The EM algorithm for Multivariable Dynamic System Estimation.

Brett Ninness[*] Stuart Gibson[†]

Abstract

This paper explores the use of the Expectation-Maximisation (EM) algorithm as an alternative to the more usual gradient search methods for the computation of Maximum-Likelihood estimates of linear dynamic systems. This new approach is shown to afford several advantages, particularly where multivariable systems are concerned, since no a-priori parameterisation of the system is required.

Technical Report EE200101, Department of Electrical and Computer Engineering,
University of Newcastle,AUSTRALIA

1 Introduction

Consider the problem of observing multivariable input-output data from a dynamic system, and on the basis of this seeking to estimate a model relating these signals. In particular, suppose that an input signal $u_t \in \mathbf{R}^m$ is related to an output signal $y_t \in \mathbf{R}^p$ according to

$$y_t = G(q)u_t + e_t \tag{1}$$

where $G(q)$ is a $p \times m$ matrix of scalar transfer functions, and $e_t \in \mathbf{R}^p$ is a zero mean i.i.d. vector process with variance $\mathbf{E}\{e_t e_t^T\} = R$.

The estimation of the dynamics $G(q)$ on the basis of N point (noise corrupted) data records $\{u_1, \cdots, u_N\}$, $\{y_1, \cdots, y_N\}$ may then be achieved in a variety of ways [9, 15]. In particular, while many non-parametric approaches are available, a large class of available estimation techniques employ the use of a model structure, and indeed this sort of description is what is commonly required in automatic control applications.

In relation to this need for a parameterised model structure, the multivariable nature of the problem suggests a state-space description of the following form [9, 10]

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t, & (2)\\ y_t &= Cx_t + Du_t + e_t. & (3) \end{aligned}$$

[*] This work was supported by the Australian Research Council and the Centre for Integrated Dynamics and Control. This author is with the Department of Electrical and Computer Engineering, University of Newcastle, Australia and can be contacted at email:`brett@ee.newcastle.edu.au` or FAX: +61 2 49 21 69 93

[†] This author is also with the Department of Electrical and Computer Engineering, University of Newcastle, Australia and can be contacted at email:`shgibson@ee.newcastle.edu.au` or FAX: +61 2 49 21 69 93

In this case, estimation of $G(q) = C(qI - A)^{-1}B + D$ becomes an issue of estimating A, B, C and D. In (2), $x_t \in \mathbf{R}^n$ is the system state, and $\{w_t\}$ is a zero mean i.i.d. vector process, independent of e_t and with variance $\mathbf{E}\{w_t w_t^T\} = Q$. Clearly, equivalence between the combination (2), (3) and (1) is achieved for $Q = 0$, but it will be useful in what follows for $Q > 0$ to be considered at the beginning of the estimation process in order to indicate initial uncertainty in the quantities A, B, C and D.

Of particular relevance to this estimation problem is the newly developed class of State-Space Subspace Identification methods[17, 16, 7, 18] which provide A, B, C, D estimates in closed form, via simply numerical procedures, and in a manner that does not require any prior parameterisation of the A, B, C, D matrices.

This is in contrast to Maximum-Likelihood methods which, although enjoying a rich theoretical basis and understanding [9, 5, 8], and in particular being statistically efficient, also require an iterative approach to solve an optimisation problem (which is possibly non-convex) over the likelihood surface.

To date, the attraction of statistical efficiency, combined with theory supporting the simple computation of confidence regions for the ensuing estimates, has preserved interest in Maximum-Likelihood-type estimates (eg. prediction-error estimates [9, 10]), with iterative search for the estimate being a gradient based method such as a Gauss–Newton type method [9, 10]. Unfortunately, computing the necessary gradient requires the imposition of a parameterisation on (2), (3).

The contribution of this paper is to examine how the Expectation Maximisation (EM) algorithm may be used (in place of a gradient based method) for the computation of Maximum Likelihood based estimates, and in such a way that all the advantages enjoyed by subspace type methods, such as lack of parameterisation, and numerical simplicity, are preserved.

