KEYNOTE

Estimation with Missing Data

Graham C. Goodwin
Department of Electrical & Computer Engineering,
The University of Newcastle, NSW 2308 Australia

Arie Feuer
Department of Electrical Engineering
Technion–Israel Institute of Technology
Technion City, Haifa 32000, Israel

Abstract This paper reviews estimation problems with missing, or hidden data. We formulate this problem in the context of Markov models and consider two interrelated issues, namely, the estimation of a state given measured data and model parameters, and the estimation of model parmeters given the measured data alone. We also consider situations where the measured data is, itself, incomplete in some sense. We deal with various combinations of discrete and continuous states and observations.

1 INTRODUCTION

We are concerned here with the description, or modelling, of physical phenomena. It is typically the case, that technological constraints imply that not all aspects of the phenomena can be measured. We will use the notion of "state" or "complete data" to describe the underlying phenomena and the notion of "output" for the actual measured data.

We will focus on phenomena which evolve in time such that the current output depends on past history and events. Such phenomena are referred to as "dynamic systems". The current "state" is used to summarize the past history so that the subsequent behaviour depends only on the current state and subsequent events. A model describing this kind of behaviour is called a "Markov model". The notion of Markov model is a very powerful tool which, by appropriate choice of state, can represent a very wide class of dynamic phenomena. (Davis, 1993).

As an (admittedly oversimplified) example, consider a hypothetical model of the height (h_t) of a river at time t that depends on rainfall (r_t) as follows:

$$h_t = \sum_{k=1}^{\infty} \alpha^k r_{t-k} \; ; \; |\alpha| < 1 \tag{1}$$

i.e., the current height is an exponentially decaying sum of the past rainfall. A corresponding Markovian representation is

$$x_{t+1} = \alpha x_{t-1} + \alpha r_{t-1} \tag{2}$$

$$h_t = x_t \tag{3}$$

Note that, in this simple example, the scalar h_t qualifies as the state at time t. However, usually the state will include other, non measured, variables.

The concept of Markov models is very powerful and can be applied to a wide range of problems arising in a variety of disciplines including ecology, engineering, manufacturing, economics, etc. Indeed, the challenge is to find a problem that cannot be put into this framework.

In general, Markov models take different forms depending on the nature of the phenomena and of the measurement system. We can thus speak of

- deterministic or stochastic depending on whether the next state is precisely determined by the current state or has some probability distribution
- continuous or discrete time depending on whether the state evolution is described by a differential or difference equation
- continuous or discrete state (or measurement) depending on whether or not the state (measurement) has a continuum of values (e.g. as in the river height example) or a finite set of values (e.g. as is the case in binary communication channels).

Of course, there exist similarities between these different descriptions and some are clearly limiting cases of others (Middleton and Goodwin (1990)). Also, it is commonly the case, that the state will evolve in continuous time whereas the measurement system will be restricted to discrete times (sampled) and/or discrete measurements (quantization). This is a large topic in its own right - see for example Feuer and Goodwin (1995), Gevers and Li (1993), Williamson (1991).

For reasons of space constraints, we limit ourselves here to a subset of the above issues and consider only linear, stochastic, discrete time and, discrete and continuous state systems.

So far, we have used the term missing (or hidden) data to describe the fact that the available measurements do not capture the complete state of the system. However, there is occasionally a need to consider another form of missing data where the measurement pattern itself may be irregular (in time) due to a number of possible mechanisms including sensor failure, lost data records, human intervention, outliers etc. Isaksson (1993). We lump these issues also under the heading of missing data. The general tools that we describe can be applied to these more general problems.

The treatment given this paper, at times, sacrifices mathematical rigor for the sake of clarity of presentation. For a complete exposition the reader is referred to the references given at the end of the paper and recent books devoted to the topic of Hidden Markov Models e.g. Elliott, Aggsum, Moore (1995).

2 DISCRETE STATE HIDDEN MARKOV MODELS

In this section we will (briefly) describe discrete time - discrete state (hidden) Markov linear models.

We assume that the state x_t at time t, takes n possible values $s_1 \; ... \; s_n$. For convenience, we describe the states by a set of indicator functions, i.e. we write

$$s_i = e_i^n \; ; \; i = 1, \; ... \; , n \tag{4}$$

where e_i^n is the i^{th} column of an $n \times n$ identity matrix. Similarly, we assume that the output y_t at time t, takes m possible values, $o_1 \ldots o_m$ where

$$o_i = e_i^m \; ; \; i = 1, \ldots, m \tag{5}$$

Let

$$a_{ij} = \text{prob}\{x_{t+1} = e_i^n | x_t = e_j^n\} \tag{6}$$

Note that the Markov property is implicit in (6). Also note that we assume a_{ij} is independent of t, i.e. we restrict attention to stationary models.

Similarly, let

$$C_{ij} = \text{prob}\{y_t = e_i^m | x_t = e_j^n\} \tag{7}$$

Clearly, we have for all $j = 1, \ldots, n$

$$\sum_{i=1}^{n} a_{ij} = 1 \; ; \; \sum_{i=1}^{m} C_{ij} = 1 \tag{8}$$

We also assume that the initial state satisfies

$$p\{x_1 = s_i\} = \pi_i \tag{9}$$

Note that the above model is characterized by a finite number of parameters consisting of the entries a_{ij}, C_{ij}, π_i in $\{A, C, \pi\}$. We denote these parameters by θ.

We can then evaluate the following conditional expectation

$$E\{x_{t+1} | x_t \; ; \; \theta\} = Ax_t \tag{10}$$

$$E\{y_t | x_t \; ; \; \theta\} = Cx_t \tag{11}$$

Note that $E\{x_{t+i} | x_t; \theta\} \neq e_i^n$ for any i although $x_{t+1} = e_j^n$ for some j. Hence writing

$$v_t = x_{t+1} - Ax_t \tag{12}$$

$$w_t = y_t - Cx_t \tag{13}$$

we can express the model in the form:

$$x_{t+1} = Ax_t + v_t \tag{14}$$

$$y_t = Cx_t + w_t \tag{15}$$

where from (10), (11)

$$E\{v_t | x_t\} = 0 \; ; \; E\{w_t | x_t\} = 0 \tag{16}$$

3 ESTIMATION PROBLEMS

In the sequel we will be interested in two problems; namely

(i) State estimation: Say we are given a time series of data $y_1, \ldots y_T$, what can we say about the corresponding state sequence $x_1, \ldots x_T$, assuming we know the model (i.e. the parameters θ)?

