International Journal of Control
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713393989

Recursive prediction error parameter estimator for non-linear models

S. Chen [a]; S. A. Billings [a]
[a] Department of Control Engineering, University of Sheffield, Sheffield, U.K

Online Publication Date: 01 February 1989
To cite this Article: Chen, S. and Billings, S. A. (1989) 'Recursive prediction error parameter estimator for non-linear models', International Journal of Control, 49:2, 569 - 594
To link to this article: DOI: 10.1080/00207178908961265
URL: http://dx.doi.org/10.1080/00207178908961265

Recursive prediction error parameter estimator for non-linear models

S. CHEN† and S. A. BILLINGS†

A recursive prediction error parameter estimation algorithm is derived for systems which can be represented by the NARMAX (non-linear ARMAX) model. A convergence analysis is presented using the differential equation approach, and the new concept of *m*-invertibility is introduced. The analysis shows that while a highly non-linear process model may be used to capture the non-linearity of the system it is advisable to fit a simple noise model. The results of applying the algorithm to both simulated and real data are included.

1. Introduction

Recursive identification of parameters in linear models is now a well-established field. Several methods of analysing recursive estimators have been proposed and an elegant cohesive theory has been developed (Ljung and Söderström 1983).

In many practical applications, however, non-linear models may be required to achieve an acceptable predictive accuracy. Subject to some mild assumptions the NARMAX model (Billings and Leontaritis 1981, Leontaritis and Billings 1985) can be used as a basis for identification of such systems, and several of the basic principles of linear recursive identification can with obvious interpretations be applied to this model (Billings and Leontaritis 1982, Billings and Voon 1984, Fnaiech and Ljung 1987).

In the present study a recursive prediction error estimator (RPEM) is derived for the polynomial NARMAX model. In order to apply the differential equation approach of convergence analysis developed by Ljung, the filter that generates the prediction should be exponentially stable and for the NARMAX model this coincides with the stability of the noise model. Whilst this is relatively easy to analyse when the noise model is linear, the new concept of *m*-invertibility is introduced for the general case of non-linear noise models. This leads to a convergence analysis of the RPEM for polynomial NARMAX models and to the development of a practical rule for the choice of noise model in non-linear system identification. The rule, which implies that to ensure *m*-invertibility the noise model should not include non-linear terms in the prediction errors, has important implications for all non-linear model fitting algorithms, recursive or batch. The results represent an extension of a previous study (Chen and Billings 1988 b), which considered non-linear output-affine models and, which by definition was therefore restricted to the special case of noise models linear in the prediction errors.

For notational simplicity the single-input single-output case is studied throughout although the results are valid for multi-input multi-output systems. The algorithms are illustrated using both simulated and real data.

Received 3 March 1988.

† Department of Control Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, U.K.

2. NARMAX model

Under some mild assumptions a discrete-time non-linear stochastic control system can be described by the NARMAX model (Leontaritis and Billings 1985)

$$y(t) = f(y(t-1), \dots, y(t-n_y), u(t-1), \dots, u(t-n_u), e(t-1), \dots, e(t-n_e)) + e(t) \tag{1}$$

where $y(t)$, $u(t)$ and $e(t)$ are the system output, input and noise respectively; n_y, n_u and n_e are the orders of the output, input and noise; $\{e(t)\}$ is assumed to be a white sequence; and $f(\cdot)$ is some non-linear function. Expanding $f(\cdot)$ as a polynomial of degree L gives the representation

$$y(t) = \sum_{i=1}^{n_\theta} \theta_i x_i(t) + \varepsilon(t, \theta) \tag{2}$$

where

$$n_\theta = \sum_{i=0}^{L} n_i; \quad n_0 = 1, \quad n_i = n_{i-1}(n_y + n_u + n_e + i - 1)/i, \quad i = 1, \dots, L \tag{3}$$

$$\theta = (\theta_1, \dots, \theta_{n_\theta})^{\mathrm{T}} \tag{4}$$

and

$$x_1(t) = 1$$

$$\left.\begin{aligned} & x_i(t) = \prod_{j=1}^{p} y(t-n_{yj}) \cdot \prod_{k=1}^{q} u(t-n_{uk}) \cdot \prod_{m=1}^{r} \varepsilon(t-n_{em}, \theta) \\ & \quad i = 2, \dots, n_\theta, \quad p, q, r \geqslant 0, \quad 1 \leqslant p+q+r \leqslant L, \quad 1 \leqslant n_{yj} \leqslant n_y \\ & \quad 1 \leqslant n_{uk} \leqslant n_u, \quad 1 \leqslant n_{em} \leqslant n_e \end{aligned}\right\} \tag{5}$$

By convention, $p = 0$, $q = 0$ or $r = 0$ indicates that $x_i(t)$ does not contain $y(\cdot)$ terms, $u(\cdot)$ terms or $\varepsilon(\cdot)$ terms, respectively. As θ ranges over D_M, a subset of $\mathbb{R}^{n_\theta}$, (2) describes the set of models within which the one that best describes the recorded data is to be selected. Denote the input–output record at time $t-1$ as

$$z^{t-1} = (z(t-1), \dots, z(0)) \tag{6}$$

where

$$z(t) = \begin{bmatrix} y(t) \\ u(t) \end{bmatrix} \tag{7}$$

Then for a given θ the one-step-ahead prediction of the output at time t is

$$\hat{y}(t \mid \theta) = g_M(\theta; t, z^{t-1}) = \theta^{\mathrm{T}} \phi_*(t, \theta) \tag{8}$$

where

$$\phi_*(t, \theta) = (x_1(t), \dots, x_{n_\theta}(t))^{\mathrm{T}} \tag{9}$$

The prediction error is given by

$$\varepsilon(t, \theta) = y(t) - \hat{y}(t \mid \theta) \tag{10}$$

The gradient of $\hat{y}(t \mid \theta)$

$$\Psi(t, \theta) = \left[\frac{d\hat{y}(t \mid \theta)}{d\theta}\right]^{\mathrm{T}} \tag{11}$$

an n_θ-dimensional column vector, plays an important role in recursive identification. Differentiating (8) with respect to θ yields

$$\Psi(t, \theta) = \phi_*(t, \theta) - \sum_{i=1}^{n_e} \left[\theta^{\mathrm{T}} \frac{\partial \phi_*(t, \theta)}{\partial \varepsilon(t - i, \theta)} \right] \Psi(t - i, \theta) \tag{12}$$

Regrouping terms in (2) gives

$$\begin{aligned} y(t) &= f^{\mathrm{p}}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u); \theta) \\ &\quad + f^{\mathrm{n}}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u), \varepsilon(t-1, \theta), \ldots, \varepsilon(t-n_e, \theta); \theta) \\ &\quad + \varepsilon(t, \theta) \end{aligned} \tag{13}$$

where $f^{\mathrm{p}}(\cdot)$ contains all terms $\theta_i x_i(t)$ with $r = 0$ and $f^{\mathrm{n}}(\cdot)$ contains all terms $\theta_i x_i(t)$ with $r \neq 0$. $f^{\mathrm{p}}(\cdot)$ is referred to as the process model and $f^{\mathrm{n}}(\cdot)$ as the noise model. A first-order NARMAX model with second-degree non-linearity would for example be given by

$$\begin{aligned} y(t) &= [\theta_1 + \theta_2 y(t-1) + \theta_3 u(t-1) + \theta_4 y^2(t-1) + \theta_5 y(t-1)u(t-1) + \theta_6 u^2(t-1)] \\ &\quad + [\theta_7 \varepsilon(t-1, \theta) + \theta_8 y(t-1)\varepsilon(t-1, \theta) + \theta_9 u(t-1)\varepsilon(t-1, \theta) \\ &\quad + \theta_{10} \varepsilon^2(t-1, \theta)] + \varepsilon(t, \theta) \end{aligned} \tag{14}$$

Notice that unlike the output-affine model (Chen and Billings 1988 b) non-linear power terms in $y(\cdot)$ and $\varepsilon(\cdot)$ are present.

An important case of the model (13) is

$$\begin{aligned} y(t) &= f^{\mathrm{p}}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u); \theta) \\ &\quad + \sum_{i=1}^{n_e} c_i \varepsilon(t-i, \theta) + \varepsilon(t, \theta) \end{aligned} \tag{15}$$

where the c_i coefficients are part of θ. The off-line identification of several industrial systems has shown that many can be modelled in the form of (15). Some examples are a 6996 bhp industrial diesel generator (Billings *et al.* 1988 b), a liquid level system (Billings 1986) and a heat exchanger (Billings and Fadzil 1985, Liu *et al.* 1987). It is obvious that the ARMAX model

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})\varepsilon(t, \theta) \tag{16}$$

where $A(q^{-1})$, $B(q^{-1})$ and $C(q^{-1})$ are the polynomials in the backward shift operator q^{-1}

$$\left.\begin{aligned} A(q^{-1}) &= 1 + \sum_{i=1}^{n_y} a_i q^{-i} \\ B(q^{-1}) &= \sum_{i=1}^{n_u} b_i q^{-i} \\ C(q^{-1}) &= \sum_{i=1}^{n_e} c_i q^{-i} \end{aligned}\right\} \tag{17}$$

is a simple case of (15).

3. Recursive prediction error estimator

A general class of recursive parameter estimators is derived by minimizing the discrepancy between the measured output and the predicted output according to a candidate model (the prediction error) over the model set. The method of deriving these estimators is generally referred to as the recursive prediction error method and its application to linear system identification has been extensively studied (Söderström 1973, Gertler and Bányász 1974, Ljung 1978, 1979, Ljung and Söderström 1983). In this section, the recursive prediction error method is applied to the NARMAX model. A quadratic criterion will be used as an illustration. Extension to the general criterion is obvious.

Based on a recursive minimization of the criterion

$$V(\theta) = E[\varepsilon^2(t, \theta)] \tag{18}$$

a recursive prediction error parameter estimator takes the form

$$\left.\begin{aligned} \varepsilon(t) &= y(t) - \hat{y}(t) \\ R(t) &= R(t-1) + \gamma(t)[\Psi(t)\Psi^{\mathrm{T}}(t) - R(t-1)] \\ \hat{\theta}(t) &= \hat{\theta}(t-1) + \gamma(t)R^{-1}(t)\Psi(t)\varepsilon(t) \end{aligned}\right\} \tag{19}$$

In the algorithm (19) $\Psi(t)\varepsilon(t)$ corresponds to the gradient of the criterion (18) and $R(t)$ is an approximation of the Hessian of the criterion. $R^{-1}(t)\Psi(t)\varepsilon(t)$ is therefore a Gauss–Newton search direction. Other search directions are also feasible, and as long as $R(t)$ is positive definite the convergence properties of the algorithm will not be changed. If the prediction error process is not stationary the criterion can be chosen as

$$\bar{V}(\theta) = \bar{E}[\varepsilon^2(t, \theta)] = \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} E[\varepsilon^2(t, \theta)] \tag{20}$$

Notice that $\varepsilon(t)$, $\hat{y}(t)$ and $\Psi(t)$ depend upon all the old estimates $\hat{\theta}(t-1)$ to $\hat{\theta}(0)$ implicitly.

