TIME SERIES RECURSIONS
AND STOCHASTIC APPROXIMATION

Victor Solo

A thesis submitted for the degree of

Doctor of Philosophy

at

The Australian National University

September 1978

The results presented in this thesis are my own, unless otherwise stated.

V. Solo

For

Emanuel, Sylvia, Julie and Paul

ACKNOWLEDGEMENTS

It is a pleasure to thank Professor E.J. Hannan for his encouragement and enthusiastic supervision. Thanks too, to Dr P.C. Young, who shared the supervision, for his guidance and persistent reminder of the 'Engineering' point of view.

I gratefully acknowledge the financial support of a Commonwealth Postgraduate Research Award.

Thanks to my friend Padmini for her good humour.

Finally, I thank Mrs B. Geary for an excellent typing job under difficult circumstances.

ABSTRACT

This thesis is concerned with the asymptotic theory, both convergence and Central Limit Theorem, for recursive estimators of time invariant parameters in Time Series models. This theory derives from a similar theory for classical Stochastic Approximation schemes so that these too, are given some consideration.

The thesis is divided into two parts. In Part I after an introductory chapter, a general approach is offered for the construction of recursions from a Prediction Error or Gaussian Likelihood viewpoint. This approach also allows other methods (for example, Model Reference) to be seen in context. Next, an heuristic argument is used to intuit the asymptotic properties of a general class of recursions (including the Prediction Error Recursions) and Time Series Stochastic Approximation schemes. The third chapter of Part I contains simulations illustrating some of the above ideas.

In Part II attention is turned to a rigorous analysis of various recursive and Stochastic Approximation schemes. In Chapter 4 a general regression problem is considered and strong convergence proved under various conditions. Chapter 5 is concerned with deriving limit laws for classical Stochastic Approximation in a general setting: in Chapter 6 this is extended to allow dependent noise. Chapter 7 contains a Central Limit Theorem and Invariance Principle for a well known Prediction Error Recursion which is seen to be asymptotically efficient. Finally, in Chapter 8, a proof is given of the convergence, without monitoring, of a slightly modified form of the RML_1 Recursion.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

NOTATION

α	vector of noise parameters
$A_i(\beta)$	Matrix of system parameters β
a.s.	almost surely
β	system parameters
CLT	Central Limit Theorem

$C_\alpha(L) = \sum_1^{n_c} C_i(\alpha)L^i$ where $C_i(\alpha)$ are matrices of noise parameters

D, $\mathcal{D}$	derivative operator
$e_k(\theta)$	one step ahead prediction error generated by a parametric model $M(\theta)$

$\tilde{\Phi}_k(\theta) = \partial y'_{k/k-}/\partial\theta = -\partial e'_k(\theta)/\partial\theta$

IP	Invariance Principle (in Chapter 2 infinite past)
K-W	Kiefer-Wolfowitz
L	lag or backwards operator
LIL	Law of the Iterated Logarithm
M-C	Monte-Carlo
m.d.s.	Martingale Difference Sequence
m.s.	mean square
o.d.e.	ordinary differential equation
p lim	limit in probability
$\xrightarrow{p}$	limit in probability
RLS	Recursive Least Squares
RML_1	Recursive Maximum Likelihood (Recursion) Scheme 1
RML_2	Recursive Maximum Likelihood (Recursion) Scheme 2
s.d.e.	Stochastic Differential Equation
SNR	Signal to Noise Ratio
w.l.o.g.	without loss of generality
w.p.1	with probability one
w.r.t.	with respect to

PART I

PREAMBLE

This thesis aims to investigate the asymptotic behaviour of Time Series Recursions and Stochastic Approximation (SA) *with particular emphasis on the second order properties, that is, asymptotic variances or Central Limit Theorem: the thesis is divided into two parts. In Part* I *the introductory chapter reviews some basic notions in recursive estimation and indicates, in an informal way, some of the basic ideas of the thesis. The remainder of Part* I *sacrifices rigour to provide a plausibility analysis of the behaviour of recursions (that is, convergence and asymptotic variances) as well as considering ways of constructing them. Some of these ideas are illustrated with simulation studies. In Part* II *a rigorous analysis is given of the asymptotic properties of some classical Stochastic Approximation schemes (allowing, for example, dependent or coloured noise) and of two simple but well known recursions. Except for Chapter* 1 *the two parts may be read separately.*

CHAPTER 1

INTRODUCTION

1.1 General

In recent years there has been an enormous interest in recursive methods of parameter estimation for Time Series models (Åström and Eykhoff, 1971, Young, 1974). Recursive estimation has been seen as a technique to go hand in hand with real time forecasting (Holst, 1977) or state estimation, to form part of an adaptive control scheme (Åström and Wittenmark, 1973, Young and Beck, 1974), to be used in pattern recognition (Kittler and Young, 1973) or to be used as an aid to model building where the parameter variation over an observation interval may suggest a cause of model inadequacy (Young, 1974).

While it seems that Gauss was the first to use recursive methods (Young, 1974), the modern use begins with Plackett (1950) and Kalman (1960). Let us therefore begin with a review of Plackett's algorithm and some of its extensions.

1.2 Recursive Least Squares

The recursive least squares (RLS) algorithm or Plackett algorithm is a sequential procedure for the computation of the least squares estimator of a regression coefficient. Suppose a sequence of noisy observations is available on a scalar process y_k together with noiseless or error free measurements of a related m dimensional process x_k . The process x_k might be chosen by the data analyst, its components consisting perhaps of sinusoids if y_k was periodic, or may be part of the environment, with, say, its components $x_{ki} = u_{k-i}$ where u_k is an input sequence into a dynamical system. The aim of regression analysis is to describe the relation between y_k and x_k in a linear fashion as

$$y_k = \mathrm{x}_k'\beta_0 + \varepsilon_k$$

where β_0 is a constant vector which would be termed a regression coefficient in the first example above or an impulse response vector in the second example. The sequence ε_k is an error or disturbance sequence. The RLS algorithm may be derived by manipulating the en-bloc estimator or by sequential optimisation of the criterion

$$\rho_N(\beta) = \tfrac{1}{2}\sum_1^N e_k^2(\beta)$$

with

$$e_k(\beta) = y_k - x_k'\beta \ .$$

The latter course is taken here. We observe

$$\rho_N(\beta) = \rho_{N-1}(\beta) + \tfrac{1}{2}e_N^2(\beta)$$

so that, noting $de_N/d\beta = x_N$,

$$d\rho_N/d\beta = d\rho_{N-1}/d\beta - x_N e_N(\beta) \ . \tag{1.1}$$

Further

$$P_N^{-1} = d^2\rho_N/d\beta d\beta' = d^2\rho_{N-1}/d\beta d\beta' + x_N x_N'$$

so

$$P_N^{-1} = \sum_1^N x_k x_k' \ . \tag{1.2}$$

Next, denoting the estimator by $\hat{\beta}_N$, a Taylor series for $d\rho_N/d\hat{\beta}_N$ (= 0) yields

$$\begin{aligned}\hat{\beta}_N &= \hat{\beta}_{N-1} - P_N d\rho_N/d\hat{\beta}_{N-1} \\ &= \hat{\beta}_{N-1} + P_N\left(de_N/d\hat{\beta}_{N-1}\right)e_N\left(\hat{\beta}_{N-1}\right) \ .\end{aligned} \tag{1.3}$$

Now (1.2), (1.3) may be put in fully recursive form by means of the Matrix Inversion Lemma (Bodewig, 1956), that is,

$$P_N = P_{N-1} - P_{N-1}x_N x_N' P_{N-1}/\left(1+x_N' P_{N-1} x_N\right) \ . \tag{1.4}$$

This expression is immediately verified by multiplying on the left by P_N^{-1} and on the right by its equivalent $P_{N-1}^{-1} + x_N x_N'$. Now note

$$P_N x_N = P_{N-1}x_N/\left(1+x_N' P_{N-1} x_N\right)$$

so that the fully recursive form is $\left(\text{denoting } \hat{e}_N = y_N - x_N'\hat{\beta}_{N-1}\right)$

$$\hat{\beta}_N = \hat{\beta}_{N-1} + P_{N-1}x_N\hat{e}_N/\left(1+x_N'P_{N-1}x_N\right) , \tag{1.5}$$

$$P_N = P_{N-1} - P_{N-1}x_Nx_N'P_{N-1}/\left(1+x_N'P_{N-1}x_N\right) . \tag{1.6}$$

Suppose the regressor sequence x_k is non stochastic and the disturbance sequence ε_k is an uncorrelated process with constant variance σ^2 . Then if we multiply (1.3) by P_N^{-1} and sum we obtain

$$\hat{\beta}_N - \beta_0 = P_N \sum_1^N x_k\varepsilon_k$$

so that

$$\operatorname{var}\left(\hat{\beta}_N\right) = \sigma^2 P_N .$$

It is immediately clear that we need only require the smallest eigenvalue of P_N to tend to ∞ with N to ensure $\hat{\beta}_N$ converges in mean square.

It is also clear that the normalising factor $1 + x_N'P_{N-1}x_N$ has a natural interpretation as the variance of the prediction error: indeed since

$$\hat{e}_N = y_N - x_N'\hat{\beta}_{N-1} = \varepsilon_N - x_N'\left(\hat{\beta}_{N-1}-\beta_0\right)$$

it follows that

$$\operatorname{var}\left(\hat{e}_N\right) = \sigma^2\left(1+x_N'P_{N-1}x_N\right)$$

(*cf.* Brown, Durban and Evans, 1975). Another way to see this is directly from (1.5) where, taking variances, gives

$$\sigma^2 P_N = \sigma^2 P_{N-1} - P_{N-1}x_Nx_N'P_{N-1}\operatorname{var}\left(\hat{e}_N\right)/\left(1+x_N'P_{N-1}x_N\right)^2 .$$

Returning to (1.3) we see that $\hat{\beta}_N$ is obtained from $\hat{\beta}_{N-1}$ plus a correction term containing the new information and consisting of a variance sequence P_N , gradient $de_N/d\hat{\beta}_{N-1}$ and an error $e_N\left(\hat{\beta}_{N-1}\right)$. Provided any linear combination of the sequence x_k has infinite energy (that is, the smallest eigenvalue of P_N^{-1} tends to ∞) then $P_N \to 0$ so that new information has less and less influence. It may be remarked that recursive estimators for time varying parameters are characterised by a variance sequence that does not decay to zero so that new information is not discounted (see Young, 1974). These algorithms are based on the well known extension of the RLS algorithm, the Kalman Filter (Kalman, 1960; Kailath, 1970; Duncan and Horn, 1972). However the emphasis of this thesis is on time invariant parameter estimation so that these important time varying recursions are mentioned only briefly (in Chapter 2).

RECURSIVE LEAST SQUARES AND TIME SERIES RECURSIONS. We now turn briefly to consider how RLS can be used to generate recursions for the parameters of a Time Series Model. For a Time Series Model the $e_k(\beta)$ sequence becomes a formula for the innovations or prediction error sequence of the Time Series Model. Thus consider the TF (transfer function) model defined by

$$e_k(\beta) = y_k - x_k(\beta) = y_k - \left(1+A(L)\right)^{-1}B(L)u_k$$

where L is the lag or backwards operator

$$\beta = \left(a_1 \ldots a_{n_a} \; b_1 \ldots b_{n_b}\right)' ,$$

$$B(L) = \sum_1^{n_b} b_i L^i ,$$

and so on, and u_k is the input or exogenous sequence. The basic idea in constructing recursions for Time Series Models has been to observe that all the usual linear models may be written as pseudo linear regressions, thus in the case just cited noting that

$$x_k(\beta) + A(L)x_k(\beta) = B(L)u_k$$

we see

$$e_k(\theta) = y_k - \varphi_k'(\theta)\theta$$

where

$$\varphi_k(\theta) = \left(-x_{k-1}(\beta) \ldots -x_{k-n_a}(\beta)u_{k-1} \ldots u_{k-n_b}\right) .$$

Thus we are led to generate the recursion

$$\beta_k = \beta_{k-1} + P_k\varphi_k e_k \ ,$$

$$P_k = P_{k-1} - P_{k-1}\varphi_k\varphi_k' P_{k-1}/\left(1+\varphi_k' P_{k-1}\varphi_k\right) \ ,$$

$$e_k = y_k - \varphi_k'\theta_{k-1} = y_k - x_k \ ,$$

$$\varphi_k = \left(-x_{k-1} \ \cdots \ -x_{k-n_a} u_{k-1} \ \cdots \ y_{k-n_b}\right) \ ,$$

$$x_k = \left(-x_{k-1} \ \cdots \ -x_{k-n_a} u_{k-1} \ \cdots \ u_{k-n_b}\right)' \varphi_{k-1} \ .$$

Related to the algorithm just given is a class of Instrumental Variable Recursions where the P_k, e_k equations are replaced by

$$P_k = P_{k-1} - P_{k-1}\varphi_k z_k' P_{k-1}/\left(1+\varphi_k' P_{k-1}\varphi_k\right) \ ,$$

$$e_k = y_k - z_k'\beta_{k-1} \ ,$$

where z_k is a regressor sequence

$$z_k = \left(-y_{k-1} \ \cdots \ -y_{k-n_a} u_{k-1} \ \cdots \ u_{k-n_b}\right) \ .$$

For this recursion the φ_k sequence can be regarded as a vector of instrumental variables chosen to be as highly correlated as possible with the noise-free vector $\varphi_k\left(\beta_0\right)$ while being uncorrelated with the possibly dependent noise: for further discussion see Young (1974).

Another way to construct recursions is to aim at sequential minimisation of a criterion such as

$$N^{-1} \sum_1^N e_k^2(\theta) \ .$$

The resulting recursions would be given by an expression such as (1.3) together with an algorithm for generating $de_k/d\hat{\theta}_{k-1}$. These recursions, which will be called Prediction Error recursions are the subject of Chapter 2. It is one of these recursions (the RML_2 recursion (Söderström *et al.*, 1974)) that is investigated rigorously in Chapter 7 and shown to be aysmptotically efficient. Some implications of using the above Prediction Error criterion for parameter estimation are explored in Section 1.5.

1.3 Stochastic Approximation

Recall that the recursive estimator is a solution to the problem of sequentially minimising the function $\rho_N(\beta)$ or alternatively to the problem of finding the zero of the equation $d\rho_N/d\beta = 0$. With this in mind the connexion with Stochastic Approximation (SA) is apparent. The Robbins-Monro (R-M) scheme for example (Robbins and Monro, 1951) is concerned with finding the solution θ of the equation

$$M(x) = 0$$

where the form of $M(x)$ is unknown but where for each x, a noisy observation $Y(x)$ is available and $E\big(Y(x)\big) = M(x)$. Then an approximating sequence X_n is generated by

$$X_{k+1} = X_k + a_k' K Y_k \tag{1.6}$$

where $Y_k = Y\big(X_k\big)$ while a_k' is a decreasing gain sequence of positive numbers and K a gain constant, both chosen by the experimenter, thus $a_k' = k^{-1}$ is a common choice. The related problem of finding the maximum of a regression function was considered by Kiefer and Wolfowitz (1952).

As originally presented SA had little to do with regression parameter estimation; the θ above being, for example, a critical dose of insectiside given to an insect that produces a given response (taken as zero). Ho (1962) pointed out the relation between RLS and SA while Albert and Gardner (1967) applied SA to the estimation of the parameters of a nonlinear regression. The Control Engineering literature abounds with examples of SA used in the system identification (that is, parameter estimation) problem (Saridis, 1974, 1977, Chapters 3, 6). Mostly these Time Series SA's are similar to the recursions introduced at the end of the last section but with the P_k matrix replaced by $Ka_k'I$ thereby cutting down the computational load. Tsypkin (1971) has given a thorough coverage of these schemes; his work is further discussed in Chapter 2.

To gain some insight into the properties of recursions and SA schemes and to expose some of the ideas of this thesis, let us consider the convergence behaviour of the R-M scheme. Write $Z_k = Y_k - M\big(X_k\big)$ so that Z_k has zero mean and is the noise on the measurement of $M\big(X_k\big)$. Also, to save notation denote $a_k = Ka_k'$ then (1.6) becomes

$$X_{k+1} = X_k + a_k M(X_k) + a_k Z_k \ . \tag{1.7}$$

Recalling the point made at the beginning of this section, that $M(x)$ should be compared with the gradient $d\rho_N/d\beta$ we see that the SA scheme differs from the recursion (1.3) in that the recursion uses an instantaneous value for the gradient $d\left(\frac{1}{2}e_N^2\right)/d\beta$.

Consider now the Taylor series

$$M(X_k) = M'(\overline{X}_k)(X_k-\theta)$$

for some $\overline{X}_k$ with $|\overline{X}_k-\theta| < |X_k-\theta|$. Now write (1.7) as

$$X_{k+1} - \theta = (X_k-\theta)\left(1-a_k M'(\overline{X}_k)\right) + a_k Z_k \ . \tag{1.7a}$$

This reveals X_k as the output of a time varying filtering operation on Z_k . It is clear that we must in general require $\sum_1^\infty a_k = \infty$ to ensure the filtering in non-degenerate.

Now if $X_k \to \theta$ then $M'(\overline{X}_k) \to M'(\theta)$ so for the purpose of theoretical analysis of the SA scheme we are lead then, to introduce the sequence U_k which is the output of the difference equation

$$U_{k+1} = U_k\left(1-a_k M'(\theta)\right) + a_k Z_k \ . \tag{1.8}$$

Now U_k , being the result of a filtering of Z_k will have relatively 'smooth' sample paths. Indeed if Z_k is a white noise (or even stationary coloured noise) $M'(\theta) = 1$, $a_k = k^{-1}$ then by the strong law of large numbers (for stationary processes)

$$U_k = k^{-1} \sum_1^k Z_s \to 0 \quad \text{w.p.1} \ .$$

In general it can be shown, for example, that $\sum_1^\infty a_k^2 < \infty$ ensures $U_k \to 0$ w.p.1 even if Z_k is a coloured noise . This is discussed in Chapters 5 and 6.

Next define the fluctuation $\hat{X}_k = X_k - U_k$ and subtract (1.8) from

(1.7) to obtain

$$\hat{X}_{k+1} = \hat{X}_k + a_k M(X_k) + M'(\theta) a_k U_k \ .$$

Consider further the Taylor series about $\hat{X}_k$;

$$M(X_k) = M(\hat{X}_k + U_k) = M(\hat{X}_k) + U_k M'(X_k^*)$$

for some X_k^* with $|X_k^* - X_k| < |U_k|$. If we add the usual assumption $|M'(x)| < \overline{K} < \infty$ for all x , $\overline{K}$ a constant, it follows we have a stochastic difference equation

$$\hat{X}_{k+1} = \hat{X}_k + a_k M(\hat{X}_k) + a_k U_k \left(M'(\theta) + M'(X_k^*)\right) \tag{1.9}$$

driven by a decaying forcing function. If we can show $\hat{X}_k \to \theta$ we may conclude $X_k \to \theta$.

Now denote

$$c_k = M'(\theta) + M'(X_k^*)$$

and notice that c_k is, by assumption, uniformly bounded in k . Set $\tau_k = \sum_1^k a_s$, $d\tau_k = \tau_k - \tau_{k-1}$, and so on, so that we can rewrite (1.9) as

$$d\hat{X}_k = M(\hat{X}_k) d\tau_k + c_k U_k d\tau_k \ . \tag{1.10}$$

With the filtering nondegeneracy condition $\sum_1^\infty a_k = \infty$ in mind, we have a time scale $\tau_k = \sum_1^k a_s \to \infty$, so that if we ensure the increments $d\tau_k = a_k$ vanish as $k \to \infty$, then, making the notational transformation $\hat{X}_k \leftrightarrow \hat{X}(\tau)$, we are led to study the differential equation with decaying forcing function

$$d\hat{X}(\tau)/d\tau = M(\hat{X}(\tau)) + c(\tau)U(\tau) \ . \tag{1.10a}$$

The stability of this equation will determine the convergence of $\hat{X}_k$. Thus it is clear, for example, that if the differential equation

$$d\hat{X}(\tau)/d\tau = M(\hat{X}(\tau))$$

is exponentially stable, then the perturbed system (1.10a) will at least be asymptotically stable, that is, $\hat{X}(\tau)$ or $\hat{X}_k \to \theta$. (Definitions of these terms are given in many texts on differential equations; a popular reference is

Lasalle and Lefschetz, 1963.)

The idea of subtracting off the U_k sequence reveals clearly why a deterministic ordinary differential equation (o.d.e.) determines the behaviour of the SA scheme and why the result holds equally well for schemes with dependent noise. The above notions and their ramifications are discussed rigorously in Chapters 5 and 6. The idea of using a deterministic o.d.e. to determine properties of SA schemes is presented (implicitly) in Khasminskii and Nevel'son (1972).

However Ljung (1974, 1975, 1977b) seems to have been the first to use this idea explicitly both for SA and recursions: see also Kushner (1976a, 1976b, 1977). Except for Kushner, the above authors have, in part, used Lyapunov methods to discuss the stability of the o.d.e.'s. The use of Stochastic Lyapunov functions (with SA) is implicit in Blum (1954) and Gladyshev (1969) but seems to have been first used explicitly by Khasminskii (1969 - reference unavailable to the author) and Nevel'son and Khasminskii (1973/1976). General discussions of SA particularly in relation to system identification are given by Young (1976), Saridis (1977). Also a wide ranging discussion of SA in an identification context was given by Tsypkin (1971): this work is further discussed in Chapter 2.

1.4 Second Order Properties of Stochastic Approximation

An important part of this thesis is the investigation of second order properties (asymptotic variances) of SA's and recursions. The basic ideas can be illustrated by pursuing the example of Section 1.3. Recall equation (1.7a),

$$X_{k+1} - \theta = (X_k-\theta)\left(1-a_k M'(\overline{X}_k)\right) + a_k Z_k .$$

To simplify this heuristic discussion take $a_k = k^{-1}$. Then if we multiply throughout (1.7a) by $\sqrt{k+1}$ and call $Q_k = \sqrt{k}\,(X_k-\theta)$, we obtain

$$Q_{k+1} - Q_k \simeq \left((\sqrt{k+1}-\sqrt{k})/\sqrt{k}-M'(\overline{X}_k)\right)Q_k/k + Z_k/\sqrt{k} . \tag{1.11}$$

Now set

$$W_k = \sum_1^k Z_s/\sqrt{s} \ ,$$

$$\tau_k = \sum_1^k s^{-1} \simeq \ln k \ .$$

If we suppose Z_k is a white noise sequence with unit variance it is clear that W_k resembles a Wiener process (McCarty, 1974 or Arnold, 1974 have very readable accounts of this and related topics) on a time scale τ_k. Denoting $dW_k = W_k - W_{k-1}$, $d\tau_k = \tau_k - \tau_{k-1} = k^{-1}$ and so on, and supposing $X_k \to \theta$ w.p.1 so that $\overline{X}_k \to \theta$ w.p.1 we can rewrite (1.11) as

$$dQ_{k+1} \simeq \left(\tfrac{1}{2}-M'(\theta)\right)Q_k d\tau_k + dW_k \ . \tag{1.12}$$

This suggests Q_k resembles the solution of the stochastic differential equation (s.d.e.)

$$dQ(\tau) = \left(\tfrac{1}{2}-M'(\theta)\right)Q(\tau)d\tau + dW(\tau) \ .$$

The limit variance, v, of this s.d.e. or diffusion is well known to satisfy the Lyapunov equation (Arnold, 1974, p. 134)

$$v\left(\tfrac{1}{2}-M'(\theta)\right) + \left(\tfrac{1}{2}-M'(\theta)\right)v = 1 \ . \tag{1.13}$$

So

$$v = \left(1-2M'(\theta)\right)^{-1}$$

which is also well known to be the correct limit variance for the R-M scheme (the basic reference here is Sacks, 1958).

This idea of representing the normalised output of the SA scheme or recursion as the solution to a s.d.e. was developed independently by the author and forms the basis of part of the heuristic discussion of Chapter 2. However it is clear that the idea is well known and has recently been discussed rigorously (for SA) by Nevel'son and Khasminskii (1973/1976) and also Kushner (1978a). In Chapters 5, 6 (SA) and 7 (recursions) this idea is given a rigorous expression. It should be remarked, that while Chapter 2 was written before the author was aware of the above two references, the relevant parts of Chapters 5, 6, 7 have been written with the benefit of these works.

1.5 Some Implications of Parameter Estimation by a Prediction Error Criterion

At the end of Section 1.2 it was pointed out that Time Series recursions may be constructed by sequential minimisation of a likelihood-like criterion namely

$$N^{-1} \sum_{1}^{N} e_k^2(\theta)$$

where $e_k(\theta)$ is a formula for the innovations sequence or prediction error sequence of the Time Series model. In this section we formulate the modelling problem and investigate some implications of using a Prediction Error Criterion.

Let u_k, y_k denote input and output measurements (scalar for ease of discussion) on a dynamic system S. System modelling is concerned, *inter alia*, with describing the input/output relation by a set of difference or differential equations. In modelling the dynamic system S one considers a class of models M and chooses a best approximation to S from among the members of M. In parametric modelling the set of models M is parametrised by a vector θ comprising a vector β of system parameters and a vector α of noise parameters. The vector β could be either or both of; a collection of physically meaningful parameters; the coefficients of a polynomial operator in a black box description of S.

A common method of choosing the approximation mentioned above is by computing the error sequence between the system output and the model output and varying θ until this sequence is white (that is, has flat spectrum) or else so as to minimise the average squared error. This is usually termed the model reference (MR) method (Landau, 1974). If the model output is properly reckoned as a *predicted output* then the resulting stochastic MR method is asymptotically the same as the Gaussian likelihood approach (as is clear by viewing the likelihood as an iterated product of conditional dnesities, each factor then having in its exponent, the squared prediction error weighted by its variance: see Schweppe, 1965). These methods have been referred to as Prediction Error (PE) methods (Ljung, 1976).

For this PE approach then, a linear model $M(\theta)$ is a rule for one step ahead prediction from past y_k, u_k values by a formula such as

$$\hat{y}_{k/k-}(\theta) = -N_\theta(L)y_k + M_\theta(L)u_k \tag{1.14}$$

where

$$N_\theta(L) = \sum_1^\infty N_i(\theta)L^i$$

and so on, and L is the lag or backwards operator. (To simplify discussion the infinite past is assumed available.) For a model that evolves in continuous time the second term would be replaced by

$$\int_0^\infty M_s(\theta)u(k-s)ds \ .$$

Defining the model innovations process or prediction error as

$$e_k(\theta) = e_{k/k-}(\theta) = y_k - \hat{y}_{k/k-}(\theta) \tag{1.15}$$

expression (1.14) may also be written in recursive form as

$$\hat{y}_{k/k-}(\theta) = T_\beta(L)u_k + K_\theta(L)e_k(\theta) \tag{1.16}$$

or

$$e_k(\theta) = \left(1+K_\theta(L)\right)^{-1}\left(y_k - T_\beta(L)u_k\right) \tag{1.17}$$

where

$$1 + K_\theta(L) = \left(1-N_\theta(L)\right)^{-1} \ ,$$

$$T_\beta(L) = \left(1+K_\theta(L)\right)M_\theta(L) \ .$$

The modelled innovations process is thus obtained by a filtering (whitening) of the output error. This is hardly surprising since $1 + K_\theta(L)$ will, by definition, be a model of a Wiener filter, that is, of a spectral factor of the spectrum of $y_k - T_\beta(L)u_k$. Note that $T_\beta(L)$ is supposedly a function only of system parameters β since clearly $T_\beta(L)$ is an approximation to the true system transfer function $dF_{yu}(\omega)/dF_u(\omega)$: where, for example, $dF_{yu}(\omega)$ is the observed cross spectrum between output and input respectively (we write $dF_u(\omega)$ rather than $f_u(\omega)d\omega$ to allow that the input sequence might consist of sums of sinusoids; thus for a sinusoid $u_k = A \sin \omega_0 k$, $\omega_0 \neq 0$, $dF_u(\omega)$ is a jump of size $\frac{1}{2}A^2$ at $\pm\omega_0$ and zero elsewhere).

Naturally $\hat{y}_{k/k-}(\theta)$ or $e_k(\theta)$ will not be computed by the formulae (1.16), (1.17). The point is however that the above description (or something like it) provides a single method for representing all the usual linear models: see Ljung (1976) for some examples.

Now the prediction error estimator will be chosen by minimising an expression such as

$$N^{-1} \sum_{1}^{N} e_k^2(\theta) \ .$$

We expect that this will converge to

$$E\left\{e_k^2(\theta)\right\} = \int_{-\pi}^{\pi} dF_e(\omega/\theta)$$

where, from expression (1.17),

$$dF_e(\omega/\theta) = \left[1+K_\theta\left(e^{i\omega}\right)\right]^{-1}\left\{dF_y(\omega)-T_\beta\left(e^{i\omega}\right)dF_{yu}(\omega)\right.$$
$$\left.-d\bar{F}_{yu}(\omega)T_\beta\left(e^{-i\omega}\right)+T_\beta\left(e^{i\omega}\right)T_\beta\left(e^{-i\omega}\right)dF_u(\omega)\right\}\left[1+K_\theta\left(e^{-i\omega}\right)\right]^{-1} .$$

We can gain some insight into this expression by reorganising it as follows: add and subtract $|dF_{yu}(\omega)|^2/dF_u(\omega)$ to the term inside the curly brackets to obtain

$$E\left\{e_k^2(\theta)\right\} = \int_{-\pi}^{\pi} \left(dF_y(\omega)/dF_n(\omega/\theta)\right)\left(1-|\gamma_{yu}(\omega)|^2\right)d\omega$$
$$+ \int_{-\pi}^{\pi} \left(dF_u(\omega)/dF_n(\omega/\theta)\right)\left|dF_{yu}(\omega)/dF_u(\omega)-T_\beta\left(e^{i\omega}\right)\right|^2 d\omega$$

where

$$dF_n(\omega/\theta) = \left|1+K_\theta\left(e^{i\omega}\right)\right|^2$$

is the spectrum of the noise whitening filter and where

$$|\gamma_{yu}(\omega)|^2 = |dF_{yu}(\omega)|^2/dF_u(\omega)dF_y(\omega)$$

is the squared coherency between output and input.

Thus by choosing a value for θ by minimising $E\left\{e_k^2(\theta)\right\}$ we minimise two terms. The first term, totals over all frequencies, the ratio of the system residual spectrum (obtained from a spectral regression of y_k on u_k) to the model residual spectrum. With the second term we minimise the difference between the true transfer function (TF) and the modelled TF with a weighting that is a signal to noise ratio: namely the ratio of the input signal spectrum to the modelled residual spectrum.

We see then that the PE or Gaussian likelihood criterion has the intuitively satisfying property of taking the input signal/output noise ratio into account when modelling the TF relation.

It is apparent, furthermore, that since $T_\beta(L)$ will be a rational polynomial operator, then $dF_{yu}(\omega)/dF_u(\omega)$ must be nonzero at a large enough number of different frequencies to ensure the second term has a well defined minimum with respect to the system parameters β. This is usually expressed as the so-called persistently exciting condition (Åström and Bohlin, 1965). Finally, the second term is small when the modelled TF follows the peaks or *dominant modes* in the true TF except that these peaks are weighted by a signal to noise ratio reflecting confidence in their existence.

CHAPTER 2

A UNIFIED APPROACH TO RECURSIVE PARAMETER ESTIMATION

2.1 Introduction

The aim of this chapter is to develop a unified approach to the Recursive, Stochastic Approximation (SA) and Model Reference methods for estimating the time invariant parameters of a lumped model of a dynamic system. It is also shown how, by sequential minimisation of an average prediction error, it is possible to construct recursive algorithms (prediction error recursions, PERs) for almost any lumped parametric model.

For a general class of recursions (including PERs and SAs) a plausibility analysis of local convergence is given by converting the recursion to a stochastic differential equation (s.d.e.). The asymptotic variances of the parameter estimators are also derived by converting the recursions to a different s.d.e. It is suggested that the PERs always satisfy the conditions for convergence and are asymptotically efficient in that they attain the Cramer-Rao lower bound. The analysis of convergence is repeated using discrete time arguments. The recursions are also analysed in the case when the system being modelled does not belong to the set of models being used to describe it.

General discussions of recursive estimation have been given by Tsypkin (1971) and Young (1976a) and some of the points made in these works are treated in a more formal analytic fashion here. In more specific terms, the literature on recursive estimation has developed in the following three directions.

(a) Model reference (MR) methods (for example, Landau, 1974). These are almost exclusively for deterministic systems and the analysis of their behaviour has been mainly by Lyapunov methods (Parks, 1966; Kudva and Narendra, 1974; also Anderson, 1977) and by hyperstability theory (Landau, 1974). Many of these Model Reference methods have rapid deterministic convergence properties but they do not necessarily function well in the presence of noise; in fact they will, in general, converge to a non-zero mean random variable. They are however suited to time-varying parameter estimation.

(b) Stochastic approximation methods (for example, Tsypkin, 1971; Saridis, 1974; and Young, 1976a). The analysis here has relied on methods of Dvoretsky (1956) and Tsypkin (1971, p. 56). Unlike the deterministic MR schemes, Stochastic Approximation was developed particularly for parameter estimation in the presence of noise (see for example, Young, 1976a). The rate of convergence of SA algorithms can, however, be rather slow with poor choice of a gain sequence.

(c) Recursive methods. This refers to those methods which produce recursions based on recursive least squares-like constructions (for example, Young, 1974; Söderström *et al.*, 1974). Two analyses have been offered here. One, due to Ljung (1977a, 1977b) which relates the behaviour of the recursion to the solution of an ordinary differential equation and then proceeds by Lyapunov methods; the other, due to Hannan (1976, 1978c) which uses a sophisticated application of probability theory to reduce the analysis to the consideration of a deterministic Lyapunov-like function. The recursive methods seem to combine the best aspects of the MR and SA approaches; they usually possess good convergence properties and function well in the presence of noise.

Almost all the existing analysis of the above methods is given under the assumption that the observed system S corresponds to the class M of models chosen to describe it. One aim of this chapter is to show how the recursions may be analysed when this is not the case. The main aim, however, is simply to consider the above three approaches from a unified point of view and so develop a general method for constructing recursions that converge and are likely to have good statistical efficiency.

In Section 2.2 the problem of recursive parameter estimation is formulated in a general fashion. As a result it becomes clear how to construct recursions (called Prediction Error Recursions (PER's)) for just about any (linear or nonlinear) lumped parameter model. Further, the relation to SA becomes clear.

In Section 2.3 the major problem in implementing these PER's is found to be the solution of a sensitivity difference or differential equation. Fortunately for discrete-time evolution, discrete-time measurements (DD) models the solution of these sensitivity equations involves only simple filtering operations. The remainder of Section 2.3 is devoted to elucidating the nature of the PER's.

In Section 2.4 their connections with recursions that have been presented previously is established. In the main body of the chapter it is shown how, by "converting" a general recursion to a stochastic differential equation (Section 2.5) the convergence properties (Section 2.6) and the asymptotic covariances of the recursive estimators (Section 2.7) can be easily derived by a plausibility argument. The analysis of convergence is repeated using discrete time arguments. Section 2.8 contains an analysis of the asymptotic behaviour of general recursions under model mis-specification.

2.2 Recursions

In the innovations description of modeling (see Ljung, 1976) a parametric model is a rule for predicting a system output y_k by a formula which is a function of the past of the system output and the system input u_k ; this function will be denoted by $\hat{y}_{k/k-}(\theta)$ * where θ consists of system parameters β , and noise parameters α . The system parameters β may be either or both of: a collection of physically meaningful parameters; the coefficients of a polynomial operator in a black box description of S . This description makes it clear that recursions for θ can be constructed by sequentially minimising the criterion

$$V_N(\theta) = N^{-1}\sum_{1}^{N} e_k'(\theta)\Sigma_0^{-1}e_k(\theta) \tag{2.1}$$

or, for measurements in continuous time

$$V_T(\theta) = T^{-1}\int_0^T e_t'(\theta)\Sigma_0^{-1}e_t(\theta)dt \tag{2.1a}$$

where

$$e_k(\theta) = y_k - \hat{y}_{k/k-}(\theta) \quad \text{and} \quad e_t(\theta) = y_t - \hat{y}_{t/t-}(\theta) \tag{2.2}$$

is the model innovations process or step ahead prediction error, and Σ_0 is an innovations covariance matrix. Alternatively, recursions may be constructed by finding the solutions to the equations,

$$dV_N(\theta)/d\theta = 0 \ , \quad dV_T(\theta)/d\theta = 0 \ . \tag{2.3}$$

These ideas apply equally well to stochastic and deterministic systems.

---

* The subscript $k-$ allows for continuous time evolution - discrete time data (CD) models where the input may be measured in continuous time. Also, for continuous time output measurements the subscript k will be replaced by t .

The above constructions can be accomplished in a number of ways.

(a) Stochastic approximation. For constant parameters, the sequential optimization could be solved by the procedure of Kiefer, Wolfowitz (1952) while the alternative (2.3) could be solved by the Robbins-Monro procedure (1951). Since gradients are computable, these lead to the same technique and for DD models they were effectively given by Tsypkin (1971); in Tsypkin's analysis $E\left[e_k'(\theta)\Sigma_0^{-1}e_k(\theta)\right]$ is termed $E\left(Q(x_k, \theta)\right)$.

(b) By analogy with recursive least squares. These methods (Young, 1974) are based on the observation that for all the usual linear time-series models the one step ahead predictor can be written as a pseudo-linear regression, namely,

$$\hat{y}_{k/k-1}(\theta) = \Phi_k'(\theta)\theta \tag{2.4}$$

where $\Phi_k(\theta) = \varphi_k(\theta) \otimes I$ and $\varphi_k(\theta)$ is a vector of 'regressors' obtained from an equation of the form

$$\varphi_k(\theta) = A(\theta)\varphi_{k-1}(\theta) + By_k + \Gamma u_k \ . \tag{2.5}$$

This is not meant to be a computational formula but a useful theoretical descriptor. Consider, for example, the dynamic adjustment (DA) model,

$$e_k(\theta) = \left(I+C_\alpha(L)\right)\left\{\left(I+A_\beta(L)\right)y_k - B_\beta(L)u_k\right\}$$

where, for example, $C_\alpha(L) = \sum_{i=1}^{n_c} C_i(\alpha)L^i$; while α denotes noise parameters, β system parameters and L is the lag operator. If H is defined as

$$H = \left(A_1 \ \dots \ A_{n_a} B_1 \ \dots \ B_{n_b} C_1 \ \dots \ C_{n_c}\right)$$

then $\theta = \text{vec } H$ is the vector obtained by stacking the columns of H one underneath another (for a full exposé of this 'vec' notation, see Neudecker, 1969).

Now we can write

$$e_k(\theta) = y_k - \hat{y}_{k/k-1}(\theta) = y_k - H\varphi_k(\theta) = y_k - \varphi_k'(\theta) \otimes I\theta \tag{2.6}$$

where

$$\varphi_k(\theta) = \left(-y'_{k-1} \;\ldots\; -y'_{k-n_a} \; u'_{k-1} \;\ldots\; u'_{k-n_b} \; \delta'_{k-1}(\theta) \;\ldots\; \delta'_{k-n_c}(\theta)\right)'$$

with

$$\Phi_k(\theta) = \varphi'_k(\theta) \otimes I \quad \text{and} \quad \delta_k(\theta) = \left(I + A_\alpha(L)\right)y_k - B_\beta(L)u_k \ .$$

Now it is straightforward to show $\varphi_k(\theta)$ obeys a recursion of the form of (2.5).

Notice that $A(\theta)$ is in general rather sparse so equation (2.5) is not used in computations; it is useful from a theoretical viewpoint, however. It should also be noted that equation (2.4) holds also for state-space (SS) models (see Kreisselmeier, 1976, p. 3).

With equation (2.4) in mind, a recursion for θ is obtained by analogy from the recursive least squares algorithm as

$$\hat{\theta}_k = \hat{\theta}_{k-1} + P_{k-1}\hat{\Phi}_k V_k^{-1}\hat{e}_k \ ,$$

$$P_k = P_{k-1} - P_{k-1}\hat{\Phi}_k V_k^{-1}\hat{\Phi}'_k P_{k-1} \ ,$$

with

$$V_k = \hat{\Phi}'_k P_{k-1}\hat{\Phi}_k + \Sigma_0 \ .$$

There are two possibilities for $\hat{\Phi}_k = \hat{\varphi}_k \otimes I$ and $\hat{e}_k$.

In practice they will be found recursively from an equivalent of

$$\hat{\varphi}_k = A\left(\hat{\theta}_{k-1}\right)\hat{\varphi}_{k-1} + By_k + \Gamma u_k \qquad (2.5a)$$

and

$$\hat{e}_k = y_k - \Phi'_k\hat{\theta}_{k-1} \ .$$

For the ensuing analysis, however, the author has found the following choices more useful: $\hat{\Phi}_k = \hat{\Phi}_k\left(\hat{\theta}_{k-1}\right)$ with $\Phi_k(\theta)$ as in (2.5) and

$$\hat{e}_k = e_k\left(\hat{\theta}_{k-1}\right) = y_k - \Phi'_k\left(\hat{\theta}_{k-1}\right)\hat{\theta}_{k-1} \ .$$

The recursions obtained with the latter choices will be called infinite past (IP) recursions since for each new θ , (2.5) must be solved for all k . See Appendix 2A for a simple example clarifying this definition. The author expects that the practical recursions and IP recursions will have the same asymptotic properties (that is, convergence behaviour and asymptotic covariance matrices).

(c) Dynamic Programming. The first problem could be solved by a dynamic programming argument. Thus, for DD models (discrete time evolution, discrete time measurement) we would minimise (2.1) subject to the conditions

$$e_k(\theta) = y_k - \hat{y}_{k/k-}(\theta) \ , \quad \theta_k = \theta_{k-1} \ .$$

Alternatively for CC models (continuous time evolution, continuous time measurements) and CD models we minimise (2.1) subject to the conditions

$$e_t(\theta) = y_t - \hat{y}_{t/t-}(\theta) \ , \quad d\theta/dt = 0 \ .$$

We can now recognise this dynamic programming problem as a non-linear filtering problem for a system with non linear observations

$$y_k = y_{k/k-}(\theta) + e_k(\theta) \quad \text{or} \quad y_t = \hat{y}_{t/t-}(\theta) + e_t(\theta)$$

and noiseless linear state equation

$$\theta_k = \theta_{k-1} \quad \text{or} \quad d\theta/dt = 0 \ .$$

We can now choose as our recursion anyone of the many approximate solutions to this problem (see Jazwinski, 1970). In the next section a particular choice has been made and the recursion will be called an IPPER.

2.3 Infinite Past Prediction Error Recursions

From now on the discussion will be mostly for IP recursions. However given the comments just prior to the last sentence of Section 2.2 (b) one approach to the construction of a practical form of the recursions should be clear.

The IPPER's which, it is stressed, are applicable to both linear and non linear models, are given first for CD and DD models, and then for CC models.

(a) CD AND DD MODELS

Parameter update.

$$\hat{\theta}_k = \hat{\theta}_{k-1} + P_{k-1}\tilde{\Phi}_k\left(\hat{\theta}_{k-1}\right)V_k^{-1}e_k\left(\hat{\theta}_{k-1}\right) \tag{2.6}$$

or

$$\hat{\theta}_k = \hat{\theta}_{k-1} + P_k\tilde{\Phi}_k(\hat{\theta}_{k-1})\Sigma_0^{-1}e_k(\hat{\theta}_{k-1}) \ . \tag{2.6a}$$

Gain update.

$$P_k = P_{k-1} - P_{k-1}\tilde{\Phi}_k(\hat{\theta}_{k-1})V_k^{-1}\tilde{\Phi}_k'(\hat{\theta}_{k-1})P_{k-1} \tag{2.7}$$

or

$$P_k^{-1} = P_{k-1}^{-1} + \tilde{\Phi}_k(\hat{\theta}_{k-1})\Sigma_0^{-1}\tilde{\Phi}_k'(\hat{\theta}_{k-1}) \tag{2.7a}$$

with

$$V_k = \tilde{\Phi}_k'(\hat{\theta}_{k-1})P_{k-1}\tilde{\Phi}_k(\hat{\theta}_{k-1}) + \Sigma_0 \tag{2.8}$$

where

$$\tilde{\Phi}_k(\theta) = \partial\hat{y}_{k/k-}/\partial\theta = -\partial e_k'(\theta)/\partial\theta \tag{2.9}$$

is an optimal instrumental variable or sensitivity matrix.

Of course, Σ_0 in (2.6a) and (2.7a) is unknown and it is logical to replace it by a recursion for the innovations covariance matrix such as

$$\hat{\Sigma}_k = \hat{\Sigma}_{k-1} + k^{-1}\left(e_k(\hat{\theta}_{k-1})e_k'(\hat{\theta}_{k-1}) - \hat{\Sigma}_{k-1}\right) \ . \tag{2.10}$$

In practice, equations (2.6), (2.7), (2.8) and (2.10) will be used. For analysis purposes, however, equations (2.6a) and (2.7a) are more useful. In relation to the above algorithm, note the following points.

(i) State Space (SS) models. With the usual case of the known observation matrix C, $\hat{y}_{k/k-}$ will be obtained from a Kalman Filter (Jazwinski, 1970) with $\hat{y}_{k/k-} = C\hat{x}_{k/k-}$ where $\hat{x}_{k/k-}$ is the output of the DD or CD Kalman Filter. It follows then that $\tilde{\Phi}_k(\theta) = S_k'(\theta)C'$ where $S_k(\theta) = \partial\hat{x}_{k/k-}(\theta)/\partial\theta'$ is a sensitivity matrix (for DD models of course $\hat{x}_{k/k-}(\theta) = \hat{x}_{k/k-1}(\theta)$).

The computation of $S_k(\theta)$ will require solving a sensitivity difference (DD) or differential (CD) equation as well as running a state estimator (Kalman Filter) or observer (Sidar, 1976; Gupta and Mehra, 1977).

Thus for a CD, SS model in innovations form, the sensitivity equations are:

For system parameters:

$$d/dt\,\partial\hat{x}_{t/k-}/\partial\beta' = F(\beta)\partial\hat{x}_{t/k-}/\partial\beta' + \hat{x}'_{t/k-} \otimes I\partial f(\beta)/\partial\beta' + u'_t \otimes I\partial b(\beta)/\partial\beta'$$

where

$$f(\beta) = \mathrm{vec}\big(F(\beta)\big) \;,\quad b(\beta) = \mathrm{vec}\big(B(\beta)\big) \;,\quad k\text{-}1 < t < k \;,$$

with initial conditions

$$\partial\hat{x}_{k-1/k-}/\partial\beta' = \partial\hat{x}_{k-1/k-1}/\partial\beta'$$

and

$$\partial\hat{x}_{k-1/k-1}/\partial\beta' = (I-KC)\partial\hat{x}_{k-1/k-1-}/\partial\beta' \;.$$

For noise parameters: denote $\alpha = \mathrm{vec}(K)$, K being the Kalman gain matrix which here consists of noise parameters

$$d/dt\;\partial\hat{x}_{t/k-}/\partial\alpha' = F(\beta)\partial\hat{x}_{t/k-}/\partial\alpha'$$

with initial conditions

$$\partial\hat{x}_{k-1/k-}/\partial\alpha' = \partial\hat{x}_{k-1/k-1}/\partial\alpha'$$

and

$$\partial\hat{x}_{k-1/k-1}/\partial\alpha' = (I-KC)\partial\hat{x}_{k-1/k-1}/\partial\alpha' + e'_k \otimes I \;.$$

In practice these equations would be solved by numerical methods.

Similar equations apply to DD models. With a mechanistic SS model where β is known, θ in (2.6) to (2.8) is replaced by α and only the second sensitivity equation above has to be solved. This in fact is just a recursion for the Kalman gain matrix. An alternative to the recursion for K would be a similar recursion for the K matrix (different notation) of Son and Anderson (1973) together with their "output statistics" Kalman Filter.

An alternative well known general method of generating recursions for SS models, is to append the unknown parameters to the state and use the Extended Kalman Filter (EKF). An interesting analysis of such EKF recursions has been given recently by Ljung (1977c) who shows that much of the trouble associated with the EKF as a recursive parameter estimator is due to not setting the model up in innovations form. Note also that the EKF as a recursive parameter estimator can be considered also as a recursive prediction error or stochastic model reference scheme (Young, 1974, p. 221).

However, the author feels that PER's are a more natural extension of prediction error ideas to the recursive sphere, since polynomial matrix description (PMD) {or differential operator form (Wolovich, 1974, p. 45)} models (with dimension reduction in the observation space) are easily handled.

(ii) PMD models. The sensitivity equation for DD, PMD models is usually relatively easily solved. Consider firstly an ARMAX model (with no dimension reduction) defined by

$$e_k(\theta) = \left(I+C_\alpha(L)\right)^{-1}\left\{\left(I+A_\beta(L)\right)y_k - B_\beta(L)u_k\right\} .$$

With H as before, it follows as in the example of Section 2.2 (b), that

$$e_k(\theta) = y_k - H\varphi_k(\theta)$$

where $\varphi_k(\theta)$ is as before but with $\delta_k(\theta)$ replaced by $e_k(\theta)$. Since PMD models are usually constructed from physically motivated sets of differential equations the elements of H will be heavily constrained (consisting say of many zero or known elements). This is important since we assume it resolves any identifiability problems.

Recalling the expression for the sensitivity matrix in (2.9), it follows that the sensitivity equation is simply

$$\left[I+C_\alpha(L)\right]\tilde{\Phi}_k'(\theta) = \varphi_k'(\theta) \otimes I . \qquad (2.11)$$

The constraints mentioned above ensure that many of the equations in (2.11) are null. For a second example take a TFARMA model (with dimension reduction)

$$e_k(\theta) = \left(I+D_\alpha(L)\right)^{-1}\left(I+C_\alpha(L)\right)\bar{\delta}_k(\beta)$$

where

$$\bar{\delta}_k(\beta) = y_k - Cx_k(\beta) \ , \quad x_k(\beta) = \left(I+A_\beta(L)\right)^{-1}B_\beta(L)u_k .$$

Now the sensitivity equation for the noise parameters

$$\alpha = \mathrm{vec}\left(D_1 \ \ldots \ D_{n_d} C_1 \ \ldots \ C_{n_c}\right)$$

is

$$\left(I+D_\alpha(L)\right)\partial e_k/\partial\alpha' = \left(e_{k-1}'(\theta) \ \ldots \ e_{k-n_d}'(\theta) - \bar{\delta}_{k-1}'(\beta) \ \ldots \ -\bar{\delta}_{k-n_c}'(\beta)\right) \otimes I .$$

The sensitivity equation for the system parameters

$$\beta = \text{vec}\left(A_1 \ \ldots \ A_{n_a} B_1 \ \ldots \ B_{n_b}\right)$$

is

$$\left(I+D_\alpha(L)\right)\partial e_k/\partial\beta' = \left(I+C_\alpha(L)\right)C\partial x_k/\partial\beta'$$

where $\partial x_k/\partial\beta'$ obeys

$$\left(I+A_\alpha(L)\right)\partial x_k/\partial\beta' = \left(-x'_{k-1}(\beta) \ \ldots \ -x'_{k-n_a}(\beta)u'_{k-1} \ \ldots \ u'_{k-n_b}\right) \otimes I \ .$$

The above sensitivity equations contain, of course, null equations corresponding to known or zero elements in $A_i(\beta)$, $B_i(\beta)$.

