An EM algorithm for the estimation of affine state-space systems with or without known inputs

Alexander W Blocker

January 2008

Abstract

We derive an EM algorithm for the estimation of affine Gaussian state-space systems with or without known inputs. We also provide comments on its numerical implementation & application and note recent extensions to predictive state representations.

1 Introduction

Linear and affine state-space systems are useful across a wide range of applications, from control theory & signal processing to speech recognition & interest rate modelling. However, despite their wide applicability and theoretical simplicity, the estimation of such systems is nontrivial. One of the most straightforward approaches is to apply the expectation-maximization (EM) algorithm originally developed by Dempster, Laird, and Rubin (1), treating the unobserved state variable as missing data. This problem was previously addressed for the linear case without control inputs in (2); here, this approach is extended to the affine case with and without known inputs. Some notes and cautions on the algorithm's implementation and usage are also provided.

2 EM Literature

As mentioned above, the EM algorithm was first formally presented by Dempster, Laird, and Rubin in 1977 (1). It is a powerful method for obtaining maximum-likelihood parameter estimates in the presence of missing data. The idea of the algorithm is to alternate between computing the expectation of the sample log-likelihood conditional on the previous parameter estimates (the expectation or "E" step) and maximizing this expectation with respect to the desired parameters to obtain parameter estimates for the next recursion (the maximization or "M" step). In (1), it was shown that this procedure is guaranteed to produce a monotonically increasing sequence of expected sample log-likelihoods which converges to a local maximum of the likelihood function (if the likelihood function is unimodal, this is clearly the unique ML estimate).

The application of the EM algorithm to state space systems has been an active area of research for over 20 years. Shumway and Stoffer presented some of the earliest work on the topic in (3), which was then extended by Ghahramani and Hintons in (2). The latter piece forms the basis for the procedure presented herein. Some examples of how this approach has been applied can be found in (4) and (5).

3 Derivation

In its most general form, the model presented here is:

$$\begin{aligned} x_t &= Fx_{t-1} + Bu_t + \varepsilon_t,\, \varepsilon_t \sim N_k(0, Q) \text{ i.i.d.} \\ z_t &= h + Hx_t + \omega_t,\, \omega_t \sim N_n(0, R) \text{ i.i.d.} \end{aligned}$$

We assume that ε_t & ω_τ are independent for all t & τ. This is the basic setup of an affine Gaussian state-space system, where z_t is the observable output variable, u_t is a known input, and x_t is the unobserved state variable (the noise variables ε_t & ω_t are also unobserved). The parameters are the transition matrix F, input matrix B, observation matrix H, affine term h, and the covariance matrices Q and R. Such a model is motivated by the problem of measurement errors. The state variable of interest x_t evolves according to some linear dynamics with known inputs. However, we can only observe z_t, a noisy, affine-transformed version of x_t. When considered in this framework, the restriction of independent errors appears quite mild, as ω_t is just measurement noise.

It is important to note that, conditional on x_t, z_t is a multivariate normally distributed random variable with mean $h + Hx_t$ and covariance matrix R. Similarly, conditional on x_{t-1} and u_t, x_t is a multivariate normally distributed random variable with mean $Fx_{t-1} + Bu_t$ & covariance matrix Q. Thus, using the fact that the given process is Markovian and adding the assumption that the initial state x_0 is unconditionally normally distributed with mean π_1 & covariance V_1, we obtain the following log-likelihood function:

$$\begin{aligned} -2\ell\left(X, Z|\theta\right) &= \sum_{t=2}^{T} \left(x_t - Fx_{t-1} - Bu_t\right)' Q^{-1} \left(x_t - Fx_{t-1} - Bu_t\right) + T \log|Q| \\ &\quad + \sum_{t=1}^{T} \left(z_t - h - Hx_t\right)' R^{-1} \left(z_t - h - Hx_t\right) + (T-1)\log|R| \\ &\quad + \left(x_1 - \pi_1\right)' V_1 \left(x_1 - \pi_1\right) + \log|V_1| + (2T-1)\log|V_1| \end{aligned}$$

where X is the T by k matrix of hidden states, Z is the T by n matrix of observations, and θ is a vector containing the given parameters.

Taking matrix derivatives, we obtain the following simplified forms for the first order conditions (FOCs):

$$F \;:\; \sum_{t=2}^{T} \left(x_t x_{t-1}' - Fx_{t-1}x_{t-1}' - Bu_t x_{t-1}'\right) = 0$$

$$
\begin{aligned}
B &: \sum_{t=2}^{T} \left(x_t u_t' - F x_{t-1} u_t' - B u_t u_t'\right) = 0 \\
Q^{-1} &: \sum_{t=2}^{T} \left(x_t - F x_{t-1} - B u_t\right)\left(x_t - F x_{t-1} - B u_t\right)' = (T-1)\,Q \\
h &: \sum_{t=1}^{T} \left(z_t - h - H x_t\right) = 0 \\
H &: \sum_{t=1}^{T} \left(z_t x_t' - h x_t' - H x_t x_t\right) = 0 \\
R^{-1} &: \sum_{t=1}^{T} \left(z_t - h - H x_t\right)\left(z_t - h - H x_t\right)' = TQ \\
x_1 &: x_1 = \pi_1 \\
V_1 &: \left(x_1 - \pi_1\right)\left(x_1 - \pi_1\right)' = V_1
\end{aligned}
$$

After taking the expectations of the above FOCs, we can obtain the formulas for the M-step of the EM algorithm by solving for the desired parameters. However, before turning to this step, we must consider the E-step.

