ITERATIVE AND RECURSIVE ESTIMATION OF TRANSFER FUNCTIONS

By Carlo Grillenzoni

University of Modena and University of Padova

First version received February 1989

Abstract. A unified treatment of non-linear estimation, pseudolinear regression and stochastic approximation for open-loop transfer function models is provided. Pseudolinear regression techniques are used to derive the recursive non-linear least-squares estimator, avoiding the methodological problems implicit in traditional derivations. Stochastic approximation analysis is used to investigate in a direct manner the conditions of convergence and consistency of both iterative and recursive algorithms. The various methods are compared using data for an industrial process.

Keywords. Non-linear estimation; pseudolinear regression; stochastic approximation; consistency and efficiency.

1. INTRODUCTION

In this paper we provide simplified techniques of derivation and analysis, sufficiently self-contained, for the recursive estimation of the parameters in transfer function (TF) models (Box and Jenkins, 1970, Part III).

This class of non-linear models has been widely used for representing, in the input–output context, dynamical stochastic systems by means of rational polynomials. Although originally designed for the control of industrial processes, it has achieved great success in forecasting economic time series and has played a crucial role in the analysis of causality. Among the estimation methods, the recursive (or on-line) technique, by working on the sequential processing of the data, has the advantage of greater computational speed and, more important can track changes of parameters (non-stationarity). In this way it constitutes the building block for adaptive predictors and regulators.

Hitherto, recursive algorithms have been extensively applied and developed for ARMAX models. Specialized approaches for TF models are those of Young (1984) and Sherif and Liu (1987), based respectively on the use of refined instrumental variables and the extended Kalman filter, neither of which are easy to implement and manage. In Section 3 we use pseudolinear regression techniques (Spliid, 1983; Hannan and McDougall, 1988) to derive the recursive non-linear least-squares estimator. This approach, which is extremely simple and natural, avoids the methodological problems which arise in the typical realization of non-linear on-line algorithms either by adapting their iterative (off-line) version (Goodwin and Sin, 1984; Ljung, 1985) or,

0143-9782/91/02 105–023 \$02.50/0

worse, by utilizing the Kalman filter framework (see Appendix A1 for a note of criticism).

The two techniques (ODE and MCT) available today for the analysis of asymptotic properties, in particular convergence and consistency, are explained in detail by Kushner and Clarke (1978), Solo (1978) and Ljung and Søderstrøm (1983), but are very difficult to follow and apply to the various algorithms. In Section 4 we propose a procedure which is a straightforward extension of the conditions developed in the analysis of the Robbins–Monro and Kiefer–Wolfowitz stochastic approximation schemes (Kashyap *et al.*, 1970). The approach, although informal, is direct and can be applied either in recursive or iterative methods.

The paper ends with an extended empirical example based on the Box–Jenkins 'gas furnace data'. We shall show the tracking capability of recursive methods in the presence of parameters that change sharply, improving the statistical fitting almost without limit.

2. INITIATION AND IDENTIFICATION

Let us consider a bivariate stochastic process $\{x_t, y_t\}$ stationary in covariance with a cross-covariance function $\gamma_{xy}(k)$ which is null for $k < b \geq 0$ and absolutely summable for $k \geq b$. In hypotheses of Gaussianity and zero mean, by the linearity of the regression and making use of rational polynomials, we obtain the parsimonious representation

$$\text{TF} \qquad y_t = \frac{\omega(B)}{\delta(B)} x_{t-b} + \frac{\theta(B)}{\phi(B)} a_t, \qquad a_t \sim \text{IN}(0, \sigma^2) \tag{2.1a}$$

$$\text{ARMA} \qquad \tilde{\phi}(B)x_t = \tilde{\theta}(B)e_t, \qquad e_t \sim \text{IN}(0, \tilde{\sigma}^2) \tag{2.1b}$$

where $(\delta, \omega, \phi, \theta, \tilde{\phi}, \tilde{\theta})$ are linear polynomials, with real coefficients of degree $(r, s, p, q, \tilde{p}, \tilde{q}) < \infty$ respectively, B is the backshift operator and b is the delay $(B^b x_t = x_{t-b})$. Some structural restrictions are needed in order to ensure the stability and the structural identifiability of the whole system:

$$[\delta(z), \phi(z), \tilde{\phi}(z)] \neq 0, \qquad |\omega(z)| < \infty \qquad \text{in } |z| < 1$$

$$[\delta(z), \theta(z), \tilde{\theta}(z)] \neq 0, \qquad |\omega(z)| < \infty \qquad \text{in } |z| \leq 1$$

$$[\delta(0), \phi(0), \theta(0), \tilde{\phi}(0), \tilde{\theta}(0)] = 1, \qquad \omega(0) \neq 1, \ b \geq 0$$

$$[\omega(z), \delta(z)], [\theta(z), \phi(z)], [\tilde{\theta}(z), \tilde{\phi}(z)] \text{ relatively prime.}$$

That is, the monic polynomials are stable and the non-monic polynomial is 'bounded'.

An equivalent representation of the TF which will be very useful in the context of the paper is the so-called pseudolinear form (Solo, 1978). To introduce it we first split (2.1a) into two subsystems $y_t = m_t + n_t$, where

$$m_t = \frac{(\omega_0 + \omega_1 B + \ldots + \omega_s B^s)}{(1 - \delta_1 B - \ldots - \delta_r B^r)} x_{t-b} \rightarrow m_t = \boldsymbol{\delta}' \boldsymbol{m}_{t-1} + \boldsymbol{\omega}' \boldsymbol{x}_{t-b} \quad (2.2a)$$

$$n_t = \frac{(1 + \theta_1 B + \ldots + \theta_q B^q)}{(1 - \phi_1 B - \ldots - \phi_p B^p)} a_t \rightarrow n_t = \boldsymbol{\phi}' \boldsymbol{n}_{t-1} + \boldsymbol{\theta}' \boldsymbol{a}_{t-1} + a_t \quad (2.2b)$$

with $\boldsymbol{m}'_{t-1} = (m_{t-1} \ldots m_{t-r})$, $\boldsymbol{\delta}' = (\delta_1 \ldots \delta_r)$ etc. Now recomposing the above, maintaining the vector notation, the compact 'linear' expression of (2.1a) becomes

$$y_t = \boldsymbol{\beta}' z_t(\boldsymbol{\beta}) + a_t, \qquad z'_t(\boldsymbol{\beta}) = [\boldsymbol{m}'_{t-1}, \boldsymbol{x}'_{t-b}, \boldsymbol{n}'_{t-1}, \boldsymbol{a}'_{t-1}] \quad (2.3)$$

where $\boldsymbol{\beta}' = [\boldsymbol{\delta}', \boldsymbol{\omega}', \boldsymbol{\phi}', \boldsymbol{\theta}']$ is the vector of parameters and $z'_t(\boldsymbol{\beta})$ is the vector of pseudolinear regressors containing lagged variables starting in general with $t - 1$.

A third representation useful in signal processing and control is the state space. This structure for TF systems has not been treated in the literature but, following standard results (Kalman, 1963), it can easily be recovered. Indeed, if we define the companion matrices $\Delta = [\boldsymbol{\delta}:\boldsymbol{I}_{k-1}]$, $k = \max(r, s+1)$, $\Phi = [\boldsymbol{\phi}:\boldsymbol{I}_{h-1}]$, $h = \max(p, q+1)$ and the vector $\tilde{\boldsymbol{\theta}}' = [1, \boldsymbol{\theta}']$, a simple Markovian form for (2.1) is given by

$$\boldsymbol{w}_{t+1} = \begin{pmatrix} \Delta & \boldsymbol{O} \\ \boldsymbol{O} & \Phi \end{pmatrix} \boldsymbol{w}_t + \begin{pmatrix} \boldsymbol{\omega} \\ \boldsymbol{o} \end{pmatrix} x_{t-b+1} + \begin{pmatrix} \boldsymbol{o} \\ \tilde{\boldsymbol{\theta}} \end{pmatrix} a_{t+1}$$

$$y_t = (1\boldsymbol{o}':1\boldsymbol{o}')\boldsymbol{w}_t$$

where $\boldsymbol{w}'_t = [w_t^1 \ldots w_t^k \ldots w_t^{k+h}]$ is the state vector.

In the next section we shall assume that the orders (r, s, b), (p, d, q) of the TF are known and we shall consider the estimation of its $(r + s + 1 + p + q)$ parameters with a bivariate sample of size $N : \{Y_t, X_t\}_1^N$. To introduce the treatment, in what follows a procedure of initiation–identification similar to that of Poskitt (1989) is given. A substantial simplification is afforded by utilizing the results of Priestley (1983) on the separability of the estimates $(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\omega}})$, $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$.

The general philosophy of the procedure is that of pseudolinearity. In practice, it is recognized that a dynamic system can be approximated by a long linear model; moreover, in the non-linear estimation of (2.1) what is really needed in passing from one iteration to another is $z_t(\cdot)$ and not $\boldsymbol{\beta}$. This framework enables calculation to be greatly simplified since only linear algorithms are involved. It works on every dynamic system (vector ARMA, simultaneous TF$_s$), and also on some models with non-linear variables such as the bilinear models.

STEP 1 (PRELIMINARY ESTIMATION OF $\boldsymbol{\phi}, \boldsymbol{\theta}$). Define the rational functions $v(B) = \omega(B)/\delta(B)$, $\pi(B) = \phi(B)/\theta(B)$ and rewrite (2.1a) as $\pi(B)y_t = v(B)\pi(B)x_{t-b} + a_t$. Since $n_t = y_t - v(B)x_{t-b}$ and $v(B) = \Sigma_{i=0}^{\infty} v_i B^i$, the autocovariance function of $\{n_t\}$ satisfies (see Appendix A3 for the proof)

$$\gamma_{nn}(k) = \gamma_{yy}(k) - \{v_k\gamma_{xy}(b) + v_{k+1}\gamma_{xy}(b+1) + \ldots\}.$$

Clearly, if $\{v_i\}$ decays rapidly (i.e. $\delta(B)$ has roots far from the unit circle), we have $\gamma_{nn}(k) \approx \gamma_{yy}(k)$ for all k; in this way $\pi(B)$ can be preliminarily identified–estimated on the observable series $\{y_t\}$ utilizing Steps 3 and 4 below.

