ROBUST AND SIMPLE ALGORITHMS FOR MAXIMUM LIKELIHOOD ESTIMATION OF MULTIVARIABLE SYSTEMS

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

[*] *Dept. of Elec. & Comp. Eng, Uni. Newcastle, Australia.*
email:`brett@ee.newcastle.edu.au`, *FAX: +61 49 21 69 93*

Abstract: This paper presents novel algorithms for the estimation of dynamic systems. These new methods offer several advantages of being parameterisation free, numerically robust, convergent to statistically optimal estimates, and applicable in a simple fashion to a wide range of multivariable, non-linear and time varying problems. The key tool underlying the new techniques presented here is the 'Expectation-Maximisation' (EM) algorithm.

Keywords: Parameter Estimation, Maximum Likelihood Estimators.

1. INTRODUCTION

In the field of system identifiction, the so-called Maximum Likelihood principle and its relations, such as prediction error techniques, play a key role. There are several reasons for this. Firstly, there is a very large and sophisticated body of theory supporting these methods (Ljung 1999, Caines 1988, Hannan and Deistler 1988, T.Söderström and P.Stoica 1989). This allows important practical issues such as error analysis and performance tradeoffs to be addressed. Secondly, via this latter theory, it is understood that Maximum Likelihood methods are provably statistically optimal in that they (at least asymptotically) achieve the Cramér–Rao Lower Bound (Ljung 1999, Caines 1988, Hannan and Deistler 1988, T.Söderström and P.Stoica 1989). That is, in some sense they provide the most accurate estimates. Finally, Maximum Likelihood type methods provide a general framework which is applicable to a very wide range of estimation problems (Ljung 1999, Caines 1988, Hannan and Deistler 1988, T.Söderström and P.Stoica 1989).

Balancing this, it should be recognised, that despite these features recommending the Maximum Likelihood approach, it is not a panacea. As a result there are also significant bodies of work directed towards alternative approaches, including non-parametric (Ljung 1999), bounded-error (Norton 1987, Milanese and Vicino 1991), and state space subspace-based estimation methodologies (van Overschee and Moor 1996, Larimore 1990).

Furthermore, despite the theoretical advantages of Maximum Likelihood methods, their practical deployment is not always straightforward. This is largely due to the non-convex optimisation problems that are often implied. Typically, these are solved via a gradient-based search strategy such as a Newton type method or one of its derivatives (Ljung 1999, T.Söderström and P.Stoica 1989, Dennis and Schnabel 1983). The success of such approaches depends on the curvature of the Maximum Likelihood cost being optimised, and this is dependent on the chosen system parameterisation. Selecting this can be difficult, particularly in the multivariable case where the cost contours resulting from natural canonical state-space parameterisations imply poor numerical conditioning during gradient-based search (Deistler 2000, McKelvey 1998).

Indeed, the possibility of avoiding these parameterisation-based difficulties is one of the key reasons for the recent intense interest in the new State Space Subspace based System Identification (4SID) methods (van Overschee and Moor 1996, Larimore 1990).

This paper, motivated by all these issues, presents new methods for gradient-search free computation of Maximum Likelihood dynamic system estimates. These new techniques can employ state space model structures (like 4SID methods), but they do not require explicit parameterisation of the system matrices. Furthermore, the numerical procedures involved here can be implemented very efficiently and reliably via well known methods such as QR decomposition. Finally, while these new algorithms are introduced here for the estimation of linear and time-invariant systems, they are very simply extended to more complicated scenarios of non-linear, time varying and missing-data estimation problems.

The central technique employed in this paper is that of the so-called Expectation-Maximisation (EM) algorithm which, for certain classes of Maximum Likelihood estimation problems, has proven to be a robust alternative to gradient-based search for the estimate (Dempster *et al.* 1977).

Despite the successful application of the these EM methods in many other fields such as image processing (Starck *et al.* 1998), speech recognition (Rabiner 1989), and various problems of applied statistics such as epidemiology (Barndorff–Nielsen *et al.* 1999), their potential utility with regard to dynamic system identification problems, particularly those with relevance to control applications, seems to have been largely unappreciated.

2. DYNAMIC SYSTEM ESTIMATION

The estimation problems considered in this paper are ones in which an observed discrete time data record of N samples $Y_N \triangleq \{y_1, y_2, \cdots, y_N\}$ is postulated to depend causally on another data record $U_N \triangleq \{u_1, u_2, \cdots, u_N\}$, and also upon external influences that will be modelled here as realisations of random variables.