As our state-space system has linear state dynamics and the measurement equation is affine in the state, a combination of the Kalman filter & smoother can be used to compute the expectations of the necessary statistics. The filtering is performed via the standard Kalman filter recursions. Define $\hat{x}_t$ as the estimated state at time t, V_t as the estimated state variance at time t, P_t as $E\left[x_t x_t'\right] = V_t + \hat{x}_t \hat{x}_t'$, and $P_{t,t-1}$ as $E\left[x_t x_{t-1}'\right] = V_{t,t-1} + \hat{x}_t \hat{x}_{t-1}'$. Any variable with a $-$ superscript indicates an initial estimate, and an f superscript indicates a filtered estimate (as opposed to the final, smoothed estimates). The forward recursion equations, as in (6), are:

$$
\begin{aligned}
\hat{x}_t^- &= F \hat{x}_{t-1}^f + B u_t \\
V_t^- &= F V_{t-1}^f F' + Q \\
K_t &= V_t^- H' \left(H V_t^- H' + R\right)^{-1} \\
\hat{x}_t^f &= \hat{x}_t^- + K_t \left(z_t - h - H \hat{x}_t^-\right) \\
V_t^f &= \left(I - K_t H\right) V_t^-
\end{aligned}
$$

These formulas are applied recursively from $t = 2$ through T, with the initial values given by π_1 & V_1. The smoother is then run in the opposite direction, from $t = T - 1$ through 1. The equations for this backward recursion, as in (4) and (3), are:

$$
\begin{aligned}
J_t &= V_t^f F' \left(V_{t+1}^-\right)^{-1} \\
\hat{x}_t &= \hat{x}_t^f + J_t \left(\hat{x}_{t+1} - F \hat{x}_t^f - B u_t\right) \\
V_t &= V_t^f + J_t \left(V_{t+1} - V_{t+1}^-\right) J_t'
\end{aligned}
$$

$$V_{t,t-1} = V_t^f J'_{t-1} + J_t \left(V_{t,t-1} - F V_t^f \right) J'_{t-1}$$

The values for time T are given by $\hat{x}_T = \hat{x}_T^f$, $V_T = V_T^f$, and $V_{T,T-1} = \left(I - K_T H \right) F V_{T-1}^f$.

Using the series $\{\hat{x}_t\}$, $\{P_t\}$, and $\{P_{t,t-1}\}$ from the E-step, we can write the solutions to the FOCs in expectation terms as:

$$\begin{aligned}
\hat{F} &= \left(\sum_{t=2}^{T} P_{t,t-1} - \hat{B} u_t \hat{x}'_{t-1} \right) \left(\sum_{t=2}^{T} P_{t-1} \right)^{-1} \\
\hat{B} &= \left[\sum_{t=2}^{T} \hat{x}_t u'_t - \left(\sum_{t=2}^{T} P_{t,t-1} \right) \left(\sum_{t=2}^{T} P_{t-1} \right)^{-1} \left(\sum_{t=2}^{T} \hat{x}_{t-1} u'_t \right) \right] \cdot \\
&\quad \left[\sum_{t=2}^{T} u_t u'_t - \left(\sum_{t=2}^{T} u_t \hat{x}'_{t-1} \right) \left(\sum_{t=2}^{T} P_{t-1} \right)^{-1} \left(\sum_{t=2}^{T} \hat{x}_{t-1} u'_t \right) \right]^{-1} \\
\hat{Q} &= \frac{1}{T-1} \sum_{t=2}^{T} P_t - \hat{F} P'_{t,t-1} - \hat{B} u_t \hat{x}'_t \\
\hat{h} &= \frac{1}{T} \left[\sum_{t=1}^{T} z_t - \left(\sum_{t=1}^{T} z_t \hat{x}'_t \right) \left(\sum_{t=1}^{T} P_t \right)^{-1} \left(\sum_{t=1}^{T} \hat{x}_t \right) \right] \cdot \\
&\quad \left[1 - \frac{1}{T} \left(\sum_{t=1}^{T} \hat{x}'_t \right) \left(\sum_{t=1}^{T} P_t \right)^{-1} \left(\sum_{t=1}^{T} \hat{x}_t \right) \right]^{-1} \\
\hat{H} &= \left(\sum_{t=1}^{T} z_t \hat{x}'_t - \hat{h} \hat{x}_t \right) \left(\sum_{t=1}^{T} P_t \right)^{-1} \\
\hat{R} &= \frac{1}{T} \sum_{t=1}^{T} z_t z'_t - \hat{h} z'_t - \hat{H} \hat{x}_t z'_t \\
\hat{\pi}_1 &= \hat{x}_1 \\
\hat{V}_1 &= V_1
\end{aligned}$$