(b) CC MODELS

Here the algorithm takes the following form:

Parameter evolution.

$$d\hat{\theta}_t/dt = P_t\tilde{\Phi}_t\left(\hat{\theta}_t\right)\Sigma_0^{-1}e_t\left(\hat{\theta}_t\right) \ . \tag{2.12}$$

Gain evolution.

$$dP_t/dt = -P_t\tilde{\Phi}_t\left(\hat{\theta}_t\right)\Sigma_0^{-1}\Phi'_t\left(\hat{\theta}_t\right)P_t \tag{2.12a}$$

or

$$dP_t^{-1}/dt = \tilde{\Phi}_t\left(\hat{\theta}_t\right)\Sigma_0^{-1}\tilde{\Phi}'_t\left(\hat{\theta}_t\right) \ . \tag{2.12b}$$

As before Σ_0 will be replaced by Σ_t with

$$d\hat{\Sigma}_t/dt = t^{-1}\left(e_t\left(\hat{\theta}_t\right)e'_t\left(\hat{\theta}_t\right)-\hat{\Sigma}_t\right) \tag{2.13}$$

with $\tilde{\Phi}_t(\theta) = -de'_t/d\theta$, as before. Again, a difficult sensitivity differential equation must, in general, be solved. Recursions of this general form have been considered by Tsypkin (1971) and Young (1976a). Recursions for CC models are, however, more usually restricted to deterministic MR schemes where the gain term in equation (2.12), which vanishes as $t \to \infty$, is replaced by a constant gain (Young, 1964; Lion, 1967; Anderson, 1977).

These latter recursions will exhibit exponential convergence in the "no noise" case (and will, therefore, be referred to as exponentially converging recursions, ECR's). In the noisy case it is well-known that they will, in general, be biased. What is not mentioned so often, however, is that even when unbiased, they *never* converge to a constant. Instead, they converge to a *random variable* whose variance depends on the noise variance, the gain and the energy of the input signal.

(c) STOCHASTIC APPROXIMATION

The SA recursions will be the same as equations (2.6), (2.8) and (2.10) above with the P_k given by $a_k I$; where a_k are a sequence of positive constants with

$$\sum_1^\infty a_k = \infty \ , \quad \sum_1^\infty a_k^2 < \infty \ .$$

As mentioned earlier for DD models, such recursions were implicit in Tsypkin (1971) though he never actually gave them since he never replaced his $Q(X, \theta)$ by a formula for the innovations sequence; *cf.* Young (1976a, p. 533).

In relation to the above recursions it is necessary to make the following remarks.

REMARK 1. For single output systems the innovations variance is a scale factor and need not appear in the recursions.

REMARK 2. The CD recursion for a non-linear SS model with no process noise should be compared with that of Sidar (1976, equation (27)), where the place of P_k in the parameter update equation is taken by the incremental matrix

$$\left(P_k^{-1} - P_{k-1}^{-1}\right)^{-1} = \left[\tilde{\Phi}_k(\hat{\theta}_{k-1})\Sigma_0^{-1}\tilde{\Phi}_k'(\hat{\theta}_{k-1})\right]^{-1} .$$

It follows that Sidar's recursion will be an ECR and so converge in the noisy case to a limit random variable (compare with discussion above). In fact Sidar gave no analysis of the noisy case. The recursion given here would seem to be preferable (see Remark 4 below).

REMARK 3. Young (1976b) showed how to construct recursions for a single input, single output transfer function (TF) model by a different solution to the sequential minimisation problem. More recently (Young and Jakeman, 1978; Jakeman and Young, 1978), PER-like recursions have been developed for multivariable TF models and evaluated for finite samples by Monte Carlo simulations.

REMARK 4. The same argument as given in (c) of Section 2.2 yields recursions for time varying parameters by analogy with the Kalman Filter. Indeed suppose the θ variation is modeled as

$$\theta_k = A\theta_{k-1} + n_k$$

where n_k is a white noise covariance matrix Q ; then for DD models, for example, the recursions will be

$$\hat{\theta}_{k/k-1} = A\hat{\theta}_{k-1/k-1} ,$$

$$\hat{\theta}_{k/k} = \hat{\theta}_{k/k-1} + P_{k/k-1}\tilde{\Phi}_k(\hat{\theta}_{k-1})V_k^{-1}e_k(\hat{\theta}_{k-1})$$

where

$$P_{k/k} = P_{k/k-1} - P_{k/k-1}\tilde{\Phi}_k(\hat{\theta}_{k-1})V_k^{-1}\Phi_k'(\theta_{k-1})P_{k/k-1}$$

and

$$P_{k+1/k} = AP_{k/k}A' + Q$$

while

$$V_k = \tilde{\Phi}_k'(\hat{\theta}_{k-1})P_{k/k-1}\tilde{\Phi}_k(\hat{\theta}_{k-1}) + \hat{\Sigma}_k$$

with $\hat{\Sigma}_k$ given by (2.10). This kind of extension is discussed by Young, Jakeman and Whitehead (1978) (for the single and multivariable TF model) and by Kaldor (1978).

If we consider this algorithm (with $A = I$) for the simple regression model

$$e_k(\beta) = y_k - u_k'\beta$$

and compare it to an equivalent ECR, we may conclude the following: the ECR stands in a similar relation to these *time-varying parameter* recursions as SA does to PER's for *time-invariant parameter* recursions.

REMARK 5. Implicit in the above PER's is a self-tuning filter (= one step ahead predictor) namely $\hat{y}_{k/k-}(\hat{\theta}_{k-1})$ or a state estimator (or observer) namely $\hat{x}_{k/k-}(\hat{\theta}_{k-1})$. If we can show $\hat{\theta}_k - \theta_0 \to 0$ then it is easy to demonstrate that $\hat{y}_{k/k-}(\hat{\theta}_{k-1})$ converges to $\hat{y}_{k/k-}(\theta_0)$ and $\hat{x}_{k/k-}(\hat{\theta}_{k-1})$ to $\hat{x}_{k/k-}(\theta_0)$ (see Appendix 2B).

REMARK 6. In relation to CD models where the recursion is for the continuous time system parameters, it is common that the input is also sampled so that the sensitivity equation can only be solved by numerical integration. Consider a TF model example

$$e_t(\beta) = y_t - x_t = y_t - \{B(D)/(D^n+A(D))\}u_t$$

where D is the derivative operator d/dt : while

$$B(D) = \sum_0^{n-1} b_s D^s$$

and so on. Then the sensitivity vector for b is

$$\partial e_t/\partial \mathrm{b} = -(D^n+A(D))^{-1}(1 \ldots D^{n-1})u_t = -\mathrm{q}_u(t)$$

while that for a is $\mathrm{q}_x(t)$ where, for example,

$$\dot{\mathrm{q}}_u(t) = \mathrm{A}\mathrm{q}_u(t) + \mathrm{c}u_t$$

with c' the $1 \times m$ vector $(0\ 0 \ldots 1)$ and

$$\mathrm{A} = \begin{bmatrix} 0 & 1 & \ldots & 0 \\ & & \ddots & \\ & & & 1 \\ -a_0 & & \ldots & -a_{n-1} \end{bmatrix}$$

is in companion form: similarly the sensitivity vector for a is $\mathrm{q}_x(t)$. Thus we see how derivatives are computed by an optimal state variable filter (*cf.* Young, 1964; Rucker, 1963, and Kohr, 1963) determined by the system. Young, Jakeman and Whitehead (1978) use this approach for the single and multivariable TF models. The important point here is not to discretise the system equations when faced with sampled input data, as is sometimes done (Phillips, 1974), rather to discretise the sensitivity equations.

2.4 Relation with Known Recursions

The computational form of the PER has appeared before in the following cases.

(i) Scalar ARMAX, PER = RML_2 (Soderstrom *et al.*, 1974).

(ii) Scalar Dynamic Adjustment (DA) model PER equals the Recursion of Gertler and Banyasz (1974) and also that of Hannan and Tanaka (1976).

(iii) PER for "Box-Jenkins" or TFARMA model

$$e_k(\theta) = \left(1+C_\alpha(L)\right)/\left(1+D_\alpha(L)\right)\left[y_k-\left(1+A_\beta(L)\right)^{-1}B_\beta(L)u_k\right]$$

is the symmetric refined IV of Young and Jakeman (1978). This equivalence is not immediately clear and further comments are offered in Appendix 2C.

2.5 Preliminary Analysis

For the next three sections, we suppose that S belongs to the model set M. Guided by the discussion of Section 2.1, consider a general recursion for CD, DD models.

$$\hat{\theta}_k = \hat{\theta}_{k-1} + P_k\Psi_k\left(\hat{\theta}_{k-1}\right)\Sigma_0^{-1}e_k\left(\hat{\theta}_{k-1}\right) , \tag{2.14}$$

$$P_k^{-1} = P_{k-1}^{-1} + \Psi_k\left(\hat{\theta}_{k-1}\right)\Sigma_0^{-1}\zeta_k'\left(\hat{\theta}_{k-1}\right) ,$$

where $\zeta_k(\theta)$ and $\Psi(\theta)$ will be matrices of instrumental variables obtained by filtering operations of the form of equation (2.5) or (2.11). For PER's $\Psi_k(\theta) = \zeta_k(\theta) = \tilde{\Phi}_k(\theta)$; for SA $P_k = a_k I$; and for method (b) of Section 2.2, $\Psi_k(\theta) = \zeta_k(\theta) = \Phi_k(\theta) = \varphi_k(\theta) \otimes I$. Note that Σ_0 has been used here rather than Σ_k (equation (2.10) of Section 2.2) in order to simplify the analysis. It is not felt that the effects of this change will be important.

Now substitute the following Taylor series*

$$e_k\left(\hat{\theta}_{k-1}\right) = e_k\left(\theta_0\right) - \tilde{\Phi}_k'\left(\theta_{k-1}^{\bullet}\right)\left(\hat{\theta}_{k-1}-\theta_0\right) \tag{2.15}$$

– for some $\theta_{k-1}^{\bullet}$ with $\|\theta_{k-1}^{\bullet}-\theta_0\| \leq \|\hat{\theta}_{k-1}-\theta_0\|$ – into the parameter update equation to give

$$\hat{\theta}_k = \hat{\theta}_{k-1} - P_k\Psi_k\left(\hat{\theta}_{k-1}\right)\Sigma_0^{-1}\tilde{\Phi}_{k-1}'\left(\hat{\theta}_{k-1}-\theta_0\right) + P_k\Psi_k\left(\hat{\theta}_{k-1}\right)\Sigma_0^{-1}e_k\left(\theta_0\right) . \tag{2.16}$$

This expression is basic to all that follows. It exhibits the recursion as a stochastic difference equation driven by a white noise. In Section 2.6 this stochastic difference equation will be analysed by converting it to a stochastic differential equation. Before proceeding further, some preliminaries are necessary.

---

* The reason for using IP recursions is to allow use of this Taylor Series expansion.

In what follows we will have occasion to deal with covariances among $\Psi_k(\theta)$, $\zeta_k(\theta)$, $\Phi_k(\theta)$ and so on. Thus consider, for θ, $\theta^* \in D_c$ (a convex stability region enclosing the true value) the temporal average

$$H_N(\theta, \theta^*) = N^{-1} \sum_1^N \Psi_k(\theta)\Sigma_0^{-1}\tilde{\Phi}_k'(\theta^*) .$$

This will converge to

$$H(\theta, \theta^*) = E\left(\hat{\Psi}_k(\theta)\hat{\tilde{\Phi}}_k'(\theta^*)\right) = \int_{-\pi}^{\pi} dF_{\hat{\Psi}\hat{\tilde{\Phi}}}(\omega|\theta, \theta^*)$$

where

$$\hat{\Psi}_k(\theta) = \Sigma_0^{-\frac{1}{2}}\Psi_k(\theta) ,$$

and so on, and $dF_{\hat{\Psi}\hat{\tilde{\Phi}}}(\omega|\theta, \theta^*)$ is the cross-sepctrum of $\hat{\Psi}_k(\theta)$ and $\hat{\tilde{\Phi}}_k(\theta^*)$. We will have occasion to compute an expression such as $H(\theta_k, \theta_k^*)$ and will denote it by $\tilde{E}\left(\hat{\Psi}(\theta_k)\hat{\tilde{\Phi}}'(\theta_k^*)\right)$.

This "twiddle" notation was used by Hannan (1976). To agree also with the notation of Ljung define

$$H(\theta) = E\left(\hat{\Psi}_k(\theta)\hat{\tilde{\Phi}}_k'(\theta)\right) = \int_{-\pi}^{\pi} dF_{\hat{\Psi}\hat{\tilde{\Phi}}}(\omega|\theta) , \tag{2.17}$$

$$G(\theta) = E\left(\hat{\Psi}_k(\theta)\hat{\zeta}_k'(\theta)\right) = \int_{-\pi}^{\pi} dF_{\hat{\Psi}\hat{\zeta}}(\omega|\theta) , \tag{2.18}$$

$$F(\theta) = E\left(\hat{\Psi}_k(\theta)\hat{\Psi}_k'(\theta)\right) = \int_{-\pi}^{\pi} dF_{\hat{\Psi}\hat{\Psi}}(\omega|\theta) .$$

REMARK. For PER's, $\Psi_k(\theta) = \zeta_k(\theta) = \tilde{\Phi}_k(\theta)$; so $H(\theta) = G(\theta) = F(\theta)$. For the SA schemes implicit in Tsypkin (1971), $\Psi_k(\theta) = \tilde{\Phi}_k(\theta)$ so $H(\theta) = F(\theta)$.

2.6 Convergence

Here, as in Section 2.5, it is assumed the system belongs to the model set. General recursions are discussed first and then SA. In this section we assume the recursion converges and use this assumption to generate *necessary* conditions for convergence, that is, conditions that must be fulfilled in order that the assumption of convergence not be contradicted. The idea then, is to convert the basic stochastic difference equation form of the recursion (equation (2.16) of Section 2.5), to a stochastic differential equation. This is done by suggesting the driving white noise term may be regarded as the increment of a certain Wiener process. From an intuitive point of view a Wiener process is the integral of white noise; see Arnold (1974) or McGarty (1974).

Despite the above conversion the discussion is repeated using discrete time arguments. However using a s.d.e. it becomes almost trivial to compute the asymptotic covariances of the recursive parameter estimates (Section 2.7). Finally, note that for the following discussion it is assumed the recursive estimator is made to lie in the convex stability region D_c containing the true value. The convexity of D_c ensures that intermediate values in a Taylor series expansion lie inside D_c .

2.6.1 GENERAL RECURSIONS

Denote $\hat{R}_k = P_k^{-1}/k$ and consider that the right most term in equation (2.16) can be written as

$$k^{-\frac{1}{2}}\hat{R}_k^{-1}F_0^{\frac{1}{2}}dW_k$$

where $dW_k = W_k - W_{k-1}$ and $F_0 = F(\theta_0)$, θ_0 being the true value; also

$$W_k = F_0^{-\frac{1}{2}} \sum_1^k \Psi_s(\hat{\theta}_{s-1})\Sigma_0^{-1}e_s(\theta_0)/\sqrt{s} \ . \tag{2.19}$$

Suppose for the moment $\Psi_k(\theta)$ is independent of θ , then

$$E(W_kW_k') \simeq I \sum_1^k s^{-1} \ .$$

Thus W_k resembles a multi-dimensional Wiener process on a time scale $\tau \simeq \ln k$. Note that $d\tau = k^{-1}$ shrinks to zero as $k \to \infty$, so τ will measure a continuous time. When $\Psi_k(\theta)$ depends on θ there will be an extra term in the above expectation, namely,

$$F_0^{-\frac{1}{2}}\left\{\sum_1^N \left[E\left(\Psi_k(\hat{\theta}_{k-1})\Sigma_0^{-1}\Psi_k'(\hat{\theta}_{k-1})\right)-F_0\right]/k\right\}F_0^{-\frac{1}{2}}$$

and provided the first term here approaches its limit value rapidly enough, this expression converges (recall $\sum_1^\infty k^{-1-\delta} < \infty$ for $\delta > 0$) and may be neglected relative to the first term above. If we make the notational transformation $\hat{\theta}_k \leftrightarrow \theta(\tau)$, $\hat{R}_k \leftrightarrow R(\tau)$ and so on, then we obtain directly from equation (2.16),

$$d\theta(\tau) = -\hat{R}^{-1}(\tau)\Psi\big(\theta(\tau),\ \tau\big)\Sigma_0^{-1}\tilde{\Phi}'\big(\theta(\tau)\big)(\theta-\theta_0)d\tau + e^{-\frac{1}{2}\tau}\hat{R}^{-1}(\tau)F_0^{-\frac{1}{2}}dW(\tau) \quad (2.20)$$

while from equation (2.15),

$$d\hat{R}(\tau)/d\tau = \Psi\big(\theta(\tau),\ \tau\big)\Sigma_0^{-1}\zeta'\big(\theta(\tau),\ \tau\big) - \hat{R}(\tau) \ . \quad (2.21)$$

The above differential equation is now analysed in two ways. Firstly, by investigating the behaviour of its first moment; this yields an analysis inspired by that of Ljung (1977a, 1977b). Secondly, by considering the s.d.e. as it is using a stochastic Lyapunov function; this yeidls an analysis related to that of Landau (1976, 1978). The conclusions in each case are the same. The aim of the second analysis is merely to illustrate the relation of the two analyses. In Section 2.6.4 the first moment analysis is repeated in discrete time; the analysis is then related to that given rigorously by Hannan (1976). Section 2.6.5 contains an analysis of SA.

2.6.2 FIRST MOMENT ANALYSIS

To first order, the mean function $m(\tau) = E\big(\theta(\tau)\big)$ of the differential equation (2.20) obeys

$$dm(\tau)/d\tau = -R^{-1}(\tau)\tilde{E}\big(\hat{\Psi}\big(m(\tau)\big)\hat{\Phi}'\big(m^{\bullet}(\tau)\big)\big)\big(m(\tau)-m_0\big) \ , \quad (2.20a)$$

$$dR(\tau)/d\tau = G\big(m(\tau)\big) - R(\tau) \ . \quad (2.21a)$$

Recall $\tilde{E}$ and $G(\theta)$ were defined at the end of Section 2.5; also $m^{\bullet}(\tau)$ corresponds to $\theta_k^{\bullet}$. The convergence of $\theta(\tau)$ can be investigated by looking at the stability of these equations. This is the type of conclusion reached by Ljung (1977b) by a rigorous argument. We can obtain an initial understanding of convergence from the Taylor series (2.15). Multiply it on the left by $\hat{\Psi}(\mathrm{m}(\tau))$ and take expectations to see

$$\tilde{E}\left(\hat{\Psi}(\mathrm{m}(\tau))\hat{\Phi}'(\mathrm{m}(\tau))\right)(\mathrm{m}(\tau)-\mathrm{m}_0) = \tilde{E}\left(\hat{\Psi}(\mathrm{m}(\tau))\hat{\mathrm{e}}(\mathrm{m}(\tau))\right) .$$

It is then clear that convergence, that is, $d\mathrm{m}/d\tau \to 0$, will ensure

$$\tilde{E}\left(\hat{\Psi}(\mathrm{m}(\tau))\hat{\mathrm{e}}(\mathrm{m}(\tau))\right) \to 0 .$$

That is the recursion can only converge by forcing a decorrelation between the model innovations $\mathrm{e}(\mathrm{m}(\tau))$ and the proxy for the gradient $\Psi(\mathrm{m}(\tau))$.

Returning to equations (2.20a), (2.21a) we search for a Lyapunov function as follows.

Consider the two cases

$$\Psi_k(\theta) = \Phi_k(\theta) \quad , \quad \Psi_k(\theta) = \tilde{\Phi}_k(\theta) .$$

(i) In the first case, take, for the moment,

$$\zeta_k(\theta) = \Psi_k(\theta) .$$

Then as suggested in Section 2.2 (b) the recursion has been constructed by supposing $\Phi_k(\theta_0)$ known and minimising

$$W_N(\theta) = N^{-1}\sum_1^N \delta_k'(\theta)\Sigma_0^{-1}\delta_k(\theta)$$

with

$$\delta_k(\theta) = y_k - \Phi_k'(\theta_0)(\theta-\theta_0) .$$

That is $W_N(\theta)$ should be suggestive of a good choice for a Lyapunov function. However recalling equations (2.2), (2.4) we have

$$W_N(\theta) = N^{-1}\sum_1^N \mathrm{e}_k'(\theta_0)\Sigma_0^{-1}\mathrm{e}_k(\theta_0) + 2N^{-1}\sum_1^N \mathrm{e}_k'(\theta_0)\Sigma_0^{-1}\Phi_k'(\theta_0)(\theta-\theta_0)$$
$$+ (\theta-\theta_0)N^{-1}\sum_1^N \Phi_k(\theta_0)\Phi_k'(\theta_0)(\theta-\theta_0) .$$

The middle term approaches zero as $N \to \infty$ leaving the choice of Lyapunov function as

$$V_\tau = V(\mathrm{m}(\tau), \tau) = (\mathrm{m}(\tau)-\mathrm{m}_0)'\mathrm{R}(\tau)(\mathrm{m}(\tau)-\mathrm{m}_0) .$$

This was the function chosen by Ljung (1977a). It will be used in the general case being considered here where $\zeta_k(\theta)$, $\Psi_k(\theta)$, $\Phi_k(\theta)$ may all differ. It is immediately clear that we must require that $G(\theta)$ be positive definite (p.d.) $\forall\theta \in D_c$. Thus we have

$$dV_\tau/d\tau = -V_\tau - \big(m(\tau)-m_0\big)'\big(H_\tau+H_\tau^\bullet-G\big(m(\tau)\big)\big)\big(m(\tau)-m_0\big)$$

with $H_\tau^\bullet = \tilde{E}\big(\hat{\Psi}\big(m(\tau)\big)\hat{\hat{\Phi}}'\big(m^*(\tau)\big)\big)$.

To conclude $V_\tau \to 0$ we require

$$H^\bullet + H^{\bullet\prime} - G(m)$$

be p.d. This can only hold for m, $m^\bullet$ in some small region D_δ (inside D_c) enclosing the true value provided

$$H(\theta_0) + H'(\theta_0) - G(\theta_0) \tag{2.22}$$

is p.d. Thus we conclude local convergence when (2.22) holds.

In summary then, the conditions for local convergence are the positive definiteness uniformly in $\theta \in D_c$ of

$$G(\theta) = \int_{-\pi}^{\pi} dF_{\hat{\Psi}\hat{\zeta}}(\omega|\theta)$$

and that

$$H(\theta_0) + H'(\theta_0) - G(\theta_0)$$

is p.d. Now since $\zeta_k(\theta)$, $\Psi_k(\theta)$ and so on depend on the input sequence, condition (2.22) in general depends on the input sequence for its validity. Ljung (1977b) has pointed out that (2.22) holds for all input sequences if

$$T_{\hat{\Psi}\hat{\hat{\Phi}}}(\omega|\theta_0) - \tfrac{1}{2}T_{\hat{\Psi}\hat{\zeta}}(\omega|\theta_0) \tag{2.22a}$$

is positive real (p.r.) where $T_{\hat{\Psi}\hat{\hat{\Phi}}}(\omega|\theta)$ is the transfer function from $\hat{\Psi}_k(\theta)$ to $\hat{\hat{\Phi}}_k(\theta)$ and so on.

The reader may wonder at the conclusion of local convergence when the discussion in Ljung (1977b) concludes global convergence. Such a result holds only for linear models and is obtainable here by replacing the Taylor series (2.15) of Section 2.5 with a similar expression that relies on linearity. This point is expanded upon with a simple example in Appendix 2D.

(ii) When $\Psi_k(\theta) = \tilde{\Phi}_k(\theta)$ the quantity being minimised is

$$\tfrac{1}{2}N^{-1} \sum_{1}^{N} e_k'(\theta)\Sigma_0^{-1}e_k(\theta)$$

which approaches

$$\tfrac{1}{2}E\left[e_k'(\theta)\Sigma_0^{-1}e_k(\theta)\right] .$$

This Lyapunov function was suggested to the author by Ljung (1978b). It follows with this choice that

$$\begin{aligned} dV_\tau/d\tau &= dV/dm' \; dm/d\tau \\ &= \tilde{E}\left(\hat{e}'\big(m(\tau)\big)\hat{\hat{\Phi}}\big(m(\tau)\big)\right)R^{-1}(\tau)H^{*\prime}\big(m(\tau)-m_0\big) \end{aligned}$$

where here,

$$H_\tau^* = \tilde{E}\left(\hat{\hat{\Phi}}\big(m(\tau)\big)\hat{\hat{\Phi}}'\big(m^\bullet(\tau)\big)\right) .$$

Applying the Taylor series once more yields

$$dV_\tau/d\tau = -\big(m(\tau)-m_0\big)'H_\tau^{*\prime}R^{-1}(\tau)H_\tau^*\big(m(\tau)-m_0\big) .$$

But this implies the term on the right hand side approaches zero. Then $R(\tau)$ p.d. ensures the conclusion $m(\tau) - m_0 \to 0$.

2.6.3 STOCHASTIC LYAPUNOV FUNCTION METHOD

In this section the analysis is given only for the case

$$\Psi_k(\theta) \neq \tilde{\Phi}_k(\theta) .$$

If (2.20) is interpreted as an Itô equation (see Arnold, 1974 or McGarty, 1974) then we may apply Itô's lemma - the stochastic equivalent of integration - to

$$V_\tau = V\big(\theta(\tau),\, \tau\big) = \big(\theta(\tau)-\theta_0\big)'R(\tau)\big(\theta(\tau)-\theta_0\big)$$

so that

$$E\big(V_t\big) - E\big(V_s\big) = \int_s^t E\big\{\mathcal{D}\big(V\big(\theta(\sigma),\, \sigma\big)\big)\big\}d\sigma$$

where the operator $\mathcal{D}$ plays the role of derivative. We find (Arnold, 1974, p. 91),

$$\mathcal{D}V_\sigma = f\big(\theta(\sigma), \sigma\big)\partial V_\sigma/\partial\theta + \tfrac{1}{2} \text{ tr } \Gamma(\sigma)\Gamma(\sigma)\partial^2 V_\sigma/\partial\theta\partial\theta' + \partial V_\sigma/\partial t$$

where $f\big(\theta(\tau), \tau\big)$ denotes the coefficient of $d\tau$ in equation (2.20), while $\Gamma(\tau)$ is the coefficient of $dW(\tau)$. In the deterministic case of course $dW(\tau) = 0$ and $\Gamma(\tau) = 0$.

Now

$$\partial V_\sigma/\partial\theta = 2R(\sigma)\big(\theta(\sigma)-\theta_0\big) ,$$

$$\partial^2 V_\sigma/\partial\theta\partial\theta' = 2R(\sigma) ,$$

$$\partial V_\sigma/\partial\sigma = \big(\theta(\sigma)-\theta_0\big)' dR(\sigma)/d\sigma\big(\theta(\sigma)-\theta_0\big) .$$

It then follows after substitution in the expression for $\mathcal{D}V_\sigma$ that

$$\mathcal{D}V_\sigma = -V_\sigma - \big(\theta(\sigma)-\theta_0\big)'\big(2\hat{\Phi}_\sigma\hat{\Psi}_\sigma - \hat{\zeta}_\sigma\hat{\Psi}'_\sigma\big)\big(\theta(\sigma)-\theta_0\big) + \tfrac{1}{2} \text{ tr}\Big[F_0 R^{-1}(\sigma)\Big]e^{-\sigma} .$$

Now, for reasons to be clear in a moment, define an 'input' and an 'output', respectively;

$$\hat{u}_k(\theta) = \hat{\Psi}'_k(\theta)\big(\theta-\theta_0\big)$$

and

$$\hat{y}_k(\theta) = \big(\hat{\Phi}_k(\theta)-\tfrac{1}{2}\hat{\zeta}_k(\theta)\big)'\big(\theta-\theta_0\big) .$$

The transfer function between them is clearly

$$T_{\hat{\Psi}\hat{\Phi}}(\omega|\theta) - \tfrac{1}{2}T_{\hat{\Psi}\hat{\zeta}}(\omega|\theta) . \tag{2.22b}$$

It follows now by Ito's lemma that

$$E\big(V_t - V_s\big) + S_t - S_s = -\int_s^t E\big(V_\sigma\big)d\sigma - \int_s^t \hat{u}'_\sigma\hat{y}_\sigma d\sigma$$

where

$$S_t \simeq \tfrac{1}{2} \text{ tr}\Big[F_0 R^{-1}(\infty)\Big]\Big(1 - \int_0^t e^{-\sigma}d\sigma\Big) .$$

Set

$$dN(\tau)/d\tau = \hat{u}'_\tau\hat{y}_\tau - N(\tau)$$

so

$$E\left(V_t+N_t+S_t\right) - E\left(V_s+N_s+S_s\right) = -\int_s^t E\left(V_\sigma+N_\sigma\right)d\sigma \ .$$

Now if the transfer function (2.22b) is p.r. uniformly in θ the cross correlation between output and input will be positive. It is reasonable to suppose then that $N(\tau)$ will be positive for τ large enough. It will then follow that $E\left(V_\tau\right) \to 0$ as required.

REMARK. This argument is related to the discussion in Landau (1976) (see also Landau, 1978). It is now clear how the p.r. condition arises in about the same way in both the discussion of this section and that of Section 2.6.2 (based on Ljung, 1977a). This partly clarifies a comment of Ljung (1977a, p. 539) that his discussion and Landau's are very different.

2.6.4 DISCRETE TIME DISCUSSION

Return now to the basic recursion (2.14). This recursion is, after all, a stochastic difference equation and we may analyse its behaviour by looking at its first moment. Set $m_k = E\left(\hat{\theta}_k\right)$ and $R_k^{-1} = kP_k$ so, to first order

$$m_k = m_{k-1} + k^{-1}R_k^{-1}\tilde{E}\left(\hat{\Psi}\left(m_{k-1}\right)\hat{e}\left(m_{k-1}\right)\right)$$

with

$$R_k = R_{k-1}\left(1-k^{-1}\right) + k^{-1}G\left(m_{k-1}\right) \ .$$

Summing up the recursion gives

$$m_k = \sum_1^k l^{-1}R_l^{-1}\tilde{E}\left(\hat{\Psi}\left(m_{l-1}\right)\hat{e}\left(m_{l-1}\right)\right)$$

whereupon we will require error decorrelation as before to ensure m_k is finite $\left(\text{since } \sum_1^\infty l^{-1} = \infty\right)$. Next apply the Taylor Series (2.15) to obtain

$$m_k = m_{k-1} - k^{-1}R_k^{-1}\tilde{E}\left(\hat{\Psi}\left(m_{k-1}\right)\hat{\Phi}\left(m_{k-1}^{\bullet}\right)\right)\left(m_{k-1}-m_0\right) \ .$$

The stability (that is convergence) of this difference equation can be analysed by means of the discrete time Lyapunov function

$$V_k = (m_k - m_0)' R_k (m_k - m_0)/k \ .$$

It follows after some simple algebra that

$$V_k < V_{k-1} - V_{k-1}/k - (m_{k-1} - m_0)' T_{k-1} (m_{k-1} - m_0)/k$$

where

$$T_k = \dot{H}_k + \dot{H}_k' - G(m_k) \ .$$

Provided, as before, we ensure (2.22) is p.d. we can conclude

$$V_k < V_{k-1} - V_{k-1}/k \ .$$

Clearly V_k is a bounded decreasing sequence and so has a limit. Also, since

$$\sum_1^\infty V_{k-1}/k < V_0 < \infty$$

this limit must be zero to avoid contradicting $\sum_1^\infty k^{-1} = \infty$.

We conclude convergence and the discussion can continue much as in Section 2.6.2.

The analysis just given paraphrases that of Hannan (1976): of course many difficult technical points have gone unmentioned here. In any case it should be clear that there is no real difference between the continuous time and discrete time analyses (at this heuristic level). Further discussion of the relation between these analyses is contained in Chapter 7.

2.6.5 STOCHASTIC APPROXIMATION

As in Section 2.6.1 we begin by introducing a "Wiener process" similar to the above but on a different time scale. Set

$$W_k = F_0^{-\frac{1}{2}} \sum_1^k \Psi_s(\hat{\theta}_{s-1}) \Sigma_0^{-1} e_s(\theta_0) \sqrt{a_s} \ .$$

As before

$$E(W_k W_k') \simeq I\tau \ , \quad \tau = \sum_1^k a_s \ .$$

Since τ represents a time scale, this gives a natural interpretation of the well-known requirement $\sum_{1}^{k} a_s = \infty$. Next, assume a_k is a strictly decreasing sequence so that we can invert the above relation to $a_k = g(\tau)$ for some monotonic $g(\tau)$; for example, with $a_k = k^{-1}$ we have $g(\tau) = e^{-\tau}$. Then, as above, we obtain a stochastic differential equation,

$$d\theta(\tau) = -\hat{\Psi}_\tau\big(\theta(\tau)\big)\hat{\Phi}_\tau\big(\theta^{\bullet}(\tau)\big)\big(\theta(\tau)-\theta_0\big)d\tau + g(\tau)F_0^{\frac{1}{2}}dW(\tau) \ .$$

As before, to first order the first moment obeys the equation

$$dm(\tau) = -\tilde{E}\big(\hat{\Psi}\big(m(\tau)\big)\hat{\Phi}\big(m^{\bullet}(\tau)\big)\big)\big(m(\tau)-m_0\big)d\tau \ . \qquad (2.20b)$$

Furthermore, to a similar approximation the second moment

$$\Sigma(\tau) = E\big(\theta(\tau)-\theta_0\big)\big(\theta(\tau)-\theta_0\big)'$$

will obey

$$d\Sigma/d\tau = -H\big(m(\tau)\big)\Sigma(\tau) - \Sigma(\tau)H'\big(m(\tau)\big) + g^2(\tau)F_0 \ . \qquad (2.23)$$

Clearly equation (2.20b) has a decaying solution if $H(\theta) + H(\theta)'$ is p.d. $\forall\theta \in D_c$. Given this condition, equation (2.23) has a decaying solution if

$$\int_0^\infty g^2(\tau)d\tau < \infty \ ,$$

that is if $\sum_{1}^{\infty} a_k^2 < \infty$, the well known variance nulling condition (compare with Young, 1976, p. 532).

It is worth noting here that the discrete time analysis of Section 2.6.4 is easily given for SA. Indeed set

$$V_k = \big(m_k-m_0\big)'R_k\big(m_k-m_0\big)a_k \ .$$

Then it follows much as in Section 2.6.4 that

$$V_k < V_{k-1} - a_kV_{k-1} - \big(m_{k-1}-m_0\big)'\overline{T}_{k-1}\big(m_{k-1}-m_0\big)a_k$$

where $\overline{T}_k = H_k^{\bullet} + H_k^{\bullet}{}'$. Then, provided the condition below equation (2.23) holds we conclude

$$V_k < V_{k-1} - a_kV_{k-1}$$

whereupon the requirement $\sum_1^\infty a_k = \infty$ is apparent to ensure $V_k \to 0$. A second moment analysis will yield the requirement $\sum_1^\infty a_k^2 < \infty$.

It is logical at this point to enquire as to the relation of all this with the conditions of Tsypkin (1971, p. 56). These apply to the case

$$\Psi_k(\theta) = \tilde{\Phi}_k(\theta) = -\partial e_k'/\partial\theta \; .$$

Tsypkin's condition (b) is

$$(\theta-\theta_0)'E\left(\Phi_k'(\theta)\Sigma_0^{-1}e_k(\theta)\right) > 0 \; .$$

Apply a Taylor Series to $e_k(\theta)$ and notice that

$$E\left(\tilde{\Phi}_k'(\theta)\Sigma_0^{-1}e_k(\theta_0)\right) = 0$$

(since an innovations process is orthogonal to functions of the past) to see that this condition is

$$(\theta-\theta_0)'E\left(\Phi_k(\theta)\Sigma_0^{-1}\tilde{\Phi}_k'(\theta^\bullet)\right)(\theta-\theta_0) > 0$$

which clearly can hold only locally. This condition is analogous to the condition below equation (2.23).

Tsypkin's condition (c) is that

$$E\left(\tilde{\Phi}_k(\theta)e_k(\theta)e_k'(\theta)\tilde{\Phi}_k'(\theta)\right) \le \alpha(1+\theta'\theta) \; .$$

This condition will hold trivially and is further discussed in Appendix 2E. We see that Tsypkin's conditions imply local convergence and come from using the Lyapunov function of case (i) of Section 2.6.2 instead of that of case (ii).

2.7 Asymptotic Variances

As in the last section we deal first with general recursions and then with SA.

2.7.1 GENERAL RECURSIONS

From the basic equation (2.16) of Section 2.5, we construct once more a stochastic differential equation but this time for $\mu_k = \sqrt{k}\left(\hat{\theta}_{k-1}-\theta_0\right)$. First, multiply both sides of equation (2.16) by $\sqrt{k}$ and subtract $\sqrt{k-1}\ \left(\hat{\theta}_{k-1}-\theta_0\right)$. Then noting that

$$\sqrt{k}\ \left(\hat{\theta}_{k\cancel{-1}}-\theta_0\right) - \sqrt{k-1}\ \left(\hat{\theta}_{k-1}-\theta_0\right) = \tfrac{1}{2}\sqrt{k-1}\ \left(\hat{\theta}_{k-1}-\theta_0\right)/\sqrt{k} \simeq \tfrac{1}{2}I\mu(\tau)d\tau \ ,$$

we obtain

$$\sqrt{k}\ \left(\hat{\theta}_k-\theta_0\right) - \sqrt{k-1}\ \left(\hat{\theta}_{k-1}-\theta_0\right) \simeq kP_k\Psi_k\left(\hat{\theta}_{k-1}\right)\Sigma_0^{-1}e_k\left(\theta_0\right)/\sqrt{k}$$
$$+\ k^{-1}\left[I/2-kP_k\Psi_k\left(\hat{\theta}_{k-1}\right)\Sigma_0^{-1}\tilde{\Phi}_k'\left(\hat{\theta}_{k-1}\right)\right]\sqrt{k-1}\ \left(\hat{\theta}_{k-1}-\theta_0\right) \ .$$

Thus, with the same "Weiner" process as before (equation (2.19))

$$d\mu(\tau) = -\left[\hat{R}^{-1}(\tau)\Psi\left(\theta(\tau),\ \tau\right)\Sigma_0^{-1}\Phi'\left(\theta^{\bullet}(\tau),\ \tau\right)-I/2\right]\mu(\tau)d\tau + \hat{R}^{-1}(\tau)F_0^{\frac{1}{2}}dW(\tau) \ .$$

Because of the convergence of $\theta(\tau)$ with probability one, the second moment Q of this equation will be approximated locally by the second moment

$$d\mu(\tau) = -A_0\mu(\tau)d\tau + G_0^{-1}F_0^{\frac{1}{2}}dW(\tau)$$

where

$$G_0 = G\left(\theta_0\right) \ , \quad H_0 = H\left(\theta_0\right)$$

and where

$$A_0 = G_0^{-1}H_0 - I/2 \ .$$

It is well known that the covariance kernel of this equation satisfies (see Arnold, 1974, p. 134)

$$A_0Q + QA_0' = G_0^{-1}F_0G_0'^{-1} \ . \tag{2.25}$$

A p.d. solution to this Lyapunov equation requires the positive-definiteness of

$$A_0 + A_0' = G_0^{-1}H_0 + H_0G_0'^{-1} - I \ . \tag{2.26}$$

This will be true if $H_0 + H_0' - G_0$ is strictly p.d. and G_0 is p.d. This follows by considering the stability of the ordinary differential equation

$$\dot{X} = -A_0 X = -\left(G_0^{-1}H - I/2\right)X$$

with Lyapunov function $V = X'G_0X$. Then

$$\dot{V} = -X'\left(H_0 + H_0' - G_0\right)X$$

and so V decays to zero when $H_0 + H_0' - G_0$ is strictly p.d. This means, however, that A_0 has no eigen-values with negative real part, that is, that $A_0 + A_0'$ is p.d.

By the discussion of Section 2.5.1 the above two conditions hold if the recursion is to converge. For PER's $G_0 = F_0 = H_0$ so $A_0 = I/2$ and the Lyapunov equation has solution

$$Q = F_0^{-1} = \left[\int_{-\pi}^{\pi} dF_{\hat{\Phi}\hat{\Phi}}(\omega|\theta_0)\right]^{-1}$$

which is the Cramer-Rao lower bound. This is, the PER's can be expected to be asymptotically efficient. Note also that

$$kP_k \equiv R^{-1}(\tau) \rightarrow G_0^{-1} = Q$$

so that, for a PER, the gain update equation yields the asymptotic covariance matrix directly.

2.7.2 STOCHASTIC APPROXIMATION

With the Wiener process of Section 2.5.2, a similar approach to that in Section 2.6.1 yields

$$d\mu(\tau) = -\left[\Psi\left(\theta(\tau),\ \tau\right)\Sigma_0^{-1}\tilde{\Phi}'\left(\theta^{\bullet}(\tau),\ \tau\right) - \varphi(\tau)I/2\right]\mu(\tau)d\tau + F_0^{\frac{1}{2}}dW(\tau)$$

where

$$\mu_k = \left(\hat{\theta}_k - \theta_0\right)/\sqrt{a_k}\ , \quad \varphi_k = a_k^{-1} - a_{k-1}^{-1}\ .$$

In deriving this it is helpful to note $a_k^{-1}\left(1-\sqrt{a_k/a_{k-1}}\right) \simeq \frac{1}{2}\varphi_k$.

As before the second moment Q will be approximately the second moment of

$$d\mu(\tau) = -\left(H_0 - \varphi I/2\right)\mu(\tau)d\tau + F_0^{\frac{1}{2}}dW(\tau)$$

where it is assumed that $\lim \varphi(\tau) = \emptyset$ exists; for example, with $a_k = A^{-1}/(a+k)$ we have $\varphi_k = A^{-1} = \emptyset$. It follows then, that Q satisfies the Lyapunov equation

$$(H_0 - \varphi I/2)Q + Q(H_0' - \varphi I/2) = F_0 \ .$$

For this equation to have a p.d. solution it is necessary that $H_0 + H_0' - \varphi I$ be p.d. Unlike the case of general recursions, this condition is different to that required for convergence. For those SA schemes with $\Psi_k(\theta) = \tilde{\Phi}_k(\theta)$ we have

$$H_0 = F_0 = \int_{-\pi}^{\pi} dF_{\hat{\tilde{\Phi}}\hat{\tilde{\Phi}}}(\omega|\theta_0)$$

so that the Lyapunov equation is

$$F_0Q + QF_0 = F_0 + \varphi Q \ .$$

2.8 Misspecification

In this section we tackle the question of convergence and asymptotic covariance properties in the situation where the system S does not belong to the class of models M. It is expected that the recursive estimator will converge to the value $\overline{\theta}_0$ which ensures error decorrelation in the sense of satisfying

$$E\left\{\Psi_k(\overline{\theta}_0)\Sigma_0^{-1}e_k(\overline{\theta}_0)\right\} = 0 \ .$$

Convergence is discussed first and then asymptotic covariances are derived. It is suggested that the PER's will converge in this circumstance. Also the results of this section are suited to the analysis of instrumental variable (IV) estimators so an analysis of the IV recursion used by Young (1974) is given.

2.8.1 CONVERGENCE

The discussion here is close to that in previous sections but must be modified a little. In particular, since $e_k(\overline{\theta}_0)$ is no longer orthogonal to $\Psi_k(\hat{\theta}_{k-1})$, a slightly different Wiener process must be chosen; namely

$$W_k = F_0^{-\frac{1}{2}} \sum_1^k \Psi_s(\bar{\theta}_0)\Sigma_0^{-1} e_s(\bar{\theta}_0)/\sqrt{s} .$$

To see this, recall that in Section 2.5, equation (2.16), a Taylor series expansion was used on $e_k(\hat{\theta}_{k-1})$ to obtain a final r.h.s. term that was recognised (Section 2.6.1) as an increment of a Wiener process. Here we use a slightly different Taylor Series, namely one for

$$\Pi_k(\theta) = -\Psi_k(\theta)\Sigma_0^{-1} e_k(\theta) .$$

It then follows as in Section 2.5, equation (2.16), that

$$\hat{\theta}_k = \hat{\theta}_{k-1} - P_k\left(\partial\Pi_k/\partial\dot{\theta}_{k-1}'\right)\left(\hat{\theta}_{k-1}-\bar{\theta}_0\right) + P_k\Psi_k(\bar{\theta}_0)\Sigma_0^{-1} e_k(\bar{\theta}_0) .$$

Now from the definition of $\Pi_k(\theta)$,

$$\partial\Pi_k/\partial\theta' = \hat{\Psi}_k(\theta)\hat{\Phi}_k'(\theta) + \left(\partial\hat{\Psi}_k/\partial\theta_1 e_k(\theta), \ldots, \partial\hat{\Psi}_k/\partial\theta_p \hat{e}_k(\theta)\right)$$

so that we will be interested in the temporal average

$$H_N = N^{-1} \sum_1^N \partial\Pi_k/\partial\dot{\theta}_{k-1}' .$$

This will converge to $\bar{H}_0 = \bar{H}(\bar{\theta}_0) \triangleq H(\bar{\theta}_0) + K(\theta_0)$ with $H(\theta)$ as in Section 2.5, while

$$K(\theta) \triangleq E\left(\partial\hat{\Psi}_k/\partial\theta_1 \hat{e}_k(\theta), \ldots, \partial\Psi_k/\partial\theta_p \hat{e}_k(\theta)\right) .$$

The analysis now proceeds as before with the condition for local convergence being that both

$$G(\theta) \quad \text{and} \quad \bar{H}(\bar{\theta}_0) + \bar{H}'(\bar{\theta}_0) - G(\bar{\theta}_0) \tag{2.27}$$

be p.d.

Once more global convergence will follow for linear models (*cf.* Section 2.6.2).

Consider now the PER's; as in Section 2.6 we use the Lyapunov function suggested by Ljung (1978b)

$$V(\theta) = \tfrac{1}{2}E\left[e_k'(\theta)\Sigma_0^{-1} e_k(\theta)\right] .$$

By the same argument as in Section 2.6 we conclude the global convergence of the PER's in this $S \notin M$ case.

Returning to the general condition (2.27) consider the scalar IV recursion defined by (*cf.* Young, 1974)

$$\Psi_k(\beta) = \left(-x_{k-1}(\beta) \;\ldots\; -x_{k-n_a}(\beta) u_{k-1} \;\ldots\; u_{k-n_b}\right)' ,$$

$$\zeta_k = \left(-y_{k-1} \;\ldots\; -y_{k-n_a} u_{k-1} \;\ldots\; u_{k-n_b}\right)' ,$$

$$e_k(\beta) = y_k - \zeta_k'\beta .$$

Assume that $y_k - x_k(\beta_0) = n_k$ is a stationary process for some value β_0. The error decorrelation condition is

$$E\left(\Psi_k(\overline{\beta}_0)\left(y_k - \zeta_k'(\overline{\beta}_0)\right)\right) = 0 .$$

However this equation is clearly satisfied by the true value β_0, since it then becomes

$$E\left(\Psi_k(\beta_0) n_k\right) = 0 .$$

To investigate the conditions for convergence, note

$$\partial e_k/\partial\beta = -\tilde{\Phi}_k(\beta) = \zeta_k \;;\quad \partial^2 e_k/\partial\beta\partial\beta' = 0 .$$

Thus $K = 0$ so that $\overline{H}(\beta) = H(\beta)$. Also

$$H(\beta) = G(\beta) = E\left(\Psi_k(\beta)\zeta_k'\right) .$$

Thus the two conditions (2.27) become simply that $E\left(\Psi_k(\beta)\zeta_k'\right)$ be p.d. uniformly in $\beta \in D_c$.

Recalling $\zeta_k = \Psi_k(\beta_0) + \left(n_k' 0'\right)'$ this may be replaced by

$$E\left(\Psi_k(\beta)\Psi_k(\beta_0)'\right) .$$

It is hardly surprising that this be required to be positive definite since this is also required for the en bloc use of the IV estimator.

2.8.2 ASYMPTOTIC COVARIANCE MATRICES

Once more the argument parallels that in Section 2.7. The covariance matrix will obey a Lyapunov equation similar to equation (2.25) of Section 2.7.1 but with two changes

(i) F_0 is the asymptotic covariance matrix of

$$N^{-\frac{1}{2}} \sum_{1}^{N} \Psi_s(\theta_0)\Sigma_0^{-1}e_s(\theta_0) .$$

If we write the matrix $\Psi_k(\theta)$ in column form as

$$\Psi_k(\theta) = \left(\Psi_{1k}(\theta) \ \ldots \ \Psi_{pk}(\theta)\right)' ,$$

so that in the example just discussed

$$\Psi_{1k}(\theta) = -x_{k-1}(\theta) \ \ldots \ \Psi_{pk}(\theta) = u_{k-n_b} , \quad p = n_a + n_b ,$$

then, *neglecting fourth cumulant terms*, the matrix F_0 has i, j element (see Hannan, 1970, p. 209)

$$\int_{-\pi}^{\pi} \mathrm{tr}\left[f_{\Psi_i\Psi_j}(\omega)\Sigma_0^{-1}f_{ee}(\omega)\right] + \mathrm{tr}\left[f_{e\Psi_i}(\omega)\Sigma_0^{-1}f_{e\Psi_j}(\omega)\right]d\omega$$

where $f_{\Psi_i\Psi_j}(\omega)$ is the cross-spectrum between $\Psi_{ik}(\theta_0)$ and $\Psi_{jk}(\theta_0)$ and so on.

(ii) H_0 is replaced by $\bar{H}_0 = \bar{H}(\bar{\theta}_0)$ of Section 8.1.

In the example above we saw $K = 0$ so $\bar{H}_0 = H_0$, also $G_0 = H_0$ so the Lyapunov equation of Section 2.7.1 for the covariance matrix Q becomes

$$Q = G_0^{-1}F_0G_0'^{-1} .$$

This result could have been guessed from an en-bloc analysis.

2.9 Conclusions

The main aim of this chapter has been to present a unified framework for viewing SA, MR and recursive methods for the estimation of the time invariant parameters of a lumped parametric model of a dynamic system. The unified view was obtained by noting that just about any model could be defined by an equation for the prediction of the system output. This is a stochastic extension of the deterministic MR idea. Then recursions could be constructed by minimising the mean squared error of prediction. The PER recursions in general required the solution of a sensitivity equation however; for DD models this poses no difficulty. The following general points have emerged:

(i) A general method has been presented for constructing prediction error recursions (PER's) for just about any linear or nonlinear lumped parametric model of a *single* or *multivariable* dynamic system. Some of the recursions have appeared before, mainly in the single output case. In particular, recursions may be written down for PMD models with a dimension reduction in the observation space.

