TECHNICAL RESEARCH REPORT

Towards a Useful Approximation of the Optimal Recursive Parameter Estimator

by D.M. Wiberg, D. Ljungquist, and T.D. Powell

T.R. 94-23

INSTITUTE FOR SYSTEMS RESEARCH

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Submitted to SYSID '94 Copenhagen
10th IFAC Symp. on System Identification.

Towards a Useful Approximation of the Optimal Recursive Parameter Estimator

D.M. Wiberg†, D. Ljungquist‡, and T.D. Powell§

A recursive parameter estimator in discrete time is presented for multiple-input, multiple-output linear stochastic systems that are bilinear in the parameters and state. It is globally convergent to minima of the asymptotic negative log likelihood function, and approximates the fast transient response of the optimal nonlinear filter used as a parameter estimator.

Key Words — Parameter estimation; nonlinear filtering; recursive estimation; system identification; optimal estimation; estimation theory; global convergence.

† Electrical Engineering Department, University of California at Los Angeles, Los Angeles, CA 90024-1594 USA In part, research performed as Visiting Professor, University of Maryland, College Park MD, and supported by NSF grant NSFD CDR-88003012

‡ Hydro Aluminium, N-5870 Øvre Årdal, NORWAY

§ Mechanical, Aerospace, and Nuclear Engineering Department, University of California at Los Angeles, Los Angeles, CA 90024-1597 USA

Abstract — Multiple-input, multiple-output (MIMO), exponentially stable, linear, stochastic systems in discrete time are considered, that are bilinear in the state and parameters to be recursively estimated, and in which no such parameters enter into the output map. A new recursive parameter estimator is developed that is globally convergent to minima of the asymptotic negative log likelihood function. Furthermore, as time t tends to infinity, the new estimator asymptotically tends to a variant of the recursive prediction error method, and as the sampling interval h tends to zero, the new estimator tends to a convergent approximation of the continuous-time minimum variance parameter estimator. This establishes a connection in discrete time between commonly used convergent parameter estimators and the optimal nonlinear filter. Numerical simulations indicate the new estimator has better transient response than commonly used convergent parameter estimators, which is to be expected because the commonly used estimators are not approximations of the optimal nonlinear filter. However, the new estimator requires about n^2p dynamic equations, where n and p are the dimensions of the state and parameter vectors, respectively.

1. Introduction

A number of recursive parameter estimation algorithms are available commercially, e.g. in the System Identification Toolbox of MATLAB. Such commonly used algorithms include the gradient, the extended least squares (ELS), the recursive prediction error (RPE), and the extended Kalman filter (EKF). These algorithms are described and analyzed in a number of books and articles, including Ljung and Söderström (1983) and Ljung (1987). For numerical implementation, all these algorithms have been developed in discrete time. They all update the parameter estimate $\hat{\theta}(t+h)$ incrementally proportional to the prediction error, i.e., the difference between the real system output $y(t)$ and that predicted by the model using the present parameter value, $\hat{y}\left(t, \hat{\theta}(t)\right)$.

In the limit as the sampling interval h tends to zero, these algorithms have continuous-time counterparts. (However, sometimes appropriate normalizations must be made, as in Gevers et al. (1991).) The connection between these algorithms and the optimal nonlinear filter used as a parameter estimator was made by DeWolf and Wiberg (1993), and Wiberg and DeWolf (1993) in continuous time. The discrete-time optimal nonlinear filter propagates the conditional probability density by integrals at each time step (Kramer and Sorensen, 1987), and no equations exist for updating the parameter estimate in terms of the prediction error. In continuous time, optimal estimates are updated using the prediction error, so it is necessary to explore the connection between the optimal nonlinear filter and these commonly used algorithms in continuous time. It is the purpose of this paper to develop a discrete-time algorithm that somehow preserves the qualities of the approximately optimal parameter estimator in continuous time that was developed in Wiberg and DeWolf (1993).

Before proceeding with this development, a quick review of the conclusions of DeWolf and Wiberg (1993), and Wiberg and DeWolf (1993) establishes the benefits of developing such a discrete-time approximation of the optimal parameter estimator.

(i) The EKF fails, with probability one, to converge to the true value of the parameter in a model whose state noise covariance is unknown. In such a case, global convergence to the true parameter value is almost impossible. However, when the EKF does converge to the true parameter value, the transient response appears close to optimal.

(ii) The ELS often needs a prefilter to satisfy a "positive real" condition that guarantees global convergence in such a case. But use of this prefilter slows the asymptotic rate of convergence.

(iii) The gradient method is globally convergent under mild conditions, but the rate of convergence can be very slow.

(iv) The RPE is globally convergent under somewhat more strict conditions than the gradient method, but is conjectured to have an optimal rate of convergence asymptotically.

(v) A globally convergent approximation of the optimal parameter estimator is developed that is asymptotically equivalent to the RPE, and thus inherits the asymptotic rate of convergence of the RPE.

(vi) Numerical simulations indicate this globally convergent approximation has faster transient response than the RPE, in the sense of minimum error variance as a function of time. This is probably because the RPE is not an approximation of an optimal (minimum variance) estimator.

It is plausible that the conclusions (i)-(vi) hold in discrete time, at least for small enough sampling interval h. This is the motivation for developing a discrete-time version of the globally convergent approximation of the optimal parameter estimator derived in Wiberg and DeWolf (1993).

Three improvements have been made in the form of the algorithm of Wiberg and DeWolf (1993). First, rather than the standard deviation, the *variance* of the process noise has been parameterized linearly in the parameter vector θ. Parameterizing the standard deviations linearly in θ gives likelihood functions with nonunique minima, such as at $\pm\theta_0$, where

θ_0 is the true value, and thus creates local maxima, such as at $\theta = 0$, to which estimate trajectories can be attracted in finite time. This phenomenon is further explained in Powell et al. (1993), who also show it can be fixed by parameterizing the variance linearly in θ.

Second, two third order moment approximations have been eliminated, in an attempt to minimize computation. Call $\tilde{x}$ and $\tilde{\theta}$ the state and parameter error vectors of dimensions n and p, respectively. Call $\otimes$ the Kronecker product as defined in Brewer (1978), where the properties of $\otimes$ are further discussed in Appendix I. Here, zero is taken to be the approximation of the conditional expectations of $\tilde{x} \otimes \tilde{x} \otimes \tilde{x}$ and $\tilde{\theta} \otimes \tilde{\theta} \otimes \tilde{x}$, rather than approximating the third order moment dynamic equations of the optimal nonlinear filter. This reduces the number of dynamic equations by n^3 and np^2. Limited numerical experimentation has shown very small loss in transient performance. Furthermore, Wiberg and DeWolf (1993) show that the approximations of the conditional expectations of $\tilde{x} \otimes \tilde{x} \otimes \tilde{x}$ and $\tilde{\theta} \otimes \tilde{\theta} \otimes \tilde{x}$ asymptotically decouple from the other estimator update dynamical equations, so that there is no asymptotic error incurred by not retaining these approximations.

Third, correlation is permitted between process and measurement noise. This permits ARMAX processes (AutoRegressive Moving Average processes with eXogenous variables) to be included in the class of systems whose parameters are to be estimated.