In practice, the algorithm (19) is implemented in the equivalent form

$$\left.\begin{aligned} \varepsilon(t) &= y(t) - \hat{y}(t) \\ P(t) &= \frac{1}{\lambda(t)}\left[P(t-1) - \frac{P(t-1)\Psi(t)\Psi^{\mathrm{T}}(t)P(t-1)}{\lambda(t) + \Psi^{\mathrm{T}}(t)P(t-1)\Psi(t)}\right] \\ \hat{\theta}(t) &= \hat{\theta}(t-1) + P(t)\Psi(t)\varepsilon(t) \end{aligned}\right\} \tag{21}$$

where

$$P(t) = \gamma(t)R^{-1}(t) \tag{22}$$

$$\lambda(t) = \gamma(t-1)[1 - \gamma(t)]/\gamma(t) \tag{23}$$

For analysis purposes, however, it is better to work with version (19).

It now remains to specify recursions $\hat{y}(t)$ and $\Psi(t)$ for the NARMAX model. From (8) and (12) it is seen that

$$\hat{y}(t) = (\hat{\theta}(t-1))^{\mathrm{T}}\phi*(t) \tag{24}$$

$$\Psi(t) = \phi*(t) - \sum_{i=1}^{n_e}\left[(\hat{\theta}(t-1))^{\mathrm{T}}\frac{\partial\phi*(t)}{\partial\varepsilon(t-i)}\right]\Psi(t-i) \tag{25}$$

where $\phi*(t)$ is obtained by replacing $\varepsilon(t-i, \theta)$ in (9) by $\varepsilon(t-i)$. If the model structure is given by (15) the recursion (25) is often written as

$$\Psi(t) = \frac{1}{\hat{C}(q^{-1})}\phi_*(t) \tag{26}$$

where $\hat{C}(q^{-1}) = 1 + \sum_{i=1}^{n_c} \hat{c}_i(t-1)q^{-i}$. Notice that this is similar to the ARMAX model. As an illustration of recursions (24) and (25), the simple example (14) in § 2 would result in the following definitions:

$$\hat{\theta}(t) = (\hat{\theta}_1(t), \ldots, \hat{\theta}_{10}(t))^{\mathrm{T}}$$

$$\Psi(t) = (\Psi_1(t), \ldots, \Psi_{10}(t))^{\mathrm{T}}$$

$$\begin{aligned}\phi_*(t) = (&1, y(t-1), u(t-1), y^2(t-1), y(t-1)u(t-1), u^2(t-1), \varepsilon(t-1),\\ &y(t-1)\varepsilon(t-1), u(t-1)\varepsilon(t-1), \varepsilon^2(t-1))^{\mathrm{T}}\end{aligned}$$

$$\begin{aligned}\hat{\theta}(t-1)^{\mathrm{T}}\frac{\partial\phi_*(t)}{\partial\varepsilon(t-1)} &= \hat{\theta}_7(t-1) + \hat{\theta}_8(t-1)y(t-1) + \hat{\theta}_9(t-1)u(t-1)\\ &\quad + 2\hat{\theta}_{10}(t-1)\varepsilon(t-1)\end{aligned}$$

4. Convergence analysis

4.1. *Results for the linear model*

In linear system identification it is assumed that the prediction $\hat{y}(t\,|\,\theta)$ is obtained by filtering the input–output data through a linear finite-dimensional filter

$$\left.\begin{aligned}\phi(t+1, \theta) &= F(\theta)\phi(t, \theta) + G(\theta)z(t)\\ \hat{y}(t\,|\,\theta) &= H(\theta)\phi(t, \theta)\end{aligned}\right\} \tag{27}$$

where $\phi(t, \theta)$ is an n-dimensional vector, $F(\theta)$ and $G(\theta)$ are matrices of appropriate dimensions. The stability region of the predictor (27) is

$$D_s = \{\theta\,|\,F(\theta) \text{ has all eigenvalues inside the unit circle}\} \tag{28}$$

The predictor should be constrained to be stable. Therefore it is necessary to require that $D_M \subset D_s$. Notice that D_s is not the stability region for the system dynamics, and that constraining θ to D_s does not impose a serious restriction on the model. All linear models can be written in the form of (27). It can easily be shown that for the ARMAX model (16)

$$D_s = \{\theta\,|\,C^*(s) \text{ has all zeros inside the unit circle}\} \tag{29}$$

where

$$C^*(s) = s^{n_c}C(s^{-1}) \tag{30}$$

Before quoting the analysis results from Ljung (1977, 1979), Ljung and Söderström (1983), it should be emphasized that the same analysis results hold for the predictor model

$$\left.\begin{aligned}\phi(t+1, \theta) &= F(\theta)\phi(t, \theta) + G(\theta)h(z(t))\\ \hat{y}(t\,|\,\theta) &= H(\theta)\phi(t, \theta)\end{aligned}\right\} \tag{31}$$

where $h(\cdot)$ is some vector-valued function. This is apparent in the proof of Theorem 1 in Ljung (1977). In the analysis it will be more convenient to include $\Psi(t, \theta)$ as the filter's output. Differentiating (27) with respect to θ and rearranging the resulting equations yields

$$\left.\begin{aligned} \xi(t+1, \theta) &= \tilde{F}(\theta)\xi(t, \theta) + \tilde{G}(\theta)z(t) \\ \begin{bmatrix} \hat{y}(t \mid \theta) \\ \Psi(t, \theta) \end{bmatrix} &= \tilde{H}(\theta)\xi(t, \theta) \end{aligned}\right\} \tag{32}$$

where

$$\xi(t, \theta) = \left[\phi^{\mathrm{T}}(t, \theta), \left[\frac{\partial \phi(t, \theta)}{\partial \theta_1}\right]^{\mathrm{T}}, \ldots, \left[\frac{\partial \phi(t, \theta)}{\partial \theta_{n_\theta}}\right]^{\mathrm{T}}\right]^{\mathrm{T}} \tag{33}$$

and $\tilde{F}(\theta)$ has the same eigenvalues as $F(\theta)$ but with higher multiplicities. The following results are from Ljung and Söderström (1983).

Consider the algorithm (19) with the recursion

$$\left.\begin{aligned} \xi(t+1) &= \tilde{F}(\hat{\theta}(t))\xi(t) + \tilde{G}(\hat{\theta}(t))z(t) \\ \begin{bmatrix} \hat{y}(t+1) \\ \Psi(t+1) \end{bmatrix} &= \tilde{H}(\hat{\theta}(t))\xi(t+1) \end{aligned}\right\} \tag{34}$$

Assume:

(*a*) D_{M} is a compact subset of $\mathbb{R}^{n_\theta}$ and $D_{\mathrm{M}} \subset D_s$.

(*b*) The matrices $\tilde{F}(\theta)$, $\tilde{G}(\theta)$ and $\tilde{H}(\theta)$ are continuously differentiable with respect to θ for $\theta \in D_{\mathrm{M}}$.

(*c*) $\lim_{t \to \infty} t \cdot \gamma(t) = \mu > 0$.

(*d*) $R(t) \geqslant \delta I \quad \forall t$ for some $\delta > 0$.

(*e*) A projection is included into the algorithm to keep $\hat{\theta}(t)$ inside D_{M}. That is,

$$\hat{\theta}(t) = [\hat{\theta}(t-1) + \gamma(t)R^{-1}(t)\Psi(t)\varepsilon(t)]_{D_{\mathrm{M}}} \tag{35}$$

with

$$[x]_{D_{\mathrm{M}}} = \begin{cases} x \text{ if } x \in D_{\mathrm{M}} \\ \text{a value strictly interior to } D_{\mathrm{M}} \text{ otherwise} \end{cases} \tag{36}$$

(*f*) The data generation is asymptotically mean stationary so that $\bar{V}(\theta)$ defined in (20) and

$$\left.\begin{aligned} \bar{E}\Psi(t, \theta)\varepsilon(t, \theta) &= f_D(\theta) \\ \bar{E}\Psi(t, \theta)\Psi^{\mathrm{T}}(t, \theta) &= G_D(\theta) \end{aligned}\right\} \tag{37}$$

exist, where as in (20)

$$\bar{E}(\cdot) = \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} E(\cdot) \tag{38}$$

and the expectation is over the stochastic process $\{z(t)\}$.

(*g*) The data generation is exponentially stable. That is, for each $t, t_1, t \geqslant t_1$, there exists a random vector $z_{t_1}^{0}(t)$ that belongs to the σ-algebra generated by z^t but is independent of z^{t_1}, such that

$$E\|z(t) - z^0_{t_1}(t)\|^4 < c\lambda^{t-t_1}, \quad c < \infty, \quad \lambda < 1$$

where $\|\cdot\|$ is a chosen norm.

Then $\{\hat\theta(t)\}$ converges with probability 1 to a local minimum of $\bar V(\theta)$.

The analysis of the algorithm is based on the associated differential equation (d.e.)

$$\left.\begin{aligned} \frac{d\theta(\tau)}{d\tau} &= R_D^{-1}(\tau) f_D(\theta(\tau)) \\ \frac{dR_D(\tau)}{d\tau} &= G_D(\theta(\tau)) - R_D(\tau) \end{aligned}\right\} \tag{39}$$

This d.e. is defined in the area D_s. The results mean that the recursive prediction error algorithm has the same convergence properties as its corresponding off-line algorithm. Nothing is assumed about the true system other than that the data generation is stable and asymptotically mean stationary (conditions (g) and (f)). The true system may be much more complex than the resulting model, but this model is the best approximation to the system within the model set in terms of the chosen criterion. The only situation where it is not realistic to assume conditions (f) and (g) *a priori* is when the generation of $\{z(t)\}$ depends upon past estimates such as in adaptive control.

To ensure condition (d) the computation of $R(t)$ can be modified to

$$\left.\begin{aligned} \bar R(t) &= R(t-1) + \gamma(t)[\Psi(t)\Psi^T(t) - R(t-1)] \\ R(t) &= \begin{cases} \bar R(t) & \text{if } \bar R(t) \geqslant \delta I \\ \bar R(t) + M_\delta(t) & \text{otherwise} \end{cases} \end{aligned}\right\} \tag{40}$$

where $M_\delta(t)$ is chosen so that $R(t) \geqslant \delta I$. This modification can easily be implemented. Many projection rules are possible. One example is

$$\left.\begin{aligned} \bar\theta(t) &= \hat\theta(t-1) + \gamma(t)R^{-1}(t)\Psi(t)\varepsilon(t) \\ \hat\theta(t) &= \begin{cases} \bar\theta(t) & \text{if } \bar\theta(t) \in D_M \\ \hat\theta(t-1) & \text{otherwise} \end{cases} \end{aligned}\right\} \tag{41}$$

For a linear model, D_s is usually known and there is no difficulty in incorporating a projection mechanism within the algorithm. Take the ARMAX model (16), for example; once n_e is given D_s is specified by (29) which can be checked by testing if the C-polynomial is stable at each stage of the estimation using a Routh scheme. If it is not, $\hat\theta(t)$ can be projected into the interior of D_s.

