MAXIMUM LIKELIHOOD NONLINEAR SYSTEM ESTIMATION

Thomas B. Schön * **Adrian Wills** ** **Brett Ninness** **

* *Division of Automatic Control*
Linköping University
SE-581 83 Linköping, Sweden
e-mail: schon@isy.liu.se
** *School of Electrical Engineering and Computer Science*
University of Newcastle
Callaghan, NSW 2308, Australia
e-mail: {Adrian.Wills, Brett.Ninness}@newcastle.edu.au

Abstract: This paper is concerned with the parameter estimation of a relatively general class of nonlinear dynamic systems. A Maximum Likelihood (ML) framework is employed in the interests of statistical efficiency, and it is illustrated how an Expectation Maximisation (EM) algorithm may be used to compute these ML estimates. An essential ingredient is the employment of so-called "particle smoothing" methods to compute required conditional expectations via a Monte Carlo approach. A simulation example demonstrates the efficacy of these techniques.

Keywords: Nonlinear Systems, System Identification, Maximum Likelihood, Expectation Maximisation Algorithm, Particle Smoother.

1. INTRODUCTION

The significance but difficulty of estimating parameterizations of nonlinear system classes is widely recognised (Ljung, 2003; Ljung and Vicino, 2005). This has led to approaches that focus on specific system classes such as those described by Volterra kernel (Bendat, 1990), neural network (Narendra and Parthasarathy, 1990), nonlinear ARMAX (NARMAX) (Leontaritis and Billings, 1985), and Hammerstein – Wiener (Rangan *et al.*, 1995) structures.

The paper here considers the estimation of a certain class of nonlinear systems that can be represented in state-space form whereby state and measurement noise enter additively and the parameter dependence is affine.

To estimate this nonlinear model structure parameterisation, a Maximum Likelihood (ML) criterion will be employed, principally in recognition of the general statistical efficiency of such an approach.

Of course, the use of an ML approach (for example, with regard to linear dynamic systems) is common, and it is customary to employ a gradient-based search technique such as a damped Gauss–Newton method to actually compute the estimates (Ljung, 1999; Söderström and Stoica, 1989). This requires the computation of a cost Jacobian which typically necessitates implementing one filter, derived (in the linear case) from a Kalman filter, for each parameter that is to be estimated.

An alternative, recently explored in (Gibson *et al.*, 2005) in the context of bilinear systems is to employ the Expectation Maximisation algorithm (Dempster *et al.*, 1977) for the computation of ML estimates.

Unlike gradient-based search, which is applicable to maximisation of any differentiable cost function, EM methods are only applicable to maximisation of likelihood functions. However, the dividend of this specialisation is that they do not require computation of gradients, and are well recognised as being particularly robust against attraction to local minima (Gibson and Ninness, 2005).

Given these recommendations, this paper develops and demonstrates an EM-based approach to nonlinear system identification. This will require the computation of smoothed state estimates that, in the linear case, could be found by standard linear smoothing methods (Gibson *et al.*, 2005). In the fairly general nonlinear (and possibly non-Gaussian) context considered in this work we propose a "particle-based" approach whereby approximations of the required smoothed state estimates are approximated by Monte Carlo based empirical averages (Doucet *et al.*, 2001).

It is important to acknowledge that there has been previous work related to this approach. In (Andrieu *et al.*, 2004), the possibility of incorporating the parameters into the state vector and employing particle *filtering* methods was discussed, but dismissed as untenable. Balancing this, the contributions (Kitagawa, 1998; Schön and Gustafsson, 2003) provide evidence to question this conclusion.

Additionally, the work (Doucet and Tadić, 2003; Andrieu *et al.*, 2004) has considered employing particle filters to compute the Jacobians necessary for a gradient-based approach. Finally, the contribution (Andrieu *et al.*, 2004) has also considered using the EM algorithm in conjunction with particle-based methods. However, by employing improved particle smoothing methods and by more careful numerical implementation of a key "maximisation" step, the present work is able to report significantly improved performance.