The paper is organized as follows. First, notation is defined and the form of the model considered is given. Second, the new parameter estimation algorithm is stated. Its ~~lengthy~~ development is postponed to Appendix II. Appendix I is a review of Kronecker products, necessary in the treatment of third order moments as used here. Third, the new estimator is shown to approach the discrete-time RPE in the limit as t tends to infinity. Fourth, global convergence is proven. Fifth, the new estimator is shown to approach the approximation of the minimum variance optimal nonlinear filter of Wiberg and DeWolf (1993) as the sampling interval h tends to zero. Sixth, application of the new estimator and the commonly used estimators are compared by simulation for simple hypothetical examples.

2. Models Considered

Let the sampling interval be h, so that discrete time t takes values in the set $T = \{0, h, 2h, \dots\}$. The n-vector state $x(t)$ at time t obeys the linear state space equation

$$x(t+h) = A(\theta)\,x(t) + B(\theta)\,u(t) + v(t) \quad , \tag{2.1}$$

$$y(t) = Cx(t) + w(t) \quad . \tag{2.2}$$

Note equation (2.2) for the output ℓ-vector $y(t)$ at time t is independent of the parameter p-vector θ. The zero mean white Gaussian random sequences defined by $v(t)$, an n-vector, and $w(t)$, an ℓ-vector, are independent of the initial condition $x(0) = x_0(\theta)$, and have a covariance

$$E\left\{ \begin{pmatrix} v(t) \\ w(t) \end{pmatrix} \begin{pmatrix} v^T(\tau) & w^T(\tau) \end{pmatrix} \right\} = \begin{pmatrix} Q(\theta) & S(\theta) \\ S^T(\theta) & R \end{pmatrix} h^{-1}\,\delta_{t\tau} \quad , \tag{2.3}$$

in which the known $\ell \times \ell$ symmetric matrix R is positive definite, E is the expectation operator, the superscript T denotes transpose, and $\delta_{t\tau}$ is the Kronecker delta defined equal to zero except in the case $\tau = t$, when $\delta_{tt} = 1$. Call the set $D_N \subset \mathbf{R}^p$ the collection of θ such that (2.3) is nonnegative definite, i.e., for $\theta \in D_N$, then $Q(\theta) \geq S(\theta)\,R^{-1}\,S^T(\theta)$.

To insure the existence of a limit measure (essentially, a steady state probability density), $A(\theta)$ must be strictly stable. Therefore, define the stability region $D_S \subset \mathbf{R}^p$ such that for

$\theta \in D_S$ the eigenvalues of $A(\theta)$ are strictly inside the unit circle. Assume that any "true value" θ_0 of θ is such that $\theta_0 \in D_S \cap D_N$.

Furthermore, to assure the existence of a limit measure, assume the input m-vector valued input process u is generated by an exponentially stable linear Gaussian stochastic system, and that u does not depend on θ.

The matrices $A(\theta)$, $B(\theta)$, $Q(\theta)$, and $S(\theta)$ are affine in θ. Define known matrices $A_0, A_1, \ldots, A_n$, and A_θ such that

$$A(\theta) = A_0 + A_\theta(I_n \otimes \theta) \quad , \tag{2.4}$$

where I_n is the $n \times n$ identity matrix, and

$$A_\theta = (A_1 \mid A_2 \mid \ldots \mid A_n) \quad , \tag{2.5}$$

in which the vertical bars denote partitioning, and $A_1, \ldots, A_n$ are each $n \times p$ matrices. Similarly, define B_0, Q_0, S_0, B_θ, etc., to obtain the most general affine form for $A(\theta)$, $B(\theta)$, $Q(\theta)$, and $S(\theta)$.

Using (2.4) and (2.5), etc., eqs. (2.1)-(2.3) represent the most general MIMO discrete-time linear stochastic system that is bilinear in the state and in the parameters to be recursively estimated, and in which C and R are independent of θ. However, the measurement noise covariance R can be estimated independently of θ using the prediction errors (Ljung and Söderström, 1983). Furthermore, the extension to known delay and time-varying systems is straightforward, as is the incorporation of forgetting factors.

The case of affine in θ $C(\theta)$ can be derived in the continuous time optimal non-linear filter. Unfortunately, the algebra becomes horribly complicated then, and the analysis is ~~greatly~~ correspondingly complicated. ~~However,~~ This is a topic of present research.

3. Statement of the Algorithm

Let the extended state be $\xi = (x^T\,\theta^T)^T$. Define

$$\tilde{\xi}(t\,|\,\tau) \;=\; \xi(t) - E\{\,\xi(t) \mid \mathcal{Y}(\tau)\,\} \quad ,$$

where $\mathcal{Y}(\tau)$ is the set of $y(t)$ for $t \in \{0, h, \ldots, h\tau\}$. Call the approximations $\hat{\xi}$, $\hat{x}$, $\hat{\theta}$, P_ξ, P_x, $P_{\theta x}$, P_θ, M to the conditional expectations of ξ, x, θ, $\tilde{\xi}\,\tilde{\xi}^T$, $\tilde{x}\,\tilde{x}^T$, $\tilde{\theta}\,\tilde{x}^T$, $\tilde{\theta}\,\tilde{\theta}^T$, and $\tilde{\theta}\,\tilde{x}^T \otimes \tilde{x}$, respectively. Also call $m = \operatorname{vec} M$, the column stacking of the matrix M. These approximations are the variables propagated by the new algorithm.

Define the ~~innovations~~ prediction error ε by

$$\varepsilon(t) \;=\; y(t) - C\hat{x}(t\,|\,t-h) \tag{3.1}$$

and the approximation of the ~~innovations~~ prediction error covariance V by

$$V(t) \;=\; h^{-1}R + C\,P_x(t\,|\,t-h)\,C^T \quad . \tag{3.2}$$

Define the state and parameter measurement update gains K_x and K_θ by

$$K_x(t) \;=\; P_x(t\,|\,t-h)\,C^T\,V^{-1}(t) \tag{3.3}$$

and

$$K_\theta(t) \;=\; P_{\theta x}(t\,|\,t-h)\,C^T\,V^{-1}(t) \quad . \tag{3.4}$$

From the state equations (2.1)-(2.5), form the quantities

$$\mathcal{A}(\theta) \;=\; A(\theta) - S(\theta)\,R^{-1}\,C \tag{3.5}$$

$$\mathcal{A}_\theta \;=\; A_\theta - S_\theta\left(R^{-1}\,C \,\otimes\, I_p\right) \tag{3.6}$$

$$\mathcal{B}(x,y,u) \;=\; \mathcal{A}_\theta(x \otimes I_p) + B_\theta(u \otimes I_p) + S_\theta\left(R^{-1}y \otimes I_p\right) \tag{3.7}$$

$$A_\xi(\theta, x, y, u) = \begin{pmatrix} \mathcal{A}(\theta) & \mathcal{B}(x, y, u) \\ & \\ O_{p,n} & I_p \end{pmatrix} . \qquad (3.8)$$

As explained in Ljung and Söderström (1983), pp. 93, 162, and 366, projections of updates must be taken. Denote the projection of the update of z to the domain D by $[z]_D$. Call the initial estimates of x, θ, P_x, and P_θ as x_0, θ_0, P_{x0}, and $P_{\theta 0}$, where P_{x0} and $P_{\theta 0}$ are positive definite. The new algorithm starts from initial conditions $\hat{x}(0 \mid -h) = x_0$, $\hat{\theta}(0 \mid -h) = \theta_0$, $P_x(0 \mid -h) = P_{x0}$, $P_{\theta x}(0 \mid -h) = 0$, $P_\theta(0 \mid -h) = P_{\theta 0}$, and $m(0 \mid -h) = 0$.