4.2. *Application to the NARMAX model with linear noise model*

An important class of NARMAX models are given by (15), where the noise model $f^n(\cdot)$ is linear. The formulations of $\hat y(t|\theta)$ and $\Psi(t, \theta)$ for this class of non-linear models are similar to those for the linear case. In fact, a non-linear model of (15) can always be rearranged into a predictor form

$$\left.\begin{aligned} \phi(t+1, \theta) &= F(\theta; z(t))\phi(t, \theta) + G(\theta)h(z(t)) \\ \hat y(t|\theta) &= H(\theta)\phi(t, \theta) \end{aligned}\right\} \tag{42}$$

where the eigenvalues of the matrix $F(\theta;\, x(t))$ are 0 (with multiplicity) and the zeros of the polynomial $C^*(s)$ given in (30). Notice that although $F(\theta;\, z(t))$ depends on $z(t)$ its eigenvalues do not. This is best illustrated using an example:

$$\begin{aligned} y(t) &= \theta_1 y(t-1) + \theta_2 y(t-2) + \theta_3 u(t-1) + \theta_4 y^3(t-1) \\ &\quad + \theta_5 y(t-1)\underline{y(t-2)u(t-2)} + \theta_6 y^2(t-1)u(t-1)u(t-2) \\ &\quad + \theta_7 \varepsilon(t-1,\theta) + \theta_8 \varepsilon(t-2,\theta) + \varepsilon(t,\theta) \end{aligned}$$

Define

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \\ \theta_8 \end{bmatrix}, \quad \phi(t,\theta) = \begin{bmatrix} y(t-1) \\ y(t-2) \\ u(t-1) \\ \underline{y(t-1)u(t-1)} \\ y^3(t-1) \\ y(t-1)y(t-2)u(t-2) \\ y^2(t-1)u(t-1)u(t-2) \\ \varepsilon(t-1,\theta) \\ \varepsilon(t-2,\theta) \end{bmatrix}, \quad h(z(t)) = \begin{bmatrix} y(t) \\ u(t) \\ y(t)u(t) \\ y^3(t) \end{bmatrix}$$

Then

$$F(\theta;\, z(t)) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & y(t) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y^2(t)u(t) & 0 & 0 & 0 & 0 & 0 & 0 \\ -\theta_1 & -\theta_2 & -\theta_3 & 0 & -\theta_4 & -\theta_5 & -\theta_6 & -\theta_7 & -\theta_8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

$$G(\theta) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and $H(\theta) = (\theta_1 \quad \theta_2 \quad \theta_3 \quad 0 \quad \theta_4 \quad \theta_5 \quad \theta_6 \quad \theta_7 \quad \theta_8)$.

The eigenvalues of $F(\theta; z(t))$ are 0 (with multiplicity 7) and the two zeros of $C^*(s) = s^2 + \theta_7 s + \theta_8$. The underlines in the above example indicate that a 'trick' has been used. The term $y(t)y(t-1)u(t-1)$, which is the one-step-ahead state of $y(t-1)y(t-2)u(t-2)$, can be represented as the product of $y(t)$ and $y(t-1)u(t-1)$, and hence, $y(t-1)u(t-1)$ is included as a 'state'. If the model includes all the possible terms, this kind of trick will not be needed. For example, representing $y^2(t)u(t)u(t-1)$ as the product of $y^2(t)u(t)$ and $u(t-1)$ does not require the introduction of any new term because $u(t-1)$ is already an element of $\phi(t, \theta)$.

Similar to the linear case, the augmented predictor model for NARMAX models of the form of (15) is

$$\left.\begin{aligned} \xi(t+1, \theta) &= \tilde{F}(\theta; z(t))\xi(t, \theta) + \tilde{G}(\theta)h(z(t)) \\ \begin{bmatrix} \hat{y}(t \mid \theta) \\ \Psi(t, \theta) \end{bmatrix} &= \tilde{H}(\theta)\xi(t, \theta) \end{aligned}\right\} \tag{43}$$

The stability region D_s of this predictor coincides with (29). It now becomes clear that analysis of the recursive prediction error estimator for NARMAX models of the form of (15) can follow the exact lines given in Ljung (1977), Ljung and Söderström (1983), and the previous convergence results for the linear case can be applied directly.

4.3. *Invertibility of noise models*

In a general NARMAX model, the noise is multiplicative with the input and output, and the convergence analysis for the linear case does not apply directly. It is important therefore to investigate if it is possible to extend the previous analysis to NARMAX models with a more complex noise structure $f^n(\cdot)$, and this leads to a new definition of invertibility called *m* (model) invertibility.

Before introducing the concept of *m*-invertibility the ideas of the differential equation method are briefly discussed. Since, for $\bar{\theta} \in D_M$, $\tilde{F}(\bar{\theta})$ is exponentially stable, exponential stability of the time-varying difference equation (34) will be guaranteed if $\hat{\theta}(k)$ varies in a sufficiently small neighbourhood of $\bar{\theta}$, and for sufficiently large t and some M, the influence of $\hat{\theta}(k)$, $k = t - M - 1, \ldots, 0$ becomes very small, that is,

$$\xi(t) = \xi(t, \hat{\theta}(t-1), \ldots, \hat{\theta}(0)) \approx \xi(t, \hat{\theta}(t-1), \ldots, \hat{\theta}(t-M)) \tag{44}$$

Furthermore, because $\gamma(t) \to 0$ as $t \to \infty$, for sufficiently large t, $\gamma(t)$ will be arbitrarily small. From the algorithm (19), it is seen that $\{\hat{\theta}(t)\}$ will change more and more slowly, and

$$\hat{\theta}(t-1) \approx \ldots \approx \hat{\theta}(t-M) \approx \bar{\theta} \tag{45}$$

As a consequence, the time-varying difference equation (34) behaves more and more like the time-invariant difference equation (32), and problems like convergence with probability 1, possible convergence points and asymptotic behaviour of the recursive algorithm can thus be studied in terms of the associated differential equation (39) (Ljung 1977). In summary, the stability of the predictor (27) is vital for the analysis using the associated differential equation.

Consider specifically the ARMAX model (16). The stability of the predictor requires that $C(q^{-1})$ is stable, that is, $C^*(s)$ has all zeros inside the unit circle. This is often referred to as $C(q^{-1})$ being invertible in the literature of time series analysis and stochastic control. It is not difficult to see why the convergence results for the NARMAX model with linear noise terms (15) is the same as that of the ARMAX model; they both have the same noise model structure. Furthermore, for NARMAX models where the noise model has an ARMA structure

$$f^n(\cdot) = \frac{C(q^{-1})}{D(q^{-1})}\varepsilon(t, \theta) \tag{46}$$

the convergence analysis for the linear case can again be applied directly. An investigation into the consequences of $C(q^{-1})$ being invertible leads to the results we seek. For a particular realization of the stochastic process $\{z(t)\}$, assume that two sequences $\{\varepsilon^{(i)}(t, \theta)\}$, $i = 1, 2$, are generated by

$$\varepsilon^{(i)}(t, \theta) = (1 - C(q^{-1}))\varepsilon^{(i)}(t, \theta) + A(q^{-1})y(t) - B(q^{-1})u(t) \tag{47}$$

with any two different initial conditions

$$\varepsilon^{(i)}(0), \ldots, \varepsilon^{(i)}(-n_e + 1), \quad i = 1, 2 \tag{48}$$

Then

$$[\varepsilon^{(1)}(t, \theta) - \varepsilon^{(2)}(t, \theta)]^2 \to 0 \quad \text{as } t \to \infty \tag{49}$$

because the influence of initial conditions decays exponentially. For almost all realizations of $\{z(t)\}$, (49) will hold with probability 1. This suggests a new definition of invertibility called m-invertibility that will also cover the general non-linear model. The concept is similar to that introduced by Granger and Andersen (1978) for time series analysis but the two definitions are not the same.

Definition: m-*invertibility*

Assume that the non-linear model (1) has been parametrized (not necessarily using a polynomial expansion) with a parameter vector θ. For given observations of $\{z(t)\}$, let $\{\varepsilon^{(i)}(t, \theta)\}$, $i = 1, 2$, be generated by

$$\begin{aligned}\varepsilon^{(i)}(t, \theta) = y(t) - f(t(t-1), \ldots, y(t-n_y), u(t-1), \ldots,\\ u(t-n_u), \varepsilon^{(i)}(t-1, \theta), \ldots, \varepsilon^{(i)}(t-n_e, \theta)), \quad i = 1, 2\end{aligned} \tag{50}$$

with initial conditions given by (48). Then model (1) is said to be m-invertible if

$$E[\Delta\varepsilon(t, \theta)]^2 = E[\varepsilon^{(1)}(t, \theta) - \varepsilon^{(2)}(t, \theta)]^2 \to 0 \quad \text{as } t \to \infty \tag{51}$$

Condition (51) guarantees that the two sequences $\varepsilon^{(i)}(t, \theta)$, $i = 1, 2$ become identical as $t \to \infty$ with probability 1 regardless of their initial conditions. If the true system is exactly described by the model, the above definition of invertibility coincides with that given by Granger and Andersen (1978), and (51) becomes

$$E[e(t) - \varepsilon(t, \theta)]^2 \to 0 \quad \text{as } t \to \infty \tag{52}$$

which implies that very good estimates of the unobserved system noise can be obtained at least for large t. Notice, however, that given an underlying process $\{z(t)\} = \{(y(t), u(t))^{\mathrm{T}}\}$ and a model parameterized by θ (which may not be anything related to $\{z(t)\}$) then m-invertibility says that if the generation of $\varepsilon(t, \theta)$ is stable then the model is invertible. This is easier to interpret if $\varepsilon(t, \theta)$ is thought of as the output of a non-linear finite-dimensional filter

$$\left.\begin{aligned}\tilde{\phi}(t+1, \theta) &= \tilde{f}(\theta; \tilde{\phi}(t, \theta), z(t))\\ \varepsilon(t, \theta) &= \tilde{h}(\theta; \tilde{\phi}(t, \theta))\end{aligned}\right\} \tag{53}$$

m-Invertibility means that this non-linear filter is exponentially stable. For the NARMAX model (13), it is clear that m-invertibility is a property of the noise model $f^{n}(\cdot)$ only. This property is concerned with the stability of the noise model.