(ii) Parameter estimation: Say we are given the data sequence $y_1, \ldots y_T$; what can we say about the model?

These questions are addressed below

4 STATE ESTIMATION FOR DISCRETE STATE-MARKOV MODELS

State estimation depends upon the criterion one employs. A common choice is to maximize the likelihood function. Say we are interested in estimating x_t given data $y_1 = e_{i_1}^m, \ldots, y_T = e_{i_T}^m$. Then two possible criteria are (Rabiner (1989)).

(i) The marginal likelihood; i.e.

$$\gamma_t(i) = \text{prob}\{x_t = e_i^n | y_1, \ldots, y_T\}; 1 \leq t \leq T \tag{17}$$

(ii) The joint likelihood; i.e.

$$\Gamma(i_1, \ldots, i_T) = \text{prob}\left\{x_1 = e_{i_1}^n, \ldots, x_T = e_{i_T}^n | y_1 \ldots y_m\right\} \tag{18}$$

In the gaussian case, these two criteria lead to the same estimate. However, this is, in general, not the case.

To generate the estimation for the first criteria, let us define

$$\alpha_t(i) = \text{prob}\{y_1, \ldots, y_t, x_t = e_i^n\}; 1 \leq t \leq T \tag{19}$$

This function can be generated recursively as follows:

$$\alpha_1(i) = \pi(i) C_{i_1 i}$$

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^{n} \alpha_t(j) a_{ij}\right] C_{i_{t+1} i}, \; 1 \leq t \leq T-1 \tag{20}$$

Note that

$$\text{prob}(y_1, \ldots, y_T) = \sum_{i=1}^{n} \alpha_T(i) \tag{21}$$

Similarly, we define

$$\beta_t(i) = \text{prob}\{y_{t+1}, \ldots, y_T | x_t = e_i^n\} \tag{22}$$

which can be generated recursively as follows:

$$\beta_T(i) = 1 \tag{23}$$

$$\beta_t(i) = \sum_{j=1}^{n} a_{ji} C_{i_{t+1} j} \beta_{t+1}(j); 1 \leq t \leq T-1 \tag{24}$$

Then, using (17), we have

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\text{prob}\{y_1 \ \ldots \ y_T\}}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{n} \alpha_t(i)\beta_t(i)} \tag{25}$$

Finally, the estimation of x_t which maximizes the marginal likelihood is

$$x_t^* = \arg \max_{1 \le i \le n} \{\gamma_t(i)\} \tag{26}$$

Optimization of the second criterion (18) can be done using Dynamic Programming and leads to the Viterbi algorithm (Forney (1973)).

Define

$$\delta_t(i) = \max_{x_1 \ldots x_{t-1}} \text{prob}\{x_1, \ \ldots \ x_{t-1}, x_t = e_i^n, y_1, \ \ldots \ y_t\} \tag{27}$$

Note that $\delta_t(i)$ satisfies the following recursion

$$\delta_1(i) = \pi_i C_{i,1} \ ; \ i = 1, \ \ldots \ , n \tag{28}$$

$$\delta_{t+1}(i) = \left[\text{Max}_j \delta_t(j) a_{ij}\right] C_{i,t+1} \tag{29}$$

In addition we need to calculate the optimal state at time t–1 given $x_t = e_j^n$, namely,

$$\hat{x}_{t-1}(e_j^n) = \arg \max_{1 \le j \le n} \{\delta_{t-1}(j) a_{ij}\} \tag{30}$$

Then, clearly, the optimal likelihood is given by

$$\Gamma^* = \frac{\max_{1 \le i \le n} [\delta_T(i)]}{\text{prob}(y_1 \ \ldots \ y_T)} \tag{31}$$

and the optimal terminal state is

$$x_T^* = \arg \max_{1 \le i \le n} \delta_T(i) \tag{32}$$

Finally, the optimal state trajectory can be computed recursively using a backward iteration as in (30)

$$x_{t-1}^* = \hat{x}_{t-1}(x_t^*) \tag{33}$$

5 PARAMETER ESTIMATION

Next, we turn to the problem of parameter estimation. In principle, this can be achieved by simply maximizing the likelihood function given in (21) with respect to θ. This is a formidable task. However, the likelihood function for the complete data ($x_1 \ \ldots \ x_T$ and $y_1 \ \ldots \ y_T$) is a relatively simple function of θ. This suggests an iteratively procedure in which, given an estimate $\theta^{(p)}$ of the parameters, we first find the expected value of the complete log likelihood function using $\theta^{(p)}$ and the data $y_1 \ \ldots \ y_T$. This is then maximized to find a new estimate $\theta^{(p+1)}$. This procedure is commonly referred to as the Expectation, Maximization (EM) algorithm. We take a brief diversion in the next section to describe this algorithm.

6 THE EM ALGORITHM

In our description of the EM procedure, we will cover general problems and thus we do not restrict ourselves to discrete data. Denote by $f(x, y|\theta)$ the density function of the complete data given the parameters θ and by $g(y|\theta)$ the density of the measurement y given θ. Generally, we would like to generate the maximum likelihood estimate of θ; i.e. the parameters that maximize $L(\theta)$ where

$$L(\theta) = \log g(y|\theta) \tag{34}$$

It is often true that any attempt to directly solve this problem is very difficult. Instead, as suggested at the end of the last section, we assume we have some estimate $\theta^{(p)}$ of θ, and evaluate

$$Q(\theta, \theta^{(p)}) = E\{ \log f(x, y|\theta) | y, \theta^{(p)}\} \tag{35}$$

as a function of θ.

We call this the E (expectation) step.