2. PROBLEM FORMULATION

This paper is concerned with the following model class, which is affinely parametrised in the (unknown) parameter $\vartheta \in \mathbf{R}^{n_\vartheta}$:

$$\underbrace{\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix}}_{z_t} = \underbrace{\begin{bmatrix} f_1(x_t, u_t, t) \\ h_1(x_t, u_t, t) \end{bmatrix}}_{\alpha_t} \vartheta + \underbrace{\begin{bmatrix} f_2(x_t, u_t, t) \\ h_2(x_t, u_t, t) \end{bmatrix}}_{\beta_t} + \underbrace{\begin{bmatrix} w_t \\ e_t \end{bmatrix}}_{\eta_t} \tag{1}$$

Here f_1, f_2, h_1 and h_2 are arbitrary (possible time-varying) nonlinear functions, $x_t \in \mathbf{R}^n$ is the underlying system state, $u_t \in \mathbf{R}^m$, $y_t \in \mathbf{R}^p$ are respectively (observed) multi-dimensional inputs and outputs. The initial state x_1 and noise terms w_t and e_t are assumed to be realisations from Gaussian stochastic processes given by,

$$x_1 \sim \mathcal{N}(\mu, P_1), \quad \eta_t \sim \mathcal{N}(0, \Pi).$$

In light of this, the model structure (1) is completely described by the parameter vector θ defined as

$$\theta^T \triangleq \left[\vartheta^T, \mathrm{vec}\,\{\Pi\}^T, \mathrm{vec}\,\{P_1\}^T, \mu^T\right].$$

With regard to this model structure, this paper will be solely concerned with a parameter estimate $\widehat{\theta}$ of θ derived via the ML criterion

$$\widehat{\theta}(Y_N) = \arg\max_\theta \, p_\theta(Y_N) \tag{2}$$

where $Y_N \triangleq [y_1, \cdots, y_N]$ is an N point record of observed system performance and $p_\theta(Y_N)$ is then the joint probability density function of Y_N implied by the model structure (1) and a parameter value θ.

In the linear, time invariant and Gaussian case, a (possibly steady state) Kalman Filter can be used to compute this cost (and required Jacobians for gradient-based search). Here, algorithms are developed to extend this principle to the more general nonlinear model class (1). In doing so, it is recognised that, especially in the nonlinear case, it is generally hard to compute (2) since it may well represent a non-convex optimisation problem. To address this issue, a central contribution of this work is the employment of the EM algorithm.

3. EXPECTATION MAXIMISATION ALGORITHM

The Expectation Maximisation (EM) algorithm introduced in (Dempster *et al.*, 1977) presents a non gradient-based approach for iteratively obtaining maximum likelihood estimates (2). Within areas of applied statistics, it is widely recognised for its robustness. The key idea underlying it is the consideration of an extension to (2); viz.

$$\widehat{\theta}(X_N, Y_N) = \arg\max_\theta \, p_\theta(X_N, Y_N). \tag{3}$$

Here, an extra data set X_N, commonly referred to as the *incomplete data* or *missing data* has been introduced. Its choice is an essential design variable, which if possible should be made so that the solution of (3) is straightforward.

The link between the two problems (2) and (3) is provided by the definition of conditional probability which implies

$$\log p_\theta(Y_N) = \log p_\theta(X_N, Y_N) - \log p_\theta(X_N|Y_N).$$

Taking expectations of both sides of this equation which are conditional on the observations Y_N and with respect to underlying density specified by θ being set at a value $\theta = \theta'$ will leave the left hand side unaltered, and hence deliver

$$L(\theta) = \underbrace{\mathsf{E}_{\theta'}\{\log p_\theta(X_N, Y_N)|Y_N\}}_{\mathcal{Q}(\theta,\theta')} - \underbrace{\mathsf{E}_{\theta'}\{\log p_\theta(X_N|Y_N)|Y_N\}}_{\mathcal{V}(\theta,\theta')}.$$

Since the logarithm is concave, Jensen's inequality establishes that $\mathcal{V}(\theta, \theta') \leq \mathcal{V}(\theta', \theta')$ and therefore choosing θ that satisfies $\mathcal{Q}(\theta, \theta') \geq \mathcal{Q}(\theta', \theta')$ implies that $L(\theta) \geq L(\theta')$. That is, values of θ that increase $\mathcal{Q}(\theta, \theta')$ beyond its value at θ' also increase the underlying log likelihood function of interest. This implies the Expectation Maximisation (EM) algorithm.

Algorithm 1. (EM Algorithm) Given an initial estimate θ_0, iterate the following until convergence.