Although not widely employed in the control-theory literature, the EM algorithm enjoys a high profile in other fields (signal processing, for example) as an iterative method for finding Maximum Likelihood estimates. The key idea is that the concavity of the logarithm function is exploited to guarantee a sequence of increasingly accurate estimates without any need for calculation of gradients (or Hessians) as are normally required by a Gauss-Newton (or Newton) search strategy.

2 Maximum Likelihood Estimation

In order to establish some notation and historical perspective, we begin be providing a précis of the Maximum–Likelihood approach and its properties when applied to the estimation problem just outlined.

Firstly, for the employment this method, it is necessary for the probability density functions $p_e(\cdot), p_w(\cdot)$ for the random variable e_t and w_t to be specified, and then based on this the joint probability

$$p(y_1, \ldots, y_N \mid A, B, C, D)$$

is calculated and known as a 'likelihood function'.

Typically, this is formulated slightly differently by parameterising A, B, C and D in a specific fashion (usually via some canonical form) with all the variables involved being collected in a vector θ. The maximum likelihood estimate (MLE) $\widehat{\theta}_N$ of θ is then defined as

$$\widehat{\theta}_N \triangleq \arg\max_{\theta} \; p(y_1, \ldots, y_N \mid \theta). \tag{4}$$

This method of estimation enjoys a wide acceptance and popularity, in large part due to the following well-known and desirable properties of the scheme which have been established in a range of works, such as [5, 8, 1, 1, 9] (for simplicity, in what follows, it has been assumed that a true parameter vector $\theta_\circ$ that can exactly describe the true data generation system (1) exists in the model set (2), (3)).

Under appropriate regularity conditions on (1) (2) and (3):

Strong Consistency

$$\widehat{\theta}_N \xrightarrow{a.s.} \theta_\circ = \arg\max_{\theta} \lim_{N\to\infty} \mathbf{E}\left\{p(y_1,\dots,y_N \mid \theta)\right\}.$$

Asymptotic Normality

$$\sqrt{N}P_N^{-\frac{1}{2}}(\widehat{\theta}_N - \theta_\circ) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I), \qquad \text{as } N \to \infty$$

where

$$N \cdot P_N^{-1} = -\left.\frac{\mathrm{d}^2}{\mathrm{d}\theta \mathrm{d}\theta^T}\right|_{\theta=\theta_\circ} \mathbf{E}\left\{\log p(y_1,\dots,y_N \mid \theta)\right\} = \mathcal{I}_N,$$

and $\mathcal{I}_N$ is the Fisher Information Matrix associated with $p(y_1,\cdots,y_N \mid \theta)$ and evaluated at $\theta = \theta_\circ$.

Asymptotic efficiency

The above formulation for P_N, due to its indicated relation to the Fischer Information matrix $\mathcal{I}_N$, allows the MLE to asymptotically achieve the Cramér-Rao lower bound

$$\mathrm{Cov}\{\widehat{\theta}_N\} \geq \mathcal{I}_N^{-1}$$

and hence it is asymptotically efficient.

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 $p(y_1,\dots,y_N \mid \theta)$ is typically smooth, any gradient-based search technique such as Steepest-Descent or Newton iteration may be employed for this purpose, and indeed this is the usual approach [9, 10].

However, this strategy, by way of requiring a gradient (with respect to a parameterisation θ), in fact *forces* the use of a parameterisation of the state-space model structure (2), (3).

In the single-input, single-output case, this has not proven to be a major difficulty [9, 10], but this is in contrast to the multi-variable case where the problems of finding a parametrisation and its numerical conditioning are well known [2, 12, 13].

For example, it is clearly desirable that the parameterisation mapping $\theta \rightarrow (A, B, C, D)$ be onto (surjective), and hence allow the description of all possible input-output responses modelled by (2), (3). In the interests of parameter identifiability, it is then important that the mapping also be 1-to-1 (bijective) so that the estimate $\widehat{\theta}_N$ is uniquely defined. However, if both input and output dimension are greater than one, then obtaining such a bijective parameterisation is impossible [2, 11, 4].