The new parameter estimation algorithm is motivated in Appendix 2. The following measurement and time updates result.

Measurement update:

$$\hat{x}(t \mid t) = \hat{x}(t \mid t-h) + K_x(t)\, \varepsilon(t) \qquad (3.9)$$

$$\hat{\theta}(t \mid t) = \left[\hat{\theta}(t \mid t-h) + K_\theta(t)\, \varepsilon(t) \right]_{D_S \cap D_N} \qquad (3.10)$$

$$P_x(t \mid t) = [\, I_n - K_x(t)\, C\,]\, P_x(t \mid t-h) \qquad (3.11)$$

$$P_{\theta x}(t \mid t) = \left[[\, P_{\theta x}(t \mid t-h) + (I_p \otimes \varepsilon^T(t)\, V^{-1}(t)\, C)\, M(t \mid t-h)\,]\, (I_n - K_x(t)\, C)^T \right]_{D_P} \qquad (3.12)$$

$$P_\theta(t \mid t) = \left[P_\theta(t \mid t-h) - P_{\theta x}(t \mid t-h)\, C^T\, V^{-1}(t)\, C\, P_{\theta x}^T(t \mid t-h) - P^\varepsilon(t) - P^{\varepsilon\, T}(t) \right]_{D_P} \qquad (3.13)$$

$$m(t \mid t) = [\, (I_n - K_x(t)\, C) \otimes I_p \otimes (I_n - K_x(t)\, C)\,]\, m(t \mid t-h) \qquad (3.14)$$

where

$$P^\varepsilon(t) = \left(I_p \otimes \varepsilon^T(t)\, V^{-1}(t)\, C \right)\, M(t \mid t-h)\, C^T\, V^{-1}(t)\, C\, P_{\theta x}(t \mid t-h) \qquad (3.15)$$

and in which D_P is the region of $P_{\theta x}$ and P_θ space in which $P_\theta \geq P_{\theta x}\, P_x^{-1}\, P_{\theta x}^T$.

To next compute the time update, adopt the convention that the subscript t means the variable in question is evaluated with its arguments at $(t\,|\,t)$, e.g. $A_t = A(\theta(t\,|\,t))$.

Time update:

$$\hat{x}(t+h\,|\,t) = A_t\,\hat{x}_t + B_t\,u(t) + S_t\,R^{-1}[y(t) - Cx_t] + \mathcal{A}_\theta\ \mathrm{vec}\,P_{\theta x t} \tag{3.16}$$

$$\hat{\theta}(t+h\,|\,t) = \hat{\theta}_t \tag{3.17}$$

$$P_\xi(t+h\,|\,t) = A_{\xi t}\,P_{\xi t}\,A_{\xi t} + H(t) + H^T(t) + \mathrm{diag}\,\left(P^A(t) + P^\Sigma(t),\, O_p\right) \tag{3.18}$$

$$\cdot\ m(t+h\,|\,t) = (\mathcal{A}_t \otimes I_p \otimes \mathcal{A}_t)\,m_t + m^A(t) + m^B(t) + m^\Sigma(t) \tag{3.19}$$

in which

$$H(t) = [\,\mathrm{diag}\,(\mathcal{A}_\theta\,U_{p,n}\,M_t,\,O_p)]\,A_{\xi t}^T \tag{3.20}$$

$$P^A(t) = \mathcal{A}_\theta\left[P_{xt} \otimes P_{\theta t} + U_{p,n}\left(P_{\theta x t} \otimes P_{\theta x t}^T\right)\right]\mathcal{A}_\theta^T \tag{3.21}$$

$$P^\Sigma(t) = h^{-1}\left[Q_t - S_t\,R^{-1}\,S_t - S_\theta\left(R^{-1} \otimes P_{\theta t}\right)S_\theta^T\right] \tag{3.22}$$

$$\begin{aligned} m^\Sigma(t) = {} & h^{-1}\,(U_{p,n} \otimes I_n)\,\big\{\,\mathrm{vec}\,[Q_\theta(I_n \otimes P_{\theta t})\,U_{p,n}] + [I_{n^2,p} + \\ & (I_p \otimes U_{n,n})]\left[I_p \otimes S_\theta\,U_{p,\ell} \otimes S_t\,R^{-1}\right]\,[\,\mathrm{vec}\,P_{\theta t} \otimes \mathrm{vec}\,I_\ell]\big\} \end{aligned} \tag{3.23}$$

and

$$\begin{aligned} m^A(t) = {} & \big\{(\mathcal{A}_t \otimes I_p \otimes \mathcal{A}_\theta) + [\,\mathcal{A}_\theta \otimes (I_p \otimes \mathcal{A}_t)\,U_{n,p}\,]\,(U_{np,n} \otimes I_p)\big\}\,(\,\mathrm{vec}\,P_{\theta x t}^T \otimes \mathrm{vec}\,P_{\theta x t}) \\ & + \big\{[\,(\mathcal{A}_t \otimes I_p)\,U_{p,n} \otimes \mathcal{A}_\theta\,]\,(U_{n^2,p} \otimes I_p) + (\mathcal{A}_\theta \otimes I_p \otimes \mathcal{A}_t)\,(I_n \otimes U_{n,p^2})\big\}\,(\,\mathrm{vec}\,P_{xt} \otimes \mathrm{vec}\,P_{\theta t}) \end{aligned} \tag{3.24}$$

$$\begin{aligned} m^B(t) = {} & \big\{[\,\mathcal{A}_\theta \otimes (I_p \otimes \mathcal{B}_t)\,(I_{p^2} + U_{p,p})\,]\,(U_{p^2,n} \otimes I_p) + [\,\mathcal{B}_t \otimes (I_p \otimes \mathcal{A}_\theta)\,U_{p,np}\,]\,(I_{p^2} \otimes U_{n,p}) \\ & + [\,\mathcal{B}_t \otimes I_p \otimes (\mathcal{A}_\theta\,U_{p,n})\,]\,U_{p^2n,p}\big\}\,(\,\mathrm{vec}\,P_{\theta t} \otimes \mathrm{vec}\,P_{\theta x t})\,. \end{aligned} \tag{3.25}$$

Computing first the measurement update (3.9) – (3.14) and then the time update (3.16) – (3.19) at each time t propagates the estimates $\hat{x}$, $\hat{\theta}$, P_ξ, and m.

There are $n+p$ first order moment approximations $\hat{x}$ and $\hat{\theta}$, there are $(n+p)\,(n+p+1)\,/\,2$ second order moment approximations P_ξ, and there are $pn(n+1)\,/\,2$ third order moment approximations M to propagate at each time step. This is the same number of dynamic equations as the state space form of the recursive prediction error method, but much more than most other recursive parameter estimation algorithms.