E: $\mathcal{Q}(\theta, \theta_k) = \mathsf{E}_{\theta_k}\{\log p_\theta(X_N, Y_N)|Y_N\}$
M: $\theta_{k+1} = \arg\max_\theta \, \mathcal{Q}(\theta, \theta_k)$

4. EM FOR PARAMETER ESTIMATION

In agreement with previous applications of EM for parameter estimation (see discussion in (Gibson *et al.*, 2005)) we define the missing data X_N to equal

the state sequence $X_N \triangleq \{x_1, \ldots, x_{N+1}\}$. With this choice in place, the next step in applying the EM algorithm involves computation of $\mathcal{Q}(\theta, \theta_k)$ which may be achieved via the following Lemma.

Lemma 4.1. With regard to system (1) and the above choice for missing data X_N, the function $\mathcal{Q}$ can be expressed as

$$\begin{aligned} -2\mathcal{Q}(\theta, \theta_k) =& N \log\det \Pi + \mathrm{Tr}\left(\Pi^{-1}\Phi(\vartheta)\right) \\ &+ \log\det P_1 + \mathrm{Tr}\left(P_1^{-1}\Psi(\mu)\right) + c, \end{aligned}$$

where c is a constant and with $e_t \triangleq z_t - \beta_t$

$$\Psi(\mu) \triangleq \mathbf{E}_{\theta_k}\left\{(x_1 - \mu)(x_1 - \mu)^T | Y_N\right\}$$
$$\Phi(\vartheta) \triangleq \sum_{t=1}^{N} \mathbf{E}_{\theta_k}\left\{(e_t - \alpha_t\vartheta)(e_t - \alpha_t\vartheta)^T | Y_N\right\}. \quad \square$$

An essential point is that both Φ and Ψ require the computation of expectations conditional on Y_N. In the case of linear systems this can be achieved by employing a linear smoother (often called a Kalman Smoother). In the nonlinear case considered in this paper, this approach is not suitable, and alternate means for computing smoothed state estimates are required. This topic is addressed in Section 5 following.

In the meantime, supposing that it is possible to compute these expectations, then the second step of the EM algorithm involves maximisation of $\mathcal{Q}$ with respect to θ, which is the subject of the following Lemma.

Lemma 4.2. The function $\mathcal{Q}(\theta, \theta_k)$ is maximised over θ by making the following choices

$$\vartheta = \Sigma^{-1}\Gamma, \qquad \mu = \mathbf{E}_{\theta_k}\{x_1 | Y_N\}$$
$$\Pi = \Phi(\Sigma^{-1}\Gamma), \qquad P_1 = \Psi(\mathbf{E}_{\theta_k}\{x_1 | Y_N\})$$

where as before $e_t \triangleq z_t - \beta_t$ and

$$\Sigma \triangleq \sum_{t=1}^{N} \mathbf{E}_{\theta_k}\left\{\alpha_t^T\alpha_t | Y_N\right\}, \; \Gamma \triangleq \sum_{t=1}^{N} \mathbf{E}_{\theta_k}\left\{\alpha_t^T e_t | Y_N\right\}. \quad \square$$

With these definitions in place, the EM algorithm for parameter estimation can be expressed in more detail as follows.

Algorithm 2. (EM algorithm for parameter estimation) Given an initial parameter vector θ_0, iterate the following steps until convergence is achieved.

(1) Calculate Σ, Γ and $\mathbf{E}_{\theta_k}\{x_1 | Y_N\}$ then ϑ_k and μ_k.
(2) Calculate $\Phi(\vartheta_k)$ and $\Psi(\mu_k)$ then Π_k and P_{1_k}.

5. MONTE CARLO BASED SMOOTHING

In this section we examine numerical solutions of nonlinear smoothing problems that employ recursive Monte Carlo techniques. In relation to this, it is worth noting that while very significant effort has been directed towards nonlinear filtering via this sort of approach (particle filters), very little has been done when it comes to solving the nonlinear smoothing problem. See (Godsill *et al.*, 2004; Kitagawa, 1996; Tanizaki, 2001) for some work in this direction.

After careful evaluation, this paper will employ the methods developed in (Tanizaki, 2001), where the *key distinguishing idea* relative to the other work mentioned above is the consideration of propagating approximations of $p(x_{t+1}, x_t | Y_N)$ rather than $p(x_t | Y_N)$.