(ii) A plausibility analysis of the local convergence behaviour of a general recursion (as well as of SA) is obtained by converting it to a stochastic differential equation. This differential equation is analysed by looking at the behaviour of its first moments as well as by a Stochastic Lyapunov function argument; a discrete-time argument is given too.

(iii) It is argued that the asymptotic covariance matrix for both general recursions and SA obeys a Lyapunov equation. It is suggested that the PER's are asymptotically efficient, that is, they attain the Cramer-Rao lower bound.

(iv) The PER's when modified slightly are shown to still converge (though they may be biased) in the mis-specification case, when the system does not belong to the model set. In this sense they may be said to possess a robustness property.

APPENDIX 2A

An Example of an Infinite Past Recursion

Consider a MA(1) model

$$e_k(\alpha) = y_k/(1-\alpha L) .$$

Suppose α_k is the output of a recursion then the innovations process for an IP recursion is found by solving

$$e_k(\alpha_k) = y_k + \alpha_k e_{k-1}(\alpha_k)$$

recursively to obtain

$$e_k(\alpha_k) = y_k + \alpha_k y_{k-1} + \alpha_k^2 y_{k-2} + \ldots + \alpha_k^m y_{k-m} + \ldots .$$

The innovations process for a practical recursion is found by solving

$$e_k = y_k + \alpha_k e_{k-1}$$

recursively to obtain

$$e_k = y_k + \alpha_k y_{k-1} + \alpha_k \alpha_{k-1} y_{k-2} + \alpha_k \alpha_{k-1} \alpha_{k-2} y_{k-3} + \ldots + \alpha_k \alpha_{k-1} \cdots \alpha_{k-m+1} y_{k-m} + \ldots .$$

APPENDIX 2B

Convergence of the Self-Tuning Filter $\hat{y}_{k/k-}(\hat{\theta}_{k-1})$ and Self-Tuning State Estimator $\hat{x}_{k/k-}(\hat{\theta}_{k-1})$

Consider the Taylor Series expansion

$$\hat{y}_{k/k-}(\hat{\theta}_{k-1}) - \hat{y}_{k/k-}(\theta_0) = -\tilde{\Phi}_k'(\theta_{k-1}^{\bullet})(\hat{\theta}_{k-1}-\theta_0)$$

where $\|\theta_{k-1}^{\bullet}-\theta_0\| \le \|\hat{\theta}_{k-1}-\theta_0\|$. Next,

$$\|\hat{y}_{k/k-}(\hat{\theta}_{k-1})-\hat{y}_{k/k-}(\theta_0)\| \le \|\tilde{\Phi}_k(\theta_{k-1}^{\bullet})\| \|\hat{\theta}_{k-1}-\theta_0\| .$$

Now $\forall \theta \in D_c$ (some region of stability) $\Phi_k(\theta)$ is obtained by a *stable* filtering operation on the past of y_k and u_k so assume

$$\|\tilde{\Phi}_k(\theta)\| < \sum_1^{\infty} \lambda^i (\|y_{k-i}\|+\|u_{k-i}\|) , \quad (2B.1)$$

for some $\lambda < 1$, $\forall \theta \in D_c$. Now suppose y_k and u_k have finite third moments; then by Hölder's inequality so does $\|\tilde{\Phi}_k(\theta)\|$. But now Chebyshev's inequality and the Borel-Cantelli lemma imply

$$\|\tilde{\Phi}_k(\theta)\|/k^{1/3+\delta} \to 0 , \quad (2B.2)$$

with probability one (w.p.1) for some $\delta > 0$ and uniformly in $\theta \in D_c$.

However the w.p.1 convergence of $\hat{\theta}_{k-1}$ and so of $\theta_{k-1}^{\bullet}$ together with the uniformity of convergence in (2B.2) ensures $\|\tilde{\Phi}_k(\theta_{k-1}^{\bullet})\|/k^{1/3+\delta} \to$ w.p.1 .

Because convergence of $\hat{\theta}_k$ will be demonstrated using Lyapunov function arguments, it will be straightforward to prove also that

$$k^{\alpha}\|\hat{\theta}_k-\theta_0\| \to 0 \quad \text{w.p.1} \ , \quad \alpha < \tfrac{1}{2} \ . \tag{2B.3}$$

(This is discussed in a moment.) The convergence of the self-tuning filter now follows. A similar argument establishes the convergence of the self-tuning state estimator.

We now give a heuristic validation of claim (2B.3). Let $V_0 = \|\hat{\theta}_k-\theta_0\|^2$ where $\tau \simeq \ln k$. In the ensuing discussion we will obtain inequalities like

$$dV_\tau/d\tau < -V_\tau \ .$$

Integrating yields

$$V_t - V_s < -\int_s^t V_\sigma d\sigma \ ;$$

thus V_t is a bounded decreasing sequence and so has a limit. However, it follows also that

$$\int_0^\infty V_\sigma d\sigma < V_0 < \infty$$

so this limit must be zero.

Now we want to show $k^{1-\delta}V_k = e^{\tau(1-\delta)}V_\tau \to 0$ for some $\delta > 0$. So consider $W_\tau = e^{\tau(1-\delta)}V_\tau$.

Then

$$\begin{aligned} dW_\tau/d\tau &= (1-\delta)W_\tau + e^{\tau(1-\delta)}dV_\tau/d\tau \\ &\le (1-\delta)W_\tau - W_\tau \\ &= -\delta W_\tau \ . \end{aligned}$$

It now follows by the same argument as given for V_τ , that $W_\tau \to 0$.

APPENDIX 2C

Equivalence of Refined IV and a Certain PER

The equivalence at the end of Section 2.4 is not immediately clear. Consider then the scalar TFARMA model

$$e_k(\theta) = \left(1+C_\alpha(L)\right)\left(1+D_\alpha(L)\right)^{-1}\left[y_k-\left(1+A_\beta(L)\right)^{-1}B_\beta(L)u_k\right] .$$

It is straightforward to see

$$e_k(\theta) = y_k - \left(-\mathrm{x}'_{k-1}(\beta),\ \mathrm{u}'_{k-1},\ \mathrm{e}'_{k-1}(\theta),\ -\bar{\delta}'_{k-1}(\beta)\right)'\theta$$

where $\mathrm{u}_{k-1} = \left(u_{k-1} \ \dots \ u_{k-n_b}\right)'$ and so on, while $\bar{\delta}_k(\beta) = y_k - x_k(\beta)$ and $x_k(\beta) = \left(1+A_\beta(L)\right)^{-1}B_\beta(L)u_k$.

The point to see is that the recursion for the system parameters β may be uncoupled from that for the noise parameters α . This uncoupling will follow if we can demonstrate that $k^{-1}\mathrm{P}_k^{-1}$ approaches a block diagonal form. Now the (1, 2) block element in this matrix is the cross covariance between $de_k/d\beta$ and $de_k/d\alpha$. It is easily verified that

$$de_k/d\beta = \left(1+A_\beta(L)\right)^{-1}\left(-\tilde{\mathrm{x}}'_{k-1}(\beta),\ \tilde{\mathrm{u}}'_{k-1}\right)' ,$$

$$de_k/d\alpha = \left(1+D_\alpha(L)\right)^{-1}\left(\mathrm{e}'_{k-1}(\theta),\ -\bar{\delta}'_{k-1}(\beta)\right)' ,$$

where

$$\tilde{x}_k = \left(1+D_\alpha(L)\right)^{-1}\left(1+C_\alpha(L)\right)x_k$$

and so on.

If the recursion converges the first expression will be a vector of linear operations on u_k and so uncorrelated with the second which will consist of linear operations on the stationary processes $e_k(\theta_0)$ and $\bar{\delta}_k(\beta_0)$.

Finally, note that the non-symmetric IV recursion (see Young and Jakeman, 1977) apparently differs from the PER recursion in that the (1, 1) block element in $k^{-1}\mathrm{P}_k^{-1}$ is replaced by the covariance between

$\Psi_k(\beta)$ and $de_k/d\beta$ where

$$\Psi_k(\beta) = \left(1+A_\beta(L)\right)^{-1}\left(-\tilde{y}'_{k-1},\ \tilde{u}'_{k-1}\right)'$$
$$= \left(1+A_\beta(L)\right)^{-1}\left(-\tilde{x}'_{k-1},\ \tilde{u}'_{k-1}\right)' + \left(1+A_\beta(L)\right)^{-1}\left(-\bar{\delta}'_{k-1},\ 0'\right)' .$$

However if the recursion converges the second term is uncorrelated with $de_k/d\beta$ (which, as mentioned above, depends on the past of u_k only) and so asymptotically the two recursions are identical. This uncoupling of noise parameter and system parameter recursions does not appear to be possible for the other common linear time series models.

APPENDIX 2D

Replacement of the Taylor Series of Section 2.5 for Linear Models

In deriving the s.d.e. (2.20) we have used the Taylor series (2.15) of Section 2.5. However, for linear time series models there is an alternative to the Taylor series. To see this consider the scalar ARMA model

$$\left(1+C_\alpha(L)\right)y_k = \left(1+D_\alpha(L)\right)e_k(\alpha) .$$

Now by definition if α_0 is the true value $\left(1+C_0(L)\right)y_k = \left(1+D_0(L)\right)e_k(\alpha_0)$. Subtracting these two yields

$$\left(C_\alpha(L)-C_0(L)\right)y_k = \left(1+D_0(L)\right)\left(e_k(\alpha)-e_k(\alpha_0)\right) + \left(D_\alpha(L)-D_0(L)\right)e_k(\alpha) .$$

Thus

$$e_k(\alpha) - e_k(\alpha_0) = \left(1+D_0(L)\right)^{-1}\left[\left(C_\alpha(L)-C_0(L)\right)y_k-\left(D_\alpha(L)-D_0(L)\right)e_k(\alpha)\right]$$
$$= \left(1+D_0(L)\right)^{-1}\Phi_k(\alpha)'\left(\alpha-\alpha_0\right) \qquad (2D.1)$$

where

$$\Phi_k(\alpha) = \left(-y_{k-1}\ \cdots\ -y_{k-n_c}\ e_{k-1}(\alpha)\ \cdots\ e_{k-n_d}(\alpha)\right) .$$

Notice that

$$-\tilde{\Phi}_k(\alpha) = de_k(\alpha)/d\alpha = \left(1+D_\alpha(L)\right)^{-1}\Phi_k(\alpha) ,$$

so that expression (2D.1) takes the place of the Taylor series. The important point however is that now the transfer function between $\tilde{\Phi}_k(\alpha)$ and $\Phi_k(\alpha)$ is $\left(1+D_0(L)\right)^{-1}$ which does not depend on α . Thus any uniformity condition is automatically satisfied. Similar results may be obtained for all the usual linear models.

APPENDIX 2E

Tsypkin's Conditions for the Convergence of a SA Scheme

As in Appendix 2B take $\tilde{\Phi}_k(\theta)$, $e_k(\theta)$ to be bounded uniformly in θ by an expression of the type given in (2B.1). If y_k has finite fourth moment and if u_k is bounded (if deterministic) or has finite fourth moment (if stationary) then by the Cauchy-Schwarz inequality twice,

$$E\left(\tilde{\Phi}_k(\theta)e_k(\theta)e_k'(\theta)\Phi_k'(\theta)\right) \leq (1-\lambda)^{-3}\sum_1^\infty \lambda^i E\left(\|y_{k-i}\|^4+\|u_{k-i}\|^4\right) \leq \text{const.}$$

CHAPTER 3

SIMULATIONS

3.1 Introduction

In this chapter some of the ideas of Chapter 2 will be illustrated by simulation exercises: also the use of control variates for variance reduction is illustrated (Example 1). Example 2a investigates the general variance formula suggested in Section 2.7.1 by Monte-Carlo (M-C) simulation of an ARMA(1, 1) model. This is pursued in Examples 2B, 2C where two ARMAX models are considered. Finally Example 3 illustrates the direct estimation of 'continuous-time parameters' from sampled data.

Before beginning however, recall the use of a fader (Söderström *et al.*, 1974) as a means of stabilising the sometimes erratic small sample behaviour of recursions. The idea is to weight down the influence of past data so as to allow, early in the recursion, more notice to be taken of new information. The basic algorithm takes the form

$$\theta_k = \theta_{k-1} + P_{k-1}\varphi_k e_k / v_k \ ,$$

$$e_k = y_k - \varphi_k' \theta_{k-1} \ ,$$

$$v_k = \lambda(k) + \varphi_k' P_{k-1} \varphi_k \ ,$$

$$\lambda(k) = \lambda_0 \lambda(k-1) + 1 - \lambda_0 \ ,$$

$$P_k = P_{k-1} - P_{k-1}\varphi_k \varphi_k' P_{k-1} / v_k \ .$$

Typical choices for initial conditions are $\lambda_0 = .95$, $\lambda(0) = .95$. The use of a fader turns out to be invaluable. Notice that $\lambda(k) \to 1$ as $k \to \infty$.

3.2 Examples

EXAMPLE 1. Here we illustrate the use of a control variate to reduce variance in simulation work. The basic idea is due to Hendry (1975). We consider the MA(1) process

$$y_k = \varepsilon_k + \alpha_0 \varepsilon_{k-1}$$

and investigate the sampling variance of the RML$_1$ recursion for the parameter α_0. Recall the recursion as

$$\alpha_k = \alpha_{k-1} + P_{k-1} e_{k-1} e_k \Big/ \left(1+e_{k-1}^2 P_{k-1}\right) ,$$

$$e_k = y_k - e_{k-1}\alpha_{k-1} ,$$

$$P_k = P_{k-1} - P_{k-1}^2 e^2 \Big/ \left(1+e_{k-1}^2 P_{k-1}\right) .$$

Notice that recursion may be summed to yield

$$\alpha_N = \left(\sum_1^N e_{k-1}^2\right)^{-1} \left(\sum_1^N e_{k-1} y_k\right) .$$

A control variate is an artificial (in the sense that it cannot be computed in practice but can be computed in a set of Monte-Carlo experiments) statistic whose statistical properties are (more or less) known and whose behaviour is expected to be close to that of the statistic under investigation. Let us introduce then the quantity

$$a_N = \left(\sum_1^N \varepsilon_{k-1}^2\right)^{-1} \left(\sum_1^N \varepsilon_{k-1} y_k\right)$$

where ε_k are the true innovations sequence (and regressors) generated in the typical Monte-Carlo (M-C) experiment: denote $\sigma^2 = E\left(\varepsilon_k^2\right)$. Let α_{Ni} be the value of α_N in the ith M-C experiment; similarly denote a_{Ni} . Define also the sample averages over m experiments

$$\overline{\alpha}_N(m) = m^{-1} \sum_1^m \alpha_{Ni} ,$$

$$\overline{a}_N(m) = m^{-1} \sum_1^m a_{Ni} .$$

Denote also $\mu_N = E\left(a_N\right)$; we might, loosely, think of μ_N as $\lim_{m \to \infty} \overline{a}_N(m)$, that is, the average over an infinite number of independent experiments: notice $E\left(\overline{a}_N(m)\right) = \mu_N$. In the present example we have, to order N^{-2} (in mean square), $\mu_N = 0$.

Consider firstly the estimation of bias. Suppose, after completing m M-C experiments we form

$$\hat{\alpha}_N(m) = \overline{\alpha}_N(m) - \overline{a}_N(m) + \mu_N .$$

Then clearly

$$E\left(\hat{\alpha}_N(m)\right) = E\left(\overline{\alpha}_N(m)\right) = E\left(\alpha_N\right) .$$

Thus $\hat{\alpha}_N(m) - \alpha_0$ is an unbiased estimator of the bias $E\left(\alpha_N\right) - \alpha_0$. However

$$\begin{aligned} \operatorname{var}\left(\hat{\alpha}_N(m)\right) &= \operatorname{var}\left(\overline{\alpha}_N(m)-\overline{a}_N(m)\right) \\ &= \operatorname{var}\left(\alpha_N-a_N\right)/m \\ &= m^{-1}\left(\operatorname{var}\left(\alpha_N\right)+\operatorname{var}\left(a_N\right)-2\operatorname{cov}\left(a_N, \alpha_N\right)\right) . \end{aligned}$$

If α_N, a_N are highly positively correlated (as is clearly the case in the present example) there is great potential for obtaining

$$\operatorname{var}\left(\hat{\alpha}_N(m)\right) << \operatorname{var}\left(\overline{\alpha}_N(m)\right) .$$

Indeed in the present example it may be shown that

$$\operatorname{var}\left(\hat{\alpha}_N(m)\right) = o(\log N/N)^2$$

which is clearly a vast improvement on

$$\operatorname{var}\left(\overline{\alpha}_N(m)\right) = O\left(N^{-1}\right) .$$

We turn now to variances. Denote the variance of α_N by $\sigma^2_{\alpha N}$ and that of a_N by σ^2_{aN}. Define the M-C variances by

$$\overline{\sigma}^2_{\alpha N}(m) = (m-1)^{-1} \sum_1^m \left(\alpha_{Ni}-\overline{\alpha}_N(m)\right)^2 ,$$

$$\overline{\sigma}^2_{aN}(m) = (m-1)^{-1} \sum_1^m \left(a_{Ni}-\overline{a}_N(m)\right)^2 .$$

We see $E\left(\overline{\sigma}^2_{aN}(m)\right) = \sigma^2_{aN}$. Now we form

$$\hat{\sigma}^2_{\alpha N}(m) = \overline{\sigma}^2_{\alpha N}(m) - \overline{\sigma}^2_{aN}(m) + \sigma^2_{aN} .$$

Once more $\hat{\sigma}^2_{\alpha N}(m)$ is an unbiased estimator of $\sigma^2_{\alpha N}$ with reduced variance over $\bar{\sigma}^2_{\alpha N}(m)$ if α_N, a_N are highly positively correlated. In the present example, we have, to order N^{-2},

$$\mathrm{var}(a_N) = \sigma^2_{aN} = \sigma^2/N .$$

Table 1 shows the results of M-C simulations illustrating the above discussion. In the table the normalised variance is $\bar{\sigma}^2_{\alpha N}(m)N$ which is the M-C estimate of $\mathrm{var}\left(\sqrt{N}\ (\alpha_N-\alpha_0)\right)$ (which converges, as $N \to \infty$, to $\sigma^2 = 1$) and the normalised bias is $\sqrt{N}\ \left(\bar{\alpha}_N(m)-\alpha_0\right)$: 'raw' standard deviation is $\bar{\sigma}_{\alpha N}(m)$, similarly for 'raw' bias: other entries have similar meaning. In each case the improved estimate of variance and bias for 10 experiments should be compared with the 'raw' estimate for 30 experiments.

EXAMPLE 2A. This example is concerned with the variance formula suggested in Section 2.7.1 (equation 25). We consider this formula for the RML_1 recursion beginning with the ARMA(1, 1) process

$$y_k + a_0 y_{k-1} = \varepsilon_k + c_0 \varepsilon_{k-1}$$

with $a_0 = -.8$, $c_0 = .7$ (*cf.* Söderström *et al.*, 1974, model (6.1.2)). We have

$$\varphi_k(\alpha_0) = (-y_{k-1}\ \varepsilon_{k-1})$$

and (recalling the definitions of Section 2.5)

$$F_0 = G_0 = E\left(\varphi_k(\alpha_0)\varphi_k'(\alpha_0)\right) ,$$

$$H_0 = E\left[\varphi_k(\alpha_0)\left(1+c_0 L\right)^{-1}\varphi_k'(\alpha_0)\right] .$$

Had we proceeded by analogy with linear linear regression we would have been led to posit the limit coraviance matrix of $\sqrt{N}\ (\alpha_N-\alpha_0)$ to be $\sigma^2 G_0^{-1}$ $\left(\sigma^2 = E\left[\varepsilon_k^2\right]\right)$; it can be shown that except in the MA(1) case this cannot be true (see the discussion in Section 4 of Chapter 8). The variance suggested in Section 2.7.1 is the solution Q of the Lyapunov Equation

$$\left(G_0^{-1}H_0 - I/2\right)Q + Q\left(G_0^{-1}H_0 - I/2\right)' = G_0^{-1} .$$

Now G_0, H_0 may be calculated by straightforward computations (by Cauchy's Residue theorem); thus the (1, 2) element of H_0 is

$$-E\left[y_k\varepsilon_k\left(1+c_0L\right)^{-1}\right] = -E\left[\left(1+a_0L\right)^{-1}\left(1+c_0L\right)\varepsilon_k\left(1+c_0L\right)^{-1}\varepsilon_k\right]$$
$$= -\sigma^2\left(1-c_0^2\right)\left(1-a_0c_0\right)^{-1} .$$

The matrices are then

$$G_0 = \begin{bmatrix} \sigma_{yy} & 0 \\ 0 & \sigma^2 \end{bmatrix} ,$$

$$H_0 = \sigma^2\begin{bmatrix} h_{11} & -\left(1-c_0^2\right)\left(1-a_0c_0\right)^{-1} \\ 1 & 1 \end{bmatrix}$$

with

$$\sigma_{yy} = \sigma^2\left(1+c_0^2-2c_0a_0\right)\left(1-a_0^2\right)^{-1} ,$$

$$h_{11} = \left(1-a_0c_0\right)\left(1-a_0^2\right)^{-1} .$$

The Lyapunov equation was solved by programming the solution given by Jameson (1968). Results are tabulated in Table 2A: the agreement with the conjectured result is evident. However this accord is not always so clear as we now illustrate in Example 2B.

EXAMPLE 2B. Here we expand Example 2A by adding a pseudo random binary noise (PRBN) input sequence to give an ARMAX model

$$y_k = x_k + \lambda n_k$$
$$= \left(1+a_0L\right)^{-1}b_0u_{k-1} + \lambda\left(1+a_0L\right)^{-1}\left(1+c_0L\right)\varepsilon_k$$

with $a_0 = -.8$, $b_0 = 1.$, $c_0 = .7$ (*cf.* Söderström *et al.*, 1974, model (6.1.2)); λ is defined in a moment.

The G_0, H_0 matrices are $\left(\sigma^2 = 1\right)$

$$G_0 = \begin{pmatrix} \sigma_{xx}+\lambda^2\sigma_{nn} & 0 & -\lambda^2 \\ 0 & 1 & 0 \\ -\lambda^2 & 0 & \lambda^2 \end{pmatrix},$$

$$H_0 = \begin{pmatrix} h_{11} & c_0\left(1-a_0c_0\right)^{-1} & -\lambda^2\left(1-c_0^2\right)\left(1-a_0c_0\right)^{-1} \\ 0 & 1 & 0 \\ -\lambda^2 & 0 & \lambda^2 \end{pmatrix}$$

where

$$\sigma_{xx} = E\left(x_k^2\right) = \left(1-a_0^2\right)^{-1},$$

$$\sigma_{nn} = E\left(n_k^2\right) = \left(1+c_0^2-2c_0a_0\right)\left(1-a_0^2\right)^{-1},$$

$$h_{11} = \left[\left(1+c_0a_0\right)\left(1-c_0a_0\right)^{-1}\left(1-c_0^2\right)^{-1}+\lambda^2\right]\left(1-a_0^2\right)^{-1},$$

$$\lambda^2 = \left(\sigma_{xx}/\sigma_{nn}\right)/\mathrm{SNR}$$

where SNR is the signal to noise ratio.

The results are given in Tables 2B1, 2B2. The agreement with the conjectured result is less clear though the results are complicated by their sensitivity to the choice of fader values (see Table 2B1) (notice the biases for each fader value are similar). One point is clear however, the Lyapunov equation solution always provides a conservative estimate of the sample variances. For this reason its use seems desirable until the result is proved or else a correct solution found.

EXAMPLE 2C. This is an extension of Example 2B to an ARMAX system of second order. We take the model (*cf.* Söderström *et al.*, 1974, model 6.1.3)

$$\begin{aligned} y_k &= x_k + \lambda n_k \\ &= \left(1+a_1L+a_2L^2\right)^{-1}\left[\left(b_1L+b_2L^2\right)u_k+\lambda\left(1+c_1L+c_2L^2\right)\varepsilon_k\right] \end{aligned}$$

with

$$\left(a_1,\ a_2,\ b_1,\ b_2,\ c_1,\ c_2\right) = (-1.5,\ .7,\ 1.,\ .5,\ -1.,\ .2)\ .$$

The G_0, H_0 matrices are $\left(\sigma^2 = 1\right)$

$$G_0 = \left[\begin{array}{ccc|ccc} g(0) & 0 & -\lambda^2 & g(1) & -b_1 & -\lambda^2\left(c_1-a_1\right) \\ 0 & 1 & 0 & 0 & 0 & 0 \\ -\lambda^2 & 0 & \lambda^2 & 0 & 0 & 0 \\ \hline g(1) & 0 & 0 & g(0) & 0 & -\lambda^2 \\ -b_1 & 0 & 0 & 0 & 1 & 0 \\ -\lambda^2\left(c_1-a_1\right) & 0 & 0 & -\lambda^2 & 0 & \lambda^2 \end{array}\right] ,$$

$$H_0 = \left[\begin{array}{ccc|ccc} a(0) & -b(0) & -\lambda^2 c(0) & a(1) & -b(1) & -\lambda^2 c(1) \\ 0 & 1 & 0 & 0 & 0 & 0 \\ -\lambda^2 & 0 & \lambda^2 & 0 & 0 & 0 \\ \hline a(-1) & -b(-1) & -\lambda^2 c(-1) & a(0) & -b(0) & -\lambda^2 c(0) \\ -b_1 & -c_1 & 0 & 0 & 1 & 0 \\ -\lambda^2 a_1 & 0 & -\lambda^2 c_1 & -\lambda^2 & 0 & \lambda^2 \end{array}\right] ,$$

where

$$g(m) = E\left(y_k\, y_{k-m}\right) ,$$

$$a(m) = E\left[y_k\left(1+c_1L+c_2L^2\right)^{-1} y_{k-m}\right] ,$$

$$b(m) = E\left[y_k\left(1+c_1L+c_2L^2\right)^{-1} u_{k-m}\right] ,$$

$$c(m) = E\left[y_k\left(1+c_1L+c_2L^2\right)^{-1} \varepsilon_{k-m}\right] .$$

For these matrices the φ_k vector

$$\left(-y_k\ -y_{k-1}\ u_k\ u_{k-1}\ \varepsilon_k\ \varepsilon_{k-1}\right)$$

has been reorganised as

$$\left(-y_k\ u_k\ \varepsilon_k\ -y_{k-1}\ u_{k-1}\ \varepsilon_{k-1}\right) .$$

The calculation of the quantities $a(m)$, $b(m)$, $c(m)$ is rather tedious and they were generated by programming the method presented recently by Hwang (1978). Results are given in Table 2C. They bear out the comments at the end of Example 2B.

EXAMPLE 3. Here we illustrate the direct estimation of the continuous time parameters of a Time Series Model (*cf.* Section 2.3, Remark (8)). Let us take a simple second order example

$$y_k = x_k + \varepsilon_k \;,$$

$$x_t = x_t(\beta_0) = \left(D^2 + a_1 D + a_0\right)^{-1} b_0 u_t$$

where D is the derivative operator. Take

$$a_0 = 2.25 \;, \quad a_1 = 1.5 \;, \quad b_0 = 1.9 \;,$$

that is, a second order system with natural frequency $\omega_n = \sqrt{a_0} = 1.5$ rad/sec , damping ratio $\zeta = \frac{1}{2}a_1/\sqrt{a_0} = \frac{1}{2}$ and gain $K = b_0/a_0 = .84$: notice the decay time constant is $\tau = (\zeta\omega_n)^{-1} = 1.33$ sec . As pointed out in Section 2.3, Remark (6) we can write x_t as

$$d/dt\ \mathrm{q}_u = \begin{pmatrix} -a_1 & 1 \\ -a_0 & 0 \end{pmatrix} \mathrm{q}_u + \binom{1}{0} u_t \;,$$

$$x_t = \left(b_1\ b_0\right)\mathrm{q}_u \;, \quad b_1 = 0 \;.$$

The input sequence was a sum of three sinusoids of unit amplitudes and frequencies $\omega_1 = .9\omega_{max}$, $\omega_2 = 1.2\omega_{max}$, $\omega_3 = .6\omega_{max}$, where $\omega_{max} = \sqrt{1-2\zeta^2}\ \omega_n$ is the resonant frequency.

The x_t sequence was generated by solving the differential equation for q_u by a standard fourth order Runge-Kutta algorithm at intervals of $\tau/20$. The data was generated (at SNR = 10 and 1) by adding white Gaussian noise to every fourth value, that is, sampling at an interval $\tau/5$. To get the correct scale factor for the SNR we need to know the power of the x_t sequence (the u_t sequence has power 3/2). This power can be calculated theoretically but is easily obtained when generating the x_t sequence as

$$\sum_1^N x_k^2 \Big/ \sum_1^N u_k^2 \ .$$

Of course the results obtained will depend on the choice of generation and sampling interval: see Goodwin and Payne (1977). The gradients for the recursion are

$$\begin{bmatrix} -de_t/db_1 \\ -de_t/db_0 \end{bmatrix} = q_u(t) \ ,$$

$$\begin{bmatrix} -de_t/da_1 \\ -de_t/da_0 \end{bmatrix} = -\left[D^2+a_1D+a_0\right]^{-2} \binom{0}{1} b_0 u_t$$

$$= -\left[D^2+a_1D+a_0\right]^{-1} \binom{0}{1} x_t$$

$$= -q_x(t) \ .$$

Two differential equations for $q_x(t)$, $q_u(t)$ were solved, once again, by a standard fourth-order Runge-Kutta. The recursion is then

$$\theta_k = \theta_{k-1} + P_{k-1}\varphi_k e_k / v_k \ ,$$

$$e_k = y_k - x_k \ ,$$

$$v_k = \lambda(k) + \varphi_k' P_{k-1} \varphi_k \ ,$$

$$\lambda(k) = \lambda_0 \lambda(k-1) + 1 - \lambda_0 \ ,$$

$$P_k = P_{k-1} - P_{k-1}\varphi_k\varphi_k' P_{k-1}/v_k \ ,$$

$$\varphi_k = \left(q_u(k) q_x(k)\right)' \ .$$

Figures 1 (SNR = 1) and 2 (SNR = 10) show (initial values in all cases $(a_0, a_1, b_0, b_1) = (1.5, 1.2, 1.5, 0.)$) the type of result obtained. Some bias is evident, particularly in the b parameters and this shows up in the MC results of Table 3. It is apparent that good small sample results depend on high quality *a priori* estimates. The bias in the b parameters is no doubt due to the choice of generation/sampling intervals and nature of the input signal. Further theoretical investigation of these matters needs to be made.

TABLE 1

Monte-Carlo simulations of RML₁ for MA(1)

$$y_k = (1+.7L)\varepsilon_k$$

illustrating use of control variates

$$(\lambda_0 = .95, \lambda(0) = .95)$$

SAMPLE SIZE	250		500	
No. experiments	10	30	10	30
Normalised variance	.56	.92	.78	1.20
Improved variance	1.05	1.36	1.26	1.18
Normalised bias	-.73	-.81	-.413	-.357
Improved bias	-.52	-.77	-.384	-.457
Raw S.D.	.035	.058	.034	.052
Improved S.D.	.066	.086	.054	.051
Raw bias	-.046	-.051	-.018	-.016
Improved bias	-.033	-.049	-.017	-.020

TABLE 2A

Covariance Matrices of RML₁ for

$$(1-.8L)y_k = (1+.7L)\varepsilon_k$$

Parameter	G_0^{-1}	Q	Monte-Carlo (30 experiments) $(N = 1500;\ \lambda_0 = .95,\ \lambda(0) = .95)$
a_0, b_0	$\begin{bmatrix} .16 & . \\ .16 & 1.16 \end{bmatrix}$	$\begin{bmatrix} .69 & . \\ .52 & 1.35 \end{bmatrix}$	$\begin{bmatrix} .74 & . \\ . & 1.51 \end{bmatrix}$

Corresponding Correlation Matrices

Parameter	G_0^{-1}	Q	Monte-Carlo
a_0, b_0	$\begin{bmatrix} 1.00 & . \\ .37 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.00 & . \\ .54 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 1.00 & . \\ .43 & 1.00 \end{bmatrix}$

TABLE 2B1 (SNR = 10.)

Covariance Matrices of RML₁ for

$$(1-.8L)y_k = 1.u_{k-1} + (1+.7L)\varepsilon_k$$

Parameters	$a_0 \quad b_0 \quad c_0$	Correlation Matrices
G_0^{-1}	$\begin{pmatrix} .013 & . & . \\ . & .038 & . \\ . & . & 1.01 \end{pmatrix}$	$\begin{pmatrix} 1.00 & & \\ .00 & 1.00 & \\ .11 & .00 & 1.00 \end{pmatrix}$
Q	$\begin{pmatrix} .055 & & \\ .01 & .038 & \\ .04 & -.01 & 1.02 \end{pmatrix}$	$\begin{pmatrix} 1.00 & & \\ .20 & 1.00 & \\ .18 & -.04 & 1.00 \end{pmatrix}$
Monte-Carlo 30 experiments N = 1500 $\lambda_0 = .99$, $\lambda(0) = .99$	$\begin{pmatrix} .036 & . & . \\ . & .057 & . \\ . & . & .924 \end{pmatrix}$	$\begin{pmatrix} 1.00 & . & . \\ -.50 & 1.00 & . \\ .47 & .14 & 1.00 \end{pmatrix}$
Normalised bias	-.007 -.020 .121	
Monte-Carlo 30 experiments N = 1500 $\lambda_0 = .995$, $\lambda(0) = .995$	$\begin{pmatrix} .045 & . & . \\ . & .072 & . \\ . & . & 1.166 \end{pmatrix}$	$\begin{pmatrix} 1.00 & . & . \\ -.61 & 1.00 & . \\ .40 & -.23 & 1.00 \end{pmatrix}$
Normalised bias	.041 .025 .100	

TABLE 2B2 (SNR = 1.)

Covariance Matrices of RML₁ for

$$(1-.8L)y_k = 1.0u_k + (1+.7L)\varepsilon_k$$

Parameters	$a_0 \quad b_0 \quad c_0$	Correlation Matrices
G_0^{-1}	$\begin{pmatrix} .074 & & \\ .00 & .383 & \\ .074 & .0 & 1.07 \end{pmatrix}$	$\begin{pmatrix} 1.00 & & \\ .00 & 1.00 & \\ .26 & .0 & 1.00 \end{pmatrix}$
Q	$\begin{pmatrix} .32 & & \\ -.05 & .38 & \\ .25 & -.05 & 1.16 \end{pmatrix}$	$\begin{pmatrix} 1.00 & & \\ -.15 & 1.00 & \\ .40 & -.08 & 1.00 \end{pmatrix}$
Monte-Carlo 30 experiments N = 2000 $\lambda_0 = .99 = \lambda(0)$	$\begin{pmatrix} .18 & . & . \\ . & .56 & . \\ . & . & .80 \end{pmatrix}$	$\begin{pmatrix} 1.00 & . & . \\ .10 & 1.00 & . \\ .16 & .19 & 1.00 \end{pmatrix}$

TABLE 2C (SNR = 10.)

Variances of RML_1 for

$$\left(1-1.5L+.7L^2\right)y_k = \left(L+.5L^2\right)u_k + \left(1-L+.2L^2\right)\varepsilon_k$$

Parameters	a_1	b_1	c_1	a_2	b_2	c_2
G_0^{-1} variances	.34	1.24	1.34	.31	1.24	1.08
Q variances	.48	1.24	.95	.32	5.98	2.40
Monte-Carlo 30 experiments N = 2000 λ_0 = .9 $\lambda(0)$ = .9	1.07	1.64	12.95†	1.03	3.90	3.48
Monte-Carlo Average	-1.49	.99	-.85	.69	.51	.09
Raw Bias	.007	-.006	.150	-.010	.010	-.100
Standard Deviation	.024	.028	.080	.022	.044	.041

† This value must be disregarded from an asymptotic point of view since the bias estimate shows clearly that parameter c_1 has not yet reached an asymptotic behaviour. This persistent bias is evident also in the examples of Söderström *et al.* (1974).

TABLE 3

Variances and biases of a Prediction Error Recursion for the continuous-time parameters of

$$y_k = x_k + \varepsilon_k \ ; \quad x_t = \left(D^2+1.5D+2.25\right)^{-1} 1.9u_t$$

with

$$u_t = \sin .9\omega_{\max} t + \sin 1.2\omega_{\max} t + \sin .6\omega_{\max} t \ ,$$

$$\omega_{\max} = 1.07 \ ; \quad \text{sampling interval } \tau/5 \ ; \quad \tau = 1.33 \text{ sec}$$

30 experiments, $\lambda_0 = .94$, $\lambda(0) = .94$

$N = 500$		SNR = 1			SNR = 10		
True Parameter Values		Monte-Carlo Average	Raw SD	Raw Bias	Monte-Carlo Average	Raw SD	Raw Bias
a_0	2.25	2.34	.25	.09	2.32	.18	.007
a_1	1.5	1.39	.15	-.10	1.37	.05	-.13
b_0	1.9	1.67	.14	-.23	1.65	.04	-.25
b_1	0.	-.40	.10	-.40	-.38	.03	-.38

$N = 1000$		SNR = 1			SNR = 10		
True Parameter Values		Monte-Carlo Average	Raw SD	Raw Bias	Monte-Carlo Average	Raw SD	Raw Bias
a_0	2.25	2.26	.22	.02	2.25	.17	.005
b_0	1.5	1.41	.11	-.09	1.38	.03	-.11
a_1	1.9	1.71	.10	-.18	1.69	.03	-.20
b_1	0.	-.38	.08	-.38	-.37	.02	-.37

FIGURE 1

PREDICTION ERROR RECURSION FOR CONTINUOUS TIME PARAMETERS(SNR=1)

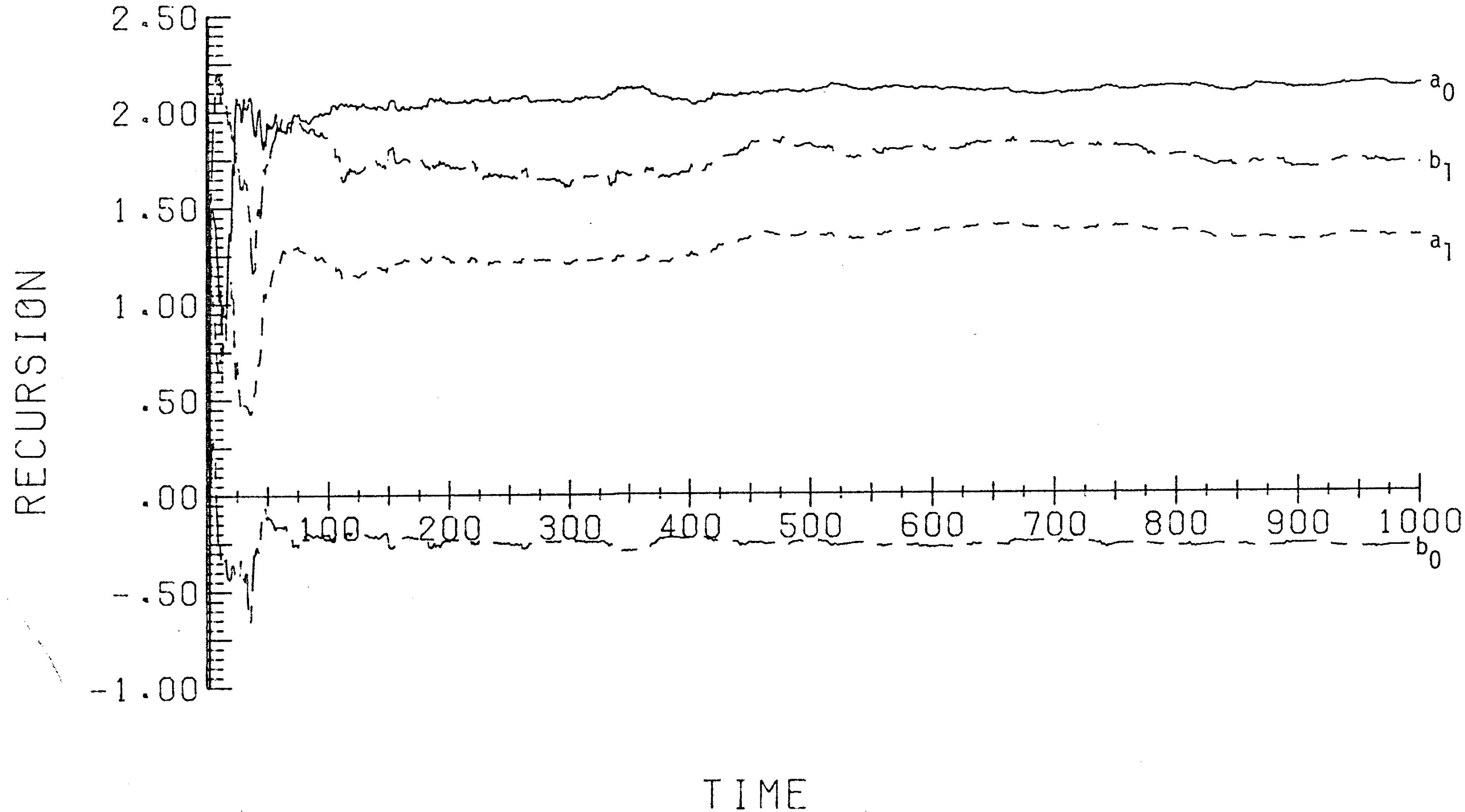
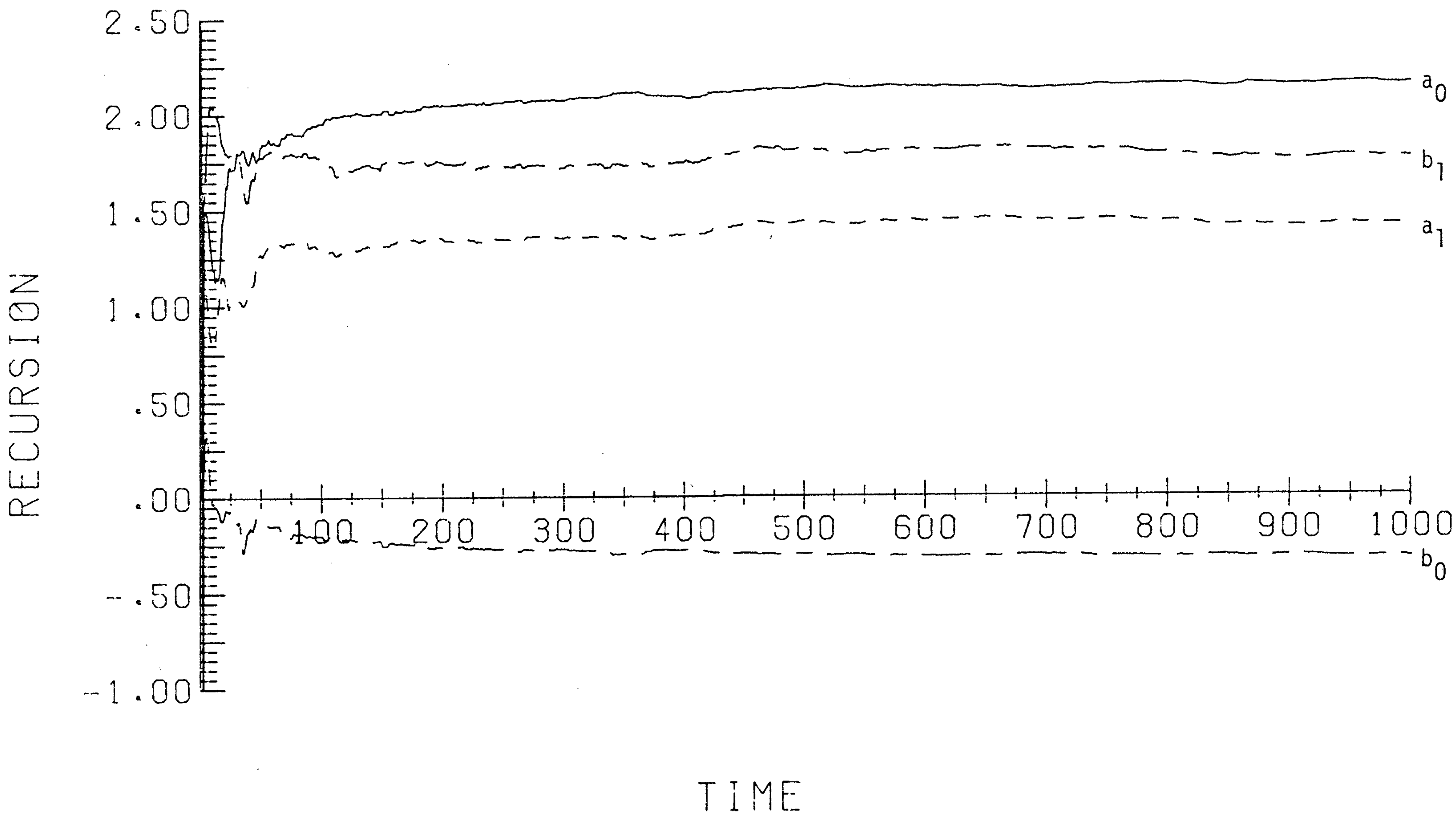


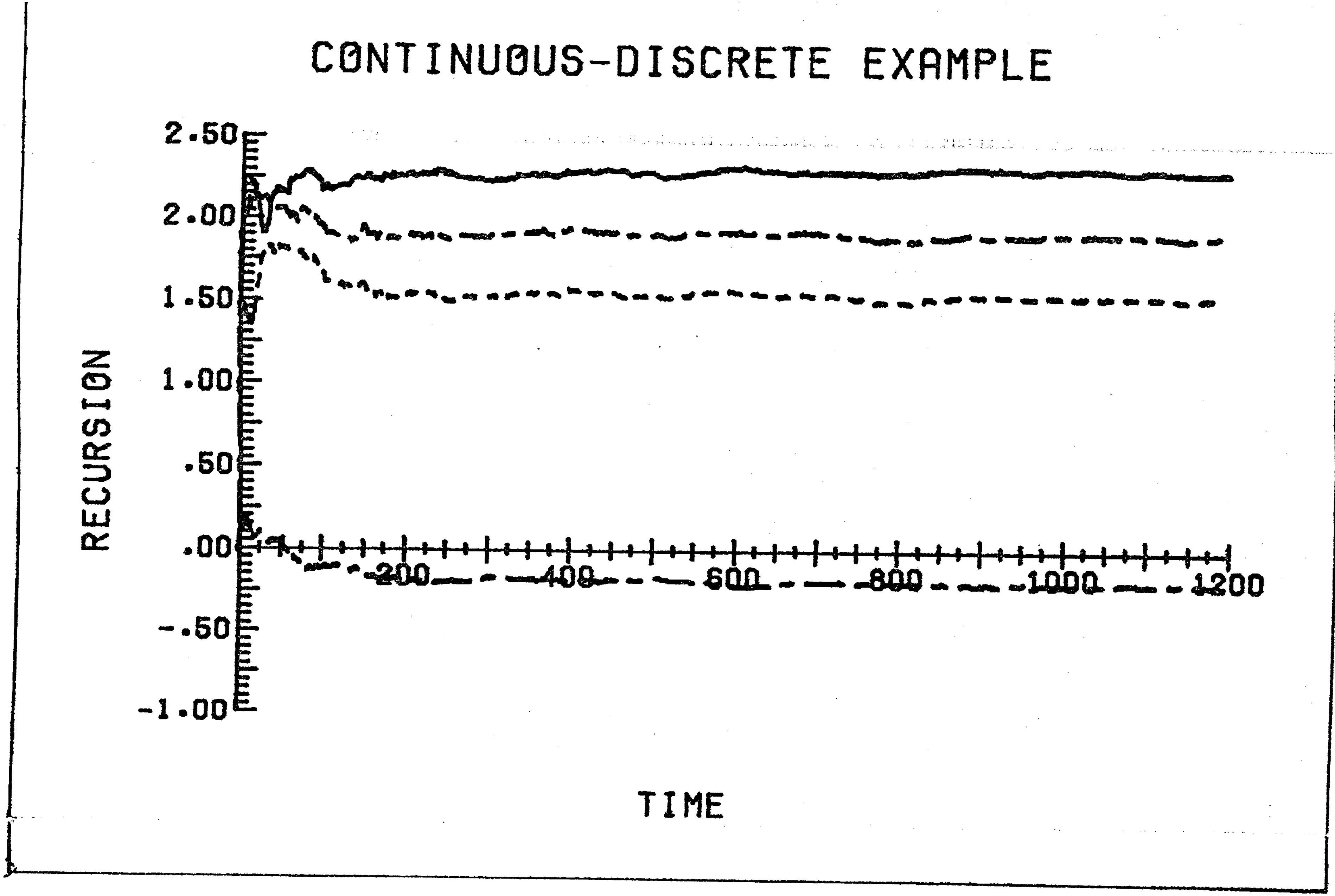
FIGURE 2

PREDICTION ERROR RECURSION FOR CONTINUOUS TIME PARAMETERS(SNR=10)



I am grateful to Dr A. Jakeman for pointing out that the bias in the continuous-discrete simulation of Chapter 3 may be reduced by more rapid sampling. A graph of the type of result obtained is attached (SNR = 10 ; sample interval $\tau/10$; $\lambda_0 = \lambda(0) = .95$). Dr Jakeman also supplied some Monte-Carlo results for the model without noise. These are given below and show the bias reduction clearly.

N = 500	TRUE VALUES		
SAMPLE INTERVALS	$q_0 = 2.25$	$a_1 = 1.5$	$b_0 = 1.9$
$\tau/5$	1.90	1.40	1.65
$\tau/30$	2.19	1.48	1.85
$\tau/90$	2.23	1.49	1.88
$\tau/270$	2.24	1.47	1.88


CONTINUOUS-DISCRETE EXAMPLE
2.50
2.00
1.50
1.00
.50
.00
-.50
-1.00
RECURSION
200
400
600
800
1000
1200
TIME

PART II

PREAMBLE

The primary aim of this part of the thesis is to address, in a rigorous fashion, some of the problems implicit in the heuristic discussion of Part I. Chapter 4 is devoted to an investigation of the w.p.1 convergence of the least squares estimator in a linear regression where, for example, the regressors may depend on past values of the observed process. With this in mind, and recalling, from Chapter 1, that the estimator can be written as a recursion via Plackett's algorithm, we see that this model provides a prototype for recursions under the special assumption that the regressors are exactly known. Indeed it is the experience gained with this problem that enabled the author to solve the problem in Chapter 8 (see later). One approach to strong convergence in Chapter 4 is by means of a Stochastic Lyapunov function (Bucy, 1965). This is usually a quadratic form (such as $(\hat{\beta}_n-\beta_0)'(\hat{\beta}_n-\beta_0)$) that obeys a supermartingale like inequality. The author developed this idea after reading the article by Caines (1975) however the same notion is implicit in the approach of Blum (1954), Gladyshev (1965) and others.