Although many practical systems can be modelled in the form of (15) it is unlikely that the linear noise model $f^{n}(\cdot)$ has sufficient generality. A convenient extension to the model in (15) is the NARMAX model with a bilinear form for $f^{n}(\cdot)$

$$\begin{aligned} y(t) = f^{p}(y(t-1), \dots, y(t-n_y), u(t-1), \dots, u(t-n_u); \theta) + \sum_{i=1}^{n_e} c_i \varepsilon(t-i, \theta) \\ + \sum_{i=1}^{n_u} \sum_{j=1}^{n_e} d_{ij} u(t-i) \varepsilon(t-j, \theta) + \sum_{i=1}^{n_y} \sum_{j=1}^{n_e} h_{ij} y(t-i) \varepsilon(t-j, \theta) + \varepsilon(t, \theta) \end{aligned} \tag{54}$$

The invertibility conditions for certain types of bilinear time series models have been investigated (e.g. Granger and Andersen 1978, Quinn 1982, Subba Rao and Gabr 1984). These results can be extended to NARMAX models with similar bilinear forms of noise model $f^{n}(\cdot)$. Consider for example a simple case of model (54):

$$\begin{aligned} y(t) = f^{p}(y(t-1), \dots, y(t-n_y), u(t-1), \dots, u(t-n_u); \theta) \\ + (\alpha + \beta u(t-1)) \varepsilon(t-1, \theta) + \varepsilon(t, \theta) \end{aligned} \tag{55}$$

Then

$$\Delta\varepsilon(t, \theta) = -(\alpha + \beta u(t-1)) \Delta\varepsilon(t-1, \theta) \quad \text{with initial condition } \Delta\varepsilon(0) \tag{56}$$

or

$$\Delta\varepsilon(t, \theta) = (-1)^{t} \Delta\varepsilon(0)(\alpha + \beta u(0)) \cdot \prod_{i=1}^{t-1} (\alpha + \beta u(i)) \tag{57}$$

Using the same analysis presented by Granger and Andersen (1978) for the time series model

$$y(t) = \beta y(t-1) e(t-1) + e(t) \tag{58}$$

it can be shown that

$$(\Delta\varepsilon(t+1, 0))^2 = \rho \cdot \prod_{i=1}^{t} (\alpha + \beta u(i))^2 \leqslant \rho \left[\frac{1}{t} \sum_{i=1}^{t} (\alpha + \beta u(i))^2 \right]^{t} \tag{59}$$

where $\rho = (\Delta\varepsilon(0))^2 (\alpha + \beta u(0))^2$. If

$$\lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} (\alpha + \beta u(t))^2 \tag{60}$$

exists w.p.1 and this limit is less than 1, then clearly

$$E[\Delta\varepsilon(t, \theta)]^2 \to 0 \quad \text{as } t \to \infty \tag{61}$$

Denote a function

$$g(t, \theta, z(t)) = g_t \tag{62}$$

where $g(\cdot)$ is differentiable with respect to θ and $z(t)$. It is known that

$$\lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} g_t = \bar{g} \quad \text{exists w.p.1} \tag{63}$$

if the following two conditions are satisfied:

$$\frac{1}{N}\sum_{t=1}^{N}(g_t - E[g_t]) \to 0 \quad \text{w.p.1 as } N \to \infty \tag{64}$$

$$\frac{1}{N}\sum_{t=1}^{N} E[g_t] \to \bar{g} \quad \text{as } N \to \infty \tag{65}$$

(see, for example, Ljung and Söderström 1983, Section 4.3.4). Condition (64) is very mild and can be assumed to hold (it corresponds to assumption (g) in Section 4.1). This leads to the conclusion that a sufficient condition for model (55) to be m-invertible is

$$\bar{E}[(\alpha + \beta u(t))^2] = \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N} E[(\alpha + \beta u(t))^2] < 1 \tag{66}$$

If $\{u(t)\}$ is stationary, $E[(\alpha + \beta u(t))^2]$ is independent of t, and condition (66) becomes

$$E[(\alpha + \beta u(t))^2] < 1 \tag{67}$$

Notice that this condition depends upon the input as well as the parameters.

Model (54) is a special case of the models in which $f^{n}(\cdot)$ is linear in $\varepsilon(t-i, \theta)$, $1 \leqslant i \leqslant n_e$:

$$\begin{aligned} y(t) &= f^{p}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u); \theta) \\ &\quad + \tilde{C}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u); q^{-1})\varepsilon(t, \theta) \end{aligned} \tag{68}$$

where

$$\begin{aligned} &\tilde{C}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u); q^{-1}) \\ &\qquad = 1 + \sum_{i=1}^{n_e} \tilde{c}_i(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u))q^{-i} \end{aligned} \tag{69}$$

and the $\tilde{c}_i(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u))$ are polynomials. The analysis developed for bilinear $f^{n}(\cdot)$ can be applied to this class of models. For example, consider the model

$$y(t) = f^{p}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u); \theta) + \alpha y^2(t-1)\varepsilon(t-1, \theta) + \varepsilon(t, \theta) \tag{70}$$

Using a similar argument as that used for (55), a sufficient m-invertibility condition is derived:

$$\alpha^2 \bar{E}[y^4(t)] < 1 \tag{71}$$

This condition depends upon the statistical properties of $\{z(t)\}$ as well as the parameter α.

If the power of $\varepsilon(t-i, \theta)$, $1 \leqslant i \leqslant n_e$, in $f^{n}(\cdot)$ of (13) is raised to a value larger than 1 it is unlikely to produce an invertible model. To demonstrate this, assume that $y(t)$ was generated by

$$y(t) = f^{p}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u); \theta) + \alpha e^2(t-1) + e(t) \tag{72}$$

Substituting $e(t)$ by $\varepsilon(t, \theta) + \Delta\varepsilon(t, \theta)$ and using

$$\varepsilon(t, \theta) = y(t) - f^{p}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u); \theta) - \alpha\varepsilon^2(t-1, \theta) \tag{73}$$

yields

$$\Delta\varepsilon(t,\theta) = -2\alpha\varepsilon(t-1,\theta)\Delta\varepsilon(t-1,\theta) - \alpha(\Delta\varepsilon(t-1,\theta))^2 \tag{74}$$

As pointed out by Granger and Andersen (1978), the solution of (74) has an explosive component. It follows that $\{\varepsilon(t,\theta)\}$ generated by (73) will diverge from $\{e(t)\}$ and thus the model (72) is non-invertible. Notice that this does not mean that systems like (72) do not exist in reality. It does imply, however, that attempts to fit non-invertible models may lead to explosive prediction errors.

So far only the invertibility conditions for polynomial NARMAX models have been discussed. Other non-linear parametric models can however be studied in a similar manner. The analysis for the non-linear output-affine model (Chen and Billings 1988 b)

$$y(t) = \frac{\sum_{i=1}^{r} \tilde{a}_i(u(t-1),\dots,u(t-r))y(t-i) + \tilde{a}_{r+1}(u(t-1),\dots,u(t-r))}{\tilde{a}_{2r+2}(u(t-1),\dots,u(t-r))} + \frac{\sum_{i=r+2}^{2r+1} \tilde{a}_i(u(t-1),\dots,u(t-r))\varepsilon(t-i+r+1,\theta)}{\tilde{a}_{2r+2}(u(t-1),\dots,u(t-r))} + \varepsilon(t,\theta) \tag{75}$$

where $\tilde{a}_i(\cdot)$, $i = 1,\dots,2r+2$ are polynomials of degree L which, for example, can follow the same lines for the model in (68) because both models are linear in $\varepsilon(\cdot)$. Consider a more general non-linear parametric model

$$y(t) = \frac{\tilde{b}(y(t-1),\dots,y(t-n_y),u(t-1),\dots,u(t-n_u),\varepsilon(t-1,\theta),\dots,\varepsilon(t-n_\varepsilon,\theta))}{\tilde{a}(y(t-1),\dots,y(t-n_y),u(t-1),\dots,u(t-n_u),\varepsilon(t-1,\theta),\dots,\varepsilon(t-n_\varepsilon,\theta))} + \varepsilon(t,\theta) \tag{76}$$

where $\tilde{b}(\cdot)$ and $\tilde{a}(\cdot)$ are polynomials of degrees L_1 and L_2, respectively. This model is referred to as the non-linear rational model in Billings and Chen (1989). The analysis for model (76) is more complex but the techniques used will be the same. For example

$$y(t) = \frac{0{\cdot}9 + \varepsilon^2(t-1,\theta)}{1 + y^2(t-1) + \varepsilon^2(t-1,\theta)}\varepsilon(t-1,\theta) + \varepsilon(t,\theta) \tag{77}$$

is m-invertible because

$$0 < \frac{0{\cdot}9 + \varepsilon^2(t-1,\theta)}{1 + y^2(t-1) + \varepsilon^2(t-1,\theta)} < 1 \tag{78}$$

but

$$y(t) = \frac{0{\cdot}9 + \varepsilon^2(t-1,\theta)}{1 + y^2(t-1) + \varepsilon^2(t-1,\theta)}\varepsilon^2(t-1,\theta) + \varepsilon(t,\theta) \tag{79}$$

is non-invertible and this can be verified using a procedure similar to that for analysing model (72).

4.4. *Extension of the convergence analysis to more complex noise models*

As discussed above, the feasibility of the associated d.e. approach depends upon the stability of the predictor model. For a NARMAX model (13), the stability of the predictor coincides with the stability of the noise model $f^n(\cdot)$. The analysis of § 4.3 reveals that a general NARMAX model may not always give a stable predictor in the sense that its noise model may not always be m-invertible. If the model is restricted to be of the form of (68), however, it is possible to extend the associated d.e. approach for convergence analysis of the recursive estimator so that it applies to this case. Define

$$D_s = \{\theta \mid \tilde{C}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u); q^{-1}) \text{ is } m\text{-invertible}\} \quad (80)$$

The analysis results for the linear case can be extended to the model (68) at least in principle.

D_s now depends upon θ as well as the statistical properties of $\{z(t)\}$. For non-adaptive control, given an underlying process the statistical properties of $\{z(t)\}$ are fixed by the experimental conditions and D_s can then be viewed as depending upon θ only. The difficulty is that in general the exact shape of D_s may not be known even though it exists. Without knowing the range of D_s it may not be possible to implement some sort of projection mechanism to ensure $\hat{\theta}(t) \in D_s$.

In some simple situations D_s can however be written down and the stability of the estimated $\hat{\tilde{C}}(\cdot; q^{-1})$ can be texted. Consider model (55) again. Assume that $u(t)$ is generated in open loop with $E[u(t)] = 0$ and $E[u^2(t)] = \sigma_u^2$. Then

$$D_s = \{\theta \mid \alpha^2 + \beta^2 \sigma_u^2 < 1\} \quad (81)$$

Compare this with a first-order MA noise model $C(q^{-1}) = 1 + \alpha q^{-1}$, whose D_s is

$$D_s = \{\theta \mid -1 < \alpha < 1\} \quad (82)$$

The implementation of a projection rule in such a situation is straightforward.

In the original theorem (Ljung 1977) it is only required that $\hat{\theta}(t)$ belongs to D_s infinitely often with probability 1. A projection is included to guarantee this boundedness condition. Many recursive algorithms nevertheless work well in practice without incorporating some kind of projection mechanism. The situation would however become serious if a stable region D_s did not exist. It is therefore necessary that the noise model is not too complex. By this it is meant that the noise model $f^n(\cdot)$ in (13) should be linear in $\varepsilon(\cdot)$. The noise model can however be non-linear in the input and the output, even non-linear in the parameters. If the noise model is allowed to be as general as possible the associated d.e. approach will not be applicable.