Step 2 (estimation of $\boldsymbol{\delta}$, $\boldsymbol{\omega}$). Having obtained $\tilde{\pi}(B)$, pre-whiten the observable series $\tilde{\pi}(B)y_t = \tilde{y}_t$, $\tilde{\pi}(B)x_t = \tilde{x}_t$, so that $\tilde{y}_t \approx v(B)\tilde{x}_{t-b} + a_t$, and form the linear system

$$\delta_r(B)\tilde{y}_t = \omega_s(B)\tilde{x}_{t-b} + \tilde{a}_t. \tag{2.4}$$

Although $\{\tilde{a}_t\}$ is weakly correlated, the estimates of the parameters in the above are consistent in the absence of feedback $y_t \rightarrow x_t$ (Priestley, 1983). Thus the identification of the orders (r, s, b) can be consistently carried out with the criterion

$$\min_b \min_{r,s} \mathrm{BIC}(r, s|b) = \log \tilde{\sigma}_a^2 + (r+s)\frac{\log(N-b)}{N-b}, \quad 0 \leq (r, s, b) < (\log N)^c$$

where $0 < c < \infty$ provides a bound and $\tilde{\sigma}_a^2$ follows from the ordinary least-squares (OLS) estimation of (2.4). The sequential minimization is introduced to simplify the calculations.

Step 3 (preliminary estimation of $z_t(\cdot)$). Having obtained $(\hat{r}, \hat{s}, \hat{b})$, $(\hat{v} = \hat{\omega}/\hat{\delta})$, compute $\hat{m}_t = \hat{v}(B)x_{t-b}$, $\hat{n}_t = y_t - \hat{m}_t$, as in Box and Jenkins (1970, p. 389) and then estimate $\{a_t\}$ with a long autoregression

$$\phi_p{}^*(B)\hat{n}_t = a_t^*, \qquad p^* < \{\log(N - \hat{b})\}^c.$$

The optimal order can be selected as in Spliid (1983) by setting $p^* = (\hat{p} + \hat{q})$, where the latter are identified on the autocorrelation function $\hat{r}(k)$ and partial autocorrelation $\hat{\pi}(k)$ of $\hat{n}_t$, or as in Hannan and Rissanen (1982) with the BIC(p^*) criterion. However, since a_t^* must simply be a white noise in statistical terms, p^* can be chosen in such a way that $[\hat{r}(k), \hat{\pi}(k)] < 2/(N - \hat{b} - k)^{1/2}$ for $k > p^*$.

Step 4 (estimation of $\boldsymbol{\phi}$, $\boldsymbol{\theta}$). Having obtained $\hat{n}_t$, $\hat{a}_t^*$, form the pseudolinear model

$$\hat{n}_t = \{\phi_p(B) - 1\}\hat{n}_t + \{\theta_q(B) - 1\}\hat{a}_t^* + a_t \tag{2.5}$$

and identify (p, q) by minimizing

$$\mathrm{BIC}(p, q) = \log \hat{\sigma}_a^2 + (p+q)\{\log(N - \hat{b})\}/(N - \hat{b})$$

with OLS.

STEP 5. If the roots of the monic polynomials of (2.1) are far from the unit circle, the estimates $\hat{\boldsymbol{\beta}}' = (\hat{\boldsymbol{\delta}}', \hat{\boldsymbol{\omega}}', \hat{\boldsymbol{\phi}}', \hat{\boldsymbol{\theta}}')$ available at this stage tend to be consistent: anyway they are not efficient. A natural way to proceed is to treat $\hat{\boldsymbol{\beta}}$ as the initial value for iterative Gauss–Newton algorithms. Indeed, under Gaussianity non-linear least-squares and maximum likelihood methods are asymptotically equivalent in single-equation models.

Some important remarks are now necessary in order to justify the method.

(i) Priestley (1983) has shown that the estimates of $v(B)$ yielded by the minimization of the two different functionals $J_1 = \Sigma_{t=1}^{N} \{y_t - v(B)x_{t-b}\}^2$, $J_2 = \Sigma_{t=1}^{N} [\pi(B)\{y_t - v(B)x_{t-b}\}]^2$ are asymptotically equivalent in the absence of feedback. Notice that Pierce (1972) had already shown that in the second estimation $\hat{\pi}(B)$, $\hat{v}(B)$ are asymptotically independent. These properties actually make Step 1 unnecessary so that its major motivation is the gain of efficiency in the estimation of Step 2. Indeed, filtering the series $\{x_t, y_t\}$ with the same operator $\pi_y(B)$ reduces the autocorrelation of the residuals $\tilde{a}_t$ on the one hand, but on the other hand leaves unchanged the impulse response function $v(B)B^b = \gamma_{xy}(B)/\gamma_{xx}(B)$.

(ii) Assuming that Steps 1 and 2 provide consistent estimation of $\hat{n}_t$, following Hannan and Kavalieris (1984, p. 274) we can show that the value $p^* < \{\log(N - \hat{b})\}^c$ that minimizes the Bhansali information criterion (BIC) at Step 3 must satisfy the relationship $p^* \approx \log(N - \hat{b})/(2 \log|\rho_0|)$ almost surely (a.s.), where ρ_0 is the 'greatest' root of $\theta_0(B)$. Now, if ρ_0 is near the unit circle and N is not very large, the size of p^* obtained in practice is underestimated and the values of (p, q) selected with BIC at Step 4 turn out to be overestimated (see also Poskitt, 1989, p. 36). Furthermore, since (2.5) is a pseudolinear regression, such that it generally converges only when ρ_0 is far from the unit circle (see Section 4), the estimates $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\theta}}$ could in turn be non-consistent. In view of these problems, Poskitt (1989), like Hannan and Kavalieris (1984), delayed the selection of the orders with BIC$(r, s, p, q|b = 0)$ until Step 5 of the Gauss–Newton estimation.

(iii) Steps 1–4 would then provide only a suitable method of initiation for TF models which were already identified. In this context, however, we emphasize that selection strategies based on information criteria (AIC, BIC, MDL, LIL etc.) strongly rely on a hypothesis that a *true* system (2.1) exists. If, on the contrary, the model generating the data contains unstable factors $(1 - B)^d$, irregular operators $\omega(B)^* = (\omega_0 + \omega_1 B^k + \omega_2 B^{k+h} + \ldots)$ or periodic filters $\phi(B^k) = (1 - \phi_1 B^k - \phi_2 B^{2k} - \ldots)$, then good identification results can be achieved with classical non-parametric methods based on the inspection of sample correlation functions.

3. ITERATIONS AND RECURSIONS

In this section we derive the algorithms for non-linear least squares (NLS) and pseudolinear regression (PLR) in the iterative (I) and recursive (R) versions, showing their algebraic connections. Acronyms will be set on the left of the corresponding formulae when appropriate.

Let us assume a loss function of the quadratic type which recalls the sample variance of the one-step-ahead prediction error:

$$\min_{\boldsymbol{\beta}} J_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{t=1}^{N} a_t^2(\boldsymbol{\beta})$$

$$\boldsymbol{\beta}' = [\delta_1 \ldots \delta_r, \omega_0, \omega_1 \ldots \omega_s, \phi_1 \ldots \phi_p, \theta_1 \ldots \theta_q] \tag{3.1}$$

In numerical analysis, the typical derivation of the iterative Gauss–Newton estimator operates through a second-order Taylor expansion of J_N and some approximations (not always acceptable) on the matrix of second-order derivatives. A more satisfactory and simple approach may instead consider a first-order expansion of a_t in $\hat{\boldsymbol{\beta}}$ and then apply the OLS method iteratively:

$$a_t(\boldsymbol{\beta}) \approx a_t(\hat{\boldsymbol{\beta}}) - (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \boldsymbol{\xi}_t(\hat{\boldsymbol{\beta}}) \tag{3.2a}$$

I – NLS $$\hat{\boldsymbol{\beta}}(k+1) - \hat{\boldsymbol{\beta}}(k) = \left\{ \sum_{t=1}^{N} \hat{\boldsymbol{\xi}}_t(k) \hat{\boldsymbol{\xi}}_t'(k) \right\}^{-1} \sum_{t=1}^{N} \hat{\boldsymbol{\xi}}_t(k) \hat{a}_t(k) \tag{3.2b}$$

It is easily seen, by substituting (3.2a) into J_N and by the properties of OLS, that this derivation is consistent with the given problem of minimization.

The implementation of the algorithm (3.2) provided by Box and Jenkins (1970) was essentially based on the numerical evaluation of the gradient $\boldsymbol{\xi}_t$ obtained (as in Kiefer–Wolfowitz schemes) with perturbation of the parameters and a three-step filtering procedure for computing the residuals a_t. An equivalent estimator based on the analytical evaluation of the derivatives follows by generalizing Stage 3 of the algorithm of Hannan and Rissanen (1982) for ARMA models. Utilizing (2.2), we can show by means of standard calculus that

$$\boldsymbol{\xi}_t(\boldsymbol{\beta}) = \begin{cases} -\dfrac{\partial a_t}{\partial \delta_i} = \dfrac{\phi(B)}{\theta(B)\delta(B)} m_{t-i} \\ -\dfrac{\partial a_t}{\partial \omega_j} = \dfrac{\phi(B)}{\theta(B)\delta(B)} x_{t-b-j} \\ -\dfrac{\partial a_t}{\partial \phi_k} = \dfrac{1}{\theta(B)} n_{t-k} \\ -\dfrac{\partial a_t}{\partial \theta_h} = \dfrac{1}{\theta(B)} a_{t-h}. \end{cases} \tag{3.3}$$

The computation of the gradient thus consists in a filtering operation on observable (x), auxiliary (m, n) and non-observable (a) quantities. This feature, allowing for the back-forecasting generation of the initial ‘regressors’ $\boldsymbol{\xi}_{-t}$, z_{-t}, guarantees that (3.2) will have the same statistical properties as the full maximum likelihood estimator (Pierce, 1972; Poskitt, 1989).

Using a vector notation, it is now readily seen from (2.3) and (3.3) that

$$\xi_t(\boldsymbol{\beta}) = \boldsymbol{G}(B)\boldsymbol{z}_t(\boldsymbol{\beta}), \qquad \boldsymbol{G}(B) = \mathrm{diag}\left[\frac{\phi(B)}{\theta(B)\delta(B)} \cdots ; \cdots \frac{1}{\theta(B)}\right].$$

In this context the PLR estimator can be formally derived from the NLS estimator by approximating $\xi_t \approx z_t$, i.e. by avoiding the filtering with $G(B)$. Indeed, since at an iterative level

$$\hat{a}_t(k) = y_t - \hat{z}_t(k)'\hat{\boldsymbol{\beta}}(k), \qquad \hat{\xi}_t(k) \approx \hat{z}_t(k),$$

substituting these quantities in (3.2b) gives the compact algorithm

$$\text{I} - \text{PLR} \qquad \hat{\boldsymbol{\beta}}(k+1) = \left\{\sum_{t=1}^{N} \hat{z}_t(k)\hat{z}_t'(k)\right\}^{-1} \sum_{t=1}^{N} \hat{z}_t(k)y_t \tag{3.4}$$

Heuristically, this estimator might also be obtained by applying OLS iteratively to (2.3) (Splidd, 1983). The crucial step, however, is taken by approximation of the gradient. It is the goodness of this approximation, expressible in terms of $G(B)$, that determines the properties of convergence (see Hannan and McDougall (1988) and the next section). Notice, moreover, that the algebraic derivation of (3.4) from (3.2) is not possible for non-linear estimators with Newton–Raphson or Marquardt steps.

