Statistica Sinica **16**(2006), 741-772

EFFICIENT RECURSIVE ESTIMATION AND ADAPTIVE CONTROL IN STOCHASTIC REGRESSION AND ARMAX MODELS

Tze Leung Lai and Zhiliang Ying

Stanford University and Columbia University

Abstract: This paper first reviews C.Z. Wei's seminal work and subsequent developments in recursive estimation and adaptive control of stochastic regression models and ARMAX systems. By using certain ideas and techniques from these developments, it then develops a new approach to efficient semiparametric estimation in linear time series models.

Key words and phrases: Adaptive control, asymptotic efficiency, least squares, M-estimators, semiparametric estimation, stochastic regression, strong consistency.

1. Introduction

As noted by Chan's article in this issue, a major area of Ching Zong Wei's research is efficient recursive estimation and adaptive control in stochastic dynamic systems. In 1977, when he was a second-year graduate student at Columbia, Wei started working in this area with Lai and Robbins. Although the initial objectives of his research were more modest, contemporary developments in the emerging field of stochastic adaptive control led to an ambitious research program shortly after he received his Ph.D. in 1980, culminating in a number of major advances within the five-year period 1982−1987. In Section 2 we review the background and significance of this work. Although Wei then turned his attention to inference in time series and stochastic processes using the results and insights from this work, and also new tools that he subsequently developed (see Chan's article), this work paved the way for the further important developments in recursive estimation and adaptive control described in Section 3. Following Wei's path, Section 4 turns to some challenging estimation problems in time series. By using certain ideas and techniques outlined in Section 3, we develop a new approach to efficient recursive estimation in linear time series models. In particular, asymptotically efficient recursive estimators are developed for the parameters of ARMAX models with i.i.d. symmetric disturbances whose common density function is unknown. Some concluding remarks are given in Section 5.

2. Stochastic Regression Models and Adaptive Control

Wei's (1980) Ph.D. thesis, and his related papers on strong consistency of least squares (LS) estimates with Lai and Robbins in 1978 and 1979 mentioned in Chan (2006), were inspired by the *multi-period control problem under uncertainty* in econometrics. His subsequent papers on stochastic regression models, stochastic approximation and adaptive prediction in the period 1982-87 were motivated by applications to recursive identification and adaptive control in the contemporary engineering literature, in particular to the problem of *self-tuning regulators*.

2.1. The multiperiod control problem and strong consistency of LS estimates

The multi-period control problem under uncertainty is to choose successive inputs $u_1, \ldots, u_n$ in the linear regression model

$$y_i = \alpha + \beta u_i + \epsilon_i, \tag{2.1}$$

where α and β are unknown parameters and the random disturbances $\epsilon_1, \epsilon_2, \ldots$ are i.i.d. with mean 0 and variance σ^2, so that the outputs $y_1, \ldots, y_n$ are as close as possible in some sense to a given target value y^*. A Bayesian formulation is often used so that the problem becomes that of minimizing

$$\begin{aligned} &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E_{\alpha,\beta}\Big[\sum_{i=1}^{n}(y_i - y^*)^2\Big] d\pi(\alpha,\beta) \\ &= n\sigma^2 + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \beta^2 E_{\alpha,\beta}\Big[\sum_{i=1}^{n}(u_i-\theta)^2\Big] d\pi(\alpha,\beta), \end{aligned} \tag{2.2}$$

where π is a prior distribution of the unknown parameters α, β and $\theta = (y^* - \alpha)/\beta$; see Zellner (1971), Prescott (1972) and Wieland (2000). The ϵ_i are often assumed to be normally distributed, and (2.2) can in principle be minimized by using dynamic programming. However, because of the "curse of dimensionality" in numerically solving the dynamic programming equations and the analytical difficulties in studying the Bayes rules, not much was known about their performance until the recent work of Han, Lai and Spivakovsky (2005), who "sandwich" the performance of the Bayes rule between (i) that of an approximate policy optimization procedure which uses Monte Carlo simulations to circumvent the curse of dimensionality, and (ii) that of an "oracle policy" which becomes tractable by assuming the value of β to be revealed after k periods in the future.

A first departure from the Bayesian approach was due to Aokoi (1974). Assuming that the sign of β is known, say $\beta > 0$, he proposed the use of a stochastic approximation scheme of the form

$$u_{t+1} = u_t - a_t(y_t - y^*), \tag{2.3}$$

where $\{a_i\}$ is a sequence of positive constants such that $\Sigma_1^\infty a_i = \infty$ and $\Sigma_1^\infty a_i^2 < \infty$. This approach has the property that $u_t \to \theta$ a.s. Another non-Bayesian approach is the certainty-equivalence rule proposed by Anderson and Taylor (1976). If α and $\beta(\neq 0)$ are both known, the optimal choice of u_t is clearly at $\theta = (y^* - \alpha)/\beta$. Without assuming α and β to be known, suppose that bounds K_1, K_2 are known such that $K_1 < \theta < K_2$. Assuming the ϵ_i to be normally distributed, the maximum likelihood estimator of of θ at stage $t \geq 2$ is

$$\widehat{\theta}_t = K_2 \wedge \{\widehat{\beta}_t^{-1}(y^* - \widehat{\alpha}_t) \vee K_1\}, \tag{2.4}$$

where $\widehat{\beta}_t = \{\Sigma_1^t(u_i - \bar{u}_t)y_i\}/\{\Sigma_1^t(u_i - \bar{u}_t)^2\}$, $\widehat{\alpha}_t = \bar{y}_t - \widehat{\beta}_t\bar{u}_t$ are the least squares estimates of β and α, $\bar{u}_t = t^{-1}\Sigma_1^t u_i$, and $\wedge$ and $\vee$ denote minimum and maximum, respectively. The initial values u_1 and u_2 are distinct but otherwise arbitrary numbers between K_1 and K_2, and for $t \geq 2$, the certainty-equivalence rule sets $u_{t+1} = \widehat{\theta}_t$. Despite its simplicity for implementation, the certainty-equivalence rule is difficult to analyze. A basic difficulty is that although $\widehat{\theta}_t$ may conceivably represent one's best current guess of θ, how good the guess is depends on how the inputs $u_1, \ldots, u_t$ are chosen. In particular, if the inputs $u_i = \widehat{\theta}_{i-1}, 1 \leq i \leq t$, tend to cluster around their mean $\bar{u}_t$, then there may not be enough information to give a reliable estimate $\widehat{\theta}_t$, even though $\widehat{\theta}_t$ may well be one's closest guess of θ at stage t.

Based on the results of simulation studies, Anderson and Taylor (1976) conjectured that the certainty-equivalence rule converges to θ a.s. and that $\sqrt{t}(\widehat{\theta}_t - \theta)$ has a limiting $N(0, \sigma^2/\beta^2)$ distribution. They also raised the question whether $\widehat{\alpha}_t$ and $\widehat{\beta}_t$ are strongly consistent. This question led Lai and Robbins to study the strong consistency of $\widehat{\alpha}_t$ and $\widehat{\beta}_t$ when the regressors u_t are sequentially determined random variables, not only in the Anderson-Taylor case $u_{t+1} = \widehat{\theta}_t$, but also in adaptive stochastic approximation for which the a_t in (2.3) is chosen to be of the form $a_t \sim (t\widehat{\beta}_t)^{-1}$ that would be asymptotically equivalent (if $\widehat{\beta}_t$ should be strongly consistent) to the optimal choice $a_t = (t\beta)^{-1}$ when β is known.

Under the regression model (2.1), the LS estimate can be written as

$$\widehat{\beta}_n - \beta = \frac{\sum_{i=1}^n (u_i - \bar{u}_n)\epsilon_i}{\sum_{i=1}^n (u_i - \bar{u}_n)^2}. \tag{2.5}$$

When the u_i are nonrandom constants, (2.5) yields $E\widehat{\beta}_n = \beta$ and $\mathrm{Var}(\widehat{\beta}_n) = \sigma^2/\Sigma_{i=1}^n(u_i-\bar{u}_n)^2$, so $\widehat{\beta}_n$ converges to β in L_2 (and therefore also in probability) if

$$\sum_{i=1}^n (u_i-\bar{u}_n)^2 \to \infty, \tag{2.6}$$

but strong consistency of $\widehat{\beta}_n$ under (2.6) is much more difficult to establish since (2.5) involves a weighted sum of the ϵ_i with a double array of weights $u_i-\bar{u}_n$. To circumvent the difficulty due to the double array structure, Lai and Robbins (1977) used the identities

$$\begin{aligned}\sum_{i=1}^n (u_i-\bar{u}_n)\epsilon_i &= \sum_{i=2}^n \Big(\frac{i-1}{i}\Big)(u_i-\bar{u}_{i-1})(\epsilon_i-\bar{\epsilon}_{i-1}),\ \sum_{i=1}^n (u_i-\bar{u}_n)^2 \\ &= \sum_{i=2}^n \Big(\frac{i-1}{i}\Big)(u_i-\bar{u}_{i-1})^2\end{aligned}$$

to convert double-array into single-array sums. Although $\{\epsilon_i-\bar{\epsilon}_{i-1},\ i\geq 1\}$ is no longer an i.i.d. sequence, it is a sequence of uncorrelated random variables and is therefore an independent sequence when the ϵ_i are normal. Therefore Kolmogorov's strong law of large numbers yields the desired strong consistency when the ϵ_i are normal, and Lai and Robbins (1977) used a strong embedding argument for $\bar{\epsilon}_{i-1}$ to prove the result when the i.i.d. ϵ_i satisfy $E\{\epsilon_i^2(\log^+|\epsilon_i|)^r\}<\infty$ for some $r>1$, under the minimal assumption (2.6) on the design with nonrandom u_i. This method does not work for general multiple regression, and around that time Wei joined the project to develop an alternative approach for multiple regression models as a graduate research assistant.