So far as proving the convergence of recursions is concerned, the present proofs (Hannan, 1976, 1978c; Ljung, 1977a, 1977b) require that the scheme be monitored - a costly affair in practice. It had been the author's hope that this Stochastic Lyapunov function approach would allow a proof of convergence of some simple recursions such as RML_1 (Söderström et al. 1974) without need of monitoring. With the RML_1 recursion as given in the above reference (see also the Introduction to Chapter 8), this has not been accomplished except in the MA(1) case (already discussed in any case by Hannan, 1976, though by a different proof): see Section 8.3 below. However by considering a modified version of RML_1, presented in Moore and Ledwich (1978) (and due to Young, 1974, in the ARMA case), it has been possible to prove convergence without monitoring: this is the subject of Section 8.2. In the last section of Chapter 8 it is pointed out that the nature of the CLT for RML_1 is unclear.

In Chapter 5 we begin an investigation of the second order properties of SA schemes. The approach in Chapter 5 forms the basis of Chapters 6 and

7. *The idea is simply to subtract off the part of the* SA *scheme that obeys the limit law (Central Limit Theorem* (CLT)*, Law of the Iterated Logarithm* (LIL) *and so on) and, using elementary arguments, show the remainder to be asymptotically negligible. In searching for a proof of a* CLT *for a recursion, the author was led naturally to consider the limit laws for* SA *and so came to the belief that these limit laws ought to be obtainable without summing up the difference equations, but by using a time varying version of the final value theorem of Z-transform theory.*

In Chapter 5 *also, we establish, inter alia, an Invariance Principle* (IP) *that makes rigorous the heuristic notion, introduced in Chapter* 1 *and pursued in Chapter* 2*, that, asymptotically the (suitably) normalised output of the* SA *obeys a stochastic differential equation. The* CLT *and* IP *are proved for general* a_n, c_n *sequences and are given without requiring a.s. convergence as is usual (in say Sacks,* 1958*): in particular the conditions* $\sum_1^\infty a_n^2/c_n^2 < \infty$ *or* $\sum_1^\infty a_n^2 < \infty$ *are not required.*

Kushner (1978a) *and Ibragimov and Khasminskii* (1973) *provide some discussion of uses to which these* IP*'s may be put: see also McLeish* (1976) *and Kersting* (1971).

Chapter 6 *is based on the observation that the same idea of subtracting off the part that obeys the limit law can be used to prove convergence for* SA *schemes with dependent noise. The essentially deterministic nature of the convergence behaviour thus becomes clear. A* CLT *and* IP *are also derived. The last part of Chapter* 6 *discusses the relation of the present approach to recent work of Kushner* (1977, 1978a) *and Ljung* (1978a) *who also consider dependent noise cases.*

In Chapter 7*, we review quickly Hannan's* (1978c) *proof of convergence of* RML_2 *(Söderström et al.,* 1974*) and use his results to obtain a* CLT *and* IP. *The intuitive notion that* RML_2 *is asymptotically efficient is substantiated. Section* 7.4 *contains the Invariance Principle.*

CHAPTER 4

STRONG CONSISTENCY OF LEAST SQUARES ESTIMATORS IN LINEAR MODELS

4.1 Introduction

In this chapter conditions are given for the almost sure convergence of a least squares estimator in a linear regression whose regressors may depend on the past of the observed process (so that there is 'feedback') and whose disturbances form a martingale difference sequence. When there is no feedback, weaker conditions are given and the disturbances are allowed to form a stationary process.

4.2 Statement of Problem and Conditions

Consider the linear model

$$y_s = B_0'x_s + u_s \ , \quad s = 1, 2, \ldots \ ,$$

where u_s is an m-vector of disturbances with uniformly bounded variances, x_s a p-vector of regressors. Let F_s denote the σ-algebra generated by y_s, y_{s-1} and so on. As has been pointed out by Anderson and Taylor (1977) the model is a regression model when the x_s are nonstochastic and a sequential control or design model when x_s is F_{s-1} measurable (i.e. there may be feedback): a special case here is the first order autoregression when $x_s = y_{s-1}$. We are interested in the a.s. convergence of the least squares estimator of B_0,

$$\hat{B}_n = V_n^{-1} \sum_{s=1}^{n} x_s y_s'$$

or

$$\hat{B}_n - B_0 = V_n^{-1} \sum_{1}^{n} x_s u_s'$$

where $V_n = \sum_1^n x_s x_s'$: provided V_p is nonsingular so is V_n , $n \geq p$.

Denote $\tau_n^{-1} = \mathrm{tr}\left(V_n^{-1}\right)$. Thus $\tau_{n-1} \leq \tau_n$. Call the largest eigenvalue of V_n , λ_n and the smallest σ_n . Note that

$$1/\sigma_n \leq 1/\tau_n \leq p/\sigma_n \ . \tag{4.1}$$

Next write

$$Q_n = \mathrm{tr}\left(\left(\hat{B}_n - B_0\right)'\left(\hat{B}_n - B_0\right)\right) \quad \text{and} \quad W_n = \mathrm{tr}\left(\left(\hat{B}_n - B_0\right)' V_n \left(\hat{B}_n - B_0\right)\right) \geq Q_n \sigma_n \ .$$

Clearly $\hat{B}_n - B_0 \to 0$ a.s., that is $Q_n \to 0$ a.s. if $W_n/\tau_n \to 0$ a.s. Unless otherwise stated assume only that x_s is F_{s-1} measurable. It will be assumed throughout that $E(Q_n)$, $E(W_n)$ exist. By the Cauchy-Schwarz inequality

$$W_n \leq n \sum_1^n \mathrm{tr}\left(u_s u_s'\right) \ .$$

Now it is also assumed throughout that u_s have uniformly bounded variances so that for some constant t , $\mathrm{tr}\, E\left(u_s u_s'\right) \leq t$. Thus $E\left(W_n\right) \leq pn^2 t < \infty$; this ensures the existence of the conditional expectation $E\left(W_n | F_{n-1}\right)$. Assume also that $\tau_p > \varepsilon$ where ε is a small positive constant. Then

$$\varepsilon Q_n \leq \tau_n Q_n \leq \sigma_n Q_n \leq W_n$$

so that $E\left(Q_n\right) < \infty$ also.

The following conditions will be of interest.

A. $E\left(u_s | F_{s-1}\right) = 0$, $E\left(u_s u_s' | F_{s-1}\right) = \Sigma$ a.s. where Σ is a constant positive definite matrix.

A$_1$. $E\left(u_s | F_{s-1}\right) = 0$, $E\left(u_s u_s' | F_{s-1}\right) = \Sigma_s$ a.s. where Σ_s is a sequence of constant positive definite (p.d.) matrices whose smallest eigenvalues are bounded below by a positive constant (this is to ensure the noise sequence is not degenerate) and there is an upper bound t with $\mathrm{tr}\left(\Sigma_s\right) \leq t$.

A$_2$. As for A$_1$ together with

$$\sup_s E\left\{\left| u_s u_s' - \mathrm{tr}\left(\Sigma_s\right)\right|^{1+\eta} \mid F_{s-1}\right\} < d < \infty \ ,$$

for some constant d and some $\eta > 0$.

A'. The disturbances u_s form a weakly stationary process with power spectrum $F(\omega)$. There is an upper bound c , a constant a.s., with $\mathrm{tr}\big(F(\omega)\big) \le c$. Thus when the x_s are non-stochastic

$$\begin{aligned} E\big(\mathrm{tr}\big(\hat{B}_n - B_0\big)'\big(\hat{B}_n - B_0\big)\big) &= \sum_1^n \sum_1^n x_s' V_n^{-2} x_t E\big(u_s' u_t\big) \\ &= \sum_1^n \sum_1^n x_s' V_n^{-2} x_t \int_{-\pi}^{\pi} \mathrm{tr}\big(F(\omega)\big) e^{i\omega(s-t)} d\omega \\ &= \int_{-\pi}^{\pi} \mathrm{tr}\big(F(\omega)\big) \left(\sum_1^n V_n^{-1} x_s e^{i\omega s}\right)' \left(\sum_1^n V_n^{-1} x_s e^{-i\omega s}\right) d\omega \\ &\le c \sum_1^n x_s' V_n^{-2} x_s = c\tau_n^{-1} . \end{aligned}$$

That is $E\big(Q_n\big) \le c\tau_n^{-1}$. Also note

$$E\big(u_s' u_s\big) = \int_{-\pi}^{\pi} \mathrm{tr}\big(F(\omega)\big) d\omega \le c2\pi .$$

B. That $\sigma_n \to \infty$ or equivalently, in view of (4.1), that $\tau_n \to \infty$.

C. $\overline{\lim}\, \lambda_n/\sigma_n < \infty$. This condition will hold when V_n can be normed by a scalar such as n : it is the condition usually employed.

Next let ν_n denote a sequence of positive numbers with

$\sum_p^\infty \big(\nu_n - \nu_{n-1}\big)/\tau_n < \infty$.

Two examples (of use below) are (given implicitly); $\tau_n = \nu_n^{1+\delta}$ for some $\delta > 0$ and $\tau_n = \nu_n \log^{1+\delta} \nu_n$ for some $\delta > 0$. Thus, for example,

$$\sum_p^\infty \big(\nu_n - \nu_{n-1}\big)/\tau_n \le \sum_p^\infty \int_{\nu_{n-1}}^{\nu_n} dt/t^{1+\delta} = \int_{\nu_{p-1}}^{\infty} dt/t^{1+\delta} < \infty .$$

C'. $\overline{\lim}\, \lambda_n/\sigma_n \nu_n < \infty$ with ν_n as above.

C". $\overline{\lim}\, \lambda_n/\sigma_n \nu_n^2 < \infty$ with ν_n as above.

D. $\overline{\lim}\, n/\sigma_n < \infty$.

D'. $\overline{\lim}\, n^{\frac{1}{2}} \log^{1+\delta} n/\sigma_n < \infty$ for some $\delta > 0$.

In view of (4.1), this is equivalent to $\overline{\lim}\, n^{\frac{1}{2}} \log^{1+\delta} n/\tau_n < \infty$.

Denote the ith component of x_n by x_{in} , $i = 1 \ldots p$; denote also $V_{in} = \sum_1^n x_{is}^2$. Let D_n be the $p \times p$ diagonal matrix whose ii element is V_{in} ; then $\mathrm{R}_n = \mathrm{D}_n^{-\frac{1}{2}} \mathrm{V}_n \mathrm{D}_n^{-\frac{1}{2}}$ is the correlation matrix of the x_n .

E. $\underline{\lim}\, \mathrm{R}_n > 0$ a.s.

Notice that C implies E.

Next denote $t_n = \mathrm{tr}(\mathrm{V}_n)$ and notice

$$\lambda_n \le t_n \le p\lambda_n \ , \quad 0 \le d\lambda_n \le dt_n \ , \tag{4.2}$$

where $d\lambda_n = \lambda_n - \lambda_{n-1}$ and so on.

F. (i) $\sum_p^\infty d\lambda_n/(\lambda_n \sigma_n) < \infty$ or equivalently

(ii) $\sum_1^\infty dt_n/(t_n \tau_n) < \infty$ or equivalently

(iii) $\sum_1^\infty dV_{in}/(V_{in} \tau_n) < \infty$, $i = 1 \ldots p$.

These equivalences are now demonstrated. From (4.1), (4.2) above we have immediately (ii) ⇒ (i). We show (i) ⇒ (ii) and then (ii) ⟺ (iii).

Now by (4.1) and (4.2), (i) implies (i)':

(i)' $\sum_1^\infty d\lambda_n/(t_n \tau_n) < \infty$.

Suppose (i)' holds yet $\sum_1^\infty dt_n/(t_n \tau_n) = \infty$, then

$$0 = \lim_{N\to\infty} \left[\sum_1^N d\lambda_n/\left(t_n\tau_n\right)\right] \Big/ \left[\sum_1^N dt_n/\left(t_n\tau_n\right)\right] = \lim_{N\to\infty} d\lambda_N/dt_N \ .$$

where the second equality follows by the discrete L'Hospital rule (a special case of the Toeplitz lemma - see Knopp (1956, p. 46)). But the discrete L'Hospital rule (in reverse) shows this limit to be

$$\lim \lambda_n/t_n \geq p^{-1} \quad \text{by (4.2).}$$

Hence (i) ⇒ (ii) by contradiction.

Next,

$$\begin{aligned} dt_n/t_n &= \sum_1^p dV_{in}/\left(\sum_1^p V_{in}\right) \\ &\leq \sum_1^p dV_{in}/V_{in} \ . \end{aligned}$$

Thus (iii) ⇒ (ii). Let (ii) hold and yet

$$\sum_1^\infty dV_{in}/\left(V_{in}\tau_n\right) = \infty$$

for at least one $i = i_0$ say. Then once more, by the discrete L'Hospital rule

$$\begin{aligned} 0 &= \lim \left[\sum_1^N dt_n/\left(t_n\tau_n\right)\right] \Big/ \left[\sum_1^N dV_{i_0n}/\left(V_{i_0n}\tau_n\right)\right] \\ &= \lim dt_n V_{i_0n}/\left(dV_{i_0n}t_n\right) = 1 + \sum_{i\neq i_0} \pi_i \ . \end{aligned}$$

where π_i (which need not be specified) are positive. Once more the equivalence follows by contradiction.

F'. $\sum_1^\infty d\lambda_k(\log k)^2/\left(\lambda_k\sigma_k\right) < \infty$.

F". $\lim_{n\to\infty} \sigma_n^{-1} \sum_1^n d\lambda_k/\lambda_k = 0$.

Notice that F implies F" via Kronecker's lemma.

F"'. $\lim_{n\to\infty} \sigma_n^{-1} \sum_1^n d\lambda_k/\lambda_k$ exists and is finite.

Consider the following two regimes.

(i) Feedback possible so that x_s may be F_{s-1} measurable. Then $Q_n \to 0$ a.s. in each case (a) - (c) below.

(a) A_1, B, C' hold. Proof by Theorem I below.

(b) A_1, B, E, F hold. This result is implicit in Hannan (1978). A different proof to Hannan's is given in Theorem I' below.

(c) A_2, B, E, F" hold. Proof by Theorem II below.

Earlier results were

(a)' A, B, C hold. Feigin (1975). Anderson and Taylor (1977).

(a)" A_1, B hold and $m = 1 = p$. Drygas (1976). This result is also a straightforward consequence of a well known strong law for Martingales: see Stout (1974, p. 159) also Neveu (1965, p. 150, Corollary).

In connexion with (a)", the usual proof proceeds via Kronecker's lemma. The results of Feigin/Anderson, Taylor (a)', used a vector form of Kronecker's lemma which apparently holds only under condition C. The proofs to be given below do not use a vector Kronecker's lemma.

(ii) No feedback so that the x_s are nonstochastic. Then $Q_n \to 0$ a.s. in cases (d) - (f).

(d) A_2, B, E, F''' hold. Proof by Theorem II' below.

(e) (A_1 or A'), B, D' hold. Proof by Theorem III below.

(f) A_1, B, C" hold. Proof by Theorem IV below.

Other results are

(d)' A, B, E, F' hold. Hannan (1978). This result follows from the method of subsequences (Stout, 1974) via a Menchoff inequality (Stout, 1974).

(e)' A, B, D hold. Anderson and Taylor (1977).

4.3 Some Basic Lemmas

The matrix inversion lemma (MIL) (Bodewig, 1956, p. 30) and its properties will be used repeatedly:

$$V_n^{-1} = V_{n-1}^{-1} - V_{n-1}^{-1}x_n x_n' V_{n-1}^{-1} \Big/ \left(1+x_n' V_{n-1}^{-1} x_n\right) \tag{4.3}$$

or

$$V_{n-1}^{-1} = V_n^{-1} + V_n^{-1}x_n x_n' V_n^{-1} \Big/ \left(1-x_n' V_n^{-1} x_n\right) . \tag{4.4}$$

It is easily proved by multiplying on the left by V_n or V_{n-1}. It follows that

$$V_n^{-1}x_n = V_{n-1}^{-1}x_n \Big/ \left(1+x_n'V_{n-1}^{-1}x_n\right) , \qquad (4.3)^1$$

$$V_{n-1}^{-1}x_n = V_n^{-1}x_n \Big/ \left(1-x_n'V_n^{-1}x_n\right) , \qquad (4.4)^1$$

$$x_n'V_n^{-1}x_n = x_n'V_{n-1}^{-1}x_n \Big/ \left(1+x_n'V_{n-1}^{-1}x_n\right) \le 1 , \qquad (4.3)^2$$

$$x_n'V_{n-1}^{-1}x_n = x_n'V_n^{-1}x_n \Big/ \left(1-x_n'V_n^{-1}x_n\right) , \qquad (4.4)^2$$

$$x_n'V_n^{-2}x_n = x_n'V_n^{-1}V_{n-1}^{-1}x_n \Big/ \left(1+x_n'V_{n-1}^{-1}x_n\right) \le x_n'V_n^{-1}V_{n-1}^{-1}x_n , \qquad (4.3)^3$$

$$x_n'V_{n-1}^{-2}x_n = x_n'V_n^{-1}V_{n-1}^{-1}x_n \Big/ \left(1-x_n'V_n^{-1}x_n\right) , \qquad (4.4)^3$$

$$x_n'V_{n-1}^{-1}V_n^{-1}x_n = x_n'V_n^{-2}x_n \Big/ \left(1-x_n'V_n^{-1}x_n\right) . \qquad (4.4)^4$$

The following lemma is basic to all that follows.

LEMMA 1. *If* ν_n *is a sequence of positive numbers with*

$$\sum_p^\infty \left(\nu_n-\nu_{n-1}\right)/\tau_n < \infty$$

then

$$\sum_p^\infty \nu_{n-1}x_n'V_n^{-1}V_{n-1}^{-1}x_n < \infty .$$

Proof. Take traces in MIL (4.4) and use MIL $(4.4)^4$ to see that

$$x_n'V_n^{-1}V_{n-1}^{-1}x_n = \tau_{n-1}^{-1} - \tau_n^{-1}$$

so that multiplying both sides by ν_{n-1} and summing gives

$$\sum_p^\infty \nu_{n-1}x_n'V_n^{-1}V_{n-1}^{-1}x_n \le \tau_{p-1}^{-1} + \sum_p^\infty \left(\nu_n-\nu_{n-1}\right)/\tau_n .$$

The following lemma will be used repeatedly.

LEMMA 2. *If* $\alpha_n \geq 0$ *then* $\sum_1^\infty \alpha_n < \infty$ *if* $\sum_1^\infty E(\alpha_n) < \infty$.

Proof. See Lukacs (1975, p. 80).

The theorems below are based on the following considerations. Recall now

$$\hat{B}_n - B_0 = \hat{B}_{n-1} - B_0 + V_n^{-1} x_n \hat{u}_n'$$

where

$$\hat{u}_n = y_n - \hat{B}_{n-1}' x_n = u_n - (\hat{B}_{n-1} - B_0)' x_n .$$

Thus recalling $Q_n = \operatorname{tr}\left((\hat{B}_n - B_0)'(\hat{B}_n - B_0)\right)$,

$$\text{(I)} \quad Q_n = Q_{n-1} + \alpha_n^2 + 2u_n' b_n - 2a_n' b_n$$

where

$$\alpha_n^2 = x_n' V_n^{-2} x_n \hat{u}_n' \hat{u}_n ,$$

$$a_n' = x_n'(\hat{B}_{n-1} - B_0) , \quad b_n = (\hat{B}_{n-1} - B_0)' V_n^{-1} x_n .$$

Further, under condition A or A_1,

$$\text{(I)}' \quad E(Q_n | F_{n-1}) \leq Q_{n-1} + x_n' V_n^{-2} x_n (t + a_n' a_n) - 2a_n' b_n .$$

Alternatively we find (this is derived in a moment)

$$\text{(II)} \quad W_n = W_{n-1} + x_n' V_n^{-1} x_n u_n' u_n - a_n' a_n \left(1 - x_n' V_n^{-1} x_n\right) + 2u_n' a_n \left(1 - x_n' V_n^{-1} x_n\right)$$

so that, under condition A or A_1,

$$\text{(II)}' \quad E(W_n | F_{n-1}) \leq W_{n-1} + x_n' V_n^{-1} x_n t .$$

To see II begin with

$$V_n(\hat{B}_n - B_0) = V_n(\hat{B}_{n-1} - B_0) + x_n \hat{u}_n' .$$

Thus

$(\hat{B}_n-B_0)'V_n(\hat{B}_n-B_0)$

$$= \left\{(\hat{B}_{n-1}-B_0)'+\hat{u}_n x_n' V_n^{-1}\right\}\{V_n(\hat{B}_{n-1}-B_0)+x_n\hat{u}_n'\}$$
$$= (\hat{B}_{n-1}-B_0)'V_n(\hat{B}_{n-1}-B_0) + a_n\hat{u}_n + \hat{u}_n a_n' + \hat{u}_n\hat{u}_n' x_n' V_n^{-1} x_n$$
$$= (\hat{B}_{n-1}-B_0)'V_{n-1}(\hat{B}_{n-1}-B_0) + a_n a_n' + a_n\hat{u}_n' + \hat{u}_n a_n' + \hat{u}_n\hat{u}_n' x_n' V_n^{-1} x_n$$

and putting $\hat{u}_n = u_n - a_n$ and taking traces yields II.

Next recall the following version of the martingale convergence theorem.

MGCT*. *If* T_{n-1}, α_n, β_n *are non negative random variables, measurable with respect to an increasing sequence of* σ*-algebras* F_{n-1}, *and satisfy*

$$E(T_n | F_{n-1}) \leq T_{n-1} + \alpha_n - \beta_n .$$

Then when $\sum_1^\infty \alpha_n < \infty$ *a.s. we have* $\sum_1^\infty \beta_n < \infty$ *a.s. and* $T_n \to T$ *a.s. a finite non-negative random variable. (In future* T *will always denote such a variate.)*

Proof. From $E(T_n | F_{n-1}) \leq T_{n-1} + \alpha_n$ and $\sum_1^\infty \alpha_n < \infty$ a.s. we deduce $T_n \to T$ a.s. (Neveu, 1975, p. 34). Next set $b_n = \sum_1^n \beta_s$ so

$$E(T_n + b_n | F_{n-1}) \leq T_{n-1} + b_{n-1} + \alpha_n .$$

The same result of Neveu ensures $T_n + b_n \to T'$ so that $b_n \to b < \infty$ a.s. as required.

Remarks. (1) The inequality involving T_n occurs also in Stochastic Approximation (Blum, 1954). Indeed one of the points of Plackett's algorithm is to see how regression can be viewed as a form of Stochastic Approximation.

(2) A variate such as T_n is also called a stochastic Lyapunov function - Bucy (1965).

4.4 Four Theorems on Strong Convergence

In Theorems I, I' below we apply MGCT* to II'. In Theorems II, II' we investigate W_n/τ_n by summing II. In Theorems III, IV we demonstrate directly the convergence of those series obtained by summing I.

THEOREM I (Feedback allowed). *If* A_1, B, C' *hold then* $Q_n \to 0$ *a.s.*

Proof. Divide II' throughout by τ_n. Set $T_n = W_n/\tau_n$,

$$E\left(T_n | F_{n-1}\right) \le T_{n-1} + t x_n' V_n^{-1} x_n / \tau_n - T_{n-1}\left(\tau_n - \tau_{n-1}\right)/\tau_n .$$

Now by MIL $(4.4)^2$,

$$\begin{aligned} x_n' V_n^{-1} x_n / \tau_n &= x_n' V_{n-1}^{-1} x_n \left(1 - x_n' V_n^{-1} x_n\right)/\tau_n \\ &= \left(x_n' V_{n-1}^{-1} x_n \Big/ \left(\tau_n \nu_{n-1} x_n' V_{n-1}^{-2} x_n\right)\right)\left(1 - x_n' V_n^{-1} x_n\right) x_n' V_{n-1}^{-2} x_n \nu_{n-1} \\ &\le \left(\lambda_{n-1}/\tau_n \nu_{n-1}\right) \nu_{n-1} x_n' V_n^{-1} V_{n-1}^{-1} x_n \end{aligned}$$

where MIL $(4.4)^3$ has also been used

$$\le \left(\lambda_{n-1}/\tau_{n-1}\nu_{n-1}\right)\nu_{n-1} x_n' V_n^{-1} V_{n-1}^{-1} x_n \quad \text{since} \quad \tau_n^{-1} \le \tau_{n-1}^{-1} .$$

Thus $\sum_p^\infty x_n' V_n^{-1} x_n / \tau_n < \infty$ by Lemma 1 and condition C'.

Thus by the MGCT*, $T_n \to T$ and $\sum_1^\infty T_{n-1}\left(\tau_n - \tau_{n-1}\right)/\tau_n < \infty$ whereupon $T = 0$ since by condition B and Kronecker's lemma we cannot have $\sum_1^\infty \left(\tau_n - \tau_{n-1}\right)/\tau_n < \infty$.

THEOREM I' (Feedback allowed). *If* A_1, B, E, F *hold then* $Q_n \to 0$ *a.s.*

Proof. According to the proof of Theorem I we need only show

$$\sum_p^\infty x_n' V_n^{-1} x_n / \tau_n < \infty .$$

Now call $u_n = D_n^{-\frac{1}{2}} x_n$ so that

$$x_n' V_n^{-1} x_n = u_n' R_n^{-1} u_n .$$

If ζ_n denotes the smallest eigenvalue of R_n then

$$x_n' V_n^{-1} x_n / \tau_n < \zeta_n^{-1} u_n' u_n / \tau_n = \zeta_n^{-1} \sum_{i=1}^{p} dV_{in} / \left(V_{in} \tau_n\right)$$

then the result follows from conditions E and F.

Before proceeding, some lemmas are required.

LEMMA 3. *Let* X_s *be a martingale difference sequence w.r.t. an increasing sequence of* σ*-algebras* F_s. *Let* J_s *be a positive non-decreasing sequence of* F_{s-1} *measurable random variables. Suppose*

$$\sum_1^\infty E\left(|X_s|^{1+\delta} \mid F_{s-1}\right) / J_s^{1+\delta} < \infty \quad \textit{for some} \quad 0 \le \delta \le 1 \tag{4.5}$$

then $J_n \uparrow \infty$ *implies* $J_n^{-1} \sum_1^n X_s \to 0$ *a.s.*

Proof. Condition (4.5) ensures $\sum_1^\infty X_s / J_s < \infty$ a.s., Stout (1974, p. 67). Then Kronecker's lemma gives the result.

LEMMA 3' (MGC$\hat{\text{T}}$). *With* X_s, F_s, J_s *as in Lemma* 3, *suppose*

$$\sum_1^\infty E\left(X_s^2 | F_{s-1}\right) / J_s^2 < \infty \quad a.s. \tag{4.6}$$

Then

$$J_n < \infty \quad a.s. \quad \textit{implies} \quad J_n^{-1} \sum_1^n X_s < \infty \quad a.s.,$$

$$J_n \to \infty \quad a.s. \quad \textit{implies} \quad J_n^{-1} \sum_1^n X_s \to 0 \quad a.s.$$

Proof. Set $T_n = J_n^{-1} \sum_1^n X_s$, $0 \le dJ_n = J_n - J_{n-1} \le J_n$, then

$$T_n = T_{n-1}\left(1 - dJ_n / J_n\right) + X_n / J_n$$

so

$$E\left(T_n^2 | F_{n-1}\right) = T_{n-1}^2\left(1-dJ_n/J_n\right)^2 + E\left(X_n^2 | F_{n-1}\right)/J_n^2$$
$$\leq T_{n-1}^2\left(1-dJ_n/J_n\right) + E\left(X_n^2 | F_{n-1}\right)/J_n^2$$

since $dJ_n \geq 0$, $dJ_n/J_n \leq 1$. Now from MGCT*, if (4.6) holds then $T_n \to T$ a.s. and $\sum_1^\infty T_{n-1}^2 dJ_n/J_n < \infty$ a.s. The first conclusion is thus established. The second conclusion follows also since $J_n \to \infty$ ensures $\sum_1^\infty dJ_n/J_n = \infty$ and we conclude $T = 0$ a.s. to avoid a contradiction. Of course the second conclusion follows also from Lemma 3 with $\delta = 1$.

Remarks. (a) With $J_n = I_n = \sum_1^n E\left(X_s^2 | F_{s-1}\right)$, Lemma 3' is given in Neveu (1972, p. 150): note that

$$\sum_2^\infty E\left(X_s^2 | F_{s-1}\right)/I_s^2 \leq \sum_2^\infty \int_{I_{s-1}}^{I_s} dt/t^2 < \int_{I_1}^\infty dt/t^2 < \infty .$$

(b) Let $J_n = f\left(I_n\right)$ where $\int_{I_1}^\infty f^{-2}(t)dt < \infty$ then

$$\sum_2^\infty E\left(X_s^2 | F_{s-1}\right)/f^2\left(I_s\right) < \sum_2^\infty \int_{I_{s-1}}^{I_1} dt/f^2(t) < \int_{I_0}^\infty dt/f^2(t) < \infty .$$

In this case the second part of Lemma 3 appears in Stout (1974, p. 147).

THEOREM II (Feedback allowed). *If* A₂, B, E, F'' *hold then* $Q_n \to 0$ *a.s.*

Proof. Denote $\omega_s = x_s' V_s^{-1} x_s$, $\delta_s = u_s' u_s - \mathrm{tr}\left(\Sigma_s\right)$, $l_n = \sum_1^n \omega_s$. Sum II and divide by τ_n to obtain

$$\tau_n^{-1} W_n = \tau_n^{-1} \sum_1^n \omega_s \delta_s + \tau_n^{-1} \sum_1^n \omega_s \,\mathrm{tr}\left(\Sigma_s\right) - \tau_n^{-1} \sum_1^n a_s' a_s \left(1-\omega_s\right)$$
$$\left\{1-2 \sum_1^n u_s' u_s \left(1-\omega_s\right) / \left(\sum_1^n a_s' a_s \left(1-\omega_s\right)\right)\right\} .$$

Next consider that condition B ensures $l_n \to \infty$ since if $l_\infty = \sum_1^\infty \omega_s < \infty$ then $\sum_1^\infty x_s'x_s/\sigma_s < \infty$ and Kronecker's lemma would imply $\sigma_n^{-1}\sum_1^n x_s'x_s \to 0$.

Now take each term in turn.

According to Lemma 3 and conditions A2, B, the first term converges a.s. to zero if $\sum_1^\infty (\omega_s/\tau_s)^{1+\eta} < \infty$. This sum is

$$\sum_1^\infty (\omega_s/l_s)^{1+\eta}(l_s/\tau_s)^{1+\eta} < \sum_1^\infty \left[\omega_s/l_s^{1+\eta}\right](l_s/\tau_s)^{1+\eta}$$

since $\omega_s = x_s'V_s^{-1}x_s \le 1$. However $F'' \Rightarrow l_s/\tau_s \to 0$ while

$$\sum_p^\infty \omega_s/l_s^{1+\eta} < \sum_p^\infty \int_{l_{s-1}}^{l_s} dt/t^{1+\eta} < \int_{l_{p-1}}^\infty dt/t^{1+\eta} < \infty .$$

The second term, by A2, is bounded by const $\times\ l_n/\tau_n$ which converges to zero by F'' **and E.**

For the third term set $J_n = \sum_1^n a_s'a_s(1-\omega_s)$, $X_n = u_n'a_n(1-\omega_n)$ and apply $\widehat{\text{MGCT}}$ (Lemma 3'). Now

$$\sum_1^\infty E\left(X_n^2 \middle| \ _{n-1}\right)/J_n^2 \le t\sum_1^\infty a_n'a_n(1-\omega_n)^2/J_n^2$$

$$\le t\sum_1^\infty a_n'a_n(1-\omega_n)/J_n^2 \quad \text{since } \omega_n = x_n'V_n^{-1}x_n \le 1 .$$

This sum is

$$t\sum_1^\infty dJ_n/J_n^2 < t\sum_1^\infty dJ_n/J_nJ_{n-1} = -t\sum_1^\infty \left(J_n^{-1}-J_{n-1}^{-1}\right) < \infty .$$

Now if $J_n < \infty$ a.s. the bracketed factor is finite a.s. while the multiplier converges to zero. If $J_n \to \infty$ a.s. the bracketed factor (and so the whole term) is negative for n 'large enough' (of course n depends on the realisation). Thus W_n/τ_n is bounded by the first two terms for n

'large enough' and so converges to zero a.s.

THEOREM II' (No feedback). *If* A_2, B, E, F''' *hold then* $Q_n \to 0$ *a.s.*

Proof. From the proof of Theorem II we conclude via F''' that $W_n/\tau_n \to T$ a finite a.s. random variable. However $E(Q_n) \le c\tau_n^{-1}$ and Fatou's lemma implies $T = 0$ a.s.

THEOREM III (No feedback). *If* (A_1 *or* A'), B, D' *hold and the* x_s *are are nonstochastic then* $Q_n \to 0$ *a.s.*

Proof. Sum I, whereupon it is clear we have to show

$$\sum_p^\infty E\left(\alpha_n^2\right) < \infty \ , \quad \sum_p^\infty u_n'b_n < \infty \ , \quad \sum_p^\infty a_n'b_n < \infty \ ,$$

to conclude $Q_n \to T$. Then, since under A_1, $E(Q_n) \le t\tau_n^{-1}$ while under A', $E(Q_n) \le c\tau_n^{-1}$, Fatou's lemma gives

$$0 \le E(T) \le \underline{\lim}\, E(Q_n) \le \underline{\lim}\, \tau_n^{-1} \text{ const.} = 0 \ ,$$

so that $T = 0$ a.s.

(a) Under A_1,

$$E\left(\alpha_n^2\right) \le x_n'V_n^{-2}x_n\left(t+x_n'V_{n-1}^{-1}x_n\right)$$

$$\le x_n'V_n^{-1}V_{n-1}^{-1}x_n(t+1) \quad \text{by MIL } (4.3)^3 \text{ twice.}$$

Under A',

$$E\left(\alpha_n^2\right) \le \left(2\pi c+cx_n'V_{n-1}^{-1}x_n\right)x_n'V_n^{-2}x_n$$

$$\le (2\pi+1)cx_n'V_n^{-1}V_{n-1}^{-1}x_n \quad \text{by MIL } (4.3)^3 \text{ twice.}$$

Thus for A_1 or A' we have for some constant c' , $E\left(\alpha_n^2\right) \le c'x_n'V_n^{-1}V_{n-1}^{-1}x_n$.

(b) Next, $\sum_p^\infty a_n'b_n < \infty$ if $\sum_p^\infty |a_n'b_n| < \infty$ and this will follow, by Lemma 2, if $\sum_p^\infty E|a_n'b_n| < \infty$.

Let θ_n be a positive increasing sequence of constants to be chosen. Take expectations in the inequality

$$\begin{aligned} 2(a_n'b_n) &= 2\left(a_n/\sqrt{\sigma_{n-1}\theta_{n-1}}\right)'\left(b_n\sqrt{\sigma_{n-1}\theta_{n-1}}\right) \\ &\le b_n'b_n\sigma_{n-1}\theta_{n-1} + a_n'a_n/\left(\sigma_{n-1}\theta_{n-1}\right) \end{aligned}$$

to obtain

$$2E|a_n'b_n| \le E\left(a_n'a_n\right)/\left(\sigma_{n-1}\theta_{n-1}\right) + \sigma_{n-1}\theta_{n-1}E\left(b_n'b_n\right)$$

$$\le c'x_n'V_n^{-1}V_{n-1}^{-1}V_n^{-1}x_n\sigma_{n-1}\theta_{n-1}/\left(1-x_n'V_n^{-1}x_n\right) + x_n'V_n^{-1}x_n/\sigma_{n-1}\theta_{n-1}$$

by MIL $(4.4)^2$ (4.7)

$$\le c'\theta_{n-1}x_n'V_n^{-1}V_{n-1}^{-1}x_n + x_n'V_n^{-1}x_n/\sigma_{n-1}\theta_{n-1} \quad \text{by MIL } (4.4)^4 \qquad (4.8)$$

$$\le c'\theta_{n-1}x_n'V_n^{-1}V_{n-1}^{-1}x_n + 1/\left(\sigma_{n-1}\theta_{n-1}\right) \quad \text{by MIL } (4.3)^2. \qquad (4.9)$$

Now choose $\theta_n = \sqrt{n}$ and σ_n (that is τ_n) in D' to see

$$\sum_1^\infty \left(\theta_n-\theta_{n-1}\right)/\tau_n < \left(\overline{\lim}\, n^{\frac{1}{2}}\log^{1+\delta} n/\tau_n\right)\sum_1^\infty (\sqrt{n}-\sqrt{n-1})/\left(\sqrt{n}\,\log^{1+\delta} n\right) < \infty$$

so by Lemma 1 the first term in (4.9) has finite sum: also

$$\sum_1^\infty 1/\sigma_{n-1}\theta_{n-1} < \left(\overline{\lim}\, n^{\frac{1}{2}}\log^{1+\delta} n/\sigma_n\right)\sum_1^\infty 1/\left(n\,\log^{1+\delta} n\right) < \infty\,.$$

(c) Noting that, under A₁ or A', $E\left(u_n'u_n\right) \le$ const. we see by almost the same argument as in (b) that

$$\sum_p^\infty E|u_n'b_n| < \infty\,.$$

THEOREM IV (No feedback). *If* A₁, B, C" *hold and the* x_s *are non-stochastic then* $Q_n \to 0$ *a.s.*

Proof. Following the proof of Theorem III we have to show

(a) $\sum_p^\infty E\left(\alpha_n^2\right) < \infty$. This follows as in that proof.

(b) According to Lemma 3', under A₁, $\sum_p^\infty u_n'b_n$ converges a.s. if

$\sum_{p}^{\infty} b_n'b_n$ converges a.s. However $\sum_{p}^{\infty} E\left(b_n'b_n\right) < \infty$ by Lemma 1: thus by Lemma 2, $\sum_{p}^{\infty} b_n'b_n < \infty$ a.s.

(c) Finally, choosing θ_n to be the ν_n of Lemma 1, consider that expression (4.8) can be bounded by

$$c'\nu_{n-1}x_n'V_n^{-1}V_{n-1}^{-1}x_n + \nu_{n-1}x_n'V_n^{-1}V_{n-1}^{-1}x_n\lambda_{n-1}/\left(\sigma_{n-1}\nu_{n-1}^2\right) .$$

These series have finite sums by condition C" and Lemma 1.

ADDITION. After this chapter was completed the author became aware of the work of Moore (1978) in which the inequality (II)' is established. Moore's main theorem, (Theorem 4.3) then states that a.s. convergence occurs if for some $\alpha > 0$,

$$n^{\alpha}V_n^{-1} \to 0 \quad \text{a.s.},$$

$$\sum_{p}^{\infty} n^{-\alpha}x_n'V_n^{-1}x_n < M < \infty \quad \text{a.s.}$$

for some constant M . However recalling $\tau_n^{-1} = \mathrm{tr}\left(V_n^{-1}\right)$ these conditions clearly imply

$$\sum_{p}^{\infty} x_n'V_n^{-1}x_n/\tau_n < \infty \quad \text{and} \quad \tau_n \to \infty$$

and so this result follows from the proof of Theorem I.

The assumption of a constant bound enables a proof of MGCT* without recourse to the result in Neveu. However it will be seen in subsequent use of MGCT* that the stronger result is vital.

CHAPTER 5

STOCHASTIC APPROXIMATION AND THE FINAL VALUE THEOREM

5.1 Introduction

The aim of this chapter is to investigate the well known limit laws for both the Robbins-Monro (RM) and Kiefer-Wolfowitz (KW) Stochastic Approximation (SA) schemes namely Central Limit Theorem (CLT), Sacks (1958), Law of the Iterated Logarithm (LIL), Heyde (1974) and Invariance Principle (IP), Kersting (1971), McLeish (1976), Kushner (1978). The idea here is to subtract out the part of the SA that obeys the limit law and use a time-varying version of the Final Value Theorem (FVT) of Z-transform theory to show the remainder is asymptotically negligible. As a result of this approach it is straightforward to state results for general a_n, c_n sequences. Further, the conditions required for the validity of the theorems, are not obscured by the summing procedure so that their nature is clearly revealed. Thus the CLT/IP are given without a.s. convergence of the SA being required (though strong assumptions on the regression functions are needed): the usual conditions $\sum_1^\infty a_n^2 < \infty$, $\sum_1^\infty a_n^2/c_n^2 < \infty$ are not required. They are needed of course for the LIL.

The chapter is in two parts. The first gives a CLT/IP/LIL for the solution to a sotchastic difference equation; the second applies these to the SA schemes.

5.2 Limit Behaviour of Stochastic Difference Equations

This section gives some lemmas concerning the asymptotic behaviour of (stochastic) difference equations.

LEMMA 1 (Final Value Theorem (FVT)). *Let* a_n *be a sequence of positive constants with* $a_n \to 0$ *and* $\sum_1^\infty a_n = \infty$. *Let* $\alpha_1 > 0$ *and*

$$\xi_{n+1} = \left(1-\alpha_1 a_n\right)\xi_n + b_n .$$

Then provided $b_0 < \infty$,

$$\lim_{n\to\infty} \xi_n = \alpha_1^{-1} \lim_{n\to\infty} b_n/a_n$$

if the right hand limit exists.

Proof. Denote $\pi_{n+1} = \prod_1^n \left(1-\alpha_1 a_s\right)$.

The conditions on a_n ensure $\pi_n^{-1} \uparrow +\infty$ (Bromwich 1947, p. 1042). Divide the ξ_n equation through by π_n and sum up

$$\xi_{n+1} = \pi_{n+1} \sum_1^n b_s/\pi_{s+1} + b_0\pi_{n+1} . \tag{5.1}$$

Next, by the discrete L'Hospital rule (a special case of the Toeplitz lemma - Bromwich (1947, pp. 414-415))

$$\lim \xi_{n+1} = \lim \frac{b_n/\pi_{n+1}}{\pi_{n+1}^{-1}-\pi_n^{-1}} = \lim \frac{b_n/\pi_{n+1}}{\alpha_1 a_n/\pi_{n+1}}$$

$$= \alpha_1^{-1} \lim b_n/a_n$$

if this limit exists.

Remarks. (1) This lemma is the time varying form of the final value theorem of Z-transform theory.

(2) If b_n is a random sequence and $E|b_0| < \infty$, then

$$\lim E|\xi_n| \le \alpha_1^{-1} \lim E|b_n|/a_n$$

if the right hand limit exists.

(3) If ν_n is a sequence of nondecreasing positive "norming" constants and calling $d\nu_n = \nu_n - \nu_{n-1}$ (this 'differential' notation will be used throughout) then provided $b_0 < \infty$ and

$$\lambda = \lim d\nu_{n+1}/\left(a_n\nu_n\right) = \lim \left(\frac{\nu_{n+1}}{\nu_n} - 1\right) \frac{1}{a_n} < \alpha_1$$

we have $\lim \nu_n\xi_n = \left(\alpha_1-\lambda\right)^{-1} \lim \nu_{n+1} b_n/a_n$.

Proof.

$$\nu_{n+1}\xi_{n+1} = \nu_n\xi_n\left\{1 - a_n\left(\alpha_1 + \alpha_1 \frac{d\nu_{n+1}}{\nu_n} - \frac{d\nu_{n+1}}{\nu_n a_n}\right)\right\} + \nu_{n+1}b_n$$

and the result follows by the same proof as in Lemma 1 since by the above condition $\lim d\nu_{n+1}/\nu_n = 0$.

(4) If $(1-\alpha_1 a_n)$ is replaced by $(1-\beta a_n+\varphi_n^2)$, Lemma 1 still holds provided $\lim \varphi_n^2/a_n = 0$.

(5) From expression (5.1) we can derive a weaker result than the FVT. If, with the conditions of Lemma 1, $\sum_1^\infty b_s < \infty$, then $\xi_n \to 0$. This follows from (5.1) by Kronecker's lemma. This result has been given by Venter (1967). In fact by the FVT we need only $b_n \to 0$ "faster" than $a_n \to 0$ to conclude $\xi_n \to 0$. The point is that the discrete L'Hospital rule is a special case of the Toeplitz lemma while the Kronecker lemma is a consequence of it.

(6) It should be remarked that the discrete L'Hospital rule holds also for probability limits. Thus if $b_n \uparrow \infty$ a.s. and

$dS_n/d\,b_n \xrightarrow{p} c$ a finite a.s. random variable and

$\sup_n E|dS_n/d\,b_n|^{1+\eta} < L < \infty$ for some $\eta > 0$ or

$\sup_n |dS_n/d\,b_n| < L < \infty$ a.s. L a constant,

then $S_n/b_n \xrightarrow{p} c$.

Proof. Take $c = 0$ w.l.o.g. by setting $S_n' = S_n - cb_n$ and dropping the dash. Now

$$S_n/b_n = \sum_1^n dS_k/\sum_1^n db_k$$
$$= \sum_1^n (dS_k/db_k)db_k I(|dS_k/db_k| < \varepsilon)/\sum_1^n db_k$$
$$+ \sum_1^n (dS_k/db_k)db_k I(|dS_k/db_k| > \varepsilon)/\sum_1^n db_k .$$

Now the modulus of the first term is bounded by ε which is arbitrarily small. Now by Hölder's inequality

$$E(|dS_k/db_k|I(|dS_k/db_k| > \varepsilon))$$
$$\leq \left(E|dS_k/db_k|^{1+\eta}\right)^{1/(1+\eta)} \left(E(I(|dS_k/db_k| > \varepsilon))\right)^{\eta/(1+\eta)}$$
$$\leq L^{1/1+\eta}P(|dS_k/db_k| > \varepsilon)$$

in view of the first condition above.

Thus the second term has its expectation bounded by

$$L^{1/(1+\eta)} \sum_1^n db_k P(|dS_k/db_k| > \varepsilon)/\sum_1^n db_k$$

which by the discrete L'Hospital rule and assumption has limit zero. A similar argument gives the result under the second condition. In the sequel it will be convenient to denote

$$dS_n/db_n \xrightarrow{p} c \quad \text{by} \quad p \lim dS_n/db_n = c .$$

LEMMA 2 (CLT for the solution to a stochastic difference equation). *Suppose* Z_n *is a martingale difference sequence w.r.t. some increasing sequence of σ-fields* F_n *and let*

$$U_{n+1} = (1-\alpha_1 a_n)U_n + (a_n/c_n)Z_n , \quad U_0 = 0 \quad a.s., \tag{5.2}$$

where α_1 *is a positive constant and where* a_n, c_n *are sequences of positive constants with*

(A) $\sum_1^\infty a_n = \infty$, $a_n \to 0$ *and either* $c_n = 1$ $\forall n$ *or else* $c_n \to 0$ *in which case* $a_n/c_n^2 \to 0$. *Suppose also*

(B) *There is a random variable* Z *with* $EZ^2 < \infty$ *and for some* $c < \infty$, $\forall n$, $\Pr\left(|Z_n| \ge x \mid F_n\right) \le c \Pr(|Z| \ge x)$ $\forall x$. *Notice that* $\sup_n E\left(Z_n^2 | F_n\right) < \infty$.

(C) $p \lim E\left(Z_n^2 | F_n\right) = \sigma^2 = \lim E\left(Z_n^2\right)$.

(D) $\lim \frac{1}{a_n}\left(\frac{c_{n+1}^2}{c_n^2}\frac{a_n}{a_{n+1}} - 1\right) = \lambda$ $\left(0 \le \lambda < 2\alpha_1\right)$; $c_n/c_{n+1} \to 1$, $a_n/a_{n+1} \to 1$.

Then $c_n U_n/\sqrt{a_n} \xrightarrow{\mathcal{D}} N\left(0, \sigma^2/K\right)$ where $K = 2\alpha_1 - \lambda$.

Remark. Condition D, with $c_n = 1$, ensures $\underline{\lim}\, na_n > 0$.

Proof. As in Lemma 1, $\pi_{n+1} = \prod_1^n \left(1-\alpha_1 a_s\right) \to 0$. Call $v_n^2 = E\left(U_n^2\right)$ so that calling $\sigma_n^2 = E\left(Z_n^2\right)$,

$$v_{n+1}^2 = \left(1-\alpha_1 a_n\right)^2 v_n^2 + a_n^2 c_n^{-2} \sigma_n^2 . \tag{5.3}$$

Then by the FVT (3) and (A), (D),

$$\begin{aligned}\lim c_n^2 v_n^2/a_n &= K^{-1} \lim c_{n+1}^2 a_{n+1}^{-2} a_n^2 c_n^{-2} \sigma_n^2 \\ &= \sigma^2/K .\end{aligned} \tag{5.4a}$$

Thus we may norm by v_n instead of $\sqrt{a_n}/c_n$.

Next call $S_n = U_n/\pi_n$ so

$$S_{n+1} - S_n = a_n c_n^{-1} Z_n/\pi_{n+1} = u_{n+1}$$

say, and S_{n+1} is a MG w.r.t. F_n. Also

$$s_n^2 = \operatorname{var}\left(S_n\right) = v_n^2/\pi_n^2 \tag{5.4b}$$

so

$$S_n/s_n = U_n/v_n ; \tag{5.5}$$

that is, a CLT for S_n/s_n is one for $U_n c_n/\sqrt{a_n}$.

To derive the CLT we must show (Scott, 1973)

(1) $s_{n+1}^{-2} \sum_1^n E\left(u_{k+1}^2 | F_k\right) \xrightarrow{p} 1$,

(2) $s_{n+1}^{-2} \sum_1^n E\left(u_{k+1}^2 I\left(|u_{k+1}| \geq \varepsilon s_{n+1}\right) \mid F_k\right) \xrightarrow{p} 0$.

Note that

$$\begin{aligned} u_{n+1}^2/s_{n+1}^2 &= Z_n^2 a_n^2 / \left(v_{n+1}^2 c_n^2\right) \\ &= \left(K/\sigma^2\right) a_n Z_n^2 \left(o_n(1)+1\right) . \end{aligned} \tag{5.6}$$

Now apply the discrete L'Hospital rule for p lim's (FVT (6)) and (B), (C) to (1) to see

$$\begin{aligned} p \lim s_{n+1}^{-2} \sum_1^n E\left(u_{k+1}^2 | F_k\right) &= p \lim E\left(u_{n+1}^2 | F_n\right) / E\left(u_{n+1}^2\right) \\ &= p \lim E\left(Z_n^2 | F_n\right) / E\left(Z_n^2\right) = 1 \text{ by (C).} \end{aligned}$$

Next, (2) is implied by

$$s_{n+1}^{-2} \sum_1^n E\left(u_{k+1}^2 I\left(|u_{k+1}| \geq \varepsilon s_{k+1}\right) \mid F_k\right) \xrightarrow{p} 0 .$$

Once more the 'p lim' discrete L'Hospital rule shows this to follow if

$$p \lim E\left(u_{k+1}^2 I\left(|u_{k+1}| \geq \varepsilon s_{k+1}\right) \mid F_k\right) / E\left(u_{k+1}^2\right) = 0 .$$

From equation (5.6) for k large enough

$$I\left(|u_{k+1}| \geq \varepsilon s_{k+1}\right) \leq I\left(Z_k^2 > \tfrac{1}{2}\varepsilon^2 / a_k\right) .$$

Thus we need only investigate (dropping the $\frac{1}{2}$) the limit of

$$E\left(Z_k^2 I\left(Z_k^2 \geq \varepsilon^2/a_k\right) \mid F_k\right)$$

$$= \varepsilon^2/a_k \Pr\left(Z_k^2 \geq \varepsilon^2/a_k \mid F_k\right) + \int_{\varepsilon^2/a_k}^{\infty} \Pr\left(Z_k^2 \geq y \mid F_k\right) dy$$

$$\leq c\varepsilon^2/a_k \Pr\left(Z^2 \geq \varepsilon^2/a_k\right) + c\int_{\varepsilon^2/a_k^2}^{\infty} \Pr\left(Z^2 \geq y\right) dy, \text{ by (B)}$$

$$= cE\left(Z^2 I\left(Z^2 \geq \varepsilon/a_k\right)\right)$$

which has a.s. limit zero since $E\left(Z^2\right) < \infty$.