The compact structure of the I-PLR algorithm makes it easy to derive its corresponding sequential version by adapting recursive least-squares (RLS) techniques (Plackett, 1950; Ljung and Söderström, 1983). For this purpose we equate the number of iterations and the number of processed data $(k = N) = t$ in (3.4) and introduce a sequence $\{\lambda^i\}$ which discounts old observations (suitable for tracking parameter changes):

$$\hat{\boldsymbol{\beta}}(t) = \left\{\sum_{\tau=1}^{t} \hat{z}_\tau(t-1)\lambda^{t-\tau}\hat{z}_\tau'(t-1)\right\}^{-1} \sum_{\tau=1}^{t} \hat{z}_\tau(t-1)\lambda^{t-\tau}y_\tau \tag{3.5a}$$

$$= \boldsymbol{R}(t)^{-1}\boldsymbol{s}(t) \quad \text{say.} \tag{3.5b}$$

Typically $0 \le \lambda \le 1$, and defining $\hat{z}_t(t-1) = \hat{z}_t$ we easily obtain

$$\boldsymbol{R}(t) = \lambda\boldsymbol{R}(t-1) + \hat{z}_t\hat{z}_t', \qquad \lambda\boldsymbol{R}(t-1) = \boldsymbol{R}(t) - \hat{z}_t\hat{z}_t'$$

$$\boldsymbol{s}(t) = \lambda\boldsymbol{s}(t-1) + \hat{z}_t y_t, \qquad \boldsymbol{s}(t-1) = \boldsymbol{R}(t-1)\hat{\boldsymbol{\beta}}(t-1).$$

Hence, substituting the last three expressions sequentially in (3.5b) we obtain

$$\text{R} - \text{PLR} \qquad \hat{\boldsymbol{\beta}}(t) = \hat{\boldsymbol{\beta}}(t-1) + \boldsymbol{R}(t)^{-1}\hat{z}_t\{y_t - \hat{\boldsymbol{\beta}}(t-1)'\hat{z}_t\} \tag{3.6a}$$

$$\nu(t) = \nu(t-1) + \{y_t - \hat{\boldsymbol{\beta}}(t)'\hat{z}_t\}^2, \qquad \boldsymbol{P}(t) = \frac{\nu(t)\boldsymbol{R}(t)^{-1}}{t} \tag{3.6b}$$

where $\boldsymbol{P}(t)$ is an approximate estimator of the dispersion matrix of $\hat{\boldsymbol{\beta}}(t)$.

The two terms in braces in (3.6a) and (3.6b) are respectively the predicted error $\tilde{a}_t$ and the estimated error $\hat{a}_t$. In the iterative version (3.4) we have, as an intermediate quantity, the residual of regression $\tilde{a}_t(k) = y_t - \hat{\boldsymbol{\beta}}(k)'\hat{z}_t(k-1)$. Used in place of $\hat{a}_t(k)$ this computationally simplifies the algorithm but introduces a substantial inefficiency that slows down the convergence. The actual cost of the recursive transformation (3.6) is now represented by the fact that, unlike (3.4), the pseudolinear regressors can no longer be computed 'exactly'. However, a simple solution can be obtained with a proper dynamic adaptation of the iterative calculations; for $\{m_t\}$ this means

I $$\hat{m}_t(k) = \sum_{i=1}^{r} \hat{\delta}_i(k)\hat{m}_{t-i}(k) + \sum_{j=0}^{s} \hat{\omega}_j(k)x_{t-b-j}$$

R $$\hat{m}(t) = \sum_{i=1}^{r} \hat{\delta}_i(t)\hat{m}(t-i) + \sum_{j=0}^{s} \hat{\omega}_j(t)x_{t-b-j}$$

Thus, using quantities like the last one, the updated vector of regressors becomes

$$\hat{z}'_{t+1}(t) = [\hat{m}(t) \ldots \hat{m}(t-r+1), x_{t-b} \ldots x_{t-b-s}, \hat{n}(t) \ldots \hat{n}(t-p+1),$$
$$\hat{a}(t) \ldots \hat{a}^*t - q + 1)]$$

which depends on all past values of $\hat{\boldsymbol{\beta}}(t)$ (indeed, we have set $k = N = t$).

At this point the sequential NLS estimator can be simply recovered by re-establishing the filtering with the 'transfer function' $G(B)$ in (3.6):

I $$\hat{\xi}_t(k) = \hat{G}_k(B)\hat{z}_t(k)$$

R $$\hat{\xi}_{t+1}(t) = \hat{G}_t(B)\hat{z}_{t+1}(t)$$

Notice, however, from (3.3) that derivatives with respect to the same kind of parameters (say α) satisfy the dynamic relationship

$$\xi^{\alpha}_{i+k_t}(\boldsymbol{\beta}) = -\frac{\partial a_t(\boldsymbol{\beta})}{\partial \alpha_{i+k}} = -\frac{\partial a_t(\boldsymbol{\beta})}{\partial \alpha_i} B^k = \xi^{\alpha}_{i_{t-k}}(\boldsymbol{\beta}), \qquad \alpha = \delta, \omega, \theta, \phi$$

so that the computation of the gradient can be compactly developed by means of recurrence formulae. For $\{-\partial a_t/\partial \phi_k\}$ we have, as in Hannan and Rissanen (1982),

I $$\ddot{n}_t(k) = \hat{n}_t(k) - \sum_{i=1}^{q} \hat{\theta}_i(k)\ddot{n}_{t-i}(k)$$

R $$\ddot{n}(t) = \hat{n}(t) - \sum_{i=1}^{q} \hat{\theta}_i(t)\ddot{n}(t-i).$$

We remark, however, that the introduction of this filtering is insignificant, from a computational point of view, only for ARMAX models. Indeed, for the TF models the calculation of the first two derivatives, such as $\{-\partial a_t/\partial \delta_i\}$, requires the three steps

R $$\tilde{m}(t) = \hat{m}(t) - \sum_{i=1}^{p} \hat{\phi}_i(t)\hat{m}(t-i)$$

$$\dot{m}(t) = \tilde{m}(t) - \sum_{i=1}^{q} \hat{\theta}_i(t)\dot{m}(t-i)$$

$$\ddot{m}(t) = \dot{m}(t) + \sum_{i=1}^{r} \hat{\delta}_i(t)\ddot{m}(t - i)$$

and these increase the length of the FORTRAN routine by more than 50%.

Finally, utilizing quantities like the above, we obtain the updated gradient

$$\hat{\boldsymbol{\xi}}'_{t+1}(t) = [\ddot{m}(t) \dots \ddot{m}(t - r + 1), \ddot{x}(t) \dots \ddot{x}(t - s), \ddot{n}(t) \dots$$
$$\ddot{n}(t - p + 1), \ddot{a}(t) \dots \ddot{a}(t - q + 1)]$$

and substituting this in place of $z_t(\cdot)$ in (3.6), by recalling (3.2), we obtain

$$\text{R} - \text{NLS} \qquad \hat{\boldsymbol{\beta}}(t + 1) = \hat{\boldsymbol{\beta}}(t) + \boldsymbol{S}(t + 1)^{-1}\hat{\boldsymbol{\xi}}_{t+1}\{y_{t+1} - \hat{\boldsymbol{\beta}}(t)'\hat{\mathbf{z}}_{t+1}\} \qquad (3.7\text{a})$$

$$\boldsymbol{S}(t + 1) = \lambda\boldsymbol{S}(t) + \hat{\boldsymbol{\xi}}_{t+1}\hat{\boldsymbol{\xi}}'_{t+1}, \qquad \boldsymbol{Q}(t + 1) = \frac{v(t + 1)\boldsymbol{S}(t + 1)^{-1}}{t + 1} \qquad (3.7\text{b})$$

where $\boldsymbol{Q}(t + 1)$ is a consistent estimator of the dispersion matrix of $\hat{\boldsymbol{\beta}}(t + 1)$.

In summary, the global strategy of derivation developed so far can be sketched as follows: I-NLS(3.2) → I-PLR(3.4) → RLS → R-PLR(3.6) → R-NLS(3.7). Following this scheme we have avoided most of the methodological problems which arise in the direct derivation of (3.7) from (3.2); furthermore, we have developed the background of the PLR techniques which are treated by many authors in a completely heuristic fashion. To explain the first point, we remark that in the prediction error method (PEM) the derivation of the recursive algorithm from the corresponding iterative version usually imposes

$$\hat{J}'_t(t - 1) = \left.\frac{\partial J_t(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(t-1)} = \hat{J}'_{t-1}(t - 1) + \hat{\boldsymbol{\xi}}_t(t - 1)\hat{a}_t(t - 1) \approx \hat{\boldsymbol{\xi}}_t\hat{a}_t,$$

i.e. $\hat{J}'_{t-1}(t - 1) \approx 0$ (Goodwin and Sin, 1984; Ljung, 1985). However, this assumption does not hold on many grounds. For example, in the case of time-varying parameters (which is the major field of application of the recursive methods) $J_{t-1}(\boldsymbol{\beta})$ may at most be minimized by the sequence $\{\hat{\boldsymbol{\beta}}(1) \dots \hat{\boldsymbol{\beta}}(t - 1)\}$ and not by its final value only.

4. CONSISTENCY AND EFFICIENCY

The asymptotic analysis of iterative and recursive estimators of dynamic models has mostly been concerned with the properties of consistency and efficiency; properties of I-NLS have been well investigated (e.g. Poskitt, 1989). We now briefly extend to TF systems the principal conclusions reached for ARMAX models in the 'recursive literature' under general assumptions of stationarity, stability and $\lambda = 1$ (e.g. Ljung and Söderström, 1983; Goodwin and Sin, 1984).

