MAXIMUM LIKELIHOOD IDENTIFICATION OF BILINEAR SYSTEMS

Stuart Gibson * **Brett M. Ninness** *,[1]

* *Department of Electrical and Computer Engineering, University of Newcastle, Australia*

Abstract: This paper considers the problem of estimating the parameters of a bilinear system from input-output measurements. A novel approach to this problem is proposed, one based upon the so-called Expectation Maximisation algorithm, wherein maximum likelihood estimates are generated iteratively without the need for a gradient-based search algorithm. This simple method is shown to perform well in simulation and it allows a multivariable bilinear system to be estimated directly in state-space form without the need for explicit parameterisation.

Keywords: Maximum Likelihood estimation, Bilinear systems, EM algorithm

1. INTRODUCTION

Due to their mathematical tractability the greater part of research in the field of system identification for control has dealt with the problem of identifying linear systems from input-output data. The use of linear models for the control of nonlinear systems has been successful in some cases in which the system of interest appears linear within some limited operating region. However, in many more scenarios it is necessary to identify a nonlinear system description.

Bilinear systems are a class of nonlinear systems in which the inputs and the states are multiplicatively coupled. They have been found to be good approximators of may types of nonlinear systems occurring especially in fields such process control.

This paper considers the problem of identifying time-invariant, MIMO bilinear systems modelled in state-space form as follows

$$x_{k+1} = Ax_k + N\left[u_k \otimes x_k\right] + Bu_k + w_k, \quad (1)$$

$$y_k = Cx_k + Du_k + v_k. \quad (2)$$

Here, $x_k \in \mathbf{R}^{n\times 1}$ is the system state vector, $u_k \in \mathbf{R}^{m\times 1}$ is the observed input and $y_k \in \mathbf{R}^{\ell\times 1}$ the observed output, while the matrices $A \in \mathbf{R}^{n\times n}$, $N \in \mathbf{R}^{n\times mn}$, $B \in \mathbf{R}^{n\times m}$, $C \in \mathbf{R}^{\ell\times n}$, and $D \in \mathbf{R}^{\ell\times \ell}$ parameterise the system dynamics. The Kronecker tensor product represented by the symbol $\otimes$ is defined in the following way. If A is an $m \times n$ -dimensional matrix such that $[A]_{k,\ell} = a_{k\ell}$ and B is an $\ell \times p$ -dimensional matrix then $A \otimes B$ is the $n\ell \times mp$ -dimensional matrix given by

$$A \otimes B \triangleq \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix}.$$

The noise corruption signals $w_k \in \mathbf{R}^{n\times 1}$ and $v_k \in \mathbf{R}^{\ell\times 1}$ are zero mean, i.i.d. random variables with a normal distribution, satisfying

$$\mathbf{E}\left\{ \begin{bmatrix} w_k \\ v_k \end{bmatrix} \begin{bmatrix} w_\ell \\ v_\ell \end{bmatrix}^T \right\} = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \delta(k-\ell). \quad (3)$$

This paper addresses the identification problem of determining the matrices A, B, C, D, N, Q and R from a finite set of input and output data.

In order to simplify the ensuing derivation somewhat, it will be assumed throughout (without loss of generality) that $S = 0$. The relaxation of this condition, while symbolically complicated, is algorithmically straightforward.

[1] This work was supported by the Australian Research Council and the Centre for Integrated Dynamics and Control.

Long-term interest in the identification of bilinear systems has ensured that there have been many contributions to this field. The approaches to the problem, too, have varied widely.

One stream has sought to approximate the bilinear system with a set of basis functions, for example Walsh functions (Karanam *et al.*, 1978) or Volterra series (Baheta *et al.*, 1979). In the latter case the approximation is typically limited to the first two kernels of the series. The place of the bilinear model structure in these contributions is, in a sense, incidental to the approach since the basis functions can be used to approximate a very large set of system classes.

Other schemes have eschewed approximation and attempted to deal directly with the bilinear model structure itself. Favoreel *et. al.* (Favoreel *et al.*, 1999) extended a well-known subspace-based identification method to handle the bilinear class. Initially, this algorithm was limited to the case of white inputs, however, the theory is such that this restriction is now able to be relaxed (Favoreel, 1999). Other workers have also adopted a subspace-based approach (Verdult and Verhaegen, 1999).