COROLLARY. *Condition* (B) *may be replaced by the stronger* (B'),

(B') $\sup_n E\left(|Z_n|^{2+\eta} \mid F_n\right) < \infty$.

Proof.

$$E\left(I\left(Z_k^2 \geq \varepsilon^2/a_k\right) \mid F_k\right) \leq \left(a_k/\varepsilon^2\right)^{\eta/2} E\left(Z_k^{2+\eta} I\left(Z_k^2 \geq \varepsilon^2/a_k\right) \mid F_k\right)$$

$$\leq \left(a_k/\varepsilon^2\right)^{\eta/2} \sup_n E\left(|Z_n|^{2+\eta} \mid F_n\right) \to 0 .$$

LEMMA 3 (LIL for a stochastic difference equation). *In the statement of Lemma* 2 *remove conditions* A, B *and add conditions*

(A') $\sum_1^\infty a_n = \infty$; $c_n = \forall n$ *or* $c_n \to 0$; $\lim a_n/c_n^2 = 0$, $\sum_1^\infty a_n^2 < \infty$.

(E) $\sum_1^\infty a_n^2 E\left(Z_n^4 I\left(|Z_n| < \delta/\sqrt{a_n}\right)\right) < \infty$ *for some* $\delta > 0$.

(F) $\sum_1^\infty a_n E\left(|Z_n| I\left(|Z_n^2| \geq \varepsilon/\sqrt{a_n}\right)\right) < \infty$ $\forall\varepsilon > 0$.

Then

$$\overline{\lim}\, c_n U_n/\left(a_n\psi_n\right)^{\frac{1}{2}} = +\sigma/\sqrt{K} \quad a.s.,$$

$$\underline{\lim}\, c_n U_n/\left(a_n\psi_n\right)^{\frac{1}{2}} = -\sigma/\sqrt{K} \quad a.s.,$$

where $\psi_n = 2 \log\log \bar{s}_n^{2}$ *and* $s_n^2 = \left(\sigma^2/K\right) a_n c_n^{-2} e^{2\alpha_1 \tau_n - 2\nu}$ *and* $\tau_n = \sum_1^n a_s$

with $\nu = \lim\left(\log \pi_n + \alpha_1\tau_n\right)$ *or* $\lim \pi_n e^{\alpha_1\tau_n} = e^{\nu}$ *and* $\pi_n = \prod_1^n \left(1-\alpha_1 a_s\right)$.

Proof. For either $c_n \to 0$ or $c_n = 1$ we have from the assumptions, that $\sum_1^\infty a_n^2 < \infty$. So take w.l.o.g. $\alpha_1 a_n < \frac{1}{2}$. Now it is well known (Bromwich, 1947, p. 181) that for $|x| < \frac{1}{2}$,

$$\left|\log(1+x)-x\right| < x^2 .$$

Thus $\sum_1^\infty \left(\log\left(1-\alpha_1 a_k\right)+\alpha_1 a_k\right)$ is absolutely convergent and so convergent: thus ν exists. Now in view of (5.4) it follows that $\bar{s}_n^{2}$ can be replaced by

$$s_n^2 = \left(\sigma^2/K\right) a_n c_n^{-2} \pi_n^{-2}\left(1+o_n(1)\right) .$$

As mentioned in the proof of Lemma 2 we need only investigate a LIL for S_n/s_n . For this we need (Heyde and Scott, 1973)

(1) $\sum_1^\infty s_n^{-4} E\left[u_n^4 I\left(|u_n| < \delta s_n\right)\right] < \infty$ for some $\delta > 0$.

(2) $\sum_1^\infty s_n^{-1} E\left(|u_n| I\left(|u_n| \geq \varepsilon s_n\right)\right) < \infty \quad \forall\varepsilon > 0$.

(3) $s_n^{-2} \sum_1^n u_k^2 \to 1$ a.s.

First these conditions are replaced. Indeed (2) is implied by

(2') $\sum_1^\infty s_n^{-2} E\left[u_n^2 I\left(|u_n| \geq \varepsilon s_n\right)\right] < \infty \; \forall\varepsilon > 0$.

This follows since (2') exceeds ε times (2) (recall an indicator function equals its square). Next it is shown that (1), (2'), (3) may be replaced by (1), (2'), (3'),

(3') $s_n^{-2} \sum_1^n E\left[u_k^2 | F_{k-1}\right] \to 1$ a.s.

Clearly (3) is impled by (3') together with $s_n^{-2} \sum_1^n \left[u_k^2 - E\left[u_k^2 | F_{k-1}\right]\right] \to 0$ a.s.

Denote

$$u_{1k} = u_k I\left(|u_k| \le \delta s_k\right) ,$$

$$u_{2k} = u_k I\left(|u_k| > \delta s_k\right) .$$

Then it is required to show

$$s_n^{-2} \sum_1^n \left(u_k^2 - u_{1k}^2\right) \to 0 \quad \text{a.s.},$$

$$s_n^{-2} \sum_1^n \left(u_{1k}^2 - E\left(u_{1k}^2 | F_{k-1}\right)\right) \to 0 \quad \text{a.s.},$$

$$s_n^{-2} \sum_1^n E\left(u_{2k}^2 | F_{k-1}\right) \to 0 \quad \text{a.s.}$$

Now the first limit holds if $P\left(u_k \ne u_{1k} \quad \text{i.o.}\right) = 0$. But by the Borel-Cantelli lemma this is equivalent to

$$\sum_1^\infty P\left(u_k \ne u_{1k}\right) = \sum_1^\infty P\left(|u_k| > \delta s_k\right) < \infty .$$

However consider that

$$\delta^2 \sum_1^\infty P\left(|u_n| > \delta s_n\right) \le \delta^2 \sum_1^\infty P\left[u_n^2 \ge \delta^2 s_n^2\right] + \sum_1^\infty s_n^{-2} \int_{\delta^2 s_n^2}^\infty P\left[u_n^2 \ge y\right] dy$$

$$= \sum_1^\infty s_n^{-2} E\left[u_n^2 I\left(|u_n| \ge \delta s_n\right)\right] < \infty \quad \text{by } (2')$$

The second limit holds by Kolmogorov's strong law for martingales if

$$\sum_1^\infty s_n^{-4} E\left(u_{1n}^4\right) < \infty$$

which is just (1). The third limit holds by Kronecker's lemma if

$$\sum_1^\infty s_n^{-2} E\left(u_{2n}^2 | F_{n-1}\right) < \infty$$

which holds, by the lemma in Lukacs (1975, p. 80) if

$$\sum_1^\infty s_n^{-2} E\left(u_{2n}^2\right) < \infty$$

which is (2').

To complete the proof recall (5.6) to see (1), (2') $\Leftrightarrow$ (E), (F); while, via (5.6) and the discrete L'Hospital rule, (3') is implied by (B).

COROLLARY. *In Lemma* 3, (E) *and* (F) *are implied by* (B)', (G).

(B)' $\sup_n E\left(|Z_n|^{2+\eta} \mid F_n\right) < \infty$ *for some* $\eta > 0$.

(G) $\sum_1^\infty a_n^{1+\eta/2} < \infty$.

Proof. $E\left(Z_n^4 I(|Z_n| < \delta/\sqrt{a_n})\right) \le \text{const.}\ E\left(Z_n^{2+\eta}\right) a_n^{-1+\eta/2}$. Thus (E) is implied by (B), (G). For (F),

$$E\left(Z_n^4 I(|Z_n| > \varepsilon/\sqrt{a_n})\right) \le E\left(Z_n^{2+\eta}\right) a_n^{\eta/2}$$

so that (F) follows from (B'), (G).

The next step will be to state an IP. Suppose we multiply throughout (5.2) by $c_{n+1}/\sqrt{a_{n+1}}$ and denote $Q_{n+1} = \left(c_{n+1}/\sqrt{a_{n+1}}\right) U_{n+1}$, we have after some manipulation and utilising condition D,

$$dQ_{n+1} \simeq \tfrac{1}{2} Q_n d\tau_n \left(\lambda - 2\alpha_1\right) + dW_n \tag{5.7}$$

where $W_n = \sum_1^n \sqrt{a_s}\, w_s Z_s$ with $w_n = c_{n+1} c_n^{-1} \sqrt{a_n a_{n+1}^{-1}}$ and $dW_n = W_n - W_{n-1}$ and so on. Notice that $w_n \to 1$ (condition D) so $E\left(W_n^2\right) \simeq \sigma^2 \tau_n$, so W_n "looks like" a Wiener process on a time scale τ_n : Q_n begins to resemble the output of a stochastic differential equation. It is these heuristic considerations that will be formalised in the IP.

There are two ways to formulate an IP for Q_n (according to choices of arrays of martingale difference sequences): one is to derive the IP from an IP for S_n/s_n (recall (5.5)) - this is the type of result stated by McLeish (1976) and Kersting (1971). The other approach follows from an IP for W_n and has been given by Kushner (1978) and Ibragimov and Khasminskii (1973) for a maximum likelihood estimator. It is the second form of the IP that will be stated here. Briefly, the author feels it to be a more natural form of IP given the difference equation nature of the SA schemes. First we give an IP for W_n and then via summation by parts and the continuous mapping theorem (Billingsley, 1968, Theorem 5.1) we obtain the IP for Q_n. The other type of IP could just as well be derived however, and must in any case be obtainable by a change of time scale.

Let $W(t)$ denote a Wiener measure of variance σ^2 on $D[0, 1]$ the space of all functions on $[0, 1]$ that are right continuous and have left hand limits (Billingsley, 1968, Chapter 3). Define

$$k_n(t) = \sup\{m : \tau_{n+m} - \tau_n \le t\} .$$

LEMMA 4. *Under the conditions of Lemma* 2,

$$W_{n+k_n(t)} - W_n \Rightarrow W(t) .$$

Proof. Consider the triangular array $v_{ni} = \sqrt{a_{n+i}}\, Z_{n+i} w_{n+i}$ and array of increasing σ-fields $F_{n,i} = F_{n+i+1}$ so that

$$W_{n+k_n(t)} - W_n = \sum_{i \le k_n(t)} v_{ni}$$

and

$$E(v_{ni} | F_{n,i-1}) = 0 .$$

Now condition C ensures $w_n \to 1$ so this multiplier will be neglected. To conclude weak convergence we have only to check the conditions of McLeish (1974, Corollary 3.8) (they are stated for $D[0, \infty)$ but apply to $D[0, 1]$)

$$\sum_{i \le k_n(t)} E\left(v_{ni}^2 I(|v_{ni}| > \varepsilon \mid F_{n,i-1})\right) \xrightarrow{p} 0 \quad \forall \varepsilon > 0 ,$$

$$\sum_{i \le k_n(t)} E\left(v_{ni}^2 | F_{n,i-1}\right) \xrightarrow{p} \sigma^2 t \quad \forall t < 1 .$$

Now the summand in the first condition is

$$a_{n+i} E\left(Z_{n+i}^2 I\left(Z_{n+i}^2 > \varepsilon / a_{n+i}\right) | F_{n,i-1}\right)$$

and under condition B this is bounded by

$$a_{n+i} E\left(Z^2 I\left(Z^2 > \varepsilon / a_{n+i}\right)\right) .$$

Now since $\sum_{i \le k_n(t)} a_{n+i} = \tau_{n+k_n(t)} - \tau_n \to t$ (since $a_n \to 0$) the result follows because $E(Z^2) < \infty$. The summand in the second condition is

$$\sum_{i \le k_n(t)} a_{n+i} E\left(Z_{n+i}^2 \mid F_{n+i}\right)$$

and $\sum_{i \le k_n(t)} a_{n+i} \to t$, together with conditions A and C ensure the result.

Remark. The result clearly also holds with condition B replaced by B'. We investigate Q_n by reorganising (5.7) as (or multiplying through (5.2) by $c_{n+1}/\sqrt{a_{n+1}}$)

$$Q_{n+1} = Q_n\left(1-\tfrac{1}{2}K_n a_n\right) + dW_n$$

where

$$K_n = 2\alpha_1 w_n - a_n^{-1}\left(w_n^2-1\right) - \tfrac{1}{2}a_n^{-1}\left(w_n-1\right)^2$$

with

$$w_n = c_{n+1}c_n^{-1}\sqrt{a_n a_{n+1}^{-1}} \,.$$

Notice that, by condition D of Lemma 2,

$$a_n^{-1}\left(w_n^2-1\right) \to \lambda$$

and $w_n \to 1$ so that $w_n + 1 \to 2$; thus

$$\tfrac{1}{2}a_n^{-1}\left(w_n-1\right)^2 = \tfrac{1}{2}\left(w_n+1\right)^{-1}a_n^{-1}\left(w_n^2-1\right)\left(w_n-1\right) \to 0$$

so that finally, $K_n \to K = 2\alpha_1 - \lambda$.

Next, denoting $P_n = \prod_1^n \left(1-\tfrac{1}{2}K_s a_s\right)$ sum up to obtain

$$Q_{N+n} = P_{N+n}P_N^{-1}Q_N + P_{N+n}\sum_1^n P_{N+s}^{-1}dW_{N+s} \,.$$

Now sum by parts noting that $P_n^{-1} - P_{n-1}^{-1} = -\tfrac{1}{2}K_n a_n P_n^{-1}$,

$$Q_{N+n} = P_{N+n}P_N^{-1}Q_N + W_{N+n} + \tfrac{1}{2}P_{N+n}\sum_1^n P_{N+s}^{-1}W_{N+s-1}K_{N+s}d\tau_{N+s} - P_{N+n}P_N^{-1}W_N .$$

Noting that

$$P_{N+n}^{-1} - P_N^{-1} = -\tfrac{1}{2}\sum_1^n K_{N+s}a_{N+s}P_{N+s}^{-1}$$

we have

$$Q_{N+n} = P_{N+n}P_N^{-1}Q_N + W_{N+n} - W_N + \tfrac{1}{2}P_{N+n}\sum_1^n K_{N+s}P_{N+s}^{-1}\left(W_{N+s-1}-W_N\right)d\tau_{N+s} .$$

Denote now

$$C_N(t) = P_{N+k_N(t)}P_N^{-1}$$

and

$$W_N(t) = W_{N+k_N(t)} - W_N$$

and rewrite this as (neglecting terms such as $W_N - W_{N-1} = \sqrt{a_N}\, Z_N w_N$ that vanish in probability in the 'sup' norm)

$$Q_N(t) = C_N(t)Q_N(0) + W_N(t) + \tfrac{1}{2}\int_0^t K_{N+k_N(s)}C_N(t)C_N^{-1}(s)W_N(s)ds . \tag{5.8}$$

We now show

$$\lim_{N\to\infty}\ \sup_{0\le t\le 1}\left|C_N(t)-e^{-\frac{1}{2}Kt}\right| \to 0 . \tag{5.9}$$

Since

$$\sup_{0\le t\le 1} e^{\frac{1}{2}Kt} < \infty$$

and

$$e^{\frac{1}{2}Kt}\exp\left(-\tfrac{1}{2}\sum_N^{N+k_N(t)} a_sK_s\right) \to 1 \quad \forall t .$$

(This follows since $\sum_N^{N+k_N(t)} a_s \to t$ (because $a_n \to 0$ and $K_n \to K$ ensure

$$\sum_{N}^{N+k_N(t)} a_s K_s \Big/ \sum_{N}^{N+k_N(t)} a_s \to K \Big)$$

it will do to show

$$C_N(t)\exp\left[\tfrac{1}{2}\sum_{N}^{N+k_N(t)} a_s K_s\right] \to 1 \quad \text{uniformly in } t \; .$$

Now $a_n \to 0$ so w.l.o.g. take $|a_n| < \frac{1}{2} \;\; \forall n$ and consider that this last expression is

$$\exp\left[\sum_{N}^{N+k_N(t)} \log\left(1-\tfrac{1}{2}K_s a_s\right)+\tfrac{1}{2}K_s a_s\right]$$

which is bounded in modulus by

$$\exp\left[\sum_{N}^{N+k_N(t)} a_s^2 K_s^2/4\right] .$$

Now since $\sum_{N}^{N+k_N(t)} a_s \to t$ and $a_n \to 0$ $\left(\text{so that} \max_{N\le k\le N+k_N(t)} a_k \to 0 \;\; \forall t\right)$ it follows that $\sum_{N}^{N+k_N(t)} a_s^2 \le \sum_{N}^{N+k_N(1)} a_s^2 \to 0$ and the result follows since $K_s \to K$. We have thus established (5.9).

Now by the continuous mapping theorem and Lemma 4 it follows for example that

$$\int_0^t e^{-\frac{1}{2}K(t-s)} W_N(s)ds \Rightarrow \int_0^t e^{-\frac{1}{2}K(t-s)} W(s)ds \; .$$

Also

$$Q_N(0) \Rightarrow Q(0) \sim N\left(0,\, \sigma^2/K\right) .$$

This is just the CLT, Lemma 2. It should be remarked that Kushner's proof is incomplete in that a CLT for $Q_N(0)$ is not proved.

Thus, from (5.9), we now obtain by simple argument

LEMMA 5. *Under the conditions of Lemma 2,*

$$Q_{N+k_N}(t) \Rightarrow Q(t)$$

where $Q(t)$ *is the continuous Gaussian process on* $D[0, 1]$ *with*

$$E\big(Q(t)\big) = 0 ,$$

$$E\big(Q(t)Q(s)\big) = K^{-1}\sigma^2 e^{-\frac{1}{2}K(t-s)} .$$

Proof. We need only establish the existence of the process. From

$$Q(t) = e^{-\frac{1}{2}Kt}Q(0) + W(t) + \tfrac{1}{2}K \int_0^t e^{-\frac{1}{2}K(t-s)}W(s)ds \tag{5.10}$$

we have

$$Q(t) = e^{-\frac{1}{2}Kt}Q(0) + X(t)$$

where

$$X(t) = e^{-\frac{1}{2}Kt} \int_0^t e^{\frac{1}{2}Ks}dW(s)$$

or else

$$E\big(X(t)\big) = 0 ,$$

$$E\big(X(t)X(s)\big) = e^{-\frac{1}{2}Kt}\big(e^{\frac{1}{2}Ks}-e^{-\frac{1}{2}Ks}\big)\sigma^2/K$$

and

$$P\big(X(0) = 0\big) = 0 .$$

However Billingsley (1968, Chapter 4) has demonstrated the existence of $X(t)$ and in view of Lemma 2 our result follows.

Before turning to a discussion of Stochastic Approximation one more lemma is needed; a version of the Martingale Convergence Theorem (MGCT), a slight extension of the one used in Chapter 4.

MGCT*. *If* F_n *is an increasing sequence of* σ*-algebras and* T_n, α_n, β_n, γ_n *are non-negative* F_n *measurable random variates with*

$$E\big(T_{n+1}\,|\,F_n\big) \le T_n\big(1+\beta_n\big) - \alpha_n + \gamma_n .$$

Then

$$\sum_1^\infty \beta_n < \infty \ , \quad \sum_1^\infty \gamma_n < \infty \quad a.s.$$

ensures $T_n \to T$ *a.s. a finite random variable and* $\sum_1^\infty \alpha_n < \infty$ *a.s.*

Proof. The conditions on β_n ensure $P_n = \prod_1^n \left(1+\beta_s\right)$ has a finite positive limit a.s. Thus we need only consider

$$E\left(T'_{n+1} \middle| F_n\right) \le T'_n - \alpha'_n + \gamma'_n$$

where dashed quantities were found by division by P_{n+1} .

Now

$$E\left(T'_{n+1} \middle| F_n\right) \le T'_n + \gamma'_n$$

and according to an exercise in Neveu (1975), p. 34), $\sum_1^\infty \gamma'_n < \infty$ a.s. ensures $W'_n \to W'$ a.s. Now call $a_n = \sum_1^n \alpha'_s$ so that

$$E\left(T'_{n+1} + a_{n+1} \middle| F_n\right) \le T'_n + a_n + \gamma'_n \ .$$

The same result ensures $W'_n + a_n \to W''$ so the result $\sum_1^\infty \alpha_n < \infty$ a.s. follows.

Remark. It is well known that $\alpha_n \ge 0$ a.s., $\sum_1^\infty E\left(\alpha_n\right) < \infty \Rightarrow \sum_1^\infty \alpha_n < \infty$ a.s. (Lukacs, 1975, p. 80).

5.3 Limit Theorems for Stochastic Approximation

Recall the R-M procedure is a recursive method of finding the solution θ of an equation $M(x) = \alpha$ when $M(x)$ is unknown. For each trial value x there is, however, a noisy measurement $Y(x)$ with $E\left(Y(x)\right) = M(x)$. The solution is found by a recursion of the form

$$X_{n+1} = X_n - a_n\left(Y_n - \alpha\right) \ , \quad Y_n = Y\left(X_n\right) \ ,$$

where a_n is a sequence of positive constants with at least $\sum_1^\infty a_n = \infty$, $a_n \to 0$. It is usually assumed also that the conditional distribution of $Y(X_s)$ given $X_1 = x_1 \ldots X_s = x_s$ is the same as the distribution of $Y(x_s)$. Rewrite the above as

$$X_{n+1} = X_n - a_n M_n - a_n Z_n$$

where $M_n = M(X_n) - \alpha$ and $Z_n = Z(X_n) = Y(X_n) - M(X_n)$ is a martingale difference sequence w.r.t. σ-algebras F_n generated by the X_n process. The following assumptions will be of interest (*cf.* Sacks, 1958).

(A1) M is a Borel-measurable function; $M(\theta) = \alpha$ and $(x-\theta)\big(M(x)-\alpha\big) > 0$, $\forall x \neq 0$. Also $|M(x)| \leq a + b|x|$, $\forall x$ and constants a, b.

(A2) For some positive constants K_1, K_2, $\forall x$,

$$K_1|x-\theta| \leq |M(x)-\alpha| \leq K_2|x-\theta| .$$

$(\overline{A3})$, (A3)* $M(x) = \alpha + \alpha_1(x-\theta) + \delta(x, \theta)$ with $\alpha_1 > 0$

for (A3)* $\delta(x, \theta) = o(|x-\theta|)$ as $x - \theta \to 0$

for $(\overline{A3})$ $\delta(x, \theta) = \alpha_2|x-\theta|^{1+\beta} + o\big(|x-\theta|^{1+\beta}\big)$

as $x - \theta \to 0$, for some finite constant α_2 and some $\beta > 0$.

(A4) (a) $\sup_x E|Z(x)|^{2+\eta} < \infty$ for some $\eta > 0$,

(b) $\lim_{x\to\theta} E\big(Z^2(x)\big) = \sigma^2$.

Next consider the KW scheme. The aim here is to find the unique maximum at $x = \theta$ of $M(x)$. There is a noisy observation $Y(x)$ with

$$E\big(Y(x)\big) = M(x) .$$

Generate a sequence of approximations by

$$X_{n+1} = X_n - a_n c_n^{-1}\big(Y_n(X_n-c_n)-Y_n(X_n+c_n)\big)$$

where a_n, c_n are sequences of positive numbers with at least, $\sum_1^\infty a_n = \infty$,

$c_n \to 0$, $a_n/c_n^2 \to 0$. As before the conditional distribution of $Y_n(X_n \pm c_n)$ given $X_1 = x_1 \ldots X_n = x_n$ is the same as the distribution of $Y(X_n \pm c_n)$. Rewrite the scheme as

$$X_{n+1} = X_n - a_n c_n^{-1} M_n - a_n c_n^{-1} Z_n$$

where

$$M_n = M(X_n - c_n) - M(X_n + c_n)$$

and

$$Z_n = Z(X_n - c_n) - Z(X_n + c_n)$$

with

$$Z(x) = Y(x) - M(x) .$$

As before Z_n is a m.d.s.w.r.t. the σ-algebras generated by X_n . The following conditions will be of interest (*cf*. Sacks, 1958, Heyde, 1974, 1975).

(B1) M is a Borel-measurable function with a unique maximum at $x = \theta$ and $\forall 0 < t_0 < t_1 < t_2 < \infty$,

$$\inf_{0<\varepsilon\le t_0,\, t_1\le|x-\theta|\le t_2} (x-\theta)\{M(x-\varepsilon)-M(x+\varepsilon)\}\varepsilon^{-1} > 0 .$$

Also $\forall x$ and some $D_1, D_2 > 0$, $|M(x+1)-M(x)| < D_1 + D_2|x|$.

$(\overline{B2})$, (B2)* $\forall x$, $M(x) = \alpha_0 - (\alpha_1/4)(x-\theta)^2 + \delta(x, \theta)$ where $\alpha_1 > 0$

for (B2)* $\delta(x, \theta) = o(|x-\theta|^2)$ as $x - \theta \to 0$

for $(\overline{B2})$ $\delta(x_1\theta) = \alpha_2|x-\theta|^{2+\beta} + o(|x-\theta|^{2+\beta})$

as $x - \theta \to 0$ for some finite α_2 and some $\beta > 0$.

(B3) For some $c_0 > 0$, K_1, K_2 $\forall c$ with $0 < c \le c_0$,

$$K_1(x-\theta)^2 \le (x-\theta)\{M(x-c)-M(x+c)\}c^{-1} \le K_2(x-\theta)^2 .$$

(B4) $\forall \eta > 0$, $\exists \varepsilon(\eta) \ni \forall c$ with $0 \le c \le \varepsilon(\eta)$ and $\forall x$ with $|x-\theta| < c$,

$$|\delta(x-c,\ \theta)-\delta(x+c,\ \theta)|c^{-1} \le \eta|x-\theta| \ .$$

(B5) (a) $\sup_x E|Z(x)|^{2+\eta} < \infty$ for some $\eta > 0$,

(b) $\lim_{x\to\theta, a\to 0} E\big(Z(x-a)-Z(x+a)\big)^2 = \sigma^2$.

We have the following results.

THEOREM 1 (Dvoretsky, 1956). *If* $\sup_x \operatorname{var}\big(Y(x)\big) < \infty$ *then for* R-M *under* (A1), (A') *of Lemma* 3, *for* K-W *under* (B1), (A') *of Lemma* 3,

$$X_n - \theta \to 0 \quad a.s.$$

To prove the CLT/IP/LIL however, some additional convergence results are needed via the stronger conditions (A2) for R-M, (B3) for K-W. Consider then the general expression

$$X_{n+1} - \theta = X_n - \theta - a_n M_n c_n^{-1} - a_n c_n^{-1} Z_n \tag{5.11}$$

where for R-M, $c_n = 1 \ \ \forall n$.

For R-M with (A2) and K-W with (B3),

$$E\Big(\big(X_{n+1}-\theta\big)^2 \big| F_n\Big) \le \big(X_n-\theta\big)^2\Big(1-2K_1 a_n+2K_2^2 a_n^2 c_n^{-2}\Big) + a_n^2 c_n^{-2} E\Big(Z_n^2 \big| F_n\Big) \ . \tag{5.12}$$

Let ν_n be an increasing sequence of positive numbers with $\nu_n \uparrow \infty$ then we have

LEMMA 6. *For* R-M *with* (A2), K-W *with* (B3) *if* $\sup_x \operatorname{var}\big(Y(x)\big) < c < \infty$ *and if* (A) *of Lemma* 2 *holds and if the* ν_n *sequence is chosen to ensure*

$$l = \lim\big(\nu_{n+1}/\nu_n - 1\big)/a_n < 2K_1 \tag{5.13}$$

then

$$\lim \nu_n E\big(X_n-\theta\big)^2 \le c\big(2K_1-l\big)^{-1} \lim \nu_n/\Big(c_n^2/a_n\Big) \ .$$

Proof. Take expectations in (5.12) and apply FVT (3). We now have

THEOREM 2. *For* (R-M) *with* (A2), K-W *with* (B3) *and with*

$\sup_x \operatorname{var}(Y(x)) < c < \infty$ *as well as* $\sum_1^\infty a_n = \infty$, $a_n \to 0$, $c_n = 1$, $\forall n$, *or* $c_n \to 0$ *and* $a_n/c_n^2 \to 0$ *we have*

$$p \lim(X_n - \theta) = 0 .$$

Proof. In Lemma 6 set $\nu_n = 1$ to conclude convergence in m.s.

Remark. A theorem close to this one has appeared in Nevel'son and Khasminskii (1973/1976, p. 94 and p. 100).

LEMMA 7. *If in addition to* (3.3) *we have* (A') *of Lemma* 3 *plus*

(A") $\sum_1^\infty \nu_n a_n^2 c_n^{-2} < \infty$

then for R-M *with* (A2), (B) *of Lemma* 2, *for* K-W *with* (B3), (B) *of Lemma* 2,

$$\nu_n (X_n - \theta)^2 \to 0 \quad a.s.$$

Proof. With conditions (A"), (B), MGCT* implies

$$\sum_1^\infty \nu_n a_n (X_n - \theta)^2 < \infty \quad \text{a.s.}$$

as well as $\nu_n (X_n - \theta)^2 \to T$ a.s. a finite r.v.

Now recall $\sum_1^\infty a_n = \infty$ and apply the discrete L'Hospital rule to see

$$\begin{aligned} 0 &= \lim \sum_1^n a_s \nu_s (X_s - \theta)^2 / \sum_1^n a_s \\ &= \lim a_n \nu_n (X_n - \theta)^2 / a_n = T . \end{aligned}$$

The basic idea now is to subtract from the SA scheme the part that obeys the CLT and use the FVT to show the remainder is asymptotically negligible.

Consider, then, reorganising (5.11) as

$$X_{n+1} - \theta = (X_n - \theta)(1 - \alpha_1 a_n) - a_n c_n^{-1} Z_n - a_n c_n^{-1} \delta_n \tag{5.11a}$$

where for R-M, by $(\overline{\text{A3}})$, $\delta_n = \delta(X_n, \theta)$ and $c_n = 1$, while for K-W, by

$(\overline{B2})$, $\delta_n = \{\delta(X_n - c_n,\ \theta) - \delta(X_n + c_n,\ \theta)\}$. Set

$$U_{n+1} = U_n(1-\alpha_1 a_n) - a_n c_n^{-1} Z_n \ , \quad U_0 = 0 \quad \text{a.s.}$$

and

$$D_{n+1} = X_{n+1} - \theta - U_{n+1} = (1-\alpha_1 a_n) D_n - a_n \delta_n c_n^{-1} \qquad (5.14)$$

For the CLT it will do to show

$$\lim E|D_n| c_n/\sqrt{a_n} \to 0 \ . \qquad (5.15)$$

The IP will certainly follow if

$$\sup_{0 \le t \le 1} |D_{n+k_n(t)}| c_{n+k_n(t)}/\sqrt{a_{n+k_n(t)}} = \max_{1 \le m \le k_n(1)} |D_{n+m}| c_{n+m}/\sqrt{a_{n+m}} \xrightarrow{p} 0 \ .$$

This will clearly follow if $c_n|D_n|/\sqrt{a_n} \xrightarrow{p} 0$ which will follow from (5.15). For the LIL we need $c_n|D_n|/\sqrt{a_n} \to 0$ a.s.

THEOREM 3. *For* R-M *under* A2, A3*, A4 *and* A, D *of Lemma* 2, *for* K-W *under* B2*, B3, B4, B5 *and* A, D *of Lemma* 2,

$$Q_n = c_n(X_n - \theta)/\sqrt{a_n} \xrightarrow{\mathcal{D}} N(0,\ \sigma^2/K) \ , \quad K = 2\alpha_1 - \lambda \ ,$$

$$Q_{n+k_n(t)} \Rightarrow Q(t) \quad \textit{of Lemma } 5,$$

$$\lambda = \lim a_n^{-1}\left(c_{n+1}^2 a_n c_n^{-2} a_{n+1}^{-1} - 1\right) \ .$$

Proof. In view of the discussion just given and of Lemma 2, Corollary and Lemma 5, it is enough to show

$$p \lim E\left(Z_n^2 | F_n\right) = \sigma^2 = \lim E\left(Z_n^2\right)$$

(where F_n are the σ-fields generated by X_n) and $\lim E|D_n| c_n/\sqrt{a_n} \to 0$. Take moduli through (5.14), then expectations, and apply FVT (2, 3) (noting $\lim c_{n+1} a_n c_n^{-1} a_{n+1}^{-1} = 1$ because of D) we see

$$\lim E|D_n| c_n/\sqrt{a_n} \le \left(\alpha_1 - \tfrac{1}{2}\lambda\right)^{-1} \lim \frac{c_n}{\sqrt{a_n}} \frac{1}{a_n} a_n E|\delta_n| c_n^{-1} \ .$$

Now take each case separately (though the arguments are similar).

(1) R-M so $c_n = 1$, $\forall n$. Consider that

$$E|\delta_n|/\sqrt{a_n} = E\big(|\delta_n|I\big(|X_n-\theta| < \varepsilon\big)\big)/\sqrt{a_n} + E\big(|\delta_n|I\big(|X_n-\theta| > \varepsilon\big)\big)/\sqrt{a_n} .$$

From A3*, given $\eta > 0$, $\exists\ \varepsilon(\eta) \ni |\delta(x, \theta)| \le \eta|x-\theta|$ for $|x-\theta| < \varepsilon(\eta)$. Also by A2, $|\delta(x, \theta)| \le \big(K_2+\alpha_1\big)|x-\theta|$ so

$$\begin{aligned} E|\delta_n|/\sqrt{a_n} &\le \eta E|X_n-\theta|/\sqrt{a_n} + \big(K_2+\alpha_1\big)E\big(|X_n-\theta|/\sqrt{a_n}\ I\big(|X_n-\theta| > \varepsilon\big)\big) \\ &\le \eta\sqrt{E\big(X_n-\theta\big)^2/a_n} + \big(K_2+\alpha_1\big)\sqrt{E\big(X_n-\theta\big)^2/a_n}\ \sqrt{P\big(|X_n-\theta| > \varepsilon\big)} . \end{aligned}$$

But η is arbitrarily small, $\lim E\big(X_n-\theta\big)^2/a_n < \infty$ by Lemma 6 and $P\big(|X_n-\theta| > \varepsilon\big) \to 0$.

(2) K-W. Consider

$$\begin{aligned} c_nE\Big(c_n^{-1}|\delta_n|\Big)/\sqrt{a_n} = E\Big[c_n^{-1}|\delta_n|I\big(|X_n-\theta| < c_n\big)\Big]c_n/\sqrt{a_n} \\ + E\Big[c_n^{-1}|\delta_n|I\big(|X_n-\theta| > c_n\big)\Big]c_n/\sqrt{a_n} . \end{aligned}$$

From B4, given $\eta > 0$, $\exists\ \varepsilon(\eta) \ni \forall c$ with $0 \le c \le \varepsilon(\eta)$,

$$|\delta(x-c, \theta)-\delta(x+c, \theta)|c^{-1} \le \eta|x-\theta| \quad \text{for} \quad |x-\theta| < c .$$

Also by B3 and the definition of $\delta(x, \theta)$,

$$|\delta(x-c, \theta)-\delta(x+c, \theta)|c^{-1} \le \big(K_2+\alpha_1\big)|x-\theta| ;$$

thus

$$\begin{aligned} c_nE\Big(c_n^{-1}|\delta_n|\Big)/\sqrt{a_n} &\le \eta E|X_n-\theta|c_n/\sqrt{a_n} + \big(K_2+\alpha_1\big)E\big(|X_n-\theta|\big(c_n/\sqrt{a_n}\big)I\big(|X_n-\theta| > c_n\big)\big) \\ &\le \eta\sqrt{E\big(X_n-\theta\big)^2c_n^2/a_n} + \big(K_2+\alpha_1\big)\sqrt{E\big(X_n-\theta\big)^2c_n^2/a_n}\ \sqrt{P\big(|X_n-\theta| > c_n\big)} \end{aligned}$$

and as above, provided $c_n \to 0$ the result follows.

We now show $p\ \lim E\Big(Z_n^2|F_n\Big) = \sigma^2 = \lim E\Big(Z_n^2\Big)$. Consider

$$E\left(Z_n^2 | \;_n\right) - \sigma^2 = E\left(Z^2(X_n)|X_n\right) - \sigma^2$$
$$= \left(E\left(Z^2(X_n)|X_n\right) - \sigma^2\right) I\left(|X_n-\theta| < \varepsilon\right)$$
$$+ \left(E\left(Z^2(X_n)|X_n\right) - \sigma^2\right) I\left(|X_n-\theta| > \varepsilon\right) \ . \qquad (5.16)$$

Now by A4 (b) or B5 (b) the first term has modulus bounded by

$$\eta(\varepsilon) I\left(|X_n-\theta| < \varepsilon\right) \leq \eta(\varepsilon)$$

where $\eta(\varepsilon) = o(\varepsilon)$. However ε is arbitrary and this term may be dispensed with. The second term is bounded by (from B4 (a) or B5 (a))

$$\left(\sigma^2+c\right) I\left(|X_n-\theta| > \varepsilon\right) \quad \text{for some constant } c \ .$$

And since $E\left(I\left(|X_n-\theta| > \varepsilon\right)\right) = P\left(|X_n-\theta| > \varepsilon\right) \to 0$ this term has p lim zero. We conclude $p \lim E\left(Z_n^2|F_n\right) = \sigma^2$. Next take expectations and then moduli in (5.16) to see

$$\left|E\left(Z^2(X_n)\right) - \sigma^2\right| \leq E\left\{\left|E\left(Z^2(X_n)|X_n\right) - \sigma^2\right| I\left(|X_n-\theta| < \varepsilon\right)\right\} + \left(c+\sigma^2\right) P\left(|X_n-\theta| > \varepsilon\right)$$
$$\leq \eta(\varepsilon) + \left(c+\sigma^2\right) P\left(|X_n-\theta| > \varepsilon\right)$$

and we conclude $\lim E\left(Z_n^2\right) = \sigma^2$ as above.

Remark. If only A1, B1 hold rather than A2, B3 then a.s. convergence of $X_n - \theta$ is required to derive the CLT. This follows by the same argument as in Sacks (1958, p. 380 and p. 386).

THEOREM 4 (LIL). *For* R-M *with* A2, $\overline{\text{A3}}$, A4, (5.17) *below, for* K-W *with* $\overline{\text{B2}}$, B3, B4, B5, (5.18), (5.19) *below, together with conditions* A', D *of Lemma* 3 *it follows that*

$$\underline{\lim}, \overline{\lim}\ c_n\left(X_n-\theta\right)/\sqrt{a_n\psi_n} = -\sigma/\sqrt{K}, \ +\sigma/\sqrt{K}$$

respectively with ψ_n *as in Lemma* 3,

$$\sum_1^\infty a_n^{(1+2\beta)/(1+\beta)} < \infty \ , \quad \beta \ \textit{as in } \overline{\text{A3}}, \qquad (5.17)$$

$$\lim c_n^2/a_n^{1/(2+\beta)} < \infty \ , \quad \beta \ \textit{as in } \overline{\text{B2}}, \qquad (5.18)$$

$$\sum_{1}^{\infty} a_n^{(3+2\beta)/(2+\beta)} c_n^{-2} < \infty \ , \quad \beta \ \textit{as in} \ \overline{\mathrm{B2}}. \tag{5.19}$$

Proof. In view of Lemma 3 it is enough to show

$$c_n |D_n| / \sqrt{a_n} \to 0 \quad \text{a.s.}$$

and

$$\lim E\left(Z_n^2 \,|\, F_n\right) = \sigma^2 = \lim E\left(Z_n^2\right) .$$

Given the a.s. convergence of $X_n - \theta$ from Theorem 1 this second requirement follows exactly as in Sacks (1958, 3.14) from A4 or B5.

Now take moduli in (5.14) and apply FVT (2, 3) (noting D implies $\lim c_{n+1} a_n c_n^{-1} a_{n+1}^{-1} = 1$) to see

$$\begin{aligned} \lim c_n \, D_n / \sqrt{a_n} &= \left(\alpha_1 - \tfrac{1}{2}\lambda\right)^{-1} \lim\left(c_n / \sqrt{a_n}\right) a_n^{-1} a_n c_n^{-1} |\delta_n| \\ &= \left(\alpha_1 - \tfrac{1}{2}\lambda\right)^{-1} \lim |\delta_n| / \sqrt{a_n} \ . \end{aligned} \tag{5.20}$$

Again take each case separately.

(1) R-M $\left(c_n = 1 \ \ \forall n\right)$ so from $(\overline{\mathrm{A3}})$ and strong convergence

$$\lim |\delta_n| / |X_n - \theta|^{1+\beta} < \infty \quad \text{a.s.}$$

so

$$\begin{aligned} \lim |\delta_n| / \sqrt{a_n} &\le \lim |\delta_n| / |X_n - \theta|^{1+\beta} \lim \left[\left(X_n - \theta\right)^2 / a_n^{1/(1+\beta)}\right]^{(1+\beta)/2} \\ &= 0 \quad \text{a.s.} \end{aligned}$$

according to Lemma 7 when (5.17) holds.

(2) For K-W from $(\overline{\mathrm{B2}})$ the limit (5.20) above vanishes if each of

$$\lim |X_n - \theta|^2 / a_n^{1/(2+\beta)} \ , \quad |X_n - \theta| c_n / a_n^{1/(2+\beta)} \ , \quad c_n^2 / a_n^{1/(2+\beta)}$$

vanish. They do if (5.18) holds and if the first limit vanishes; it does by Lemma 7 when (5.~~18~~19) holds.

Remark. If $a_n = A/n$, $c_n = n^{-\gamma} = 1$. Then (5.18) holds if $\gamma > 1/6$ while (5.19) holds if $\gamma < 1/3$ (*cf.* Heyde, 1974).

5.4 Some Extensions

The above limit laws do not hold when $2\alpha_1 - \lambda \le 0$ and Major and Révész (1973) gave some results in this case (for the Robbins-Monro scheme with $a_n = 1/n$). It is not hard to intuit the results for general sequences. If we return to the stochastic difference equation of Section 5.2, the variance will obey the difference equation $\left(v_n = E\left(U_n^2\right)\right)$

$$v_{n+1} = v_n\left(1-\alpha_1 a_n\right)^2 + a_n^2 c_n^{-2}\sigma_n^2 .$$

Suppose firstly that $2\alpha_1 = \lambda$; we look for a norming sequence of the form $\nu_n = c_n^2 a_n^{-1}\rho_n^{-1}$ where $\rho_n \to \infty$ is to be chosen. Consider that

$$\nu_{n+1}v_{n+1} = \nu_n v_n\left[1-2\alpha_1 a_n + d\nu_n/\nu_{n+1} + \varphi_n^2\right] + a_n\left(c_{n+1}/c_n\right)^2\rho_n^{-1}\sigma_n^2$$

where $d\nu_n = \nu_{n+1} - \nu_n$ and φ_n^2 need not be specified.

Now calling $w_n = c_n^2/a_n$ we have

$$d\nu_n/\nu_{n+1} = dw_n/w_{n+1} - d\rho_n/\rho_{n+1}$$

so the bracketed term above is

$$1 - a_n\left(2\alpha_1 - dw_n/\left(w_{n+1}a_n\right)\right) - d\rho_n/\rho_{n+1} + \varphi_n^2 .$$

Now when

$$\lim 2\alpha_1 - dw_n/w_{n+1}a_n = 2\alpha_1 - \lambda = 0$$

then, since $\rho_n \to \infty$ implies $\sum_1^\infty d\rho_n/\rho_{n+1} = \infty$, the FVT will still apply with

$$\lim c_{n+1}^2 v_{n+1}/a_{n+1}\rho_{n+1} = \sigma^2 \lim a_n\rho_n^{-1}\,\rho_n/d\rho_n .$$

If we choose $d\rho_n = a_n$ or $\rho_n = \sum_1^n a_s = \tau_n$ then the limit exists and is finite. We would thus expect to conclude $c_n U_n/\sqrt{a_n\tau_n} \xrightarrow{\mathcal{D}} N\left(0, \sigma^2\right)$.

Next suppose $\lim 2\alpha_1 \, dw_n/w_{n+1}a_n = -\mu < 0$. Then choose $\mu a_n = d\rho_n/\rho_{n+1}$ so that $\rho_n \simeq e^{\mu\tau}e^{-\nu'}$ (provided $\sum_1^\infty a_n^2 < \infty$) where $\nu' = \lim(\log P_n + \mu\tau_n)$ with $P_n = \prod_1^n (1-\mu a_s)$.

Next, consider that the forcing term in the difference equation for $\nu_n v_n$ has a finite sum. Indeed this sum is effectively

$$\sum_1^\infty \mu a_n \rho_n^{-1} = \sum_1^\infty d\rho_n/\rho_{n+1}\rho_n = \sum_1^\infty \left(\rho_{n+1}^{-1} - \rho_n^{-1}\right) < \infty .$$

Thus

$$\sum_1^\infty \nu_n v_n = \sum_1^\infty \nu_n E\left(U_n^2\right) < \infty .$$

It follows then by the lemma in Lukacs (1975, p. 80) that

$$\sqrt{\nu_n}\, U_n = e^{-\mu\tau_n/2} c_n U_n/\sqrt{a_n} \to 0 \quad \text{a.s.}$$

Of course to prove these results for the SA schemes we would, in particular have to weaken conditions A2 and B3, to A2' and B3' respectively of Sacks (1958, p. 380 and p. 386).

So far as Multidimensional SA is concerned the results obtained will hold under the natural multidimensional extensions of the conditions A1-A4, B1-B5 as in Sacks (1958, Section 5). However in expressions such as

$$M(x) = B(x-\theta) + \delta(x)$$

Sacks assumed B to be positive definite. We can relax this assumption by proceeding as follows. Let us first try to find a vector FVT beginning with the special case

$$x_{n+1} = \left(I - a_n A\right)x_n + a_n p$$

where x_n, p are vectors and the matrix A has an inverse. Defining

$$\pi_n = \prod_1^n \left(I - a_s A\right)$$

we see that

$$\pi_{n+1}^{-1} - \pi_n^{-1} = a_n \pi_{n+1}^{-1} A .$$

Next, summing the expression for x_{n+1} we obtain

$$
\begin{aligned}
x_{n+1} &= \pi_{n+1} \sum_{1}^{n} \pi_{s+1}^{-1} a_s p + \pi_{n+1}\pi_1^{-1} x_1 \\
&= \pi_{n+1} \sum_{1}^{n} \left(\pi_{s+1}^{-1} - \pi_s^{-1}\right) A^{-1} p + \pi_{n+1}\pi_1^{-1} x_1 \\
&= A^{-1} p + \pi_{n+1}\pi_1^{-1}\left(x_1 - A^{-1} p\right)
\end{aligned}
$$

we will obtain

$$x_n \to A^{-1} p$$

if we can ensure

$$w_{n+1} = \pi_n w = \left(I - a_n A\right) w_n \to 0 .$$

We can guess the condition required on A as follows. Introduce the time scale $\tau_n = \sum_{1}^{n} a_s$ and write $dw_{n+1} = w_{n+1} - w_n$ to see

$$dw_{n+1} = -A d\tau_n :$$

it is clear we will need $-A$ to be a stability matrix, that is, to have all its eigenvalues negative. Let K be a positive definite (p.d.) matrix to be chosen, and consider the quadratic form

$$
\begin{aligned}
\|w_{n+1}\|_K &= w'_{n+1} K w_{n+1} \\
&\le \|w_n\|_K \left(1 + l a_n^2\right) - a_n w'_n (AK + KA') w_n
\end{aligned}
$$

where l is the largest eigenvalue of

$$K^{-\frac{1}{2}} A' K A K^{-\frac{1}{2}} .$$

Now if we require $-A$ to be a stability matrix, then according to the celebrated Lyapunov stability theorem (Barnett, 1971, Chapter 4) we can always find a K so that

$$AK + KA'$$

is strictly p.d. Thus we conclude that for some $c > 0$,

$$\|w_{n+1}\|_K \le \|w_n\|_K \left(1 - c a_n + l a_n^2\right)$$

and $\sum_1^\infty a_n = \infty$ ensures $\|w_n\|_K \to 0$; so the simplest vector FVT is established.

Next, if we consider

$$\bar{x}_{n+1} = (I-a_nA)\bar{x}_n + a_np_n$$

where $p_n \to p$, then we can, by subtracting off the equation for x_{n+1} reduce consideration to

$$y_{n+1} = (I-a_nA)y_n + a_n\tilde{p}_n$$

where $\tilde{p}_n \to 0$: it is required to show $y_n \to 0$. This follows upon considering that

$$\|y_{n+1}\|_K \le \|y_n\|_K\left(1-ca_n+la_n^2\right) + a_n\|\tilde{p}_n\|_K$$

and applying the scalar FVT. Finally notice that we can replace A by a sequence $A_n \to A$ since then

$$y_{n+1} = \left(I-a_nA+a_n(A-A_n)\right)\bar{y}_n + a_n\tilde{p}_n$$

and

$$\|y_{n+1}\|_K \le \|y_n\|_K\left(1-(c-\rho_n)a_n+la_n^2\right) + a_n\|\tilde{p}_n\|_K$$

where

$$\rho_n = \|A-A_n\|_K\|K^{-\frac{1}{2}}\|^2 \to 0$$

once more the result follows by the scalar FVT; we have then, a vector FVT.