R-NLS. Since this estimator is a simple algebraic transformation of its corresponding I-NLS, it not only has the same general asymptotic properties (normality and consistency) but also shares an identical limiting dispersion matrix:

$$t^{1/2}\{\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0\} \underset{t\to\infty}{\xrightarrow{d}} \mathrm{N}[\mathbf{0}, \sigma_0^2 E\{\boldsymbol{\xi}_t(\boldsymbol{\beta}_0)\boldsymbol{\xi}_t'(\boldsymbol{\beta}_0)\}^{-1}]. \tag{4.1}$$

R-PLR. Since this algorithm approximates the above in the gradient, it maintains the properties of convergence only if the matrix $G(z)$ behaves like a passive filter. This feature is expressible in terms of the positive real conditions:

$$\text{necessary} \qquad \mathrm{Re}\{G(z)\} > 0 \qquad |z| = 1 \tag{4.2a}$$

$$\text{sufficient} \qquad \mathrm{Re}\{G(z) - 1/2\} > 0 \qquad |z| = 1 \tag{4.2b}$$

(see Appendix A2 for details). Moreover, unlike (4.1), we generally have

$$\lim_{t\to\infty} E[t^{1/2}\{\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0\}]^2 \neq \sigma_0^2 E\{z_t(\boldsymbol{\beta}_0)z_t'(\boldsymbol{\beta}_0)\}^{-1} \tag{4.3}$$

Recently, these conclusions have been extended to the iterative case by Stoica *et al.* (1985), as regards the conditions of convergence, and by Hannan and McDougall (1988) for the general asymptotic properties.

The formal proofs of the results (4.1)–(4.3) for ARMAX models have required complex and refined mathematical apparatus which is difficult to follow and apply to the various cases. In particular, it is not easy to check whether both (4.2a) and (4.2b) are needed in I-PLR, why they are not required in NLS, or when (4.3) may hold with an equals sign. In the sequel we reconsider the approach of analysis by recognizing that the algorithms available today have a strong background in the methods of stochastic approximation developed in the period 1950–70.

A typical problem of stochastic approximation consists in finding the value $\boldsymbol{\beta}_0$ that minimizes a mean value $J(\boldsymbol{\beta})$. For every $\boldsymbol{\beta}$ a random variable $x(\boldsymbol{\beta})$ is observed, such that $E\{x(\boldsymbol{\beta})|\boldsymbol{\beta}\} = J(\boldsymbol{\beta})$, or its gradient $\boldsymbol{y}(\boldsymbol{\beta})$, such that $E[\boldsymbol{y}(\boldsymbol{\beta})|\boldsymbol{\beta}] = \partial J(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \nabla(\boldsymbol{\beta})$. The derivation of the basic scheme of calculation parallels the steepest descent method:

$$\hat{\boldsymbol{\beta}}(n) = \hat{\boldsymbol{\beta}}(n) + \alpha(n)\boldsymbol{y}(n), \qquad \boldsymbol{y}(n) = \left.\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(n)} = \hat{\nabla}(n) \tag{4.4}$$

where $\alpha(n)$ is the so-called gain sequence. Many authors have dealt with the analysis of the above; one of the more convincing results is that of Gladyshev (1965).

THEOREM. *If the stochastic approximation scheme (4.4) fulfils the conditions*

$$(i) \qquad \alpha(n) > 0, \qquad \sum_{n=0}^{\infty} \alpha(n) = \infty, \qquad \sum_{n=0}^{\infty} \alpha(n)^2 < \infty \tag{4.5a}$$

$$(ii) \qquad \inf_{\varepsilon<\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|<\varepsilon^{-1}} E\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\boldsymbol{y}(\boldsymbol{\beta})\} > 0 \qquad \forall\varepsilon > 0 \tag{4.5b}$$

(*iii*) $$E\{\|y(\boldsymbol{\beta})\|^2\} \leq \eta(1 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2) \qquad \eta > 0 \tag{4.5c}$$

then $\hat{\boldsymbol{\beta}}(n)$ converges almost surely and globally to the unique value $\boldsymbol{\beta}_0$ which makes $\nabla(\boldsymbol{\beta}_0) = 0$: $\lim p\{\|\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}_0\| = 0\} = 1$, $n \to \infty$, $\forall \hat{\boldsymbol{\beta}}(0)$ initial.

Instead of reviewing the proof of the theorem (an excellent survey is given by Kashyap *et al.* (1970, p. 350)), we now briefly discuss the meaning of assumptions (4.5). Condition (4.5a) requires that the rate of decrease in $\{\alpha\}$ is such that the variance of the estimate of $J(\boldsymbol{\beta})$ is reduced to zero; in particular, the second requirement provides unlimited correction effort and the third guarantees mutual cancellation of individual errors for a large number of steps. The harmonic sequence $\alpha(n) = 1/n$ satisfies all three requirements. Condition (4.5b), i.e. $\inf(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\nabla(\boldsymbol{\beta}) > 0$, means that $\nabla(\boldsymbol{\beta})$ behaves like a linear function of $\boldsymbol{\beta}$ in a neighbourhood of $\boldsymbol{\beta}_0$. It also implicitly recalls the principle of dynamic programming which states that the direction $\hat{\Delta}(k) = \hat{\boldsymbol{\beta}}(k) - \hat{\boldsymbol{\beta}}(k-1)$ of an iterative algorithm is admissible only if it forms an acute angle with the gradient of the objective function evaluated at that point (Tsypkin, 1971, p. 28):

$$\hat{J}(k) < \hat{J}(k-1) \Leftrightarrow \hat{\Delta}(k)'\hat{\nabla}(k) > 0 \tag{4.6}$$

Furthermore, for $\alpha(n) = \alpha_0$ constant, $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\nabla(\boldsymbol{\beta})$ is equivalent to a Lyapunov function (Tsypkin, 1971, p. 34) and this ensures the global character of the convergence. Finally, condition (4.5c) guarantees that the variance of $y(\boldsymbol{\beta})$ is finite and that $\|y(\boldsymbol{\beta})\|^2$ is bounded above by a quadratic function of $\Delta = \boldsymbol{\beta} - \boldsymbol{\beta}_0$ for all $\boldsymbol{\beta}$. It is worth noting that the independence of the sample realizations $\{y(n)\}$ is not required here.

Now, in order to utilize the theorem for the analysis of the algorithms of Section 3, we must first put them in the form of stochastic approximation schemes and then we must check if and how the conditions (4.5) hold. This task can be accomplished by equating $(k = N) = t$, introducing $\alpha(t)$ and then evaluating the 'angle'

$$\inf_{0<t<\infty} E\{\Delta(t)'\nabla(t)|\boldsymbol{\beta}\} > 0 \tag{4.7}$$

for suitable objective functions. The importance of the measure (4.7), which actually combines conditions (4.4), (4.5b) and (4.6), is twofold. It can work directly on estimators in both recursive and iterative form; moreover, it treats jointly the problems of the convergence in numerical sense (k) and in statistical sense (N). In contrast, many statistical analyses of non-linear estimators are unsatisfactory since they usually assume the existence of a consistent initial estimator and only investigate what happens in the first iteration. In what follows we analyse three cases.

I-PLR. Consider (3.4) in step-variable form and equate $(k = N) = t$:

$$\hat{\boldsymbol{\beta}}(t) = \hat{\boldsymbol{\beta}}(t-1) + \alpha(t)\left\{\sum_{\tau=1}^{t} \hat{\boldsymbol{z}}_\tau(t-1)\hat{\boldsymbol{z}}'_\tau(t-1)\right\}^{-1} \sum_{\tau=1}^{t} \hat{\boldsymbol{z}}_\tau(t-1)\hat{a}_\tau(t-1)$$

with $\alpha(t)$ satisfying (4.5a). Now referring to $J_t = \Sigma_{\tau=1}^{t} a_\tau^2/2t$, we have $/t$, with $\xi_\tau = \boldsymbol{G}(B)z_\tau$, so that (4.7) becomes

$$\inf_t E\left[\alpha(t)\left(\sum_{\tau=1}^{t} z_\tau a_\tau\right)'\left(\sum_{\tau=1}^{t} z_\tau z_\tau'\right)^{-1}\left\{\frac{1}{t}\sum_{\tau=1}^{t} \boldsymbol{G}(B) z_\tau a_\tau\right\}\right] > 0$$

Clearly this holds only if the passivity condition (4.2a) is satisfied. To prove this in detail, define the random variable $w_t = \sqrt{\boldsymbol{w}_t'\boldsymbol{w}_t}$ where $\boldsymbol{w}_t = [\Sigma_1^t z_\tau z_\tau']^{-1/2}\Sigma_1^t z_\tau a_\tau$ and let $\alpha(t) = 1/t$; thus the previous expression becomes

$$\begin{aligned}\operatorname*{Inf}_t E[\alpha(t)^2 w_t^2 \boldsymbol{G}(B)] &= \operatorname*{Inf}_t \int_{-\pi}^{+\pi} \alpha(t)^2 \lambda_{ww}(e^{-i\omega})\boldsymbol{G}(e^{-i\omega})\,\mathrm{d}\omega \\ &= \operatorname*{Inf}_t \int_{-\pi}^{+\pi} \alpha(t)^2 \lambda_{ww}(e^{-i\omega})\mathrm{Re}[\boldsymbol{G}(e^{-i\omega})]\mathrm{d}\omega \geq 0\end{aligned}$$

R-PLR. Let $\lambda = 1$, $\alpha(t) = 1/t$ and substitute $\boldsymbol{R}(t)$ with $\tilde{\boldsymbol{R}}(t) = \boldsymbol{R}(t)/t$ in (3.6):

$$\hat{\boldsymbol{\beta}}(t) = \hat{\boldsymbol{\beta}}(t-1) + \alpha(t)\tilde{\boldsymbol{R}}(t)^{-1}\hat{\boldsymbol{z}}_t(t-1)\tilde{a}_t(t-1)$$

where $\tilde{a}_t(t-1) = y_t - \hat{\boldsymbol{\beta}}(t-1)'\hat{\boldsymbol{z}}_t(t-1)$. Hence considering $J_t = a_t^2/2$, (4.7) becomes

$$\inf E[\{\alpha(t)z_t' a_t \tilde{\boldsymbol{R}}(t)^{-1}\}\{\boldsymbol{G}(B)z_t a_t\}] > 0$$

which again holds only if $\boldsymbol{G}(B)$ is positive in the sense (4.2a).

REMARK. As regards the restrictive condition (4.2b), it is apparent, by combining the arguments of Ljung and Söderström (1983, p. 457) and Goodwin and Sin (1984, p. 345), that it fundamentally serves to assure the positive definiteness of $\tilde{\boldsymbol{R}}(t)$ for all t. Notice, in fact, that $\boldsymbol{R}(t)$, computed recursively, depends on all the values $\{\hat{\boldsymbol{\beta}}(1)\ldots\hat{\boldsymbol{\beta}}(t-1)\}$, and thus it might be negative definite. Furthermore, we observe that in the analysis of the stochastic gradient version of the R-PLR—in which $\boldsymbol{R}(t)$ is replaced by its trace $\Sigma_\tau^t\|\hat{\boldsymbol{z}}_\tau(\tau-1)\|^2$ which is always positive—the necessary condition (4.2a) is sufficient for the convergence (Ljung and Söderström, 1983, p. 214).