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

A very general way of formulating this scenario (for finite dimensional systems) is via a state space description such as

$$x_t = f(x_t, u_t, w_t) \tag{1}$$

$$y_t = g(x_t, u_t, e_t). \tag{2}$$

Here, $y_t \in \mathbf{R}^p$ and $u_t \in \mathbf{R}^m$ are the vector valued data records just mentioned, while $x_t \in \mathbf{R}^n$ is the system state sequence, and $\{w_t\}$, $\{e_t\}$ are independent and identically distributed vector stochastic process such that $\mathsf{E}\{w_t\} = 0$, $\mathsf{E}\{e_t\} = 0$, $\mathsf{E}\{w_t w_t^T\} \triangleq Q \geq 0$ and $\mathsf{E}\{e_t e_t^T\} \triangleq R \geq 0$ where $\mathsf{E}\{\cdot\}$ denotes expectation over the probability space that w_t and e_t are defined on. Together, the unknown functions $f(\cdot,\cdot,\cdot)$ and $g(\cdot,\cdot,\cdot)$ together with Q and R constitute the model that is to be determined on the basis of the data records Y_N and U_N.

A Maximum Likelihood solution to this estimation problem requires the specification of the probability density functions $p_e(\cdot), p_w(\cdot)$ for the random variables e_t and w_t. Based on this the joint probability

$$p(Y_N, U_N \mid f, g) \tag{3}$$

which is dependent on f and g is calculated, and known as a 'likelihood function'. The Maximum Likelihood estimates of f and g are then defined as those which maximise (3). That is, they are such that they maximise the probability that the observed data is consistent with the estimated model.

Typically, this process is formulated slightly differently by proposing specific forms for f and g that depend on some vector of parameters $\theta \triangleq [\theta_1, \theta_2, \cdots, \theta_\ell]$. In this case, the Maximum Likelihood estimate based on the N observations is defined as

$$\widehat{\theta}_N \triangleq \arg\max_{\theta} \; p(Y_N, U_N \mid \theta). \tag{4}$$

This method of system estimation enjoys a wide acceptance and popularity, in large part due to its well-known and desirable properties of consistency, asymptotic normality and statistical efficiency that have been established in a range of works, such as (Hannan and Deistler 1988, Lehmann 1983, Caines 1988, Caines 1988, Ljung 1999) and apply under fairly mild regularity assumptions on f, g, w_t, e_t

Balancing these attractive features that recommend a Maximum Likelihood approach, there is the significant disadvantage that the equation (4) defining the Maximum Likelihood estimate $\widehat{\theta}_N$ is, in general, a non-convex optimisation problem. As a result, calculation of $\widehat{\theta}_N$ requires some sort of numerical search technique.

Since $p(Y_N, U_N|\theta)$ is typically smooth, any gradient-based search technique such as Steepest-Descent or Newton iteration (Dennis and Schnabel 1983, Nocedal and Wright 1999) may be employed for this purpose, and indeed this is the usual approach (Ljung 1999, Ljung 2000*a*). In this case, an approximation θ_k for $\widehat{\theta}_N$ is repeatedly updated to a new approximation θ_{k+1} according to

$$\theta_{k+1} = \theta_k - \mu_k J_k \left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(Y_N, U_N \mid \theta) \bigg|_{\theta=\theta_k} \right] \tag{5}$$

where μ_k is a scalar 'step-length' and J_k is a matrix that may be chosen in various ways (Dennis and Schnabel 1983), but is often related to the Hessian of the cost function, and hence also related to its curvature relative to its parameterisation.

Importantly though, the search strategy (5), by way of requiring a gradient (with respect to a parameterisation θ), in fact also *forces* the use of a parameterisation of the state-space model structure (1), (2). This can lead to important difficulties.

For example, in the case where f and g describe a linear, time invariant (LTI), and multivariable system, it is well known that no surjective mapping exists (hence allowing the description of all possible input-output responses) that is also bijective and therefore ensures that the estimate $\widehat{\theta}_N$ is uniquely defined (Deistler 2000, McKelvey 1998).

Furthermore, in this same LTI case, it is also well known that any simple parameterisation based on canonical forms leads to problems in which Hessian-based choices for J_k become ill-conditioned and lead to slow convergence of the search (5) ().

These difficulties, combined with the fact that subspace-based system identification methods do not require parametrisation of the system matrices (van Overschee and Moor 1996, Larimore 1990), are one of the key features leading to the recent intense interest in them. However, the price paid there is that it is not yet clear what cost function is being optimised by subspace-based estimates. As a result, the theory supporting such approaches is still developing (Deistler *et al.* 1995, Bauer *et al.* 1999).

The contribution of this paper is to show how the theoretical advantages of a Maximum Likelihood approach may be combined with the parameterisation free advantages of a subspace-based method by employing the so-called Expectation Maximisation (EM) algorithm.