A definitive solution to the strong consistency problem for nonrandom regressors was obtained within a year and was published in Lai, Robbins and Wei (1978, 1979). The basic idea is to use the concept of convergence systems in orthogonal series theory (cf., Banach (1930) and Gaposhkin (1966)) and the property that certain linear transformations associated with the information matrices of LS estimates preserve convergence systems. Specifically, a sequence of random variables ϵ_i is called a *convergence system* if

$$\sum_{i=1}^\infty a_i\epsilon_i \text{ converges a.s. for all nonrandom } \{a_i, i\geq 1\} \text{ such that } \sum_{i=1}^\infty a_i^2<\infty. \tag{2.7}$$

Let $\mathbf{x}_i$ be a k-dimensional nonrandom vector and let $\mathbf{H}_n = \Sigma_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$. Suppose $\mathbf{H}_n$ is positive definite for all $n\geq m$ and $\{\epsilon_i\}$ is a convergence system. Then

Theorem 2 of Lai, Robbins and Wei (1979) yields

$$\{\mathbf{x}_{n+1}^T \mathbf{H}_n^{-1}\Big(\sum_{i=1}^{n} \mathbf{x}_i \epsilon_i\Big)\Big/(1+\mathbf{x}_{n+1}^T \mathbf{H}_n^{-1} \mathbf{x}_{n+1})^{1/2},\ n \geq m\} \text{ is a convergence system.} \tag{2.8}$$

The proof of (2.8) uses an induction argument and an inequality for certain sequences of real numbers; see Lemma 1 of Lai, Robbins and Wei (1979, pp.346-348) whose Sections 3 and 4 use (2.8) to derive the strong consistency of LS estimates in the multiple regression model $y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i$ under (2.7) and the minimal assumption $(\Sigma_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)^{-1} \to 0$ on the design constants. Chen, Lai and Wei (1981) subsequently extended this approach to provide a unified treatment of all previous results in the literature on the strong consistency of LS estimates in multiple regression models with nonrandom regressors, and in particular for the Gauss-Markov model.

Lai, Robbins and Wei (1978) also considered strong consistency of LS estimates under weak moment assumptions on the random errors ϵ_i. As pointed out later by Lai and Wei (1984, Section 3), these consistency theorems can be restated more generally in terms of certain linear transformations of lacunary systems in orthogonal series theory. Given $p > 0$, a sequence of random variables ϵ_i is called a *lacunary system of order* p, or S_p *system*, if there exists a positive constant K_p such that for all nonrandom c_i,

$$E\Big|\sum_{i=m}^{n} c_i \epsilon_i\Big|^p \leq K_p \Big(\sum_{i=m}^{n} c_i^2\Big)^{\frac{p}{2}} \text{ for all } n \geq m \geq m_0. \tag{2.9}$$

Móricz (1976) has shown that if Z_i are random variables for which there exist $\alpha > 1$ and nonnegative constants d_i such that $E|Z_n - Z_m|^p \leq (\Sigma_{i=m}^n d_i)^\alpha$ for $n \geq m \geq m_0$, then there exists an absolute constant $C_{p,\alpha}$ such that

$$E\Big(\max_{m \leq i \leq n} |Z_i - Z_m|^p\Big) \leq C_{p,\alpha} \Big(\sum_{i=m}^{n} d_i\Big)^\alpha \text{ for all } n \geq m \geq m_0. \tag{2.10}$$

Making use of (2.9) and the maximal inequality (2.10), Corollary 2 of Lai and Wei (1984) generalizes Lai, Robbins and Wei (1978), who considered the special case $p = 4$ for S_p systems, to derive rates of a.s. convergence of the LS estimate to $\boldsymbol{\beta}$ under the minimal assumption $(\Sigma_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)^{-1} \to 0$ on the design constants.

These strong consistency results assume nonrandom $\mathbf{x}_i$ and are therefore not applicable to the multiperiod control problem with $\mathbf{x}_i = (1, u_i)^T$, for which Lai and Robbins (1979, 1981) used adaptive stochastic approximation schemes of the form (2.3) with $ta_t = \zeta_t \vee (\widehat{\beta}_t \wedge \xi_t)$, where $\zeta_t \to 0$ and $\xi_t \to \infty$. When the u_i are sequentially determined random variables, as in stochastic approximation or in the certainty-equivalence rule, Lai and Robbins (1981) proved strong consistency of $\widehat{\beta}_n$ under the condition

$$\frac{\sum_{i=1}^{n}(u_i - \bar{u}_n)^2}{\log n} \to \infty \quad \text{a.s.} \tag{2.11}$$

Although (2.11) is stronger than (2.6) for nonrandom u_i, they gave an example in which u_i is $\mathcal{F}_{i-1}$-measurable and $\Sigma_{i=1}^{n}(u_i - \bar{u}_n)^2/\log n \to c$ a.s for some constant $c > 0$, but in which $\widehat{\beta}_n \to \beta - c^{-1}$ a.s. They also provided a set of sufficient conditions on $\Sigma_{i=1}^{n}(u_i - \theta)^2$ and $\bar{u}_n - \theta$ that can be verified for adaptive stochastic approximation, thereby proving the desired strong consistency of $\widehat{\beta}_n$ and establishing the desired properties

$$\sum_{i=1}^{n}(u_i - \theta)^2 \sim \left(\frac{\sigma^2}{\beta^2}\right)\log n \quad \text{a.s.}, \tag{2.12a}$$

$$\sqrt{n}(u_n - \theta) \Rightarrow N(0, \frac{\sigma^2}{\beta^2}), \tag{2.12b}$$

for adaptive stochastic approximation schemes; see Lai and Robbins (1979, 1981). For the Anderson-Taylor certainty-equivalence rule, however, Lai and Robbins (1982) showed that (2.4) does not converge to θ a.s. by exhibiting an event which has positive probability and on which u_n gets stuck at one of the endpoints K_1, K_2 for $n \geq 2$, giving little information to estimate β.

2.2. Related limit theorems and multivariate stochastic approximation

Wei (1987a) generalized (2.12a, b) to multivariate adaptive stochastic approximation schemes of the form $\mathbf{u}_{t+1} = \mathbf{u}_t - (t\mathbf{B}_t)^{-1}\mathbf{y}_t$ for the regression model $\mathbf{y}_i = \mathbf{M}(\mathbf{u}_i) + \boldsymbol{\epsilon}_i$, in which the regression function $\mathbf{M} : \mathbf{R}^d \to \mathbf{R}^d$ has a unique zero at $\boldsymbol{\theta}$ and satisfies certain regularity conditions. Some basic results of this work were already obtained in the second part of his thesis, Wei (1980), whose first part focused on limit theorems for weighted sums of independent random variables and more general martingale difference sequences ϵ_i that arise in regression and time series models. After completing his thesis, he continued working with Lai in this direction, and developed laws of the iterated logarithm (LIL) for double arrays of independent random variables in Lai and Wei (1982a), and limit theorems for generalized linear processes in Lai and Wei (1983a, 1985). In particular, Theorem 2 of Lai and Wei (1982a) gives the following LIL of the jth component $\widehat{\beta}_{nj}$ of the LS estimate in the multiple regression model $y_i = \boldsymbol{\beta}^T\mathbf{x}_i + \epsilon_i$ with nonrandom regressors:

$$\limsup_{n\to\infty} \frac{|\widehat{\beta}_{nj} - \beta_j|}{\{2\ \mathrm{Var}(\widehat{\beta}_{nj})\log\log(\mathrm{Var}\widehat{\beta}_{nj})^{-1}\}^{\frac{1}{2}}} = 1 \quad \text{a.s.} \tag{2.13}$$

when the ϵ_i are independent random variables with $E\epsilon_i = 0, E\epsilon_i^2 = \sigma^2 > 0$ and $\sup_i E|\epsilon_i|^r < \infty$ for some $r > 2$, and the regressors satisfy $\mathrm{Var}(\widehat{\beta}_{nj}) \to 0$ and some other weak regularity conditions.

