On The relationship between State-Space-Subspace-Based and Maximum-Likelihood System Identification methods.

Brett Ninness* Stuart Gibson†

Abstract

State-Space Subspace Identification methods obtain system estimates in closed form, and this is in contrast to Maximum Likelihood methods which, although provably consistent and statistically efficient, require an iterative approach to solve an optimisation problem (which is possibly non-convex) over the likelihood surface. Particularly in signal processing and pattern recognition, the so-called Expectation Maximisation (EM) method is a popular way of performing these latter iterations. This paper establishes that a subspace identification method can, in fact, be viewed as one iteration of the EM algorithm. As such, a link between subspace and Maximum Likelihood methods is established.

Technical Report EE200021, Department of Electrical and Computer Engineering, University of Newcastle,AUSTRALIA

1 Introduction

The study of State Space Subspace System Identification (4SID) [13, 11, 14, 15, 7] has been of enormous recent interest in the area of estimation methods targeted at control and signal processing applications.

Despite the recent nature of this activity, the ideas involved actually go back many years, at least to Akaike [1] whose approach was targeted at types of stochastic estimation problems pertinent to this paper.

For the purposes of explaining this (see[13, 11] for an explanation of the same ideas, but from a geometric rather than an innovations point of view), suppose one is presented with observations $\{y_t\}$ of a stationary stochastic process and is faced with the task of estimating a state-space representation of this process in *innovations* form:

$$x_{t+1} = Ax_t + Be_t, \quad (1)$$

$$y_t = Cx_t + De_t \quad (2)$$

where $\{e_t\}$ is an i.i.d. zero mean and unit variance white noise process. Exploiting the idea of a 'predictor space', Akaike [1] made clear for the first time that such a representation always exists,

*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

and in doing so suggested a way that it may be estimated from observations of $\{y_t\}$. This method (with modifications involving user chosen weighting matrices) is now known as State-Space Subspace System Identification.

The great advantage of the approach lies in the numerical simplicity and reliability of its implementation. The key operation required is one of projection which may be performed with Singular Value Decomposition or even QR factorisation.

Unfortunately, a feature of these advantages is that the operations providing them are also non-linear in the data and this renders difficult any analysis of the statistical performance of the approach.

This is in contrast to the well established maximum likelihood (ML) method for estimation where the consistency, distributional and statistical efficiency properties are well-known for a range of experimental scenarios [9, 6, 8]. Again, these benefits come at a price, this time in the form of the difficulty of actually calculating the estimate. Normally, some sort of iterative procedure is required to numerically solve the non-convex optimisation problem that is involved.

Although not well recognised in the control-theory literature, the Expectation Maximisation (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.

The contribution of this paper is to establish that subspace identification methods are closely related to a single iteration of the EM algorithm.

The precise details of the equivalence depend upon the weighings used in the subspace methods as well as the initial parameter guesses made in the EM algorithm implementation. However, a unifying principle is that both methods involve projections onto estimates of the state space. It is then only how the state estimates are obtained that discriminates between the EM and subspace methods.

This discovery therefore also establishes a link between the Subspace and Maximum Likelihood methods, and hence provides further credence to observations that in many cases Subspace methods are almost as efficient as Maximum Likelihood techniques [2].

2 An Overview of Subspace Identification Methods

This overview is drawn from that in [2, 3], and hence the notation (with minor simplifications) is also copied. In essence, what follows here is also a restatement of the fundamental ideas in [1].

Suppose that the system of (1) and (2) involves matrices (A, B, C, D) of dimensions $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $C \in \mathbf{R}^{m\times n}$, $D \in \mathbf{R}^{m\times m}$ and that a distinction is made between the past Y_t^- and future Y_t^+ of the process $\{y_t\}$ as follows

$$\begin{aligned} Y_t^- &\triangleq [y_{t-1}^T, y_{t-2}^T, \dots]^T, \\ Y_t^+ &\triangleq [y_t^T, y_{t+1}^T, \dots]^T, \\ E_t^+ &\triangleq [e_t^T, e_{t+1}^T, \dots]^T. \end{aligned}$$

In this case, (1) and (2) yield

$$Y_t^+ = \mathcal{O}x_t + \mathcal{E}E_t^+,$$

where $\mathcal{O}$ and $\mathcal{E}$ are infinite dimensional matrices defined as

$$\mathcal{O} \triangleq \begin{bmatrix} C \\ CA \\ CA^2 \\ CA^3 \\ \vdots \end{bmatrix}, \qquad \mathcal{E} \triangleq \begin{bmatrix} D & 0 & 0 & \cdots \\ CB & D & 0 & \cdots \\ CAB & CB & D & \\ CA^2B & \ddots & \ddots & \ddots \\ \vdots & \ddots & & \end{bmatrix}.$$