One of the great strengths of subspace algorithms is their ability to identify systems directly in state-space without the need to explicitly parameterise the state-space matrices. The result is that the bilinear subspace algorithm mentioned above is able to perform multivariable identification just as simply as for the single-input single-output case.

Another set of notable contributions, this time motivated by the well-known philosophy of Maximum Likelihood, formulates the problem as finding the set of parameters which maximise the relevant likelihood function for the observed data (Balakrishnan, 1972; Bruni *et al.*, 1972; Fnaiech and Ljung, 1987; Gabr, 1985). One drawback with these endeavours is that the likelihood function for this problem is a non-convex function of the parameters and therefore a gradient-based search algorithm is typically required. Unlike the case of the subspace-based algorithm the modifications required for multivariable identification are by no means trivial.

The algorithm proposed in this paper is based on the Expectation Maximisation algorithm (Dempster *et al.*, 1977; Shumway, 1982), and identifies directly in state-space form without requiring an explicit parameterisation.

This latter advantage is a key reason for the interest in subspace-based algorithms , but, unlike that approach, where it is not clear exactly what cost function is being optimised, the algorithm proposed here seeks a Maximum Likelihood estimate.

2. MAXIMUM LIKELIHOOD ESTIMATION

Consider the problem of estimating the parameters of a particular model structure, collected in a vector $\theta \in \mathbf{R}^p$, from a set of measured data $\{Y_k\}_{k=1}^N$. The Maximum Likelihood approach to this problem seeks the value of θ that maximises the probability of the observed data. That is, an estimate, $\widehat{\theta}_N$, based on the N data observations is given as

$$\widehat{\theta}_N = \arg\max_{\theta} p(Y_1, Y_2, \ldots, Y_N \mid \theta). \qquad (4)$$

In this context the joint probability density function $p(Y_1, Y_2, \ldots, Y_N \mid \theta)$ is known as the 'Likelihood Function' and once the values of the observations $\{Y_k\}$ are specified this is a function purely of the model parameters θ.

This method of estimation, first introduced by Fisher (Fisher, 1912), has become popular and widely used, in large part due to the well-known and desirable features of consistency, asymptotic normality and statistical efficiency (Caines, 1988; Hannan and Deistler, 1988; Lehmann, 1983; Ljung, 1999).

While these theoretical underpinnings are very attractive, it is in the implementation of the scheme that difficulties arise.

Specifically, the likelihood function $p(Y_1, Y_2, \ldots, Y_N \mid \theta)$ is frequently a non-convex function of the model parameters, requiring solution by some form of iterative optimisation algorithm. In the event that the likelihood function is smooth, a gradient-based numerical search can be employed (Ljung, 1999; Ljung, 2000). However, computation of these gradients can be involved, particularly in the multivariable case. Furthermore, they depend on a parameterisation choice which can imply poor numerical conditioning.

The contribution of this paper is to present an algorithm for likelihood maximisation that is suited to estimating the parameters of the bilinear model (1), (2) directly in state-space form, which does not require the computation of likelihood gradients, and is numerically robust.

3. THE EM ALGORITHM

The Expectation Maximisation (EM) algorithm (Dempster *et al.*, 1977) is a method for obtaining Maximum Likelihood estimates that has attracted a large amount of attention in the signal processing and mathematical statistics literature (Bock and Aitken, 1981; Fessler *et al.*, 1993) but one which is relatively unknown in the automatic control community.

An essential feature of the EM algorithm is the postulate of an unobserved 'complete' data set, Z, that contains what is actually observed Y, plus other observations X, which one might wish were available, but in fact are not, and are termed the 'missing' data. That is, $Z = (X, Y)$, so that by Bayes' rule

$$p(Z \mid Y) = \frac{p(Z, Y)}{p(Y)} = \frac{p(Z)}{p(Y)}.$$

Therefore,

$$p(Y) = \frac{p(Z)}{p(Z \mid Y)}. \qquad (5)$$

In what follows conditional dependence will sometimes be denoted by subscripting. For example $p_\theta(Y) \equiv$

