i_2} +$$
$$+ \sum_{i_1} \dots \sum_{i_T} \frac{f^{(k)}(y_1,\dots,y_T,i_1,\dots,i_T)}{f^{(k)}(y_1,\dots,y_T)} \ln \gamma_{i_2,i_3} + \dots +$$
$$+ \sum_{i_1} \dots \sum_{i_T} \frac{f^{(k)}(y_1,\dots,y_T,i_1,\dots,i_T)}{f^{(k)}(y_1,\dots,y_T)} \ln \gamma_{i_{T-1},i_T} +$$
$$+ \sum_{i_1} \dots \sum_{i_T} \frac{f^{(k)}(y_1,\dots,y_T,i_1,\dots,i_T)}{f^{(k)}(y_1,\dots,y_T)} \ln f(y_1 \mid i_1) +$$
$$+ \sum_{i_1} \dots \sum_{i_T} \frac{f^{(k)}(y_1,\dots,y_T,i_1,\dots,i_T)}{f^{(k)}(y_1,\dots,y_T)} \ln f(y_2 \mid i_2) + \dots +$$
$$+ \sum_{i_1} \dots \sum_{i_T} \frac{f^{(k)}(y_1,\dots,y_T,i_1,\dots,i_T)}{f^{(k)}(y_1,\dots,y_T)} \ln f(y_T \mid i_T).$$

Marginalizing with respect to the states of the Markov chain, we have

$$Q(\phi;\phi^{(k)}) = \sum_{i_1} \frac{f^{(k)}(y_1,\dots,y_T,i_1)}{f^{(k)}(y_1,\dots,y_T)} \ln \delta_{i_1} +$$
$$+ \sum_{i_1}\sum_{i_2} \frac{f^{(k)}(y_1,\dots,y_T,i_1,i_2)}{f^{(k)}(y_1,\dots,y_T)} \ln \gamma_{i_1,i_2} +$$
$$+ \sum_{i_2}\sum_{i_3} \frac{f^{(k)}(y_1,\dots,y_T,i_2,i_3)}{f^{(k)}(y_1,\dots,y_T)} \ln \gamma_{i_2,i_3} + \dots +$$
$$+ \sum_{i_{T-1}}\sum_{i_T} \frac{f^{(k)}(y_1,\dots,y_T,i_{T-1},i_T)}{f^{(k)}(y_1,\dots,y_T)} \ln \gamma_{i_{T-1},i_T} +$$
$$+ \sum_{i_1} \frac{f^{(k)}(y_1,\dots,y_T,i_1)}{f^{(k)}(y_1,\dots,y_T)} \ln f(y_1 \mid i_1) +$$
$$+ \sum_{i_2} \frac{f^{(k)}(y_1,\dots,y_T,i_2)}{f^{(k)}(y_1,\dots,y_T)} \ln f(y_2 \mid i_2) + \dots +$$

$$+\sum_{i_T}\frac{f^{(k)}(y_1,\ldots,y_T,i_T)}{f^{(k)}(y_1,\ldots,y_T)}\ln f(y_T\mid i_T)=$$

$$=\sum_{i_1}\frac{f^{(k)}(y_1,\ldots,y_T,i_1)}{f^{(k)}(y_1,\ldots,y_T)}\ln\delta_{i_1}+$$
$$+\sum_{t=1}^{T-1}\left[\sum_{i_t}\sum_{i_{t+1}}\frac{f^{(k)}(y_1,\ldots,y_T,i_t,i_{t+1})}{f^{(k)}(y_1,\ldots,y_T)}\ln\gamma_{i_t,i_{t+1}}\right]+$$
$$+\sum_{t=1}^{T}\left[\sum_{i_t}\frac{f^{(k)}(y_1,\ldots,y_T,i_t)}{f^{(k)}(y_1,\ldots,y_T)}\ln f(y_t\mid i_t)\right]=$$

$$=\sum_{i}\frac{f^{(k)}(y_1,\ldots,y_T,X_1=i)}{f^{(k)}(y_1,\ldots,y_T)}\ln\delta_i+$$
$$+\sum_i\sum_j\frac{\sum_{t=1}^{T-1}f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)}{f^{(k)}(y_1,\ldots,y_T)}\ln\gamma_{i,j}+$$
$$+\sum_i\frac{\sum_{t=1}^{T}f^{(k)}(y_1,\ldots,y_T,X_t=i)}{f^{(k)}(y_1,\ldots,y_T)}\ln f(y_t\mid i)\ \square$$

Proof of Proposition 2 - According to Basawa and Prakasa Rao (1980), to get the estimators of $\gamma_{i,j}$, we assume that for large T the initial distributions δ_i are non-informative about the transition probabilities $\gamma_{i,j}$: the function $Q(\phi;\phi^{(k)})$ (formula 5) must be differentiated with respect to the parameter $\gamma_{i,j}$, under the m constraints

$$\sum_{j\in S_X}\gamma_{i,j}=1,\ \text{for}\ 1\le i\le m, \tag{11}$$

setting the derivative equal to zero.