5. Pseudo-linear regression

Another class of recursive parameter estimators is based on the pseudo-linear regression. If the prediction can be written as

$$\hat{y}(t \mid \theta) = \theta^{\mathrm{T}} \phi(t, \theta) \quad (83)$$

a pseudo-linear regression estimator is given as

$$\left.\begin{aligned} \varepsilon(t) &= y(t) - (\hat{\theta}(t-1))^{\mathrm{T}} \phi(t) \\ R(t) &= R(t-1) + \gamma(t)[\phi(t)\phi^{\mathrm{T}}(t) - R(t-1)] \\ \hat{\theta}(t) &= \hat{\theta}(t-1) + \gamma(t) R^{-1}(t) \phi(t) \varepsilon(t) \end{aligned}\right\} \quad (84)$$

A typical example is the recursive extended least squares (RELS) algorithm applied to the ARMAX model (16). More general algorithms involving filtered $\phi(t)$ and $\varepsilon(t)$ have also widely been used in practice. Algorithm (84) can be interpreted as an approximate prediction error method in the following way. If the implicit θ-dependence in $\phi(t, \theta)$ is neglected, an approximate gradient of the prediction is obtained as

$$\left[\frac{d\hat{y}(t\,|\,\theta)}{d\theta}\right]^{\mathrm{T}} \approx \phi(t, \theta) \tag{85}$$

Replacing $\Psi(t)$ in algorithm (19) by $\phi(t)$ yields the estimator (84).

In linear system identification, it is well known that convergence conditions for pseudo-linear regression algorithms are more restricted than those for prediction error estimators. For RELS, for example, two further assumptions are required:

(i) The true system belongs to the model set, that is, there exists a θ^0 such that the data is generated according to the model

$$A^0(q^{-1})y(t) = B^0(q^{-1})u(t) + C^0(q^{-1})e(t) \tag{86}$$

(ii) $(1/C^0(q^{-1})) - \frac{1}{2}$ is strictly positive real, that is,

$$\mathrm{Re}\left\{\frac{1}{C^0(e^{i\omega})} - \frac{1}{2}\right\} > 0, \quad -\pi < \omega \leqslant \pi \tag{87}$$

Then $\hat{\theta}(t)$ converges with probability 1 to the set

$$D_{\mathrm{c}} = \{\theta\,|\,\bar{E}[\varepsilon(t, \theta) - e(t)]^2 = 0\} \tag{88}$$

as $t \rightarrow \infty$. Notice however that a means of projecting $\hat{\theta}(t)$ into D_{s} (condition (*e*) in § 4.1) is not required as proved by Solo (1979).

Because the prediction $\hat{y}(t\,|\,\theta)$ for the general polynomial NARMAX model is linear in the parameters (§ 2) the RELS algorithm can be readily extended to this case (Billings and Voon 1984). It is of interest to investigate whether the convergence results for the linear case can be carried over for polynomial NARMAX models. For model (15), the answer is obviously yes. Assume that there exists a θ^0, such that the true system is described by

$$y(t) = f^{\mathrm{P}}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u); \theta^0) + C^0(q^{-1})e(t) \tag{89}$$

If condition (87) regarding the true system is satisfied, $\hat{\theta}(t)$ will converge to D_{c} defined in (88) with probability 1. This can easily be verified using the associated d.e. approach as given in Ljung and Söderström (1983). The only difference will be that whereas for the ARMAX model

$$\bar{E}[\phi(t, \theta)\phi^{\mathrm{T}}(t, \theta)] = G_{\mathrm{D}}(\theta) \tag{90}$$

will exist provided the limits

$$\left.\begin{aligned}&\bar{E}[u(t)u(t-k)]\\&\bar{E}[e(t)e(t-k)]\end{aligned}\right\} \tag{91}$$

exist (since $\phi(t, \theta)$ consists of elements obtained by filtering u^t and e^t through constant linear filters) for the NARMAX model (15), the existence of limits (91) will not be enough to guarantee the existence of limit (90) because $\phi(t, \theta)$ is non-linear in the input and output. For the more general model (68), convergence analysis using the associated d.e. method is feasible in principle. This will be investigated in a future study.

6. Simulation study

All the simulation studies assume that, as in the linear case, the structure of the model has been determined by some preliminary analysis on the system. There are several ways of achieving this when the system is non-linear (Billings and Fadzil 1985, Billings *et al.* 1988 a).

Example 1

This is a simulated first-order example. In order to show the robustness of the recursive prediction error method, the system to be identified was chosen to be open loop unstable, and was operated in closed loop:

$$\begin{aligned} y(t) = {} & 1{\cdot}2y(t-1) + 0{\cdot}2u(t-1) - 0{\cdot}8e(t-1) + 0{\cdot}1y^3(t-1) \\ & - 0{\cdot}1y(t-1)u^2(t-1) - 0{\cdot}2y(t-1)u(t-1)e(t-1) + e(t) \end{aligned}$$

The feedback law used was given by

$$u(t) = w(t) - 2{\cdot}0y(t)$$

where $w(t)$ was an independent sequence of uniform distribution with mean zero and variance 1·0. The system noise $e(t)$ was a gaussian white sequence with mean zero and variance 0·04. An input–output sequence of 1000 points were generated. A calculation gives

$$\frac{1}{999}\sum_{t=2}^{1000}(-0{\cdot}8 - 0{\cdot}2y(t-1)u(t-1))^2 = 0{\cdot}55$$

This indicates that the noise model in the data generation is invertible for the particular realization of $\{z(t)\}$ obtained (see the derivation of a sufficient invertibility condition for model (55) in § 4.3). This realization of $\{z(t)\}$ is plotted in Fig. 1.

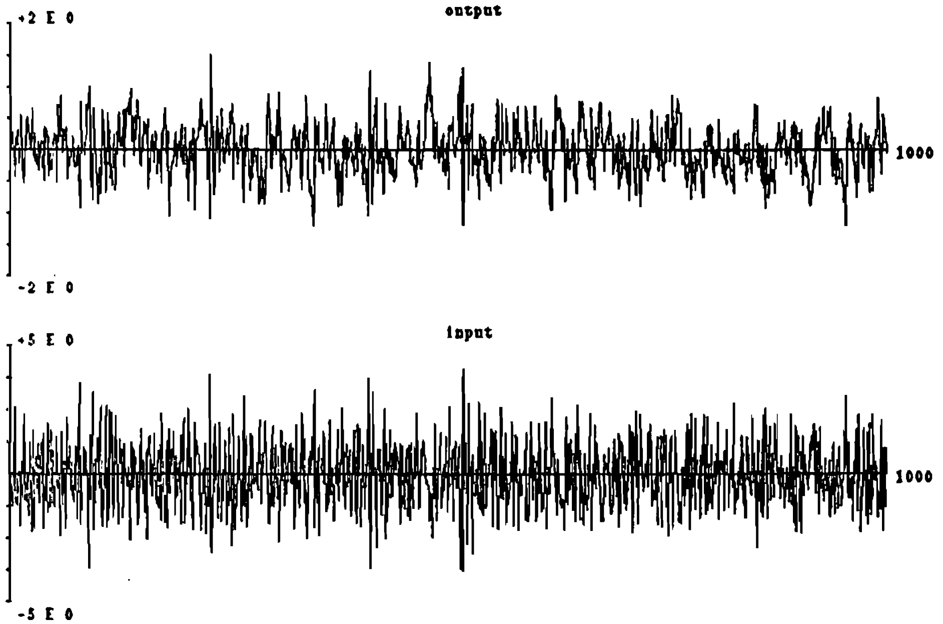


Figure 1. Inputs and outputs of Example 1.

Using $\lambda(t) = \lambda_0 \lambda(t-1) + (1-\lambda_0)$ with initial conditions $\lambda_0 = 0{\cdot}99$, $\lambda(0) = 0{\cdot}95$, $P(0) = 1000{\cdot}0I$ and $\hat{\theta}(0) = 0$, the recursive prediction error algorithm discussed in § 3 was used to estimate the parameters. The results obtained are given in Table 1. The values of a normalized loss function

$$V_1(\hat{\theta}(t)) = \frac{E[\varepsilon^2(t, \hat{\theta}(t))]}{\sigma_e^2}, \quad \sigma_e^2 = E[e^2(t)]$$

are shown in Fig. 2. The dashed curve in Fig. 2 is the asymptotically expected loss under the assumption that $\hat{\theta}(t)$ is asymptotically gaussian distributed with mean θ^0 and covariance equal to the Cramer–Rao lower bound. The evolution of $\hat{\theta}(t)$ is shown in Fig. 3, where the dashed lines indicate the true values of the parameters. As in the recursive identification of linear models, it is seen that the convergence of parameters in the noise model is slower compared with that of the other parameters.

Terms	Parameters	Estimates	True values
$y(t-1)$	θ_1	0·12075E + 1	1·2
$u(t-1)$	θ_2	0·19718E + 0	0·2
$e(t-1)$	θ_3	−0·71458E + 0	−0·8
$y^3(t-1)$	θ_4	0·16287E + 0	0·1
$y(t-1)u^2(t-1)$	θ_5	−0·10835E + 0	−0·1
$y(t-1)u(t-1)e(t-1)$	θ_6	−0·14298E + 0	−0·2

Table 1. Results of Example 1.