4. Asymptotic Equivalence to RPE

Here, the unprojected new algorithm (3.9) – (3.25) is shown to approach a variant of the state space form of the unprojected recursive prediction error method (RPE) of Ljung and Söderström (198~~6~~3) in the limit as time t tends to infinity. From page 428 of Ljung and Söderström (198~~6~~3), in terms of the system (2.1) – (2.5) of this paper, the RPE is

$$R(t) = R(t-h) + t^{-1}\left[W^T(t)\,C^T\,V^{-1}(t)\,CW(t) - h\,R(t-h)\right] \tag{4.1}$$

$$\hat{\theta}(t) = \hat{\theta}(t-h) + t^{-1}\,R^{-1}(t)\,W^T(t)\,C^T\,V^{-1}(t)\,\varepsilon(t) \tag{4.2}$$

$$\hat{x}(t+h) = A_t\,\hat{x}(t) + B_t\,u(t) + K(t)\,\varepsilon(t) \tag{4.3}$$

$$W(t+h) = [A_t - K(t)\,C]\,W(t) + A_\theta(\hat{x}(t)\otimes I_p) + \cancel{K(t)\,\varepsilon(t)} \tag{4.4}$$

$$B_\theta(u(t)\otimes I_p) + [\mathcal{K}_1(t)|\mathcal{K}_2(t)|\cdots|\mathcal{K}_p(t)][\varepsilon(t)\otimes I_p]$$

$$P(t+h) = A_t\,P(t)\,A_t^T + h^{-1}\,Q_t - K(t)\,V(t)\,K^T(t) \tag{4.5}$$

and for $i = 1, 2, \ldots, p$,

$$\Pi_i(t+h) = A_\theta(I_n\otimes e_i)\,P(t)\,A_t^T + A_t\,\Pi_i(t)\,A_t^T + A_t\,P(t)\,(I_n\otimes e_i^T)\,A_\theta^T$$

$$+\,Q_\theta\,(I_n\otimes e_i) - \mathcal{K}_i(t)\,V(t)\,K^T(t) - K(t)\,C\,\Pi_i(t)\,C^T\,K^T(t) - K(t)\,V(t)\,\mathcal{K}_i(t) \tag{4.6}$$

in which

$$\varepsilon(t) = y(t) - C\hat{x}(t) \tag{4.7}$$

$$V(t) = CP(t)\,C^T + h^{-1}\,R \tag{4.8}$$

and

$$K(t) = \left(A_t\,P(t)\,C^T + h^{-1}\,S_t\right)\,V^{-1}(t)\quad, \tag{4.9}$$

and for $i = 1, 2, \ldots, p$,

$$\mathcal{K}_i(t) = \left[A_\theta(I_n \otimes e_i)\, P(t)\, C^T + A_t\, \Pi_i(t)\, C^T + S_\theta(I_\ell \otimes e_i) \right] V^{-1}(t) - K(t)\, C\, \Pi_i(t)\, C^T\, V^{-1}(t) \quad . \tag{4.10}$$

Consider a variant of the RPE (4.1) – (4.9) in which eqn. (4.1) is replaced by

$$R(t+h) = R(t) + t^{-1} \left[W^T(t)\, C^T\, V^{-1}(t)\, C W(t) - h R(t) \right] \quad . \tag{4.1A}$$

The solution for $R(t)$ in (4.1) is

$$R(t) = t^{-1} \sum_{\tau=0 \;\; h}^{t} W^T(\tau)\, C^T\, V^{-1}(t)\, C W(t) + R(0) \, , \qquad \tau = h$$

whereas the solution for $R(t)$ in (4.1A) is the same, except that the most recent term, $W^T(t)\, C^T\, V^{-1}(t)\, C W(t)$, is not included in the time average. There is very little difference between eqns. (4.1) and (4.1A) asymptotically as t tends to infinity, because in the proofs of global convergence of the RPE and its variant with (4.1A), the associated ordinary differential equations are exactly the same. This implies that $R(t)$, as defined in (4.1) or (4.1A), approaches the same constant limiting value.

To make the correspondence of the variant of the RPE, eqns. (4.1A), (4.2) – (4.10), and the new algorithm (3.9) – (3.25), identify $\hat{x}(t)$, $\hat{\theta}(t)$, and $P(t)$ above with $\hat{x}(t\,|\,t-h)$, $\hat{\theta}(t\,|\,t)$, and $P_x(t\,|\,t-h)$, respectively. Then the definitions of ε and V in (4.7) and (4.8) correspond to the definitions of ε and V in (3.1) and (3.2). Furthermore, change variables as

$$R(t) = t^{-1}\, P_\theta^{-1}(t\,|\,t-h) \quad , \tag{4.11}$$

$$W^T(t) = t\, R(t)\, P_{\theta x}(t\,|\,t-h) \quad , \tag{4.12}$$

$$\text{vec}\,(\Pi_1(t)\,|\,\ldots\,|\,\Pi_p(t)) = t\,(I_n \otimes V(t) \otimes I_n)\; m(t\,|\,t-h) \quad . \tag{4.13}$$

Combining (3.9) and (3.16), and also (3.10) and (3.17), and using eqns. (4.7), (4.8), (4.9) and (4.11), gives

$$\hat{x}(t+h\,|\,t) \;=\; A_t\,\hat{x}(t\,|\,t-h) + B_t u(t) + K(t)\,\varepsilon(t) + O(t^{-1}) \tag{4.14}$$

$$\hat{\theta}(t+h\,|\,t) \;=\; \hat{\theta}(t\,|\,t-h) + t^{-1}\,R^{-1}(t)\,W^T(t)\,C^T\,V^{-1}(t)\,\varepsilon(t) \tag{4.15}$$

Therefore, the first order moment approximations obeying (3.9), (3.10), (3.16), and (3.17) are within $O(t^{-1})$ of (4.2) and (4.3), respectively.

Now investigate the correspondence of P_θ of (3.13) and (3.18) with (4.1A), using eqn. (4.11). First, notice that (3.18) implies

$$P_\theta(t+h\,|\,t) \;=\; P_\theta(t\,|\,t) \quad . \tag{4.16}$$

Using (4.11), (4.12), and (4.16) above in the unprojected version of (3.13), and suppressing the dependence on t, gives

$$R^{-1}(t+h) = (t+h)\,\left[t^{-1}\,R^{-1} - t^{-2}\,R^{-1}\,W^T\,C^T\,V^{-1}\,CWR^{-1} - P^\varepsilon - P^{\varepsilon\,T}\right] \quad . \tag{4.17}$$

Referring to the definition (3.15) of $P^\varepsilon(t)$, note that $M \sim t^{-1}$ by (4.13), and $P_{\theta x} \sim t^{-1}$ by (4.12), so that $P^\varepsilon(t) \sim t^{-2}$. Therefore, denote $\overline{P}(t) = t^2\,P^\varepsilon(t)$, and rewrite (4.17) above as

$$R^{-1}(t+h) \;=\; R^{-1} - t^{-1}\,R^{-1}\,\left[W^T\,C^T\,V^{-1}\,CW - hR + R\left(\overline{P}+\overline{P}^T\right)R\right]\,R^{-1} + O(t^{-2}) \tag{4.18}$$

To $O(t^{-2})$, the inverse can then be taken, to obtain

$$R(t+h) \;=\; R + t^{-1}\,\left[W^T\,C^T\,V^{-1}\,CW - hR + R\left(\overline{P}+\overline{P}^T\right)R\right] + O(t^{-2}) \tag{4.19}$$

Except for the term $R(\overline{P}+\overline{P}^T)R$, this is to within $O(t^{-2})$ of eqn. (4.1A). To explain the term $R\left(\overline{P}+\overline{P}^T\right)R$, note P^ε, and consequently $\overline{P}$, is driven by the innovations ε. A corresponding innovations-driven term in P_θ is also found in continuous time (Wiberg and DeWolf, 1993). There, an averaging theory argument shows this P term does not affect the asymptotic dynamics of the averaged parameter estimate, and has only a slight effect on the asymptotic dynamics of the averaged value of $R(t)$. The same technique applies in discrete time, and therefore this $\overline{P}$ term can be omitted here for purposes of comparison.