From this remark it is clear that (4.2b) does not serve in the step-variable form of the I-PLR, since in this case $\boldsymbol{R}(k)$ depends on $\hat{\boldsymbol{\beta}}(k)$ only. Rather, it is required in the lagged residual version of the I-PLR, which substitutes $\hat{a}_t(k)$ with $\tilde{a}_t(k) = y_t - \hat{\boldsymbol{\beta}}(k)'\hat{\boldsymbol{z}}_t(k-1)$, because $\boldsymbol{R}(k)$ would become a function of $\{\hat{\boldsymbol{\beta}}(1)\ldots\hat{\boldsymbol{\beta}}(k)\}$. These conclusions agree with those of Stoica *et al.* (1985) and Hannan and McDougall (1988).

As a final check on the reliability of the measure (4.7), we can assess why passivity conditions (4.2) are not necessary for NLS estimators.

R-NLS. Consider, for example, the stochastic gradient version of (3.7) with $\lambda = 1$:

$$\hat{\boldsymbol{\beta}}(t) = \hat{\boldsymbol{\beta}}(t-1) + \frac{1}{\operatorname{tr} S(t)}\, \hat{\xi}_t(t-1)\tilde{a}_t(t-1).$$

Now if we utilize the functional $J_t = a_t^2/2$ (proper for sequential optimizations), (4.7) becomes

$$\inf E\left[\frac{1}{\operatorname{tr} S(t)}\, \{G(B^{-1})z_t a_t\}' G(B) z_t a_t\right] > 0$$

and this holds for any $G(B)$ stable, positive real or not. The same conclusions can be drawn for (3.7) and for I-NLS, so that optimal methods always converge.

Proceeding in this way, we might still apply (4.7) to other algorithms; however, we now wish to investigate the problem of efficiency. Here, unlike the convergence, there is no general framework of analysis and every case must be treated with *ad hoc* techniques. We heuristically deal with two cases.

R-NLS. A very simple way to show (4.1) is that of 'solving' the recursions for $\hat{\boldsymbol{\beta}}(t)$ and then applying standard off-line calculations. Let $\lambda = 1$, $\hat{\boldsymbol{\beta}}(0) = \boldsymbol{0}$ and subtract $\boldsymbol{\beta}_0$ from (3.7a):

$$\begin{aligned} S(t)\{\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0\} &= S(t)\{\hat{\boldsymbol{\beta}}(t-1) - \boldsymbol{\beta}_0\} + \hat{\xi}_t(t-1)\tilde{a}_t(t-1) \\ &= S(t-1)\{\hat{\boldsymbol{\beta}}(t-1) - \boldsymbol{\beta}_0\} + \hat{\xi}_t\hat{\xi}_t'\{\hat{\boldsymbol{\beta}}(t-1) - \boldsymbol{\beta}_0\} + \hat{\xi}_t\tilde{a}_t. \end{aligned}$$

Now, summing both sides from $\tau = 0$ to $\tau = t$, we obtain

$$S(t)\{\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0\} = \sum_{\tau=1}^{t} \hat{\xi}_\tau(\tau-1)[\hat{\xi}_\tau(\tau-1)'\{\hat{\boldsymbol{\beta}}(\tau-1) - \boldsymbol{\beta}_0\} + \tilde{a}_\tau(\tau-1)]$$

but the expression in square brackets is just the Taylor expansion of $a_\tau(\boldsymbol{\beta}_0)$ in $\hat{\boldsymbol{\beta}}(\tau-1)$; hence

$$\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0 \approx tS(t)^{-1}\frac{1}{t}\sum_{\tau=1}^{t}\hat{\xi}_\tau(\tau-1)a_\tau.$$

Finally, since $\hat{\boldsymbol{\beta}}(t) \rightarrow \boldsymbol{\beta}_0$ a.s., the Slutsky theorem implies $\hat{\xi}_t(t-1) \rightarrow \xi_t$ a.s., and (4.1) can be shown as in the iterative case.

I-PLR. Given the non-linearity (in the parameters) of the TF model and the approximation $\xi_t \approx z_t$, the implicit objective function of the PLR method becomes $H_N(\boldsymbol{\beta}) = \Sigma_{t=1}^{N} z_t a_t/N$, concerning the correlation residuals–regressors. Following the approach of Spliid (1983), this may be very useful in proving (4.3). Indeed, assume that (3.4) converges and expand the empirical $\hat{H}(t)$, $t = (k = N)$, in $\boldsymbol{\beta}_0$:

$$\hat{H}(t) - H_0 \approx \frac{1}{t}\sum_{\tau=1}^{t}\left(\frac{\partial z_\tau}{\partial \boldsymbol{\beta}'}a_\tau + z_\tau\xi_\tau'\right)\{\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0\} \underset{t\rightarrow\infty}{\overset{p}{\rightarrow}} E(z_t\xi_t')\{\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0\}.$$

The result follows by assuming t large and ergodicity. In this situation, in fact, $E(\partial z_\tau/\partial \boldsymbol{\beta}' a_\tau) = 0$ because z_t contains lagged regressors and differentiation introduces a multiplication with a backward filter. Now, since

$$\lim E[t^{1/2}\{\hat{H}(t) - H_0\}]^2 = E(z_t z_t')\sigma^2$$

the asymptotic dispersion becomes

$$\lim_{t\to\infty} E[t^{1/2}\{\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0\}]^2 = E(z_t \xi_t')^{-1} E(z_t z_t')\sigma^2 E(\xi_t z_t')^{-1}$$

and the loss of efficiency in passing from (3.2) to (3.4) is thus well established. Note, however, that the conditions of convergence (4.2), under which the above analysis has been carried out, roughly means $G(B) \approx \boldsymbol{I}$; therefore in the recursive context an approximate estimator for the dispersion may be $\boldsymbol{P}(t)$, at least whenever $|\boldsymbol{P}(t)| > |\boldsymbol{Q}(t)|$.

Let us finally summarize the conclusions of this section. Despite the analysis of Spliid (1983), pseudolinear regression methods do not always converge; however, by utilizing algorithms in stochastic approximation form the severe conditions of convergence (4.2b) can be avoided. The previous theorem and the related measure (4.7) are concerned with global convergence, but for a certain initial value $\hat{\boldsymbol{\beta}}(0)$ the limit $\boldsymbol{\beta}_0$ may be a relative minimum. In practical terms, however, if $\hat{\boldsymbol{\beta}}(0)$ is yielded by the initiation of Section 2, we can reasonably think that $\hat{\boldsymbol{\beta}}(t)$, $t = (k = N)$, converges toward the absolute minimum $\boldsymbol{\beta}_0{}^*$. Indeed, the estimates $\hat{\boldsymbol{\delta}}(0)$, $\hat{\boldsymbol{\omega}}(0)$ of Step 2 are completely linear, and under regularity conditions $\hat{\boldsymbol{\phi}}(0)$, $\hat{\boldsymbol{\theta}}(0)$ of Step 4 are strongly consistent (Hannan and Rissanen, 1982). Lastly, the approach of Spliid (1983) is suitable for evaluating the efficiency of non-linear estimators without gradient.

5. AN INDUSTRIAL APPLICATION

In this section we apply and compare the algorithms described in Section 3 to the 'gas furnace' data of Box and Jenkins (1970). The example is concerned with a real industrial process described by the TF system

$$Y_t = \frac{\omega_0 + \omega_1 B + \omega_2 B^2}{1 - \delta_1 B} X_{t-3} + \frac{1}{1 - \phi_1 B - \phi_2 B^2} a_t.$$

The I-PLR algorithm (3.4) can easily be implemented on standard statistical software with OLS. Without Step 1 the initiation procedure of Section 2 is sketched as follows:

Estimate $y_t = \boldsymbol{\delta}' \boldsymbol{y}_{t-1} + \boldsymbol{\omega}' \boldsymbol{x}_{t-b} + n_t^*$	Generate $\hat{m}_t = \hat{\boldsymbol{\delta}}' \hat{\boldsymbol{m}}_{t-1} + \hat{\boldsymbol{\omega}}' \boldsymbol{x}_{t-b}$
Estimate $\hat{n}_t^* = \boldsymbol{\alpha}' \hat{\boldsymbol{n}}_{t-1}^* + a_t^*$	Generate$\hat{n}_t = y_t - \hat{m}_t$
Estimate $\hat{n}_t^* = \boldsymbol{\phi}' \hat{\boldsymbol{n}}_{t-1}^* + \boldsymbol{\theta}' \hat{\boldsymbol{a}}_{t-1}^* + \tilde{a}_t$	Generate $\hat{a}_t = \hat{n}_t - \hat{\boldsymbol{\phi}}' \hat{\boldsymbol{n}}_{t-1} - \hat{\boldsymbol{\theta}}' \hat{\boldsymbol{a}}_{t-1}$
Estimate $y_t = \boldsymbol{\beta}' \hat{z}_t + \tilde{a}_t$	Generate $\hat{z}_t = [\hat{m}_t, x_t, \hat{n}_t, \hat{a}_t]$

TABLE I

ITERATIVE OFF-LINE ESTIMATES AND *t* Ratios

Method	ω_0	ω_1	ω_2	δ_1	ϕ_1	ϕ_2	RSS
I-NLS	−0.531 (−7.1)	−0.378 (−3.6)	−0.518 (−4.8)	+0.550 (15.4)	1.533 (32.1)	−0.634 (−12.5)	16.6
I-PLR	−0.509 (−6.8)	−0.462 (−3.2)	−0.364 (−3.1)	+0.583 (31.0)	1.531 (32.4)	−0.633 (−12.7)	16.8

RSS, residual sum of squares.

In the generation steps, initial values of pseudolinear regressors can be obtained via back-forecasting or simply set equal to zero. As mentioned in the previous section, the mild condition of convergence (4.2a) requires I-PLR in step-variable form; therefore, after every iteration, the estimate $\hat{\boldsymbol{\beta}}(k)$ must be replaced by the convex combination $\tilde{\boldsymbol{\beta}}(k) = \alpha_k \hat{\boldsymbol{\beta}}(k-1) + (1-\alpha_k)\hat{\boldsymbol{\beta}}(k)$ $(0 \le \alpha_k \le 1)$.

Table I summarizes the iterative estimates obtained with the Pack package and a TSP program. We can see that if the former coincides with that of Box and Jenkins, the latter is not statistically different if we use some asymptotically normal tests, although a greater number of iterations is required to converge (α_k was 1/2).