These parameterisation-based difficulties, combined with Subspace-based system identification methods[17, 16, 7, 18] not requiring a parametrisation of the model structure (2), (3), are one of the key features leading to the recent intense interest in them. However, the price paid here 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.

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 EM Algorithm for Likelihood Maximisation

As already mentioned this method arose in the mathematical statistics community [3] but has found wide engineering application in areas such as signal processing and pattern recognition.

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

In what follows, a complete set (say $\{y_1, \dots, y_N\}$) will be abbreviated to an uppercase letter (say Y) and conditional dependence on θ will be noted by subscripting; for example, $p(y_1, \dots, y_N \mid \theta) \equiv p_\theta(Y)$.

Now, 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. That is,

$$Z = (Y, X)$$

so that by Bayes rule

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

Therefore

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

which implies that

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

As a consequence, by taking expectations with respect to probabilities defined by a guess at the parameters θ', and conditional on the observed data $Y = Y_N$ leads to

$$\begin{aligned} L(\theta) \triangleq \log p_\theta(Y_N) &= \mathbf{E}_{\theta'}\left\{\log p_\theta(Y) \mid Y = Y_N\right\} \\ &= \mathbf{E}_{\theta'}\left\{\log p_\theta(Z) \mid Y = Y_N\right\} - \mathbf{E}_{\theta'}\left\{\log p_\theta(Z \mid Y) \mid Y = Y_N\right\} \\ &\triangleq Q(\theta, \theta') - V(\theta, \theta'). \end{aligned} \tag{5}$$

with the obvious definitions for $Q(\theta, \theta')$ and $V(\theta, \theta')$.

The key point now is that since $V(\theta, \theta') \leq V(\theta', \theta')$ with equality if and only if $\theta = \theta'$ (this follows by the concavity of the logarithm and the fact that the area under p_θ is one for any θ), then a strategy of finding θ such that $Q(\theta, \theta') \geq Q(\theta', \theta')$ ensures that $L(\theta) \geq L(\theta')$. This leads to the EM algorithm:

1. **E Step**

$$\text{Calculate } Q(\theta, \widehat{\theta}_n).$$

2. **M Step**
   Maximise:

$$\widehat{\theta}_{n+1} = \arg\max_\theta Q(\theta, \widehat{\theta}_n).$$

As an alternative perspective on the EM algorithm, note that it is possible to think of X as the “incomplete” data at hand, and $Y = (X, Z)$ as the “complete” data set, that if available, would make the expectation problem easier. Since the complete data is not available, the best L_2 approximant, formed by taking conditional expectation with respect to the best guess at $\theta = \theta'$, is used instead:

$$\log P_\theta(Y) \approx \mathbf{E}_{\theta'}\{\log P_\theta(Y)|X = X_N\} = Q(\theta, \theta').$$

This leads to a procedure of maximising $Q(\theta, \widehat{\theta}_n)$ to get $\widehat{\theta}_{n+1}$ which leads to new conditional expectation and so on. Of course, there is only any sense in this scheme if maximising

$$Q(\theta, \theta') = \mathbf{E}_{\theta'}\left\{\log P_\theta(Y)|X = X_N\right\}$$

is easier than maximising $L(\theta)$ directly.

4 Application of the EM Algorithm for State-Space Estimation

For the purpose of applying the EM algorithm to the problem of estimating A, B, C, D, Q and R in (2) and (3), the most obvious choice for the incomplete data set X is one in which it is taken as the unobserved state sequence $\{x_1, \ldots, x_N\}$ (hence the use of the symbol X). That is,

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

Attention is then focussed on the calculation of

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

Now by repeated application of Bayes’ Rule

$$\begin{aligned} p_\theta(y_t, \ldots, y_N, x_{t-1}, \ldots, x_N \mid \theta) &= p_\theta(y_t, \ldots, y_N \mid x_{t-1}, \ldots, x_N) p_\theta(x_{t-1}, \ldots, x_N) \\ &= p_\theta(x_{t-1}) \prod_{k=t}^{N} p_\theta(x_k \mid x_{k-1}) \prod_{k=t}^{N} p_\theta(y_k \mid x_k). \end{aligned}$$