In order to explain the ideas, the paper begins by addressing the general problem of random number generation with respect to a given, possibly complicated distribution.

5.1 Random Sampling

Consider the problem of generating random numbers distributed according to some *target density* $t(x)$ which potentially is rather complex. One way of doing this would be to employ an alternate density that is simple to draw from, say $s(x)$, referred to as the *sampling density*, and then calculate the probability that the sample was in fact generated from the target density. That is, a sample $x^{(i)} \sim s(x)$ is drawn, and then the following ratio is calculated

$$a(x^{(i)}) \propto t(x^{(i)})/s(x^{(i)}),$$

which indicates how probable it is that $x^{(i)}$ is in fact generated from the target density $t(x)$.

The probability of accepting $x^{(i)}$ as a sample from $t(x)$ is referred to as the *acceptance probability*, and typically it is computed via consideration of $a(x^{(i)})$. This is the case, for example, for all of the so-called "rejection sampling", "importance sampling/resampling" and "Metropolis–Hastings independence sampling" methods (Tanizaki, 2001; Liu, 1996). Here, as will be detailed presently, importance resampling will be employed.

5.2 Monte Carlo based Filtering

In the case of filtering, the target density referred to above becomes $t(x_t) = p(x_t | Y_t)$, and it is then necessary to also choose an appropriate sampling density $s(\cdot)$ and acceptance probability. This is in fact quite simple, since from Bayes' theorem and the Markov property

$$\begin{aligned} p(x_t | Y_t) = p(x_t | y_t, Y_{t-1}) &= \frac{p(y_t | x_t) p(x_t | Y_{t-1})}{p(y_t | Y_{t-1})} \\ &\propto p(y_t | x_t) p(x_t | Y_{t-1}) \end{aligned}$$

which suggests, since $t(x) \propto a(x)s(x)$, the following choices

$$\underbrace{p(x_t | Y_t)}_{t(x_t)} \propto \underbrace{p(y_t | x_t)}_{a(x_t)} \underbrace{p(x_t | Y_{t-1})}_{s(x_t)}.$$

Via the principle of importance resampling the acceptance probabilities, $\{\tilde{a}^{(i)}\}_{i=1}^{M}$, are calculated according to

$$\tilde{a}^{(i)} = \frac{a(x_{t|t-1}^{(i)})}{\sum_{j=1}^{M} a(x_{t|t-1}^{(j)})} = \frac{p(y_t | x_{t|t-1}^{(i)})}{\sum_{j=1}^{M} p(y_t | x_{t|t-1}^{(j)})},$$

where $x^{(i)}_{t|t-1} \sim p(x_t|Y_{t-1})$. That is, acceptance probabilities $\tilde{a}^{(i)}$ depend upon computation of $p(y_t|x_{t||t-1})$. Via the assumption of additive noise e_t, the model (1) makes this straightforward to obtain.

The algorithm then proceeds by obtaining samples from $p(x_t|Y_t)$ by resampling the particles $\{x^{(i)}_{t|t-1}\}^M_{i=1}$ from the sampling density, $p(x_t|Y_{t-1})$, according to the corresponding acceptance probabilities, $\{\tilde{a}^{(i)}\}^M_{i=1}$. If this procedure is recursively repeated over time the following approximation

$$p(x_t|Y_t) \approx \sum_{i=1}^{M} \frac{1}{M}\delta(x_t - x^{(i)}_{t|t}) \tag{4}$$

is obtained, and we have in fact derived the *particle filter* algorithm, which is given in Algorithm 3. It was first introduced in (Gordon *et al.*, 1993).

Algorithm 3. Particle filter

(1) Initialise the particles, $\{x^{(i)}_{0|-1}\}^M_{i=1} \sim p_{x_0}(x_0)$.
(2) Calculate weights $\{q^{(i)}_t\}^M_{i=1}$ according to

$$q^{(i)}_t = p(y_t|x^{(i)}_{t|t-1})$$

and normalize $\tilde{q}^{(i)}_t = q^{(i)}_t / \sum^M_{j=1} q^{(j)}_t$.
(3) Resample N particles according to

$$\Pr(x^{(i)}_{t|t} = x^{(j)}_{t|t-1}) = \tilde{q}^{(j)}_t$$

(4) For $i = 1, \ldots, M$, predict new particles according to $x^{(i)}_{t+1|t} \sim p(x_{t+1|t}|x^{(i)}_{t|t})$.
(5) Set $t := t + 1$ and iterate from step 2.