These are the key equations for the M-step of the algorithm. To modify them for a version of the model without inputs or an affine measurement term, the only change required is setting $\hat{B}$ or $\hat{h}$ to zero in the relevant equations. Applying both of these restrictions leads to the same M-step equations found in (2).

4 Implementation

When writing a numerical implementation of this algorithm, setting the proper halting condition is of great importance. In (1), Dempster, Lair, & Rubin show that an EM algorithm, such as the one given above, is guaranteed to converge to at least a local maximum. Thus, in theory, one must simply tell the algorithm to stop when the expected log-likelihoods produced by two sequential iterations are identical. In practice, this is not feasible due to numerical imprecision and the sheer number of iterations that is often required to reach such a fixed point. Therefore, the typical halting condition used for applications such as this is a measure of the relative change in expected log-likelihood. Define ℓ_k as the expected log-likelihood of the k^{th} iteration. A typical halting condition would be:

$$\text{Stop if } \frac{\ell_k - \ell_{k-1}}{\frac{1}{2}\left|\ell_k + \ell_{k-1} + \varepsilon\right|} < c$$

The averaging in the denominator is used to increase the stability of the condition, and the ε term is a small non-zero value is used to keep the condition well-behaved in the event a fixed point is reached. A typical value for c in such a condition is between 1×10^{-4} & 1×10^{-6}.

Initialization of the algorithm is also an important consideration. As mentioned earlier, an EM algorithm is only guaranteed to converge to a local maximum of the likelihood function, not the global maximum. For problems in low dimensions with relatively few parameters to be estimated, this is not as large a concern. In particular, if the likelihood function is unimodal, initialization will only affect the speed of convergence, not the final result. Unfortunately, for higher-dimensional problems (such as many state space estimation problems), initialization has a significant impact on the estimates produced by the EM algorithm. Thus, using the EM algorithm on such systems with little or no prior knowledge is frequently unproductive.

If one has some prior knowledge about the system's parameters, the best course of action is the initialize based on this information. One of the best examples of this situation would be estimating the parameters for a physical system with partially understood system & observation matrices. Another possible method relies on an assumption of independence. If one believes that the final observations should be independent, independent components analysis could be used to obtain an initial estimate for the observation matrix H. However, this is a relatively strong assumption to make. Overall, the local maximization property of the EM algorithm makes it more significantly more useful for tuning a relatively known model than for blind estimation of an unknown system.

5 Conclusion

We have presented a tractable EM algorithm for the estimation of affine state-space systems with known inputs and provided some notes & cautions on its usage. For lower-dimensional systems, this method is quite effective; however, for higher-dimensional systems, it's performance deteriorates significantly due to the local maximization property of EM algorithms. To get past this limitation, some recent work has focused on predictive state representations (PSRs), which define systems using statistics over future observations. In (7), Ruday, Singh, and Wingate show that any uncontrolled linear state-space system has an equivalent PSR. They also demonstrate an algorithm for the estimation of the latter form, which is extended to controlled systems in (8). Their approach appears promising, but the basis for comparison (and, in some cases, the preferred approach), remain expectation maximization algorithms of the type presented here.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[2] Z. Ghahramani and G. E. Hintons, “Parameter estimation for linear dynamical systems,” Tech. Rep. CRG-TR-96-2, Department of Computer Science, University of Toronto, 1996.

[3] R. H. Shumway and D. S. Stoffer, “An approach to time series smoothing and forecasting using the EM algorithm,” *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[4] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, “ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 431–442, October 1993.

[5] A. Hero, H. Messer, J. Goldberg, D. Thomson, M. Amin, G. Giannakis, A. Swami, J. Tugnait, A. Nehorai, A. Swindlehurst, *et al.*, “Highlights of statistical signal and array processing,” *IEEE Signal Processing Magazine*, vol. 15, no. 5, pp. 21–64, 1998.

[6] G. Welch and G. Bishop, “An introduction to the Kalman filter.” SIGGRAPH Course Packet, 2001.

[7] M. Rudary, S. Singh, and D. Wingate, “Predictive linear-gaussian models of stochastic dynamical systems.” Working paper.

[8] M. Rudary and S. Singh, “Predictive linear-gaussian models of controlled stochastic dynamical systems,” in *ICML ’06: Proceedings of the 23rd international conference on Machine learning*, (New York, NY, USA), pp. 777–784, ACM, 2006.