2.3. The self-tuning regulator and efficient adaptive prediction and control

A widely used stochastic model in the time series and stochastic control literature is the ARMAX system (autoregressive moving average system with exogenous inputs) defined by the linear difference equation

$$A(q^{-1})y_n = B(q^{-1})u_{n-d} + C(q^{-1})\epsilon_n, \tag{2.14}$$

where $\{y_n\}, \{u_n\}$ and $\{\epsilon_n\}$ denote the output, input and disturbance sequences, respectively, $d \geq 1$ represents the delay and $A(q^{-1}) = 1 + a_1 q^{-1} + \cdots + a_p q^{-p}$, $B(q^{-1}) = b_1 + \cdots + b_k q^{-(k-1)}$, $C(q^{-1}) = 1 + c_1 q^{-1} + \cdots + c_h q^{-h}$ are scalar polynomials in the backward shift operator q^{-1}. Because of its theoretical interest and practical importance, the problem of determining the inputs u_n, based on current and past observations $y_n, y_{n-1}, u_{n-1}, \ldots$ (i.e., u_n is $\mathcal{F}_n$-measurable), to keep the outputs y_{n+d} as close as possible to certain target values y_{n+d}^* when the system parameters are not known in advance but have to be estimated "on-line" (i.e., during the operation of the system) has been one of the major topics in the subject of stochastic adaptive control. Let $x_0 = (y_0, \ldots, y_{1-p}, u_0, \ldots, u_{2-d-k}, \epsilon_0, \ldots, \epsilon_{1-h})^T$ denote the "initial condition" of (2.14). The polynomial $B(z) = b_1 + \cdots + b_k z^{k-1}$ is called *stable* if all its zeros lie outside the unit circle, and two or more polynomials are said to be *relatively prime* if their greatest common divisors have degree 0.

In principle, given a probability distribution of the random sequence $\{x_0, \epsilon_1, \epsilon_2, \ldots\}$ and a prior distribution π of the unknown parameter vector

$$\boldsymbol{\beta} = (-a_1, \ldots, -a_p, b_1, \ldots, b_k, c_1, \ldots, c_h)^T, \tag{2.15}$$

we can use backward induction to solve the dynamic programming equations defining the inputs $u_1, \ldots, u_{N-d}$ that minimize

$$\int E_{\boldsymbol{\beta}}\Big\{ \sum_{i=d+1}^{N} (y_i - y_i^*)^2 \Big\} d\pi(\boldsymbol{\beta}), \tag{2.16}$$

for every given horizon N, where the y_i^* are given target values for the outputs. Despite the analytical and computational difficulties in the implementation of the Bayesian approach, Bayesian analysis of some very simple examples has provided important insights into the structure of optimal control rules. In particular, Feldbaum (1961) and subsequent authors have shown that Bayes rules have the "dual control" function of both probing the system for information about its parameters and trying to drive the outputs towards their target values.

A seminal paper of Åström and Wittenmark (1973) proposed a certainty-equivalence approach as a practical alternative to the Bayesian approach. To begin with, consider the regulation problem (with $y_t^* \equiv 0$) in the case of unit delay ($d = 1$) and white noise ($C(q^{-1}) = 1$). Replacing b_1 in (2.15) by a prior guess $b \neq 0$, they proposed to estimate the other parameters at stage t by $\alpha_1^{(t)}, \ldots, \alpha_p^{(t)}, \beta_2^{(t)}, \ldots, \beta_k^{(t)}$ that minimize

$$\sum_{i=1}^{t}(y_i + \alpha_1 y_{i-1} + \cdots + \alpha_p y_{i-p} - b u_{i-1} - \beta_2 u_{i-2} - \cdots - \beta_k u_{i-k})^2, \tag{2.17}$$

and to determine the input u_t by the certainty-equivalence rule

$$u_t = \frac{\alpha_1^{(t)} y_t + \cdots + \alpha_p^{(t)} y_{t-p+1} - \beta_2^{(t)} u_{t-1} - \cdots - \beta_k^{(t)} u_{t-k+1}}{b}. \tag{2.18}$$

They also showed that if the estimates should converge as $t \to \infty$, then $\alpha_1^{(t)}/b, \ldots, \alpha_p^{(t)}/b, \beta_2^{(t)}/b, \ldots, \beta_k^{(t)}/b$ must necessarily converge to the coefficients $a_1/b_1, \ldots, a_p/b_1, b_2/b_1, \ldots, b_k/b_1$ in the optimal regulator $u_t^* = \{a_1 y_t + \cdots + a_p y_{t-p+1} - b_2 u_{t-1} - \cdots - b_k u_{t-k+1}\}/b_1$ that assumes knowledge of the system parameters, and therefore the adaptive regulator u_t "self-tunes" itself in the sense that its defining equation has asymptotically negligible difference from that of u_t^*. Moreover, instead of adhering to a prior guess b of b_1 in the rule, they also considered updating this guess with the current and past data, which amounts to a LS certainty-equivalence rule. By reparameterizing the system (2.14) and using least squares or extended least squares to directly estimate the transformed parameters, they also suggested natural extensions of (2.18) to general delay and colored noise.

An open problem with the Åström-Wittenmark approach is whether, with positive probability, the parameter estimates may fail to converge. Using the insights from adaptive stochastic approximation in Lai and Robbins (1979, 1981) and its multivariate extension in Wei (1980), Lai and Wei began studying this problem in 1981. They were also inspired by a landmark paper by Goodwin, Ramadge and Caines (1981) who circumvented the difficulties in the analysis of sequential LS estimates in a feedback control environment by using stochastic gradient (SG) estimates instead of LS estimates, and who showed that under certain assumptions the SG certainty-equivalence control rule has the "self-optimizing" (or "globally convergent") property that

$$n^{-1}\sum_{i=1}^{n}(y_i - y_i^* - \epsilon_i)^2 \to 0 \quad \text{a.s.} \tag{2.19}$$

In the case of white noise and unit delay (i.e., $C(q^{-1}) = 1$ and $d = 1$), the SG estimates are given recursively by

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \frac{\alpha}{r_t}\mathbf{x}_t(y_t - \boldsymbol{\beta}_{t-1}^T\mathbf{x}_t), \;\; r_t = r_{t-1} + \|\mathbf{x}_t\|^2, \tag{2.20}$$

where $\alpha > 0$ is a tuning parameter and $\mathbf{x}_t = (y_{t-1}, \ldots, y_{t-p}, u_{t-1}, \ldots, u_{t-k})^T$. Since (2.14) can be written as the stochastic regression model $y_n = \boldsymbol{\beta}^T\mathbf{x}_n + \epsilon_n$, using LS estimates $\widehat{\boldsymbol{\beta}}_t$ is expected to be more efficient than using SG estimates. Moreover, like (2.20), LS estimates can be implemented on-line via the recursions

$$\widehat{\boldsymbol{\beta}}_t = \widehat{\boldsymbol{\beta}}_{t-1} + \mathbf{P}_t\mathbf{x}_t(y_t - \widehat{\boldsymbol{\beta}}_{t-1}\mathbf{x}_t), \tag{2.21a}$$

$$\mathbf{P}_t = \mathbf{P}_{t-1} - \frac{\mathbf{P}_{t-1}\mathbf{x}_t\mathbf{x}_t^T\mathbf{P}_{t-1}}{(1 + \mathbf{x}_t^T\mathbf{P}_{t-1}\mathbf{x}_t)}, \tag{2.21b}$$

in which (2.21b) is the recursion (or *Riccati equation*) for $(\Sigma_{i=1}^t \mathbf{x}_t\mathbf{x}_i^T)^{-1}$. Thus, (2.20) simply replaces $\mathbf{P}_t$ in (2.21a) by $\alpha/tr(\mathbf{P}_t^{-1})$. To demonstrate that using LS instead of SG estimates indeed leads to a more efficient control rule, Lai and Wei proceeded to show that a suitably chosen adaptive control rule using LS estimates can achieve

$$R_n = O(\log n) \text{ a.s., where } R_n = \sum_{i=1}^{n}(y_i - y_i^* - \epsilon_i)^2, \tag{2.22}$$

which is much sharper than the self-optimizing property (2.19) for the SG-based control rule. Note that if $\boldsymbol{\beta}$ is known and $b_1 \neq 0$, then the minimum-variance controller u_{t-1} is given by $\boldsymbol{\beta}^T\mathbf{x}_t = y_t^*$, yielding the output $y_t = y_t^* + \epsilon_t$ at time t. When $\boldsymbol{\beta}$ is unknown, the *regret* R_n of an adaptive control rule at stage n is defined by (2.22). In this connection, note that (2.12a) gives a logarithmic regret of the Lai-Robbins adaptive stochastic approximation rule in the multiperiod control problem with $y_t^* = 0$, for which $y_t - y_t^* - \epsilon_t = \beta(u_t - \theta)$.