The M (maximization) step then chooses

$$\theta^{(p+1)} = \arg \max_{\theta} Q(\theta, \theta^{(p)}) \tag{36}$$

and hence

$$Q(\theta^{(p+1)}, \theta^{(p)}) \ge Q(\theta^{(p)}, \theta^{(p)}) \tag{37}$$

To appreciate this algorithm, we note that

$$f(x, y|\theta) = k(x|y, \theta) g(y|\theta) \tag{38}$$

where $k(x|y, \theta)$ is the conditional distribution of x given y.

Then clearly from (35), (38) we have

$$Q(\theta, \theta^{(p)}) = L(\theta) + E\{ \log k(x|y, \theta) | y, \theta^{(p)}\} \tag{39}$$

Using Jensen's inequality, it can readily be shown (see Lemma A.3 of Appendix A) that

$$E\{\log k(x|y, \theta) | y, \theta^{(p)}\} \le E\{\log k(x|y, \theta^{(p)}) | y, \theta^{(p)}\} \tag{40}$$

with equality if and only if

$$k(x|y, \theta^{(p)}) = k(x|y, \theta) \quad \text{a.e.} \tag{41}$$

Using (40) and (37) it is clear that

$$L(\theta^{(p+1)}) \ge L(\theta^{(p)}) \tag{42}$$

with equality if and only if equality holds in (37), (41).

This means that any generated sequence $\{L(\theta^{(p)})\}$ will converge in view of (42) and, under some regularity assumptions, this will

imply that $\{\theta^{(p)}\}$ will converge to same value θ^* which will be a local maximum of the likelihood function (Dempster et.al (1977), Boyles(1983), Wu (1983)).

7 PARAMETER ESTIMATION FOR DISCRETE STATE MARKOV MODELS

In order to apply the EM algorithm to the case at hand we calculate

$$Q(\theta, \theta^{(p)}) = E\Big\{\log \text{prob}\Big\{x_1 = e^n_{i_1}, \ \dots \ , x_T = e^n_{i_T}, \tag{43}$$

$$y_1 = e^m_{j_1}, \ \dots \ y_T = e^m_{i_T}\Big\}\Big|y_1 = e^m_{j_1}, \ \dots \ y_T = e^m_{j_T}; \ \theta^{(p)}\Big\}$$

To facilitate this, we first evaluate

$$\xi_t(i,j) = \text{prob}\{x_t = e^n_i, x_{t+1} = e^n_j, y_1, \ \dots \ , y_T|\theta\} \tag{44}$$

$$= \alpha_t(i) a_{ij} C_{i_{t+1}j} \beta_{t+1}(j); \ 1 \le t \le T-1$$

where we have used (19), (22).

It follows that (see also Rabiner (1983))

$$Q(\theta, \theta^{(p)}) = E\{\log[\pi_{i_1} a_{i_2 i_1} a_{i_3 i_2} \dots a_{i_T i_{T-1}} C_{j_1 i_1} \dots C_{j_T i_T}]|$$

$$y_1 = e^m_{j_1}, \ \dots \ , y_T = e^m_{j_T}; \theta^{(p)}\}$$

$$= E\{\log \pi_{i_1} + \sum_{t=1}^{T-1} \log a_{i_{t+1} i_t}$$

$$+ \sum_{t=1}^{T} \log C_{j_t i_t} | y_1 \ \dots \ y_T; \theta^{(p)}\}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{t=1}^{T-1} \xi^{(p)}(i,j) \log a_{ij}$$

$$+ \sum_{i=1}^{n} \gamma_1^{(p)}(i) \log \pi_i$$

$$+ \sum_{t=1}^{T} \sum_{i=1}^{n} \gamma_t^{(p)}(i) \log C_{j_t i} \tag{45}$$

Note that, in equation (46), the superscript (p) denotes that the term is evaluated using $\theta^{(p)}$

Maximizing (46) leads to the standard Baum-Welch (Baum et.al (1970)) estimates for θ, namely

$$a_{ij}^{(p+1)} = \frac{\sum_{t=1}^{T-1} \xi_t^{(p)}(i,j)}{\sum_{t=1}^{T-1} \sum_{j=1}^{n} \xi_t^{(p)}(i,j)} \tag{46}$$

$$\pi_i^{(p+1)} = \gamma_1^{(p)}(i) \tag{47}$$

$$C_{k,j}^{(p+1)} = \frac{\sum_{t=1}^{T} \gamma_t^{(p)}(j) \ \text{(such that } y_t = e^m_k)}{\sum_{t=1}^{T} \gamma_t^{(p)}(j)} \tag{48}$$

Equations (46) to (48) constitute the re-estimation (or M) step of the EM algorithm. As stated earlier, subject to reasonable conditions, this will converge to a (local) maximum of the likelihood function.

Further details on the application of the EM algorithm to Discrete State Hidden Markov Models may be found in Baum and Petrie (1966), Baum et.al (1970), Rabiner (1989).

8 CONTINUOUS STATE MARKOV MODELS

Next we consider the situation where the state can take a continuum of values. It is now convenient to let x_t take values in $\mathbb{R}^n$.

We define a continuous state (stationary) Markov model as follows:

$$x_{t+1} = Ax_t + v_t \ ; \tag{49}$$

$$y_t = Cx_t + w_t \tag{50}$$

where $x_t \in \mathbb{R}^n, y_t \in \mathbb{R}^m$ and $\{v_t\}$ and $\{w_t\}$ are mutually independent iid sequences. Note that A and C are assumed to be constant matrices of appropriate dimensions. The model given in (49), (50) is deceptively similar to the one in (14), (15). Note, however, that the interpretations are quite different. For example, in (49), (50) the matrices A and C do not have a probabilistic interpretation as in the discrete state case.

As before, we assume that $\{y_t\}$ constitutes the measured data and typically $n > m$.

9 STATE ESTIMATION FOR CONTINUOUS STATE MARKOV MODELS

Adding the assumptions that the distributions for $\{v_t\}$, $\{w_t\}$ and x_1 are independent gaussian with means $0, 0, \mu$ and covariances Q, R and Σ_1 respectively then the optimal state estimate problem becomes relatively easy because one can readily compute the conditional distributions. This leads to the celebrated Kalman Filter (Anderson and Moore (1979)).