To solve this constrained maximization problem, we resort to Lagrange's multipliers method. Let $G^{(k)}$ be the Lagrangean function of $Q(\phi;\phi^{(k)})$ with respect to the constraints (11), being λ_i the Lagrange's multipliers:

$$G^{(k)}=Q(\phi;\phi^{(k)})-\lambda_1\left(\sum_{j\in S_X}\gamma_{1,j}-1\right)-\ldots-\lambda_m\left(\sum_{j\in S_X}\gamma_{m,j}-1\right)=$$

$$=Q(\phi;\phi^{(k)})-\sum_{i\in S_X}\lambda_i\left(\sum_{j\in S_X}\gamma_{i,j}-1\right).$$

Differentiating $G^{(k)}$ with respect to $\gamma_{i,j}$ and setting the derivative equal to zero, we have

$$\frac{\partial G^{(k)}}{\partial\gamma_{i,j}}=\frac{\partial}{\partial\gamma_{i,j}}\left[Q(\phi;\phi^{(k)})-\sum_{i\in S_X}\lambda_i\left(\sum_{j\in S_X}\gamma_{i,j}-1\right)\right]=$$

$$=\frac{\partial}{\partial\gamma_{i,j}}\left\{\sum_i\sum_j\left[\frac{\sum_{t=1}^{T-1}f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)}{f^{(k)}(y_1,\ldots,y_T)}\ln\gamma_{i,j}\right]\right\}-\lambda_i=$$

$$=\frac{\sum_{t=1}^{T-1}f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)}{f^{(k)}(y_1,\ldots,y_T)}\frac{1}{\gamma_{i,j}}-\lambda_i=0,$$

from which it follows

$$\frac{\sum_{t=1}^{T-1} f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)}{f^{(k)}(y_1,\ldots,y_T)} = \lambda_i \gamma_{i,j}. \tag{12}$$

Summing over j both sides of (12), we obtain

$$\sum_{j\in S_X} \frac{\sum_{t=1}^{T-1} f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)}{f^{(k)}(y_1,\ldots,y_T)} = \lambda_i \sum_{j\in S_X} \gamma_{i,j} = \lambda_i. \tag{13}$$

Replacing (13) in (12), we have

$$\begin{aligned}&\frac{\sum_{t=1}^{T-1} f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)}{f^{(k)}(y_1,\ldots,y_T)} = \\ &= \gamma_{i,j} \sum_{j\in S_X} \frac{\sum_{t=1}^{T-1} f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)}{f^{(k)}(y_1,\ldots,y_T)},\end{aligned}$$

from which it follows

$$\gamma_{i,j} = \frac{\sum_{t=1}^{T-1} f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)}{\sum_{j\in S_X}\sum_{t=1}^{T-1} f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)},$$

that, by definition, is the value $\gamma_{i,j}^{(k+1)}$.

Hence, marginalizing with respect to j, we have

$$\gamma_{i,j}^{(k+1)} = \frac{\sum_{t=1}^{T-1} f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)}{\sum_{t=1}^{T-1} f^{(k)}(y_1,\ldots,y_T,X_t=i)}. \tag{14}$$

Let $A_{i,j}^{(k)}$ be the vector whose entries are the *pdf*s $f^{(k)}(y_1,\ldots,y_T,X_t=i,X_{t+1}=j)$, for any $t=1,\ldots,T-1$, and $B_i^{(k)}$ be the vector whose entries are the *pdf*s $f^{(k)}(y_1,\ldots,y_T,X_t=i)$, for any $t=1,\ldots,T-1$. Using the matrix notation introduced in Subsections 2.2 and 2.3, the numerator and the denominator of the (14) may be written as

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

and

$$\sum_{t=1}^{T-1} f^{(k)}(y_1, \ldots, y_T, X_t = i) =$$

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

$$= 1'_{(T-1)} B_i^{(k)},$$

which ends the proof □

Proof of Proposition 3 - To obtain the expressions of the estimators $\mu_i^{(k+1)}, \sigma_i^{2(k+1)}$, the function $Q\left(\phi; \phi^{(k)}\right)$, as in (5), must be differentiated with respect to μ_i, σ_i^2 $(i = 1, \ldots, m)$, setting the derivatives equal to zero.