In a similar manner, equations (3.11) and (3.12) with (3.18) can be shown asymptotically equivalent to (4.5) and (4.4), respectively, under the correspondence of $P_x(t\,|\,t-h)$ to $P(t)$ and (4.12). Finally, using the change of variables (4.13), eqns. (3.14) and (3.19) can also be shown to be within $O(t^{-1})$ of (4.6). This shows that the unprojected algorithm (3.9) – (3.19) is within $O(t^{-1})$ of the unprojected state space form of the variant of the RPE (4.1A), (4.2) – (4.10).

5. Proof of Global Convergence

Assume that the $O(t^{-1})$ terms of the previous section are kept finite by projection, and thus decay at least as fast as t^{-1}. Assume that trajectories of the parameter estimates are projected away from any stationary points of the asymptotic negative log likelihood function that are not exponentially stable. Assume that all slow variables are projected into a compact domain. Then, under the same conditions as the RPE (Ljung, 1977), the algorithm (eqns. (3.9) – (3.19)) is globally convergent to a minimum of the asymptotic negative log likelihood function.

Proof is immediate from the preceding section, since (3.9) – (3.19) is asymptotically equivalent to the RPE. The projections away from finite escape and from non-stable equilibrium permit the use of averaging theory. Then the averaged equations of (3.9) – (3.19) and the RPE are identical, and Ljung's (1977) proof holds here, also.

6. Limit as h Tends to Zero

As the sampling interval h tends to zero, the algorithm (3.9) – (3.19) approaches a continuous time algorithm. Here, that continuous time algorithm is shown to be the approximation of the optimal recursive parameter estimator of Wiberg and DeWolf (1993).

To do this for the general case is an algebraic mess. Furthermore, the correlated process/measurement noise case is not developed in continuous time. Consequently, consider the special case of estimating the parameters a and q in the scalar system

$$x(t+h) = a\,x(t) + v(t) \tag{6.1}$$

$$y(t) = x(t) + w(t) \quad , \tag{6.2}$$

where

$$E\left\{v^2(t)\right\} = qh^{-1} \quad ,$$

$$E\left\{\omega^2(t)\right\} = h^{-1} \quad ,$$

and

$$E\left\{v(t)\,w(t)\right\} = 0 \quad .$$

The limiting continuous time system corresponding to (6.1) and (6.2) above as h tends to zero is

$$dx = f\,x\,dt + dv \tag{6.3}$$

$$dy = x\,dt + dw \tag{6.4}$$

in which v and w are Wiener processes having incremental variance σ and 1, respectively, so that $q = \sigma h^2$, and f is defined by $a = \exp(fh)$, so that

$$a = 1 + fh + O(h^2) \quad . \tag{6.5}$$

From Wiberg and DeWolf (1993), the continuous-time approximately optimal parameter for eqns. (6.3) and (6.4) is, for the case m_1 and m_3 set to zero and the process noise incremental variance σ linear in the parameter $\phi^T = (f\ \sigma)$,

$$\begin{aligned}
d\hat{x} &= \left(\hat{f}\,\hat{x} + P_{fx}\right) dt + P_x\, d\varepsilon && (6.6)\\
d\phi &= P_{\phi x}\, d\varepsilon && (6.7)\\
dP_x/dt &= 2\hat{f}\, P_x - P_x^2 + 2\hat{x}\, P_{fx} + \hat{\sigma} + 2m^T e_1 && (6.8)\\
dP_{\phi x} &= \left[(f - P_x)\, P_{\phi x} + \hat{x}\, P_f\, e_1\right] dt + m d\varepsilon && (6.9)\\
dP_\phi/dt &= -P_{\phi x}\, P_{\phi x}^T && (6.10)\\
dm_2\,/\,dt &= 2(f - P_x)\, m_2 + 2\left(P_f\, P_x + P_{fx}^2\right) e_1 + P_\sigma\, e_2 \quad , && (6.11)
\end{aligned}$$

where $e_1^T = (1\ 0)$ and $e_2^T = (0\ 1)$.

Now the discrete time algorithm (3.9) – (3.19) can be written for the system (6.1), (6.2), and the limit taken as h tends to zero. First, from eqns. (3.1) – (3.8), note that in the limit as h tends to zero, for this case,

$$\begin{aligned}
\varepsilon(t) &= d\varepsilon + O(h) && (6.12)\\
V(t) &= h^{-1} + O(1) && (6.13)\\
K_x(t) &= h\, P_x(t\,|\,t-h) + O(h^2) && (6.14)\\
K_\theta(t) &= h\, P_{\theta x}(t\,|\,t-h) + O(h^2) && (6.15)\\
\mathcal{A}(\theta) &= a \;=\; e_1^T\, \theta && (6.16)\\
\mathcal{A}_\theta &= e_1^T && (6.17)\\
\mathcal{B}(x) &= x\, e_1^T && (6.18)\\
A_\xi &= \begin{pmatrix} a & x\, e_1^T \\ 0 & I_2 \end{pmatrix} \quad . && (6.19)
\end{aligned}$$

In continuous time, the incremental variance is σ, implying that at time h the variance is $\sigma h = q h^{-1}$ by (2.3). This fact and (6.5) give

$$\theta = e_1 + [\operatorname{diag}(h, h^2)]\,[\phi + O(h)] \quad . \tag{6.20}$$

For the system (6.1), (6.2), the measurement update (3.9) – (3.14) then becomes, to the leading terms in the h expansion,

$$\hat{x}(t\,|\,t) = \hat{x}(t\,|\,t-h) + h\varepsilon(t)\,P_x(t\,|\,t-h) \tag{6.21}$$

$$\hat{\theta}(t\,|\,t) = \hat{\theta}(t\,|\,t-h) + h\varepsilon(t)\,P_{\theta x}(t\,|\,t-h) \tag{6.22}$$

$$P_x(t\,|\,t) = [1 - h\,P_x(t\,|\,t-h)]\;P_x(t\,|\,t-h) \tag{6.23}$$

$$P_{\theta x}(t\,|\,t) = [P_{\theta x}(t\,|\,t-h) + h\varepsilon(t)\,m(t\,|\,t-h)]\;[1 - h\,P_x(t\,|\,t-h)] \tag{6.24}$$

$$P_\theta(t\,|\,t) = P_\theta(t\,|\,t-h) - h\,P_{\theta x}(t\,|\,t-h)\,P_{\theta x}^T(t\,|\,t-h) \tag{6.25}$$

$$m(t\,|\,t) = [1 - 2h\,P_x(t\,|\,t-h)]\;m(t\,|\,t-h) \tag{6.26}$$

The time update (3.16) – (3.19) becomes, to the leading terms in the h expansion,

$$\hat{x}(t+h\,|\,t) = \hat{a}_t\,\hat{x}_t + P_{ax} \tag{6.27}$$

$$\hat{\theta}(t+h\,|\,t) = \hat{\theta}_t \tag{6.28}$$

$$P_\xi(t+h\,|\,t) = A_{\xi t}\,P_{\xi t}\,A_{\xi t}^T + H(t) + H^T(t) + \left(P_{xt}\,P_{at} + P_{axt}^2 + h^{-1}\hat{q}_t\right)\operatorname{diag}(1,0,0) \tag{6.29}$$

$$m(t+h\,|\,t) = \hat{a}_t^2\,m_t + e_2\,h^{-1}\,P_{qt} + 2\hat{a}_t\left(P_{axt}^2 + P_{at}\,P_{xt}\right)e_1 \quad , \tag{6.30}$$

where
$$H(t) = \begin{pmatrix} \hat{a}_t & \hat{x}_t\,e_1^T \\ 0 & 0 \end{pmatrix} e_1^T\,m_t \quad . \tag{6.31}$$