Denote by $(\hat{x}_{t|k}, \Sigma_{t|k})$ the condition mean and covariances of x_t given data up to time k. The filtered estimates are then generated recursively by:

$$\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + \Sigma_{t+1|t} C^T (C\Sigma_{t+1|t} C^T + R)^{-1} \tag{51}$$

$$(y_{t+1} - C\hat{x}_{t+1|t})$$

$$\Sigma_{t+1|t+1} = \Sigma_{t+1|t} - \Sigma_{t+1|t} C^T (C\Sigma_{t+1|t} C^T + R)^{-1} C\Sigma_{t+1|t} \tag{52}$$

$$\hat{x}_{t+1|t} = A\hat{x}_{t|t} \tag{53}$$

$$\Sigma_{t+1|t} = A\Sigma_{t|t} A^T + Q \tag{54}$$

with $\Sigma_{1|0} = \Sigma_1, \hat{x}_{1|0} = \mu$.

Note that the above filter uses the complete model, i.e. we have used A, C, Q, R, $\Sigma_1\mu$. Although we have stated this result under a gaussian assumption, it turns out (Anderson and Moore (1979) that the filter is also the Best Linear Estimator under others distributions.

It is also possible to obtain "smoothed" estimates by evaluating the condition distribution of x_t given the whole data (y_1 ... y_T). Using a procedure, analogous to that which we found in the discrete data case, it is possible to use forward (i.e. filtered) recursions and backward recursions (see (33)) to evaluate (Anderson and Moore (1979), p.189), Shumway and Stoffer (1982)) the smoothed values:

$$\hat{x}_{t-1|T} = \hat{x}_{t-1|t-1} + J_{t-1}(\hat{x}_{t|T} - \hat{x}_{t|t-1}) \tag{55}$$

$$\Sigma_{t-1|T} = \Sigma_{t-1|t-1} + J_{t-1}(\Sigma_{t|M} - \Sigma_{t|t-1})J_{t-1}^T \tag{56}$$

where

$$J_{t-1} = \Sigma_{t-1|t-1}A^T\Sigma_{t|t-1}^{-1} \tag{57}$$

and with final boundary conditions, $\Sigma_{T|T}, \hat{x}_{T|T}$ given by (51) to (54).

Also, for future use we need the covariance, $\Sigma_{t,t-1|T}$, between x_t and x_{t-1} given y_1 ... y_T. This quantity satisfies the following backward recursion

$$\Sigma_{t-1,t-2|T} = \Sigma_{t-1|t-1}J_{t-2}^T + J_{t-1}(\Sigma_{t,t-1|T} - A\Sigma_{t-1|t-1})J_{t-2}^T \tag{58}$$

with

$$\Sigma_{T,T-1|T} = \left(I - \Sigma_{T|T-1}C^T(C\Sigma_{T|T-1}C^T + R)^{-1}C\right) A\Sigma_{T-1|T-1} \tag{59}$$

10 PARAMETER ESTIMATION FOR CONTINUOUS STATE MARKOV MODELS

The unknown parameters are the entries in A, C, Q, R, μ, Σ_1. Using system theory principles, and since we do not have the probabilistic interpretation of C as in the discrete state case, then C can always be taken as a known matrix.

Thus, the model depends upon the parameters in $\theta = (A, Q, R, \mu, \Sigma_0)$. There are many possible criteria one could employ, to estimate the parameters in the model (see Ljung (1987)). For reasons of space, we again restrict ourselves to maximum likelihood. As in the discrete state case, direct maximization of the likelihood function is, in most cases, virtually very difficult. Hence, we briefly describe the application of the EM algorithm in the context of the model as given in Section 8 and 9.

Again based on a Gaussian assumption, the log-likelihood for the complete data is

$$\begin{aligned}\log[\text{prob}\{x_1, \ldots, x_T, y_1 \ldots y_T|\theta\}] &= -\frac{1}{2}\log|\Sigma_1| \\ &\quad -\frac{1}{2}(x_1 - \mu)^T\Sigma_1^{-1}(x_1 - \mu) \\ &\quad -\frac{T}{2}\log|Q| - \frac{1}{2}\sum_{t=2}^{T}(x_t - Ax_{t-1})^TQ^{-1}(x_t - Ax_{t-1}) \\ &\quad -\frac{T}{2}\log|R| - \frac{1}{2}\sum_{t=1}^{T}(y_t - Cx_t)^TR^{-1}(y_t - Cx_t)\end{aligned} \tag{60}$$

The E step consists of finding the expectation of the expression in (60) leading to (see also Shumway and Stoffer (1982), Shumway (1984))

$$\begin{aligned}Q(\theta, \theta^{(p)}) &= -\frac{1}{2}\log|\Sigma_1| \\ &\quad -\frac{1}{2}\text{trace}\left\{\Sigma_1^{-1}\left(\Sigma_{1|T}^{(p)} + \left(\hat{x}_{1|T}^{(p)} - \mu\right)\left(\hat{x}_{1|T}^{(p)} - \mu\right)^T\right)\right\} \\ &\quad -\frac{T}{2}\log|Q| \\ &\quad -\frac{1}{2}\text{trace}\{Q^{-1}\left(S_{11}^{(p)}\right) - A\left(S_{10}^{(p)}\right)^T - S_{10}^{(p)}A^T \\ &\qquad + AS_{00}^{(p)}A^T\} \\ &\quad -\frac{T}{2}\log|R| \\ &\quad -\frac{1}{2}\text{trace}\{R^{-1}\sum_{t=1}^{T}[\left(y_t - C\hat{x}_{t|T}^{(p)}\right)\left(y_t - C\hat{x}_{t|T}^{(p)}\right)^T \\ &\qquad + A\Sigma_{t|T}^{(p)}A^T]\}\end{aligned} \tag{61}$$

where

$$S_{00}^{(p)} = \sum_{t=2}^{T}\left(\Sigma_{t-1|T}^{(p)} + \hat{x}_{t-1|T}^{(p)}\hat{x}_{t-1|T}^{(p)T}\right) \tag{62}$$

$$S_{10}^{(p)} = \sum_{t=2}^{T}\left(\Sigma_{t,t-1|T}^{(p)} + \hat{x}_{t|T}^{(p)}\hat{x}_{t-1|T}^{(p)T}\right) \tag{63}$$

$$S_{11}^{(p)} = \sum_{t=1}^{T}\left(\Sigma_{t|T}^{(p)} + \hat{x}_{t|T}^{(p)}\hat{x}_{t|T}^{(p)T}\right) \tag{64}$$

and the superscript (p) denotes filtered and smoothed estimates generated by equations (51) to (57) using $\theta^{(p)}$