3. THE EXPECTATION MAXIMISATION (EM) ALGORITHM

The Expectation Maximisation (EM) algorithm is a technique that, in certain circumstances, can be used to compute Maximum Likelihood estimates without resort to gradient-based search. The method arose in the mathematical statistics community (Dempster *et al.* 1977, Titterington 1984) but has found wide engineering application in areas such as signal processing, pattern recognition and speech recognition (Rabiner 1989, Starck *et al.* 1998).

The key feature of the technique is to exploit the concavity of the $\log$ function (together with the fact that the area under a probability density function is one) so as to guarantee iterations of non-decreasing likelihood whilst avoiding the need to calculate derivatives of the likelihood.

To explain these ideas, note that 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 'incomplete' data. That is $Z = (Y, X)$ so that by Bayes' rule

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

which implies that

$$\log p(Y \mid \theta) = \log p(Z \mid \theta) - \log p(Z \mid Y, \theta) \tag{6}$$

where we note that since $\log x$ is monotonic in x, then finding θ which maximises $\log p(Y \mid \theta)$ is equivalent to finding θ maximising $p(Y \mid \theta)$.

As a consequence of (6), by taking expectations with respect to probabilities defined by an approximation of the parameters θ', and conditional on the observed data $Y = Y_N$, then leads to ($L(\theta) \triangleq \log p(Y_N \mid \theta)$)

$$\begin{aligned} L(\theta) &= \mathbf{E}\left\{\log p(Y \mid \theta) \mid Y = Y_N, \theta = \theta'\right\} \\ &= \mathbf{E}\left\{\log p(Z \mid \theta) \mid Y = Y_N, \theta = \theta'\right\} - \\ &\quad \mathbf{E}\left\{\log p(Z \mid Y, \theta) \mid Y = Y_N, \theta = \theta'\right\} \\ &= \mathcal{Q}(\theta, \theta') - \mathcal{V}(\theta, \theta') \end{aligned}$$

where the following definitions have clearly been made

$$\mathcal{Q}(\theta, \theta') \triangleq \mathbf{E}\left\{\log p(Z \mid \theta) \mid Y = Y_N, \theta = \theta'\right\}, \quad (7)$$

$$\mathcal{V}(\theta, \theta') \triangleq \mathbf{E}\left\{\log p(Z \mid Y, \theta) \mid Y = Y_N, \theta = \theta'\right\}. \quad (8)$$

In this case, the difference in log-likelihood corresponding to two different parameter vectors θ and θ' may be written as

$$\begin{aligned} L(\theta) - L(\theta') = &\left[\mathcal{Q}(\theta, \theta') - \mathcal{Q}(\theta', \theta')\right] + \\ &\left[\mathcal{V}(\theta', \theta') - \mathcal{V}(\theta, \theta')\right]. \end{aligned} \quad (9)$$

The key point now is the following inequality for $\mathcal{V}(\theta, \theta')$ that guarantees non-negativity of the second term in (9).

Lemma 3.1.

$$\mathcal{V}(\theta', \theta') \geq \mathcal{V}(\theta, \theta')$$

with equality if, and only if, $p(Z|Y, \theta) = p(Z|Y, \theta')$ for all Z.

Therefore, as a consequence of this, the decomposition (9) shows that if a value for θ is found that increases $\mathcal{Q}(\theta, \theta')$, then this must also increase the log-likelihood $L(\theta)$. This suggests the following algorithm for iteratively updating an approximation θ_k of the Maximum Likelihood estimate $\widehat{\theta}_N$, to a better one θ_{k+1}.

(1) **E Step** (Compute Expectation)

$$\mathcal{Q}(\theta, \theta_k) \triangleq \mathbf{E}\left\{\log p(Z \mid \theta) \mid Y = Y_N, \theta = \theta_k\right\} \quad (10)$$

(2) **M Step** (Maximise)

$$\text{Compute: } \theta_{k+1} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta_k) \quad (11)$$

This is the the Expectation-Maximisation (EM) algorithm, in which the iterative procedure (11) replaces the gradient based one (5) as a method for finding Maximum Likelihood estimates. The principle underlying it is shown in figure 1. There it is illustrated that the function $\mathcal{Q}(\theta, \theta_k)$ acts as an approximant to the likelihood $L(\theta)$ which is exact at $\theta = \theta_k$, and also (locally) follows the contours of $L(\theta)$ in that $L(\theta)$ increases in directions that $\mathcal{Q}(\theta, \theta_k)$ increases.