Using eqn. (6.5) gives $P_{ax} = h\,P_{fx}$, $P_a = h^2\,P_f$, and $m = [\operatorname{diag}(h, h^2)]\,m_2$. Then substitution of (6.21) – (6.26) into (6.27) – (6.30), and use of eqn. (6.20), gives (6.6) – (6.11) in the limit as h tends to zero. Therefore, for this special case, the discrete-time algorithm

(3.9) – (3.19) approaches the approximation of the optimal recursive parameter estimator of Wiberg and DeWolf (1993). This can be shown in general.

7. Simulation

The first example is (6.1) – (6.2) of the previous section, estimating both a and q in a first order system. Figure 1 shows a typical response.

The second example estimates only the state noise variance q in (6.1) – (6.2) for known $a = 0.3$. Also considered is the case where the standard deviation s is estimated, such that $s^2 = q$. In this case, the algorithm not only converges to $\hat{s} = \pm\sqrt{q_0}$, but also very quickly to $\hat{s} = 0$. The value $\hat{s} = 0$ is a local maximum of the likelihood function. Instead, the parameterization in terms of q prevents convergence to $\hat{q} = 0$ and only converges to $\hat{q} = q_0$. Figure 2 shows this behavior for the RPE algorithm, similar to that occuring for (3.9) – (3.19).

The third example shows the effect of using zero as the approximation of the third order moments, $E\{\,\tilde{x}^3(t)\,|\,\mathcal{Y}(t-1)\,\}$ and $E\{\,\tilde{x}(t)\,\tilde{a}^2(t)\,|\,\mathcal{Y}(t-1)\,\}$, as contrasted to retaining their dynamic equations. Figure 3 shows very little difference for the simple example, in which it appears that keeping the extra third order moment dynamic equations is not worthwhile. However, this effect is still under investigation for more complicated cases.

~~A fourth example is intended to be included in the conference proceedings, but is not ready at this time. The fourth example is to compare the algorithm here with other algorithms from MATLAB, to evaluate the improvement in transient response.~~

Presently, the algorithm is being compared with extended least squares (pseudo-linear regression) and constant gain recursive prediction error method in ARMA form. These methods are both found in the MATLAB System Identification Toolbox Results should be available soon.

8. Conclusions

A discrete-time parameter estimator is derived that is globally convergent and that is an approximation of the optimal parameter estimator. It is globally convergent with probability one to stationary points of the asymptotic negative log likelihood function, and appropriate parameterization can assure convergence to true parameter values in many cases. Correlation is permitted between process and measurement noise, so that ARMAX models can be estimated. In the case of significant process noise, the estimator derived has much better transient response than commonly used algorithms.

Much work remains to make the algorithm practical. The update equations, especially for the third order moment, can be significantly compressed. Numerically stable coding should be employed for the propagation of the second order moments. Appropriate projection rules need to be researched.

The class of models can be widened to include parameter dependence in the output map (i.e. θ dependence in C). Forgetting factors, time-~~varying~~ variation and known delay models can be easily incorporated, however. The continuous-time optimal estimator is easily written for the case $C(\theta)$, but an algebraic nightmare results that must be carefully analyzed.

References

[1] Anderson, B.D.O., and J.B. Moore (1979). *Optimal Filtering.* Prentice-Hall, Englewood Cliffs, N.J.

[2] Brewer, J.W. (1978). Kronecker products and matrix calculus in system theory. *IEEE Trans. Circuits and Systems*, **25**, 9, 772-781.

[3] DeWolf, D.G., and D.M. Wiberg (1993). An ordinary differential equation technique for continuous-time parameter estimation. *IEEE Trans. Auto. Control*, **38**, 4, 514-528.

[4] Gevers, M., G.C. Goodwin, and V. Wertz (1991). Continuous-time stochastic adaptive control. *SIAM J. Contr. Optimiz.*, **29**, 2, 264-278.

[5] Henderson, H.V., and S.R. Searle (1981). The vec permutation matrix, the vec operator, and Kronecker products: a review. *Linear and Multilinear Algebra*, **9**, 271-281.

[6] Kramer, S.C., and H.W. Sorensen (1987). Bayesian parameter estimation. *Proc. 1987 Amer. Contr. Conf.*, **2**, 786-790.

[7] Lancaster, P. (1969). *Theory of Matrices*, Ch. 8. Academic Press, New York.

[8] Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Auto. Control*, **22**, 4, 551-575.

[9] Ljung, L. (1987). *System Identification Theory for the User.* Prentice-Hall, Englewood Cliffs, NJ.

[10] Ljung, L. and T. Söderström (1983). *Theory and Practice of Recursive Identification.* M.I.T. Press, Cambridge, MA.

[11] Powell, T.D., D.M. Wiberg, and D. Ljungquist (1993). A counterexample of proof of convergence using asymptotic averaging. *Proc. 1993 Conf. on Decision and Control* (in press).

[12] Wiberg, D.M. (1992). The MIMO discrete-time convergent approximation of the optimal recursive parameter estimator. *Proc. 1992 Am. Contr. Conf.*, **4**, 3355-3359.

[13] Wiberg, D.M. and D.G. DeWolf (1993). A convergent approximation of the continuous-time optimal estimator. *IEEE Trans. Auto. Control*, **38**, 4, 529-545.

Appendix I: KRONECKER PRODUCTS

Kronecker products are the appropriate tool to handle moments of third order (and higher, if necessary), similar to vectors for first order moments (estimates) and matrices for second order moments (Riccati equations). Here, a brief review is given. More complete treatments can be found in Brewer (1978), Henderson and Searle (1981), and Lancaster (1969), although Lancaster uses a different definition of the vec operation. Here, for an $m \times n$ matrix A with column m-vectors $a_1, \ldots, a_n$, $\operatorname{vec} A$ is defined as the mn-vector formed by stacking a_1, then a_2, etc., and finally, a_n on the bottom. Let a_{ik} denote the ik element of A for $i = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, n$. The Kronecker product of $A(m \times n)$ and $B(p \times q)$ is denoted $A \otimes B$, and is an $mp \times nq$ matrix defined by

$$A \otimes B = \begin{pmatrix} a_{11} B \ldots & a_{1n} B \\ \vdots & \vdots \\ a_{m1} B \ldots & a_{mn} B \end{pmatrix} \tag{AI.1}$$

From this definition of $A \otimes B$, it follows that if A is partitioned into A_{11}, A_{12}, A_{21}, and A_{22}, then

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \otimes B = \begin{pmatrix} A_{11} \otimes B & A_{12} \otimes B \\ A_{21} \otimes B & A_{22} \otimes B \end{pmatrix} , \tag{AI.2}$$

but there is no such property for partitions of B.