It is quite straightforward to maximize the expression given in (61) with respect to θ leading to:–

$$A^{(p+1)} = \left(\sum_{t=2}^{T}\Sigma_{t,t-1|T}^{(p)} + \hat{x}_{t|T}^{(p)}x_{t-1|T}^{(p)T}\right)\left(\sum_{t=2}^{T}\Sigma_{t-1|T}^{(p)} + \hat{x}_{t-1|T}^{(p)}\hat{x}_{t-1|T}^{(p)T}\right)^{-1} \tag{65}$$

$$Q^{(p+1)} = \frac{1}{T} \cdot \left\{ \sum_{t=2}^{T} \left(\Sigma_{t|T}^{(p)} - A^{(p+1)} \left(\Sigma_{t,t-1|T}^{(p)} + \hat{X}_{t|T}^{(p)} \hat{X}_{t-1|T}^{(p)T} \right)^T \right) \right\} \quad (66)$$

$$R^{(p+1)} = \frac{1}{T} \{ \sum_{t=1}^{T} [\left(y_t - C\hat{X}_{t|T}^{(p)} \right) \left(y_t - C\hat{X}_{t|T}^{(p)} \right)^T + C\Sigma_{t|T}^{(p)} C^T] \} \quad (67)$$

$$\mu^{(p+1)} = \hat{X}_{1|T}^{(p)} \quad (68)$$

$$\Sigma_1^{(p+1)} = \Sigma_1^{(p)} \quad (69)$$

Actually, the above expressions also apply when C is time-varying. Thus missing values of y_t can be treated by simply setting the corresponding value C_t to zero. (see Shumway and Stoffer (1982)).

For further details of the application of the EM algorithm to continuous state models see Kuczera (1987), Abraham and Chuang (1993), Tanaka and Katayama (1990), Ansley and Khan (1983), Harvey and McKenzie (1984), Jones(1980), Khon and Ansey (1986), Little and Rubin (1987), McGiffin and Murthy (1980), (1982), Miller and Ferreiro (1984), Rosen and Porat (1986), Shumway and Stoffer (1982), Shumway (1984), Isaksson (1993), Feder and Weinstein (1985).

11 INNOVATIONS MODELS

In the stationary data case, the signal $\{y(t)\}$ that was generated by the general Markov model (49), (50) has an alternative form which, whilst preserving the generality of the description, depends upon a significantly fewer number of unknown parameters. This, and the resulting structure, will turn out to be helpful in simplifying the parameter estimation problem. We will generate this alternative representation using the Kalman filter. Thus, consider equations (51) to (54).

It can be readily seen that the "innovations sequence", $\eta_t = \left(y_t - C\hat{x}_{t|t-1} \right)$ is orthogonal to the data $y_1 \ldots y_{t-1}$, i.e. it is a Martingale difference sequence. It is also known (Goodwin and Sin (1983)) that subject to mild regulatory conditions, the covariance $\Sigma_{t+1|t}$ converges to a constant matrix $\bar{\Sigma}$ for large t. Hence, putting $K = A\bar{\Sigma}C^T\left(C\bar{\Sigma}C^T + R\right)^{-1}$, using z_t to denote $\hat{x}_{t|t-1}$ and assuming steady state, (51) to (54) can be rewritten as

$$z_{t+1} = Az_t + K\eta_t \quad (70)$$

$$y_t = Cz_t + \eta_t \quad (71)$$

This can be viewed as an alternative Markov model for the system. For simplicity we now restrict attention to the single output case. However, similar constructions apply to the multivariable case (Kailath (1980)). Without loss of generality we may take (C, A) to be in observer canonical form (Goodwin and Sin (1983)), i.e. we may take

$$A = \begin{bmatrix} -a_1 & 1 & 0 \\ \vdots & & \ddots \\ -a_n & 0 \cdots 0 & 1 \end{bmatrix}; \quad K = \begin{bmatrix} k_1 \\ \vdots \\ k_n \end{bmatrix}, \quad (72)$$

$$C = [1\ 0\ \ldots\ 0] \quad (73)$$

Using (72), (73) it is quite straightforward to successively eliminate $\{z_t\}$ from (70), (71) leading to the following simpler ARMA model:

$$\mathcal{A}(q^{-1})y_t = \mathcal{C}(q^{-1})\eta_t\ ;\ E\{\eta_t^2\} = \sigma^2 \quad (74)$$

where

$$\mathcal{A}(q^{-1}) = 1 + a_1 q^{-1} + \ldots + a_n q^{-n} \quad (75)$$

$$\mathcal{C}(q^{-1}) = 1 + c_1 q^{-1} + \ldots + c_n q^{-n} \quad (76)$$

and where q^{-1} is the unit delay operator and

$$c_i = a_i + k_i\ ;\ i = 1, \ldots, n \quad (77)$$

A useful trick is now to replace the original system parameterization by the above model which is directly parameterized using the Kalman gain rather than indirectly via (51) to (54).

The unknown parameters θ in this case will consist of $\theta_1 = (a_1, \ldots, a_n, c_1, \ldots, c_n)^T, \sigma^2$ and the initial state $\varphi_0 = [-y_{-1}, \ldots, -y_{-n}, \eta_{-1}, \ldots, \eta_{-n}]^T$. For simplicity, in the sequel, we assume φ_0 is known and fixed.

With Gaussian noise, the log-likelihood function then turns out to be

$$L(\theta) = \text{constant} - \frac{1}{2\sigma^2} \sum_{t=1}^{T} \left(\frac{A(q^{-1})}{C(q^{-1})} y_t \right)^2 \quad (78)$$

It is possible to maximize (78) directly in terms of θ (see Ljung (1983)). However, we see from (78) that $L(\theta)$ is a non-quadratic function of θ. This leads us to ask if it might be possible to define the complete data sequence in such a way that the M step in the EM algorithm is straightforward.