5.3 Particle Smoother

In solving the smoothing problem the target density becomes $t(x_{t+1}, x_t) = p(x_{t+1}, x_t|Y_N)$. Similarly to what was discussed in the previous section we have to find a suitable sampling density and the corresponding acceptance probabilities to solve the smoothing problem. Again, using Bayes' theorem we have

$$p(x_{t+1}, x_t|Y_N) = p(x_t|x_{t+1}, Y_N)p(x_{t+1}|Y_N) \tag{5}$$

where

$$\begin{aligned} p(x_t|x_{t+1}, Y_N) &= p(x_t|x_{t+1}, Y_t, Y_{t+1:N}) \\ &= \frac{p(Y_{t+1:N}|x_t, x_{t+1}, Y_t)p(x_t|x_{t+1}, Y_t)}{p(Y_{t+1:N}|x_{t+1}, Y_t)} \\ &= p(x_t|x_{t+1}, Y_t) = \frac{p(x_{t+1}|x_t)p(x_t|Y_t)}{p(x_{t+1}|Y_t)}. \end{aligned} \tag{6}$$

Inserting (6) into (5) gives

$$\underbrace{p(x_{t+1}, x_t|Y_N)}_{t(x_{t+1},x_t)} = \underbrace{\frac{p(x_{t+1}|x_t)}{p(x_{t+1}|Y_t)}}_{a(x_{t+1},x_t)} \underbrace{p(x_t|Y_t)p(x_{t+1}|Y_N)}_{s(x_{t+1},x_t)}$$

At time t the sampling density can be used to generate samples. In order to find the acceptance probabilities $\{a^{(i)}\}^M_{i=1}$ we have to calculate

$$a(x_{t+1}, x_t) = \frac{p(x_{t+1}|x_t)}{p(x_{t+1}|Y_t)},$$

where $p(x_{t+1}|x_t)$ is calculated using the model (1), and $p(x_{t+1}|Y_t)$ can be approximated according to

$$\begin{aligned} p(x_{t+1}|Y_t) &= \int p(x_{t+1}|x_t)p(x_t|Y_t)dx_t \\ &\approx \sum_{j=1}^{M} \frac{1}{M} p(x_{t+1}|x^{(j)}_{t|t}), \end{aligned}$$

where (4) has been used. The particles can now be resampled according to the normalised acceptance probabilities $\{\tilde{a}^{(i)}\}^M_{i=1}$ in order to generate samples from $p(x_{t+1}, x_t|Y_N)$. The above discussion can be summarised in the following algorithm (first introduced in (Tanizaki, 2001)),

Algorithm 4. Particle smoother

(1) Run the particle filter (Algorithm 3) and store the filtered particles, $\{x^{(i)}_{t|t}\}^M_{i=1}, t = 1, \ldots, N$.
(2) Initialise the smoothed particles and importance weights at time N according to $\{x^{(i)}_{N|N} = x^{(i)}_{N|N}, \tilde{q}^{(i)}_{N|N} = 1/M\}^M_{i=1}$ and set $t := t - 1$.
(3) Calculate weights $\{q^{(i)}_{t|N}\}^M_{i=1}$ according to

$$q^{(i)}_{t|N} = \frac{p(x^{(i)}_{t+1|N}|x^{(i)}_{t|t})}{\frac{1}{M}\sum^M_{j=1} p(x^{(i)}_{t+1|N}|x^{(j)}_{t|t})}$$

and normalise $\tilde{q}^{(i)}_{t|N} = q^{(i)}_{t|N} / \sum^M_{j=1} q^{(j)}_{t|N}$.
(4) Resample the smoothed particles according to

$$\Pr(x^{(i)}_{t+1|N}, x^{(i)}_{t|N}) = (x^{(j)}_{t+1|N}, x^{(j)}_{t|t}) = \tilde{q}^{(j)}_{t|N}$$

(5) Set $t := t - 1$ and iterate from step 3.

5.4 Using a particle smoother with EM

In Lemmas 4.1 and 4.2 we require the computation of various expectations that are conditional on the data Y_N. In the following Lemma we provide explicit formulations of these expectations in terms of smoothed particles as calculated in Algorithm 4.