Now, (as will hold throughout the paper) it will be assumed that (1), (2) represents a strictly stable ($|\lambda_{\max}(A)| < 1$) and minimum phase ($|\lambda_{\max}(A - BD^{-1}C)| < 1$) system so that with

$$\mathcal{K} \triangleq [BD^{-1}, (A - BD^{-1}C)BD^{-1}, (A - BD^{-1}C)^2BD^{-1}, \dots]$$

then $x_t = \mathcal{K}Y_t^-$ and hence

$$Y_t^+ = \mathcal{O}\mathcal{K}Y_t^- + \mathcal{E}E_t^+. \tag{3}$$

Now, the key contribution of [1] was to consider the space spanned by the state $x_t = \mathcal{K}Y_t^-$ as being a 'predictor space' and to define this with projections that, in turn, depend on defining an inner product $\langle \cdot, \cdot \rangle$ between random variables X and Y as $\langle X, Y \rangle = \mathbf{E}\{XY\}$.

More specifically, the predictor space represents the interface between the future and the past in that it is the projection of (the space spanned by) Y_t^+ onto (the space spanned by) Y_t^- which leaves as a residual an error that is orthogonal to Y_t^-. Therefore, since this projection lives in Y_t^- and hence may be written as PY_t^- for some linear operator P, then

$$\langle Y_t^+ - PY_t^-, PY_t^- \rangle = 0. \tag{4}$$

Let the dimension of the space spanned by Y_t^+ be denoted by n. Then the projection P may be factored into 2 rank n operators $\widehat{\mathcal{O}}$, $\widehat{\mathcal{K}}$ as $P = \widehat{\mathcal{O}}\widehat{\mathcal{K}}$, and then $x_t = \widehat{\mathcal{K}}Y_t^-$ is an element in the predictor space (or state space) with respect to a basis defined by $\widehat{\mathcal{K}}$.

Now, equation (4) allows elements of Y_t^+ to be formed as the sum of $\widehat{C}\widehat{x}_t$ (ie. linear combinations of $\widehat{x}_t$) and an error term $\widehat{D}e_t$ that is orthogonal to Y_t^-. That is

$$y_t = \widehat{C}\widehat{x}_t + \widehat{D}e_t; \qquad \mathbf{E}\left\{y_t e_t^T\right\} = 0$$

and hence

$$\widehat{C} = \mathbf{E}\left\{y_t \widehat{x}_t^T\right\} \mathbf{E}\left\{\widehat{x}_t \widehat{x}_t^T\right\}^{-1}. \tag{5}$$

Increasing $t \mapsto t+1$ and repeating the argument would lead to a conclusion that any element $\widehat{x}_{t+1}$ of the predictor space spanned by the orthogonal projections of Y_{t+1}^+ onto Y_{t+1}^- must be expressible as a linear transformation of $\widehat{x}_t$ (spanning the orthogonal projection of Y_t^+ on Y_t^-) by $\widehat{A}$ plus the linear transformation of e_t (itself orthogonal to the space spanned by Y_{t+1}^-) by $\widehat{B}$. That is,

$$\widehat{x}_{t+1} = \widehat{A}\widehat{x}_t + \widehat{B}e_t,$$

and hence,

$$\widehat{A} = \mathbf{E}\left\{\widehat{x}_t \widehat{x}_{t+1}^T\right\} \mathbf{E}\left\{\widehat{x}_t \widehat{x}_t^T\right\}^{-1} \tag{6}$$

and

$$\widehat{B} = \mathbf{E}\left\{\widehat{x}_{t+1}\widehat{e}_t^T\right\}\mathbf{E}\left\{\widehat{e}_t\widehat{e}_t^T\right\}^{-1}. \tag{7}$$

This argument by Akaike [1] therefore establishes in a fundamental manner the suitability of a Markovian (or state-space) representation such as expressions (1) and (2) for a stochastic process.

However, it goes beyond this by also suggesting (via relations (5), (6) and (7) that arose from orthogonality considerations) a means for estimating A, B, C from observed data, provided that $\widehat{x}_t$ can also be estimated from this data.