$p(Y \mid \theta)$. Taking the logarithm of (5) and making all densities conditional upon the value of θ leads to

$$\log p_\theta(Y) = \log p_\theta(Z) - \log p_\theta(Z \mid Y),$$

which is

$$\log p_\theta(Y) = \log p_\theta(X,Y) - \log p_\theta(X \mid Y).$$

Up to this point in the derivation it has been assumed that the complete data set has been available. Unfortunately since the missing data is not available, the calculations above can not be made. However, if the missing data is estimated as its expected value conditioned on a guess at the parameters of the system, θ', and the observed data, Y, then

$$\begin{aligned}\mathbf{E}_{\theta'}\{\log p_\theta(Y) \mid Y\} = &\mathbf{E}_{\theta'}\{\log p_\theta(X,Y) \mid Y\} \\ &- \mathbf{E}_{\theta'}\{\log p_\theta(X \mid Y) \mid Y\}.\end{aligned}$$

The fact that the left hand side of this equation is independent of the missing data and that it holds for all X allows the density function to be pulled through the expectation integral to yield

$$\begin{aligned}L(\theta) \triangleq \log p_\theta(Y) &= \mathbf{E}_{\theta'}\{\log p_\theta(X,Y) \mid Y\} \\ &\quad -\mathbf{E}_{\theta'}\{\log p_\theta(X \mid Y) \mid Y\} \\ &= \mathcal{Q}(\theta,\theta') - \mathcal{V}(\theta,\theta'),\end{aligned}$$

where the obvious definitions for $\mathcal{Q}(\cdot,\cdot)$ and $\mathcal{V}(\cdot,\cdot)$ apply.

Now, by virtue of the concavity of the logarithm, an application of Jensen's Inequality ensures that $\mathcal{V}(\theta,\theta') \leq \mathcal{V}(\theta',\theta')$ with equality if and only if $\theta = \theta'$. Thus, if θ is chosen so that

$$\mathcal{Q}(\theta,\theta') \geq \mathcal{Q}(\theta',\theta'), \tag{6}$$

then

$$L(\theta) \geq L(\theta'). \tag{7}$$

The previous discussion allows the EM algorithm to be stated as follows:

(1) Initialisation: Procure an initial estimate of the parameters θ_0.
(2) E Step: Calculate $\mathcal{Q}(\theta, \widehat{\theta}_n)$.
(3) M Step: $\widehat{\theta}_{n+1} = \arg\max_\theta \mathcal{Q}(\theta, \widehat{\theta}_n)$.
(4) Upon convergence, terminate, otherwise go to step 2.

Equation (7), subject to (6), ensures that this algorithm generates a sequence of parameter estimates with non-decreasing likelihoods.

The test for convergence required by step 4 could be implemented in many ways. One obvious one is to monitor the log likelihood function, $L(\cdot)$, at each step and terminate when the rate of increase between iterations drops below some predetermined value.

4. APPLICATION OF THE EM ALGORITHM TO BILINEAR SYSTEMS

Let us re-examine the bilinear system model,

$$x_{k+1} = Ax_k + N[u_k \otimes x_k] + Bu_k + w_k, \tag{8}$$

$$y_k = Cx_k + Du_k + v_k. \tag{9}$$

In a conventional system identification experiment the system's inputs are known and its outputs are measured. Therefore the measured outputs will make up the observed data, Y. The set of known inputs will be denoted by the symbol U. However, in order to apply the EM algorithm to this system it is necessary first to decide what constitutes the missing data.

Note that the matrices A, B, N and Q could be extracted by linear regression from the equation (8) if the state sequence was known. Similarly, given the state sequence the matrices C, D and R could be calculated by applying linear regression techniques to equation (9). Since knowledge of the state sequence would be of such enormous help in the estimation process we define it to be the missing data. i.e. $X \triangleq \{x_k\}_{k=1}^N$.

Given this choice, at each iteration of the EM algorithm the function $\mathcal{Q}(\theta,\theta')$ must be maximised over θ. This function, defined as

$$\mathcal{Q}(\theta,\theta') = \mathbf{E}_{\theta'}\{\log p_\theta(X,Y) \mid Y, U\}, \tag{10}$$

may be calculated using Lemma 1.