Using this and the model structure (2), (3) (and excluding terms that do not depend on quantities to be estimated),

$$\begin{aligned} -2\log p_\theta(Y_N, X_N) = \log|P_0| \quad + \quad & (x_{t-1} - \mu)^T P_0^{-1}(x_{t-1} - \mu) + N\log|Q| + N\log|R| + \\ & \sum_{k=t}^{N}(x_k - Ax_{k-1} - Bu_{k-1})^T Q^{-1}(x_k - Ax_{k-1} - Bu_{k-1}) + \\ & \sum_{k=t}^{N}(y_k - Cx_k - Du_k)^T R^{-1}(y_k - Cx_k - Du_k) \end{aligned}$$

where we have assumed initial distribution on x_{t-1} of

$$x_{t-1} \sim \mathcal{N}(\mu, P_0).$$

In this case, the definition (5) of $Q(\theta, \theta')$ leads to,

$$\begin{aligned}
-2Q(\theta,\theta') &= \log|P_0| + N\log|Q| + N\log|R| + \\
&\quad \mathrm{Tr}\left\{P_0^{-1}\mathbf{E}_{\theta'}\{(x_{t-1}-\mu)(x_{t-1}-\mu)^T \mid Y_N\}\right\} + \\
&\quad \sum_{k=t}^{N}\mathrm{Tr}\left\{Q^{-1}\mathbf{E}_{\theta'}\{(x_k - Ax_{k-1} - Bu_{k-1})(x_k - Ax_{k-1} - Bu_{k-1})^T \mid Y_N\}\right\} + \\
&\quad \sum_{k=t}^{N}\mathrm{Tr}\left\{R^{-1}\mathbf{E}_{\theta'}\{(y_k - Cx_k - Du_k)(y_k - Cx_k - Du_k)^T \mid Y_N\}\right\}.
\end{aligned} \tag{6}$$

Now introduce the notation

$$\widehat{x}_{k|N} \triangleq \mathbf{E}_{\theta'}\{x_k \mid Y_N\}, \qquad z_t \triangleq \begin{bmatrix} x_t \\ u_t \end{bmatrix}, \tag{7}$$

$$\Lambda \triangleq \sum_{k=t}^{N} y_k \mathbf{E}_{\theta'}\{z_k^T \mid Y_N\}, \qquad \Omega \triangleq \sum_{k=t}^{N} y_k y_k^T, \qquad \Pi \triangleq \sum_{k=t}^{N}\mathbf{E}_{\theta'}\{x_k x_k^T \mid Y_N\}, \tag{8}$$

$$\Phi \triangleq \sum_{k=t}^{N}\mathbf{E}_{\theta'}\{z_k z_k^T \mid Y_N\}, \qquad \Psi \triangleq \sum_{k=t}^{N}\mathbf{E}_{\theta'}\{x_k z_{k-1}^T \mid Y_N\}, \qquad \Gamma \triangleq \sum_{k=t}^{N}\mathbf{E}_{\theta'}\{z_{k-1} z_{k-1}^T \mid Y_N\}. \tag{9}$$

Then (6) may be more compactly expressed as

$$\begin{aligned}
-2Q(\theta,\theta') &= \log|P_0| + N\log|Q| + N\log|R| + \\
&\quad \mathrm{Tr}\left\{P_0^{-1}\left[(\widehat{x}_{t-1|N}-\mu)(\widehat{x}_{t-1|N}-\mu)^T + P_0\right]\right\} + \\
&\quad \mathrm{Tr}\left\{Q^{-1}\left[\Phi - \Psi[A,B]^T - [A,B]\Psi^T + [A,B]\Gamma[A,B]^T\right]\right\} + \\
&\quad \mathrm{Tr}\left\{R^{-1}\left[\Omega - \Lambda[C,D]^T - [C,D]\Lambda^T + [C,D]\Pi[C,D]^T\right]\right\}
\end{aligned} \tag{10}$$