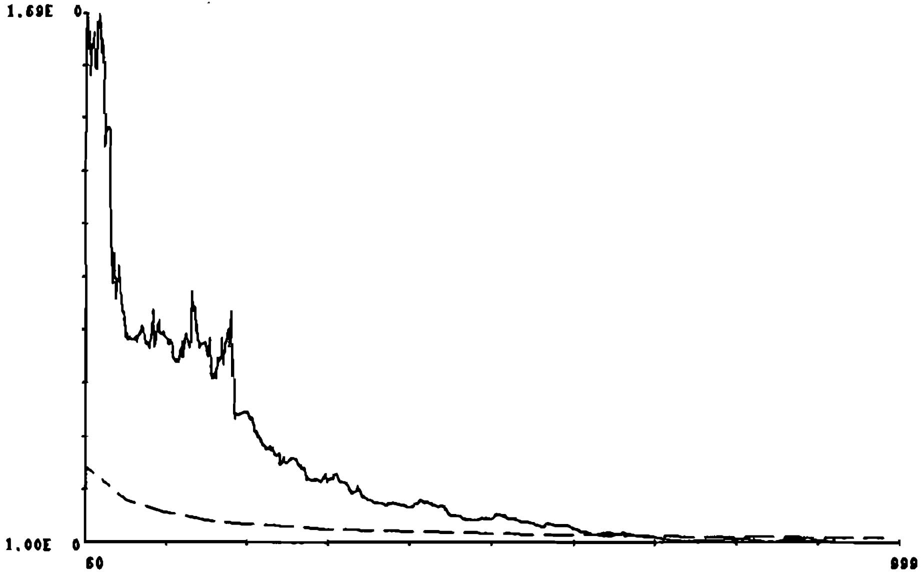
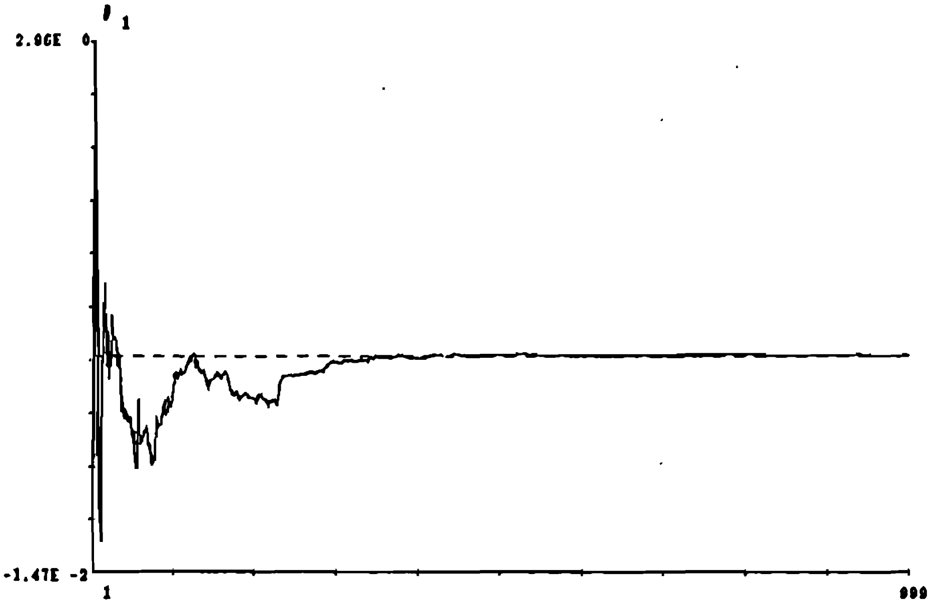
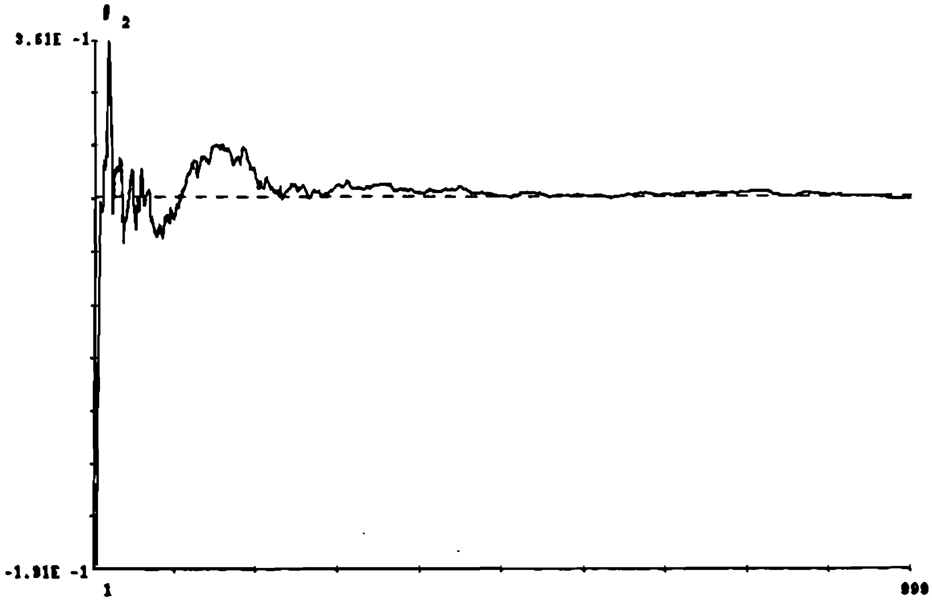


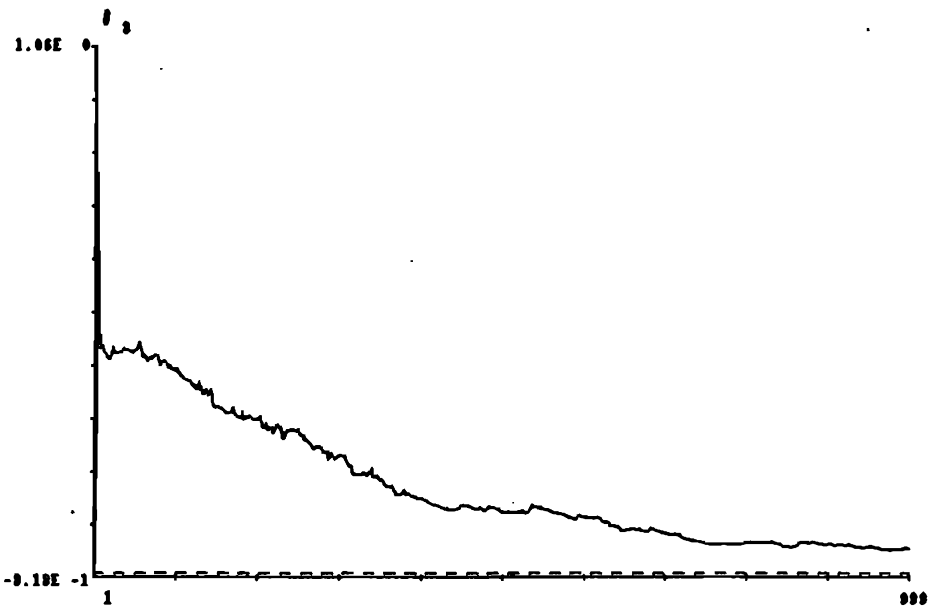
Figure 2. Loss function (normalized) of Example 1.



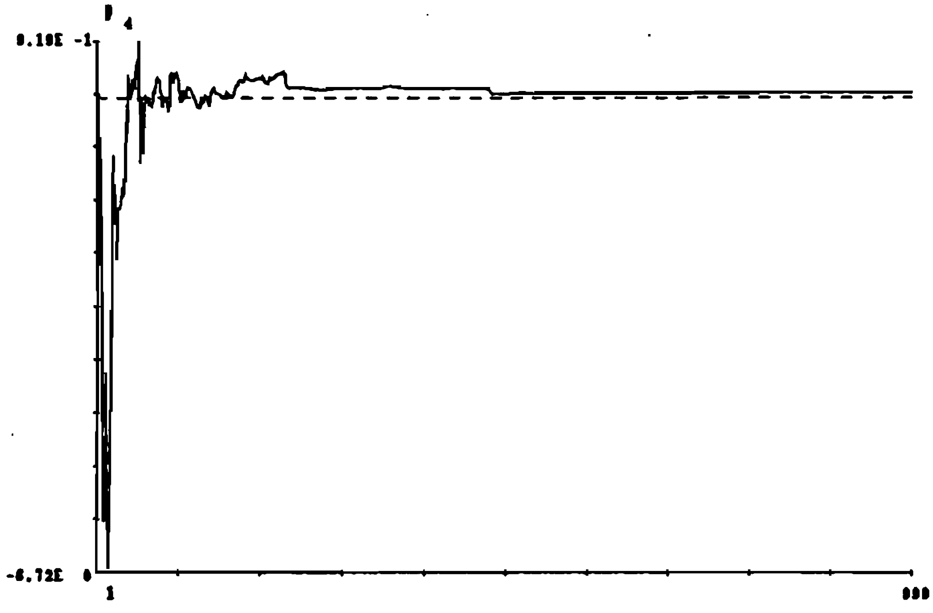
3 (*a*)



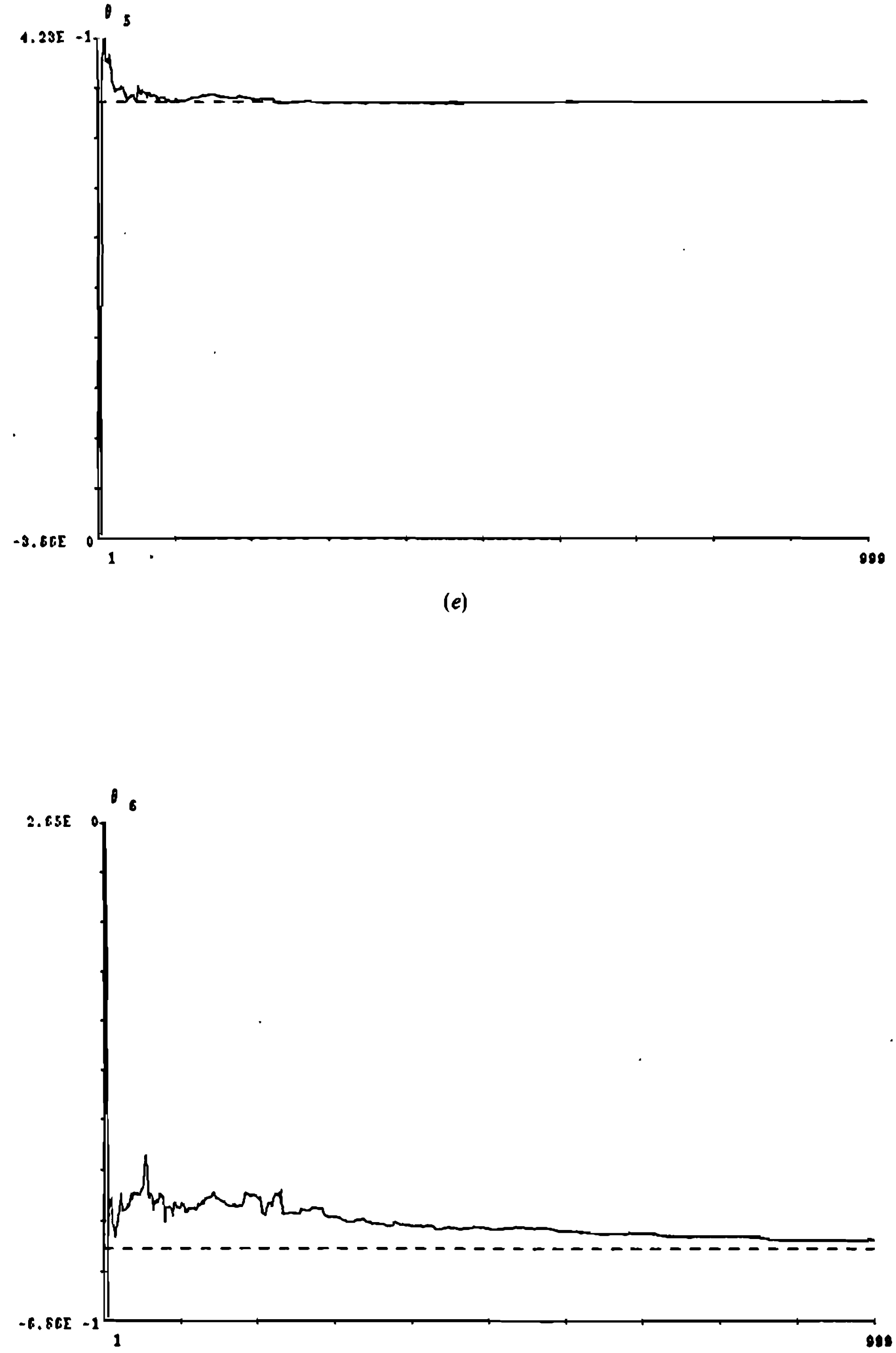
3 (*b*)



(*c*)



(*d*)



(*e*)

(*f*)

Figure 3. Evolution of estimates (Example 1).