Now we can deduce the vector version of the CLT of Theorem 2 by showing it for $\alpha'U_n$, for an arbitrary fixed vector α and noting that provided $-(B-\lambda I/2)$ is a stability matrix then

$$K_n = \left(c_n^2/a_n\right)E\left(U_nU_n'\right) \to K$$

where K is the p.d. covariance matrix satisfying

$$(B-\lambda I/2)K + K(B-\lambda I/2)' = \pi$$

with π given as in Sacks (1958, Section 5) as

$$\pi = \lim_{x\to\theta} E\left(Z(x)Z'(x)\right) .$$

To see this note that $V_n = E(U_n U_n')$ obeys

$$V_{n+1} = (I-a_n B) V_n (I-a_n B') + a_n^2 \pi$$

so that after a little reorganisation

$$c_{n+1}^2 / a_{n+1} V_{n+1} = K_{n+1} = O\left(a_n^2\right) + (1+o_n(1)) a_n \pi$$
$$+ K_n - a_n\left\{(B-\lambda_n I/2) K_n + K_n (B-\lambda_n I/2)'\right\}$$

where

$$\lambda_n = a_n^{-1}\left(c_{n+1}^2 c_n^{-2} a_n a_{n+1}^{-1} - 1\right) \to \lambda \ .$$

Now introduce $k_n = \text{vec}(K_n)$ that is, the vector obtained by stacking the columns of K_n one underneath another (see Neudecker, 1969 for an expose of this 'vec' operator) then

$$k_{n+1} = (I-a_n A_n) k_n + O\left(a_n^2\right) + a_n \text{ vec}(\pi)(1+o_n(1))$$

where

$$A_n = (B-\lambda_n I/2) \otimes I + I \otimes (B-\lambda_n I/2)'$$

where $\otimes$ is the Kronecker product. Now if (for n large enough) $B - \lambda_n I/2$ is a stability matrix, so is A_n since they have the same eigenvalues, thus

$$k_n \to \text{vec}(K) = A^{-1} \text{ vec}(\pi) \ .$$

For the IP we need to know, that, uniformly in t,

$$\prod_{N}^{N+k_N(t)} \left(I-a_s(B-\lambda_s I/2)\right) - e^{-\frac{1}{2}(B-\lambda I/2)t} \to 0$$

where the matrix exponential is defined by the obvious series expansion. The proof follows much as before by using obvious properties of the matrix exponential as defined in Gantmacher (1960, p. 113) provided we take s large enough that all eigenvalues of $a_s(B-\lambda_s I/2)$ have modulus less than unity.

For the SA results notice that an expression of the form (5.12) will hold with $(X_n-\theta)^2$ replaced by $(X_n-\theta)'(X_n-\theta)$. Also we can investigate the vector analogue of (5.14) successfully by considering $\|D_n\|_K$ and using the scalar FVT. Similar results to the above have been obtained for R-M with $a_n = n^{-1}$ by Nevelson and Khasminskii (1973/1976) and also Walk (1976) by rather less clear argument and under more restrictive conditions.

CHAPTER 6

STOCHASTIC APPROXIMATION WITH DEPENDENT NOISE

6.1 Introduction

In Chapter 5 the usual limit laws for Stochastic Approximation (SA) - Central Limit Theorem (CLT), Invariance Principle (IP), and Law of the Iterated Logarithm were obtained by *first* showing consistency by usual arguments and *then* subtracting out the part of the scheme that obeys the limit law and showing the remainder to be asymptotically negligible by a time varying version of the Z-transform Final Value Theorem (FVT). The point to be shown now is that this idea of subtracting out the part that obeys the limit law can in fact be used to show convergence. Indeed the error term is shown to approach zero by essentially deterministic arguments so that, in essence no probability theory is used.

We begin by specifying the noise model to be considered. It will be recalled that in classical SA the variance of the observation error is allowed to depend on the value of the SA scheme. Here we assume the variance is independent of this level. This much stronger assumption is used since the author has had difficulty in specifying a suitable model that would allow the statement of 'tidy' results. In particular this assumption allows 'clean' formulae to be given for the asymptotic variance of the normalised SA scheme. We begin as in the earlier work by investigating the behaviour of the part of the SA scheme that obeys limit laws.

6.2 Limit Theory for a Stochastic Difference Equation Driven by Stationary Noise

Suppose U_n is the output of the difference equation

$$U_{n+1} = U_n\left(1-\alpha_1 a_n\right) + a_n c_n^{-1} Z_n \tag{6.1}$$

where α_1 is a positive constant and a_n, c_n are sequences of positive constants satisfying

(A) $\sum_1^\infty a_n = \infty$

(this ensures the filtering operation in (6.1) is nondegenerate) and $c_n = 1$ $\forall n$ or else $c_n \to 0$ and $a_n c_n^{-2} \to 0$.

The following further conditions will be of interest: some being motivated by similar conditions in Chapter 5.

(A') As for A plus $\sum_1^\infty a_n^2 c_n^{-2} < \infty$.

(B) Z_n is a stationary process. Let F_n be the increasing σ-algebras generated by Z_n and consider

$$u_{n,j} = E(Z_n | F_j) - E(Z_n | F_{j-1})$$

so that

$$Z_n = \sum_0^\infty u_{n,j} + E(Z_n | F_{-\infty}) \ .$$

Now stationarity ensures we can write

$$\alpha_j = \sqrt{E}\left(u_{n,n-j}^2\right)$$

so put $u_{n,n-j} = \xi_j\ {}_{n,n-j}$ with $E\left(\xi_{n,n-j}^2\right) = 1$. Assume now (as in Hannan, 1978b) that

$$E(Z_n | F_{-\infty}) = 0 \ , \quad \sum_0^\infty \alpha_j < \infty \ .$$

This assumption implies immediately a Doob-like maximal inequality for Z_n that allows strong convergence to be established for partial sums (via the method of subsequences (Stout, 1974) as well as tightness for the IP (see later)).

(B') As for B together with: Z_n is mixing (Rosenblatt, 1974, p. 110) that is for any F_∞ measurable events A, B ,

$$\lim_{n\to\infty} P\left(A \cap T^{-n}B\right) = P(A)P(B) = 0 \ .$$

This asymptotic independence assumption will allow establishment of a CLT and IP.

(C)

$$\lim a_n^{-1}\left(c_{n+1}^2 c_n^{-2} a_n a_{n+1}^{-1} - 1\right) = \lambda < 2\alpha_1 ,$$

$$\lim c_n c_{n+1}^{-1} = 1 = \lim a_n a_{n+1}^{-1} .$$

In connexion with an IP for $Q_n = c_n U_n / \sqrt{a_n}$ the following considerations enter. Denote

$$k_n(t) = \sup\{m : \tau_{n+m} - \tau_n \le t\} ,$$

$$\tau_n = \sum_1^n a_k .$$

Multiply through (6.1) and reorganise it to obtain

$$dQ_{n+1} = -\tfrac{1}{2} K_n a_n d\tau_n + dW_n \tag{6.2}$$

where $d\tau_n = \tau_n - \tau_{n-1} = a_n$ and so on,

$$W_n = \sum_1^n \sqrt{a_s}\, Z_s w_n ,$$

$$K_n = 2\alpha_1 w_n - a_n^{-1}\left(\omega_n^2 - 1\right) - \tfrac{1}{2} a_n^{-1}\left(\omega_n - 1\right)^2 ,$$

with

$$w_n = c_{n+1} c_n^{-1} \sqrt{a_n a_{n+1}^{-1}} ,$$

so that by condition (C), $K_n \to K = 2\alpha_1 - \lambda$ (*cf.* Chapter 5). On $D[0, 1]$ the space of all right continuous functions on $[0, 1]$ with left hand limits take $W(t)$ to be Wiener measure of variance

$$\sigma^2 = \sum_0^\infty \alpha_j .$$

Define $Q(t)$ to be the continuous Gaussian process on $D[0, 1]$ with

$$E\big(Q(t)\big) = 0 , \quad 0 \le t \le 1 ,$$

$$E\big(Q(t)Q(s)\big) = \sigma^2 e^{-\frac{1}{2}K(t-s)} , \quad 0 \le s \le t \le 1 , \tag{6.3}$$

(*cf.* Chapter 5).

The following limit laws are then of interest: proofs are given after the statement of the theorems.

THEOREM 1. *Under* A, B, C, $E\left(U_n^2\right) \to 0$ *and* $\lim c_n^2 E\left(U_n^2\right)/a_n < \infty$.

THEOREM 2. *Under* A', B, C, $U_n \to 0$ *a.s.* *and* $\sum_1^\infty a_n U_n^2 < \infty$ *a.s.*

THEOREM 3. *Under* A, B', C,

$$Q_n = c_n U_n / \sqrt{a_n} \xrightarrow{\mathcal{D}} N\left(0, \sigma^2/K\right)$$

where

$$\sigma^2 = \sum_0^\infty \alpha_j$$

and

$$K = 2\alpha_1 - \lambda \ .$$

THEOREM 4. *Under* A, B', C,

$$W_{n+k_n(t)} - W_n \Rightarrow W(t) \ .$$

THEOREM 5. *Under* A, B', C,

$$Q_{n+k_n(t)} \Rightarrow Q(t)$$

defined in (6.3).

Before giving the proofs first note that since under B or B', $\sum_0^\infty \alpha_j < \infty$, Z_n has a bounded spectrum. Indeed

$$f(\omega) = \left|\sum_0^\infty \alpha_j e^{ji\omega}\right|^2 \le \left(\sum_0^\infty \alpha_j\right)^2 = M \text{ say.}$$

Proof of Theorem 1. Summing (6.1) gives

$$U_n = \sum_1^n a_s c^{-1} Z_s \pi_s^{-1} \pi_n$$

(initial condition effects will not effect the asymptotic results) where $\pi_n = \pi_1^n\left(1-\alpha_1 a_s\right)$. Thus

$$\operatorname{var}\left(U_n\right) = \pi_n^2 \sum_1^n \sum_1^n a_s a_{s'} c_s^{-1} c_{s'}^{-1} \pi_s^{-1} \pi_{s'}^{-1} \rho_{s-s'}$$

where

$$\rho_{s-t} = E\left(Z_s Z_t\right)$$
$$= \int_{-\pi}^{\pi} f(\omega) e^{i\omega(s-t)} d\omega \ .$$

Thus

$$\operatorname{var}\left(U_n\right) = \int_{-\pi}^{\pi} \pi_n^2 \left| \sum_1^n a_s c_s^{-1} \pi_s^{-1} e^{i\omega s} \right|^2 f(\omega) d\omega$$
$$\leq M\pi_n^2 \sum_1^n a_s^2 c_s^{-2} \pi_s^{-2} \ .$$

Denoting the right hand side (less the factor M) by v_n we see

$$v_{n+1}^2 = v_n^2 \left(1-\alpha_1 a_n\right)^2 + a_n^2 c_n^{-2}$$

and the final Value Theorem FVT of Chapter 5 yields

$$\lim v_n = \lim a_n^{-1} a_n^2 c_n^{-2} = 0 \quad \text{by C.}$$

The FVT also yields

$$\lim c_n^2 v_n / a_n < \infty \ .$$

Proof of Theorem 2. We appeal to a theorem of Hannan (1978a) which may be restated as follows. If x_s, J_s are a nonstochastic sequence of numbers and Z_s obeys B then

$$\sum_1^\infty x_s Z_s / J_s < \infty \quad \text{a.s.}$$

if

$$\sum_1^\infty x_s^2 / J_s^2 < \infty \ .$$

This theorem is of course well known if the Z_s are a m.d.s. (Hannan's proof uses the method of subsequences together with a Doob-like maximal inequality for the Z_s). If $J_s \uparrow \infty$ then Kronecker's lemma implies

$$J_n^{-1} \sum_1^n x_s Z_s \to 0 \quad \text{a.s.}$$

In the present case take $x_s = a_s c_s^{-1} \pi_s^{-1}$, $J_s = \pi_s^{-1}$, then

$$\sum_1^\infty x_s^2/J_s^2 = \sum_1^\infty a_s^2 c_s^{-2} \pi_s^{-2} \pi_s^2 = \sum_1^\infty a_s^2 c_s^{-2} < \infty \quad \text{by A}'$$

so that

$$J_n^{-1} \sum_1^n x_s Z_s = \pi_n \sum_1^n a_s c_s^{-1} \pi_s^{-1} = U_n \to 0 \quad \text{a.s.}$$

On the other hand $\sum_1^\infty a_n U_n^2 < \infty$ if $\sum_1^\infty a_n E\left(U_n^2\right) < \infty$ (Lukacs, 1975, p. 80) but by the comment at the end of the proof of Theorem 1 ~~Theorem 2~~ and by condition A' this holds.

Remark. By assuming the existence of higher moments Ljung (1978a) has given other conditions under which $U_n \to 0$ a.s.

Proof of Theorem 3. Summing (6.2) yields, neglecting initial conditions

$$\begin{aligned} Q_N &= P_{N+1} \sum_1^N P_{n+1}^{-1} \sqrt{a_n}\, Z_n w_n \\ &= \sum_1^N y_{Nn} Z_n / d(N) \end{aligned}$$

where

$$P_n = \pi_1^n \left(1 - \tfrac{1}{2} K_s a_s\right) ,$$

$$y_{Nn} = \sqrt{P_{N+1}}\, P_{n+1}^{-1} \sqrt{a_n}\, w_n , \quad d(N) = P_{N+1}^{-\frac{1}{2}} .$$

Notice

$$P_n^{-2} - P_{n-1}^{-1} = P_n^{-2} K_n a_n \left(1 + o_n(1)\right) \tag{6.4}$$

so that, recalling $w_n \to 1$, $K_n \to K$, we see

$$\sum_1^N y_{Nn}^2 / d^2(N) \to K .$$

The result will follow from Hannan (1978b) (see also Hannan, 1973, where regularity is imposed rather than mixing) if we show

(i) $d^2(N) \to \infty$,

(ii) $\lim_{N\to\infty} \max_{1\le n\le N} \frac{|y_{Nn}|}{d(N)} = 0$.

(In Hannan, 1973 condition (ii) is incorrectly stated and should be the one given here: *cf*. Hannan, 1978b.)

(iii) $\lim_{N\to\infty} \sum_{1}^{N-n} y_{Nm} y_{Nn+m}/d^2(N) = 1$, $m \ge 0$.

For (i) we require $P_N^{-1} \to \infty$ or $P_N \to 0$ which follows from the theory of infinite products (Bromwich, 1947) if

$$\sum_{1}^{\infty} K_n a_n = \infty$$

since $K_n \to K$ this follows from A.

For (ii) we have to investigate

$$\max_{1\le n\le N} P_N P_n^{-1} \sqrt{a_n} \ .$$

Let M_N be the sequence such that

$$\max_{1\le n\le N} P_N P_n^{-1} \sqrt{a_n} = P_N P_{M_N}^{-1} \sqrt{a_{M_N}} \ .$$

If $\{M_N\}$ is bounded (b) follows since $P_N \to 0$. If $M_N \to \infty$ then since P_N is a decreasing sequence $\left(P_N P_{N-1}^{-1} = 1 - \tfrac{1}{2} K_N a_N\right)$

$$P_N P_{M_N}^{-1} \sqrt{a_{M_N}} < P_{M_N} P_{M_N}^{-1} \sqrt{a_{M_N}} = \sqrt{a_{M_N}} \to 0 \ .$$

For (iii). It follows immediately from condition C and expression (6.4) that (iii) holds.

Proof of Theorem 4. Recall

$$W_N = \sum_{1}^{N} \sqrt{a_n}\, z_n w_n = \sum_{0}^{\infty} \alpha_j \sum_{1}^{N} \sqrt{a_n}\, \xi_{n,n-j} w_n \ .$$

Now $w_n \to 1$ and so 1 can replace it. Consider the truncated process

$$W_N^u = \sum_0^u \alpha_j \sum_1^N \sqrt{a_n}\, \xi_{n,n-j} \ .$$

Define

$$W_N^u(t) = W_{N+k_N(t)}^u - W_N^u \ .$$

Theorem 4 will follow via Theorem 4.2 of Billingsley (1968) if we show (α), (β), (γ).

(α) $W_N^u(t) \Rightarrow W^u(t)$, $\forall u$ as $N \to \infty$, where $W^u(t)$ is Brownian motion on $D[0, 1]$ with

$$\sigma_u^2 = \sum_0^u \alpha_j \ .$$

(β) $W^u(t) \Rightarrow W(t)$ as $u \to \infty$. This follows already from (α) since $\sum_0^u \alpha_j \to \sum_0^\infty \alpha_j < \infty$.

(γ) $\lim_{u\to\infty} \overline{\lim_{N\to\infty}} P\left\{ \sup_{0\le t\le 1} \left| W_N^u(t) - W_N(t) \right| \ge \varepsilon \right\} \to 0$, $\forall \varepsilon$.

Now (γ) follows if

$$\lim_{u\to\infty} \overline{\lim_{N\to\infty}} E\left\{ \max_{1\le k\le k_N(1)} \left| \sum_N^{N+k} \sqrt{a_n}\, Z_n^u \right|^2 \right\} \to 0$$

where

$$Z_n^u = \sum_u^\infty \alpha_j \xi_{n,n-j}$$

and

$$\tau_{N+k_N(1)} - \tau_N \to 1 \ .$$

However the expectation is

$$E_{Nu} = E\left\{\max_{1\le k\le k_N(1)} \left(\sum_u^\infty \alpha_j \sum_N^{N+k} \sqrt{a_n}\, \xi_{n,n-j}\right)^2\right\}$$

$$\le \sum_u^\infty \alpha_j \sum_u^\infty \alpha_j E\left\{\max_{1\le k\le k_N(1)} \left(\sum_N^{N+k} \sqrt{a_n}\, \xi_{n,n-j}\right)^2\right\}$$

$$\le \left(\sum_u^\infty \alpha_j\right)\left(\sum_u^\infty \alpha_j\right)\left(\sum_N^{N+k_N(1)} a_n\right)$$

by Doob's maximal inequality (Stout, 1974).

Thus

$$\lim_{u\to\infty} \overline{\lim_{N\to\infty}} E_{Nu} = \lim_{u\to\infty} \left(\sum_u^\infty \alpha_j\right)^2 = 0 \ .$$

To demonstrate (α) we rewrite $W_N^{u}(t)$ as a sum of martingale difference sequences and apply the triangular array IP of Scott (1973).

$$W_N^{u}(t) = \sum_{n\le k_N(t)} \xi_{N_n} + \mu_{N-1} - \mu_{N+k_N(t)} \tag{6.5}$$

where

$$\mu_m = \sum_0^u \alpha_j \sum_{n=m}^{m+j} \sqrt{a_n}\, \xi_{n,n-j}$$

and

$$\xi_{Nn} = \sum_0^u \alpha_j \sqrt{a_{N+n+j}}\, \xi_{N+n+j,N+n} \ .$$

However proceeding as for (γ) above

$$E\left\{\sup_{0\le t\le 1} \mu_{N+k_N(t)}\right\} = E\left\{\max_{N\le M\le N+k_N(1)} \mu_m^2\right\}$$

$$\le \sum_0^u \alpha_j \sum_0^u \alpha_j \sum_{N+k_N(1)}^{N+k_N(1)+j} a_n$$

$$\le \left(\sum_0^u \alpha_j\right)^2 \sum_{N+k_N(1)}^{N+k_N(1)+u} a_n \to 0 \tag{6.6}$$

since

$$\sum_{N}^{N+k_N(1)} a_n \to 1 .$$

We have now to show

$$\sup_{1\le k\le N} \xi_{Nn}^2 \xrightarrow{p} 0 , \tag{6.7}$$

$$\sum_{n\le k_N(t)} \xi_{Nn}^2 \xrightarrow{p} t\sigma_u^2 . \tag{6.8}$$

Now (6.7) holds by the type of argument leading to (6.6). Furthermore, in view of (6.5), (6.6) we see that (6.8) can be replaced by

$$\sum_{N}^{N+k_N(t)} a_n\left(Z_n^u\right)^2 \xrightarrow{p} t\sigma_u^2 .$$

We now follow Hannan (1978b). Put $\eta(n) = \left(Z_n^u\right)^2 - \sigma_u^2$ and show

$$\sum_{N}^{N+k_N(t)} a_n\eta(n) \xrightarrow{p} 0 . \tag{6.9}$$

Set

$$\eta_1(n) = \eta(n)I\left(|\eta(n)| < M\right) , \quad \eta_2(n) = \eta(n)I\left(|\eta(n)| > M\right) .$$

Then $\eta_i(n) - E\left(\eta_i(n)\right)$, $i = 1, 2$, are stationary, mixing. Now

$$E\left|\sum_{N}^{N+k_N(t)} a_n\left(\eta_2(n)-E\left(\eta_2(n)\right)\right)\right| \le \sum_{N}^{N+k_N(t)} a_n E\left|\eta_2(0)-E\left(\eta_2(0)\right)\right|$$
$$\to tE\left|\eta_2(0)-E\left(\eta_2(0)\right)\right|$$

which can be made arbitrarily small by taking M large. Thus we need only show (6.9) with $\eta(n)$ replaced by $\eta_1(n)$.

Let b_j be a finite set of nondecreasing numbers with $b_{j+1} - b_j < \varepsilon$ and set

$$A_j = \{\omega | b_j \le \eta_1(0, \omega) \le b_{j+1}\} .$$

Then denoting the indicator function of A_j by $\chi_j(\omega)$,

$$\left|\eta_1(n) - \sum b_j\chi_j\left(T^{-n}\omega\right)\right| < \varepsilon$$

and it is enough to show (6.9) with $\eta_1(n)$ replaced by $\sum a_j\chi_j\left(T^{-n}\omega\right)$ and so with $\chi_j\left(T^{-n}\omega\right)$.

However

$$E\left(\sum_{N}^{N+k_N(t)} a_n\chi_j\left(T^{-n}\omega\right)\right) = \sum_{N}^{N+k_N(t)} a_nP\left(T^{-n}A_j\right) \to tP\left(A_j\right)$$

by B' and the Toeplitz lemma (Stout, 1974). Once more this may be made arbitrarily small.

6.3 Stochastic Approximation

We turn now to the R-M, K-W schemes and take conditions A1, A2, A3*/B2*, B3, B4 of Chapter 5. Also B1 of Chapter 5 is strengthened slightly to B1'.

(B1') M is a Borel-measurable function with a unique maximum at $x = \theta$ and $\forall t_0 < t_1 < \infty$,

$$\inf_{\substack{t_1\le|x-\theta|<\infty \\ 0<\varepsilon<t_0}} (x-\theta)\{M(x-\varepsilon)-M(x+\varepsilon)\}\varepsilon^{-1} > 0 \ .$$

Also $\forall x$ and some positive constants D_1, D_2 ,

$$|M(x+1)-M(x)| < D_1 + D_2|x| \ .$$

First weak laws are discussed and then a.s. convergence.

THEOREM 6. *For* R-M *under* A2, A3* *of Chapter* 5, *for* K-W *under* B2*, B3, B4 *of Chapter* 5, and with A, B, C, $X_n - \theta \to 0$ m.s.

Proof. Recall the basic expression of Chapter 5 (5.11)

$$X_{n+1} - \theta = X_n - \theta - a_nc_n^{-1}M_n - a_nc_n^{-1}Z_n \qquad (6.10)$$

and subtract (6.1) from it to obtain

$$D_{n+1} = X_{n+1} - \theta - U_{n+1}$$
$$= D_n - a_n c_n^{-1} M_n + \alpha_1 a_n U_n \ . \qquad (6.11)$$

Squaring yields

$$D_{n+1}^2 = \left(D_n - a_n c_n^{-1} M_n\right)^2 + \alpha_1^2 a_n^2 U_n^2 + 2\alpha_1 a_n U_n \left(D_n - a_n c_n^{-1} M_n\right)$$
$$\leq \left(D_n - a_n c_n^{-1} M_n\right)^2 + \alpha_1^2 a_n^2 U_n^2 + \varepsilon^{-1} \alpha_1^2 a_n U_n^2 + \varepsilon a_n \left(D_n - a_n c_n^{-1} M_n\right)^2$$

with ε to be chosen. Now apply A2 for R-M or B3 for K-W to see

$$D_{n+1}^2 \leq D_n^2\left(1-K_1 a_n\right)\left(1+\varepsilon a_n\right) + K_2 a_n^2 \left(X_n - \theta\right)^2 \left(1+\varepsilon a_n\right) + \alpha_1^2 a_n^2 U_n^2 + \varepsilon^{-1} \alpha_1^2 a_n U_n^2 \ .$$

Now

$$\left(X_n-\theta\right)^2 = \left(D_n + U_n\right)^2 \leq 2D_n^2 + 2U_n^2 \ ; \quad 1 + \varepsilon a_n \leq 2 \ .$$

So

$$D_{n+1}^2 \leq D_n^2\left[1 - \left(K_1 - \varepsilon\right) a_n + 4K_2 a_n^2\right] + 2\varepsilon^{-1} \alpha_1^2 a_n U_n^2 \ .$$

Choose $\varepsilon = \frac{1}{2} K_1$ and take expectations to see

$$E\left(D_{n+1}^2\right) \leq E\left(D_n^2\right)\left(1 - \tfrac{1}{2}K_1 a_n + 4K_2 a_n^2\right) + 2\varepsilon^{-1} \alpha_1^2 a_n E\left(U_n^2\right) \qquad (6.12)$$

and it follows by the FVT (Chapter 5) that

$$\lim E\left(D_{n+1}^2\right) = \left(\tfrac{1}{2}K_1\right)^{-1} 2\varepsilon^{-1} \alpha_1^2 \lim E\left(U_n^2\right)$$
$$= 0 \quad \text{by Theorem 1.}$$

Thus $E\left(X_n - \theta\right)^2 \leq 2E\left(D_n^2\right) + 2E\left(U_n^2\right) \to 0$.

Remark. It is clear from the proof that A, B, C are required only to ensure $E\left(U_n^2\right) \to 0$. In other words under the first two sets of conditions $E\left(X_n - \theta\right)^2 \to 0$ if $E\left(U_n^2\right) \to 0$.

THEOREM 7. *For* R-M *under* A2, A3* *of Chapter* 5, *for* K-W *under* B2*, B3, B4 *of Chapter* 5, *and conditions* A, B, C,

(i) $\hat{Q}_n = c_n\left(X_n - \theta\right)/\sqrt{a_n} \xrightarrow{\mathcal{D}} N\left(0, \sigma^2/K\right)$,

(ii) $\hat{Q}_{n+k_n}(t) \Rightarrow Q(t)$ *of Theorem* 5.

Proof. Now, by Theorem 1, A, B, C ensure

$$\lim c_n^2 E\left(U_n^2\right)/a_n < \infty .$$

Thus from (6.12), it follows, by the FVT (Chapter 5) that

$$\lim c_n^2 E\left(D_n^2\right)/a_n = \text{const.} \lim E\left(U_n^2\right)c_n^2/a_n < \infty ;$$

thus

$$\lim c_n^2 E\left(X_n-\theta\right)^2/a_n < \infty . \qquad (6.13)$$

Now (i) will follow from Theorem 3 if we show

$$\lim E|D_n|c_n/\sqrt{a_n} = 0 . \qquad (6.14)$$

However given (6.13) this follows from (A3) exactly as in Chapter 5 (Proof of Theorem 3).

Again, given (6.13), (iii) follows from (6.14) by the same argument as in Chapter 5 (Proof of Theorem 3).

THEOREM 8. *For* R-M *under* A1, *for* K-W *under* B1', *and if*

$$U_n \to 0 \quad a.s., \qquad (6.15)$$

$$\sum_1^\infty a_n U_n^2 < \infty \quad a.s. \qquad (6.16)$$

Then $X_n - \theta \to 0$ *a.s.*

THEOREM 8'. *For* R-M *under* A1, *for* K-W *under* B1', *and conditions* A', B', C, *then* $X_n - \theta \to 0$ *a.s.*

Proof of Theorem 8'. This follows from Theorem 8 since (Theorem 2) A', B', C imply the validity of (6.15) and (6.16).

Proof of Theorem 8. We deal first with the K-W scheme and then with the R-M scheme.

(i) K-W. We argue realisation-wise. Given $0 < \rho < 1$,

$$\exists\, n_0'(\omega) \ni \forall n > n_0'(\omega)$$

$$|U_n(\omega)| < \rho \quad \text{since} \quad U_n \to 0 \quad \text{a.s.}$$

Also

$$\exists N_0 \ni \forall n > N_0 , \quad c_n < \rho ;$$

so take $n_0(\omega) = \max\left(N_0, n_0'(\omega)\right)$ and let $n > n_0(\omega)$.

If $|X_n(\omega)-\theta| < 2\rho$ then $|D_n(\omega)| < 3\rho$ so that, by B1',

$$\left|c_n^{-1}M_n\right| < A\rho + B \quad \text{for some constants } A, B .$$

Thus, from the basic expression (6.11),

$$\begin{aligned} |D_{n+1}(\omega)| &\le 3\rho + a_n(A\rho+B) + \alpha_1 a_n \\ &= 3\rho + B(\rho)a_n , \end{aligned} \tag{6.17}$$

$$B(\rho) = \alpha_1\rho + A\rho + B .$$

If $|X_n(\omega)-\theta| > 2\rho$ rewrite (6.11) as

$$\begin{aligned} D_{n+1}(\omega) = D_n(\omega) + a_n\left(X_n(\omega)-\theta\right)c_n^{-1}M_n(\omega)/\left(X_n(\omega)-\theta\right) + \alpha_1 a_n U_n(\omega) \\ + a_n U_n(\omega)\left[\alpha_1+c_n^{-1}M_n(\omega)\right]/\left(X_n(\omega)-\theta\right) . \end{aligned} \tag{6.11a}$$

Now denote

$$\gamma(\rho) = \inf_{\substack{2\rho<|x-\theta|<\infty \\ 0<\varepsilon<\rho}} \frac{\{M(x-\varepsilon)-M(x+\varepsilon)\}\varepsilon^{-1}}{(x-\theta)} .$$

This is positive by B1'.

We find, on squaring (6.11a) and using B1 (which implies $\left|c_n^{-1}M_n(\omega)/\left(X_n(\omega)-\theta\right)\right|^2 \le c(\rho)$ for some continuous function $c(\cdot)$) that

$$\begin{aligned} D_{n+1}^2(\omega) &\le D_n^2(\omega)\left(1-\gamma(\rho)a_n\right)^2 + a_n^2U_n^2(\omega)\left[2\alpha_1^2+2c^2(\rho)\right] \\ &\qquad + 2a_n|U_n(\omega)|\,|D_n(\omega)|\left(1-\gamma(\rho)a_n\right)\left(\alpha_1+c(\rho)\right) \\ &\le D_n^2(\omega)\left(1-\tfrac{1}{2}\gamma(\rho)a_n\right) + a_n^2U_n^2(\omega)\left[2\alpha_1^2+2c^2(\rho)\right] \\ &\qquad + a_nU_n^2(\omega)\left(\alpha_1+c(\rho)\right)^2\left(\tfrac{1}{2}\gamma(\rho)\right)^{-1} . \end{aligned}$$

However in the present case we must have $|D_n(\omega)| > \rho$ so that

$$D_{n+1}^2(\omega) \le D_n^2(\omega) - \tfrac{1}{2}\rho^2\gamma(\rho)a_n + D(\rho)a_nU_n^2(\omega) \tag{6.18}$$

for some continuous function $D(\rho)$.

Now square (6.17) and combine it with (6.18) as

$$D^2_{n+1}(\omega) \le \max\left\{9\rho^2+6\rho B(\rho)a_n+B^2(\rho)a_n^2,\ D_n^2(\omega)-\tfrac{1}{2}\rho^2\gamma(\rho)a_n+D(\rho)a_nU_n^2(\omega)\right\} .$$

Finally, since $\sum_1^\infty a_n = \infty$, $a_n \to 0$, $\sum_1^\infty a_nU_n^2 < \infty$ a.s. and since ρ is arbitrary we conclude

$$D_n(\omega) \to 0$$

by Lemma 2 of Derman and Sacks (1958).

(ii) R-M. If we notice that A1 is equivalent to

$$\inf_{\varepsilon<|x-\theta|<\infty} \frac{\{M(x)-M(\theta)\}}{(x-\theta)} > 0$$

then the proof follows almost exactly as for (i).

6.4 Relation with Other Recent Work

It seems appropriate here to comment on recent work of Ljung (1978a) and Kushner (1977, 1978a, 1978b) who have emphasized the relation of the behaviour of SA schemes to the behaviour of differential equations and have also presented analyses dealing with dependent noise. We begin by sketching the conclusions of these works in terms of the present approach and then discuss their approaches.

Return to the basic expression (6.4) and recall for the R-M scheme

$$M_n = M(X_n) - \alpha \tag{6.19}$$

while for the K-W scheme

$$M_n = M(X_n-c_n) - M(X_n+c_n) . \tag{6.20}$$

Let, for example, condition B1' be strengthened by supposing $M(x)$ to be differentiable and $M'(x)$ to obey a Lipschitz condition; then with (6.20) in mind

$$\left|\tfrac{1}{2}c^{-1}\big(M(x-c)-M(x+c)\big)-M'(x)\right| \le K_3c$$

for some suitable constant K_3 . Now (6.4) can be written

$$
\begin{aligned}
X_{n+1} &= X_n - a_n c_n^{-1} M_n + a_n c_n^{-1} Z_n \\
&= X_n - a_n\left(c_n^{-1} M_n - M'(X_n)\right) - a_n M'(X_n) + a_n c_n^{-1} Z_n \\
&= X_n - a_n f(X_n) + a_n \beta_n + a_n c_n^{-1} Z_n
\end{aligned}
\tag{6.21}
$$

where

$$f(x) = M'(x) ,$$

$$\beta_n = c_n^{-1} M_n - M'(X_n)$$

so that $|\beta_n| \le K_3 c_n$.

A vector version of (6.21) is the basic scheme considered by Kushner (1977) and Ljung (1978a). Let X_n, $f(\cdot)$, Z_n, β_n be p-vectors and $f(\cdot)$ to obey a Lipschitz condition with constant L .

We introduce the following conditions.

C1. There exists a twice differentiable function $V(x) > 0$, $x \neq 0$, with $\dot{V}'(x)f(x) > 0$ where $\dot{V}(x) = dV/dx$ outside a set

$$D_s = \{x \mid \dot{V}'(x)f(x) = 0\} .$$

Also for some constant K ,

$$\sup_x \|\ddot{V}(x)\| \le K < \infty ,$$

$$\|\dot{V}(x)\|^2 \le KV(x) , \quad \forall x .$$

This condition might be compared with B1' with $D_s = \{0\}$, $V(x) = x^2$; *cf*. also Blum (1954, conditions A_2-A_5).

C2. There exist constants K_1, K_2 $\ni$ $K_1 x'x \le x'f(x) \le K_2 x'x$.

Now define a p-vector sequence U_n ,

$$U_{n+1} = U_n - a_n B U_n + a_n c_n^{-1} Z_n , \quad U_0 = 0 ,$$

where $B = df/dx'|_{x=0}$.

Now it follows from the discussion at the end of Chapter 5 that if B has all its eigenvalues positive we can certainly show $U_n \xrightarrow{p} 0$ as in Theorem 1. To conclude $U_n \to 0$ a.s. as in Theorem 2 we need a vector

Kronecker's Lemma which states

$$\sum_1^\infty a_s c_s^{-1} Z_s < \infty \quad \text{a.s.} \Rightarrow \pi_{n+1} \sum_1^n \pi_{s+1}^{-1} a_s c_s^{-1} Z_s \to 0$$

where $\pi_n = \prod_1^n (I - a_s B)$. To see this holds (when B is a stability matrix) put

$$g_n = \sum_1^n a_s c_s^{-1} Z_s \to g = \sum_1^\infty a_s c_s^{-1} Z_s$$

and consider

$$\begin{aligned}
\pi_{n+1} \sum_1^n \pi_{s+1}^{-1} a_s c_s^{-1} Z_s &= \pi_{n+1} \sum_1^n \pi_{s+1}^{-1} (g_{s+1} - g_s) \\
&= g_{n+1} - \sum_1^n \pi_{n+1} \left(\pi_{s+1}^{-1} - \pi_s^{-1} \right) g_s \\
&= g_{n+1} - \pi_{n+1} \sum_1^n \pi_{s+1}^{-1} a_s g_s \\
&\to g - g = 0
\end{aligned}$$

by the vector FVT established at the end of Chapter 5.

Subtract the U_n equation from the vector version of (6.21) to obtain

$$\begin{aligned}
\hat{X}_{n+1} &= X_{n+1} - U_{n+1} \\
&= \hat{X}_n - a_n f(\hat{X}_n) + a_n \beta_n + a_n U_n^* \qquad (6.22)
\end{aligned}$$

where

$$U_n^* = BU_n + f(\hat{X}_n) - f(X_n) .$$

Notice that the Lipschitz condition ensures

$$\|U_n\| \le (\|B\| + K_4) \|U_n\|$$

for some constant K_4 : thus $U_n^* \to 0$.

Now as pointed out in Chapter 1 the "smoothness" of the sample paths of U_n^* allows (6.22) to be treated naturally, more or less deterministically, that is realisation-wise (as has indeed been done in Theorem 8). Thus, once more introducing the shrinking time scale $\tau_n = \sum_1^n a_k$ we may write (6.22) as

$$d\hat{X}_{n+1} = -f(\hat{X}_n)d\tau_n + (\beta_n+U_n^*)d\tau_n \tag{6.22a}$$

with $d\tau_n = \tau_n - \tau_{n-1}$ and so on. It seems reasonable then to expect the convergence behaviour of X_n to be related to the asymptotic stability properties of the ordinary differential equation (o.d.e.)

$$d\hat{X}(\tau)/d\tau = -f(\hat{X}(\tau)) \tag{6.23}$$

when perturbed by decaying forcing functions $\beta(\tau) \to 0$ or $U(\tau) \to 0$. Such stability is often analysed by means of a Lyapunov function $V(x) > 0$, $x \neq 0$ as follows. Now

$$dV(\hat{X}(\tau))/d\tau = dV_\tau/d\tau = -dV/d\hat{X}(\tau')f(\hat{X}(\tau)) .$$

Suppose the right hand side vanishes at $x = 0$ and is otherwise negative then V_τ is a strictly decreasing function and so converges to zero: thus $\hat{X}(\tau) \to 0$. However it is usually difficult to find such a $V(x)$; it is more usual to be able to find a $V(x)$ satisfying C1 (often obtained from investigating an approximation to the o.d.e. of interest). The conclusion under C1 is that $\hat{x}(\tau) \to M$, the largest invariant set contained in D_s (see Hahn, 1967, p. 108). An invariant set I is one such that if $\hat{x}(0) \in I$ then $\hat{x}(\tau) \in I$, $\forall\tau$. It may then be possible, by geometric argument to show $M = \{0\}$ (La Salle and Lefschetz, 1962).

With this in mind we obtain a result similar to Theorem 6 as follows:

Under conditions A, B, C, C2, $X_n \xrightarrow{p} 0$. Condition C2 is close to the requirement that (6.23) be uniformly asymptotically stable. The above result may be compared with that of Kushner (1977) who demonstrates convergence in probability to an invariant set of the o.d.e. (6.23). However Kushner's point of view is rather different since his results are obtained under the assumption

$$\sup_n \|X_n\| < \infty \quad \text{a.s.} \tag{6.24}$$

The point of this assumption consists in the interesting observation that, for schemes with dependent noise, it allows the result

$$X_{n+k_n(t)} - X_n \Rightarrow X(t)$$

where $X(t)$ is the solution to the equation

$$X(t) = -\int_0^t f\big(X(s)\big)ds \ .$$

However Kushner gives no conditions under which (6.24) might hold. It could be made to hold if the SA scheme were monitored though no analysis of such a monitored scheme is given.

So far as a.s. convergence is concerned we can obtain an analogue of Theorem 8 as: Under conditions A', B, C, C1 and $\sum a_n c_n^{-2} < \infty$, $X_n \to M$ a.s. The proof is similar to that of Theorem 8 if we note the following three points.

(i) Instead of the inequalities between D_{n+1} and D_n we use the Taylor series

$$\begin{aligned} V_{n+1} = V\big(\hat{X}_{n+1}\big) &= V\big(\hat{X}_n\big) + \big(\hat{X}_{n+1}-\hat{X}_n\big)'\dot{V}\big(\hat{X}_n\big) + \tfrac{1}{2}\big(\hat{X}_{n+1}-\hat{X}_n\big)'\ddot{V}\big(X_n^\bullet\big)\big(\hat{X}_{n+1}-\hat{X}_n\big) \\ &\le V_n - a_n\big(U_n^*+f\big(\hat{X}_n\big)\big)'\dot{V}\big(\hat{X}_n\big) + O\big(a_n^2\big) \end{aligned}$$

where U_n^* has been redefined as $U_n^* + \beta_n$.

(ii) We define $\gamma(\rho)$ by

$$\gamma(\rho) = \inf_{X \ni d(X,M)>\rho} f'(X)\dot{V}(X)/V(X)$$

where

$$d(X, M) = \inf_{\theta \in M} \|X-\theta\| \ .$$

(iii) Finally we observe that

$$a_n U_n^{*\prime}\dot{V}\big(\hat{X}_n\big) \le a_n U_n^{*\prime} U_n^*\big(\tfrac{1}{2}\gamma(\rho)/K\big)^{-1} + \tfrac{1}{2}\gamma(\rho)a_n V\big(\hat{X}_n\big) \ .$$

This result should be compared with that of Ljung (1978a) who similarly proceeds with an observation that the assumption $\underline{\lim}\, \|X_n\| < \infty$ allows a.s. convergence to be proved for schemes with dependent noise. However Ljung goes on to give conditions under which this assumption is satisfied. His conditions are similar to C1. above though $\|\dot{V}(X)\|^2 \le KV(X)$ is not required. However Ljung stresses the fact that his method of proof handles more complex cases; especially schemes where the statistics of the noise sequence depends on the history of the X_k process. The present approach depends on having the noise sequence statistics not dependent in such a fashion and it is not clear to what extent the present approach may be modified to handle such situations.

CHAPTER 7

SECOND ORDER ASYMPTOTIC PROPERTIES OF RML_2

7.1 General

This chapter is concerned with proving a Central Limit Theorem (CLT) and Invariance Principle (IP) for the RML_2 recursion of Söderström *et al.* (1974) (defined below). In Section 2 we review the convergence behaviour of this algorithm based on the work of Ljung (1977b) and Hannan (1978c). Section 3 contains the CLT for RML_2 which is seen to be asymptotically efficient. The approach in Section 3 is to subtract off the part that obeys the CLT and show the remainder is asymptotically negligible; however the ensuing argument though straightforward is by no means as simple as for the SA of Chapters 5 and 6. Finally in Section 4 we derive also an IP where again the argument is rather more subtle than for SA. In the remainder of this section we define RML_2 and introduce some assumptions.

Consider then, an ARMAX model for a scalar time series with stationary exogenous sequence or input sequence, u_n and output or measured sequence y_n. Following Ljung (1976) (*cf.* also Chapters 1, 2) the model is defined by a formula for its innovations or prediction error sequence as

$$e_n(\theta) = \big(1+\nu(L)\big)^{-1}\Big(\big(1+\alpha(L)\big)y_n-\gamma(L)u_n\Big) \tag{7.1}$$

where

$$\nu(L) = \sum_{1}^{n_\nu} \nu_i L^i$$

and so on, and L is the lag or backwards operator; also $\theta = (\alpha', \nu', \gamma')'$ with $\alpha = \big(\alpha_1 \dots \alpha_{n_\alpha}\big)'$ and so on, while $\theta \in R$ a compact subset of the open set

$S = \{\theta \mid 1+\alpha(L),\ 1+\nu(L)$ have all zeroes strictly outside the unit circle in the complex L plane$\}$.

(The above notation differs a little from that used in Chapters 1, 2, thus $\alpha(L)$ is used here rather than $A(L)$. This is so as to make it comparable with that in Hannan (1978c): this last reference will be referred to repeatedly and will be denoted by (H).)

It is assumed there is a true value $\theta_0 \in R \ni e_n(\theta_0)$, denoted ε_n, is a martingale difference sequence w.r.t. F_n the increasing sequence of σ-algebras, generated by ε_n : ε_n is also then called a nonlinear innovations process so that the best least squares predictor of y_n from its own past is a linear predictor. This is a minimal assumption if the linear model is to make sense: *cf.* Hannan and Heyde (1972). Denote $e_n(\theta_0)$ by ε_n and assume that

$$E\left(\varepsilon_n^2 | F_{n-1}\right) = \sigma^2 . \tag{7.2}$$

Assume also $\exists$ a small $\delta > 0$ with

$$\sup_n E\left(|\varepsilon_n|^{4/(1-3\delta)}\right) < K < \infty , \tag{7.3a}$$

$$\sup_n E\left(|u_n|^{4/(1-3\delta)}\right) < K < \infty , \tag{7.3b}$$

where u_n is taken to be a stationary ergodic stochastic process.

Recalling the view of the construction of recursions expressed in Chapter 1, namely by sequential minimisation of an approximate likelihood such as

$$N^{-1} \sum_1^N e_n^2(\theta)$$

we can understand the RML2 recursion as a Prediction Error Recursion (PER) that utilises the gradient vector

$$\begin{aligned}\tilde{\varphi}_n(\theta) &= -de_n(\theta)/d\theta \\ &= -\left(1+\nu(L)\right)^{-1}\varphi_n(\theta) .\end{aligned}$$

This gradient has three components

$$\left(-\tilde{y}_{n-1}(\theta) \ \ldots \ -\tilde{y}_{n-n_\alpha}(\theta)\tilde{e}_{n-1}(\theta) \ \ldots \ \tilde{e}_{n-n_\nu}(\theta)\tilde{u}_{n-1}(\theta) \ \ldots \ \tilde{u}_{n-n_\gamma}(\theta)\right)$$

that obey

$$\left(1+\nu(L)\right)\tilde{y}_n(\theta) = y_n , \tag{7.4a}$$

$$\left(1+\nu(L)\right)\tilde{e}_n(\theta) = e_n(\theta) , \tag{7.4b}$$

$$\big(1+\nu(L)\big)\tilde{u}_n(\theta) = u_n \; . \tag{7.4c}$$

Notice that

$$\big(1+\alpha_0(L)\big)\tilde{y}_n\big(\theta_0\big) = \varepsilon_n + \gamma_0(L)\tilde{u}_n\big(\theta_0\big) \; . \tag{7.4d}$$

We will need to know that stationarity and ergodicity (which we assume for ε_n) ensure that the limit

$$\mathrm{R} = \lim n^{-1}\mathrm{P}_n^{-1}\big(\theta_0\big) = \lim n^{-1}\sum_1^n \tilde{\varphi}_s\big(\theta_0\big)\tilde{\varphi}_s'\big(\theta_0\big) \tag{7.5}$$

exists w.p.1.

The recursion will be then

$$\theta_n = \theta_{n-1} + \mathrm{P}_n\tilde{\varphi}_n e_n \; , \tag{7.6}$$

$$\mathrm{P}_n^{-1} = \mathrm{P}_{n-1}^{-1} + \tilde{\varphi}_n\tilde{\varphi}_n' \; , \tag{7.7}$$

$$e_n = y_n - \varphi_n'\theta_{n-1} \tag{7.8}$$

where

$$\varphi_n = \big(-y_{n-1} \cdots -y_{n-n_\alpha}\; e_{n-1} \cdots e_{n-n_\nu}\; u_{n-1} \cdots u_{n-n_\gamma}\big) \tag{7.9}$$

while $\tilde{\varphi}_n$ has components $\tilde{y}_n$, $\tilde{e}_n$, $\tilde{u}_n$ that obey

$$\big(1+\nu_{n-1}(L)\big)\tilde{y}_n = y_n \; , \tag{7.10a}$$

$$\big(1+\nu_{n-1}(L)\big)\tilde{e}_n = e_n \; , \tag{7.10b}$$

$$\big(1+\nu_{n-1}(L)\big)\tilde{u}_n = u_n \tag{7.10c}$$

(with $\nu_{n-1}(L) = \sum_1^{n_\nu} \nu_{n-1,i}L^i$ and so on).

With (7.4d) in mind, and for an ARMA model (u_n absent) (7.10a) is replaced in (H) by

$$\big(1+\alpha_{n-1}(L)\big)\tilde{y}_n = e_n \; . \tag{7.10a)'}$$

Actually (H) only explicitly derives results for an ARMA case but it is pointed out there that the same results will go through for the ARMAX case (with an equivalent of (7.4d) replacing (7.10a) once more).

7.2 A Review of the Convergence Behaviour of RML_2

The RML_2 recursion as given may not converge and must be monitored. Two schemes have been suggested, one used in (H) and the other in Ljung (1977b). This latter scheme is similar to that used by Albert and Gardner (1967) and also Nevel'son and Khasminskii (1973/1976; Chapter 7). Only the first scheme is discussed here.

The region of stability is separated into an inner region R_2 and an outer region R_1 with

$$R_2 \subset R_1 \subset R \, .$$

Here R_1 is such that $\forall\theta \in R_1$, $1 + \alpha(L)$, $1 + \beta(L)$ have no zeroes for $|L| \le 1 + \eta$, $\eta > 0$. Also any pair of zeroes one from $1 + \alpha(L)$, one from $1 + \beta(L)$ are at least η apart in the complex L plane. R_2 is described in a moment.

Denote the output of the modified recursion by $\tilde{\theta}_n$.

Let $\bar{\theta}$ be a fixed value inside R_2. Define an auxiliary sequence $\bar{\theta}_n$ which is $\tilde{\theta}_n$ until $\tilde{\theta}_n$ first exits from R_1 and is $\bar{\theta}$ until it returns to R_2 when $\bar{\theta}_n = \tilde{\theta}_n$ again and so on. R_2 will be chosen to ensure the event $\{\tilde{\theta}_n$ remains outside R_1 indefinitely long$\}$ has zero probability. The recursion is now

$$\tilde{\theta}_n = \tilde{\theta}_{n-1} + P_n\tilde{\varphi}_n e_n \, , \tag{7.11a}$$

$$P_n^{-1} = P_{n-1}^{-1} + \tilde{\varphi}_n\tilde{\varphi}_n' \, , \tag{7.11b}$$

$$e_n = y_n - \varphi_n'\tilde{\theta}_{n-1} \, . \tag{7.11c}$$

However e_n in (7.9), (7.10) is replaced by

$$\bar{e}_n = y_n - \varphi_n'\bar{\theta}_{n-1}$$

and $\nu_{n-1}(L)$ in (7.10) is $\bar{\nu}_{n-1}(L)$ and so on.