Furthermore, Brewer (1978) lists (among many others) the following properties:

$$(A \otimes B) \otimes C = A \otimes (B \otimes C) \tag{AI.3}$$

$$(A + H) \otimes (B + R) = A \otimes B + A \otimes R + H \otimes B + H \otimes R \tag{AI.4}$$

$$(A \otimes B)^T = A^T \otimes B^T \tag{AI.5}$$

$$(A \otimes B)\,(D \otimes G) \;=\; (A\,D) \otimes (B\,G) \tag{AI.6}$$

$$(A \otimes B)^{-1} \;=\; A^{-1} \otimes B^{-1} \tag{AI.7}$$

$$\operatorname{vec}(A\,D\,B) \;=\; (B^T \otimes A)\operatorname{vec} D \tag{AI.8}$$

The permutation matrix $U_{m,n}$ is defined by

$$\operatorname{vec} A \;=\; U_{m,n}\,\operatorname{vec} A^T \quad . \tag{AI.9}$$

Explicitly, U can be computed as

$$U_{m,n} \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} e_i\,e_j^T \otimes e_j\,e_i^T \;=\; \sum_{i=1}^{m} e_i^T \otimes I_n \otimes e_i \;=\; \sum_{j=1}^{n} e_j \otimes I_m \otimes e_j^T \quad , \tag{AI.10}$$

where e_i is the unit vector of zeros, except for a "1" in the ith position. Brewer (1978) describes $U_{m,n}$ as "square $(mn \times mn)$ and has precisely a single '1' in each row and in each column." Computationally, sometimes it may be easier to compute $U_{m,n}\operatorname{vec} A^T$ by unstacking $\operatorname{vec} A^T$ to form A^T, taking the transpose, and then forming $\operatorname{vec} A$ rather than using eq. (AI.10) above.

Some properties of $U_{m,n}$ are

$$U_{m,n}^{-1} \;=\; U_{m,n}^T \;=\; U_{n,m} \tag{AI.11}$$

$$U_{m,1} \;=\; U_{1,m} \;=\; I_m \tag{AI.12}$$

$$B \otimes A \;=\; U_{m,p}\,(A \otimes B)\,U_{q,n} \tag{AI.13}$$

This final formula (AI.13) is especially helpful in the derivation in the following Appen- pendix II.

Appendix II: DEVELOPMENT OF THE ALGORITHM

This appendix is based on the work found in Wiberg (1992), with the following exceptions:

(i) In (2.1), $E\left\{v(t)\,v^T(t)\right\} = h^{-1}\,Q(\theta)\,\delta_{t\tau}$ where $Q(\theta)$ is affine in θ, instead of a term $\Sigma(\theta)\,v(t)$ appearing in place of $v(t)$ in (2.1), where $\Sigma(\theta)$ is affine in θ. This creates a unique global minimum in the likelihood function, as explained in Powell et al. (1993), and is the difference between a successful and an unsuccessful algorithm.

(ii) In Wiberg (1992), the third order conditional moments $m_1 = E\left\{\tilde{\theta} \otimes \tilde{\theta} \otimes \tilde{x}\right\}$, $m_2 = E\{\tilde{x} \otimes \tilde{\theta} \otimes \tilde{x}\}$, and $m_3 = E\{\tilde{x} \otimes \tilde{x} \otimes \tilde{x}\}$ are all approximated by dynamic equations. Here, m_1 and m_3 are set to zero, saving significant computation at apparently little cost in performance.

(iii) The model is extended to correlated measurement and process noise. This is easily accomplished by setting

$$\overline{v}(t) = v(t) - S(\theta)R^{-1}\,w(t) \quad , \tag{AII.1}$$

so that $\overline{v}$ and w are uncorrelated in the model. Then the uncorrelated noise algorithm of Wiberg (1992) can be used on the model with process noise $\overline{v}$.

The reparameterization of (i) has only slight effect on the resulting algorithm. Only terms containing $\Sigma(\theta)$ and Σ_θ are affected. Compare eqs. (3.22) and (3.23) with the corresponding equations in Wiberg (1992) numbered there as (38) and (76). Unfortunately, there are a number of algebraic and typographical errors in Wiberg (1992), and it is hoped that these are all corrected here.

The following derivation is merely for purposes of motivation. The algorithm (3.9) – (3.23), regardless of its derivation, has been proven to have desirable properties in Sections 4 and 5. Therefore, the following merely indicates the thought processes involved in its creation.

First, use of the substitution (AII.1) in the model (2.1) gives the new model

$$x(t+h) = \mathcal{A}(\theta)\, x(t) + z(t) + \overline{v}(t) \quad , \tag{AII.2}$$

where $\mathcal{A}(\theta)$ is defined by (3.5), and in which the "input" $z(t)$ is defined by

$$z(t) = B(\theta)\, u(t) + S(\theta)\, R^{-1}\, y(t) \quad . \tag{AII.3}$$

Then following the derivation of the EKF in Anderson and Moore (1979) gives eqns. (3.9), (3.10), and (3.17). This also gives (3.16), except for the term $\mathcal{A}_\theta \ \mathrm{vec}\, P_{\theta x}$. Since (3.16) is the conditional expectation of the right-hand side of (AII.2), substitution of $\theta = \hat{\theta} + \tilde{\theta}$ and $x = \hat{x} + \tilde{x}$ in that expectation shows the appearance of $\mathcal{A}_\theta \ \mathrm{vec}\, P_{\theta x}$.

Having (3.9), (3.10), (3.16), and (3.17), subtract them from the model (AII.2) to obtain the error equations for $\tilde{x}$ and $\tilde{\theta}$. Form the second and third order products, and take conditional expectations to get eqs. (3.11), (3.12), (3.13), (3.14), (3.18), and (3.19). This is not quite straightforward, because moments higher than third need to be approximated by Gaussians to close the moment equations, and because the measurement updates are the best linear unbiased estimator for $P_\xi(t\,|\,t)$ and $m(t\,|\,t)$ in terms of the innovations. Furthermore, the algebra with the Kronecker products gets quite involved. The interested reader is referred to Wiberg (1992) for further details.