Therefore, since

$$\begin{aligned}
\Phi - \Psi[A,B]^T - [A,B]\Psi^T + [A,B]\Gamma[A,B]^T &= ([A,B] - \Psi\Gamma^{-1})\Gamma([A,B]-\Psi\Gamma^{-1})^T + \\
&\quad \Phi - \Psi\Gamma^{-1}\Psi^T
\end{aligned}$$

then the second last term in (10) in combination with the $N\log|Q|$ term is clearly minimised by the choices

$$[A,B] = \Psi\Gamma^{-1} \qquad \text{and} \qquad Q = N^{-1}(\Phi - \Psi\Gamma^{-1}\Psi^T). \tag{11}$$

The latter equation for Q follows by application of Lemma A.1 and the chain rule to compute

$$\frac{\mathrm{d}}{\mathrm{d}Q} N\log|Q| + \mathrm{Tr}\{Q^{-1}(\Phi - \Psi\Gamma^{-1}\Psi^T)\} = NQ^{-1} - Q^{-2}(\Phi - \Psi\Gamma^{-1}\Psi^T)$$

which is clearly zero for the choice of Q in (11). Note that by construction, this formulation (11) for Q is positive semi-definite since it is a Schur complement of

$$\sum_{k=t}^{N} \mathbf{E}_{\theta'} \left\{ \left[\begin{array}{c} z_k \\ z_{k-1} \end{array} \right] \left[\begin{array}{cc} z_k^T & z_{k-1}^T \end{array} \right] \right\} \geq 0.$$

In a similar fashion, the last term in (10) in combination with the $N \log |R|$ term is minimised by the choices

$$[C, D] = \Lambda\Phi^{-1} \qquad \text{and} \qquad R = N^{-1}(\Omega - \Lambda\Pi^{-1}\Lambda^T). \tag{12}$$

Finally, via the same arguments, the terms involving P_0 imply the following choices for the maximisation of $Q(\theta|\theta')$:

$$\mu = \mathbf{E}_{\theta'}\{x_{t-1} \mid Y_N\}, \qquad P_0 = \mathbf{E}_{\theta'}\{x_{t-1}x_{t-1}^T \mid Y_N\}.$$

5 Calculation of Kalman–Smoothed Quantities

Clearly, to implement this algorithm, it is necessary to be able to compute the quantities

$$\widehat{x}_{t|N} = \mathbf{E}\left\{x_t \mid Y_N\right\}, \qquad \mathbf{E}\left\{x_t x_t^T \mid Y_N\right\}, \qquad \mathbf{E}\left\{x_t x_{t-1}^T \mid Y_N\right\}$$

In the case considered in this paper where the distributions on the random components e_t and w_t are Gaussian, then the Kalman Smoother recursions may be used for this purpose. Specifically, with the further definition

$$P_{t|s} = \mathbf{E}\left\{(\widehat{x}_{t|s} - x_t)(\widehat{x}_{t|s} - x_t)^T |Y_s\right\}$$

then it is first necessary to run the measurement and time update Kalman Filter recursions [6]

$$\begin{aligned}
P_{t|t-1} &= AP_{t-1|t-1}A^T + Q && (13)\\
K_t &= P_{t|t-1}C^T\left(CP_{t|t-1}C^T + R\right)^{-1} && (14)\\
P_{t|t} &= P_{t|t-1} - K_tCP_{t|t-1} &&\\
&= P_{t|t-1} - P_{t|t-1}C^T\left(CP_{t|t-1}C^T + R\right)^{-1}CP_{t|t-1} && (15)\\
\widehat{x}_{t|t-1} &= A\widehat{x}_{t-1|t-1} + Bu_{t-1} && (16)\\
\widehat{x}_{t|t} &= \widehat{x}_{t|t-1} + K_t\left(y_t - Du_t - C\widehat{x}_{t|t-1}\right) && (17)
\end{aligned}$$