A first step by Lai and Wei towards generalizing the logarithmic regret in (2.12) to the present setting was to develop new techniques to prove strong consistency of LS estimates in stochastic regression models. This led to the paper by Lai and Wei (1982b) mentioned in Section 2.1 of Chan's article in this issue. As pointed out in Sections 3 and 5 of Lai (2003), a major technical breakthrough of Lai and Wei (1982b) was the use of "extended stochastic Liapounov functions", whereas previous authors relied on Liapounov functions of associated ordinary differential equations (cf., Ljung (1977)) or on stochastic Liapounov functions (cf., Gladyshev (1965), Robbins and Siegmund (1970), Solo (1979) and Goodwin, Ramage and Caines (1981)). Let $\{\epsilon_n, \mathcal{F}_n, n \geq 1\}$ be a martingale difference such that $\sup_n E(\epsilon_n^2|\mathcal{F}_{n-1}) < \infty$ a.s. An *extended stochastic Liapounov function* V_n is a nonnegative $\mathcal{F}_n$-measurable random variable satisfying

$$V_n \leq (1 + a_{n-1})V_{n-1} + b_n - c_n + w_{n-1}\epsilon_n \quad \text{a.s.,} \tag{2.23}$$

where $a_n \geq 0, b_n \geq 0, c_n \geq 0$ and w_n are $\mathcal{F}_n$-measurable random variables such that $\Sigma a_n < \infty$. Although Lai and Wei (1982b) did not introduce this terminology, as they only considered the recursions (2.21a, b), their arguments can be used to show that (2.23) implies

$$\max\Big(V_n, \sum_{i=1}^n c_i\Big) = O\Big(\sum_{i=1}^n b_i + \Big(\sum_{i=1}^n w_i^2\Big)^{\frac{1}{2}+\delta}\Big) \quad \text{a.s.} \tag{2.24}$$

for every $\delta > 0$. In fact, the recursions (2.21a) and (2.21b) lead to the recursive inequality (2.23) with $V_n = (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^T \mathbf{P}_n^{-1}(\widehat{\boldsymbol{\beta}}_n - \beta), a_n = 0, b_n = \mathbf{x}_n^T \mathbf{P}_n \mathbf{x}_n \epsilon_n^2$, $c_n = \{(\widehat{\boldsymbol{\beta}}_{n-1} - \boldsymbol{\beta})^T \mathbf{x}_n\}^2 (1 - \mathbf{x}_n^T \mathbf{P}_n \mathbf{x}_n),\ w_{n-1} = 2\{(\widehat{\boldsymbol{\beta}}_{n-1} - \boldsymbol{\beta})^T \mathbf{x}_n\}(1 - \mathbf{x}_n^T \mathbf{P}_n \mathbf{x}_n)$. Assuming that $\{\epsilon_n, \mathcal{F}_n, n \geq 1\}$ is a martingale difference with $\sup_n E(\epsilon_n^2|\mathcal{F}_{n-1}) < \infty$ a.s., Lai and Wei (1982b) showed that (2.24) holds in this case. Moreover, they also showed under the stronger moment condition $\sup_n E(|\epsilon_n|^{2+\delta}|\mathcal{F}_{n-1}) < \infty$ a.s. for some $\delta > 0$, that

$$\sum_{i=1}^n \mathbf{x}_i^T \mathbf{P}_i \mathbf{x}_i \epsilon_i^2 = O\Big(\sum_{i=1}^n \mathbf{x}_i^T \mathbf{P}_i \mathbf{x}_i\Big) = O\Big(\log \lambda_{\max}\Big(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T\Big)\Big) \quad \text{a.s.,}$$

thereby strengthening (2.24) in this case into

$$\max\Big(V_n, \sum_{i=1}^n c_i\Big) = O\Big(\log \lambda_{\max}\Big(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T\Big)\Big) \quad \text{a.s.} \tag{2.25}$$

Here and in the sequel, we use $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ to denote the maximum and minimum eigenvalues, respectively.

Stochastic Liapounov functions V_n, used earlier by Gladyshev (1965) and Robbins and Siegmund (1970) for the convergence analysis of stochastic approximation schemes, and by Solo (1979) and Goodwin, Ramage and Caines (1981) in their analysis of the AML and SG algorithms, are $\mathcal{F}_n$-measurable random variables satisfying

$$E(V_n|\mathcal{F}_{n-1}) \leq (1 + a_{n-1})V_{n-1} + b_{n-1} - c_{n-1},$$

in contrast with (2.23) that defines extended stochastic Liapounov functions. Whereas stochastic Liapounov functions converge a.s. because they are nonnegative almost supermartingales, extended stochastic Liapounov functions offer much greater flexibility as they need not converge a.s.; convergence of V_n is only ensured on $\{\Sigma_{i=1}^{\infty} E(b_i|\mathcal{F}_{i-1}) < \infty\}$. Since

$$V_n = (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^T \Big(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \Big) (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \geq \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|^2 \lambda_{\min} \Big(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \Big),$$

(2.25) yields the "local convergence" result

$$\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta} \text{ a.s. on } \Big\{ \frac{\lambda_{\min}\left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right)}{\log \lambda_{\max}\Big(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \Big)} \to \infty \Big\}. \tag{2.26}$$

Moreover, since $c_t = \{(\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta})^T \mathbf{x}_t\}^2 (1 - \mathbf{x}_t^T \mathbf{P}_t \mathbf{x}_t)$, (2.25) also yields

$$\sum_{t=1}^{n} (\widehat{\boldsymbol{\beta}}_{t-1}^T \mathbf{x}_t - \boldsymbol{\beta}^T \mathbf{x}_t)^2 I_{\{\mathbf{x}_t^T \mathbf{P}_t \mathbf{x}_t \leq \delta\}} = O\Big(\log \lambda_{\max} \Big(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \Big) \Big) \quad \text{a.s.} \tag{2.27}$$

for every $0 < \delta < 1$. Note that (2.27) gives an asymptotic bound on the cumulative squared difference between the minimum-variance $\mathcal{F}_{t-1}$-measurable predictor $\boldsymbol{\beta}^T \mathbf{x}_t$ of y_t, when $\boldsymbol{\beta}$ is known, and the LS adaptive predictor $\widehat{\boldsymbol{\beta}}_{t-1}^T \mathbf{x}_t$. Hence, if it can be shown that

$$\log(\|\mathbf{x}_n\|^2 \vee 1) = O(\log n) \text{ and } \mathbf{x}_n^T \Big(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \Big)^{-1} \mathbf{x}_n \to 0 \quad \text{a.s.}, \tag{2.28}$$

then (2.27) yields $\Sigma_{t=1}^{n} (\widehat{\boldsymbol{\beta}}_t^T \mathbf{x}_{t+1} - \boldsymbol{\beta}^T \mathbf{x}_{t+1})^2 = O(\log n)$, which would give the desired logarithmic order (2.22) of the LS certainty equivalence rule $\widehat{\boldsymbol{\beta}}_t^T \mathbf{x}_{t+1} = y_{t+1}^*$ (that can be used to define u_t when $\widehat{\beta}_{1,t} \neq 0$), noting that $\boldsymbol{\beta}^T \mathbf{x}_{t+1} = y_{t+1} - \epsilon_{t+1}$. To ensure (2.28), and to handle the stages t at which $\widehat{\beta}_{1,t}$ vanishes or is close to 0, required much more work, culminating in Lai and Wei (1986a, 1987).