Subtract θ_0 from (7.11a), multiply through by P_n^{-1} and sum up to obtain

$$\tilde{\theta}_n - \theta_0 = P_n^{-1} \sum_1^n \tilde{\varphi}_s \left(e_s - \tilde{\varphi}_s' \left(\tilde{\theta}_{s-1} - \theta_0\right)\right) .$$

If we notice the Taylor series

$$e_n(\theta) = \varepsilon_n + \tilde{\varphi}_n(\theta^{\bullet})'\left(\theta-\theta_0\right)$$

then it is intuitively clear that we can expect ultimately to be dealing with an expression like

$$P_n^{-1} \sum_1^n \tilde{\varphi}_s \varepsilon_s$$

and so expect convergence and asymptotic efficiency. In any case, by repeated use of the martingale convergence theorem (in the form presented by Neveu (1975, p. 150) *cf.* also Chapter 4, Lemma 3') Hannan (1976) is able to evaluate $\tilde{\theta}_n$ as

$$\tilde{\theta}_n = \hat{\theta}_n + o_n(1) \quad \text{a.s.} \tag{7.12}$$

where

$$\hat{\theta}_n = n^{-1} K_n^{-1} \sum_1^n a\left(\overline{\theta}_{s-1}\right) \tag{7.13}$$

where

$$K_n = n^{-1} \sum_1^n G\left(\overline{\theta}_{s-1}\right) \tag{7.14}$$

with

$$G(\theta) = E\left(\tilde{\varphi}_n(\theta)\tilde{\varphi}_n'(\theta)\right)$$

(in (H) this is denoted by $K(\theta)$) and

$$a(\theta) = E\left(\tilde{\varphi}_n(\theta)\left(e_n(\theta) - \tilde{\varphi}_n'(\theta)\theta\right)\right) \tag{7.15}$$

(*cf.* expression (2.14) of (H) where this is written in a different form; recall also that from (7.11a)

$$\tilde{\theta}_n = P_n^{-1} \sum_1^n \tilde{\varphi}_s \left(e_s - \tilde{\varphi}_s' \tilde{\theta}_{s-1}\right)).$$

We reorganise these expressions as follows. Write

$$a(\theta) = f(\theta) - G(\theta)\theta$$

with

$$f(\theta) = E\left(\tilde{\varphi}_n(\theta)e_n(\theta)\right) .$$

Now we may write (7.13) as

$$\hat{\theta}_n = \hat{\theta}_{n-1} + n^{-1}K_n^{-1}\left(a\left(\bar{\theta}_{n-1}\right)+G\left(\bar{\theta}_{n-1}\right)\hat{\theta}_{n-1}\right) .$$

Once it is known that the recursion converges so that $\tilde{\theta}_n = \bar{\theta}_n + o_n(1)$ then this expression may be evaluated to $o_n(1)$ a.s. as

$$\hat{\theta}_n = \hat{\theta}_{n-1} + n^{-1}K_n^{-1}f\left(\bar{\theta}_{n-1}\right) . \tag{7.16}$$

Also (7.14) may be reorganised as

$$K_n - K_{n-1} = n^{-1}\left(G\left(\bar{\theta}_{n-1}\right)-K_{n-1}\right) . \tag{7.17}$$

Now setting $\tau_n = \sum_1^n s^{-1}$ and making the notational transformation $\hat{\theta}_n \leftrightarrow \theta(\tau)$ we are led to consider the behaviour of the ordinary differential equations

$$d\theta(\tau)/d\tau = R^{-1}(\tau)f\left(\theta(\tau)\right) ,$$

$$dR(\tau)/d\tau = G\left(\theta(\tau)\right) - R(\tau) ,$$

which are, of course, the pair presented by Ljung (1977b): see also Söderström *et al.* (1974).

The monitoring of the recursion enables the following conclusion to be drawn

PROPOSITION 1 (P1). $\underline{\lim}\, n^{-1}P_n > 0$ *a.s.*

Now we can see how the region R_2 is defined. Suppose $\tilde{\theta}_n$ remains outside of R_2 indefinitely long then clearly

$$\hat{\theta}_n \to \hat{\theta} = G^{-1}(\bar{\theta})a(\bar{\theta})$$

$$= G^{-1}(\bar{\theta})f(\bar{\theta}) + \theta .$$

Recall

$$f(\theta) = E\left(\tilde{\varphi}_n(\theta)e_n(\theta)\right) .$$

Straightforward manipulation shows the 'pseudo' Taylor series

$$e_n(\theta) = \varepsilon_n - \varphi_n'\left(\theta, \theta_0\right)\left(\theta-\theta_0\right)$$

where

$$\varphi_n\left(\theta, \theta_0\right) = \left(1+\nu_0(L)\right)^{-1}\varphi_n(\theta) .$$

Thus

$$f(\theta) = -L\left(\theta, \theta_0\right)\left(\theta-\theta_0\right) \tag{7.18}$$

where

$$L\left(\theta, \theta_0\right) = E\left(\tilde{\varphi}_n(\theta)\varphi_n'\left(\theta, \theta_0\right)\right)$$

so that

$$\hat{\theta} - \theta_0 = \left\{I-G^{-1}(\bar{\theta})L\left(\bar{\theta}, \theta_0\right)\right\}\left(\bar{\theta}-\theta_0\right)$$

($cf.$(H) expression (2.16); also expression (2.4) for $L\left(\theta, \theta_0\right)$ and the present one are easily seen to be identical).

Now R_2 is defined so as to ensure

$$\left\| I-G^{-1}(\bar{\theta})L\left(\bar{\theta}, \theta_0\right)\right\| < 1$$

so that

$$\|\hat{\theta}-\theta_0\| \leq \|\bar{\theta}-\theta_0\|$$

which is a contradiction.

The following result is obtained.

P2. $\tilde{\theta}_n - \theta_0 \to 0$ *a.s.*

Actually in (H) a stronger result is proved. It is not required that $e_n\left(\theta_0\right)$ be a martingale difference sequence, only that y_n be a stationary process. Then ($cf.$ Chapter 2) θ_0 is defined as a value which minimises $E\left[e_n^2(\theta)\right]$ or else satisfies $E\left(\tilde{\varphi}_n(\theta)e_n(\theta)\right) = 0$.

For the CLT/IP a stronger convergence result is needed namely (Theorem 2 of (H)) when $\sup_n E\left(\varepsilon_n^4\right) < \infty$ then,

P3. $n^{\frac{1}{2}-\delta'}\left(\tilde{\theta}_n-\theta_0\right) \to 0$ *a.s.* $\forall\delta' > 0$.

In particular we will take $\delta' = \delta$ of (7.3a). The result is intuitively clear for the following reason. Consider the quadratic form $V_\tau = \left(\theta(\tau)-\theta_0\right)'R(\tau)\left(\theta(\tau)-\theta_0\right)$ (equivalent to $V_n = \left(\hat{\theta}_n-\theta_0\right)'K_n\left(\hat{\theta}_n-\theta_0\right)$) then it follows from the differential equations below (7.16) and (7.17), that

$$dV_\tau/d\tau = -V_\tau - 2\left(\theta(\tau)-\theta_0\right)'f\left(\theta(\tau)\right) + \left(\theta(\tau)-\theta_0\right)'G\left(\theta(\tau)\right)\left(\theta(\tau)-\theta_0\right)\right) .$$

Recalling (7.18) we see

$$dV_\tau/d\tau = -V_\tau - \left(\theta(\tau)-\theta_0\right)'\left(L_\tau+L_\tau'-G\left(\theta(\tau)\right)\left(\theta(\tau)-\theta_0\right)$$

with

$$L_\tau = L\left(\theta(\tau),\ \theta_0\right) .$$

Now if $\theta(\tau) \to \theta_0$ a.s. then $L_\tau \to G\left(\theta_0\right)$ and we may expect the second term to be negative for τ large enough: we are left with

$$dV_\tau/d\tau \le -V_\tau$$

or

$$V_n \le V_{n-1}(1 - 1/n) .$$

It clearly follows that $n^{1-\delta'}V_n \to 0$ for any $\delta' > 0$.

It will be naturally necessary in the ensuing discussion to dispose of the difference $\tilde{\theta}_n - \overline{\theta}_n$. It follows from the definition of $\tilde{\theta}_n$ and P2 that

P4. *There exists a random variable* n_0 *with* $n_0 < \infty$ *a.s. and* $\tilde{\theta}_n \equiv \overline{\theta}_n \quad \forall n > n_0$.

We will also need

P5. $\lim n\left(P_n-P_n\left(\theta_0\right)\right) = 0$ *a.s.* or $\lim n^{-1}\left(P_n^{-1} - P_n^{-1}(\theta_0)\right) = 0$ a.s.

Recall from (7.5), $\lim n^{-1}P_n^{-1}\left(\theta_0\right) = R$. The proof of this intuitively reasonable proposition is postponed.

The following rate of convergence result is also needed subsequently.

P6. *Under condition* (7.3a),

$$n^{\frac{1}{2}+\delta}\left(\bar{\theta}_n-\bar{\theta}_{n-1}\right) \to 0 \quad a.s.$$

Proof. Now

$$\tilde{\theta}_n - \tilde{\theta}_{n-1} = nP_n n^{-1}\tilde{\varphi}_n\left(y_n-\varphi_n'\tilde{\theta}_{n-1}\right) .$$

In view of P4, P5 and expression (7.9) it is enough to show

$$n^{\frac{1}{2}+\delta}n^{-1}\tilde{\varphi}_n\left(y_n-\varphi_n'\bar{\theta}_{n-1}\right) = n^{-\frac{1}{2}+\delta}\tilde{\varphi}_n\bar{e}_n = n^{-\frac{1}{2}+\delta}\tilde{\varphi}_n\varepsilon_n + n^{-\frac{1}{2}+\delta}\tilde{\varphi}_n\left(\bar{e}_n-\varepsilon_n\right) \to 0 \quad \text{a.s.}$$

In Appendix 7A ((i), (ii)) it is shown that for some $\lambda < 1$,

$$\left|\bar{e}_n-\varepsilon_n\right| \le \sum_1^n \lambda^{n-s}\varphi_s s^{-\frac{1}{2}+\delta}o_s(1)$$

where $\varphi_s = \|\varphi_s\|$. Now the second term above has modulus $\left(\tilde{\varphi}_n = \|\tilde{\varphi}_n\|\right)$

$$\tilde{\varphi}_n n^{-\frac{1}{2}+\delta}\left|\bar{e}_n-\varepsilon_n\right| .$$

Next the fact that $\sup_n E\left(\varphi_n^{4/(1-3\delta)}\right) < \infty$ (Appendix 7A (iii)) ensures by the Bienamé-Chebyshev inequality and then the Borel-Cantelli Lemma† that

$$\varphi_n n^{-\frac{1}{4}+\delta/2} \to 0 \quad \text{a.s.} \qquad (7.~~18~~19a)$$

Next the discrete L'Hospital rule (Bromwich 1947, p. 404) implies

$$\lim \left|\bar{e}_n-\varepsilon_n\right| = \lim \varphi_n n^{-\frac{1}{2}+\delta}o_n(1)/(1-\lambda) = 0 \quad \text{a.s.}$$

(For future reference we note in passing that clearly $\lim(\ln n)\left|\bar{e}_n-\varepsilon_n\right| = 0$ a.s.) Thus we have to show

$$\lim \tilde{\varphi}_n n^{-\frac{1}{2}+\delta} = 0 \quad \text{a.s.} \qquad (7.~~19~~19b)$$

and

$$n^{-\frac{1}{2}+\delta}\tilde{\varphi}_n\left|\varepsilon_n\right| \to 0 \quad \text{a.s.}$$

---

† This sequence of argument will be used repeatedly and denoted BC^2 .

This second will follow from

$$\tilde{\varphi}_n n^{-\frac{1}{4}+\delta/2} \to 0 \quad \text{a.s.,} \tag{7.20}$$

$$|\varepsilon_n| n^{-\frac{1}{4}+\delta/2} \to 0 \quad \text{a.s.} \tag{7.21}$$

Now (7.21) follows from $\sup_n E|\varepsilon_n|^{4/(1-3\delta)} < \infty$ and BC^2 .

Consider one component of $\tilde{\varphi}_n$ say $\tilde{e}_n$ with

$$\left(1+\bar{\nu}_{n-1}(L)\right)\tilde{e}_n = \bar{e}_n \ . \tag{7.10b}$$

Recalling the definition of $\bar{\theta}_n$ (that is such as to ensure $\|F(\bar{\nu}_{n-1})\| < \lambda$ where $F(\nu)$, λ are defined in Appendix 7A) it follows exactly as in Appendix 7A (ii) that

$$|\tilde{e}_n| \le \sum_1^n \lambda^{n-k}|\bar{e}_k| \ .$$

Now the discrete L'Hospital rule implies

$$\begin{aligned} \lim|\tilde{e}_n|n^{-\frac{1}{4}+\delta/2} &\le \lim|\bar{e}_n|n^{-\frac{1}{4}+\delta/2}/(1-\lambda) \\ &= 0 \quad \text{a.s.} \end{aligned}$$

in view of (7.19a), (7.19b) and the remarks following (7.11). Collecting these results together (including case (7.10a)') we conclude (7.20)

$$\lim \tilde{\varphi}_n n^{-\frac{1}{4}+\delta/2} = 0 \quad \text{a.s.}$$

The proof of P5 is now given. The norm of the difference in P5 is bounded by

$$2N^{-1}\sum_1^N \tilde{\varphi}_n^0\delta_n + N^{-1}\sum_1^N \delta_n^2 \le 2\left(N^{-1}\sum_1^N \left(\tilde{\varphi}_n^0\right)^2 N^{-1}\sum_1^N \delta_n^2\right)^{\frac{1}{2}} + N^{-1}\sum_1^N \delta_n^2$$

where

$$\tilde{\varphi}_n^0 = \|\tilde{\varphi}_n(\theta_0)\| \ , \quad \delta_n = \|\tilde{\varphi}_n - \tilde{\varphi}_n'(\theta_0)\| \ .$$

If we show $\lim \delta_n = 0$ then, by the discrete L'Hospital rule we can conclude $\lim N^{-1}\sum_1^N \delta_n^2 = 0$. Consider one component of $\tilde{\varphi}_n$ say $\tilde{e}_n$ with

$$\left(1+\bar{\nu}_{n-1}(L)\right)\tilde{e}_n = \bar{e}_n ;$$

also

$$\left(1+\nu_0(L)\right)\tilde{e}_n(\theta_0) = \varepsilon_n .$$

Thus

$$\left(1+\bar{\nu}_{n-1}(L)\right)\left(\tilde{e}_n-\tilde{e}_n(\theta_0)\right) = \bar{e}_n - \varepsilon_n + \left(\nu_0(L)-\bar{\nu}_{n-1}(L)\right)\tilde{e}_n(\theta_0) .$$

Once more following Appendix 7A (ii) and P3,

$$|\tilde{e}_n-\tilde{e}_n(\theta_0)| \le \delta_{1n} + \delta_{2n}$$

with

$$\delta_{1n} = \sum_1^n \lambda^{n-s}|\tilde{e}_s(\theta_0)|s^{-\frac{1}{2}+\delta}o_s(1) ,$$

$$\delta_{2n} = \sum_1^n \lambda^{n-s}|\bar{e}_s-\varepsilon_s| .$$

Now by the discrete L'Hospital rule

$$\lim \delta_{1n} = \lim|\tilde{e}_n(\theta_0)|n^{-\frac{1}{2}+\delta}o_n(1) = 0$$

which holds via $\sup_n E|\tilde{e}_n(\theta_0)|^{4/(1-3\delta)} < \infty$ and BC^2. A similar argument disposes of the term δ_{2n}. For future reference we note in passing that $\delta_n \ln n \to 0$. The argument is easily repeated for the other components of $\tilde{\varphi}_n$ including case (7.10a)'.

7.3 A Central Limit Theorem for RML_2

The following proof is straightforward though tedious. We begin by summing the recursion to obtain

$$\tilde{\theta}_n - \theta_0 = P_n \sum_1^n \tilde{\varphi}_s\left(\bar{e}_s+\tilde{\varphi}_s'\left(\tilde{\theta}_{s-1}-\theta_0\right)\right) .$$

Thus

$$\sqrt{n}\left(\tilde{\theta}_n-\theta_0\right) = \sqrt{n}\ P_n \sum_1^n \tilde{\varphi}_s\left(\bar{e}_s-\varepsilon_s+\tilde{\varphi}_s'\left(\tilde{\theta}_{s-1}-\theta_0\right)\right) + Q_n$$

with

$$Q_n = \sqrt{n}\ P_n \sum_1^n \tilde{\varphi}_s\varepsilon_s \ . \tag{7.22}$$

It will do to show (a), (b):

(a) $Q_n \xrightarrow{\mathcal{D}} N\left(0,\ R^{-1}\right)$

where R was defined in equation (7.5). In view of P5 this may be replaced by

(a') $n^{-\frac{1}{2}} \sum_1^n \lambda'\tilde{\varphi}_s\varepsilon_s \xrightarrow{\mathcal{D}} N(0,\ \lambda' R\lambda)$

where λ is an arbitrary fixed vector. This follows by MGCLT (Scott, 1974).

Once more by P5 we require

(b) $n^{-\frac{1}{2}} \sum_1^n \tilde{\varphi}_s\left(\bar{e}_s-\varepsilon_s+\tilde{\varphi}_s'\left(\tilde{\theta}_{s-1}-\theta_0\right)\right) \to 0$ a.s.

From P4 this may be replaced by

(b)' $n^{-\frac{1}{2}} \sum_1^n \tilde{\varphi}_s\left(\bar{e}_s-\varepsilon_s+\tilde{\varphi}_s'\left(\bar{\theta}_{s-1}-\theta_0\right)\right) \to 0$ a.s.

We deal firstly with (b)'. From Appendix 7A (i),

$$\left(1+\nu_0(L)\right)\left(\bar{e}_n-\varepsilon_n\right) = -\varphi_n'\left(\bar{\theta}_{n-1}-\theta_0\right) \ . \tag{7.23}$$

Further more it is shown in Appendix 7B that

$$\left(1+\bar{\nu}_{n-1}(L)\right)\left(\tilde{\varphi}_n'\left(\bar{\theta}_{n-1}-\theta_0\right)\right) = \varphi_n'\left(\bar{\theta}_{n-1}-\theta_0\right) + d_n \tag{7.23a}$$

where

$$d_n = O\left(\|\bar{\theta}_n-\bar{\theta}_{n-1}\|\tilde{\varphi}_n\right) \ .$$

It follows then that

$$\left(1+\bar{\nu}_{n-1}(L)\right)\left(\tilde{\varphi}_n'\left(\bar{\theta}_{n-1}-\theta_0\right)+\bar{e}_n-\varepsilon_n\right) = \left(\bar{\nu}_{n-1}(L)-\nu_0(L)\right)\left(\bar{e}_n-\varepsilon_n\right) + d_n . \quad (7.23b)$$

Now in view of P6 and (7.21),

$$|d_n| = o_n(1)n^{-\frac{1}{2}-\delta}\tilde{\varphi}_n .$$

Also it follows as in Appendix 7A (ii) that

$$f_n = |\tilde{\varphi}_n'\left(\bar{\theta}_{n-1}-\theta_0\right)+\bar{e}_n-\varepsilon_n| \le \sum_1^n \lambda^{n-s}|d_s| + \sum_1^n \lambda^{n-s}|\bar{e}_s-\varepsilon_s|o_s(1)s^{-\frac{1}{2}+\delta} .$$

We have to show

$$N^{-\frac{1}{2}}\sum_1^N \tilde{\varphi}_s f_s \to 0 \quad \text{a.s.}$$

The idea is to show

$$\sum_1^\infty \tilde{\varphi}_s f_s s^{-\frac{1}{2}} < \infty \quad \text{a.s.}$$

so that the result will follow from Kronecker's lemma.

Now recalling the notation $\tilde{\varphi}_n^0 = \|\tilde{\varphi}_n(\theta_0)\|$ we have to show

$$\sum_1^\infty \tilde{\varphi}_s^0 f_s s^{-\frac{1}{2}} < \infty , \quad \sum_1^\infty \delta_s f_s s^{-\frac{1}{2}} < \infty .$$

We have already shown $\delta_s \to 0$ so we have to show

$$\sum_1^\infty f_s s^{-\frac{1}{2}} < \infty .$$

So let $\zeta_s = \tilde{\varphi}_s^0$ or 1 ; thus $\sup_s E\left(\zeta_s^4\right) < \infty$. Now consider

$$\sum_1^N \zeta_s f_s s^{-\frac{1}{2}} = \sum_1^N \zeta_s s^{-\frac{1}{2}} \sum_1^s \lambda^{s-t} o_t(1)\tilde{\varphi}_t t^{-\frac{1}{2}-\delta}$$

$$+ \sum_1^N \zeta_s s^{-\frac{1}{2}} \sum_1^s \lambda^{s-t} |\bar{e}_t - \varepsilon_t| o_t(1) \quad . \tag{7.24}$$

The first term may be rewritten as

$$\sum_1^N o_t(1)\tilde{\varphi}_t t^{-\frac{1}{2}-\delta} \sum_{t+1}^N \zeta_s s^{-\frac{1}{2}} \lambda^{s-t} \leq \sum_1^N o_t(1)\tilde{\varphi}_t t^{-1-\delta} \sum_{t+1}^N \zeta_s \lambda^{s-t}$$

$$\leq \sum_1^N o_t(1)\tilde{\varphi}_t t^{-1-\delta} \psi_t \tag{7.25}$$

where

$$\psi_t = \sum_1^\infty \zeta_{s+t} \lambda^s \ .$$

Now the Cauchy-Schwarz inequality implies

$$\sup_t E\left(\psi_t^2\right) \leq \sup_t (1-\lambda)^{-1} \sum_1^\infty \lambda^s E\left(\zeta_{s+t}^2\right)$$

$$\leq (1-\lambda)^{-2} \text{ const.}$$

Also

$$\sup_t E\left(\psi_t^4\right) \leq (1-\lambda)^{-2} \sup_t \sum_1^\infty \lambda^s E\left(\zeta_{s+t}^4\right)$$

$$\leq (1-\lambda)^{-3} \text{ const.} \tag{7.26}$$

Now from (7.25) we need only show

$$\sum_1^\infty \tilde{\varphi}_t t^{-1-\delta} \psi_t < \infty$$

that is

$$\sum_1^\infty \tilde{\varphi}_t^0 \psi_t t^{-1-\delta} < \infty$$

and

$$\sum_1^\infty \delta_t \psi_t t^{-1-\delta} < \infty$$

that is

$$\sum_{1}^{\infty} \psi_t t^{-1-\delta} < \infty .$$

These both follow by the lemma in Lukacs (1975, p. 80) if

$$\sup_t E\left(\tilde{\varphi}_t^0 \psi_t\right) \quad , \quad E\left(\psi_t\right) < \infty ,$$

which they are by Cauchy-Schwarz and (7.26).

The second term in (7.24) can similarly be bounded by an expression such as

$$\sum_{1}^{N} o_t(1)|\bar{e}_t - \varepsilon_t|\tilde{\varphi}_t t^{-1+\delta}\psi_t \tag{7.27}$$

so that we need to show

$$\sum_{1}^{\infty} |\bar{e}_t - \varepsilon_t|\tilde{\varphi}_t t^{-1+\delta}\psi_t < \infty \tag{7.27a}$$

that is

$$\sum_{1}^{\infty} |\bar{e}_t - \varepsilon_t|\zeta_t t^{-1+\delta}\psi_t < \infty$$

with

$$\zeta_t = \tilde{\varphi}_t^0 \quad \text{or} \quad 1 .$$

Recall from Appendix 7A that

$$|\bar{e}_t - \varepsilon_t| \le \sum_{1}^{n} \lambda^{t-s} o_s(1)\varphi_s s^{-\frac{1}{2}+\delta} .$$

Now apply the same argument as led to (7.25) to see (7.27) is bounded by

$$\sum_{1} \varphi_t o_t(1) t^{-3/2+\delta} \psi_t^{\bullet}$$

where

$$\psi_t^{\bullet} = \sum_{1}^{\infty} \zeta_{t+s}\psi_{t+s}\lambda^s .$$

Now

$$\sup_t E\left(\dot{\psi}_t\right)^2 \le (1-\lambda)^{-1} \sup_t \sum_1^\infty \lambda^s E\left[\zeta_{t+s}^2 \psi_{t+s}^2\right]$$

$$\le (1-\lambda)^{-1} \sup_t \sum_1^\infty \lambda^s \sqrt{E}\left(\zeta_{t+s}^4\right) \sqrt{E}\left(\psi_{t+s}^4\right)$$

$$\le (1-\lambda)^{-2} \text{ const.}$$

and the result follows as before.

7.4 An Invariance Principle for RML_2

We turn now to the IP. Denote $\tau_n = \sum_1^n s^{-1}$ and define

$$k_n(t) = \sup\{m : \tau_{n+m} - \tau_n \le t\} .$$

Now (7.22) may be reorganised as

$$dQ_n = \left\{\left(\tfrac{1}{2} + o\left(n^{-1}\right)\right) I - R_n^{-1} \tilde{\varphi}_n \tilde{\varphi}_n'\right\} Q_{n-1} d\tau_n + dW_n$$

where

$$W_n = R_n^{-1} \sum_1^n \tilde{\varphi}_s \varepsilon_s / \sqrt{s} ,$$

$R_n^{-1} = nP_n$ and $d\tau_n = \tau_n - \tau_{n-1}$ and so on. The idea is first to show

$$Q_{n+k_n(t)} \Rightarrow Q(t) \tag{7.28}$$

where for arbitrary fixed vector λ, $Q(t) = \lambda' Q(t)$ is the continuous Gaussian process on $D[0, 1]$ with

$$Q(t) = e^{-\frac{1}{2}t} Q(0) + \int_0^t e^{-\frac{1}{2}(t-s)} dW(s)$$

(*cf.* equation (5.10)) and

$$Q(0) \sim N\left(0, \sigma^2\right)$$

and where $W(t)$ is the Wiener process on $D[0, 1]$ with variance $\sigma^2 = \lambda' R^{-1} \lambda$. Define $\hat{Q}_n = \sqrt{n}\left(\tilde{\theta}_n - \theta_0\right)$. Once (7.28) is established the conclusion

$$\hat{Q}_{n+k_n(t)} \Rightarrow Q(t)$$

follows provided

$$\sup_{0 \le t \le 1} \|Q_{n+k_n(t)} - \hat{Q}_{n+k_n(t)}\| \xrightarrow{p} 0 .$$

This will hold if $\|Q_n - \hat{Q}_n\| \xrightarrow{p} 0$ which is, however, a consequence of the proof of the CLT (that is, of Condition (b) of Section 7.3).

We begin now, as in Chapter 5 by showing

$$W_{n+k_n(t)} - W_n \Rightarrow W(t)$$

where

$$W_n = \lambda' \mathrm{W}_n = \lambda' \mathrm{R}_n^{-1} \sum_1^n \tilde{\varphi}_s \varepsilon_s / \sqrt{s} .$$

Since $\mathrm{R}_n \to \mathrm{R}$ we can replace R_n here by R. Further recalling the remark at the end of Section 7.2 that $\ln n \|\tilde{\varphi}_n - \tilde{\varphi}_n(\theta_0)\| \to 0$ we see that $\tilde{\varphi}_n$ may be replaced by $\tilde{\varphi}_n(\theta_0)$.

Consider the triangular array

$$v_{ns} = \lambda' \mathrm{R}^{-1} \tilde{\varphi}_{n+s}(\theta_0) \varepsilon_{n+s} / \sqrt{n+s}$$

with associated increasing σ-fields $F_{n,s} = F_{n+s}$. We have to show (Scott, 1973)

$$\max_{n \le k \le n+k_n(1)} \left(\lambda' \mathrm{R}^{-1} \tilde{\varphi}_k(\theta_0)\right)^2 / k \xrightarrow{p} 0$$

and this follows from

$$\left(\lambda' \mathrm{R}^{-1} \tilde{\varphi}_k(\theta_0)\right)^2 / k \xrightarrow{p} 0$$

which holds trivially since $\tilde{\varphi}_k(\theta_0)$ has finite variance. Also we need

$$\sum_n^{n+k_n(t)} \left(\lambda' \mathrm{R}^{-1} \tilde{\varphi}_s(\theta_0)\right)^2 / s \xrightarrow{p} \lambda' \mathrm{R}^{-1} \lambda t .$$

This will follow from

$$\sum_{n}^{n+k_n(t)} \left[\left(\lambda' R^{-1} \tilde{\varphi}_s(\theta_0)\right)^2 - \lambda' R^{-1}\lambda\right]/s \xrightarrow{p} 0 \tag{7.29a}$$

which we now establish using a truncation argument similar to that used in Chapter 6: *cf*. Hannan (1973).

Define

$$\eta(n) = \left(\lambda' R^{-1} \tilde{\varphi}_n(\theta_0)\right)^2 - \lambda' R^{-1}\lambda$$

so that $\eta(n)$ is a zero mean stationary ergodic stochastic process. Set

$$\eta_1(n) = \eta(n) I\left(|\eta(n)| \le A\right) ,$$

$$\eta_2(n) = \eta(n) I\left(|\eta(n)| > A\right)$$

so $\eta_1(n) - E\left(\eta_1(n)\right)$ is stationary ergodic. Now

$$E\left\{\sum_{n}^{n+k_n(t)} \left|\eta_1(s) - E\left(\eta_1(s)\right)\right|/s\right\}^2 = \sum_{n}^{n+k_n(t)} \sum_{n}^{n+k_n(t)} s^{-1}k^{-1}c(k-s)$$

where $c(s)$ is the lag s autocovariance of $\eta_1(s) - E\left(\eta_1(s)\right)$. The above expression is

$$\sum_{n}^{n+k_n(t)} c(s) \sum_{l=n+s}^{n+k_n(t)-s} \left(l(l+s)\right)^{-1} \le \max_{n \le s \le n+k_n(t)} |c(s)| \left(\sum_{n}^{n+k_n(t)} s^{-1}\right)^2 \to 0$$

since stationarity and ergodicity ensure $|c(n)| \to 0$ while $\sum_{n}^{n+k_n(t)} s^{-1} \to t$. Next,

$$\left|E \sum_{n}^{n+k_n(t)} \left|\eta_2(s) - E\left(\eta_2(s)\right)\right|/s\right| \le \sum_{n}^{n+k_n(t)} s^{-1} E\left|\eta_2(0) - E\left(\eta_2(0)\right)\right|$$

$$\to tE\left|\eta_2(0) - E\left(\eta_2(0)\right)\right|$$

which may be made arbitrarily small by taking A large.

Now rewrite the expression for dQ_n as

$$Q_n = \left(I - I_n/2n - E_n\right)Q_{n-1} + dW_n$$

where

$$I_n = 2\left(R_n^{-1}R - I\left(\tfrac{1}{2}+o\left(n^{-1}\right)\right)\right) \to I \ ,$$

$$E_n = R_n^{-1}\left(\tilde{\varphi}_n\tilde{\varphi}_n' - R\right)/n \ .$$

Now recalling the derivation of an IP in Section 5.2 and its extension to Multidimensional SA at the end of Chapter 5 we need only show

$$\sup_{0\le t\le 1}\left\| \prod_{n}^{n+k_n(t)} \left(I - I_s/2s - E_s\right) - e^{-It/2}\right\| \xrightarrow{p} 0 \ .$$

Now below it will be shown that

$$\|E_n\| \to 0 \quad \text{a.s.,} \tag{7.29b}$$

$$\sum_{1}^{\infty} \|E_n\|^2 < \infty \quad \text{a.s.} \tag{7.29c}$$

so that

$$\sum_{n}^{n+k_n(t)} \|E_s\|^2 \to 0 \quad \text{a.s.}$$

uniformly in t. It then follows that

$$\exp\left\{ \sum_{n}^{n+k_n(t)} \left[\log\left(I - I_s/2s - E_s\right) + I_s/2s + E_s\right]\right\} \to I \quad \text{a.s.}$$

uniformly in t , so that

$$\sup_{0\le t\le 1}\left\|\exp\left\{ \sum_{n}^{n+k_n(t)} \log\left(I - I_s/2s - E_s\right)\right\} - \exp\left\{ \sum_{n}^{n+k_n(t)} - I_s/2s - E_s\right\}\right\| \to 0 \quad \text{a.s.}$$

However we have already shown (7.29a)

$$\sum_{n}^{n+k_n(t)} E_s \xrightarrow{p} 0$$

uniformly in t so the result now follows. We have then to show (7.29b), (7.29c). Now (7.29c) will clearly follow if we show

$$\sum_{1}^{\infty} \tilde{\varphi}_n'\tilde{\varphi}_n/n^2 < \infty \quad \text{a.s.}$$

However

$$\sum_1^\infty \tilde{\varphi}_n'\tilde{\varphi}_n/n^2 \le \sum_1^\infty \tilde{\varphi}_n' P_n^2 \varphi_n \left(\text{tr } P_n^{-1}/n\right)^2 < \infty \quad \text{a.s.}$$

since $n^{-1}P_n^{-1} \to R$ while as pointed out in Chapter 4, $\sum_1^\infty \tilde{\varphi}_n' P_n^2 \tilde{\varphi}_n < \infty$ a.s. always. Finally $\|E_n\| \to 0$ a.s. if we show $\tilde{\varphi}_n'\tilde{\varphi}_n/n \to 0$ a.s. But this follows from Kronecker's lemma since

$$\lim_{N\to\infty} \sum_1^N \left(\tilde{\varphi}_n'\tilde{\varphi}_n - \tilde{\varphi}_{n-1}'\tilde{\varphi}_{n-1}\right)/n = \lim_{N\to\infty} \sum_1^N \left(\tilde{\varphi}_n'\tilde{\varphi}_n/n - \tilde{\varphi}_{n-1}'\tilde{\varphi}_{n-1}/(n-1)\right) + \sum_1^N \tilde{\varphi}_{n-1}'\tilde{\varphi}_{n-1}/n(n-1)$$

$$\le \lim_{N\to\infty} \tilde{\varphi}_N' P_N \tilde{\varphi}_N \left(\text{tr } P_N^{-1}/N\right) + \sum_1^\infty \tilde{\varphi}_n'\tilde{\varphi}_n/n^2$$

$$\le \text{tr } R + \sum_1^\infty \tilde{\varphi}_n'\tilde{\varphi}_n/n^2 < \infty$$

since $\tilde{\varphi}_N' P_N \tilde{\varphi}_N \le 1$.

APPENDIX 7A

SOME USEFUL RELATIONS AND BOUNDS

(i) Consider that equations (1) and (8) imply

$$\left(1+\bar{\nu}_{n-1}(L)\right)\bar{e}_n = \left(1+\bar{\alpha}_{n-1}(L)\right)y_n - \bar{\gamma}_{n-1}(L)\, u_n,$$

$$\left(1+\nu_0(L)\right)\varepsilon_n = \left(1+\alpha_0(L)\right)y_n - \gamma_0(L)u_n .$$

Thus

$$\left(1+\nu_0(L)\right)\left(\bar{e}_n-\varepsilon_n\right) = \left(\bar{\alpha}_{n-1}(L)-\alpha_0(L)\right)y_n - \left(\bar{\nu}_{n-1}(L)-\nu_0(L)\right)u_n - \left(\bar{\gamma}_{n-1}(L)-\gamma_0(L)\right)u_n$$

$$= -\varphi_n'\left(\bar{\theta}_{n-1}-\theta_0\right) .$$

(ii) We can write

$$\bar{e}_n - \varepsilon_n = -c'x_n$$

with

$$c = (1\ 0\ \dots\ 0)$$

and

$$x_n = F(\nu_0)x_{n-1} + c\varphi_n'(\overline{\theta}_{n-1}-\theta_0)$$

with

$$F(\nu) = \begin{pmatrix} -\nu_1 & 1 & 0 & \dots & 0 \\ -\nu_2 & 0 & & \dots & 0 \\ & & & \ddots & \\ & & & & 1 \\ -\nu_m & 0 & & \dots & 0 \end{pmatrix} .$$

It follows that for some $\lambda < 1$,

$$\|x_n\| \le \lambda\|x_{n-1}\| + \varphi_n\|\overline{\theta}_{n-1}-\theta_0\|$$

so

$$|\overline{e}_n-\varepsilon_n| \le \sum_1^n \lambda^{n-k}\varphi_k o_k(1)k^{-\frac{1}{2}+\delta}$$

(since by P3, P4, $\|\overline{\theta}_n-\theta_0\| = o_n(1)n^{-\frac{1}{2}+\delta}$) with $\varphi_k = \|\varphi_k\|$.

Notice in this last expression initial condition effects have been left out; in this and all subsequent similar expression they decay exponentially and so may be omitted.

(iii) Since $\overline{\theta}_n$ which is used to construct φ_n lies inside R_1 then φ_n is formed by stable filtering operations on y_n, u_n . Thus we must have for some $\lambda < 1$,

$$\varphi_n = \|\varphi_n\| \le C \sum_1^\infty \lambda^i(|\varepsilon_{n-i}|+|u_{n-i}|)$$

so that $\sup_n E\left(\varphi_n^{4/(1-3\delta)}\right) < \infty$.

APPENDIX 7B

DERIVATION OF RELATION (7.23)

The argument is tedious but simple and is therefore illustrated for a MA(2) model only. Now

$$\left(1+\bar{\nu}_{n-1}(L)\right)\left(\tilde{\varphi}_n'\left(\bar{\theta}_{n-1}-\theta_0\right)\right) = \left(1+\bar{\nu}_{1,n-1}L+\bar{\nu}_{2,n-1}L^2\right)\left(\tilde{e}_n\left(\bar{\nu}_{1,n-1}-\nu_1\right)\right)$$
$$+ \left(1+\bar{\nu}_{1,n-1}L+\bar{\nu}_{2,n-1}L^2\right)\left(\tilde{e}_n\left(\bar{\nu}_{2,n-1}-\nu_2\right)\right) .$$

Now the first term is

$$\tilde{e}_n\left(\bar{\nu}_{1,n-1}-\nu_1\right) + \bar{\nu}_{1,n-1}\tilde{e}_{n-1}\left(\bar{\nu}_{1,n-2}-\nu_1\right) + \bar{\nu}_{2,n-1}\tilde{e}_{n-2}\left(\bar{\nu}_{1,n-3}-\nu_1\right)$$
$$= \tilde{e}_n\left(\bar{\nu}_{1,n-1}-\nu_1\right) + \bar{\nu}_{1,n-1}\tilde{e}_{n-1}\left(\bar{\nu}_{1,n-1}-\nu_1\right) + \bar{\nu}_{2,n-1}\tilde{e}_{n-2}\left(\bar{\nu}_{1,n-1}-\nu_1\right)$$
$$+ \bar{\nu}_{1,n-1}\tilde{e}_{n-1}\left(\bar{\nu}_{1,n-2}-\bar{\nu}_{1,n-1}\right) + \bar{\nu}_{2,n-1}\tilde{e}_{n-2}\left(\bar{\nu}_{1,n-3}-\bar{\nu}_{1,n-1}\right)$$
$$= \left(\bar{\nu}_{1,n-1}-\nu_1\right)\left[\left(1+\bar{\nu}_{1,n-1}L+\bar{\nu}_{2,n-1}L^2\right)\tilde{e}_n\right] + O\left(\|\bar{\theta}_n\|\|\tilde{\varphi}_n\|\|\bar{\theta}_n-\bar{\theta}_{n-1}\|\right)$$
$$= \left(\bar{\nu}_{1,n-1}-\nu_1\right)e_n + O\left(\|\tilde{\varphi}_n\|\|\bar{\theta}_n-\bar{\theta}_{n-1}\|\right)$$

in view of (7.10b) and the fact that $\bar{\theta}_n$ is a.s. bounded.

CHAPTER 8

LIMIT LAWS FOR RML_1

8.1 Introduction

This chapter is devoted to a consideration of the convergence and CLT for the RML_1 recursion in two forms (discussed below). We begin once more with the ARMAX model

$$e_n(\theta) = \left(1+\nu(L)\right)^{-1}\left(\left(1+\alpha(L)\right)y_n-\gamma(L)u_n\right) \tag{8.1}$$

where

$$\nu(L) = \sum_1^{n_\nu} \nu_i L^i$$

and so on, with L the lag or backwards operator; also $\theta = (\alpha', \nu', \gamma')$ with $\alpha = \left(\alpha_1 \ldots \alpha_n\right)'$ and so on, while $\theta \in R$ a compact subset of the open set

$S = \{\theta \mid 1+\alpha(L), 1+\nu(L)$ have all zeroes strictly outside the unit circle in the complex L plane$\}$.

It is assumed there is a true value $\theta_0 \in R \ni e_n(\theta_0)$, denoted ε_n, is a stationary ergodic martingale difference sequence or nonlinear innovations sequence w.r.t. the increasing sequence of σ-algebras F_n generated by ε_n. We assume further

$$E\left(\varepsilon_n^2 \middle| F_{n-1}\right) = \sigma^2 \quad \text{(an a.s. constant)}$$

we then have by the ergodicity assumption

$$N^{-1}\sum_1^N \varepsilon_n^2 \to \sigma^2 .$$

The u_k sequence may be deterministic or stationary ergodic stochastic (independent of the ε_n sequence) and is assumed to have finite non zero power, that is

$$\lim N^{-1} \sum_1^N u_k^2 = \sigma_u^2 .$$

Before stating the final condition we notice that (8.1) may be written as a pseudo regression

$$e_n(\theta) = y_n - \varphi_n'(\theta)\theta \tag{8.2}$$

where

$$\varphi_n(\theta) = \left(-y_{n-1} \cdots -y_{n-n_\alpha} e_{n-1}(\theta) \cdots e_{n-n_\nu}(\theta) u_{n-1} \cdots u_{n-n_\gamma}\right) .$$

We assume that the limit

$$\lim N^{-1} \sum_1^N \varphi_k(\theta_0)\varphi_k(\theta_0)' = \mathrm{R}$$

exists and is positive definite.

With the expression (8.2) in mind the definition of the first RML₁ recursion for θ is immediate from Plackett's algorithm as

$$\theta_n = \theta_{n-1} + \mathrm{P}_n \varphi_n e_n , \tag{8.3a}$$

$$\mathrm{P}_n^{-1} = \mathrm{P}_{n-1}^{-1} + \varphi_n \varphi_n' , \tag{8.3b}$$

or

$$\mathrm{P}_n = \mathrm{P}_{n-1} - \mathrm{P}_{n-1}\varphi_n\varphi_n'\mathrm{P}_{n-1}/\left(1+\varphi_n'\mathrm{P}_{n-1}\varphi_n\right) ,$$

$$e_n = y_n - \varphi_n'\theta_{n-1} ,$$

with

$$\varphi_n = \left(-y_{n-1} \cdots -y_{n-n_\alpha} e_{n-1} \cdots e_{n-n_\nu} u_{n-1} \cdots u_{n-n_\gamma}\right) . \tag{8.3c}$$

The second form of the RML₁ recursion is obtained by replacing the prediction errors in the regression vector φ_n by the residuals. The resulting recursion is

$$\theta_n = \theta_{n-1} + \mathrm{P}_n \psi_n e_n , \tag{8.4a}$$

$$\mathrm{P}_n^{-1} = \mathrm{P}_{n-1}^{-1} + \psi_n \psi_n' , \tag{8.4b}$$

$$e_n = y_n - \psi_n'\theta_{n-1} \tag{8.5a}$$

where

$$\psi_n = \left(-y_{n-1} \cdots -y_{n-n_\alpha} \eta_{n-1} \cdots \eta_{n-n_\nu} u_{n-1} \cdots u_{n-n_\gamma}\right)$$

and

$$\begin{aligned}\eta_n &= y_n - \psi_n'\theta_n \\ &= y_n - \psi_n'\theta_{n-1} - \psi_n'P_n\psi_n e_n \\ &= e_n\left(1-\psi_n'P_n\psi_n\right) .\end{aligned} \tag{8.5b}$$

This recursion was presented as one of the examples in the recent article by Moore and Ledwich (1978). We will refer to this recursion as RML_1'.

Turning to consider the convergence of these recursions, we find Ljung (1977a) and Hannan (1976) have provided analyses of RML_1 by requiring the recursions be monitored (with this requirement both analyses would clearly extend to RML_1'). However Hannan (1976) also pointed out that for a MA(1) model, RML_1 converges without the need of monitoring. In a first draft of this chapter the author had tried to find a general proof of convergence of RML_1 *without monitoring* by means of a Stochastic Lyapunov function approach, but had succeeded only in repeating Hannan's result: that work is the subject of Section 8.3. Subsequently the author became aware of the (earlier) work of Moore and Ledwich (1978) in which, *inter alia*, RML_1' is considered also by a Stochastic Lyapunov function approach; they were however unable to complete the proof of convergence providing instead two conditions whose establishment would complete the proof. In any case it immediately became clear that the proof could be completed using the techniques developed to tackle RML_1. With this in mind the analysis of RML_1' which is, in any case easier to follow, has been presented first. Finally in Section 8.4 we discuss briefly the unusual second order behaviour of RML_1: the status of the CLT for RML_1 is uncertain.

Before leaving this section we recall some consequences (and present some new ones) of the Matrix Inversion Lemma (MIL), that will be basic to the ensuing argument. Suppose we have a sequence of p-dimensional vectors x_k and denote

$$M_n = \left(\sum_1^n x_k x_k'\right)^{-1} .$$

The MIL states

$$M_n - M_{n-1} = -M_{n-1}x_n x_n' M_{n-1} / \left(1 + x_n' M_{n-1} x_n\right)$$

so that

$$M_n x_n = M_{n-1} x_n / \left(1 + x_n' M_{n-1} x_n\right)$$

and

$$x_n' M_n x_n = x_n' M_{n-1} x_n / \left(1 + x_n' M_{n-1} x_n\right) \le 1 \ .$$

Also

$$\sum_p^\infty x_n' M_n^2 x_n \le \sum_p^\infty x_n' M_n M_{n-1} x_n$$
$$= \sum_p^\infty \operatorname{tr}\left(M_{n-1} - M_n\right) < \infty \ .$$

Suppose $\overline{\lim}\ \operatorname{tr} M_n^{-1}/n < \infty$ then

$$x_n' x_n / n \to 0 \ ,$$

$$x_n' M_n x_n \to 0 \ ,$$

$$\sum_p^\infty x_n' M_n x_n / n < \infty \ ,$$

$$\sum_p^\infty x_n' x_n / n^2 < \infty \ .$$

To see these results consider that

$$\sum_p^\infty x_n' x_n / n^2 \le \sum_p^\infty \left(x_n' M_n x_n / n\right)\left(\operatorname{tr} M_n^{-1}/n\right)$$
$$\le \sum_p^\infty \left(x_n' M_n^2 x_n\right)\left(\operatorname{tr} M_n^{-1}/n\right)^2 < \infty \ .$$

Next $x_n' M_n x_n \to 0$ follows by Kronecker's lemma since

$$\sum_p^\infty \left(n x_n' M_n x_n - (n-1) x_{n-1}' M_{n-1} x_{n-1}\right)/n = \sum_p^\infty \left(x_n' M_n x_n - \left(1 - n^{-1}\right) x_{n-1}' M_{n-1} x_{n-1}\right)$$
$$\le 1 + \sum_p^\infty x_n' M_n x_n / n < \infty \ .$$

Finally

$$x_n'x_n/n \le x_n'M_nx_n\left(\mathrm{tr}\ M_n^{-1}/n\right) \to 0 \ .$$

The above results will be used repeatedly and referred to as MIL.

8.2 The Convergence of RML$_1'$

Let us begin by noticing that we can rewrite (8.4a) as

$$\theta_n = \theta_{n-1} + P_{n-1}\psi_n\eta_n \ . \tag{8.6}$$

This follows from (8.5b) and the fact that MIL implies

$$\begin{aligned} P_n\psi_n e_n &= P_{n-1}\psi_n e_n/\left(1+\psi_n'P_{n-1}\psi_n\right) \\ &= P_{n-1}\psi_n e_n\left(1-\psi_n'P_n\psi_n\right) \ . \end{aligned}$$

Next reorganise (8.6) as

$$P_{n-1}^{-1}\left(\theta_{n-1}-\theta_0\right) = P_{n-1}^{-1}\left(\theta_n-\theta_0\right) - \psi_n\eta_n \ .$$

Now denoting $T_n = \left(\theta_n-\theta_0\right)'P_n^{-1}\left(\theta_n-\theta_0\right)$ we can find

$$T_{n-1} = T_n - \left(\psi_n'\left(\theta_n-\theta_0\right)\right)^2 - 2\psi_n'\left(\theta_n-\theta_0\right)\eta_n + \psi_n'P_{n-1}\psi_n\eta_n^2$$

and, noting, from MIL,

$$\psi_n'P_{n-1}\psi_n = \psi_n'P_n\psi_n/\left(1-\psi_n'P_n\psi_n\right)$$

we rewrite this as

$$\begin{aligned} T_n = T_{n-1} &+ \psi_n'\left(\theta_n-\theta_0\right)\left[\psi_n'\left(\theta_n-\theta_0\right)+2\left(\eta_n-\varepsilon_n\right)\right] \\ &+ 2\varepsilon_n\left(\psi_n'\left(\theta_{n-1}-\theta_0\right)+\psi_n'P_n\psi_n\left(e_n-\varepsilon_n\right)+\psi_n'P_n\psi_n\varepsilon_n\right) \\ &- \psi_n'P_n\psi_n\left(1-\psi_n'P_n\psi_n\right)\left[\left(e_n-\varepsilon_n\right)^2+2\varepsilon_n\left(e_n-\varepsilon_n\right)+\varepsilon_n^2\right] \ . \end{aligned} \tag{8.7}$$

Next consider that, by definition

$$\left(1+\nu_n(L)\right)\eta_n = y_n - \left(-y_{n-1}\ \cdots\ -y_{n-n_\alpha}\ u_{n-1}\ \cdots\ u_{n-n_\gamma}\right)'\beta_n \ ,$$

$$\left(1+\nu_0(L)\right)\varepsilon_n = y_n - \left(-y_{n-1}\ \cdots\ -y_{n-n_\alpha}\ u_{n-1}\ \cdots\ u_{n-n_\gamma}\right)'\beta_0 \ ,$$

with

$$\beta = (\alpha\gamma)'$$

and so on,

$$\nu_{n-1}(L) = \sum_{1}^{n_\nu} \nu_{i,n-1} L^{i} .$$

Thus

$$\left(1+\nu_0(L)\right)\left(\eta_n - \varepsilon_n\right) = -\psi_n'\left(\theta_n - \theta_0\right) . \tag{8.8}$$

Now denote

$$\hat{u}_n = -\psi_n'\left(\theta_n - \theta_0\right) ,$$

$$\hat{y}_n = \eta_n - \varepsilon_n + \tfrac{1}{2}\psi_n'\left(\theta_n - \theta_0\right) .$$

Thus neglecting initial conditions

$$\hat{y}_n = \left[\left(1+\nu_0(L)\right)^{-1} - \tfrac{1}{2}\right]\hat{u}_n .$$

With this notation, and recognising for example that $e_n - \varepsilon_n$ is F_{n-1} measurable take conditional expectations in (8.7) to obtain

$+ 2\hat{u}_n\hat{y}_n$

$$E\left(T_n \mid F_{n-1}\right) = T_{n-1} \cancel{-2\hat{u}_n\hat{y}_n} + 2\sigma^2\psi_n' P_n \psi_n - \psi_n' P_n \psi_n \left(1-\psi_n' P_n \psi_n\right)^2 \left[\left(e_n - \varepsilon_n\right)^2 + \sigma^2\right]$$

$$\leq T_{n-1} \cancel{-2\hat{u}_n\hat{y}_n} + 2\sigma^2\psi_n' P_n \psi_n . \tag{8.9a}$$

Bearing in mind always that we are aiming to use MGCT* of Chapter 4 let us now give a condition that will ensure

$$C_N = 2\sum_{1}^{N} \hat{u}_n \hat{y}_n \geq 0 \quad \text{a.s.}$$

We require

$$h(L) = \sum_{0}^{\infty} h_s L^s = \left(1+\nu_0(L)\right)^{-1} - \tfrac{1}{2}$$

to be positive real, that is,

$$\text{Re } h\left(e^{i\omega}\right) \geq 0 .$$

Then, neglecting initial conditions

$$C_N = 2 \sum_1^N \hat{u}_s \hat{y}_s = 2 \sum_1^N \hat{u}_s \sum_1^s h_{s-t} \hat{u}_t$$
$$= \pi^{-1} \int_{-\pi}^{\pi} \left| \sum_1^N \hat{u}_s e^{is\omega} \right|^2 h\left(e^{i\omega}\right) d\omega$$
$$= \pi^{-1} \int_{-\pi}^{\pi} \left| \sum_1^N \hat{u}_s e^{is\omega} \right|^2 \operatorname{Re} h\left(e^{i\omega}\right) d\omega \tag{8.10}$$
$$\geq 0 \ .$$

The positive real condition has been associated with RML_1 by Ljung (1977a) and in another form by Hannan (1976). It was also used in the analysis of a related deterministic recursion (Landau, 1976, 1978). Indeed from this last point of view the above conclusion $C_N \geq 0$ can be obtained by another argument that has appeared often in the Engineering literature (Narendra and Kudva, 1974) in connexion with some deterministic recursions related to RML_1'. This other, rather longer approach, via the positive-real lemma is given for completeness, in Appendix 8A. The remainder of the proof can be followed without perusing the appendix. Setting $T_n' = T_n + C_n$ we can reform (8.9a) as

$$E\left(T_n' | F_{n-1}\right) \leq T_{n-1}' + 2\sigma^2 \psi_n' P_n \psi_n \ . \tag{8.9b}$$

An inequality similar to (8.9b) was obtained by Moore and Ledwich (1978, equation (3.7)) by a rather longer argument: an extension of the one used in Landau, (1974, 1978). That argument, on the other hand, has an attractive physical interpretation since it shows how the recursion can be viewed as a feedback system.