For this purpose, it is necessary to be explicit about how finite data records will be accommodated. For this purpose define finite data amendments to Y_t^-, Y_t^+ as

$$Y_{t,p}^- \triangleq \begin{pmatrix} y_{t-1} & y_t & \cdots & \cdots & \cdots & \cdots & y_{N-1} \\ y_{t-2} & y_{t-1} & \cdots & \cdots & \cdots & \cdots & y_{N-2} \\ y_{t-3} & y_{t-2} & \ddots & & & & \vdots \\ \vdots & \vdots & & \ddots & & & \vdots \\ y_{t-p} & y_{t+1-p} & \cdots & \cdots & y_{t-1} & \cdots & y_{N-p} \end{pmatrix} \tag{8}$$

$$Y_{t,f}^+ \triangleq \begin{pmatrix} y_t & y_{t+1} & \cdots & \cdots & \cdots & \cdots & y_{N-f} \\ y_{t+1} & y_{t+2} & & & & & y_{N-f+1} \\ y_{t+2} & & & & & & \vdots \\ \vdots & & & & & & \vdots \\ y_{t+f} & y_{t+1+f} & \cdots & \cdots & \cdots & \cdots & y_N \end{pmatrix} \tag{9}$$

so that (3) becomes

$$Y_{t,f}^+ = \mathcal{O}_f\mathcal{K}_pY_{t,p}^- + \mathcal{O}_f(A - BD^{-1}C)^p\mathcal{K}Y_{t,p}^- + \mathcal{E}_fE_t^+ \tag{10}$$

where

$$\begin{aligned} \mathcal{O}_f &\triangleq \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{f-1} \end{bmatrix} \\ \mathcal{K}_p &\triangleq \left[BD^{-1}(A - BD^{-1}C)BD^{-1}\ldots(A - BD^{-1}C)^{p-1}BD^{-1}\right]. \end{aligned} \tag{11}$$

Assuming then that p is large enough that $(A - BD^{-1}C)^p$ is negligible, the subspace identification methods proceed through the following steps.

1. Derive an estimate $\widehat{\beta}$ of $\mathcal{O}_f\mathcal{K}_p$ by regressing $Y_{t,f}^+$ on $Y_{t,p}^-$:

$$\widehat{\beta} = R_{fp}(R_{pp})^{-1} \tag{12}$$

where

$$R_{fp} \triangleq \frac{1}{N}Y_{t,f}^+(Y_{t,p}^-)^T \qquad \text{and} \qquad R_{pp} \triangleq \frac{1}{N}Y_{t,p}^-(Y_{t,p}^-)^T. \tag{13}$$

2. Decompose $\widehat{\beta}$ into the product $\widehat{\beta} = \widehat{\mathcal{O}}_f \widehat{\mathcal{K}}_p$ of two rank n matrices $\widehat{\mathcal{O}}_f$, $\widehat{\mathcal{K}}_p$ as follows.

   (a) Choose weighting matrices W_f, W_p and perform the SVD factorisation

$$W_f \widehat{\beta} W_p = USV^T = [U_n \mid u] \begin{bmatrix} S_n & 0 \\ 0 & s \end{bmatrix} \begin{bmatrix} V_n^T \\ v^T \end{bmatrix} \tag{14}$$

   where $S \in \mathbf{R}^{n \times n}$ and U_n, u, V_n, v are chosen conformally.

   (b) Take as estimates:

$$\widehat{\mathcal{O}}_f = W_f^{-1} U_n S_n^{\frac{1}{2}} \qquad \widehat{\mathcal{K}}_p = S_n^{\frac{1}{2}} V_n^T W_p^{-1}. \tag{15}$$

3. Use the state estimate

$$\widehat{X}_{t,p}^- = [\widehat{x}_t, \widehat{x}_{t+1} \ldots \widehat{x}_N] = \widehat{\mathcal{K}}_p Y_{t,p}^- \tag{16}$$

   to estimate C using (7) as the least squares solution to

$$[y_t^T, y_{t+1}^T, \ldots, y_N^T] = C \widehat{X}_{t,p}^-. \tag{17}$$

   Use this to estimate the residuals $\{e_t\}$ as

$$\widehat{De_t} = [y_t^T, y_{t+1}^T, \ldots, y_N^T] - \widehat{C} \widehat{X}_{t,p}^- \tag{18}$$

   and then use (5) and (6) to estimate A and BD^{-1} by finding the least squares solution to

$$\widehat{X}_{t+1,p}^- = [A,\ BD^{-1}] \begin{bmatrix} \widehat{X}_{t,p}^- \\ \\ \widehat{De_t} \end{bmatrix} \tag{19}$$

4. Estimate D as the lower triangular Cholesky factor of the sample covariance of $\widehat{De_t}$.

The fact that steps 1-4 are features of all the various documented subspace estimation methods means that they can all be viewed as being versions of a single basic estimation technique [12]. However, these same estimation schemes do vary according to the exact method by which the weighting matrices in step 2 are chosen.

For example, Akaike's original work [1] used the following weightings.

$$W_f = \left(\frac{1}{N} Y_{t,f}^+ (Y_{t,f}^+)^T \right)^{-1/2} \qquad \text{and} \qquad W_p = \left(\frac{1}{N} Y_{t,p}^- (Y_{t,p}^-)^T \right)^{1/2}. \tag{20}$$

These were chosen so as to provide

$$\frac{1}{N} \widehat{X}_{t,p}^- (\widehat{X}_{t,p}^-)^T = I_n. \tag{21}$$

The resulting algorithm, now known as Canonical Correlation Analysis (CCA), has been pursued in engineering System Identification scenarios by Larimore [7].