Lemma 5.1. Using the smoothed state particles as calculated in Algorithm 4 we have the following approximations

$$\mathsf{E}_{\theta_k}\left\{\alpha^T_t \alpha_t|Y_N\right\} \approx \frac{1}{M}\sum_{i=1}^{M}\left(\alpha^{(i)}_t\right)^T\left(\alpha^{(i)}_t\right)$$

$$\mathsf{E}_{\theta_k}\left\{\alpha^T_t e_t|Y_N\right\} \approx \frac{1}{M}\sum_{i=1}^{M}\left(\alpha^{(i)}_t\right)^T\left(\mathrm{e}^{(i)}_t\right)$$

$$\mathsf{E}_{\theta_k}\left\{x_1|Y_N\right\} \approx \frac{1}{M}\sum_{i=1}^{M} x^{(i)}_{1|N}.$$

Similarly,

$$\mathsf{E}_{\theta_k}\left\{(x_1 - \mu)(x_1 - \mu)^T|Y_N\right\} \approx \frac{1}{M}\sum_{i=1}^{M}(x^{(i)}_{1|N} - \mu)(x^{(i)}_{1|N} - \mu)^T$$

$$\mathbf{E}_{\theta_k}\left\{(e_t-\alpha_t\vartheta)(e_t-\alpha_t\vartheta)^T|Y_N\right\}\approx$$
$$\frac{1}{M}\sum_{i=1}^{M}(e_t^{(i)}-\alpha_t^{(i)}\vartheta)(e_t^{(i)}-\alpha_t^{(i)}\vartheta)^T$$

where $e_t^{(i)}$ and $\alpha_t^{(i)}$ are simply the respective functions evaluated at the i'th particle $x_{t|N}^{(i)}$.

6. SIMULATION EXAMPLE

This section profiles the performance of the EM-based estimation methods just presented by way of considering the following nonlinear system.

$$x_{t+1}=ax_t+b\frac{x_t}{1+x_t^2}+c\cos(1.2t)+w_t, \quad (7a)$$

$$y_t=dx_t^2+e_t, \quad (7b)$$

where $a=0.5$, $b=25$, $c=8$, $d=0.05$, $w_t\sim\mathcal{N}(0,10^{-2})$ and $e_t\sim\mathcal{N}(0,10^{-2})$. In terms of the structure in (1) we make the following associations

$$\alpha_t=\begin{bmatrix}x_t & \dfrac{x_t}{1+x_t^2} & \cos(1.2t) & 0\\ 0 & 0 & 0 & x_t^2\end{bmatrix}, \qquad \beta_t=0,$$
$$\vartheta^T=[a\ b\ c\ d].$$

This system has been extensively studied in the context of *state* estimation (Gordon *et al.*, 1993; Kitagawa, 1996; Kitagawa, 1998; Doucet *et al.*, 2000; Godsill *et al.*, 2004). However, it has not been the subject of great attention from the *parameter* estimation viewpoint of this paper.

As is well recognised (Ljung, 2003), a particularly important aspect of nonlinear system estimation is the difficulty of finding appropriate initial parameter values with which to initialise an iterative search.

To address this issue, and in so doing illustrate the inherent robustness of the EM-based approach presented here, each of the 200 simulation runs was initialised at a randomly chosen initial estimate $\widehat{\theta}_0$ which itself was formed using perturbations from the true values.

Using $N=1000$ data samples, and despite only using a very modest number of $M=50$ particles in the smoothing calculations, the empirical estimation results shown in Fig. 1 are encouraging. In particular, note that despite quite widely varying initialisations, convergence to the true parameters occurred in most cases. Further simulations were conducted with $M=100$ and higher number of particles, but without any significant performance benefit. This suggests a robustness of the EM-based approach to inaccuracies in computation in the E-step.

In relation to this, note that the method requires $O(NM^2)$ floating point operations per iteration. The computational load is sensitive to the number of particles chosen, but scales well with increasing observed data length. To provide a reference point for these scaling comments, each simulation required to present the Monte–Carlo presentation in Fig. 1 completed within 3 minutes on a Pentium IV running at 3GHz.

By way of comparison, alternative methods, including Newton-based gradient search were also tried, but proved very unsuccessful.