Lemma 1. Let the initial state x_1 be distributed as $x_1 \sim \mathcal{N}(\mu, P_1)$, then

$$\begin{aligned}-2\mathcal{Q}(\theta,\theta') &= \log|P_1| + N\log|Q| + N\log|R| \\ &+\mathrm{Tr}\left\{P_1^{-1}\left(\Delta - \widehat{x}_1\mu^T - \mu\widehat{x}_1^T + \mu\mu^T\right)\right\} + \\ &\mathrm{Tr}\left\{Q^{-1}\left(\Phi - \Pi\left[\begin{array}{ccc}A & N & B\end{array}\right]^T - \left[\begin{array}{ccc}A & N & B\end{array}\right]\Pi^T\right.\right. \\ &\left.\left.+\left[\begin{array}{ccc}A & N & B\end{array}\right]\Lambda\left[\begin{array}{ccc}A & N & B\end{array}\right]^T\right)\right\} + \mathrm{Tr}\left\{R^{-1}\right. \\ &\left(\Omega - \Psi\left[\begin{array}{cc}C & D\end{array}\right]^T - \left[\begin{array}{cc}C & D\end{array}\right]\Psi^T + \left[\begin{array}{cc}C & D\end{array}\right]\times\right. \\ &\left.\left.\Gamma\left[\begin{array}{cc}C & D\end{array}\right]^T\right)\right\} + (n(N+1)+N)\log(2\pi)\end{aligned} \tag{11}$$

where

$$\widehat{x}_t \triangleq \mathbf{E}_{\theta'}\{x_t \mid Y, U\}), \quad \Delta \triangleq \mathbf{E}_{\theta'}\{x_1 x_1^T \mid Y, U\},$$

$$\Phi \triangleq \sum_{t=2}^{N} \mathbf{E}_{\theta'}\{x_t x_t^T \mid Y, U\}, \quad \Psi \triangleq \sum_{t=1}^{N} y_t\left[\begin{array}{cc}\widehat{x}_t^T & u_t^T\end{array}\right],$$

$$\Pi \triangleq \sum_{t=2}^{N} \mathbf{E}_{\theta'}\left\{x_t \Xi^T \mid Y, U\right\}, \quad \Omega \triangleq \sum_{t=1}^{N} y_t y_t^T,$$

$$\Gamma \triangleq \sum_{t=1}^{N} \mathbf{E}_{\theta'}\left\{\left[\begin{array}{c}x_t \\ u_t\end{array}\right]\left[\begin{array}{cc}x_t^T & u_t^T\end{array}\right] \mid Y, U\right\},$$

$$\Lambda \triangleq \sum_{t=2}^{N} \mathbf{E}_{\theta'}\left\{\Xi\Xi^T \mid Y, U\right\}, \qquad \text{and}$$

$$\Xi \triangleq \left[\begin{array}{c}x_{t-1} \\ u_{t-1} \otimes x_{t-1} \\ u_{t-1}\end{array}\right].$$

PROOF. Noting that (8) is Markovian, repeated applications of Bayes' Rule allows the joint likelihood function $p_\theta(X, Y \mid U)$ to be decomposed as

$$\begin{aligned} &p_\theta(X, Y \mid U) = p_\theta(X \mid U) p_\theta(Y \mid X, U) \qquad (12)\\ &= p_\theta(x_1, x_2, \dots x_N \mid u_1, u_2, \dots u_N) \times \\ &\quad p_\theta(y_1, y_2, \dots, y_N \mid x_1, x_2, \dots x_N, u_1, u_2, \dots u_N) \\ &= p_\theta(x_1) \prod_{t=2}^{N} p_\theta(x_t \mid x_{t-1}, u_{t-1}) \prod_{t=1}^{N} p_\theta(y_t \mid x_t, u_t). \end{aligned}$$

From equation (3) with $S = 0$, it is known that the measurement and state corruptions are i.i.d. and mutually independent with a normal distribution. Via equations (8) and (9) we thus get $p_\theta(x_t \mid x_{t-1}, u_{t-1}) = \mathcal{N}(Ax_{t-1} + Nu_{t-1} \otimes x_{t-1} + Bu_{t-1}, Q)$, and $p_\theta(y_t \mid x_t, u_t) = \mathcal{N}(Cx_t + Du_t, R)$.

Inserting these densities into (12), taking the logarithm of both sides, and then applying the conditional expectation operator as required by (10), yields