Another popular choice of weightings is a variant of the so-called N-4SID approach of Van Overschee and De Moor wherein the following choices are made.

$$W_f = I_f, \qquad \text{and} \qquad W_p = \left(\frac{1}{N} Y_{t,p}^- (Y_{t,p}^-)^T \right)^{1/2}. \tag{22}$$

3 Maximum Likelihood Estimation

An older and more well-known approach for tackling the estimation of systems described by (1), (2) is the so-called 'Maximum Likelihood' method wherein the probability density function $p_e(\cdot)$ for the random variable e_t is 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, D in a specific fashion 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

$$\widehat{\theta}_N \triangleq \arg\max_{\theta} \; p(y_1, \ldots, y_N \mid \theta). \tag{23}$$

This method of estimation enjoys a wide acceptance and popularity, and this is in large part due to the following well-known and desirable properties of the scheme which have been established in [6, 8].

Consistency

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

Asymptotic Normality

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

where

$$P_N = -\frac{\mathrm{d}^2}{\mathrm{d}\theta \mathrm{d}\theta^T}\bigg|_{\theta=\theta_\circ} \mathbf{E}\left\{\log p(y_1, \ldots, y_N \mid \theta)\right\}.$$

Asymptotic efficiency

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

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

and hence it is asymptotically efficient.

Balancing these features that recommend a Maximum Likelihood approach is the significant disadvantage that (23) defining the MLE is, in general, a non-convex optimisation problem. As a result, calculation of the MLE requires some sort of numerical search technique, and hence the problem of $\widehat{\theta}_N$ mistakenly being a local rather than a global minimum (which is the true MLE) is impossible to avoid absolutely.

Since $p(y_1, \ldots, y_N \mid \theta)$ is typically smooth, any gradient-based search technique such as Steepest Descent or Newton iteration may be employed in the calculation of $\widehat{\theta}_N$. However, this involves the computational burden of calculating the gradient (and possibly also the Hessian) of the likelihood function.

As a means of avoiding this overhead, whilst still using an iterative search method, the so-called Expectation Maximisation (EM) method is a useful alternative.

4 The EM Algorithm for Likelihood Maximisation

This method arose in the mathematical statistics community [4] but has found wide engineering application in areas such as signal processing and pattern recognition.

The key feature of the method 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 $\{y_1, \ldots, y_N\}$) will be abbreviated to an uppercase letter (say Y) and conditional dependence on θ will be noted by subscripting; for example, $p(y_1, \ldots, y_N \mid \theta) \equiv p_\theta(Y)$.

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

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

and hence

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

which implies that

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

Therefore, 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

$$\begin{aligned} L(\theta) \triangleq \log p_\theta(Y_N) &= \mathbf{E}_{\theta'} \{\log p_\theta(Y) \mid Y = Y_N\} \\ &= \mathbf{E}_{\theta'} \{\log p_\theta(Z) \mid Y = Y_N\} - \mathbf{E}_{\theta'} \{\log p_\theta(Z \mid Y) \mid Y = Y_N\} \\ &= Q(\theta, \theta') - V(\theta, \theta'). \end{aligned} \tag{24}$$

However, 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**

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

2. **M Step**

   Maximize:

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

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 maximizing $Q(\theta, \widehat{\theta}_n)$ 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 maximizing

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

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

5 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 and D in (1) and (2), the most obvious choice for the incomplete data set X is the one made by Shumway [10] in which it is taken as the unobserved state sequence $\{x_1, \ldots, x_N\}$ (hence the use of the symbol X). ie.

$$Z = (X_N, Y_N),$$

and then attention is 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

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

For the purpose of actually formulating what this density is, it is necessary to specify the density of the innovations e_t in (1) and (2). Here it will be taken to be Gaussian, $e_t \sim \mathcal{N}(0, I_m)$. A difficulty now arises because $B \in \mathbf{R}^{n \times m}$ will typically have $n > m$ and thus Be_t will have a singular Gaussian distribution for which no closed form expression exists.

To circumvent this, consider a system related to (1) and (2), and given by:

$$x_{t+1} = Ax_t + Be_t + w_t, \tag{25}$$

$$y_t = Cx_t + De_t \tag{26}$$

where $\{w_t\}$ is an n-dimensional vector i.i.d. process that is independent of $\{e_t\}$ and distributed as $w_t \sim \mathcal{N}(0, \epsilon I)$ where $\epsilon > 0$ is arbitrarily small, and hence (in some sense) the system given by (25) and (26) is arbitrarily close to that of (1) and (2).