Example 2

This is a large pilot-scale liquid level system. The input was a zero-mean gaussian signal. A description of this process is given in Billings and Voon (1986). The inputs and outputs of the system are illustrated in Fig. 4. The data set consists of 1000 points.

Using a combined procedure of forward-regression orthogonal and prediction error estimation coupled with correlation and chi-squared model validity tests (Billings *et al.* 1988 c, Billings and Chen 1989, Billings *et al.* 1988 a), the off-line identification shows that the system can be represented adequately by the following NARMAX model:

$$\begin{aligned} y(t) = {} & 0{\cdot}56276y(t-1) + 0{\cdot}40016y(t-2) + 0{\cdot}41581u(t-1) - 0{\cdot}061813u(t-2) \\ & + 0{\cdot}20941e(t-1) - 0{\cdot}028464e(t-2) + 0{\cdot}042886e(t-3) - 0{\cdot}050201y(t-1)y(t-2) \\ & - 0{\cdot}37836y(t-1)u(t-1) + 0{\cdot}15928y(t-1)u(t-2) - 0{\cdot}037551y^2(t-2)y(t-3) \\ & - 0{\cdot}27654y(t-2)y(t-3)u(t-2) + 0{\cdot}065076y(t-2)y(t-3)u(t-3) \\ & + 0{\cdot}10562y^2(t-3)u(t-2) - 0{\cdot}12220u(t-1)u^2(t-2) + e(t) \end{aligned}$$

The purpose of the present study is to compare the performance of the recursive prediction error algorithm with its corresponding off-line algorithm using the same model structure. With initial conditions $\lambda_0 = 0{\cdot}99$, $\lambda(0) = 0{\cdot}95$, $P(0) = 1000{\cdot}0I$ and $\hat{\theta}(0) = 0$, the recursive prediction error estimator produced the estimates of the parameters very close to those given by its off-line counterpart, as can be seen in Table 2. The loss function $V(\hat{\theta}(t)) = E[\varepsilon^2(t, \hat{\theta}(t))]$ and some of the parameters in the on-line case are shown in Figs 5 and 6. The two models obtained in off-line and on-line identification both have an invertible $C(q^{-1})$ (Table 3).

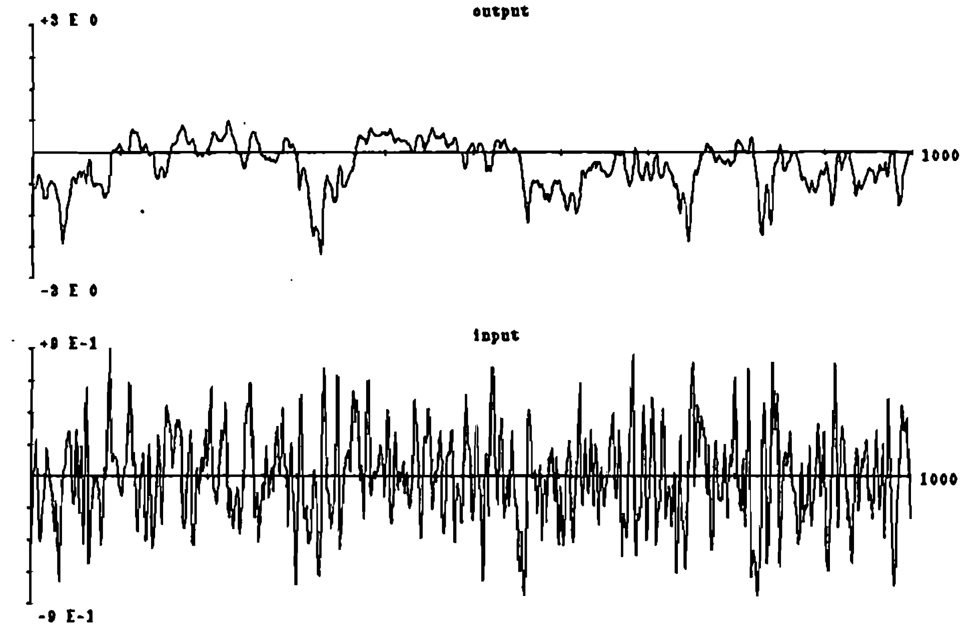


Figure 4. Inputs and outputs of Example 2.

Terms	Parameters	On-line	Off-line
$y(t-1)$	θ_1	0·53832E + 0	0·56276E + 0
$y(t-2)$	θ_2	0·42759E + 0	0·40016E + 0
$u(t-1)$	θ_3	0·41146E + 0	0·41581E + 0
$u(t-2)$	θ_4	−0·50026E − 1	−0·61813E − 1
$e(t-1)$	θ_5	0·22269E + 0	0·20941E + 0
$e(t-2)$	θ_6	0·89100E − 2	−0·28464E − 1
$e(t-3)$	θ_7	0·55900E − 1	0·42886E − 1
$y(t-1)y(t-2)$	θ_8	−0·50813E − 1	−0·50201E − 1
$y(t-1)u(t-1)$	θ_9	−0·37157E + 0	−0·37836E + 0
$y(t-1)u(t-2)$	θ_{10}	0·15142E + 0	0·15928E + 0
$y^2(t-2)y(t-3)$	θ_{11}	−0·37312E − 1	−0·37551E − 1
$y(t-2)y(t-3)u(t-2)$	θ_{12}	−0·26366E + 0	−0·27654E + 0
$y(t-2)y(t-3)u(t-3)$	θ_{13}	0·40384E − 1	0·65076E − 1
$y^2(t-3)u(t-2)$	θ_{14}	0·10379E + 0	0·10562E + 0
$u(t-1)u^2(t-2)$	θ_{15}	−0·12399E + 0	−0·12220E + 0
variance of residuals		0·20184E − 2	0·20046E − 2

Table 2. Results of Example 2.

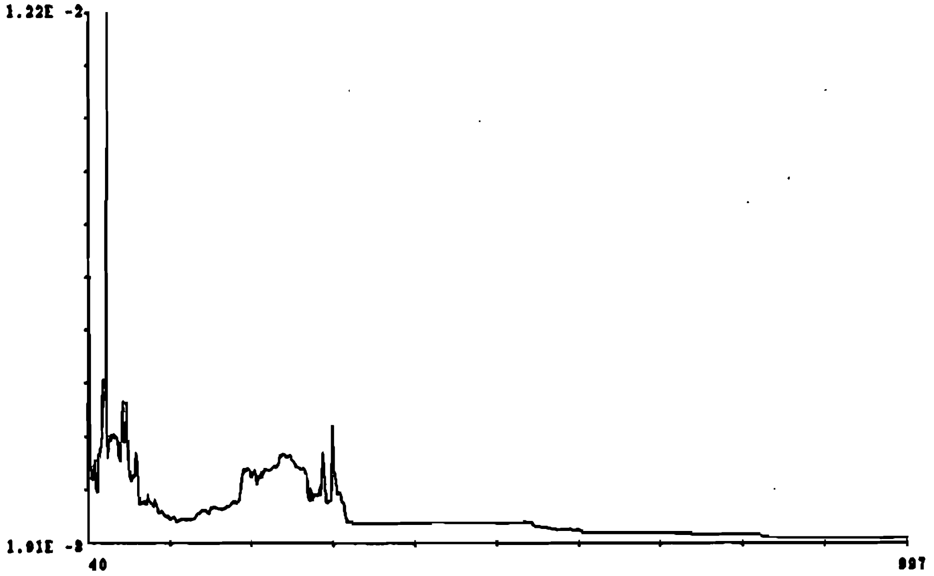
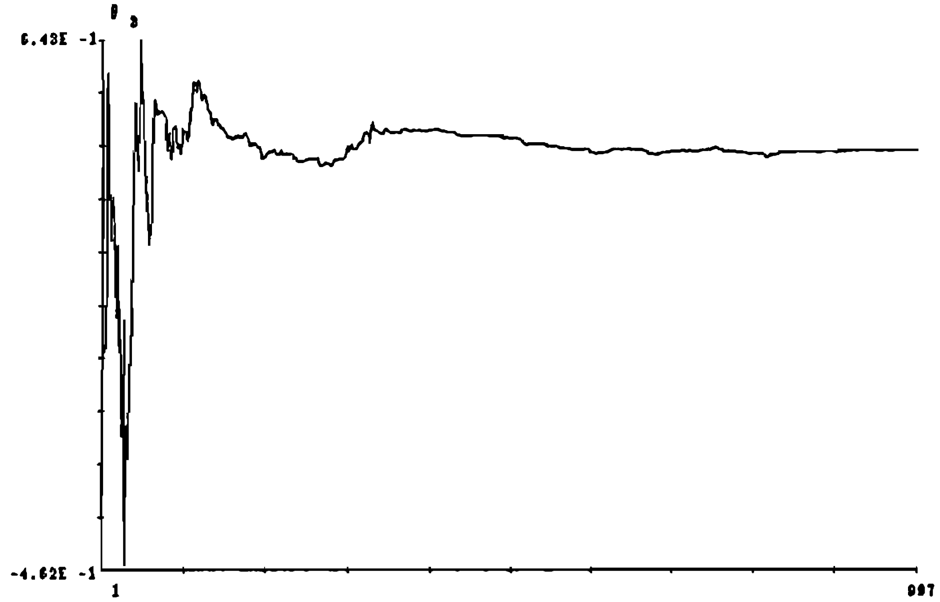
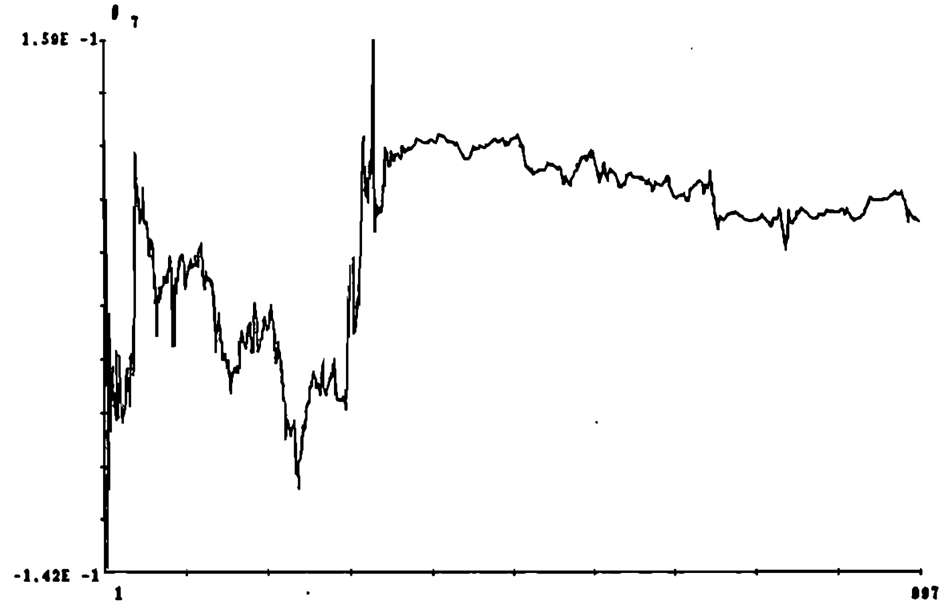