$$\begin{aligned} &-2\mathcal{Q}(\theta, \theta') = \log|P_1| + \mathrm{Tr}\left\{P_1^{-1} + N\log|Q| + \right.\\ &N\log|R| + \mathbf{E}_{\theta'}\left\{(x_1 - \mu)(x_1 - \mu)^T \mid Y, U\right\} + \\ &\sum_{t=1}^{N} \mathrm{Tr}\left\{R^{-1}\mathbf{E}_{\theta'}\left\{(y_t - Cx_t - Du_t)(y_t - Cx_t - \right.\right.\\ &\left.\left.Du_t)^T \mid Y, U\right\}\right\} + (n(N+1) + N)\log(2\pi) + \\ &\sum_{t=2}^{N} \mathrm{Tr}\left\{Q^{-1}\mathbf{E}_{\theta'}\left\{(x_t - Ax_{t-1} - Nu_{t-1} \otimes x_{t-1} - \right.\right.\\ &Bu_{t-1})(x_t - Ax_{t-1} - Nu_{t-1} \otimes x_{t-1} - \\ &\left.\left.Bu_{t-1})^T \mid Y, U\right\}\right\} \qquad (13) \end{aligned}$$

where $\mathrm{Tr}\{\cdot\}$ is the trace operator, and μ and P_1 are now considered part of θ.

Equation (11) follows directly from (13) on substituting the expressions $\widehat{x}_t$, Δ, Φ, Ψ, Π, Ω, Γ and Λ.

5. COMPUTATION OF KALMAN SMOOTHED QUANTITIES

Lemma 1 has shown that in order to compute $\mathcal{Q}(\theta, \theta')$ it is necessary to obtain the Kalman Smoothed quantities $\mathbf{E}_{\theta'}\{x_t x_t^T \mid Y, U\}$, $\mathbf{E}_{\theta'}\{x_t x_{t-1}^T \mid Y, U\}$ and $\mathbf{E}_{\theta'}\{x_t \mid Y, U\}$. Fortunately, by reformulating the bilinear model structure (1), (2) as linear but time-varying and noting that, by (3) the noise processes are normally distributed, most of the required matrices may be calculated directly by a standard Kalman Smoother.

That is, assuming that the system parameters are given by $\theta' \triangleq \{A', N', B', C', D', Q', R', \mu', P_1'\}$, equations (1) and (2) may be rewritten as

$$x_{t+1} = A_t' x_t + B' u_t + w_t, \qquad (14)$$

$$y_t = C' x_t + D' u_t + v_t, \qquad (15)$$

where $A_t' \triangleq A' + N'(u_t \otimes I_n)$ and I_n is the $n \times n$ identity matrix.

Thus, by introducing the following notation,

$$\begin{aligned} \widehat{x}_{t|N} &\triangleq \mathbf{E}_{\theta'}\{x_t \mid Y, U\}, \\ P_{t|N} &\triangleq \mathbf{E}_{\theta'}\{(x_t - \widehat{x}_t)(x_t - \widehat{x}_t)^T \mid Y, U\}, \end{aligned}$$

the Kalman smoother may be implemented with the backward recursion (Jazwinski, 1970),

$$J_{t-1} = P_{t-1|t-1}(A_{t-1}')^T (P_{t|t-1})^{-1} \qquad (16)$$

$$\widehat{x}_{t-1|N} = \widehat{x}_{t-1|t-1} + J_{t-1}(\widehat{x}_{t|N} - A_{t-1}'\widehat{x}_{t-1|t-1}) \qquad (17)$$

$$P_{t-1|N} = P_{t-1|t-1} + J_{t-1}(P_{t|N} - P_{t|t-1})J_{t-1}^T \qquad (18)$$

for $t = N, N-1, \dots, 1$ with the obvious definitions for $\widehat{x}_{t-1|t-1}$ and $P_{t-1|t-1}$.

In turn, the expectations $\widehat{x}_{t-1|t-1}$ and $P_{t-1|t-1}$ may be calculated using the well-known Kalman filter recursion (Jazwinski, 1970)

$$\widehat{x}_{t|t-1} = A_{t-1}'\widehat{x}_{t-1|t-1} + Bu_{t-1} \qquad (19)$$

$$P_{t|t-1} = A_{t-1}' P_{t-1|t-1}(A_{t-1}')^T + Q' \qquad (20)$$

$$K_t = P_{t|t-1}(C')^T (C' P_{t|t-1}(C')^T + R')^{-1} \qquad (21)$$

$$\widehat{x}_{t|t} = \widehat{x}_{t|t-1} + K_t(y_t - C'\widehat{x}_{t|t-1} - D'u_t) \qquad (22)$$

$$P_{t|t} = P_{t|t-1} - K_t C' P_{t|t-1} \qquad (23)$$

for $t = 1 \dots N$.