The derivative of $Q\left(\phi; \phi^{(k)}\right)$ with respect to μ_i is:

$$\frac{\partial}{\partial \mu_i} Q\left(\phi; \phi^{(k)}\right) = \frac{\partial}{\partial \mu_i} \left[\sum_i \left(\frac{\sum_{t=1}^{T} f^{(k)}(y_1, \ldots, y_T, X_t = i)}{f^{(k)}(y_1, \ldots, y_T)} \ln f(y_t \mid i) \right) \right] =$$

$$= \frac{\sum_{t=1}^{T} f^{(k)}(y_1, \ldots, y_T, X_t = i)}{f^{(k)}(y_1, \ldots, y_T)} \frac{y_t - \mu_i}{\sigma_i^2} = 0,$$

from which it follows

$$\sum_{t=1}^{T} f^{(k)}(y_1, \ldots, y_T, X_t = i) \mu_i = \sum_{t=1}^{T} f^{(k)}(y_1, \ldots, y_T, X_t = i) y_t$$

$$\mu_i = \frac{\sum_{t=1}^{T} f^{(k)}(y_1, \ldots, y_T, X_t = i) y_t}{\sum_{t=1}^{T} f^{(k)}(y_1, \ldots, y_T, X_t = i)}. \qquad (15)$$

Let $C_i^{(k)}$ be the vector whose entries are the *pdf*s $f^{(k)}(y_1, \ldots, y_T, X_t = i)$, for any $t = 1, \ldots, T$. Using the matrix notation introduced in Subsections 2.2, the numerator and the denominator of the (15) may be written as

$$\sum_{t=1}^{T} f^{(k)}(y_1,\ldots,y_T,X_t=i)y_t =$$

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

where the symbol $\odot$ denotes the Hadamard product, and

$$\sum_{t=1}^{T} f^{(k)}(y_1,\ldots,y_T,X_t=i) =$$

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

The derivative of $Q\left(\phi;\phi^{(k)}\right)$ with respect to σ_i^2 is:

$$\frac{\partial}{\partial\sigma_i^2}Q\left(\phi;\phi^{(k)}\right) = \frac{\partial}{\partial\sigma_i^2}\left[\sum_i\left(\frac{\sum_{t=1}^{T} f^{(k)}(y_1,\ldots,y_T,X_t=i)}{f^{(k)}(y_1,\ldots,y_T)}\ln f(y_t\mid i)\right)\right] =$$

$$= \frac{\sum_{t=1}^{T} f^{(k)}(y_1,\ldots,y_T,X_t=i)}{f^{(k)}(y_1,\ldots,y_T)}\frac{-\sigma_i^2+(y_t-\mu_i)^2}{2\sigma_i^4} = 0,$$

from which it follows

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

$$\sigma_i^2 = \frac{\sum_{t=1}^{T} f^{(k)}(y_1,\ldots,y_T,X_t=i)\,(y_t-\mu_i)^2}{\sum_{t=1}^{T} f^{(k)}(y_1,\ldots,y_T,X_t=i)}. \tag{16}$$

Using the matrix notation introduced in Subsections 2.2 and the vectors $C_i^{(k)}$ previously defined, the numerator of the (16) may be written as

$$\sum_{t=1}^{T} f^{(k)}(y_1,\ldots,y_T,X_t=i)\,(y_t-\mu_i)^2 =$$

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

$$\left. \odot \begin{bmatrix} (y_1-\mu_i)^2 \\ \vdots \\ (y_t-\mu_i)^2 \\ \vdots \\ (y_T-\mu_i)^2 \end{bmatrix} \right) = 1'_{(T)} \left(C_i^{(k)} \odot \left(y - \mu_i 1_{(m)} \right)^2 \right).$$

Hence

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

and

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

References

Basawa I. V. and Prakasa Rao B. L. S. (1980). *Statistical Inference for Stochastic Processes*. Academic Press, London.

Chatfield C. (1996). *The Analysis of Time Series: an introduction*, fifth edition. Chapman & Hall, London.

Dempster A. P., Laird N. M., 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.

Hamilton J. D. (1994). *Time Series Analysis*, Princeton University Press, Princeton.

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

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

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