Using this new system (and excluding terms that do not depend on A, B, C and D),

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

where

$$R \triangleq DD^T, \qquad Q \triangleq BB^T + \epsilon I, x_{t-1} \sim \mathcal{N}(\mu, \Sigma).$$

In this case, the definition (24) provides,

$$\begin{aligned}-2Q(\theta,\theta') &= \log|\Sigma| + N\log|Q| + N\log|R| + \\ &\quad \mathbf{E}_{\theta'}\{(x_{t-1}-\mu)^T\Sigma^{-1}(x_{t-1}-\mu) \mid Y_N\} + \\ &\quad \sum_{k=t}^{N}\mathbf{E}_{\theta'}\{(x_k - Ax_{k-1})^T Q^{-1}(x_k - Ax_{k-1}) \mid Y_N\} + \\ &\quad \sum_{k=t}^{N}\mathbf{E}_{\theta'}\{(y_k - Cx_k)^T R^{-1}(y_k - Cx_k) \mid Y_N\} \\ &= \log|\Sigma| + N\log|Q| + N\log|R| + \\ &\quad \mathrm{Tr}\left\{\Sigma^{-1}\mathbf{E}_{\theta'}\{(x_0-\mu)(x_0-\mu)^T \mid Y_N\}\right\} + \\ &\quad \sum_{k=t}^{N}\mathrm{Tr}\left\{Q^{-1}\mathbf{E}_{\theta'}\{(x_k - Ax_{k-1})(x_k - Ax_{k-1})^T \mid Y_N\}\right\} + \\ &\quad \sum_{k=t}^{N}\mathrm{Tr}\left\{R^{-1}\mathbf{E}_{\theta'}\{(y_k - Cx_k)(y_k - Cx_k)^T \mid Y_N\}\right\}\end{aligned} \tag{27}$$

Now introduce the notation

$$\widehat{x}_k^N \triangleq \mathbf{E}_{\theta'}\{x_k \mid Y_N\}, \qquad P_k^N \triangleq \mathrm{Cov}_{\theta'}\{x_k x_k^T \mid Y_N\} = \mathbf{E}_{\theta'}\{x_k x_k^T \mid Y_N\} - \widehat{x}_k(\widehat{x}_k)^T,$$

$$P_{k,k-1}^N \triangleq \mathrm{Cov}_{\theta'}\{x_k x_{k-1}^T \mid Y_N\} = \mathbf{E}_{\theta'}\{x_k x_{k-1}^T \mid Y_N\} - \widehat{x}_k(\widehat{x}_{k-1})^T.$$

so that (27) may be more compactly expressed as

$$\begin{aligned}-2Q(\theta,\theta') &= \log|\Sigma| + N\log|Q| + N\log|R| + \\ &\quad \mathrm{Tr}\left\{\Sigma^{-1}\left[(\widehat{x}_{t-1}^N - \mu)(\widehat{x}_{t-1}^N - \mu)^T + P_0\right]\right\} + \\ &\quad \mathrm{Tr}\left\{Q^{-1}\left[\Phi - \Psi A^T - A\Psi^T + A\Gamma A^T\right]\right\} + \\ &\quad \mathrm{Tr}\left\{R^{-1}\left[\Omega - \Lambda C^T - C\Lambda^T + C\Phi C^T\right]\right\}\end{aligned} \tag{28}$$

where

$$\Gamma \triangleq \sum_{k=t}^{N} \left(P_{k-1}^{N} + \widehat{x}_{k-1}(\widehat{x}_{k-1})^T\right) = \sum_{k=t}^{N} \mathbf{E}_{\theta'}\{x_{k-1}x_{k-1}^T \mid Y_N\} \tag{29}$$

$$\Psi \triangleq \sum_{k=t}^{N} \left(P_{k,k-1}^{N} + \widehat{x}_{k}(\widehat{x}_{k-1})^T\right) = \sum_{k=t}^{N} \mathbf{E}_{\theta'}\{x_{k}x_{k-1}^T \mid Y_N\} \tag{30}$$

$$\Phi \triangleq \sum_{k=t}^{N} \left(P_{k}^{N} + \widehat{x}_{k}(\widehat{x}_{k})^T\right) = \sum_{k=t}^{N} \mathbf{E}_{\theta'}\{x_{k}x_{k}^T \mid Y_N\} \tag{31}$$

$$\Omega = \sum_{k=t}^{N} y_k y_k^T \tag{32}$$

$$\Lambda = \sum_{k=t}^{N} y_k(\widehat{x}_k^N)^T \tag{33}$$

Now, since

$$\Phi - \Psi A^T - A\Psi^T + A\Gamma A^T = (A - \Psi\Gamma^{-1})\Gamma(A - \Psi\Gamma^{-1})^T + \Phi - \Psi\Gamma^{-1}\Psi^T$$

then the second last term in (28) in combination with the $N \log |L|$ term is clearly minimised by the choices

$$A = \Psi\Gamma^{-1} \qquad \text{and} \qquad Q = N^{-1}(\Phi - \Psi\Gamma^{-1}\Psi^T). \tag{34}$$