We note that (74) can be expanded as

$$y_t = \varphi_{t-1}^T \theta_1 + \eta_t \quad (79)$$

where

$$\varphi_{t-1}^T = [-y_{t-1}, -y_{t-2} \ldots, -y_{t-n}, \eta_{t-1}, \ldots, \eta_{t-n}] \quad (80)$$

We now define the complete data log likelihood as

$$\log[\text{prob}\{y_1, \ldots, y_T, \eta_1, \ldots, \eta_T\}] \quad (81)$$

$$= \text{constant} + \log[\text{prob}(y_1, \eta_1)\text{prob}(y_2, \eta_2 | y_1, \eta_1) \ldots \text{prob}(y_T, \eta_T | y_{T-1}, \ldots y_1, \eta_{T-1}, \ldots \eta_1)]$$

$$= \text{constant} - \frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - \varphi_{t-1}^T \theta_1)^2 - \frac{T}{2} \log \sigma^2 \quad (82)$$

The E-step in the EM algorithm then gives

$$Q(\theta, \theta^{(p)}) = \text{constant} - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_t - \left(\varphi_{t-1}^{(p)}\right)^T\theta_1\right)^2 - \frac{T}{2}\log\sigma^2 \tag{83}$$

where $\{\varphi_{t-1}^{(p)}\}$ is given by

$$\left(\varphi_{t-1}^{(p)}\right)^T = \left[-y_{t-1}, \ \ldots \ -y_{t-n}, \eta_{t-1}^{(p)}, \ \ldots \ , \eta_{t-n}^{(p)}\right] \tag{84}$$

and where $\{\eta_t^{(p)}\}$ is generated recursively by

$$\eta_t^{(p)} = y_t - \left(\varphi_{t-1}^{(p)}\right)^T\theta_1^{(p)} \tag{85}$$

The expression in (83) is quadratic in θ_1 and simple for σ^2. Hence, we readily get from the M step

$$\theta_1^{(p+1)} = \left[\sum_{t=1}^{T}\varphi_{t-1}^{(p)}\left(\varphi_{t-1}^{(p)}\right)^T\right]^{-1}\left[\sum_{t=1}^{T}\varphi_{t-1}^{(p)}y_t\right] \tag{86}$$

$$\left(\sigma^2\right)^{(p+1)} = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \left(\varphi_{t-1}^{(p)}\right)^T\theta_1^{(p+1)}\right)^2 \tag{87}$$

The remarkably simple result in (83) holds because $\eta_t^{(p)}$ is completely determined by $\{\eta_k^{(p)}, y_k \ ; \ k = 1, \ \ldots \ t-1\}$. This leads to a particularly simple E step.

12 ESTIMATION USING MEASUREMENTS HAVING RANDOMLY MISSING VALUES

In the case of missing measurements, we could proceed by augmenting the E-step so as to "fill-in" the missing measurements. In this case, we would need to employ fixed interval smoothing techniques to complete the E-step. We note that $\{\varphi_t\}$ satisfies the following state space model

$$\varphi_t = F\varphi_{t-1} + G\eta_t \tag{88}$$

$$F = \left[\begin{array}{c|c} -a_1 \ \ldots \ -a_n & c_1 \ \ldots \ c_n \\ \hline \begin{array}{ccc} 1 & & \\ & \ddots & \\ 0 & & 1\cdot 0 \end{array} & \begin{array}{c} \\ 0 \\ \\ \end{array} \\ \hline \begin{array}{ccc} & 0 & \\ & & \end{array} & \begin{array}{ccc} 0 & \ldots & 0 \\ 1 & \ddots & \\ & 0 & \ddots 1\cdot 0 \end{array} \end{array}\right] \ ; \ G = \left[\begin{array}{c} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{array}\right] \tag{89}$$

and hence, the complete data $\{y_t\}$ and the measured data $\{y_t^m\}$ satisfy

$$y_t = e_1^T\varphi_t \tag{90}$$

$$y_t^m = f_t y_t = f_t\theta_1^T\varphi_{t-1} + f_t\eta_t \tag{91}$$

where $\{f_t\}$ is such that $f_t = 1$ if the data is available and is 0 otherwise. We can then apply the fixed interval smoothing techniques of Section 9 to evaluate the appropriate conditional distributions needed to compute the expected value of the complete data likelihood. One additional feature is that in (89), (91), we see that the process noise and measurement noise are correlated whereas this was not the case in Section 7. However, the correlation can be removed by substituting (91) into (89) leading to

$$\varphi_t = F\varphi_{t-1} + G(1 - f_t)\eta_t + y_t^m - f_t\theta_1^T\varphi_{t-1} \tag{92}$$

$$y_t^m = f_t\theta_1^T\varphi_{t-1} + f_t\eta_t \tag{93}$$

Then, in the E-step we obtain the following result in lieu of (83):

$$Q(\theta, \theta^{(p)}) = \text{constant} - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(e_1^T\hat{\varphi}_{t|T}^{(p)} - \left(\hat{\varphi}_{t-1|T}^{(p)}\right)^T\theta_1\right)^2 + \frac{1}{\sigma^2}\sum_{t=1}^{T}e_1^T\Sigma_{t,t-1|T}^{(p)}\theta_1 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\theta^T\Sigma_{t-1|T}^{(p)}\theta_1 \tag{94}$$

We see that $Q(\theta, \theta^{(p)})$ is also quadratic in θ_1 and hence the M step is again straightforward.

Applications of this algorithm can be found in many different areas - see for example Kuczera (1987) and Isaksson (1993).

13 ESTIMATION USING MEASUREMENTS HAVING PERIODICALLY MISSING VALUES

In the above treatment we have suggested that the EM algorithm can be employed to deal with arbitrary patterns of missing data. However, the generality of this approach can lead to difficulties in certain cases. For example, the convergence rate of the algorithm can be quite slow in the final interations. There is thus motivation to examine special missing data patterns to see if simpler parameter estimation strategies can be developed.