Unlike recursive methods based on the Kalman filter framework (Bayesian approach, see Appendix A1), the implementation of the recursive algorithms of Section 3 is much simpler and the initial values to be specified are less awkward. $\hat{\boldsymbol{\beta}}(0)$ can be set equal to zero, $\boldsymbol{R}(0)$ and $\boldsymbol{S}(0)$ must be strictly positive definite, e.g. diagonal $\boldsymbol{I}/\rho$, and λ usually belongs to [0.9–1.0].

The specification of the pair (ρ, λ) depends on the aims of estimation and the nature of the system. As a general remark, we note from (3.6b) that these coefficients play a similar role on $\boldsymbol{P}(t)$ (both prevent it from vanishing). Moreover, the action of λ in discounting old observations enables (3.6a) to track changes of system parameters. These features establish the well-known trade-off between noise sensitivity and tracking capability typical of the Kalman filter framework; in terms of (3.6), this means in practice that $(\rho, \lambda) \rightarrow \boldsymbol{R}(t)^{-1} \rightarrow \Delta(t) = \hat{\boldsymbol{\beta}}(t) - \hat{\boldsymbol{\beta}}(t-1)$. Hence we have the following.

(i) Assuming a stationary system, if we want to estimate $\boldsymbol{\beta}$ consistently, initializing $\hat{\boldsymbol{\beta}}(0) = \boldsymbol{0}$ for example, we must have $\lambda = 1$ so that $\boldsymbol{P}(t) \rightarrow \boldsymbol{O}$. The value of ρ must be large enough to 'forget' $\hat{\boldsymbol{\beta}}(0)$ rapidly, but without hindering the vanishing of $\boldsymbol{P}(t)$.

(ii) Assuming an evolving system, if we desire to track the changes $\Delta(t)$, the action of the pair must be properly integrated, because if ρ is the initial variance (or energy), λ is the factor that spreads it into the sample. Intuitively the coefficients must be designed so as to have mild sampling error conditions uniformly in the record of data. Empirical experience has suggested $[0.01 \le \rho \le 0.30]$, $[0.95 \le \lambda \le 0.99]$.

TABLE II

FINAL ON-LINE ESTIMATES AND ($t = 296$)

ρ	ω_0	ω_1	ω_2	δ_1	ϕ_1	ϕ_2	RSS
0.01	−0.426	−0.392	−0.301	0.591	0.521	−0.167	159
0.05	−0.502	−0.433	−0.237	0.622	0.852	−0.068	61.7
0.1	−0.534	−0.475	−0.223	0.611	1.043	−0.199	43.4
0.3	−0.563	−0.581	−0.148	0.597	1.303	−0.423	27.3
0.6	−0.557	−0.682	−0.045	0.601	1.411	−0.526	22.5
1	−0.543	−0.766	+0.046	0.608	1.465	−0.573	20.5
2	−0.520	−0.871	+0.150	0.616	1.512	−0.624	18.8

In the following tables and figures we shall check these considerations empirically. Table II gives final values of $\hat{\boldsymbol{\beta}}(t)$ in PLR recursions with $[\hat{\boldsymbol{\beta}}(0) = \boldsymbol{0},\ \lambda = 1]$ and different values of ρ; the aim is to check the performance of on-line algorithms as off-line estimators. The results show the absence of convergence, in particular towards the corresponding values of Table I, which must be intended as average values of the sample. The behaviour is probably due to a latent non-stationarity (see the effect of ρ) but not to the non-passivity of the system; otherwise, I-PLR estimates should also diverge.

Table III provides final values of the residual sum of squares (RSS), corresponding to different pairs (ρ, λ) and $\hat{\boldsymbol{\beta}}(0) = \boldsymbol{0}$. It substantially confirms (i) the similarity of the role of (ρ, λ) in adapting the model to the data (i.e. on tracking capability), (ii) the range in which the pair must take on values to have mild variability of the estimates and (iii) practically unlimited improvement of the statistical performance outside this range.

Figures 1 and 2 show the recursive estimates obtained with $\hat{\boldsymbol{\beta}}(0) = \hat{\boldsymbol{\beta}}$ of Table I and the values $[\rho = 0.15, \lambda = 0.97]$ which best seem to satisfy the requirement of mild–uniform variability. They show the strongly non-stationary nature of the system; in particular, at $t = 260$ the parameters of $\omega(B)$ change sharply, followed by the others. This jump is difficult to understand; however, its persistence excludes accidental factors (e.g. measurement errors).

In comparing Figures 1 and 2 we note the greater smoothness of the NLS

TABLE III

FINAL ON-LINE ESTIMATES OF THE RESIDUAL SUM OF SQUARES

ρ	$\lambda = 0.98$	$\lambda = 0.96$	$\lambda = 0.94$	$\lambda = 0.92$	$\lambda = 0.90$	$\lambda = 0.80$	$\lambda = 0.70$
0.01	75.2	50.6	37.9	29.4	23.3	9.1	4.1
0.05	28.4	20.6	16.0	12.7	10.4	4.4	2.1
0.1	20.3	14.9	11.7	9.5	7.8	3.4	1.5
0.3	15.8	11.7	9.2	7.5	6.2	2.7	1.2
0.6	12.5	9.2	7.3	5.9	4.9	2.1	0.9
1	11.7	8.6	6.8	5.5	4.6	2.0	0.8
2	10.9	7.9	6.3	5.1	4.2	1.8	0.7

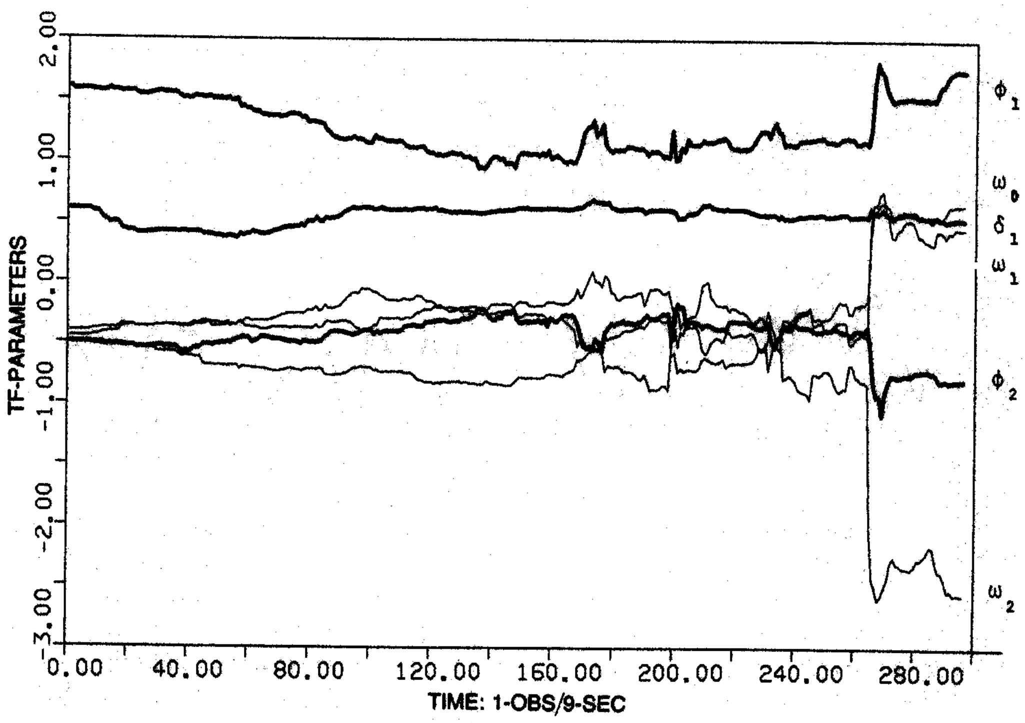

FIGURE 1. Time path of R-PLR estimates.

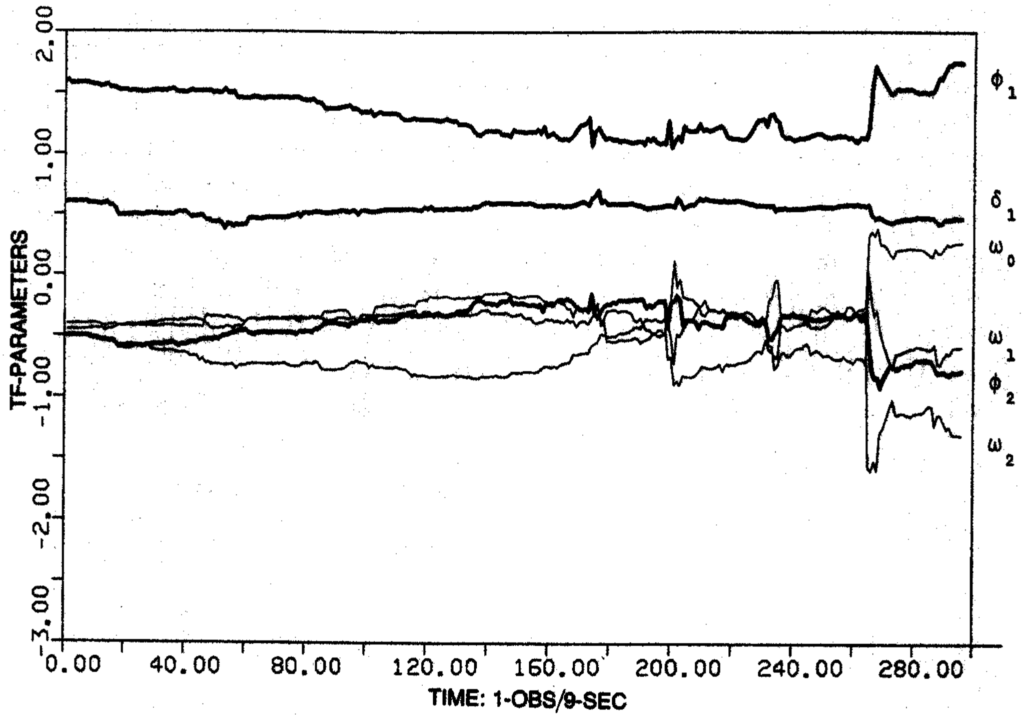

FIGURE 2. Time path of R-NLS estimates.