For colored noise and unit delay, (2.14) can still be written as a stochastic regression model $y_t = \boldsymbol{\beta}^T \boldsymbol{\psi}_t + \epsilon_t$, in which $\boldsymbol{\beta}$ is given by (2.15) and

$$\boldsymbol{\psi}_t = (y_{t-1}, \ldots, y_{t-p}, u_{t-1}, \ldots, u_{t-k}, \epsilon_{t-1}, \ldots, \epsilon_{t-h})^T \tag{2.29}$$

contains the unobservable $\epsilon_{t-1}, \ldots, \epsilon_{t-h}$. The SG estimates are given recursively by (2.20) with

$$\mathbf{x}_t = (y_{t-1}, \ldots, y_{t-p}, u_{t-1}, \ldots, u_{t-k}, \widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-h})^T, \text{ where } \widehat{\epsilon}_i = y_i - \widehat{\boldsymbol{\beta}}_{i-1}^T \mathbf{x}_i. \tag{2.30}$$

The extended least squares (ELS) estimates are given recursively by (2.21a, b) with $\mathbf{x}_t$ defined by (2.30), or by a variant that uses $\widehat{\epsilon}_i = y_i - \widehat{\boldsymbol{\beta}}_i^T \mathbf{x}_i$ in (2.30). By making use of extended stochastic Liapounov functions, Lai and Wei (1986a) generalized (2.25) to ELS. Under stability assumptions on $A(z), B(z)$ and $C(z)$ and certain regularity conditions, they also developed an ELS-based certainty-equivalence rule whose regret has the logarithmic order (2.22).

For white noise and unit delay, Lai (1986) studied Bayes control rules in a more tractable setting that assumes $b_1(\neq 0)$ to be known, $y_i^* \equiv 0$, and the ϵ_i to be normal with mean 0 and variance σ^2. Putting a truncated normal prior distribution π on $\boldsymbol{\lambda} := b_1^{-1}(-a_1, \dots, -a_p, b_2, \dots, b_k)^T$, which is the restriction of a standard multivariate normal distribution to the region of λ values for which $A(z)$ and $B(z)$ are stable and the polynomials $z^p\{A(z^{-1}) - 1\}$ and $z^{k-1}B(z^{-1})$ are relatively prime, he showed that

$$\int E_{\boldsymbol{\lambda}}\Big\{\sum_{i=1}^{n}(y_i - \epsilon_i)^2\Big\}d\pi(\boldsymbol{\lambda}) \geq (1 + o(1))\sigma^2(p + k - 1)\log n \tag{2.31}$$

for all input sequences $\{u_n\}$ satisfying (2.19) and the additional growth condition $u_n^2 = O(n^\delta)$ a.s. for some $0 < \delta < 1$. This led him to call an input sequence *asymptotically efficient* if its regret satisfies

$$R_n \leq (1 + o(1))\sigma^2(p + k - 1)\log n \quad \text{a.s.} \tag{2.32}$$

Without assuming b_1 to be known, Lai and Wei (1987) showed how LS certainty-equivalence rules can be modified to make them asymptotically efficient. They also considered the general delay case and constructed adaptive regulators that satisfy

$$\sum_{i=1}^{n}(y_i - \widetilde{\epsilon}_i)^2 \leq (1 + o(1))vd(p + k + d - 2)\log n \quad \text{a.s.}, \tag{2.33}$$

where $\widetilde{\epsilon}_{n+d} = y_{n+d} - E(y_{n+d}|\mathcal{F}_n)$ and $v = \limsup_{n\to\infty} E(\widetilde{\epsilon}_{n+d}^2|\mathcal{F}_n)$, thereby providing a natural extension of (2.32) to the general delay case.

An important feature of the adaptive control rules in Lai and Wei (1986a, 1987) is the use of occasional white-noise probing inputs to resolve the apparent dilemma between the control objective and the need of information for parameter estimation. These white-noise probing inputs yield strongly consistent auxiliary estimates of $\boldsymbol{\beta}$ which can be used to monitor and modify the LS estimates. Further discussion of this is given in Section 3. Whereas the adaptive control rules of Lai and Wei (1986a, 1987) and those of Lai and Robbins (1979, 1981) for the multiperiod control problem also yield strongly consistent estimates of $\boldsymbol{\beta}$, the SG-based controller of Goodwin, Ramage and Caines (1981) do not yield SG estimates $\boldsymbol{\beta}_t$ that converge to $\boldsymbol{\beta}$ a.s. Becker, Kumar and Wei (1985) made use of the self-optimizing property (2.19) of the SG-based control law to show that $\lim_{t\to\infty} \boldsymbol{\beta}_t = \gamma\boldsymbol{\beta}$ a.s., where γ is an a.s. nonzero random variable, thereby establishing the "self-tuning" property that the coefficients of the control law converge a.s. to the optimal values $a_1/b_1, \ldots, a_p/b_1, b_2/b_1, \ldots, b_k/b_1$.

2.4. Weak excitation in stochastic regression models and further applications

The condition (2.26) on the strong consistency of the LS estimate $\widehat{\boldsymbol{\beta}}_n$ in stochastic regression models generalizes (2.11) which, as shown by Lai and Robbins (1981), is in some sense the weakest possible. Its extension to the ELS estimate in the ARMAX system (2.14) paved way for subsequent developments concerning the Åström-Wittenmark regulator that will be reviewed in Section 3. Previous work on strong consistency of the ELS estimate requires a *persistent excitation* condition of the form

$$n^{-1}\sum_{t=1}^{n}\mathbf{x}_t\mathbf{x}_t^T \text{ converges a.s. to a positive definite matrix;} \tag{2.34}$$

see Ljung (1977) or its variant due to Solo (1979) that replaces $\mathbf{x}_t$ by $\boldsymbol{\psi}_t$ defined in (2.29). Lai and Wei (1986b) developed some basic tools for excitation analysis via $\lambda_{\min}(\Sigma_{t=1}^n \boldsymbol{\psi}_t\boldsymbol{\psi}_t^T)$ in ARMAX models. For LS estimates in stochastic regression models, Lai and Wei (1982c, 1983b, 1986b), Lai (1986) and Wei (1985, 1987b) gave further refinements and a variety of statistical applications of (2.25) and related results in Lai and Wei (1982b).

3. Asymptotic Theory of Recursive Estimation and Adaptive Control in ARMAX Systems

While the self-optimizing adaptive control rules of Goodwin, Ramadge and Caines (1981), Lai and Wei (1986a, 1987), and Guo and Chen (1988) were constructed by modifying the certainty equivalence approach proposed by Åström and Wittenmark (1973), the problem concerning the convergence of the original Åström-Wittenmark regulator remained open until Guo and Chen (1991) proved that (2.19) holds and the ELS estimate $\widehat{\boldsymbol{\beta}}_t$ is strongly consistent for the Åström-Wittenmark regulator, provided that $\widehat{b}_{t,1}^2$ is truncated at $1/\log r_{t-1}$ if it falls below that threshold, where $r_n = \Sigma_{i=1}^n \|\mathbf{x}_i\|^2$. Important tools in their analysis are (2.26) and (2.27), besides the Bellman-Gronwall inequality which they applied in an ingenious way to bound $\|\mathbf{x}_t\|^2$ via the dynamics of the controlled system. Such truncation on $\widehat{b}_{t,1}$ is reminiscent of a similar truncation $\zeta_t \vee (\widehat{\beta}_t \wedge \xi_t)$ described in the last paragraph of Section 2.1 that reviews the work of Lai and Robbins (1979, 1981) related to the multiperiod control problem. On the other hand, whereas the logarithmic regret (2.12a) holds for adaptive stochastic approximation schemes that use a truncated LS estimate of the slope β in (2.1), only the self-optimizing property (2.19) was established for the Åström-Wittenmark regulator by Guo and Chen (1991).

Because the adaptive control rules of Lai and Wei (1986a, 1987) involve occasional white-noise inputs to probe the system, they have to assume that $A(z)$ is a stable polynomial, which is a restrictive assumption since an important goal of control theory is to choose inputs suitably to stabilize a system that becomes unstable without such control inputs. Since the Åström-Wittenmark regulator does not require these probing inputs, it can handle the case of unstable $A(z)$. In the white-noise case ($C(z) = 1$), Guo (1995, Thm. 6.3) showed that the Åström-Wittenmark regulator that uses certain modifications of LS estimates can achieve a regret of the order

$$R_n = O(\{1 + \max_{1\le i\le n}[i^\rho d_i \mathrm{tr}(\mathbf{P}_{i-1} - \mathbf{P}_i)]\} \log n) \quad \text{a.s.} \tag{3.1}$$

for every $\rho > 0$, where $(d_i)_{i\ge1}$ is a nonrandom sequence such that $\epsilon_i^2 = O(d_i)$ a.s. and $\mathbf{P}_{i-1} - \mathbf{P}_i$ is given by (2.21b).