Clearly, it is only sensible to employ this approach in cases where maximising $\mathcal{Q}(\theta, \theta')$ is straightforward, and certainly easier than maximising $L(\theta)$ directly. In turn, this will depend on what is chosen as the incomplete data set X. As such, an EM algorithm approach is not always appropriate, but as will now be demonstrated, it is very suitable for a wide range of dynamic system estimation problems of engineering relevance.

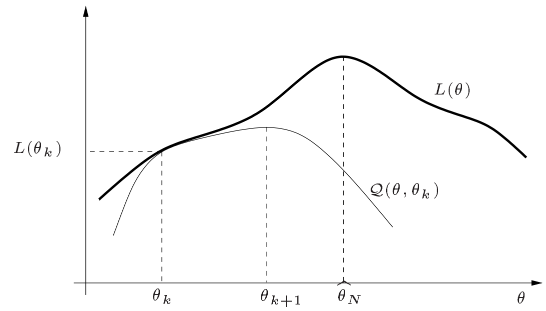


Fig. 1. *Illustration of the principle underlying the EM-algorithm. The function $\mathcal{Q}(\theta, \theta_k)$ acts as a local approximant of the likelihood $L(\theta)$.*

4. APPLICATION TO LINEAR TIME INVARIANT SYSTEMS

This section illustrates the application of the preceding methods by deriving a new algorithm for the estimation of linear, time invariant and multivariable systems that may be represented via the following specialisation of (1), (2).

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad (12)$$

$$y_t = Cx_t + Du_t + e_t. \quad (13)$$

The estimation of the system model $f(x_t, u_t, w_t) = Ax_t + Bu_t + w_t$ and $g(x_t, u_t, e_t) = Cx_t + Du_t + e_t$ then amounts to estimation of the constant matrices $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $C \in \mathbf{R}^{p\times n}$, $D \in \mathbf{R}^{p\times m}$, $Q \in \mathbf{R}^{n\times n}$ and $R \in \mathbf{R}^{p\times p}$. That is, the collected vector of quantities to be estimated is $\theta = \text{Vec}\{A, B, C, D, Q, R\}$ where the $\text{Vec}\{\cdot\}$ operator creates a vector from a matrix by stacking its columns on top of one another.

For the purpose of estimating θ, notice that if x_t where observed, then (12) and (13) would be linear regressions in $[A, B]$ and $[C, D]$ (respectively) which could then be very directly estimated via least-squares. This suggests, that in the interests of defining the incomplete data X so that $\mathcal{Q}(\theta, \theta')$ is simple to maximise, then X should be taken as the unobserved state sequence $X_N \triangleq \{x_0, x_1, \ldots, x_N\}$ (hence the use of the symbol X). That is, the complete data could be taken as

$$Z \triangleq (X_N, Y_N).$$

According to (10), the expection step of the EM-algorithm then requires the calculation of

$$\mathcal{Q}(\theta, \theta') = \mathbf{E}\left\{\log p_\theta(X_N, Y_N) \mid Y = Y_N, \theta = \theta'\right\}$$

which in turn, as in all Maximum Likelihood estimation scenarios, requires the specification of the probability density functions governing the random disturbances w_t and e_t. Here we will assume these are Gaussian as follows

$$w_t \sim \mathcal{N}(0, Q), \qquad e_t \sim \mathcal{N}(0, R) \quad (14)$$

which then allows the computation of $\mathcal{Q}(\theta, \theta')$ via the following Lemma.

Lemma 4.1. For the model structure (12), (13) and the Gaussian assumptions (14) the function $-2\mathcal{Q}(\theta, \theta')$ defined in (7) may be computed as

$$\log|P_0| + N\log|Q| + N\log|R| +$$
$$\text{Tr}\left\{P_0^{-1}\left[(\widehat{x}_{0|N} - \mu)(\widehat{x}_{0|N} - \mu)^T + P_0\right]\right\} +$$

$$\mathrm{Tr}\left\{Q^{-1}\left[\Phi-\Psi[A,B]^T-[A,B]\Psi^T+[A,B]\Gamma[A,B]^T\right]\right\}+ \mathrm{Tr}\left\{R^{-1}\left[\Omega-\Lambda[C,D]^T-[C,D]\Lambda^T+[C,D]\Pi[C,D]^T\right]\right\} \tag{15}$$

with the following definitions applying

$$\widehat{x}_{t|N} \triangleq \mathbf{E}\left\{x_t \mid Y_N,\theta'\right\}, \quad z_t \triangleq \begin{bmatrix} x_t \\ u_t \end{bmatrix}, \tag{16}$$

$$\Lambda \triangleq \sum_{t=1}^{N} y_t\, \mathbf{E}\left\{z_t^T \mid Y_N,\theta'\right\}, \quad \Omega \triangleq \sum_{t=1}^{N} y_t\, y_t^T$$