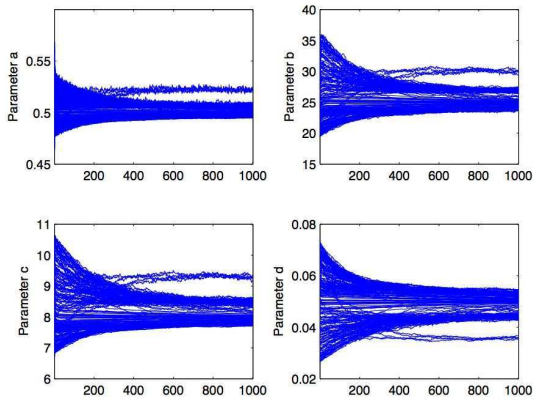

Fig. 1. *Parameter estimates for each of the 200 simulation runs as they evolve over 1000 iterations of the EM method. The true parameter values are* $a=0.5$, $b=25$, $c=8$ *and* $d=0.05$.

To explore the reason behind this, and also to emphasise the surprising robustness to initial starting point just presented, consider the simpler estimation problem which involves estimating only $\vartheta=[a,b]^T$ with c and d fixed to their true values, and with the additive noise w_t and e_t set to zero. The former is done so that the cost surface implied by the likelihood can be visualised, and the latter is considered so that attention is focused solely on how the nonlinear dynamics affects the difficulty of the estimation problem.

The resulting mean square error (the dominating component of the likelihood computation) cost surface is shown in Fig. 2. Clearly, it is very far from convex. Note that the very irregular cost function, even if due

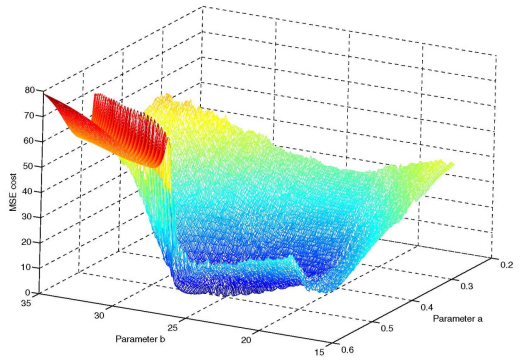

Fig. 2. *Surface plot of the MSE versus parameters* a *and* b *only.*

to finite precision effects and not intrinsic, is still an obstacle to gradient based methods but not, as will be illustrated, to an EM-based approach. The perhaps surprising complexity from such a simple example underlines the particular difficulties of nonlinear system estimation.

The MSE cost function associated with the present problem contains quite a few local minima. It is therefore not surprising that gradient-based search was found to perform so poorly on the preceding example. To emphasise this, Fig. 3 shows a contour plot of the

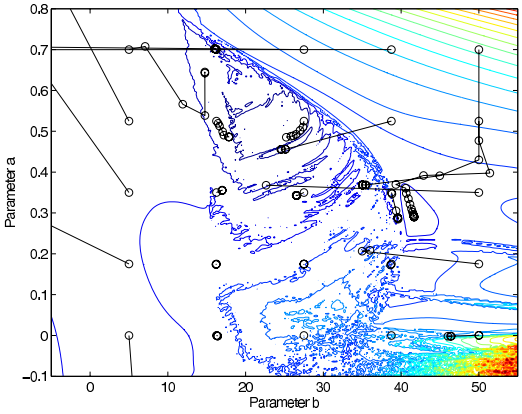

Fig. 3. *Contour plot of MSE cost for the case of identifying parameters a and b only, together with Gauss–Newton gradient-based search estimate trajectories overlaid. Note that, presumably due to the very large number of local minima, no trajectories converge to the global minimum.*

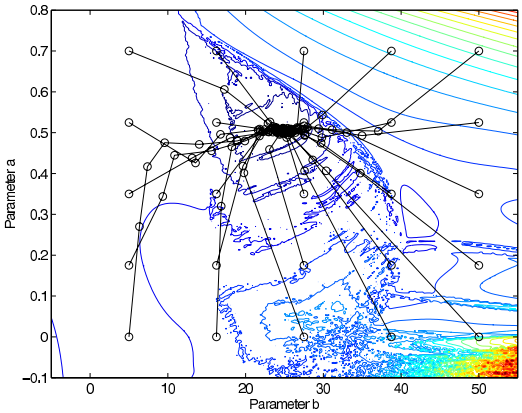

Fig. 4. *Same as previous plot, but with EM-based estimate trajectories for 25 different starting points. Note that all converge to the global minimum.*

the MSE cost function. Clearly, and as suggested in the previous figure, there seem to be a large number local minima, any of which may attract gradient-based approaches. Indeed, the black lines shown in that diagram are Gauss–Newton gradient-based search trajectories for 25 different starting points, and all become locked in local minima.