For colored noise and general delay, Lai and Ying (1991a) used another way of introducing occasional white-noise probing inputs without requiring $A(z)$ to be stable and thereby constructed an asymptotically efficient adaptive control rule which not only has logarithmic regret but also attains the minimal multiplier of $\log n$ as in (2.32) and (2.33), thus providing an analog of (2.12a) that was established by Lai and Robbins for the multiperiod control problem. In this connection Lai and Ying (1991b) also developed an asymptotic theory of efficient recursive estimation and adaptive prediction for ARMAX systems. The subject of recursive estimation had been described by Åström and Eykhof (1971) as a "fidder's paradise," with a long and growing list of proposed methods and potential possibilities whose statistical properties were relatively unexplored except for a few particular algorithms. Although the theory of stochastic approximation provides important tools to derive and analyze recursive parameter estimates based on i.i.d. observations from parametric families (cf. Chapters 8 and 9 of Nevel'son and Has'minskii (1973)), it is unable to handle the complexity of the interplay between the recursive estimates and system dynamics in ARMAX models. The extended stochastic Liapounov function approach of Lai and Wei described in Section 2.3 paved way for a more powerful approach that led to a unified treatment of recursive estimation by Lai and Ying (1991b) who also introduced parallel recursions to develop asymptotically efficient recursive estimates of the system parameter vector (2.15).

Noting that the classical theory of efficient estimation in ARMAX models involves maximum likelihood, the basic idea of Lai and Ying (1991b) was to modify the traditional off-line algorithms for maximum likelihood estimation into on-line recursive algorithms. Suppose the ϵ_i are i.i.d. $N(0, \sigma^2)$; the case of non-normal ϵ_i can be treated similarly and details are given in the next section. In the white noise case $C(z) = 1$, the maximum likelihood estimates coincide with the LS estimates that have the recursive representation (2.21a, b). For colored noise, ELS uses the same recursions with $\mathbf{x}_t$ given by (2.30), where the $\widehat{\epsilon}_i$ replaces the unobservable ϵ_i in (2.29). The recursive identification literature calls this algorithm RML1 (recursive maximum likelihood of the first kind), and the variant that uses $\widehat{\epsilon}_i = y_i - \widehat{\boldsymbol{\beta}}_i^T \mathbf{x}_i$ in (2.30) is called "approximate maximum likelihood" (AML). Although AML and RML1 have been referred to as recursive versions of the "maximum likelihood" method, their statistical properties are actually unrelated to those of the off-line maximum likelihood estimator, and it is more appropriate to regard them simply as formal extensions of the recursive least squares algorithm (2.21a, b). There is, however, some resemblance between them and the iterative EM algorithm that is sometimes used to compute the off-line maximum likelihood estimator when the ϵ_i are i.i.d. $N(0, \sigma^2)$ random variables. The jth iteration in the EM algorithm to compute the maximum likelihood estimator based on $y_1, u_1, \ldots, y_n, u_n$ consists of a E-step that replaces the unobservable ϵ_i in (2.29) by $\epsilon_i^{(j-1)} := E(\epsilon_i | \boldsymbol{\beta}_{(j-1)}) = y_i - \boldsymbol{\beta}_{(j-1)}^T \boldsymbol{\psi}_{i,j-1}$, where $\boldsymbol{\beta}_{(j-1)}$ denotes the estimate of $\boldsymbol{\beta}$ after $j-1$ iterations and

$$\boldsymbol{\psi}_{i,j-1} = (y_{i-1}, \ldots, y_{i-p}, u_{i-d}, \ldots, u_{i-d-k+1}, \epsilon_{i-1}^{(j-1)}, \ldots, \epsilon_{i-h}^{(j-1)})^T. \tag{3.2}$$

It then revises the estimate by an M-step that gives the least squares solution $\boldsymbol{\beta}_{(j)} = (\Sigma_{i=1}^n \boldsymbol{\psi}_{i,j-1} \boldsymbol{\psi}_{i,j-1}^T)^{-1} \Sigma_{i=1}^n \boldsymbol{\psi}_{i,j-1} y_i$. The recursions for AML or RML1 simply replace $\boldsymbol{\psi}_{i,j-1}$ in (3.2) by $\mathbf{x}_i$, without any iterative refinement and without updating the estimates of ϵ_i when new data are available. This explains why the AML or RML1 algorithm fails to inherit the statistical properties of the maximum likelihood estimator, and can even run into serious difficulties when $C(z)$ differs so much from 1 (the white-noise case) that the positive real assumption

$$C(z) \text{ is stable and } \min_{|t| \leq \pi} \mathrm{Re}\{\frac{1}{C(e^{it})} - \frac{1}{2}\} > 0, \tag{3.3}$$

which has been assumed for the convergence analysis of ELS in Solo (1979) and Lai and Wei (1986a), is violated.

Given the initial condition $\boldsymbol{\psi}_0$ and the parameter value $\boldsymbol{\beta}$, the ϵ_i in (2.29) can be expressed as a function $\epsilon_i(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$, which yields (2.29) with $\boldsymbol{\psi}_t$ being a vector-valued function $\boldsymbol{\psi}_t(\boldsymbol{\beta})$. The log-likelihood function at stage n is

$$\ell_n(\boldsymbol{\beta}) = -\sum_{i=1}^n \frac{(y_i - \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta}))^2}{2\sigma^2} + \text{ constant}. \tag{3.4}$$

Differentiation of (3.4) and (2.14) with respect to $\boldsymbol{\beta}$ yields

$$\nabla \ell_n(\boldsymbol{\beta}) = \sigma^{-2} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta})) \nabla(\boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta})), \tag{3.5}$$

$$\nabla(\boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta})) + c_1 \nabla(\boldsymbol{\beta}^T \boldsymbol{\psi}_{i-1}(\boldsymbol{\beta})) + \cdots + c_h \nabla(\boldsymbol{\beta}^T \boldsymbol{\psi}_{i-h}(\boldsymbol{\beta})) = \boldsymbol{\psi}_i(\boldsymbol{\beta}). \tag{3.6}$$

The off-line maximum likelihood estimator is determined by numerical solution of the equation $\nabla \ell_n(\boldsymbol{\beta}) = \mathbf{0}$ using iterative schemes. From (3.5) it follows that

$$\sigma^2 \nabla^2 \ell_n(\boldsymbol{\beta}) = -\sum_{i=1}^{n} (\nabla \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta}))(\nabla \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta}))^T + \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta})) \nabla^2(\boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta})). \tag{3.7}$$

The Gauss-Newton scheme to solve $\nabla \ell_n(\boldsymbol{\beta}) = \mathbf{0}$ starts with an initial guess $\boldsymbol{\beta}_n^{(0)}$ of $\boldsymbol{\beta}$ and defines the jth iterative step by setting $\boldsymbol{\beta}_n^{(j)} - \boldsymbol{\beta}_n^{(j-1)}$ equal to

$$\Big[\sum_{i=1}^{n} (\nabla \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta}))(\nabla \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta}))^T \Big]^{-1}_{\boldsymbol{\beta} = \boldsymbol{\beta}_n^{(j-1)}} \times \Big[\sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta}))(\nabla \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta})) \Big]_{\boldsymbol{\beta} = \boldsymbol{\beta}_n^{(j-1)}}.$$

Define $\mathbf{x}_n$ by (2.30) and let $c_{n-1,1}, \ldots, c_{n-1,h}$ be the estimates of $c_1, \ldots, c_h$ given by $\widehat{\boldsymbol{\beta}}_{n-1}$. A one-step implementation of the iterative scheme which initializes at $\boldsymbol{\beta}_n^{(0)} = \widehat{\boldsymbol{\beta}}_{n-1}$ and which approximates $\nabla \ell_{n-1}(\widehat{\boldsymbol{\beta}}_{n-1})$ by $\mathbf{0}$, $y_n - \widehat{\boldsymbol{\beta}}_{n-1}^T \psi_n(\widehat{\boldsymbol{\beta}}_{n-1})$ by $\widehat{\epsilon}_n$ that is defined in (2.30), and $\nabla(\boldsymbol{\beta}^T \boldsymbol{\psi}_n(\boldsymbol{\beta}))|_{\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{n-1}}$ by $\boldsymbol{\phi}_n$ that is defined like (3.6) via the recursion

$$\boldsymbol{\phi}_n + c_{n-1,1} \boldsymbol{\phi}_{n-1} + \cdots + c_{n-1,h} \boldsymbol{\phi}_{n-h} = \mathbf{x}_n, \tag{3.8a}$$

yields the RML2 algorithm (recursive maximum likelihood of the second kind; see Ljung and Söderström (1983, pp.26-30)):

$$\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\beta}}_{n-1} + \mathbf{P}_n \boldsymbol{\phi}_n \widehat{\epsilon}_n, \tag{3.8b}$$

$$\mathbf{P}_n = \mathbf{P}_{n-1} - \frac{\mathbf{P}_{n-1} \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \mathbf{P}_{n-1}}{(1 + \boldsymbol{\phi}_n^T \mathbf{P}_{n-1} \boldsymbol{\phi}_n)}, \tag{3.8c}$$

where $\mathbf{P}_0$ is a positive definite matrix.