$$\Pi \triangleq \sum_{t=1}^{N} \mathbf{E}\left\{z_t\, z_t^T \mid Y_N,\theta'\right\}, \Phi \triangleq \sum_{t=1}^{N} \mathbf{E}\left\{x_t\, x_t^T \mid Y_N,\theta'\right\}$$

$$\Psi \triangleq \sum_{t=1}^{N} \mathbf{E}\left\{x_t\, z_{t-1}^T \mid Y_N,\theta'\right\}, \Gamma \triangleq \sum_{t=1}^{N} \mathbf{E}\left\{z_{t-1}\, z_{t-1}^T \mid Y_N,\theta'\right\}$$

and where it has been assumed that the initial distribution on x_0 is

$$x_0 \sim \mathcal{N}(\mu, P_0). \tag{17}$$

This takes care of the Expectation step (10). The particular choice $X = X_N = \{x_0,\cdots,x_N\}$ of the incomplete data that is made here then allows the Maximisation step (11) to be achieved via the expressions of the following Lemma.

Lemma 4.2. The function $\mathcal{Q}(\theta,\theta')$ defined in (15) of Lemma 4.1 is maximised ($-2\log\mathcal{Q}(\theta,\theta')$ is minimised) by the choices

$$[A,B] = \Psi\Gamma^{-1}, \qquad [C,D] = \Lambda\Phi^{-1}, \tag{18}$$

$$Q = N^{-1}(\Phi-\Psi\Gamma^{-1}\Psi^T), R = N^{-1}(\Omega-\Lambda\Pi^{-1}\Lambda^T) \tag{19}$$

$$\mu = \mathbf{E}\left\{x_0 \mid Y_N,\theta'\right\}, P_0 = \mathbf{E}\left\{x_{t-1}\, x_{t-1}^T \mid Y_N,\theta'\right\} \tag{20}$$

There are several points to note here. Firstly, the computations (18) for updating of the estimates of the system matrices A,B,C,D may, via their relationship to least-squares solutions, be computed in very efficient and numerically robust fashions (Golub and Loan 1989).

Secondly, note that the estimate (19) for Q is positive semi-definite by construction since it is a Schur complement of

$$\sum_{t=1}^{N} \mathbf{E}_{\theta'}\left\{\begin{bmatrix} z_t \\ z_{t-1}\end{bmatrix}\begin{bmatrix} z_t^T & z_{t-1}^T\end{bmatrix}\right\} \geq 0.$$

A similar argument indicates that the estimate update (19) is also guaranteed to always yield an $R \geq 0$.

Finally, Lemma 4.2 indicates the the implementation of the Maximisation step of the EM-algorithm for the lineary time-invariant scenario requires the computation of the quantities $\widehat{x}_{t|N} = \mathbf{E}\left\{x_t \mid Y_N,\theta'\right\}$ and

$$\mathbf{E}\left\{x_t x_t^T \mid Y_N,\theta'\right\}, \mathbf{E}\left\{x_t x_{t-1}^T \mid Y_N,\theta'\right\} \tag{21}$$

which are essential to the definition of Λ,Π,Φ,Ψ and Γ.

4.1 *Computation of Conditional Expectations*

In the case considered in this paper where the distributions on the random components e_t and w_t are Gaussian, then recursive expressions exist for the computation of the quantities in (21), as specified in the following Lemma.

Lemma 4.3. For the system (12), (13) and with the definition $\widehat{x}_{t|N} \triangleq \mathbf{E}\left\{x_t \mid Y_N,\theta'\right\}$ together with

$$P_{t|s} \triangleq \mathbf{E}\left\{(\widehat{x}_{t|s}-x_t)(\widehat{x}_{t|s}-x_t)^T|Y_s\right\}, S_t \triangleq P_{t|t}A^T P_{t+1|t}^{-1}$$

then the first two quantities in equation (21) may be computed via the (reverse time) recursions

$$\widehat{x}_{t|N} = \widehat{x}_{t|t} + S_t\left[\widehat{x}_{t+1|N} - Bu_t - A\widehat{x}_{t|t}\right] \tag{22}$$

$$P_{t|N} = P_{t|t} + S_t\left[P_{t+1|N} - P_{t+1|t}\right]S_t^T, \tag{23}$$

$$\mathbf{E}\left\{x_t x_t^T \mid Y_N,\theta'\right\} = P_{t|N} + \widehat{x}_{t|N}\widehat{x}_{t|N}^T \tag{24}$$

where the quantities $\widehat{x}_{t|t}, P_{t|t}, P_{t|t-1}$ involved in these expressions are pre-computed from the (forward in time) Kalman Filter recursions