The one object still required to evaluate equation (11) is the cross-covariance term

$$M_{t|N} \triangleq \mathbf{E}_{\theta'}\{(x_t - \widehat{x}_t)(x_{t-1} - \widehat{x}_{t-1})^T \mid Y, U\},$$

and this may be calculated using the recursion

$$\begin{aligned} M_{t-1,N} &= P_{t-1|t-1}J_{t-2}^T \qquad (24)\\ &\quad + J_{t-1}(M_{t|N} - A_{t-1}'P_{t-1|t-1})J_{t-2}^T \end{aligned}$$

for $t = N, N-1, \dots, 2$ and where $M_{N|N}$ is initialised as

$$M_{N,N} = (I - K_N C')A_{N-1}' P_{N-1|N-1}. \qquad (25)$$

Now the quantities of interest, $\mathbf{E}_{\theta'}\{x_t x_t^T \mid Y, U\}$ and $\mathbf{E}_{\theta'}\{x_t x_{t-1}^T \mid Y, U\}$, may be calculated as

$$\begin{aligned} \mathbf{E}_{\theta'}\{x_t x_t^T \mid Y, U\} &= \widehat{x}_t \widehat{x}_t^T + P_{t|N}, \\ \mathbf{E}_{\theta'}\{x_t x_{t-1}^T \mid Y, U\} &= \widehat{x}_t \widehat{x}_{t-1}^T + M_{t|N}. \end{aligned}$$

6. THE MAXIMISATION STEP

The next step of the EM algorithm is to maximise the function $\mathcal{Q}(\theta, \theta')$ over θ. The following lemma provides the means to do that.

Lemma 2. The function $\mathcal{Q}(\theta, \theta')$ given by (11) is maximised by choosing

$$\mu = \widehat{x}_1, \tag{26}$$

$$P_1 = \Delta - \widehat{x}_1 \widehat{x}_1^T, \tag{27}$$

$$\begin{bmatrix} A & N & B \end{bmatrix} = \Pi \Lambda^{-1}, \tag{28}$$

$$Q = N^{-1}(\Phi - \Pi \Lambda^{-1} \Pi^T), \tag{29}$$

$$\begin{bmatrix} C & D \end{bmatrix} = \Psi \Gamma^{-1} \tag{30}$$

and

$$R = N^{-1}(\Omega - \Psi \Gamma^{-1} \Psi^T). \tag{31}$$

PROOF. Rather than maximise $\mathcal{Q}(\theta, \theta')$ directly its negative shall be minimised.

Define the functions

$$\mathcal{Q}_1(\mu, P_1) \triangleq \mathrm{Tr}\left\{ P_1^{-1} \left(\Delta - \widehat{x}_1 \mu^T - \mu \widehat{x}_1^T + \mu \mu^T \right) \right\} + \log |P_1|,$$

$$\mathcal{Q}_2([A, N, B], Q) \triangleq N \log |Q| + \mathrm{Tr}\left\{ Q^{-1} \left(\Phi - \Pi \begin{bmatrix} A & N & B \end{bmatrix}^T \begin{bmatrix} A & N & B \end{bmatrix} \Pi^T + \begin{bmatrix} A & N & B \end{bmatrix} \Lambda \begin{bmatrix} A & N & B \end{bmatrix}^T \right) \right\},$$

and $\mathcal{Q}_3([C, D], R) \triangleq N \log |R| + \mathrm{Tr}\left\{ R^{-1} \left(\Omega - \Psi \begin{bmatrix} C & D \end{bmatrix}^T - \begin{bmatrix} C & D \end{bmatrix} \Psi^T + \begin{bmatrix} C & D \end{bmatrix} \Gamma \begin{bmatrix} C & D \end{bmatrix}^T \right) \right\}$.

so that, by equation (11),

$$-2\mathcal{Q}(\theta, \theta') = \mathcal{Q}_1(\mu, P_1) + \mathcal{Q}_2([A, N, B], Q) + \mathcal{Q}_3([C, D], R) + (n(N+1) + N\ell) \log(2\pi).$$

Since the last term on the right hand side of this equation is constant it may be ignored in any maximisation operation. Maximising $\mathcal{Q}(\theta, \theta')$ is made much easier by the fact that $\mathcal{Q}_1(\cdot)$ depends only upon the distribution of the initial state x_1, $\mathcal{Q}_2(\cdot)$ is a function purely of A, N, B and Q, and $\mathcal{Q}_3(\cdot)$ depends solely upon the remaining parameters. Thus $\mathcal{Q}(\theta, \theta')$ may be maximised by minimising each of the subfunctions $\mathcal{Q}_x(\cdot)$ separately.