which are initialised at

$$\widehat{x}_{t-1|t-1} = \mu, \qquad P_{t-1|t-1} = P_0.$$

Once these computations are made, then the so-called Rauch–Tung–Striebel [6] reverse time recursions are performed to calculate the smoothed estimates which are expectations conditional on the whole date set Y_N:

$$\begin{aligned}
\widehat{x}_{t|N} &= \widehat{x}_{t|t} + S_t\left[\widehat{x}_{t+1|N} - Bu_t - A\widehat{x}_{t|t}\right] && (18)\\
P_{t|N} &= P_{t|t} + S_t\left[P_{t+1|N} - P_{t+1|t}\right]S_t^T, && (19)
\end{aligned}$$

where

$$S_t \triangleq P_{t|t} A^T P_{t+1|t}^{-1}.$$

These latter recursions are initialised with the terminal values $\widehat{x}_{N|N}$ and $P_{N|N}$ of the ‘forward–pass’ Kalman Filter recursions. All these calculations provide sufficient material for the computation

$$\mathsf{E}\left\{x_t x_t^T \mid Y_N\right\} = P_{t|N} + \widehat{x}_{t|N}\widehat{x}_{t|N}^T.$$

However, the quantity $\mathsf{E}\{x_t x_{t-1}^T \mid Y_N\}$ is also required, and for this purpose define the quantity

$$M_{t|s} = \mathsf{E}\left\{(\widehat{x}_{t|s} - x_t)(\widehat{x}_{t-1|s} - x_{t-1})^T|Y_s\right\}.$$

In this case, with the initialisation

$$M_{N|N-1} = (I - K_N C)AP_{N-1|N-1} \tag{20}$$

then the following reverse time recursion

$$M_{t|t-1} = P_{t|t}S_{t-1}^T + S_t(M_{t+1|t} - AP_{t|t})S_{t-1}^T \tag{21}$$

may be used to permit the computation

$$\mathsf{E}\left\{x_t x_{t-1}^T \mid Y_N\right\} = M_{t|N} + \widehat{x}_{t|N}\widehat{x}_{t-1|N}^T.$$

6 Estimation Algorithm

The previous developments may now be summarised in the following estimation algorithm definition.

1. Initialise estimates at $\theta_k = [A, B, C, D, Q, R]$. For example, a subspace-based estimation method could be employed.
2. Run Kalman-Filter recursions (13)-(17) followed by the Kalman Smoother recursions (18), (19), (20) (21) in order to compute the quantities defined in (7), (8) (9).
3. Maximise $Q(\theta, \theta_k)$ over θ via the choices (11) and (12) in order to provide an improved estimate θ_{k+1}.
4. Repeat until convergence.

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 (2), (3) 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.

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.

7 Simulation Examples

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 (1) with (a sampling period of 1 second used in the following zero-order-hold computation)

$$\begin{aligned} G(q) &= \mathcal{ZOH}\left\{\begin{bmatrix} \dfrac{1}{(s+1)(s+0.1)} & \dfrac{1}{(s+2)(s+0.5)} \\ \dfrac{1}{(s+0.7)(s+0.3)} & \dfrac{1}{(s+0.8)(s+0.4)} \end{bmatrix}\right\} \\ &= \begin{bmatrix} \dfrac{0.0355q+0.02465}{(q-0.3679)(q-0.9084)} & \dfrac{0.2364q+0.1038}{(q-0.1353)(q-0.6065)} \\ \dfrac{0.07601q+0.05447}{(q-0.4966)(q-0.7408)} & \dfrac{0.1087q+0.07286}{(q-0.4493)(q-0.6703)} \end{bmatrix} \end{aligned}$$