Moore and Ledwich (1978) point out that convergence will follow if we can show (their conditions (iv), (v))

$$\sum_0^\infty n^{-\gamma} \psi_n' P_n \psi_n < M \text{ (a constant)} < \infty \ ,$$

$$n^\gamma P_n \to 0 \quad \text{a.s. for some} \quad \gamma > 0 \ .$$

They are however unable to do this. We now complete the proof. Let us divide through (8.9b) by n to obtain

$$E\left(T_n'/n | F_{n-1}\right) \leq T_{n-1}'/(n-1) - n^{-1} T_{n-1}'/(n-1) + 2\sigma^2 \psi_n' P_n \psi_n / n \ . \tag{8.9c}$$

We will now set about establishing $\overline{\lim}\, n^{-1} \sum_1^n \eta_s^2 < \infty$ a.s. Recalling that the diagonal elements of P_n^{-1} are of the form

$$N^{-1} \sum_1^n y_n^2 \,, \quad N^{-1} \sum_1^N u_n^2 \,, \quad N^{-1} \sum_1^N \eta_n^2 \tag{8.11}$$

this will then imply $\overline{\lim}\, N^{-1} \operatorname{tr} P_N^{-1} < \infty$ but then MIL will imply

$$\sum_1^\infty \psi_n' P_n \psi_n / n < \infty \tag{8.12}$$

and we will be able to apply MGCT* to (8.9c). Return, then, to (8.4), form $\theta_n' P_n^{-1} \theta_n$ so that

$$\begin{aligned} \theta_n' P_n^{-1} \theta_n &= \theta_{n-1}' P_{n-1}^{-1} \theta_{n-1} + \left(\psi_n' \theta_{n-1}\right)^2 + 2e_n \psi_n' \theta_{n-1} + \psi_n' P_n \psi_n e_n^2 \\ &= \theta_{n-1}' P_{n-1}^{-1} \theta_{n-1} + y_n^2 - e_n^2 \left(1 - \psi_n' P_n \psi_n\right) . \end{aligned}$$

Summing up yields

$$\theta_n' P_n^{-1} \theta_n + \sum_1^n e_s^2 \left(1 - \psi_s' P_s \psi_s\right) = \sum_1^n y_s^2 \tag{8.13a}$$

or, in view of (8.5b)

$$\theta_n' P_n^{-1} \theta_n + \sum_1^n \eta_s^2 / \left(1 - \psi_s' P_s \psi_s\right) = \sum_1^n y_s^2 \,. \tag{8.13b}$$

This is just the decomposition of total sum of squares into regression sum of squares plus error sum of squares. From (8.13b) we conclude

$$\overline{\lim}\, n^{-1} \sum_1^n \eta_s^2 / \left(1 - \psi_s' P_s \psi_s\right) < \infty \quad \text{a.s.}$$

so that

$$\overline{\lim}\, n^{-1} \sum_1^n \eta_s^2 / \left(1 - \underline{\lim}\, \psi_N' P_N \psi_N\right) < \infty \,. \tag{8.14}$$

Now suppose

$$\underline{\lim}\, \psi_N' P_N \psi_N = 1$$

then we must have

$$\overline{\lim}\, n^{-1} \sum_{1}^{n} \eta_s^2 = 0$$

but then, recalling the diagonal elements of P_n^{-1} in (8.11) we conclude, via MIL that $\psi_n' P_n \psi_n \to 0$ which contradicts our supposition: thus $\underline{\lim}\, \psi_N' P_N \psi_N < 1$. Now, however (8.14) ensures

$$\overline{\lim}\, n^{-1} \sum_{1}^{n} \eta_s^2 < \infty \quad \text{a.s.}$$

Now via MIL we have established (8.12). Thus MGCT* applied to (8.9c) yields $T_n'/n \to T'$ a.s. a finite positive random variable and

$$\sum_{1}^{\infty} n^{-1}\left(T_{n-1}'/(n-1)\right) < \infty \quad \text{a.s.}$$

so that $T' = 0$ a.s. to avoid contradiction. Thus, by positivity we conclude

$$T_n/n \to 0 \quad \text{a.s.}$$

$$n^{-1} \sum_{1}^{n} \hat{u}_s \hat{y}_s = C_n/n \to 0 \quad \text{a.s.}$$

Now if we can establish

$$\underline{\lim}\, n^{-1} P_n^{-1} > 0 \quad \text{a.s.}$$

we may conclude, from $T_n/n \to 0$ a.s. that $\theta_n - \theta_0 \to 0$ a.s. as required.

First we show

$$\lim n^{-1} \sum_{1}^{n} \eta_s^2 = \sigma^2 \quad \text{a.s.} \tag{8.15a}$$

Now

$$n^{-1} \sum_{1}^{n} \eta_s^2 = n^{-1} \sum_{1}^{n} \left[\left(\eta_s - \varepsilon_s\right)^2 + 2\varepsilon_s\left(\eta_s - \varepsilon_s\right) + \varepsilon_s^2\right] .$$

In a moment we show

$$n^{-1}\sum_1^n \left(\eta_s-\varepsilon_s\right)^2 \to 0 \quad \text{a.s.} \tag{8.15b}$$

this dispenses with the first term. For the second term apply the Cauchy-Schwarz inequality to see

$$\left|\sum_1^n \varepsilon_s\left(\eta_s-\varepsilon_s\right)/n\right| \le \left(\sum_1^n \varepsilon_s^2 \sum_1^n \left(\eta_s-\varepsilon_s\right)^2\right)^{\frac{1}{2}}/n \to 0$$

since $n^{-1}\sum_1^n \varepsilon_s^2 \to \sigma^2$. We have then only to show (8.15b).

We have already established

$$n^{-1}C_n = n^{-1}\sum_1^n \hat{u}_s\hat{y}_s \to 0 \quad \text{a.s.} \tag{8.16a}$$

and recalling the definitions of $\hat{y}_n$, $\hat{u}_n$, we see (8.15b) will follow from

$$n^{-1}\sum_1^n \hat{u}_s^2 \to 0 , \tag{8.16b}$$

$$n^{-1}\sum_1^n \hat{y}_s^2 \to 0 . \tag{8.16c}$$

We can use standard methods to deduce these zero power conditions from (8.16a).

To show these we have to strengthen the positive real condition to one of strict positive realness; that is, for some small $\sigma' > 0$,

$$\text{Re } h\left(e^{i\omega}\right) \ge \sigma' .$$

Then we obtain from (8.10)

$$\sum_1^N \hat{u}_s\hat{y}_s \ge (\sigma'/2\pi)\sum_1^N \hat{u}_s^2 \tag{8.17}$$

so that (8.16b) follows from (8.16a). We now bound the energy $\sum_1^N \hat{y}_s^2$.

Define

$$\bar{u}_n = \hat{u}_n , \quad 1 < n < N ,$$
$$= 0 \qquad \text{otherwise.}$$

Set

$$\bar{y}_n = \sum_1^n h_{n-s}\bar{u}_s \ .$$

Now consider

$$\begin{aligned}\sum_1^N \hat{y}_s^2 &= \sum_1^N \bar{y}_n^2 \le \sum_1^\infty \bar{y}_n^2 \\ &= \sum_1^\infty \left(\sum_1^n h_{n-s}\bar{u}_s\right)^2 \\ &= (2\pi)^{-1}\int_{-\pi}^{\pi}\left|\sum_0^\infty e^{in\omega}\sum_1^n h_{n-s}\bar{u}_s\right|^2 d\omega \\ &= (2\pi)^{-1}\int_{-\pi}^{\pi}\left|\sum_0^\infty h_m e^{im\omega}\right|^2\left|\sum_0^\infty e^{is\omega}\bar{u}_s\right|^2 d\omega \\ &\le b\int_{-\pi}^{\pi}\left|\sum_0^\infty e^{is\omega}\bar{u}_s\right|^2 d\omega\end{aligned}$$

where

$$b = \max_{\omega}.\left|h\left(e^{i\omega}\right)\right|^2/2\pi$$

and $b < \infty$ since $h(L)$ is rational and stable. Continuing

$$\sum_1^N \hat{y}_s^2 \le b\sum_0^\infty \bar{u}_s^2 = b\sum_1^N \hat{u}_s^2 \tag{8.18}$$

so that (8.16c) is established.

We may now conclude that all the diagonal elements of P_n^{-1}, such as in (8.11) have finite positive limits. We will conclude

$$\underline{\lim}\; n^{-1} P_n^{-1} > 0 \quad \text{a.s.}$$

if we show to be positive definite the correlation matrix whose typical elements are of the form

$$\sum_1^N \eta_n\eta_{n-m}\Big/\left(\sum_1^N \eta_n^2 \sum_1^N \eta_{n-m}^2\right)^{\frac{1}{2}} , \tag{8.19a}$$

$$\sum_1^N \eta_n y_{n-m}\Big/\left(\sum_1^N \eta_n^2 \sum_1^N y_{n-m}^2\right)^{\frac{1}{2}} , \tag{8.19b}$$

$$\sum_1^N \eta_n u_{n-m} \Big/ \left(\sum_1^N \eta_n^2 \sum_1^N u_{n-m}^2 \right)^{\frac{1}{2}} , \tag{8.19c}$$

$$\sum_1^N y_n u_{n-m} \Big/ \left(\sum_1^N y_n^2 \sum_1^N u_{n-m}^2 \right)^{\frac{1}{2}} , \tag{8.19d}$$

$$\sum_1^N y_n y_{n-m} \Big/ \left(\sum_1^N y_n^2 \sum_1^N y_{n-m}^2 \right)^{\frac{1}{2}} , \tag{8.19e}$$

$$\sum_1^N u_n u_{n-m} \Big/ \left(\sum_1^N u_n^2 \sum_1^N u_{n-m}^2 \right)^{\frac{1}{2}} . \tag{8.19f}$$

We show this correlation matrix converges to the limit that is obtained if η_n in (8.19) is replaced by ε_n : the resulting matrix is positive definite by assumption. The results follow immediately for (8.19d), (8.19e), (8.19f). For (8.19b) consider

$$\sum_1^N \eta_n y_{n-m} \Big/ \left(\sum_1^N \eta_n^2 \sum_1^N y_{n-m}^2 \right)^{\frac{1}{2}} = \sum_1^N (\eta_n - \varepsilon_n) y_{n-m} \Big/ \left(\sum_1^N \eta_n^2 \sum_1^N y_{n-m}^2 \right)^{\frac{1}{2}}$$

$$+ \sum_1^N \varepsilon_n y_{n-m} \Big/ \left(\sum_1^N \eta_n^2 \sum_1^N y_{n-m}^2 \right)^{\frac{1}{2}} .$$

Apply the Cauchy-Schwarz inequality to the first term to conclude via (8.15a) and (8.15b) that it converges to zero. The second term converges to the required limit (zero in this case). Similar arguments apply to terms (8.19a), (8.19c). The proof is complete.

Remark. Having thus established a.s. convergence we can in fact show more. In (8.9c) the divisor n may be replaced by n^δ for any $\delta > 0$ whereupon

$$\sum_1^\infty \psi_n' P_n \psi_n / n^\delta < \infty \tag{8.20}$$

will ensure (since we already know $\underline{\lim}\, n^{-1} P_n^{-1} > 0$)

$$n^{\frac{1}{2}-\delta} (\theta_n - \theta_0) \to 0 \quad \text{a.s.}$$

as well as

$$n^{-\delta} \sum_1^n \hat{u}_s^2 \to 0 \quad \text{a.s.,}$$

$$n^{-\delta} \sum_1^n \hat{y}_s^2 \to 0 \quad \text{a.s.},$$

$$n^{-\delta} \sum_1^n \hat{u}_s \hat{y}_s \to 0 \quad \text{a.s.}$$

However since the smallest eigenvalue of P_n^{-1} is of order n, Lemma 1 of Section 4.3 implies

$$\sum_1^\infty \psi_n' P_n^2 \psi_n n^{1-\delta} < \infty$$

whence

$$\overline{\lim}\ \mathrm{tr}\ P_n^{-1}/n < \infty$$

and

$$\psi_n' P_n \psi_n / n \le \mathrm{tr}\ P_n^{-1}/n\ \psi_n' P_n^2 \psi_n$$

imply (8.20).

8.3 The Convergence of RML_1

We turn now to consider RML_1 using a very similar approach to that used for RML_1'. However it will become clear that the proof cannot be completed quite as easily as for RML_1' (except in the MA(1) case). Since it is the author's feeling that a proof can be found, the argument has been presented for interest. This section can however be skipped over quickly without losing continuity.

Recall, then, the RML_1 recursion for an ARMAX model

$$\theta_n = \theta_{n-1} + P_n \varphi_n e_n , \tag{8.21a}$$

$$e_n = y_n - \varphi_n' \theta_{n-1} , \tag{8.21b}$$

$$P_n^{-1} = P_{n-1}^{-1} + \varphi_n \varphi_n' , \tag{8.21c}$$

with φ_n as in (8.3c). Now premultiply (8.21a) by P_n^{-1} and use (8.21c) to see

$$P_n^{-1}(\theta_n-\theta_0) = P_{n-1}^{-1}(\theta_{n-1}-\theta_0) + \varphi_n\varphi_n'(\theta_{n-1}-\theta_0) + \varphi_n e_n . \qquad (8.22)$$

Next denote $T_n = (\theta_n-\theta_0)'P_n^{-1}(\theta_n-\theta_0)$; it follows from (7.21a), (7.21c) that

$$T_n = T_{n-1} + (\theta_{n-1}-\theta_0)'\varphi_n\varphi_n'(\theta_{n-1}-\theta_0) + 2(\theta_{n-1}-\theta_0)'\varphi_n e_n + \varphi_n'P_n\varphi_n e_n^2$$

which we rewrite as

$$T_n = T_{n-1} + (\theta_{n-1}-\theta_0)'\varphi_n[\varphi_n'(\theta_{n-1}-\theta_0)+2(e_n-\varepsilon_n)]$$
$$+ 2\varepsilon_n(\varphi_n'(\theta_{n-1}-\theta_0)+(e_n-\varepsilon_n)\varphi_n'P_n\varphi_n) + \varphi_n'P_n\varphi_n\{(e_n-\varepsilon_n)^2+\varepsilon_n^2\} .$$

Next consider that, by definition,

$$(1+\nu_{n-1}(L))e_n = y_n - (y_{n-1} \cdots y_{n-n_\alpha} u_{n-1} \cdots u_{n-n_\gamma})'\beta_{n-1} ,$$
$$(1+\nu_0(L))\varepsilon_n = y_n - (y_{n-1} \cdots y_{n-n_\alpha} u_{n-1} \cdots u_{n-n_\gamma})'\beta_0 ,$$

with

$$\beta' = (\alpha\gamma) : \nu_{n-1}(L) = \sum_1^{n_\nu} \nu_{in-1}L^i .$$

Thus (*cf*. (8.8))

$$(1+\nu_0(L))(e_n-\varepsilon_n) = -\varphi_n'(\theta_{n-1}-\theta_0) . \qquad (8.23)$$

Now denote

$$\hat{u}_n = -\varphi_n'(\theta_{n-1}-\theta_0) ,$$
$$\hat{y}_n = e_n - \varepsilon_n + \tfrac{1}{2}\varphi_n'(\theta_{n-1}-\theta_0)$$

so that, neglecting initial conditions

$$\hat{y}_n = \{(1+\nu_0(L))^{-1}-\tfrac{1}{2}\}\hat{u}_n .$$

The above expression for T_n may be written

$$T_n = T_{n-1} + 2\varepsilon_n\{-\hat{u}_n+\varphi_n'P_n\varphi_n(e_n-\varepsilon_n)\} - 2\hat{u}_n\hat{y}_n$$
$$+ \varepsilon_n^2\varphi_n'P_n\varphi_n + (e_n-\varepsilon_n)^2\varphi_n'P_n\varphi_n . \qquad (8.24)$$

Before continuing we are going, once more, to state a condition that ensures $C_N = 2\sum_1^{\infty} \hat{u}_n\hat{y}_n > 0$ a.s. Let us require

$$h(L) = \sum_0^{\infty} h_s L^s = \left(1+\nu_0(L)\right)^{-1} - \tfrac{1}{2}$$

to be strictly positive real, that is, for some $\sigma' > 0$,

$$\mathrm{Re}\ h\left(e^{i\omega}\right) > \sigma' .$$

Then, neglecting initial conditions we obtain as in (8.10),

$$C_N = 2\sum_1^N \hat{u}_s\hat{y}_s \geq \sigma \sum_1^N \hat{u}_s^2 \tag{8.25}$$

where $\sigma = \sigma'/\pi$.

Also we state for future reference the bound on input/output energy ratios similar to (8.18),

$$\sum_1^N \left(e_n-\varepsilon_n\right)^2 \leq b \sum_1^N \hat{u}_s^2 \tag{8.26}$$

where this time

$$b = \max_w \left|g\left(e^{i\omega}\right)\right|^2/2\pi$$

with

$$g(L) = \left(1+\nu_0(L)\right)^{-1} .$$

Next set

$$C_N' = 2\sum_1^N \hat{u}_s\hat{y}_s - \sum_1^N \left(e_n-\varepsilon_n\right)^2 \varphi_n' P_n \varphi_n$$

then after taking conditional expectations in (8.24) we can find either

$$E\left(T_n/n + C_n'/n \mid F_{n-1}\right)$$

$$= \left(T_{n-1}/(n-1)+C_{n-1}'/(n-1)\right)(1 - 1/n) + \sigma^2\varphi_n' P_n \varphi_n/n \tag{8.27a}$$

or else, recalling $\varphi_n' P_n \varphi_n \leq 1$,

$$E\left(T_n/n + C_n/n \mid F_{n-1}\right) \le \left(T_{n-1}/(n-1)+C_{n-1}/(n-1)\right)(1 - 1/n)$$
$$+ \sigma^2\varphi_n'P_n\varphi_n/n + \left(e_n-\varepsilon_n\right)^2/n \ . \quad (8.27b)$$

For reasons to be made clear *suppose* it can be shown that

$$\overline{\lim}\ \varphi_n'P_n\varphi_n < 1 \quad \text{a.s.} \tag{8.28}$$

Then it follows that $\sum_1^\infty \varphi_n'P_n\varphi_n/n < \infty$ a.s. To see this multiply (8.21a) by P_n^{-1}, form $\theta_n'P_n^{-1}\theta_n$ and sum to obtain

$$\theta_n'P_n^{-1}\theta_n + \sum_1^n e_s^2\left(1-\varphi_s'P_s\varphi_s\right) = \sum_1^n y_s^2 \tag{8.29}$$

which is of course expression (8.13b). Now we see that (8.28) implies

$$\overline{\lim}\ n^{-1}\sum_1^n e_s^2 < \infty \quad \text{a.s.}$$

Recalling that the diagonal elements of P_n^{-1} are of the form

$$n^{-1}\sum_1^n e_s^2\ , \quad n^{-1}\sum_1^n y_s^2\ , \quad n^{-1}\sum_1^n u_s^2$$

we conclude $\overline{\lim}\ \mathrm{tr}\left(P_n^{-1}/n\right) < \infty$ a.s. so that $\sum_1^\infty \varphi_n'P_n\varphi_n/n < \infty$ follows by MIL.

There are now two cases.

In case $\sum_1^\infty \left(e_n-\varepsilon_n\right)^2 < \infty$ then MGCT* of Chapter 4 applied to (8.27b) implies

$$T_n/n \to 0 \quad \text{a.s.}$$

and

$$C_n/n \to 0 \quad \text{a.s.}$$

so that, in view of (8.26),

$$N^{-1}\sum_1^N \hat{u}_n^2 \to 0\ .$$

If we can show $\underline{\lim}\, N^{-1}P_N^{-1} > 0$ a.s. then we may conclude a.s. convergence of θ_n to θ_0. We postpone the proof but note that it will depend only on knowing

$$N^{-1}\sum_1^N (e_n-\varepsilon_n)^2 \to 0\ ,\quad N^{-1}\sum_1^N \hat{u}_n^2 \to 0\ ,\quad \overline{\lim}\, N^{-1}\sum_1^N e_k^2 < \infty\ . \tag{8.30}$$

In case $\sum_1^N (e_n-\varepsilon_n)^2 \to \infty$, then rewrite C_N' as

$$\begin{aligned} C_N' &\geq \sum_1^N \hat{u}_n^2\left(\sigma-\xi_n \sum_1^n (e_n-\varepsilon_n)^2/\sum_1^N \hat{u}_n^2\right) \\ &\geq \sum_1^N \hat{u}_n^2(\sigma-\xi_n b) \end{aligned}$$

in view of (8.26) and where

$$\xi_n = \sum_1^n (e_s-\varepsilon_s)^2\varphi_s'P_s\varphi_s/\sum_1^n (e_s-\varepsilon_s)^2\ .$$

However recall that, by MIL, $\overline{\lim}\, n^{-1}\sum_1^n e_s^2 < \infty$ ensures $\varphi_n'P_n^{-1}\varphi_n \to 0$ so that ξ_n can be made arbitrarily small, thus ultimately $C_N' > 0$. MGCT* applied to (8.27a) now enables the conclusion $T_n/n \to 0$ a.s. as well as (8.30).

We now show how (8.30) implies $\underline{\lim}\, N^{-1}P_N^{-1} > 0$ a.s. First notice that each diagonal element of P_N^{-1} say $v_{ii}(N)$ satisfies

$$\overline{\lim}\, v_{ii}(N)/N < \infty\ ,\quad \underline{\lim}\, v_{ii}(N)/N > 0\ .$$

This second condition follows by considering that

$$N^{-1}\sum_1^N e_n^2 = N^{-1}\sum_1^N (e_n-\varepsilon_n)^2\left\{1+2\sum_1^N \varepsilon_n(e_n-\varepsilon_n)/\sum_1^N (e_n-\varepsilon_n)^2\right\} + N^{-1}\sum_1^N \varepsilon_n^2$$

and Lemma 3' of Chapter 4 implies the second term in brackets converges if $\sum_1^\infty (e_n-\varepsilon_n)^2 < \infty$ or converges to 0 a.s. if $\sum_1^N (e_n-\varepsilon_n)^2 \to \infty$: in either case clearly

$$\underline{\lim}\, N^{-1} \sum_1^N e_n^2 \geq \lim N^{-1} \sum_1^N \varepsilon_n^2 = \sigma^2 .$$

The required result will now follow if we investigate the correlation matrix whose typical terms are

$$\sum_1^N e_n e_{n-m} \Big/ \left(\sum_1^N e_n^2 \sum_1^N e_{n-m}^2\right)^{\frac{1}{2}} , \tag{8.31}$$

$$\sum_1^N e_n y_{n-m} \Big/ \left(\sum_1^N e_n^2 \sum_1^N y_{n-m}^2\right)^{\frac{1}{2}} , \tag{8.32a}$$

$$\sum_1^N e_n u_{n-m} \Big/ \left(\sum_1^N e_n^2 \sum_1^N u_{n-m}^2\right)^{\frac{1}{2}} , \tag{8.32b}$$

$$\sum_1^N y_n y_{n-m} \Big/ \left(\sum_1^N y_n^2 \sum_1^N y_{n-m}^2\right)^{\frac{1}{2}} , \tag{8.32c}$$

and show these terms converge to the same ones that would be obtained were e_n replaced everywhere by ε_n. The resulting correlation matrix is positive definite by assumption. This follows immediately for (8.32c). For (8.32a) consider

$$\left|\sum_1^N e_n y_{n-m} \Big/ \left(\sum_1^N e_n^2 \sum_1^N y_{n-m}^2\right)^{\frac{1}{2}}\right| = \left|\left(\sum_1^N (e_n - \varepsilon_n) y_{n-m} + \sum_1^N \varepsilon_n y_{n-m}\right) \Big/ \left(\sum_1^N e_n^2 \sum y_{n-m}^2\right)^{\frac{1}{2}}\right|$$

$$\leq \sum_1^N (e_n - \varepsilon_n)^2 \Big/ \sum_1^N e_n^2 + \left|\sum_1^N \varepsilon_n y_{n-m}\right| \Big/ \sum_1^N y_{n-m}^2$$

$$\to 0 + 0$$

by (8.30) and Lemma 3' of Chapter 4. A similar argument works for the other terms.

It thus remains to show (8.28). At this point the proof collapses and the author has only been able to demonstrate (8.28) for the MA(1) case.

To do this we proceed as in Hannan (1976). We have

$$\nu_N = \sum_1^N e_{n-1} y_n \Big/ \sum_1^N e_n^2$$

$$= \sum_1^N e_{n-1} \varepsilon_n \Big/ \sum_1^N e_n^2 + \nu_0 \sum_1^N (e_{n-1} - \varepsilon_{n-1}) \varepsilon_{n-1} \Big/ \sum_1^N e_n^2 + \nu_0 \sum_1^N \varepsilon_{n-1}^2 \Big/ \sum_1^N e_n^2 .$$

However by MGC$\hat{\text{T}}$ (Lemma 3' of Chapter 4) the first two terms vanish a.s. so that

$$\overline{\lim}\ |\nu_N| \le |\nu_0|$$

which immediately implies $\overline{\lim}\ N^{-1} \sum_1^N e_n^2 < \infty$ so that MIL $\Rightarrow \varphi_n' P_n \varphi_n \to 0$ which implies (8.28). The unfortunate aspect of finishing the proof in this way is that it does not provide an argument that generalises to higher order cases.

Remark. Having thus established a.s. convergence *conditional on* (8.28) we can in fact show more. In (8.27) the divisor can be replaced by n^δ (for any $\delta > 0$) whereupon

$$\sum_1^\infty \varphi_n' P_n \varphi_n / n^\delta < \infty \tag{8.33}$$

will ensure (since we already know $\underline{\lim}\ n^{-1} P_n^{-1} > 0$) $n^{\frac{1}{2}-\delta}(\theta_n - \theta_0) \to 0$ a.s. as well as

$$n^{-\delta} \sum_1^n \hat{u}_s^2 \to 0 \quad \text{a.s.,}$$

$$n^{-\delta} \sum_1^n (e_s - \varepsilon_s)^2 \to 0 \quad \text{a.s.}$$

However since the smallest eigenvalue of P_n^{-1} is of order n, Lemma 1 of Section 4.3 implies

$$\sum_1^\infty \varphi_n' P_n^2 \varphi_n n^{1-\delta} \to \infty$$

whereupon $\overline{\lim}$ tr $P_n^{-1}/n < \infty$ and

$$\varphi_n' P_n \varphi_n / n \le \text{tr}\left(P_n^{-1}/n\right) \varphi_n' P_n^2 \varphi_n$$

implies (8.33).

8.4 The Unusual Second Order Properties of RML_1

We begin by investigating the CLT for RML_1 applied to a MA(1) model. The remark at the end of the last section will be referred to repeatedly. Consider then

$$\theta_n - \theta_0 = \left(\sum_1^n \varphi_k^2\right)^{-1} \sum_1^n \varphi_k\left(y_k - \varphi_k\theta_0\right)$$

$$= \left(\sum_1^n e_{k-1}^2\right)^{-1} \sum_1^n e_{k-1}\varepsilon_k - \left(\sum_1^n e_{k-1}^2\right)^{-1} \sum_1^n e_{k-1}\left(e_{k-1}-\varepsilon_{k-1}\right)$$

$$= \left(\sum_1^n e_{k-1}^2\right)^{-1} \sum_1^n e_{k-1}\varepsilon_k - \left(\sum_1^n e_{k-1}^2\right)^{-1} \sum_1^n \left(e_{k-1}-\varepsilon_{k-1}\right)^2$$

$$\times \left(1-\left(\sum_1^n \left(e_{k-1}-\varepsilon_{k-1}\right)^2\right)^{-1} \sum_1^n \varepsilon_{k-1}\left(e_{k-1}-\varepsilon_{k-1}\right)\right) .$$

Clearly when the second term is normed by $\sqrt{n}$ it still converges to zero: the first term will yield an obvious CLT. For any other model however the situation changes. Consider a MA(2) model. We need only deal with

$$n^{-1} \sum_1^n \varphi_k\left(y_k - \varphi_k'\nu_0\right) = n^{-1} \sum_1^n \varphi_k\varepsilon_k + n^{-1} \sum_1^n \varphi_k\left(\varphi_k\left(\nu_0\right) - \varphi_k\right)'\theta_0 .$$

The first term, when normed up by $\sqrt{n}$ will obey a CLT. We take each component of the second term separately. Denote $\nu_0 = \left(a_0, b_0\right)$ and consider minus the second component

$$n^{-1} \sum_1^n e_{k-2}\left(a_0\left(e_{k-1}-\varepsilon_{k-1}\right) + b_0\left(e_{k-2}-\varepsilon_{k-2}\right)\right) = n^{-1} \sum_1^n \left(e_{k-2}-\varepsilon_{k-2}\right)\left(e_{k-1}-\varepsilon_{k-1}\right)a_0$$

$$+ n^{-1} \sum_1^n \left(e_{k-2}-\varepsilon_{k-2}\right)^2 b_0 + n^{-1} \sum_1^n \varepsilon_{k-2}\left(e_{k-1}-\varepsilon_{k-1}\right)a_0 + b_0 n^{-1} \sum_1^n \left(e_{k-2}-\varepsilon_{k-2}\right)\varepsilon_{k-2} .$$

Since we expect, according to the end of Section 8.3, that

$$n^{-\delta} \sum_1^n \left(e_k - \varepsilon_k\right)^2 \to 0$$

all the terms, except the third one will vanish a.s. when normed up by $\sqrt{n}$. In fact this term cannot be expected to vanish. Now recall

$$\left(1+\nu_0(L)\right)\left(e_k - \varepsilon_k\right) = -\varphi_k'\left(\theta_{k-1} - \theta_0\right)$$

so

$$e_{k-1} - \varepsilon_{k-1} = -a_0(e_{k-2}-\varepsilon_{k-2}) - b_0(e_{k-3}-\varepsilon_{k-3}) - e_{k-2}(a_{k-2}-a_0) - e_{k-3}(b_{k-2}-b_0)$$

minus the third term will reduce itself to (norming up)

$$n^{-\frac{1}{2}} \sum_1^n \varepsilon_{k-2} e_{k-2}(a_{k-2}-a_0)$$

and so by MGC$\hat{\text{T}}$ to

$$n^{-\frac{1}{2}} \sum_1^n \varepsilon_{k-2}^2 (a_{k-2}-a_0)$$

which clearly cannot be expected to vanish though it might converge in distribution. In any case the author has been unable to find an argument to prove such a result. In Chapter 2 an heuristic argument was given to suggest a form for the CLT for a general recursion and indeed that argument suggests a form for this CLT.

If we rewrite the recursion (8.3a) as

$$\theta_n - \theta_0 = \theta_{n-1} - \theta_0 + P_n\varphi_n(e_n-\varepsilon_n) + P_n\varphi_n\varepsilon_n \ ,$$

recall expression (8.23)

$$(1+\nu_0(L))(e_n-\varepsilon_n) = -\varphi_n'(\theta_{n-1}-\theta_0)$$

and notice that, from (8.1),

$$\tilde{\varphi}_n(\theta) = de_n/d\theta = -(1+\nu(L))^{-1}\varphi_n(\theta)$$

then we are led, as in Chapter 2 to conjecture an Invariance Principle for

$$Q_n = \sqrt{n}\ (\theta_n-\theta_0)$$

as

$$Q_{n+k_n}(t) - Q_n \Rightarrow Q(t)$$

where $Q(t)$ is the solution to the stochastic differential equation (Arnold, 1974)

$$dQ(t) = (\tfrac{1}{2}I-G^{-1}H)Q(t)dt + dW(t)$$

with initial condition $Q(0) \sim N(0,\ K)$ where

$$H = E(\varphi_n(\theta_0)de_n/d\theta_0') \ ,$$

$$G = E\left(\varphi_n(\theta_0)\varphi_n'(\theta_0)\right)$$

and $W(t)$ is a Brownian motion with variance G^{-1} : while K obeys the Liapunov equation (Arnold, 1974)

$$\left(\tfrac{1}{2}I-G^{-1}H\right)K + K\left(\tfrac{1}{2}I-G^{-1}H\right)' = G^{-1} .$$

This result is only partially confirmed by the simulations of Chapter 3; in any case the author has been unable to find a proof for it.

APPENDIX 8A

THE POSITIVE REAL LEMMA

Let us rewrite $h(L)$ as

$$h(L) = \frac{-\nu_0(L)}{1+\nu_0(L)} + \tfrac{1}{2}$$

so that it is obvious that $h(L)$ has a state space representation as

$$x_n = F_0 x_{n-1} + b\hat{u}_{n-1} ,$$

$$\hat{y}_n = h'x_n + \tfrac{1}{2}\hat{u}_n$$

where b is the vector of noise parameters; F_0 is in observable canonical form, that is

$$F_0 = \begin{pmatrix} -\nu_1 & 1 & \dots & 0 \\ -\nu_2 & & \ddots & \\ \vdots & & & 1 \\ -\nu_{n_\nu} & 0 & \dots & 0 \end{pmatrix}$$

and h is the n_ν vector $(1\ 0\ \dots\ 0)'$: $h(L) = h'\left(L^{-1}I-F_0\right)^{-1}b$. Now it it is known (the discrete positive real lemma, see Hitz and Anderson, 1969; Anderson *et al.* 1974) that iff the above state space system is positive real there exists a positive definite matrix P , a vector l , and scalar w such that

$$F_0'PF_0 = P - ll' ,$$

$$F_0'Pb = h - l\omega \ ,$$

$$\omega^2 = \tfrac{1}{2} + \tfrac{1}{2} - b'Pb \ .$$

Consider now the quadratic form

$$\begin{aligned}
E_n = x_n'Px_n &= \left(x_{n-1}'F_0'+b'\hat{u}_{n-1}\right)P\left(F_0x_{n-1}+b\hat{u}_{n-1}\right) \\
&= E_{n-1} - \left(x_{n-1}'l\right)^2 + 2b'PF_0x_{n-1}\hat{u}_{n-1} + b'Pb\hat{u}_{n-1}^2 \\
&= E_{n-1} - \left(x_{n-1}'l\right)^2 - 2l'x_{n-1}\hat{u}_{n-1}\omega - \omega^2\hat{u}_{n-1}^2 \\
&\qquad + 2\left(\hat{y}_{n-1}-\tfrac{1}{2}\hat{u}_n\right)\hat{u}_{n-1} + 2(\tfrac{1}{2})\hat{u}_{n-1}^2 \\
&= E_{n-1} - \left(x_{n-1}'l-\omega\hat{u}_n\right)^2 + 2\hat{u}_{n-1}\hat{y}_{n-1} \ .
\end{aligned}$$

Thus we conclude

$$0 \le E_n \le 2\sum_{1}^{n-1} \hat{u}_s\hat{y}_s + E_0$$

as required.

REFERENCES

Albert, A.E., and Gardner, L.A. (1967), *Stochastic Approximation and Nonlinear Regression,* Cambridge, Mass., M.I.T. Press.

Anderson, B.D.O. (1977), An approach to multivariable system identification, *Automatica*, 13, 401-408.

Anderson, B.D.O., Hitz, K.L., and Diem, N.D. (1974), Recursive algorithm for spectral factorization, *IEEE Trans. Circ. and Syst.*, 6, 742-750.

Anderson, T.W. and Taylor, J.B. (1977), Strong consistency of least squares estimates in dynamic models. Unpublished manuscript.

Arnold, L. (1974), *Stochastic Differential Equations: Theory and Applications*, J. Wiley

Aström, K.J. and Bohlin, T. (1965), Numerical identification of linear dynamic systems from normal operating records, in *Theory of Self Adaptive Systems*, ed. P.N. Hammond, Plenum Press.

Aström, K.J. and Eykhoff, P. (1971), System identification - A survey, *Automatica*, 7, 123-162.

Aström, K.J. and Wittenmark, B. (1973), On self-tuning regulators, *Automatica*, 9, 185-199.

Barnett, S. (1971), *Matrices in Control Theory*, Van Nostrand.

Billingsley, P. (1968), *Convergence of Probability Measures,* J. Wiley.

Blum, J.R. (1954), Multidimensional stochastic approximation methods, *Ann. Math. Stats.*, 25, 737-744.

Bodewig, E. (1956), *Matrix Calculus*, Interscience.

Bromwich, T.J. (1947), *An Introduction to the Theory of Infinite Series*, MacMillan, London, 2nd ed.

Brown, R.L., Durbin, J., and Evans, J.M. (1975), Techniques for testing the constancy of regression relationships over time, with comments, *J. Royal Stat. Soc.*, *Ser. B*, 37, 149-192.

Bucy, R.S. (1965), Stability and positive super-martingales, *J. Differential Equations*, 1, 151-155.

Caines, P.E. (1975), A note on the consistency of maximum likelihood estimates for finite families of stochastic processes, *Ann. Statist.*, 3, 539-546.

Derman, C. and Sacks, J. (1959), On Dvoretsky's stochastic approximation theorem, *Ann. Math. Statist.*, 30, 601-606.

Drygas, H. (1976), Weak and strong consistency of the least squares estimators in regression models, Z. *Wahrscheinlichkeitstheorie verw. Gebiete*, 34, 119-127.

Duncan, D.B., and Horn, S.D. (1972), Linear dynamic recursive estimation from the viewpoint of regression analysis, *J. Amer. Stat. Assoc.*, 67, 815-821.

Dvoretsky, A. (1956), On stochastic approximation, in *Proc.* 3rd *Berkeley Symp. on Mathematical Statistics and Probability*, Vol. 1, 39-56.

Feigin, P. (1975), Maximum Likelihood Estimation for Stochastic Processes - a Martingale Approach, PhD Thesis, Australian National University.

Gantmacher, F.R. (1960), *The Theory of Matrices*, Volume I, Chelsea Publishing Co., New York.

Gertler, J., and Banyasz, Cs. (1974), A recursive (on-line) maximum likelihood identification method, *IEEE Trans. Autom. Control*, 19, 816-819.

Gladyshev, E.G. (1965), On stochastic approximation, *Theory Prob. Applic.*, 10, 275-278.

Goodwin, G.C. and Payne R.L. (1977), *Dynamic System Identification: Experiment Design and Analysis*, Academic Press.

Gupta, N.K. Mehra, R.K. (1977), Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations, *IEEE Trans. Autom. Control*, 19, 774-783.

Hahn, S (1967), Stability of Motion, Springer-Verlag.

Hannan, E.J. (1970), *Multiple Time Series*, J. Wiley.

Hannan, E.J. (1973), Central Limit Theorems for Time Series Regression, Z. *Wahrscheinlichkeitstheorie verw. Gebiete*, 26, 157-170.

Hannan, E.J. (1976), The convergence of some recursions, *Ann. Statist.*, 4, 1258-1270.

Hannan, E.J. (1978a), Rates of convergence for Time Series Regression, submitted for publication.

Hannan, E.J. (1978b), The Central Limit Theorem for Time Series Regression, to be published.

Hannan, E.J. (1978c), Recursive estimation based on ARMA models, to be published.

Hannan, E.J. and Heyde, C.C. (1972), On Limit Theorems for quadratic functions of discrete time series, *Ann. Math. Statist.*, 43, 2058-2066.

Hannan, E.J., and Tanaka, K. (1976), ARMAX models and recursive calculations, in *Proc. Conf. on Systems Dynamics and Control in Quantitative Economics*, Kinokuniya (to be published), Nagoya. H. Myoken (Nagoya City University), ed.

Hendry, D. (1975), Monte-Carlo methods for investigating finite sample distributions, London School of Economics, discussion paper.

Heyde, C.C. (1974), On martingale limit theory and strong convergence results for stochastic approximation procedures, *Stoc. Proc. and Applic.*, 2, 371-389.

Heyde, C.C. (1975), Stochastic approximation, unpublished notes.

Heyde, C.C., and Scott, D.J. (1973), Invariance principles for the law of the iterated logarithm for martingales and processes with stationary increments, *Ann. Prob.*, 1, 428-436.

Hitz, K.L., and Anderson, B.D.O. (1969), Discrete positive real functions and their application to system stability, *Proc. IEE* 116, 153-155.

Ho, Y.C. (1963), On the stochastic approximation method and optimal filtering theory, *J. Math. Anal. and Applic.*, 6, 152-154.

Holst, J. (1977), Adaptive Prediction and Recursive Estimation, PhD Thesis, Lund Institute of Technology.

Hwang, S.Y. (1978), Solution of complex integrals using the Laurent expansion, *IEEE*, ASSP-26, 263-265.

Ibragimov, I.A. and Khas'minskii, R.Z. (1973), On the approximation of statistical estimators by sums of independent variables, *Soviet Math. Dokl.*, 14, 883-887.

Jakeman, A.J., and Young, P.C. (1978), Refined Instrumental Variable Methods of Recursive Time-Series Analysis; Part II, Multivariable Systems, Report AS/R13, Centre for Resource and Environmental Studies, ANU (to appear *Int. J. Control*).

Jameson, A. (1968), Solution of the equation $AX + XB = C$ by inversion of an $M \times M$ or $N \times N$ matrix, *SIAM J. Appl. Math.*, 16, 1020-1023.

Jazwinski, A.H. (1970), *Stochastic Processes and Filtering Theory*, Academic Press.

Kailath, T. (1970), The innovations approach to detection and estimation theory, *Proc. IEEE*, 58, 680-695.

Kaldor, J.M. (1978), Estimation of Parametric Change in Time Series Models, MA Thesis, Centre for Resource and Environmental Studies, ANU.

Kalman, R.E. (1960), A new approach to linear filtering and prediction problems, *Trans ASME Ser. D*, 82, 35-44.

Kersting, G. (1971), Almost sure approximation of the Robbins-Monro process by sums of independent random variables, *Ann. Prob.*, 5, 954-965.

Khasminskii, R.Z. (1969), *Stability of systems of differential equations under random perturbations of their parameters*, "Nauka", Moscow (Russian).

Keifer, J., and Wolfowitz, J. (1952), Stochastic estimation of the maximum of a regression function, *Ann. Math. Statist.*, 23, 462-466.

Kittler, J. and Young, P.C. (1973), A new approach to feature selection based on the Karhunen-Loeve expansion, *Pattern Recognition*, 5, 335.

Knopp, K. (1956), *Infinite Sequences and Series*, New York, Dover.

Kohr, R.H. (1963), A method for the determination of the differential equation model for simple non-linear systems, *IEEE Tr. Elect.*, 12, 394-400.

Kreisselmeier, G. (1977), Adaptive observers with exponential rate of convergence, *IEEE Trans. Autom. Control*, 22, 2-8.

Kushner, H.J. (1977), General convergence results for stochastic approximations via weak convergence theory, *J. Math. Anal. Applic.*, 61, 490-503.

Kushner, H.J. (1978a), Rates of convergence for sequential Monte Carlo optimization methods, *SIAM J. Control and Opt.*, 10, 150-168.

Kushner, H.J. (1978b), Convergence of recursive adaptive and identification procedures via weak convergence theory, *IEEE Trans. Autom. Control*, 22, 921-930.

Landau, I.D. (1974), A survey of model reference adaptive techniques - theory and applications, *Automatica*, 10, 353-379.

Landau, I.D. (1976), Unbiased recursive identification using model reference adaptive techniques, *IEEE Trans. Autom. Control*, 21, 194-202.

Landau, I.D. (1978), An addendum to "Unbiased recursive identification using model reference adaptive techniques", *IEEE Trans. Autom. Control*, 25, 97-98.

La Salle, J., and Lefschefz, S. (1961), *Stability by Liapunov's Direct Method*, Academic Press.

Lion, P.M. (1967), Rapid identification of linear and nonlinear systems, *AIAA Journal*, 5, 1835-1842.

Ljung, L. (1974), Convergence of Recursive Stochastic Algorithms. Rep. 7403. Department of Automatic Control, Lund Inst. of Technology, Lund, Sweden.

Ljung, L. (1975), Theorems for the Asymptotic Analysis of Recursive Stochastic Algorithms. Rep. 7522. Department of Automatic Control, Lund Inst. of Technology, Lund, Sweden.

Ljung, L. (1976), Prediction error identification methods, in *System Identification - Advances and Case Studies*, Mehra, R.K., Lainiotis, P.G., eds.

Ljung, L. (1977a), On positive real transfer functions and the convergence of some recursive schemes, *IEEE Trans. Autom. Control*, 22, 539-551.

Ljung, L. (1977b), Analysis of recursive stochastic algorithms, *IEEE Trans. Autom. Control*, 22, 551-575.

Ljung, L. (1977c), The Extended Kalman Filter as a Parameter Estimator for Linear Systems. Report Li TH-ISY-I-0154, Department of Electrical Engineering, Linköping University, Sweden.

Ljung, L. (1978a), Strong convergence of a stochastic approximation algorithm, *Ann. Stats.*, 6, 680-696.

Ljung, L. (1978b), Personal Communication.

Lukacs, E. (1975), *Stochastic Convergence*, Academic Press, 2nd ed.

Major, P. and Révész, P. (1973), A limit theorem for the Robbins-Monro approximation, *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 27, 79-86.

McGarty, T.P. (1974), *Stochastic Systems and State Estimation*, New York, J. Wiley.

McLeish, D.L. (1974), Dependent Central Limit Theorems and Invariance Principles, *Ann. Prob.*, 2, 620-628

McLeish, D.L. (1976), Functional and random Central Limit Theorems for the Robbins-Monro process, *J. Appl. Prob.*, 13, 148-154.

Moore, J.B. (1978), On strong consistency of least squares identification algorithms, *Automatica*, September 1978.

Moore, J.B. and Ledwich, G. (1978), Multivariable adaptive parameter and state estimators with convergence analysis, to be published.

Narendra, K.S., and Kudva, P. (1974), Stable adaptive schemes for system identification and control - Part I, *IEEE Trans. Syst. Man. Cybernetics*, 4, 542-551.

Neudecker, H. (1969), Some theorems on matrix differentiation with specific reference to Kronecker matrix products, *J. Amer. Stat. Assn.*, 64, 953-963.

Nevel'son, M.B., and Khas'minskii, R.Z. (1973/1976), *Stochastic Approximation and Recursive Estimation*. Translations of Math. Monographs, Vol. 47. Providence, Rhode Island, American Mathematical Society.

Neveu, J. (1965), *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco.

Neveu, J. (1975), *Discrete Parameter Martingales*, North-Holland, New York.

Parks, P.E. (1966), Liapunov redesign of model reference adaptive control systems, *IEEE Trans. Autom. Control*, 11, 362-367.

Phillips, P.C.B. (1974), The estimation of some continuous time models, *Econometrica*, 42, 803-823.

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

Robbins, H., and Monro, S. (1951), A stochastic approximation method, *Ann. Math. Statist.*, 22, 400-407.

Rosenblatt, M. (1974), *Random Processes*, Springer-Verlag, 2nd ed.

Rucker, R.A. (1963), Real Time System Identification in the Presence of Noise. Proc. Western Electronics Conv. Paper 2, 3, August (1963).

Sacks, J. (1958), Asymptotic distribution of stochastic approximation procedures, *Ann. Math. Statist.* 29, 373-405.

Saridis, G.N. (1974), Stochastic approximation methods for identification and control - A survey, *IEEE Trans. Autom. Control*, 19, 798-809.

Saridis, G.N. (1977), *Self-Organizing Control of Stochastic Systems*, Marcel Dekker, New York.

Schweppe, F.C. (1965), Evaluation of likelihood functions for Gaussian signals, *IEEE Trans. Info. Theory*, 11, 61-69.

Scott, D.J. (1973), Central limit theorems for martingales and for processes with stationary increments using a Skorokhod representation approach, *Adv. Appl. Prob.*, 5, 119-137.

Sidar, M. (1976), Recursive identification and tracking of parameters for linear and nonlinear multivariable systems, *Int. J. Control*, 24, 361-378.

Söderström, T., Ljung, L., and Gustavsson, I. (1974), A Comparative Study of Recursive Identification Methods. Report 7427, Div. of Automatic Control, Lund Inst. of Technology, Lund, Sweden.

Son, L.H., Anderson, B.D.O. (1971), Design of Kalman filters using signal-model output statistics, *Proc. IEE*, 120, 312-318.

Stout, W.F. (1974), *Almost Sure Convergence*, Academic Press.

Tsypkin, Ya.Z. (1971), *Adaptation and Learning in Automatic Systems*, Academic Press.

Venter, J.H. (1967), An extension of the Robbins-Monro procedure, *Ann. Math. Statist.*, 38, 181-190.

Walk, H. (1977), An Invariance Principle for the Robbins-Monro process in a Hilbert space, *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 39, 135-150.

Wolovich, W.A. (1974), *Linear Multivariable Systems*, Springer, New York.

Young, P.C. (1964), Discussion in *IEEE Trans. Aerosp. Systems*, 2, 1106-1111.

Young, P.C. (1974), Recursive approaches to time-series analysis, *Bull. Inst. Math. Applic.*, 10, Nos. 5/6, 209-224.

Young, P.E. (1976a), Optimisation in the presence of noise - a guided tour, Chapter in *Optimisation in Action*, L.C.W. Dixon, ed., Academic Press, London.

Young, P.C. (1976b), Some observations on instrumental variable methods of time-series analysis, *Int. J. Control*, 23, 593-612.

Young, P.C. and Beck, B.M. (1974), The modelling and control of water quality in a river system, *Automatica*, 10, 455-468.

Young, P.C. and Jakeman, A.J. (1978), Refined Instrumental Variable Methods of Recursive Time-Series Analysis; Part I, Single Input - Single Output Systems, Report Number AS/R12 (L978), Centre for Resource and Environmental Studies, ANU (to appear *Int. J. Control*).

Young, P.C., Jakeman, A.J., and Whitehead, P.G., 1978, Report Number AS/R17, Centre for Resource and Environmental Studies, ANU.