Lemma 3 may be employed to differentiate $\mathcal{Q}_1(\cdot)$ with respect to μ. Setting the result to zero and trivially simplifying yields $\mu = \widehat{x}_1$. Substituting this back into $\mathcal{Q}_1(\cdot)$ then provides

$$\mathcal{Q}_1(\mu, P_1) \triangleq \log |P_1| + \mathrm{Tr}\left\{ P_1^{-1} \left(\Delta - \widehat{x}_1 \widehat{x}_1^T \right) \right\}.$$

Differentiating this expression, this time with respect to P_1 and again equating to zero gives

$$P_1^{-1} - P_1^{-2} \left(\Delta - \widehat{x}_1 \widehat{x}_1^T \right) = 0,$$

which has the solution $P_1 = \Delta - \widehat{x}_1 \widehat{x}_1^T$.

Identical arguments may be applied to minimise the functions $\mathcal{Q}_2(\cdot)$ and $\mathcal{Q}_3(\cdot)$.

7. THE OVERALL ALGORITHM

This section summarises the developments of the previous discussion and provides the overall estimation algorithm for the identification of bilinear systems.

(1) Initialise the parameter estimate θ_0;
(2) Formulate the system as shown in (14) and (15);
(3) **(E-Step)** Use the recursion equations (16) through (25) to compute (11);
(4) **(M-Step)** Maximise $\mathcal{Q}(\theta, \theta_k)$ over θ by choosing θ_{k+1} according to equations (26) through (31);
(5) If converged, terminate, otherwise return to step 2.

8. SIMULATION EXAMPLE

This section provides a simulation example in order to demonstrate the utility of the EM algorithm approach to Maximum Likelihood estimation proposed in this paper.

Consider the following bilinear system in state-space form

$$x_{t+1} = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.3 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 & 0.2 & 0 \\ 0 & 2 & 0 & 0.5 \end{bmatrix} u_t \otimes x_t + \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} u_t, \tag{32}$$

$$y_t = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} x_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} u_t + v_t, \tag{33}$$

where v_t are i.i.d. random variables satisfying $v_t \sim \mathcal{N}\left(0, 10^{-4} I\right)$.

The observed data was generated by applying an input signal $\{u_t\}$ consisting of i.i.d. random variables distributed as $u_t \sim \mathcal{N}(0, I)$ to this system and sampling $N = 400$ data points.

An estimate for the system parameters of a second-order bilinear system was determined from this data using the subspace identification software from `ftp.esat.kuleuven.ac.be`, implementing the results of (Favoreel *et al.*, 1999). Two block rows and 380 columns were specified for the data Hankel matrix used in that identification.

A series of Maximum-Likelihood estimates was generated via the EM algorithm described in this paper, starting from the parameter estimate provided by the subspace algorithm in addition to $Q = 1.2 \times 10^{-2} I_n$, $R = 0.8 \times 10^{-2} I_\ell$.

Table 1 shows how well the algorithms captured the dynamic modes of the true system by profiling the eigenvalues of the estimated A and N matrices, as calculated by the subspace and EM algorithms, against their true values. The values produced by the EM algorithm are very close to those of the true system - much closer than those of the initial system produced by a subspace-based algorithm. The EM algorithm's good performance is despite the number of observations being fairly small.

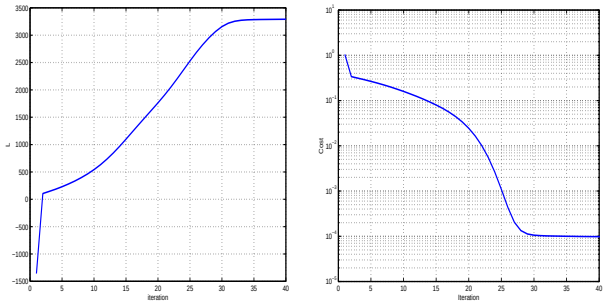
Table 1. Eigenvalues of parameter matrices

	True System	Initial Estimate	EM Estimate
A	0.3, 0.5	0.2457, 0.4694	0.3001, 0.4999
$N(:, 1:2)$	0.4, 0.6	0.2704, 0.3978	0.4000, 0.6000
$N(:, 3:4)$	0.2, 0.5	0.0808, 0.3462	0.2010, 0.5004

The plots in Figure 1 demonstrate the speed with which convergence was achieved by the EM algorithm. While the speed of convergence appears to have been quite high - requiring only about 30 iterations - the algorithm has converged to the Maximum Likelihood estimate since the final value of the (per output) prediction error, defined as

$$\sum_{t=1}^{N}(y_t - \widehat{C}_k\widehat{x}_{t|t-1} - \widehat{D}_k u_t)^2,$$

is equal to $\mathbf{E}\{v_t^2\} = 1e^{-4}$.