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

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

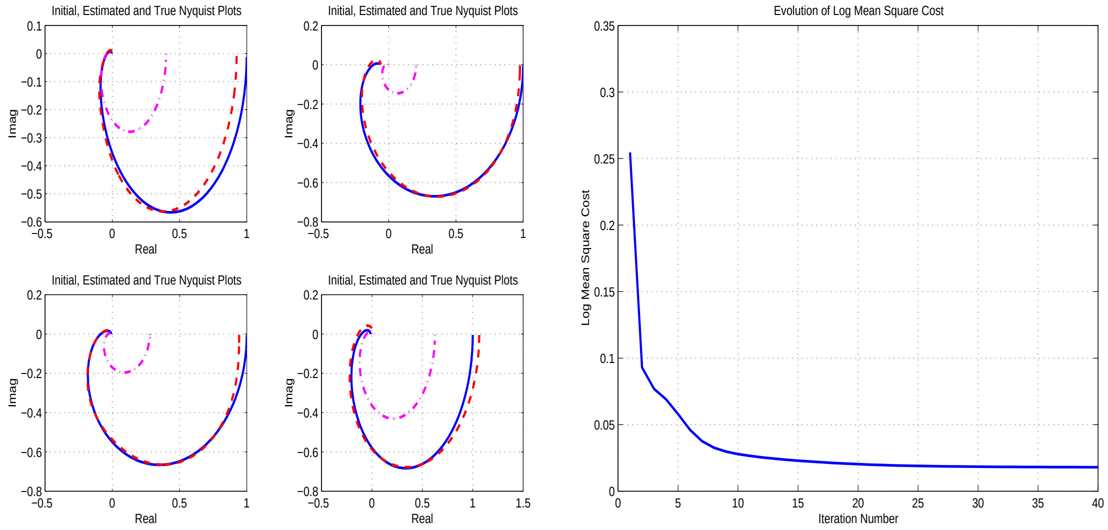
The results of this estimation experiment are shown in figure 1. On the left, the relationship between initial and EM-derived ML estimates is shown together with the true response. On the right, the evolution of the log mean-square cost

$$\frac{1}{N}\sum_{t=1}^{N}\left(y_t - C\widehat{x}_{t|t-1}\right)^2$$

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 2. There, Overschee and DeMoor's N4SID variant [17, 16] 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 2 and final estimate shown as the dashed line on the left in 2, 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 evolution

Figure 1: *Results when using the EM-algorithm method for computing Maximum-Likelihood estimates. The left hand figure shows initial estimates (via Nyquist plots) as dash-dot line, true systems as solid lines, and EM-derived ML estimates as dashed lines. The right hand figure shows the evolution of the means square cost as the EM-algorithm is iterated.*

8 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.

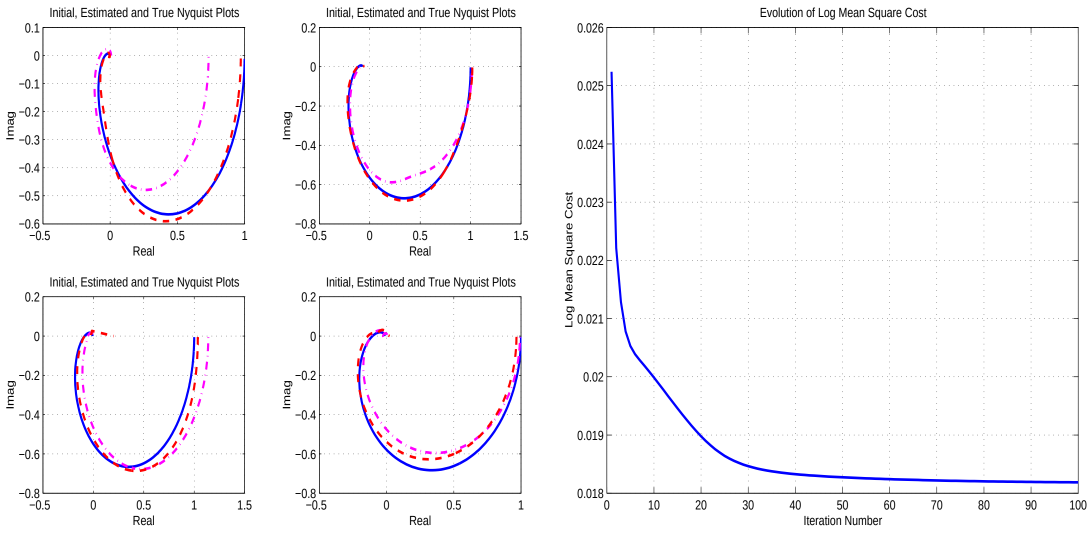
Although this method is novel in the context considered here, the EM-algorithm itself is quite old, being very well known in (for example) the speech-recognition community as the Baum–Welch method for Hidden Markov Model estimation [14].