Similarly, the last term in (28) in combination with the $N \log |R|$ term is clearly minimised by the choices

$$C = \Lambda\Phi^{-1} \qquad \text{and} \qquad R = N^{-1}(\Omega - \Lambda\Phi^{-1}\Lambda^T). \tag{35}$$

6 Relationship between EM and Subspace Methods

The key issue in understanding the relationship between the EM and subspace methods is the recognition of the fact that both involve a common component of regression on an estimated state sequence.

In particular, note that for the EM method, the estimate $\widehat{C}$ at each iteration is given by (32), (33) and (35) as

$$\widehat{C} = \left(\sum_{k=t}^{N} y_k \widehat{x}_k^T\right)\left(\sum_{k=t}^{N} \mathbf{E}_{\theta'}\left\{x_k x_k^T \mid Y_N\right\}\right)^{-1} \tag{36}$$

where $\widehat{x}_k$ and $\mathbf{E}_{\theta'}\left\{x_k x_k^T \mid Y_N\right\}$ may be extracted from the "Kalman Smoother" inherent in (46).

On the other hand, from (17), a subspace method finds $\widehat{C}$ *exactly as per (36)* except that instead of using (46), $\widehat{x}_k$ is estimated from the columns of $\widehat{X}_{t,p}^- = \widehat{\mathcal{K}}_p Y_{t,p}^-$ and $\sum_{k=t}^{N} \mathbf{E}_{\theta'}\left\{x_k x_k^T \mid Y_N\right\}$ is estimated as $\widehat{X}_{t,p}^-(\widehat{X}_{t,p}^-)^T$.

There are, therefore, two unifying themes

1. Both the EM and subspace methods rely on an estimate of the state sequence $\{x_t\}$. For the EM algorithm, this is found by the Kalman smoothed estimate $\widehat{x}_t = \mathbf{E}_{\theta'}\{x_t \mid Y_N\}$ while for the subspace method, it is (an approximation of) a Kalman filtered estimate $\widehat{x}_t = \mathbf{E}\{x_t \mid Y_{t-1}\}$.

2. Once $\widehat{x}_t$ is estimated, then both the EM and subspace methods estimate C by regressing y_t on $\widehat{x}_t$.

The same comments also apply for the estimation of A. For the EM method, the estimate $\widehat{A}$ of A is given by (30), (31) and (34) as

$$\widehat{A} = \left(\sum_{k=t}^{N} \mathbf{E}_{\theta'} \left\{ x_k x_{k-1}^T \mid Y_N \right\} \right) \left(\sum_{k=t}^{N} \mathbf{E}_{\theta'} \left\{ x_{k-1} x_{k-1}^T \mid Y_N \right\} \right)^{-1} \tag{37}$$

while for the subspace method (19) is used which, denoting $Y_t \triangleq [y_t, \ldots, y_{t+N}]$ and $\overline{B} \triangleq BD^{-1}$, involves the solution of

$$\widehat{X}_{t+1,p}^{-} = \left[A,\ \overline{B} \right] \begin{bmatrix} \widehat{X}_{t,p}^{-} \\ \\ Y_t \widehat{X}_{t,p}^{\perp} \end{bmatrix} \tag{38}$$

where

$$\widehat{X}_{t,p}^{\perp} = I - (\widehat{X}_{t,p}^{-})^T \left[\widehat{X}_{t,p}^{-} (\widehat{X}_{t,p}^{-})^T \right]^{-1} \widehat{X}_{t,p}^{-} \tag{39}$$

so that since $\widehat{X}_{t,p}^{\perp}$ is idempotent and an annihilator of $\widehat{X}_{t,p}^{-}$

$$\widehat{A} = \widehat{X}_{t+1,p}^{-} (\widehat{X}_{t,p}^{-})^T \left[\widehat{X}_{t,p}^{-} (\widehat{X}_{t,p}^{-})^T \right]^{-1} \tag{40}$$

and

$$\widehat{\overline{B}} = \widehat{X}_{t+1,p}^{-} \widehat{X}_{t,p}^{\perp} Y_t^T \left[Y_t \widehat{X}_{t,p}^{\perp} Y_t^T \right]^{-1}. \tag{41}$$

Comparing (37) with (40) shows that the same underlying theme emerges in that both the subspace and EM methods perform an identical regression to find $\widehat{A}$, but differ in how state estimates are derived.