$$P_{t|t-1} = AP_{t-1|t-1}A^T + Q \tag{25}$$

$$K_t = P_{t|t-1}C^T\left(CP_{t|t-1}C^T + R\right)^{-1} \tag{26}$$

$$P_{t|t} = P_{t|t-1} - K_t C P_{t|t-1} \tag{27}$$

$$\widehat{x}_{t|t-1} = A\widehat{x}_{t-1|t-1} + Bu_{t-1} \tag{28}$$

$$\widehat{x}_{t|t} = \widehat{x}_{t|t-1} + K_t\left(y_t - Du_t - C\widehat{x}_{t|t-1}\right) \tag{29}$$

which are initialised at

$$\widehat{x}_{0|0} = \mu, \qquad P_{0|0} = P_0$$

and where the system matrices A,B,C,D,Q,R used in (25)- (23) are those corresponding to θ'. The final quantity $\mathbf{E}\{x_t x_{t-1}^T \mid Y_N,\theta'\}$ in (21) may be computed via the (reverse time) recursions

$$M_{t|N} = P_{t|t}S_{t-1}^T + S_t(M_{t+1|N} - AP_{t|t})S_{t-1}^T \tag{30}$$

$$\mathbf{E}\left\{x_t x_{t-1}^T \mid Y_N,\theta'\right\} = M_{t|N} + \widehat{x}_{t|N}\widehat{x}_{t-1|N}^T \tag{31}$$

where

$$M_{t|N} \triangleq \mathbf{E}\left\{(\widehat{x}_{t|s}-x_t)(\widehat{x}_{t-1|s}-x_{t-1})^T|Y_N,\theta'\right\}$$

and (30) is initialised at

$$M_{N|N} = (I - K_N C)AP_{N-1|N-1}. \tag{32}$$

4.2 *Estimation Algorithm*

The previous developments may now be summarised in the estimation procedure defined in 1.

At the risk of over-emphasis, the key point of the above algorithm for finding Maximum Likelihood estimates is that, in contrast to the more common gradient based approach, *no* parameterisation of the state-space model structure (12), (13) is required.

Notice too, that from a computational point of view, the above algorithm is comparable to a gradient based approach in that the Recursive Kalman Smoothing operations take the place of the recursive filtering operations necessary for gradient computation.

(1) Initialise estimates at $\theta_k = [A, B, C, D, Q, R]$. For example, a subspace-based estimation method could be employed.
(2) Using the system specification $\theta_k = [A, B, C, D, Q, R]$, run the Run Kalman-Filter recursions (25)-(29) followed by the Kalman Smoother (type) recursions (22), (23), (32) (30) in order to compute the quantities defined in Lemma 4.2.
(3) Maximise $\mathcal{Q}(\theta, \theta_k)$ over θ via the choices (18) and (19) in order to provide an improved estimate θ_{k+1}.
(4) Return to step 2 and repeat until termination.

Algorithm 1. *EM-based Estimation Algorithm*

Finally, on the issue of judging convergence, and hence terminating the above iterative search, an immediately obvious strategy is to monitor the likelihood function $p(y_1, \cdots, y_N|\theta_k)$, and when its rate of increase drops below a threshold, convergence can be declared. This is the method used in the simulation examples of the following section.

5. SIMULATION EXAMPLE

This section provides two brief simulation examples in order to illustrate the utility of the EM-algorithm approach to Maximum–Likelihood estimation proposed in this paper.

In both cases, the observed data is generated according to a system

$$y_t = G(q)u_t + e_t$$

with $G(q)$ given by

$$\begin{bmatrix} \dfrac{0.0355q + 0.02465}{(q-0.3679)(q-0.9084)} & \dfrac{0.2364q + 0.1038}{(q-0.1353)(q-0.6065)} \\[2ex] \dfrac{0.07601q + 0.05447}{(q-0.4966)(q-0.7408)} & \dfrac{0.1087q + 0.07286}{(q-0.4493)(q-0.6703)} \end{bmatrix}$$

and u_t is an i.i.d. zero mean and unit variance Gaussian process while e_t is also i.i.d. zero mean and Gaussian, but has variance $\mathbf{E}\{e_t^2\} = \sigma^2 = 0.01$.