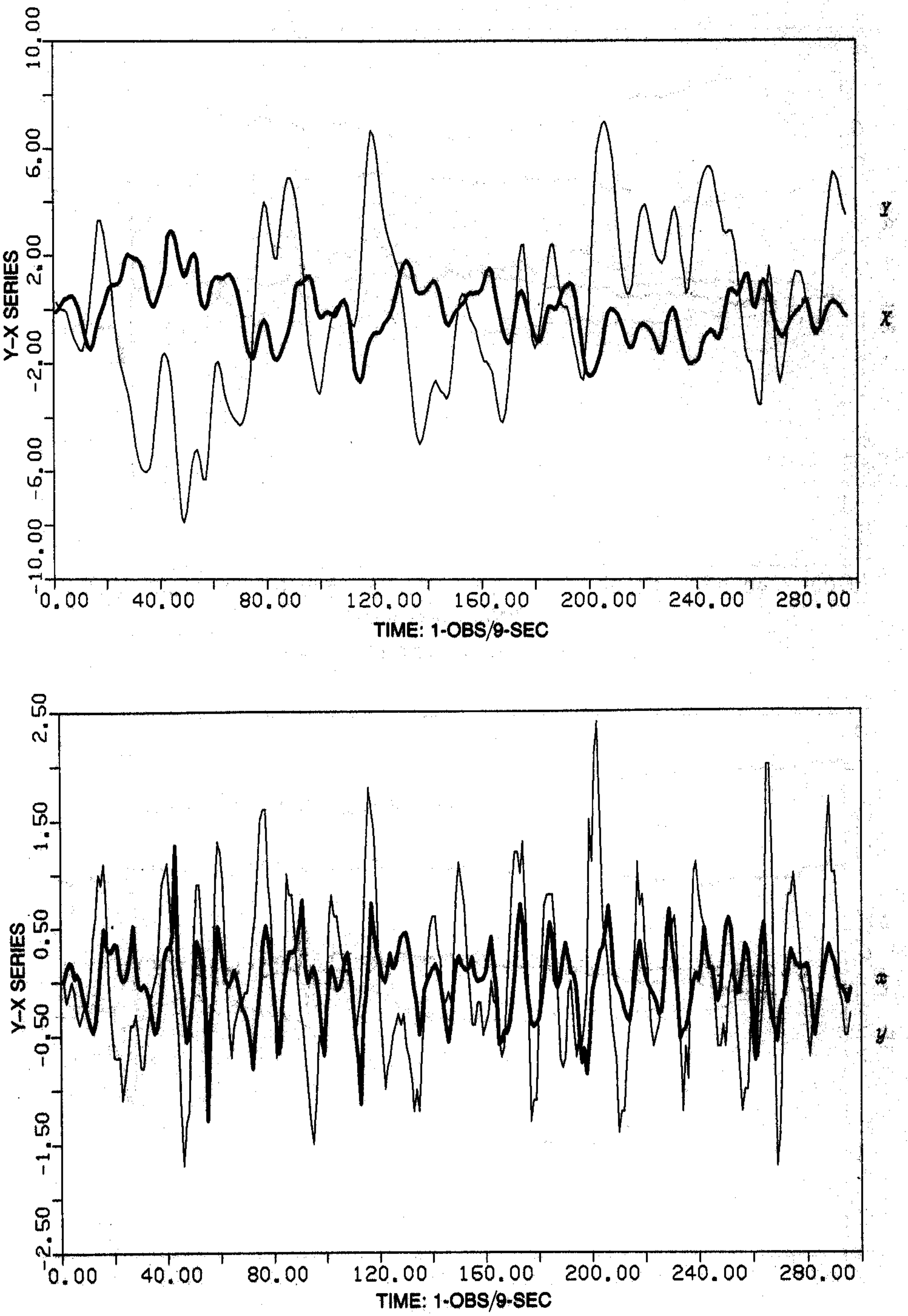

FIGURE 3. Original and differenced series.

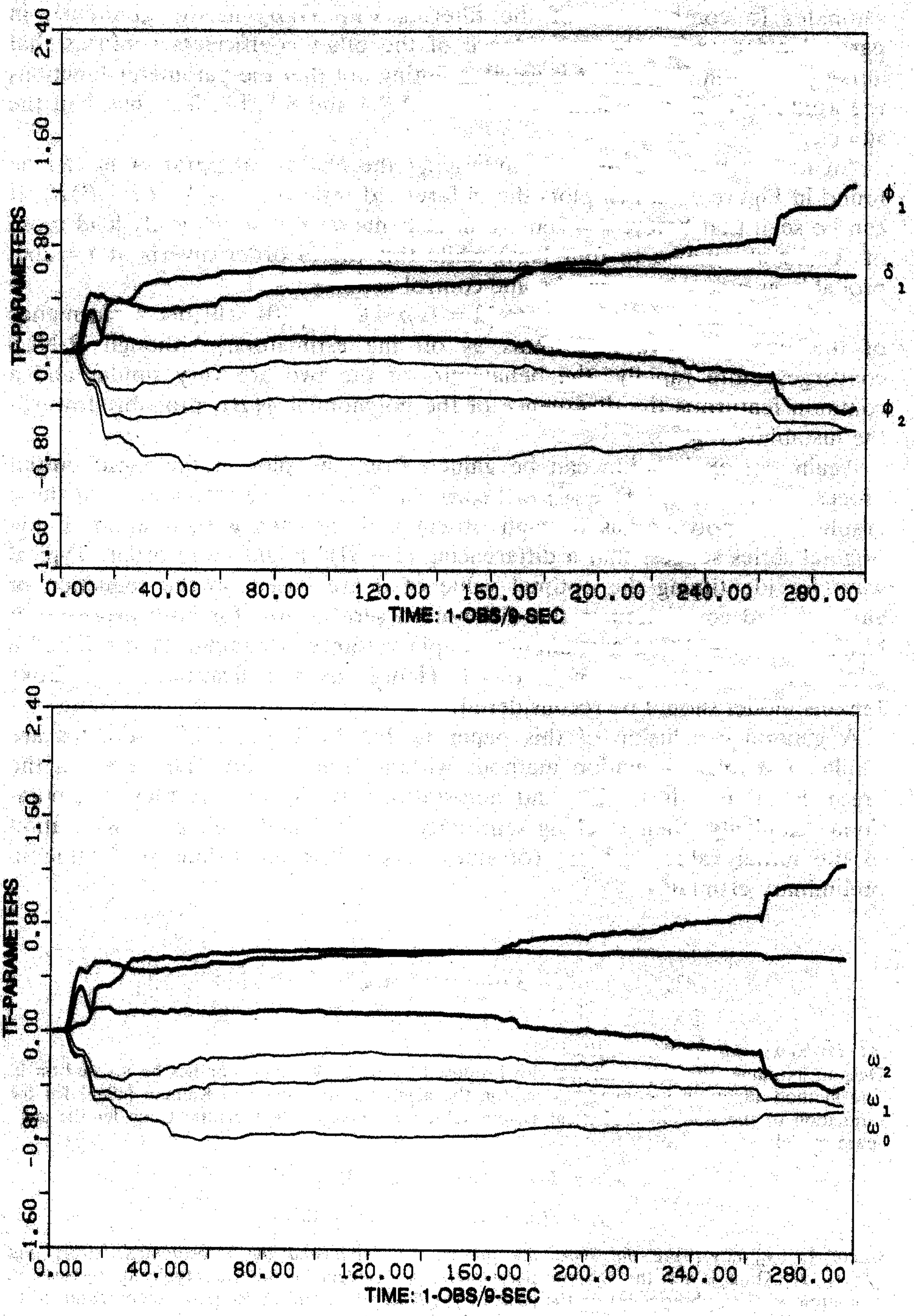

FIGURE 4. On-line as off-line estimates (PLR and NLS).

estimates (a consequence of the filtering with $G(B)$ in the gradient), in particular for $\omega(B)$; the coincidence of the other coefficients confirms that already found in Table I. It is worth pointing out that the parameter functions reported in the figures yield RSS values of 7.4 and 8.5, i.e. less than half the static case.

An insight into the reasons underlying the change of parameters can be found in Figure 3b which plots the differenced series $(1-B)Y_t$, $(1-B)X_t$. It can be seen that before $t=260$ the increments of Y systematically lead those of X (by about two to four lags) while this causal order inverts at $t=260$, probably as a consequence of some control action.

Lastly, in Figure 4 we evaluate $[\lambda=1, \rho=0.25, \hat{\boldsymbol{\beta}}(0)=\mathbf{0}]$ the performance of the two recursive algorithms as off-line estimators. Although R-NLS converges more rapidly, the behaviours of the two are very similar and a common feature is the divergence of the polynomial $\hat{\phi}(B)$, probably towards the instability region.

Again, an explanation can be gained from the plot of the input–output processes $(Y_t-\bar{Y})$, $(X_t-\bar{X})$ in Figure 3a. The relative smoothness of these graphs (like polynomials of high order) and the strong correlation of the original series suggest that a differencing $(1-B)^d$ might be in order. Typical ways for identifying the optimal value of d are based on the reduction of variance and correlation in the differenced series. Now for both processes it has been assessed that a minimum sample variance is attained at $d=2$ and a minimum serial correlation at $d=3$. Hence the specification of the Box–Jenkins model should be reconsidered.

A general conclusion of this paper is that PLR and NLS estimates are similar and thus estimation methods without gradient are valid even in the presence of near instability and non-stationarity. Moreover, their computational simplicity, their tracking sensitivity (on-line) and their general solution of the initial value problem (off-line) make PLR algorithms preferable in preliminary estimates.

APPENDICES

A1. Problems with the Kalman filter

Despite its name, this approach was not proposed by R. E. Kalman, nor has he insisted on it. The method adapts to parameter estimation the algorithm developed to Kalman (1963) for the estimation of the state in a physical system subject to measurement errors. Consider the state space model

$$\boldsymbol{x}_{t+1} = \boldsymbol{F}_t\boldsymbol{x}_t + \boldsymbol{G}_t\boldsymbol{a}_t, \qquad \boldsymbol{a}_t \sim \mathrm{IN}(\boldsymbol{o}, \boldsymbol{\Omega}_t)$$

$$\boldsymbol{y}_t = \boldsymbol{H}_t\boldsymbol{x}_t + \boldsymbol{e}_t, \qquad \boldsymbol{e}_t \sim \mathrm{IN}(\boldsymbol{o}, \boldsymbol{\Sigma}_t)$$

where $\{\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{a}_t, \boldsymbol{e}_t\}$ *are the state, the output, the input and the measurement error, and* $\{\boldsymbol{F}_t, \boldsymbol{G}_t, \boldsymbol{H}_t; \boldsymbol{\Omega}_t, \boldsymbol{\Sigma}_t\}$ are the system parameters; a problem of filtering typically consists in estimating $\boldsymbol{x}_t$ (non-observable) on the basis of $\boldsymbol{y}_t$ (observable) and of the parameters (assumed to be known from physical laws). The sequential solution derived by Kalman (1963) was based on the joint use of two predictors of the observed output:

$$\hat{x}_{t|t-1} = E[x_t|y_{t-1}, y_{t-2}, \ldots] = F_{t-1}\hat{x}_{t-1|t-1}$$
$$\hat{y}_{t|t-1} = E[y_t|y_{t-1}, y_{t-2}, \ldots] = H_t\hat{x}_{t|t-1}.$$

Next, Mayne (1963) and Harrison and Stevens (1976) proposed an adaptation of the Kalman algorithm to the parameter estimation of a system with unknown stochastic coefficients ϕ:

$$\phi_t = \phi_{t-1} + a_t, \qquad a_t \sim \text{IN}(o, \Omega)$$
$$y_t = z_t'\phi_t + e_t, \qquad e_t \sim \text{IN}(o, \sigma^2).$$

Treating ϕ as an unobservable state x and the regressors z as the observation matrix H, straightforward application of the Kalman equations gives

$$\hat{\phi}_t = (I - k_t z_t')\hat{\phi}_{t-1} + k_t y_t$$
$$k_t = (P_{t-1} + \Omega)z_t\{z_t'(P_{t-1} + \Omega)z_t + \sigma\}^{-1}$$
$$P_t = (I - k_t z_t')(P_{t-1} + \Omega).$$

Apart from the complexity of the solution, the forcing of the interpretation (parameters and state have a very different nature) and the practical problems of implementation (Ω, σ are not readily available), the derivation of the above, following the same steps as Kalman, is possible only in models with fixed regressors. In the case of stochastic z_t, the derivation fails at the step of computing the predictor

$$\hat{y}_{t|t-1} = E(z_t'\phi_t|y_{t-1}, y_{t-2}, \ldots) \neq z_t'\hat{\phi}_{t|t-1}.$$

This occurs whenever z_t has unknown mean, z_t, ϕ_t are correlated or a feedback $y_t \rightarrow z_t$ exists.