(a) Log Likelihood

(b) Prediction error (per output)

Fig. 1. *Convergence of the EM algorithm.*

9. CONCLUSIONS

The contribution of this paper was to present a novel Expectation Maximisation-based approach to Maximum Likelihood estimation of bilinear systems. A simulation study demonstrated that the algorithm performs well.

Further work is required to investigate fully the features of this algorithm. At present its properties are known only from simulation studies. In particular, an analysis of it convergence properties is essential.

Appendix A. TECHNICAL LEMMATA

Lemma 3. Suppose $M, N \in \mathbf{R}^{n\times n}$ and M is invertible. Then

$$\frac{\mathrm{d}}{\mathrm{d}M}\ln|M| = M^{-T}, \qquad \frac{\mathrm{d}}{\mathrm{d}M}M^{-1} = -M^{-2},$$
$$\frac{\mathrm{d}}{\mathrm{d}M}\mathrm{Tr}\{MN\} = N^T.$$

Appendix B. REFERENCES

Baheta, R.S., R.R. Mohler and H.A. Spang (1979). A new cross correlation algorithm for volterra kernel estimation of bilinear systems. *IEEE Transactions* **AC-24**, 661–664.

Balakrishnan, A.V. (1972). *Theory and applications of variable structure systems*. Chap. Modelling and identification theory - a flight control application, pp. 1–21. Academic Press.

Bock, R.D. and M. Aitken (1981). Marginal maximum likelihood estimation of item parameters: an application of an em algorithm. *Psychometrika* **46**, 443–459.

Bruni, C., G. di Pillo and G. Koch (1972). *Theory and applications of variable structure systems*. Chap. Mathematic models and identification of bilinear systems, pp. 137–152. Academic Press.

Caines, P.E. (1988). *Linear Stochastic Systems*. John Wiley and Sons. New York.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Favoreel, W. (1999). Subspace methods for identification and control of linear and bilinear systems. PhD thesis. Departement Elektrotechniek, Katholieke Universiteit Leuven.

Favoreel, Wouter, Bart de Moor and Peter van Overschee (1999). Subspace identification of bilinear systems subject to white inputs. *IEEE Transactions on Automatic Control* **44**(6), 1157–1165.

Fessler, J.A., N.H. Clinthorne and W.L. Rogers (1993). On complete data spaces for pet reconstruction algorithms. *IEEE. Transactions on Nuclear Science* **40**, 1055–1061.

Fisher, R.A. (1912). On an absolute criterion for fitting frequency curves. *Mess. Math.* **41**, 155.

Fnaiech, F. and L. Ljung (1987). Recursive identification of bilinear systems. *International Journal of Control* **45**(2), 453–470.

Gabr, M.M. (1985). A recursive (on-line) identification of bilinear systems. *International Journal of Control* **44**(4), 911–917.

Hannan, E.J. and Manfred Deistler (1988). *The Statistical Theory of Linear Systems*. John Wiley and Sons. New York.

Jazwinski, Andrew (1970). *Stochastic Processes and Optimal Filtering Theory*. Academic Press.

Karanam, V.P., P.A. Frick and R.R. Mohler (1978). Bilinear system identification by walsh functions. *IEEE Transactions* **AC-23**, 709–713.

Lehmann, E.L. (1983). *Theory of Point Estimation*. John Wiley & Sons.

Ljung, Lennart (1999). *System Identification: Theory for the User, (2nd edition)*. Prentice-Hall, Inc.. New Jersey.

Ljung, Lennart (2000). *MATLAB System Identification Toolbox Users Guide, Version 5*. The Mathworks.

Shumway, R.H. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis* **3**(4), 253–264.

Verdult, V. and M. Verhaegen (1999). Subspace identification of mimo bilinear systems. In: *Proceedings of the European Control Conference*. Karlsruhe, Germany.