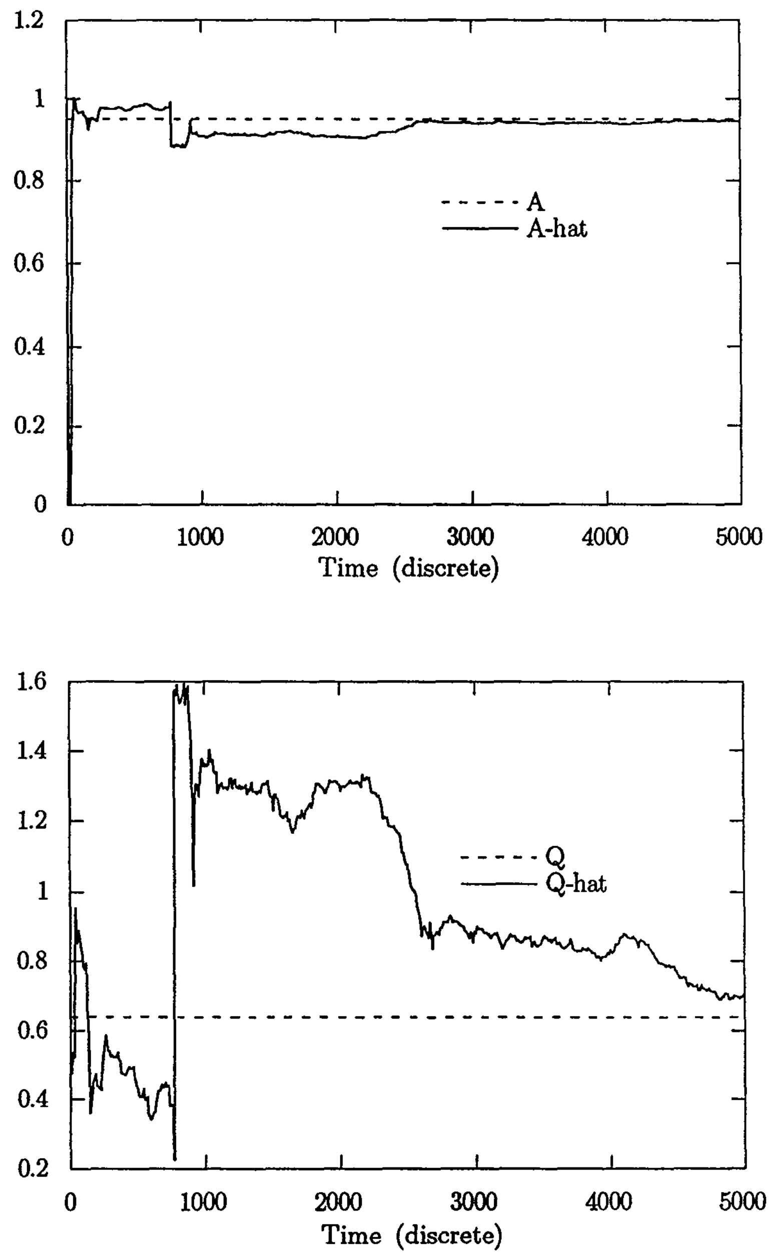

Figure 1: Estimation of both a and q in a first order system.

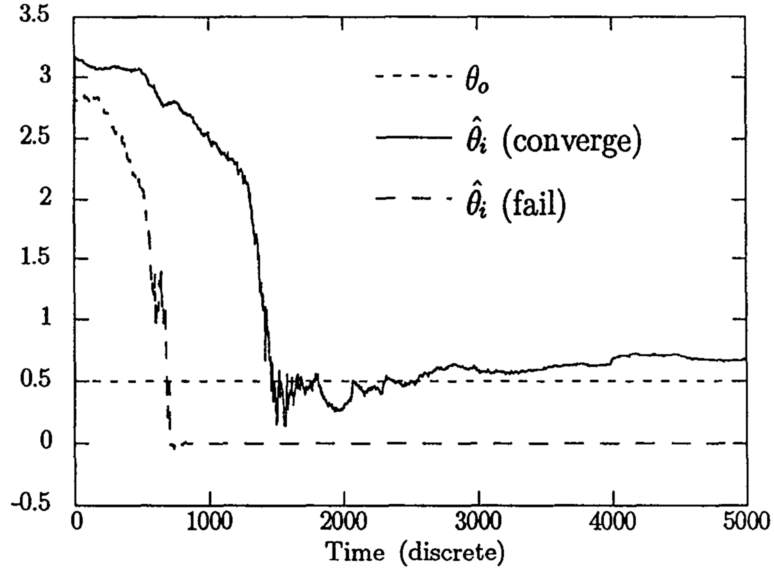

Figure 2: Convergence of $\hat{\theta}_i$ to θ_0, the true value, and of $\hat{\theta}_i$ to 0, a local maximum of the likelihood function for the RPE.

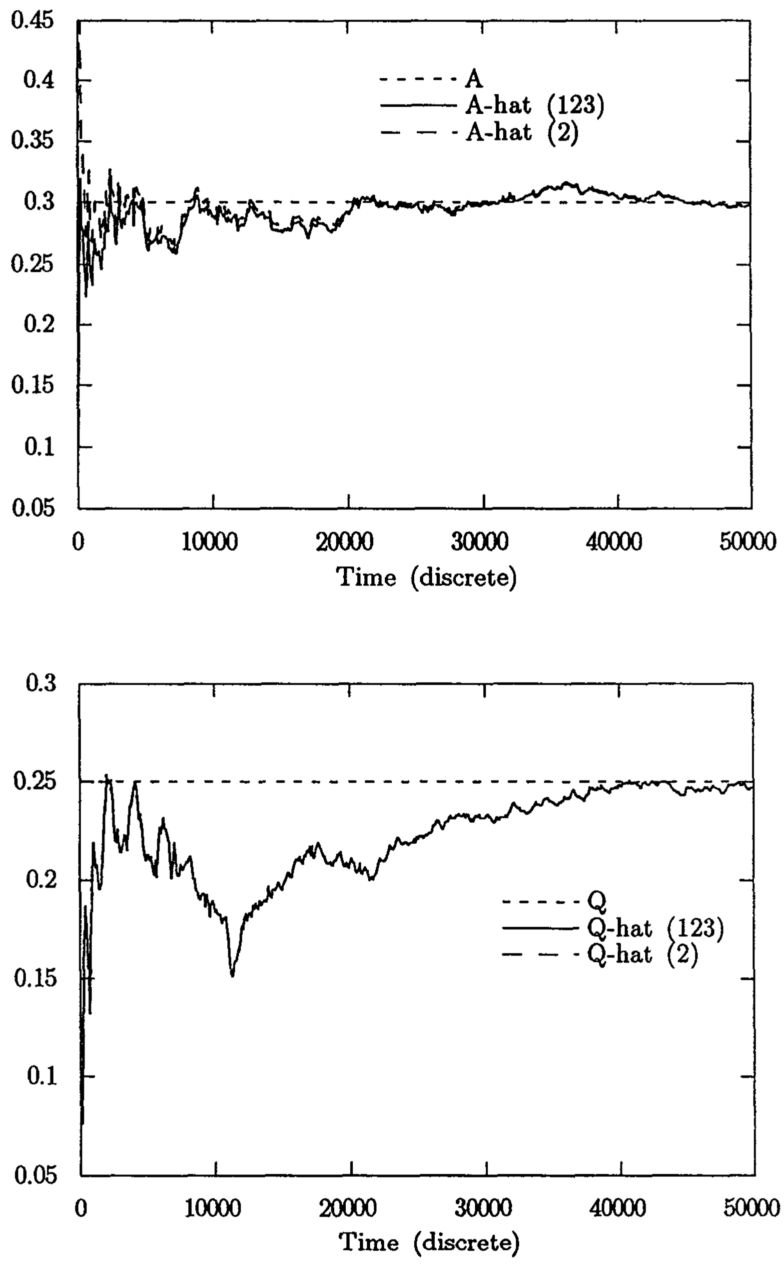

Figure 3: Parameter estimates keeping three third order moments (1 2 3), compared with keeping only one (2).