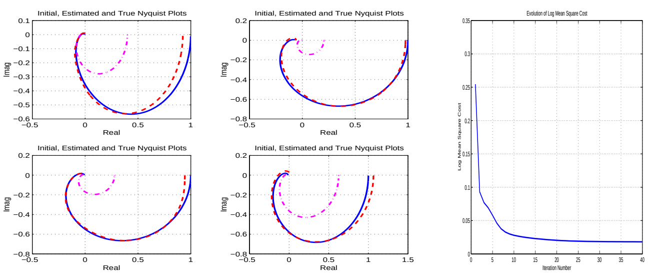
For this scenario, $N = 200$ data samples were collected and Maximum–Likelihood estimates were computed via the EM-algorithm described in this paper and initialised with the starting estimate

$$G(q) = \begin{bmatrix} \dfrac{0.1}{(q-0.5)^2} & \dfrac{0.1}{(q-0.7)^2} \\[2ex] \dfrac{0.1}{(q-0.6)^2} & \dfrac{0.1}{(q-0.4)^2} \end{bmatrix}.$$

The results of this estimation experiment are shown in figure 2. On the left, the relationship between initial and EM-derived ML estimates is shown together with the true response. On the right, the evolution of the log mean-square cost $N^{-1}\sum_{t=1}^{N}\left(y_t - C\widehat{x}_{t|t-1}\right)^2$ is shown as the EM-algorithm iteration progresses. Clearly, the algorithm converges to estimates close to the true system.

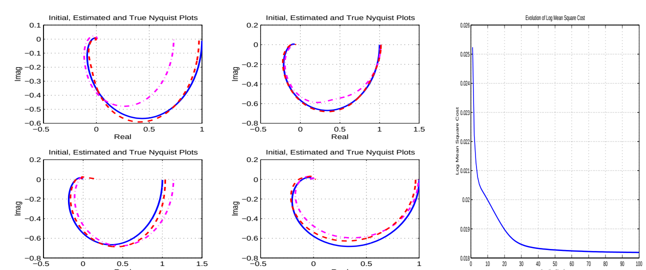
In relation to this simulation, the previous section has raised the possibility of initialising the EM iterations with a subspace-based method, and the results of this strategy for the experimental conditions just outlined are shown in figure 3. There, Overschee and DeMoor's N4SID variant (van Overschee and Moor 1996) of the general class of subspace-based methods is used to provide the initial estimate shown as the dash-dot line. The EM-algorithm of this paper is then used to refine this to be closer to the Maximum Likelihood estimate, with concomitant cost function evolution shown in the right hand diagram of figure 3 and final estimate shown as the dashed line on the left in 3, together (again) with the true system shown as a solid line.

Clearly, the final estimate is significantly improved from the initial subspace-based one, and the key point is that this is achieved in a very simple manner by the parameterisation free method proposed here, while it would be very difficult to implement using a more standard gradient based method that imposed a parameterisation.



(a) Nyquist Plots (b) LS cost

Fig. 2. *EM-algorithm computed ML estimates. Left figure shows initial estimates as dash-dot line, true systems as solid lines, and EM-derived ML estimates as dashed lines. Right figure shows the evolution of the means square cost as the EM-algorithm is iterated.*



(a) Nyquist Plots (b) LS cost

Fig. 3. *Same scenario as previous figure but with initial estimate being found via a subspace-based method.*

6. ERRORS IN VARIABLES

The prediction-error methods presented in (Ljung 1999) and embodied in the pre-eminent software package (Ljung 2000*a*) have become a dominant force in the science of system identifiation. Although the structure of this framework is very general, it is typically applied by means of a θ-parameterised model structure

$$y_t = G_\theta(q)u_t + H_\theta(q)e_t \tag{33}$$

and associated steady state Kalman-Filter innovations

$$\varepsilon_t(\theta) = H_\theta^{-1}(q)\left[y_t - G_\theta(q)u_t\right].$$

However, note that the state-space model structure (12),(13) is more general than the steady state one (33) by virtue of the state disturbance $w_t \sim \mathcal{N}(0, Q)$.

One benefit of this generality arises in the common case where there are noise corruptions $\nu_t \sim \mathcal{N}(0, \Sigma)$ on the observed input (the so-called 'Errors in Variables' scenario) as follows

$$x_{t+1} = Ax_t + B(u_t + \nu_t) + w_t \quad (34)$$
$$y_t = Cx_t + Du_t + e_t. \quad (35)$$

However, this is equivalent to the model structure (12), (13) with $\{w_t\}$ i.i.d. and $w_t \sim \mathcal{N}(0, B\Sigma B^T + Q)$.

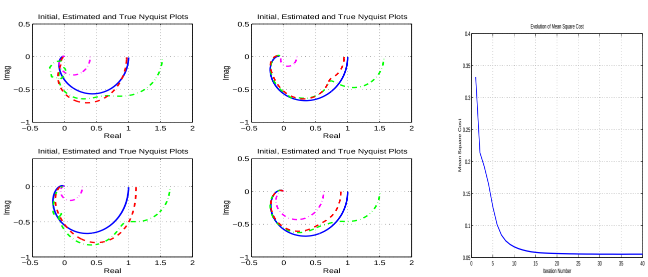
Therefore, the model structure (12), (13) is able to encompass the errors in variables scenario (34), (35) and hence the estimation scheme in algorithm 1 applies to such problems without any modification.