One such simplification occurs when the missing data has a regular time pattern. This can occur, for example, when there exists a cyclic (or periodic) interruption to the flow of data. In this case, other tools are available to deal with the missing data problem. We briefly outline one such possibility below.

Consider again (49), (50), and assume that the available data has a periodic time pattern. In this case, the matrix C will vary periodically with time between a constant value and zero. An interesting fact (Feuer and Goodwin (1995)) is that, in this case, the Kalman filter given in (51) to (54) settles to a form in which the covariance matrix $\Sigma_{t+1|t}$ is a periodic sequence. Hence, the innovations model (70), (71) remains valid save that in this case K, C and σ^2 vary periodically with time. This suggests that an alternative, finite parameterization of the model can be achieved. Each point of one period in the above sequences becomes a parameter to be estimated. This leads to a clear choice on the part of the modeller; namely, either work with the original (low dimensional) parameter vector assumed for the model with no measurements missing, or work with the resultant periodic parameters describing the innovation model. The latter will have more parameters to estimate but will lead to simpler estimation algorithms.

Using $\{\hat{y}_t\}$ to denote the predicted output at time t, then by a similar argument leading to (74), it can be shown that $\{\hat{y}_t\}$ satisfies

$$\mathcal{A}(q^{-1})\hat{y}_t = \mathcal{K}_t(q^{-1})(y_t - \hat{y}_t) \tag{95}$$

where $\mathcal{A}(q^{-1})$ is as in (75) and $\mathcal{K}_t(q^{-1})$ is

$$\mathcal{K}_t(q^{-1}) = k_{1t}q^{-1} + \ldots + k_{nt}q^{-n} \tag{96}$$

where $\{k_{1t}, \ldots, k_{nt}\}$ vary periodically. Also, note that k_{it} turns out to be zero whenever the corresponding data point is missing. Hence, if the period of the data pattern is L and, in each period, n_d values are missing, then direct parameterization of (95) leads to $n + n(L - n_d)$ unknown constant parameters.

Equation (95) will be recognized as a (special case) of a periodic ARMA (PARMA) model. Once this fact has been recognized, there are a plethora of parameter estimation methods that can be used - see for example Jones and Brelsford (1967), Gardner and Franks (1975), Pagano (1978), Tiao and Grupe (1980), Pankratz (1983), Vecchia (1985), Anderson and Vecchia (1993). It is also possible to develop recursive algorithms that process the data sequentially as it arrives - see for example Adams and Goodwin (1995). Of course it is also possible to use the EM algorithm yet again but this will be simpler in this case due to the reparameterization in terms of the (periodically varying) innovations model applicable to the (periodic) missing data case.

14 CONCLUSIONS

We have briefly outlined methods for state and parameter estimation in Hidden Markov models. We have only touched upon discrete and continuous state problems in discrete time. However, similar approaches hold for continuous time equivalents and mixed type systems, e.g. mixed continuous and discrete state and or mixed continuous and discrete time.

REFERENCES

Abraham, B and A. Chuang, 'Estimation of time series models in presence of outliers,' *Journal of Time Series Analysis*, Vol.14, No.3, pp.221-234, 1993.

Adams, G. and G.C. Goodwin, 'Parameter estimation for periodic ARMA models,' *Journal of Time Series Analysis*, Vol.16, No.2, pp.127-145, March 1995.

Anderson, B.D.O. and J.B. Moore, *Optimal Filtering*, Englewood Cliffs, NJ, Prentice-Hall, 1979.

Anderson, P.L. and A.V. Vecchia, 'Asymptotic results for periodic autoregressive moving-average processes,' *J. Time Ser. Anal*, Vol.1, pp.1-18, 1993.

Ansley, C.F. and R. Kohn, 'Exact likelihood of vector autoregressive-moving average process with missing or aggregated data,' *Biometrika*, Vol.70, No.1, pp.275-278, 1983.

Baum, L.E. and T. Petrie, 'Statistical inference for probabilistic functions of finite state Markov chains,' *Ann Math. Stat.*, Vol.37, pp.1554-1563, 1966.

Baum, L.E., T. Petrie, G. Soules and N. Weiss, 'A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,' *Ann. Math. Stat.*, Vol.41, No.1, pp.164-171, 1970.

Boyles, R.A., 'On the convergence of the EM algorithm,' *J. Roy. Stat. Soc. B.*, Vol.45, No.1, pp.47-50, 1983

Davis, M.H.A., '*Markov models and optimization*,' Chapman and Hall, London, 1993.

Dempster, A.P., N.M. Laird and D.B. Rubin, 'Maximum likelihood from incomplete data via the EM algorithm,' *J. Roy. Stat. Soc. B*, Vol.39, No.1, pp.1-38, 1977.

Elliott, R.J., L. Aggoun and J.B. Moore, '*Hidden Markov model : estimation and control*,' Springer-Verlag 1995.

Feder, M. and E. Weinstein, 'Optimal multiple source location estimation via the EM algorithm,' Proc. of the 1985 Int. Conf. on Acoust., Speech & Signal Processing (ICASSP'85), Vol.4, pp.1762-1765, 1985.

Feuer, A. and G.C. Goodwin, '*Sampling in digital signal processing and control*,' CRC Press, 1995.

Forney, D., 'The Viterbi algorithm,' *Proc. IEEE.*, Vol.61, pp.258-278, 1973.

Gardner, W.A. and L.E. Franks, 'Characterization of cyclostationary random signal processes,' *IEEE Trans. Inf. Theory*, Vol.21, pp.4-14, 1975.

Gevers, M. and G. Li, '*Parameterizations in control, estimation and filtering problems*,' Springer Verlag, Berlin, 1993.

Goodwin, G.C. and K.S. Sin, '*Adaptive filtering predictions and control*,' Prentice Hall, Englewood Cliffs, 1983.

Harvey, A.C. and C.R. McKenzie, 'Missing observations in dynamic econometric models: A partial synthesis,' in E. Parzen, Ed., *Time Series Analysis of Irregularly Observed Data, pp.108-133, College Station, TX, 1984.*

Isaksson, A., 'Identification of ARX models subject to missing data,,' *IEEE Trans. Auto. Control*, Vol.38, No.5, pp.813-819, 1993.