While the RML2 algorithm circumvents the computational complexity in updating the off-line maximum likelihood estimator by using the preceding approximations, these approximations may be poor if $\widehat{\boldsymbol{\beta}}_{n-1}$ differs substantially from $\boldsymbol{\beta}$, and no consistency or asymptotic normality results have been established for it. To ensure that $\widehat{\boldsymbol{\beta}}_{n-1}$ is eventually close to $\boldsymbol{\beta}$, Lai and Ying (1991b) suggested the following modification of (3.8b). Let $\widetilde{\boldsymbol{\beta}}_n$ be a consistent recursive estimate of $\boldsymbol{\beta}$ such that $\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = o(\delta_n)$ a.s., where δ_n can be computed from the input-output data up to stage n and converges to 0 a.s. In particular, we can use the generalized method of moments (GMM) that will be described in Section 4.3 to construct such $\widetilde{\boldsymbol{\beta}}_n$. Let S_n be a cube with center $\widetilde{\boldsymbol{\beta}}_n$ and width δ_n. Since $\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\| = o(\delta_n)$ a.s.,

$$P\{\boldsymbol{\beta} \in S_n \text{ for all large } n\} = 1. \tag{3.9}$$

The consistent estimators $\widetilde{\boldsymbol{\beta}}_n$ and the associated confidence sets S_n are used in Lai and Ying (1991b) to monitor the RML2 algorithm (3.8a, b, c) and need only be updated occasionally at times $m_1 < m_2 < \ldots$ The basic idea is to constrain (monitor) the algorithm so that it lies inside S_{m_j} for $m_j \le n < m_{j+1}$ by using at stage n the projection $\boldsymbol{\pi}_n$ with respect to the norm of the matrix $\mathbf{P}_n^{-1} = \mathbf{P}_0^{-1} + \Sigma_{i=1}^n \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T$. Specifically, for $\mathbf{x} \in \mathbf{R}^{p+k+h}$ and $m_j \le n < m_{j+1}$, let $\boldsymbol{\pi}_n(\mathbf{x})$ denote the unique solution of the quadratic programming problem

$$(\boldsymbol{\pi}_n(\mathbf{x}) - \mathbf{x})^T \mathbf{P}_n^{-1} (\boldsymbol{\pi}_n(\mathbf{x}) - \mathbf{x}) = \min_{\mathbf{y} \in S_{m_j}} \{(\mathbf{y} - \mathbf{x})^T \mathbf{P}_n^{-1} (\mathbf{y} - \mathbf{x})\}. \tag{3.10}$$

It is convenient to choose S_{m_j} to be a cube so that we have linear constraints for (3.10). The *monitored recursive maximum likelihood estimate* introduced by Lai and Ying (1991b) replaces (3.8b) in RML2 by

$$\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\pi}_n(\widehat{\boldsymbol{\beta}}_{n-1} + \mathbf{P}_n \boldsymbol{\phi}_n \widehat{\epsilon}_n), \tag{3.11}$$

for which they showed that on $\{\lambda_{\max}(\mathbf{P}_n^{-1}) \to \infty \text{ and } \boldsymbol{\phi}_n^T \mathbf{P}_n \boldsymbol{\phi}_n \to 0\}$,

$$\begin{aligned} &\sum_{i=1}^n (\widehat{\boldsymbol{\beta}}_{i-1}^T \mathbf{x}_i - \boldsymbol{\beta}^T \boldsymbol{\psi}_i)^2 \\ &\le (\sigma^2 + o(1)) \log \det\Big(\sum_{i=1}^n \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T\Big) + o\Big(\sum_{j: m_j \le n} \sum_{r=1}^h (\|\boldsymbol{\phi}_{m_j - r}\| \vee 1)^2\Big) \quad \text{a.s.} \end{aligned} \tag{3.12}$$

Note that (3.12) gives an asymptotic bound on the cumulative squared difference between the minimum-variance $\mathcal{F}_{i-1}$-predictor $\boldsymbol{\beta}^T \boldsymbol{\psi}_i$ of y_i (when $\boldsymbol{\beta}$ is known and $\boldsymbol{\psi}_i$ is observable) and the adaptive predictor $\widehat{\boldsymbol{\beta}}_{i-1}^T \mathbf{x}_i$, analogous to (2.27).

In the case of unit delay ($d = 1$), if it can be shown that with probability $1, \Sigma_{i=1}^n \|\boldsymbol{\phi}_i\|^2 \to \infty$ but is bounded by $n^{1+o(1)}$, while $\boldsymbol{\phi}_n^T \mathbf{P}_n \boldsymbol{\phi}_n \to 0$ and

$$\sum_{j: m_j \le n} \sum_{r=1}^h (\|\boldsymbol{\phi}_{m_j - r}\|^2 \vee 1)^2 = O(\log n). \tag{3.13}$$

Then, in view of (3.12), the certainty equivalence regulator $\widehat{\boldsymbol{\beta}}_t^T \mathbf{x}_{t+1} = 0$ (that can be used to define u_t when $\widehat{b}_{1,t} \neq 0$) would have a regret satisfying

$$R_n \leq (1 + o(1))\sigma^2\{(p \vee h) + k - 1\} \log n \quad \text{a.s.}, \tag{3.14}$$

which agrees with (2.32) in the case $h = 0$. Lai and Ying (1991a) proved (3.14) for a modified version of the RML2-based certainty equivalence rule. An important ingredient of this asymptotically efficient adaptive regulator is an auxiliary consistent recursive estimate $\widetilde{\boldsymbol{\beta}}_t$ which can be used to monitor RML2, and which is constructed by the generalized method of moments (see Section 4.3) applied to occasional blocks of well-excited input-output data. To generate these well excited blocks, the stochastic gradient (SG) algorithm (2.20) is also run in parallel, with $\mathbf{x}_t$ defined by (2.30) in which $\widehat{\boldsymbol{\beta}}_{i-1}$ should be replaced here by the SG estimate $\boldsymbol{\beta}_{i-1}$. Since SG has been shown by Goodwin, Ramage and Caines (1981) to be able to stabilize the system even though $A(z)$ may be unstable, occasional blocks of white noise probing inputs can be introduced via the *disturbed control* scheme of Caines and Lafortune (1984):

$$\boldsymbol{\beta}_t^T(y_t, \ldots, y_{t-p+1}, u_t, \ldots, u_{t-k+1}, w_t, \ldots, w_{t-h}) = w_n \tag{3.15}$$

for $n_j \leq n < m_j$, where w_i are i.i.d. bounded random variables with $Ew_1 = 0, Ew_1^2 > 0$, and $n_1 < m_1 < n_2 < m_2 \ldots$ are suitably chosen stopping time such that $m_j - n_j \to \infty$ and $\Sigma_{j=1}^t (m_j - n_j) = o(\log t)$. For $t \notin \cup_j\{n : n_j \leq n < m_j\}$, the RML2-based certainty equivalence rule $\widehat{\boldsymbol{\beta}}_t^T \mathbf{x}_{t+1} = 0$ is used if $\widehat{b}_{1,t} \neq 0$, setting $u_t = w_t$ if $\widehat{b}_{1,t} = 0$.

Lai and Ying (1991a) also extended this idea to general target values y_t^* and delay $d > 1$ by using the so-called "implicit approach" in the adaptive control literature and still obtained (3.14) but with a somewhat larger multiplicative constant. The preceding approach for unit delay is called *explicit* (or *indirect*); see Section 7 of Lai and Ying (1991a).

4. Efficient Semiparametric Estimation in Linear Time Series Models

Although the monitored RML2 algorithm in the preceding section is motivated by maximum likelihood estimation when the ϵ_i are i.i.d. normal, the same algorithm and its associated adaptive control rule can still be applied to more general random disturbances ϵ_i. In fact, Lai and Ying (1991a, b) assume that $\{\epsilon_n, \mathcal{F}_n, n \geq 1\}$ is a martingale difference sequence such that $E(\epsilon_n^2|\mathcal{F}_{n-1}) = \sigma^2 > 0$ and $\sup_n E(|\epsilon_n|^r|\mathcal{F}_{n-1}) < \infty$ a.s. for some $r > 2$, and derive results such as (3.14) in this setting. On the other hand, because of the complexity of the feedback control scheme, they have not provided traditional asymptotic normality results that show the asymptotic efficiency of the recursive approximations to the off-line maximum likelihood estimates. In this section we consider estimation in simpler designs for ARMAX models, e.g., estimation in open loop for which the $\boldsymbol{\psi}_t$ in (2.29), with $u_{t-1}, \ldots, u_{t-k}$ replaced by $u_{t-d}, \ldots, u_{t-d-k+1}$ for general delay, have good excitation properties to yield $\sqrt{n}$-consistency and asymptotic normality of the recursive estimates.