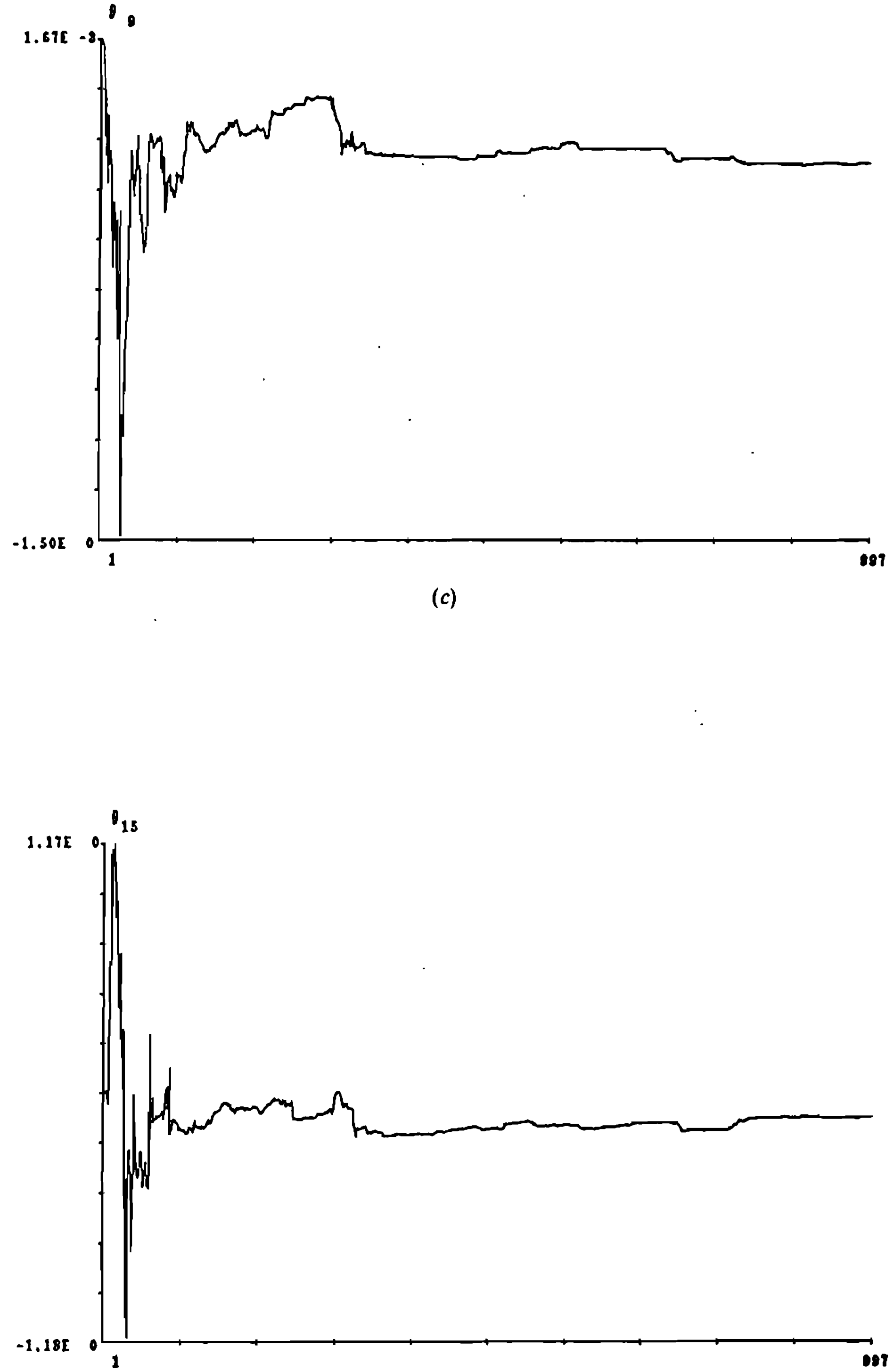
Figure 5. Loss function of Example 2.



6 (*a*)



6 (*b*)



(*c*)

(*d*)

Figure 6. Evolution of some estimates (Example 2).

	Zeros $\bar{s} = \omega + \phi i$ ($i = \sqrt{-1}$)	$\lvert\bar{s}\rvert = (\omega^2 + \phi^2)^{1/2}$
	$-0{\cdot}464$	0·464
On-line	$0{\cdot}120 + 0{\cdot}326i$	0·347
	$0{\cdot}120 - 0{\cdot}326i$	0·347
	$-0{\cdot}467$	0·467
Off-line	$0{\cdot}129 + 0{\cdot}274i$	0·303
	$0{\cdot}129 - 0{\cdot}274i$	0·303

Table 3. Zeros of $C^*(s)$ (Example 2).

7. Conclusions

A recursive prediction error estimator for on-line identification of parameters in NARMAX models has been presented. It has been shown that convergence analysis for the linear model can be extended to NARMAX models. The application to both simulated and real data has been demonstrated.

In order to apply the associated d.e. approach for convergence analysis, the filter that generates the prediction should be exponentially stable. For the NARMAX model, the stability of the filter coincides with the stability of the noise model. m-Invertibility has been introduced to define the stability of the noise model. The analysis shows that while terms which are non-linear functions of the input and output are necessary to describe the dynamics of highly non-linear systems the noise model should be restricted to be linear in the prediction errors.

Although a polynomial expansion is used to provide a parametric representation of the NARMAX model in the present study, alternative expansions such as the rational parametric model (Sontag 1979, Billings and Chen 1989) are also possible. Furthermore the output-affine model (Sontag 1979, Chen and Billings 1988 a) can be thought of as a special parametric case of the NARMAX model. These two non-linear models are essentially non-linear in the parameters. The recursive prediction errror method can readily be applied to these two parametric models where the predicted output can be viewed as the output of a non-linear finite-dimensional filter. Because the output-affine model is linear in the prediction errors the stability of the filter is equivalent to the invertibility of the noise model, and the associated d.e. approach can be employed to analyse the convergence properties of the estimator (Chen and Billings 1988 b). For the rational model, the situation is more complex and further research is required to analyse this estimator.

Acknowledgment

This work was supported by the Science and Engineering Research Council (Grant Ref. GR/D/30587).

References

Billings, S. A., 1986, *Signal Processing for Control*, edited by K. Godfrey and P. Jones (Berlin: Springer-Verlag), pp. 261–294.

Billings, S. A., and Chen, S., 1989, Identification of non-linear rational systems using a prediction-error estimation algorithm, *Int. J. Systems Sci.*, to be published.

Billings, S. A., Chen, S., and Backhouse, R. J., 1988 a, Identification of linear and non-linear models of a turbocharged automotive diesel engine, *Mech. Systems Signal Process.*, to be published.

Billings, S. A., and Fadzil, M. B., 1985, The practical identification of systems with nonlinearities. *Proc. 7th IFAC Symp. on Identification and System Parameter Estimation*, York, U.K., pp. 155–160.

Billings, S. A., Fadzil, M. B., Sulley, J., and Johnson, P. M., 1988 b, *Mech. Systems Signal Process.*, **2**, 59.

Billings, S. A., Korenberg, M. J., and Chen, S., 1988 c, *Int. J. Systems Sci.*, **19**, 1559.

Billings, S. A., and Leontaritis, I. J., 1981, Identification of nonlinear systems using parameter estimation techniques. *Proc. I.E.E. Conf. on Control and its Applications*, Warwick, U.K., pp. 183–187; 1982, Parameter estimation techniques for nonlinear systems. *Proc. 6th IFAC Symp. on Identification and System Parameter Estimation*, Washington D.C., U.S.A., pp. 505–510.

Billings, S. A., and Voon, W. S. F., 1984, *Int. J. Systems Sci.*, **15**, 601; 1986, *Int. J. Control*, **44**, 803.

Chen, S., and Billings, S. A., 1988 a, *Int. J. Control*, **47**, 309; 1988 b, *Ibid.*, **48**, 1605.

Fnaiech, F., and Ljung, L., 1987, *Int. J. Control*, **45**, 453.

Gertler, J., and Bányász, C., 1974, *I.E.E.E. Trans. autom. Control*, **19**, 816.

Granger, C. W. J., and Andersen, A. P., 1978, *Stochastic Proc. Applic.*, **8**, 87.

Leontaritis, I. J., and Billings, S. A., 1985, *Int. J. Control*, **41**, 303.

Liu, Y. P., Korenberg, M. J., Billings, S. A., and Fadzil, M. B., 1987, The nonlinear identification of a heat exchanger. Presented at the *26th I.E.E.E. Conf. on Decision and Control*, Los Angeles, U.S.A.

Ljung, L., 1977, *I.E.E.E. Trans. autom. Control*, **22**, 551; 1978, On recursive prediction error identification algorithms. Report LiTH-ISY-I-0226, Department of Electrical Engineering, Linköping University, Linköping, Sweden; 1979, Convergence of recursive estimators. *Proc. 5th IFAC Symp. on Identification and System Parameter Estimation*, Darmstadt, FRG, pp. 24–28.

Ljung, L., and Söderström, T., 1983, *Theory and Practice of Recursive Identification* (Cambridge, Mass: MIT Press).

Quinn, B. G., 1982, *Stochastic Proc. Applic.*, **12**, 225.

Söderström, T., 1973, An on-line algorithm for approximate maximum likelihood identification of linear dynamic systems. Report 7308, Department of Automatic Control, Lund Institute of Technology, Lund, Sweden.

Solo, V., 1979, *I.E.E.E. Trans. autom. Control*, **24**, 958.

Sontag, E. D., 1979, *Polynomial Response Maps.* Lecture Notes in Control and Information Sciences, Vol. 13 (Berlin: Springer-Verlag).

Subba Rao, T., and Gabr, M. M., 1984, *An Introduction to Bispectral Analysis and Bilinear Time Series Models.* Lecture Notes in Statistics (New York: Springer-Verlag).