At the same time, translating (34), (35) to transfer function form with $G_1(q) = C(qI - A)^{-1}B + D$, $G_2(q) = C(qI - A)^{-1}B$, $w_t = 0$ implies innovations associated with the model structure (33) of

$$\varepsilon_t(\theta) = H_\theta^{-1}[G_1 - G_\theta u_t] + H_\theta^{-1}[G_2\nu_t + e_t]$$

implying coupling between estimates of noise model and dynamics model parameters, which is well known to lead to possible estimate bias (Ljung 1999).

To illustrate these observations, consider the previous multivariable simulation example repeated but with input measurement noise corruptions of the form $\nu_t \sim \mathcal{N}(0, 0.01I)$. The results of applying the EM based procedure of this paper are then shown in figure 4, and clearly show the efficacy of the approach, in contrast with the bias problems associated with using the model structure (33).



(a) Nyquist Plots (b) LS cost

Fig. 4. *EM-algorithm applied to Errors-in-Variables problem. Line labellings are as for previous figures save that the outer dash-dot nyquist plots on the left represent estimates obtained by fitting an Output-Error Prediction-Error Model structure.*

7. CONCLUSIONS

The contribution of this paper was to suggest a novel, EM-algorithm based approach to Maximum Likelihood estimation of dynamic systems. The key features recommending the approach are that it avoids the need for a particular parameterisation of a state-space model structure, and it is simple to implement.

Although this method is novel in the context considered here, the EM-algorithm itself is quite old, being very well known in (for example) the speech-recognition community as the Baum–Welch method for Hidden Markov Model estimation (Rabiner 1989).

This paper represents only a very preliminary study of this whole topic, and there is much more that needs to be studied in terms of (again, only for example) convergence analysis and extension to more sophisticated model structures

8. REFERENCES

Åström, K. and P. Eykhoff (1971). 'System identification - a survey'. *Automatica* **7**, 123–162.

Barndorff–Nielsen, O. E., D.R. Cox & C. Klüppelberg (1999). *Complex Stoch. Sys.*. Chapman & Hall.

Bauer, D., M. Deistler & W. Scherrer (1999). 'Consist. & asymptotic Normality subspace alg. for sys. without observed inputs'. *Automatica* **35**, 1243–1254. L

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

Deistler, M. (2000). *Model Identification and Adaptive Control*. Springer-Verlag. chapter System Identification - General Aspects and Structure.

Deistler, M., K. Peternell and W. Scherrer (1995). 'Consistency and relative efficiency of subspace methods'. *Automatica* **31**(12), 1865–1875.

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

Dennis, J. and Robert B. Schnabel (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice Hall.

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

Golub, G. and Charles Van Loan (1989). *Matrix Computations*. Johns Hopkins University Press.

Goodwin, G. and R.L. Payne (1977). *Dynamic System Identification*. Academic Press.

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

Larimore, W. (1990). Canonical variate analysis in identification, filtering and adaptive control. In 'Proceedings of the 29th IEEE Conference on Decision and Control, Hawaii'. pp. 596–604.

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

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

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

McKelvey, T. (1998). 'Discussion: 'on the use of min. param. in multivariable ARMAX identification' by R.P. Guidorzi'. *European J. Control* **4**, 93–98.

Milanese, M. and A. Vicino (1991). 'Optimal estimation theory for dynamic systems with set membership uncertainty:An overview'. *Automatica* **27**(6).

Nocedal, J. and Stephen Wright (1999). *Numerical Optimization*. Springer-Verlag, New York.

Norton, J. (1987). 'Identification & application of bounded param. models'. *Automatica* **23**, 497–507.

Norton, J. P. (1985). *An Introduction to Identification*. Academic Press.

Rabiner, L. (1989). 'A tutorial on hidden Markov models and selected applications in speech recognition'. *Proceedings of the IEEE* **77**(2), 257–285.

Starck, J.-L., F. Murtagh and A. Bijaoui (1998). *Image processing and data analysis*. Cambridge University Press. Cambridge. The multiscale approach.

Titterington, D. (1984). 'Recursive parameter estimation using incomplete data'. *Journal of the Royal Statistical Society, Series B* **46**(2), 257–267.

T.Söderström and P.Stoica (1989). *System Identification*. Prentice Hall. New York.

van Overschee, P. and Bart De Moor (1996). *Subspace Identification for Linear Systems*. Kluwer Academic Publishers.