By way of contrast, Fig. 4 shows the estimate trajectories of the EM-based algorithm of this paper. Note that from the same starting points, all cases converge to the global maximum.

7. CONCLUSION

This paper has explored an approach to nonlinear dynamic system estimation whose key distinguishing features include the use of EM-based methods as opposed to more traditional gradient-based search, a fairly general model structure, the use of Monte Carlo based "particle" methods for the computation of required smoothed state estimates, and a capacity for simply encompassing multivariable problems.

By way of example, the resulting approach has been demonstrated to be (perhaps) surprisingly robust to attraction to local minima, even in cases where the underlying cost is extremely "irregular" and non-convex. Further work is required to understand the mechanisms underlying this robustness, and to test the ideas on more substantial problem sizes.

REFERENCES

Andrieu, C., A. Doucet, S. Singh and V. Tadić (2004). Particle methods for change detection, system identification, and contol. *Proceedings of the IEEE* **92**(3), 423–438.

Bendat, J. (1990). *Nonlinear System Analysis and Identification from Random Data*. Wiley Interscience.

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

Doucet, A. and V. Tadić (2003). Parameter estimation in general state-space models using particle methods. *Annals of the Institute of Statistical Mathematics* **55**, 409–422.

Doucet, A., de Freitas, N. and Gordon, N. (Eds.) (2001). *Sequential Monte Carlo Methods in Practice*. Springer Verlag.

Doucet, A., S. J. Godsill and C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing* **10**(3), 197–208.

Gibson, S., A. Wills and B. Ninness (2005). Maximum-likelihood parameter estimation of bilinear systems. *IEEE Transactions on Automatic Control* **50**(10), 1581–1596.

Gibson, S. and B. Ninness (2005). Robust maximum-likelihood estimation of multivariables dynamic systems. *Automatica* **41**(10), 1667–1682.

Godsill, S. J., A. Doucet and M. West (2004). Monte Carlo smoothing for nonlinear time series. *Journal of the American Statistical Association* **99**(465), 156–168.

Gordon, N. J., D. J. Salmond and A. F. M. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In: *IEE Proceedings on Radar and Signal Processing*. Vol. 140. pp. 107–113.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics* **5**(1), 1–25.

Kitagawa, G. (1998). A self-organizing state-space model. *Journal of the American Statistical Association* **93**(443), 1203–1215.

Leontaritis, I. and S. Billings (1985). Input-output parametric models for non-linear systems. part ii: stochastic non-linear systems. *International Journal of Control* **41**(2), 329–344.

Liu, J. (1996). Metropolized independent sampling with comparison to rejection ampling and importance sampling. *Statistics and Computing* **6**, 113–119.

Ljung, L. (1999). *System identification, Theory for the user*. System sciences series. 2nd ed. Prentice Hall. Upper Saddle River, NJ.

Ljung, L. (2003). Bode Lecture: Challenges of Nonlinear System Identification. *Proceedings of the IEEE Conference on Decision and Control*. Hawaii, USA.

Ljung, L. and Vicino, A. (Eds.) (2005). Special Issue on System Identification. Vol. 50. *IEEE Transactions on Automatic Control*.

Narendra, K. and K. Parthasarathy (1990). Identification and control of dynamical systems using neural networks. *IEEE Transactions on Neural Networks* **1**, 4–27.

Rangan, S., G. Wolodkin and K. Poolla (1995). New results for Hammerstein system identification. In: *Proceedings of the 34th IEEE Conference on Decision and Control*. New Orleans, USA. pp. 697–702.

Schön, T. and F. Gustafsson (2003). Particle filters for system identification of state-space models linear in either parameters or states. In: *proceedings of the 13th IFAC Symposium on System Identification*. Rotterdam, The Netherlands. pp. 1287–1292.

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

Tanizaki, H. (2001). Nonlinear and non-Gaussian state space modeling using sampling techniques. *Annals of the Institute of Statistical Mathematics* **53**(1), 63–81.