Jones, R.H., 'Maximum likelihood fitting of ARMA models to times series with missing observations,' *Technometrics*, Vol.22, pp.389-395, August 1980.

Jones, R.H. and W.M. Brelsford, 'Time series with periodic structure,' *Biometrika*, Vol.54, pp.403-407, 1967.

Kailath, T., '*Linear sytems*,' Prentice-Hall, Englewood Cliffs, N.J. 1980.

Kohn, R. and C.F. Ansley, 'Estimation, prediction, and interpolation for ARMA models with missing data,' *J. Amer. Stat. Assoc.*, Vol.81, pp.751-761, September 1986.

Kuczera, G., 'On maximum likelihood estimators for the multisite log-one streamflow model complete and incomplete data cases,' *Water Resources Research*, Vol.23, No.4, pp.641-645, April 1987.

Levinson, S.E., L.R. Rabiner and M.M. Sondhi, 'An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition,' *The Bell System Technical Journal*, Vol.62, No.4, pp.1035-1074, 1983.

Little, R.J.A. and D.B. Rubin, *Statistical Analysis with Missing Data*, New York: Wiley, 1987.

Ljung, L., '*System identification: Theory for the user*,' Prentice Hall, Englewood Cliffs, 1983.

McGiffin, P.B. and D.N. Murthy, 'Parameter estimation for auto-regressive systems with missing observations,' *Int. J. Syst. Sci.*, Vol.11, No.9, pp.1021-1034, 1980.

Middleton, R.H. and G.C. Goodwin, '*Digital control and estimation; A unified approach*,' Englewood Cliffs, New Jersey, 1990.

Miller, R.B. and O. Ferreiro, 'A strategy to complete a time series with missing observations,' in E. Parzen, Ed., *Time Series Analysis of Irregularly Observed Data*, pp.251-275, College Station, TX, 1984.

Pagano, M., 'On periodic and multiple autoregressions.' *Ann. Statist.*, Vol.4, pp.396–399, 1978.

Pankratz, A., 'Forecasting with univariate box–Jenkins models: Concepts and cases, New York: Wiley, 1978.

Rabiner, L.R., 'A tutorial on hidden Markov models and selected applications in speech recognition,' *Proc. IEEE*, Vol.77, No.2, pp.257-286, February, 1989.

Rao, C.R., *'Linear statistical inference and its application,'* Wiley, New York, 1965.

Rosen, Y. and B. Porat, 'ARMA parameter estimation based on sample covariances for missing data,' *Proc. IEEE ICASSP*, Tokyo, Japan, pp.5.11.1-5.11.4, 1986.

Shumway, R.H., 'Some applications of the EM algorithm to analyzing incomplete time series data,' In Brillinger et.al. *'Time series analysis of irregularly observed data,'* Springer Verlag pp.290-324, 1984.

Shumway, R.H. and D.S. Stoffer, 'An approach to time series smoothing and forecasting using the EM algorithm,' *J. Time Series Analysis*, Vol.3, No.4, pp.253-264, 1982.

Tanaka, M. and T. Katayama, 'Robust identification and smoothing for linear system with outliers and missing data,' in *Preprints 11th IFAC World Congress*, pp.160-165, Taliin, Estonia, 1990.

Tiao, G.C. and M.R. Grupe, 'Hidden periodic autoregressive moving average models in time series data.' *Biometrika*, Vol.67, pp.365-373, 1980.

Vecchia, A.V., 'Maximum likelihood estimation for periodic autoregressive moving average models.' *Technometrics*, Vol.27, pp.375-384, 1985.

Williamson, D., *'Digital control and implementation - Finite word length considerations,'* Prentice Hall Int., London, 1991.

Wu, C.F.J., 'On the convergence properties of the EM algorithm,' *The Ann. Stats.*, Vol.11, No.1, pp.95-103, 1983.

APPENDIX A

Lemma A.1 – Jensens Inequality: If x is a random variable such that $E(x) = \mu$ and $f(x)$ is a convex function, then

$$E\{f(x)\} \geq f(E\{x\}) \tag{97}$$

with equality if and only if x is a degenerate distribution at μ.

Proof: See Rao (1965)

Lemma A.2: Let f and g be non-negative and integrable functions with respect to a measure μ and S be the region in which $f > 0$. If $\int_S (f - g)d\mu \geq 0$, then

$$\int_S f \log \frac{f}{g} d\mu \geq 0 \tag{98}$$

with equality only when $f = g(a.e.\mu)$

Proof: Let us choose the convex function $-\log(\cdot)$. Then, by (97) with $\frac{f}{\int_S f d\mu}$ as a distribution of $\left(\frac{g}{f}\right)$ we have

$$-\frac{\int_S f \log\left(\frac{g}{f}\right)}{\int_S f d\mu} \geq -\log\left[\frac{\int_S f \frac{g}{f} d\mu}{\int_S f d\mu}\right] = -\log \frac{\int_S g d\mu}{\int_S f d\mu} \tag{99}$$

which implies, since $\int_S f d\mu \geq \int_S g d\mu$,

$$\int_S f \log\left(\frac{f}{g}\right) d\mu \geq \int_S f d\mu \log \frac{\int_S f d\mu}{\int_S g d\mu} \geq 0$$

with equality when $\int_S (f - g)d\mu = 0$ ❑

Lemma A.3:

$$E\left\{\log[k(x|y, \theta)|y, \theta^{(p)}]\right\} \leq E\left\{\log k(x|y, \theta^{(p)})|y, \theta^{(p)}\right\} \tag{100}$$

Proof: Let $f(x) = k(x|y, \theta^{(p)})$

$$g(x) = k(x|y, \theta)$$

then from Lemma A.2

$$\int k(x|y, \theta^{(p)}) \log\left[\frac{k(x|y, \theta^{(p)})}{k(x|y, \theta)}\right] dx \geq 0$$

or

$$\int k(x|y, \theta^{(p)}) \log[k(x|y, \theta^{(p)})] dx - \int k(x|y, \theta^{(p)}) \log[k(x|y, \theta)] dx \geq 0$$

From which the result follows immediately

❑