To complete the picture, note that via (35) the EM method estimates $R = DD^T$ as

$$\widehat{D}\widehat{D}^T = \widehat{R} = \frac{1}{N} Y_t \left[I - \widehat{X}_t^T \left(\sum_{k=t}^{N} \mathbf{E}_{\theta'} \left\{ x_k x_k^T \mid Y_N \right\} \right)^{-1} \widehat{X}_t \right] Y_t^T, \tag{42}$$

where $\widehat{X}_t \triangleq [\widehat{x}_t, \widehat{x}_{t+1} \ldots, \widehat{x}_{t+N}]$. An estimate of D can then be obtained by Cholesky factorisation of $\widehat{R}$. At the same time, the subspace method also finds D by factoring an estimate of R, this time formed as the sample covariance of $\widehat{De_t}$. That is,

$$\begin{aligned} \widehat{D}\widehat{D}^T &= \widehat{R} = \frac{1}{N} Y_t \left[I - \left(\widehat{X}_{t,p}^{-} \right)^T \left[\widehat{X}_{t,p}^{-} \left(\widehat{X}_{t,p}^{-} \right)^T \right]^{-1} \widehat{X}_{t,p}^{-} \right] Y_t^T \\ &= \frac{1}{N} Y_t \left(\widehat{X}_{t,p}^{\perp} \right) Y_t^T. \end{aligned} \tag{43}$$

Again, a comparison of (42) and (43) shows that the subspace and EM methods are equivalent, modulo how $\{x_t\}$ itself is estimated. Finally, for the EM algorithm,

$$\widehat{B}\widehat{B}^T + \epsilon I = \frac{1}{N} \left[\sum_{k=t}^{N} \mathbf{E}_{\theta'} \left\{ x_k x_k^T \mid Y_N \right\} - \left(\sum_{k=t}^{N} \mathbf{E}_{\theta'} \left\{ x_k x_{k-1}^T \mid Y_N \right\} \right) \left(\sum_{k=t}^{N} \mathbf{E}_{\theta'} \left\{ x_{k-1} x_{k-1}^T \mid Y_N \right\} \right)^{-1} \left(\sum_{k=t}^{N} \mathbf{E}_{\theta'} \left\{ x_k x_{k-1}^T \mid Y_N \right\} \right)^T \right] \tag{44}$$

while a subspace method estimates $\widehat{\overline{B}} = \widehat{B}\widehat{D}^{-1}$ as per (41)

$$\widehat{B}\widehat{D}^{-1} = \widehat{X}^{-}_{t+1,p}\widehat{X}^{\perp}_{t,p}\left(\widehat{X}^{\perp}_{t,p}Y_t^T\right)\left[Y_t\widehat{X}^{\perp}_{t,p}Y_t^T\right]^{-1}. \tag{45}$$

However, from (42) $\widehat{D}$ is estimated as a square root of $N^{-1}Y_t\widehat{X}^{\perp}_{t,p}Y_t^T$ and hence (regardless of the precise factorisation chosen) the idempotency $X^{\perp}_{t,p}$ yields

$$\widehat{X}^{\perp}_{t,p}Y_t^T\left[Y_t\widehat{X}^{\perp}_{t,p}Y_t^T\right]^{-1} = \frac{1}{\sqrt{N}}\widehat{D}^T\left(\widehat{D}\widehat{D}^T\right)^{-1} = \frac{1}{\sqrt{N}}\widehat{D}^{-1}$$

and hence for a subspace method

$$\widehat{B}\widehat{B}^T = \frac{1}{N}\widehat{X}^{-}_{t+1,p}\widehat{X}^{\perp}_{t,p}\left(\widehat{X}^{-}_{t+1,p}\right)^T$$

which, again, is identical to (44) except for the particular method of state estimation.

7 Calculation of Kalman smoothed estimates

It remains to calculate the quantities $\mathbf{E}_{\theta'}\left\{x_{t+k}x_{t+k}^T \mid Y_N\right\}$, $\mathbf{E}_{\theta'}\left\{x_{t+k}x_{t+k-1}^T \mid Y_N\right\}$ and $\mathbf{E}_{\theta'}\left\{x_{t+k} \mid Y_N\right\}$. These may be derived by noting that from equations (25) and (26)

$$\begin{aligned} X &= \Pi x_{t-1} + \Delta_1 E + \Delta_2 W \\ Y &= (I_N \otimes C)X + (I_N \otimes D)E \end{aligned}$$

where

$$\begin{aligned} X^T &\triangleq \left[x_t^T, x_{t+1}^T, \cdots, x_{t+N-1}^T\right], \\ Y^T &\triangleq \left[y_t^T, y_{t+1}^T, \cdots, y_{t+N-1}^T\right], \\ V^T &\triangleq \left[\nu_t^T, \nu_{t+1}^T, \cdots, \nu_{t+N-1}^T\right], \\ W^T &\triangleq \left[w_t^T, w_{t+1}^T, \cdots, w_{t+N-1}^T\right], \\ E^T &\triangleq \left[e_t^T, e_{t+1}^T, \cdots, e_{t+N-1}^T\right], \end{aligned}$$