A2. Passivity conditions for PLR$_s$

Considering the ending polynomials of $G(B)$, we can summarize the requirements (4.2) as

$$\text{Re}\left[\frac{1}{\theta(z)} - \frac{c}{2}\right] > 0 \qquad |z| = 1, \qquad (0 \leq c \leq 1)$$

where, for $c = 0.1$, we have the necessary and sufficient conditions of convergence. The latter can also be expressed as

$$\left|\frac{1}{\theta(z)} - \frac{1}{2}\right| = \frac{|2 - \theta(z)|}{|2\theta(z)|} > 0 \Leftrightarrow |\theta(z) - 1| < 1 \qquad \forall |z| = 1.$$

Hence if $\theta(z)$ has first degree, the above holds on the whole invertibility region $[-1, +1]$. Instead, for second-order polynomials the condition becomes more severe (see Figure 5). (half the invertibility region) such that in many situations (4.2b) may be violated. Thus the proposal to

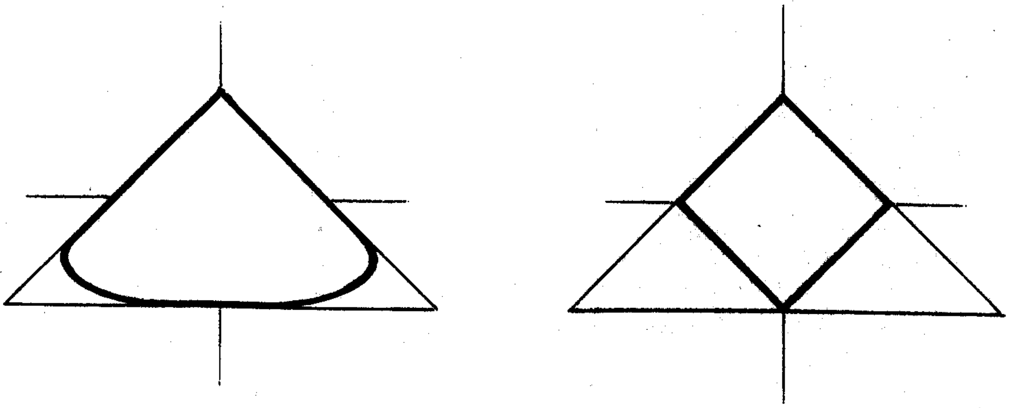

FIGURE 5. Regions (4.2a), (4.2b) for a second order polynomial.

use the PLR algorithms in stochastic approximation form (in which necessary and sufficient conditions coincide) has great practical value.

A3. Proof of formula at Step 1

From the assumptions of Section 2 we have $\gamma_{xy}(k) = E[y_t x_{t-k}] = 0$, $k < b$, $b \geq 0$; now since $n_t = y_t - \nu(B)x_{t-b}$, $\nu(B) = \Sigma_0^\infty \nu_i B^i$ it follows that

$$\begin{aligned}\gamma_{nn}(k) = E(n_t n_{t-k}) = E\{&(y_t - \nu_0 x_{t-b} - \nu_1 x_{t-b-1} - \nu_2 x_{t-b-2} - \ldots - \nu_k x_{t-b-k} - \ldots)\\ &(y_{t-k} - \nu_0 x_{t-b-k} - \nu_1 x_{t-b-k-1} - \ldots - \nu_k x_{t-b-2k} - \ldots)\}.\end{aligned}$$

Hence

$$\begin{aligned}\gamma_{nn}(k) = \gamma_{yy}(k) &- \nu_k\gamma_{xy}(b) - \nu_{k+1}\gamma_{xy}(b+1) - \ldots && \text{(A0)}\\ &- \nu_0\gamma_{xy}(b+k) - \nu_1\gamma_{xy}(b+k+1) - \nu_2\gamma_{xy}(b+k+2) - \ldots && \text{(A1)}\\ &+ \nu_0\nu_0\gamma_{xx}(k) + \nu_0\nu_1\gamma_{xx}(k-1) + \nu_0\nu_2\gamma_{xx}(k-2) + \ldots && \text{(A2)}\\ &+ \nu_1\nu_0\gamma_{xx}(k+1) + \nu_1\nu_1\gamma_{xx}(k) + \nu_1\nu_2\gamma_{xx}(k-1) + \ldots && \text{(A3)}\\ &+ \nu_2\nu_0\gamma_{xx}(k+2) + \ldots\end{aligned}$$

But $y_t = \nu(B)x_{t-b} + \psi(B)a_t$ so that

$$\begin{aligned}\nu_0\gamma_{xy}(b+k) &= \nu_0 E[\{\nu_0 x_{t-b} + \nu_1 x_{t-b-1} + \nu_2 x_{t-b-2} + \ldots + \psi(B)a_t\}x_{t-b-k}]\\ &= \nu_0\{\nu_0\gamma_{xx}(k) + \nu_1\gamma_{xx}(k-1) + \nu_2\gamma_{xx}(k-2) + \ldots\}\\ \nu_1\gamma_{xy}(b+k+1) &= \nu_1\{\nu_0\gamma_{xx}(k+1) + \nu_1\gamma_{xx}(k) + \ldots\}\end{aligned}$$

Hence, the first term of (A1) cancels with row (A2); the second term of (A1) cancels with row (A3), and so on. Only row (A0) does not vanish, being indeed the required formula.

ACKNOWLEDGEMENTS

I am grateful to the editor and the referee for their prompt and constructive reviews.

REFERENCES

Box, G. E. P. and Jenkins, G. M. (1970) *Time Series Analysis: Forecasting and Control.* San Francisco, CA: Holden-Day.

Gladyshev, E. A. (1965) On stochastic approximation. *Theory Prob. Appl. (USSR)* 10, 236–45.

Goodwin, C. G. and Sin, K. S. (1984) *Adaptive Filtering Prediction and Control.* Englewood Cliffs, NJ: Prentice-Hall.

Hannan, E. J. and Kavalieris, L. (1984) A method for ARMA estimation. *Biometrika* 72, 273–80.

—— and McDougall, A. J. (1988) Regression procedures for ARMA estimation. *J. Am. Statist. Assoc.* 83, 490–98.

—— and Rissanen, J. (1982) Recursive estimation of mixed autoregressive moving average order. *Biometrika* 69, 81–94.

Harrison, P. J. and Stevens, C. F. (1976) Bayesian forecasting. *J. R. Statist. Soc., Ser. B* 38, 205–48.

Kalman, R. E. (1963) New methods in Wiener filtering theory. In *Engineering Application of Random Functions Theory* (eds J. L. Bogdanoff and F. Kozin). New York: Wiley.

Kashyap, R. L., Blaydon, S. and Fu, K. S. (1970) Stochastic approximation. In *Adaptive Learning and Pattern Recognition, Methods and Applications* (eds J. M. Mendel and K. S. Fu). New York: Academic Press.

KUSHNER, H. J. and CLARK, D. S. (1978) *Stochastic Approximation Methods for Constrained and Unconstrained systems*. New York: Springer-Verlag.

LII, K. S. (1985) Transfer function model order and parameter estimation. *J. Time Series Anal.* 6, 153–69.

LJUNG, L. (1985) Estimation of parameter in dynamical systems. In *Handbook of Statistics*, Vol 5 (eds E. J. Hannan, P. R. Krishnaiah and M. M. Rao). Amsterdam: North-Holland.

—— and SØDERSTRØM, T. (1983) *Theory and Practice of Recursive Identification*. Cambridge, MA: MIT Press.

MAYNE, D. Q. (1963) Optimal non-stationary estimation of the parameters of a linear system with Gaussian inputs. *J. Electron Control* 14, 107–12.

PIERCE, D. A. (1972) Least squares estimation of dynamic-disturbance time series models. *Biometrika* 65, 297–303.

PLACKETT, R. L. (1950) Some theorems in least squares. *Biometrika* 37, 149–57.

POSKITT, D. S. (1989) A method for the estimation and identification of transfer function models. *J. R. Statist. Soc. Ser. B* 51, 29–46.

PRIESTLEY, M. B. (1983) The frequency domain approach to the analysis of closed-loop systems. In *Handbook of Statistics*, Vol. 3 (eds D. R. Brillinger and P. R. Krishnaiah). Amsterdam: North-Holland.

SHERIF, M. H. and LIU, L. M. (1987) Estimation of transfer function models using mixed recursive and nonrecursive methods. In *Control and Dynamic Systems*, Vol. 26 (ed. C. T. Leondes). Orlando, FL: Academic Press.

SOLO, V. (1978) Time series recursions and stochastic approximation. Ph.D. Dissertation Australian National University, Canberra.

SPLIID, H. (1983) Fast estimation of vector ARMAX models. *J. Am. Statist. Assoc.* 78, 843–49.

STOICA, P., SÖDERSTRÖM, T., AHLEN, A. and SOLBRAND, G. (1985) On the convergence of pseudolinear regression algorithms. *Int. J. Control* 40, 1429–44.

TSYPIN, Y. Z. (1971) *Adaptation and Learning in Automatic Systems*. New York: Academic Press.

YOUNG, P. (1984) *Recursive Estimation and Time Series Analysis*. New York: Springer-Verlag.