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

A Technical Lemmata

Lemma A.1. *Suppose* $M, N \in \mathbf{R}^{n\times n}$ *and* M *is invertible. Then*

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



(a) Nyquist Plots

(b) LS cost evolution

Figure 2: *Results when using the EM-algorithm method for computing Maximum-Likelihood estimates. The left hand figure shows initial estimates (via Nyquist plots) as dash-dot line, true systems as solid lines, and EM-derived ML estimates as dashed lines. The right hand figure shows the evolution of the means square cost as the EM-algorithm is iterated.*

References

[1] P. CAINES, *Linear Stochastic Systems*, John Wiley and Sons, New York, 1988.

[2] M. DEISTLER, *Model Identification and Adaptive Control*, Springer-Verlag, 2000, ch. System Identification - General Aspects and Structure.

[3] A. DEMPSTER, N. LAIRD, AND D. RUBIN, *Maximum likelihood from incomplete data via the em algorithm*, Journal of the Royal Statistical Society, Series B, 39 (1977), pp. 1–38.

[4] R. GUIDORZI, *Canonical structures in the identification of multivariable systems*, Automatica— J. IFAC, 11 (1975), pp. 361–374.

[5] E. HANNAN AND M. DEISTLER, *The Statistical Theory of Linear Systems*, John Wiley and Sons, New York, 1988.

[6] A. JAZWINSKI, *Stochastic Processes and Optimal Filtering Theory*, Academic Press, 1970.

[7] W. LARIMORE, *Canonical variate analysis in identification, filtering and adaptive control*, in Proceedings of the 29th IEEE Conference on Decision and Control, Hawaii, 1990, pp. 596–604.

[8] E. LEHMANN, *Theory of Point Estimation*, John Wiley & Sons, 1983.

[9] L. LJUNG, *System Identification: Theory for the User*, Prentice-Hall, Inc., New Jersey, 1987.

[10] ——, *MATLAB System Identification Toolbox Users Guide, Version 5*, The Mathworks, 2000.

[11] D. G. LUENBERGER, *Canonical forms for linear multivariable systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 290–293.

[12] T. MCKELVEY, *Discussion: 'on the use of minimal parametrizations in multivariable armax identification' by r.p. guidorzi*, European Journal of Control, 4 (1998), pp. 93–98.

[13] T. MCKELVEY AND A. HELMERSSON, *A dynamical minimal parametrization of multivariable linear systems and its application to optimization and system identification*, in Proc. of the 14th World Congress of IFAC, H. Chen and B. Wahlberg, eds., vol. H, Beijing, P. R. China, July 1999, IFAC, Elsevier Science, pp. 7–12.

[14] L. RABINER, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of the IEEE, 77 (1989), pp. 257–285.

[15] T.SÖDERSTRÖM AND P.STOICA, *System Identification*, Prentice Hall, New York, 1989.

[16] P. VAN OVERSCHEE AND B. DE MOOR, *N4SID:Subspace algortihms for the identification of combined deterministic-stochastic systems*, Automatica, 30 (1994), pp. 75–93.

[17] P. VAN OVERSCHEE AND B. D. MOOR, *Subspace Identification for Linear Systems*, Kluwer Academic Publishers, 1996.

[18] M. VERHAEGEN, *Identification of the deterministic part of mimo state space models in innovations form from input-output data*, Automatica, 30 (1994), pp. 61–74.