and

$$\Pi \triangleq \begin{bmatrix} A \\ A^2 \\ A^3 \\ \vdots \\ A^N \end{bmatrix}, \qquad \Delta_1 \triangleq \begin{bmatrix} B & 0 & 0 & \cdots & 0 \\ AB & B & 0 & \cdots & 0 \\ A^2B & AB & B & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ A^{N-1}B & A^{N-2} & \cdots & AB & B \end{bmatrix}, \qquad \Delta_2 \triangleq \begin{bmatrix} I \\ A \\ A^2 \\ \vdots \\ A^{N-1} \end{bmatrix},$$

so that

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} P_x & P_x\overline{C}^T \\ \overline{C}P_x & P_y \end{bmatrix}\right)$$

where

$$\begin{aligned} P_x &\triangleq \Pi\Sigma\Pi^T + \Delta_1\Delta_1^T + \epsilon\Delta_2\Delta_2^T, \\ P_y &\triangleq \begin{bmatrix} \overline{C} & \overline{D} \end{bmatrix} \begin{bmatrix} P_x & \Delta_1 \\ \Delta_1^T & I \end{bmatrix} \begin{bmatrix} \overline{C}^T \\ \overline{D}^T \end{bmatrix}, \\ \overline{C} &\triangleq I_N \otimes C, \\ \overline{D} &\triangleq I_N \otimes D \end{aligned}$$

Therefore, (see, for example, Lemma D.6.3 of [5]):

$$\begin{aligned} (X \mid Y) &\sim \mathcal{N}\left(P_x\overline{C}^T P_y^{-1} Y, P_x - P_x\overline{C}^T P_y^{-1}\overline{C}P_x\right) \\ &= \mathcal{N}\left(P_x\overline{C}^T P_y^{-1} Y, \left[P_x^{-1} + \overline{C}^T(\overline{D}\overline{D}^T)^{-1}\overline{C}\right]^{-1}\right) \end{aligned} \tag{46}$$

so that the vector $\widehat{x}_{t+k} = \mathbf{E}_{\theta'}\{x_{t+k} \mid Y_N\}$ may be taken as the k'th block element of the composite vector

$$P_x\overline{C}^T P_y^{-1} Y \tag{47}$$

while the conditional moments $\mathbf{E}_{\theta'}\left\{x_{t+k}x_{t+k}^T \mid Y_N\right\}$, $\mathbf{E}_{\theta'}\left\{x_{t+k}x_{t+k-1}^T \mid Y_N\right\}$ may be calculated from the k, k'th and $k, k-1$'th sub-blocks (respectively) of the covariance matrix

$$P_x - P_x\overline{C}^T P_y^{-1}\overline{C}P_x. \tag{48}$$

8 Conclusion

This paper has established the following link between Maximum-Likelihood and Subspace-Based estimation methods: the application of a subspace-based method can be viewed as one iteration of the EM-algorithm method for calculating the Maximum Likelihood solution.

However, as was commented on, this viewpoint is not exact in that although the above two steps derive system estimates from state estimates in an identical fashion, the state estimates employed are in fact different between the two perspectives.

To gain a deeper understanding of these links between subspace-based and maximum likelihood estimates therefore requires a investigation of how the Kalman-smoothed quantities (47), (48) relate to the state estimate (16), and this will be a topic of further investigation by the authors.

References

[1] H. AKAIKE, *Markovian representation of stochastic processes by canonical variables,* SIAM Journal of Control, 13 (1975), pp. 162–173.

[2] D. BAUER, M. DEISTLER, AND W. SCHERRER, *Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs,* Automatica, 35 (1999), pp. 1243–1254.

[3] M. DEISTLER, K. PETERNELL, AND W. SCHERRER, *Consistency and relative efficiency of subspace methods,* Automatica, 31 (1995), pp. 1865–1875.

[4] 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.

[5] G. GOODWIN AND K. SIN, *Adaptive Filtering Prediction and Control*, Prentice-Hall, Inc., New Jersey, 1984.

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

[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, (2nd edition)*, Prentice-Hall, Inc., New Jersey, 1999.

[10] R. SHUMWAY, *An approach to time series smoothing and forecasting using the em algorithm*, Journal of Time Series Analysis, 3 (1982), pp. 253–264.

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

[12] ——, *A unifying theorem for subspace identification algorithmsand its interpretation*, in IFAC International Symposium on System Identification, 1994, pp. 145–150, Vol2.

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

[14] 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.

[15] M. VIBERG, *Subspace methods for the identification of linear time invariant systems*, Automatica, 31 (1995), pp. 1835–1852.