To begin with, we replace the projection $\boldsymbol{\pi}_n$ in (3.11) by a simpler modification of RML2 when the input-output data have good excitation properties. In Lai and Ying (1991b), the projection $\boldsymbol{\pi}_n$ (with respect to the norm of $\mathbf{P}_n^{-1}$) is used to bound the extended stochastic Liapounov function $V_n := (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^T \mathbf{P}_n^{-1} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ by

$$V_n \leq (\widehat{\boldsymbol{\beta}}_{n-1} - \boldsymbol{\beta} + \mathbf{P}_n \boldsymbol{\phi}_n \widehat{\epsilon}_n)^T \mathbf{P}_n^{-1} (\widehat{\boldsymbol{\beta}}_{n-1} - \boldsymbol{\beta} + \mathbf{P}_n \boldsymbol{\phi}_n \widehat{\epsilon}_n). \tag{4.1}$$

Instead of the projection $\boldsymbol{\pi}_n$, it is simpler to directly apply the auxiliary estimate $\widetilde{\boldsymbol{\beta}}_n$ that is used to monitor the RML2 algorithm as follows:

$$\widehat{\boldsymbol{\beta}}_n = \begin{cases} \widehat{\boldsymbol{\beta}}_{n-1} + \mathbf{P}_n \boldsymbol{\phi}_n \widehat{\epsilon}_n \text{ if } \|\widehat{\boldsymbol{\beta}}_{n-1} + \mathbf{P}_n \boldsymbol{\phi}_n \widehat{\epsilon}_n - \widetilde{\boldsymbol{\beta}}_n\| \leq \delta_n, \\ \widetilde{\boldsymbol{\beta}}_n \text{ otherwise.} \end{cases} \tag{4.2}$$

With probability 1, since $\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = o(\delta_n)$, it follows that for all large n, (4.1) still holds if $\|\widehat{\boldsymbol{\beta}}_{n-1} + \mathbf{P}_n \boldsymbol{\phi}_n \widehat{\epsilon}_n - \widetilde{\boldsymbol{\beta}}_n\| > \delta_n$ and $\lambda_{\max}(\mathbf{P}_n^{-1}) = O(\lambda_{\min}(\mathbf{P}_n^{-1}))$. In particular, (4.1) still holds for all large n if $\widehat{\boldsymbol{\beta}}_n$ is defined by (4.2) and

$$n^{-1} \sum_{i=1}^{n} \boldsymbol{\zeta}_i \boldsymbol{\zeta}_i^T \text{ converges a.s. to a positive definite nonrandom matrix } \mathbf{A}, \tag{4.3}$$

where $\boldsymbol{\zeta}_i + c_1 \boldsymbol{\zeta}_{i-1} + \ldots c_h \boldsymbol{\zeta}_{i-h} = \boldsymbol{\psi}_i$. In this section, the excitation condition (4.3) is assumed and the "monitored RML2 algorithm" refers to (4.2). In addition, $A(z)$ and $C(z)$ are assumed to be stable and the ϵ_n in (2.14) are assumed to be i.i.d. with mean 0 and variance σ^2.

We next extend the monitored RML2 algorithm to i.i.d. ϵ_i that have a common density function f, not necessarily normal, with respect to Lebesgue measure. In this case the log-likelihood function (3.4) has the more general form

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i - \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta})) + \text{ constant.} \tag{4.4}$$

Assuming f to be continuously differentiable and letting $g = -f'/f$, the off-line maximum likelihood estimator solves the estimating equation

$$\sum_{i=1}^{n} g(y_i - \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta})) \nabla(\boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta})) = 0, \tag{4.5}$$

and $\nabla(\boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta}))$ still satisfies (3.6). Instead of maximum likelihood assuming pre-specified f, Martin and Yohai (1985) proposed to use more robust M-estimators in which g in (4.5) is a given score function, not necessarily of the form $-f'/f$.

4.1. Monitored recursive M-estimators

In the case where $g' \geq 0$, the Gauss-Newton method to solve (4.5) starts with an initial guess $\boldsymbol{\beta}_n^{(0)}$ of $\boldsymbol{\beta}$ and defines the $(j+1)$st iterative step by

$$\boldsymbol{\beta}_n^{(j+1)} = \boldsymbol{\beta}_n^{(j)} + \Big\{ \sum_{i=1}^n g'(e_{i,n}^{(j)})[(\nabla \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta}))(\nabla \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta}))^T] \Big\}^{-1} \times \Big\{ \sum_{i=1}^n g(e_{i,n}^{(j)}) \nabla(\boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta})) \Big\} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n^{(j)}}, \tag{4.6}$$

where $e_{i,n}^{(j)} = y_i - \boldsymbol{\beta}^T \boldsymbol{\psi}_i(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n^{(j)}}$. When $g(x) = x$ (so $g' \equiv 1$), the iterative scheme (4.6) reduces to the Gauss-Newton scheme described in Section 3. When $g = -f'/f, Eg(\epsilon_i) = 0$ and $Eg'(\epsilon_i) = Eg^2(\epsilon_i) > 0$, and one usually replaces g' in (4.6) by g^2 in this case. We therefore assume that

$$Eg(\epsilon_1) = 0 \text{ and } E\widetilde{g}(\epsilon_1) = Eg'(\epsilon_1) > 0, \tag{4.7}$$

where $\widetilde{g}$ is some known nonnegative function. Note that if g is nondecreasing, then we can simply set $\widetilde{g} = g'$.

To develop an M-estimation version of (4.2), we recursify the Gauss-Newton algorithm (4.6) as before, but with $\mathbf{P}_n^{-1}$ replaced by

$$\widetilde{\mathbf{P}}_n^{-1} = \widetilde{\mathbf{P}}_{n-1}^{-1} + \widetilde{g}(\widehat{\epsilon}_{n-1}) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T. \tag{4.8}$$

The monitored recursive M-estimator is defined by (4.2), in which $\mathbf{P}_n^{-1}$ is replaced by $\widetilde{\mathbf{P}}_n^{-1}$ and $\widehat{\epsilon}_n$ by $g(\widehat{\epsilon}_n)$. Specifically,

$$\widehat{\boldsymbol{\beta}}_n = \begin{cases} \widehat{\boldsymbol{\beta}}_{n-1} + \widetilde{\mathbf{P}}_n \boldsymbol{\phi}_n g(\widehat{\epsilon}_n) \text{ if } \|\widehat{\boldsymbol{\beta}}_{n-1} + \widetilde{\mathbf{P}}_n \boldsymbol{\phi}_n g(\widehat{\epsilon}_n) - \widetilde{\boldsymbol{\beta}}_n\| \leq \delta_n, \\ \widetilde{\boldsymbol{\beta}}_n \text{ otherwise.} \end{cases} \tag{4.9}$$

Clearly $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}$. From (4.3), it follows that

$$n^{-1} \widetilde{\mathbf{P}}_n^{-1} \to \widetilde{\boldsymbol{A}} = \boldsymbol{A} E\widetilde{g}(\epsilon_1) = \boldsymbol{A} Eg'(\epsilon_1). \tag{4.10}$$

We next use the extended stochastic Liapounov function approach to analyze the monitored recursive M-estimator. Define $V_n = (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^T \widetilde{\mathbf{P}}_n^{-1} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. As pointed out in the paragraph containing (4.1), the recursion (4.9) yields that with probability 1, for all large n,

$$V_n \le (\widehat{\boldsymbol\beta}_{n-1} - \boldsymbol\beta + \widetilde{\mathbf P}_n \boldsymbol\phi_n g(\widehat\epsilon_n))^T \widetilde{\mathbf P}_n^{-1} (\widehat{\boldsymbol\beta}_{n-1} - \boldsymbol\beta + \widetilde{\mathbf P}_n \boldsymbol\phi_n g(\widehat\epsilon_n))$$
$$= V_{n-1} + 2\boldsymbol\phi_n^T(\widehat{\boldsymbol\beta}_{n-1} - \boldsymbol\beta) g(\widehat\epsilon_n) + [\boldsymbol\phi_n^T(\widehat{\boldsymbol\beta}_{n-1} - \boldsymbol\beta)]^2 \widetilde g(\widehat\epsilon_n) + \boldsymbol\phi_n^T \widetilde{\mathbf P}_n \boldsymbol\phi_n g^2(\widehat\epsilon_n). \quad (4.11)$$

Since $g(\widehat\epsilon_n) = g(\widehat\epsilon_n - \epsilon_n + \epsilon_n)$ and $\widehat\epsilon_n - \epsilon_n \in \mathcal F_{n-1}$, it follows that

$$E[g(\widehat\epsilon_n)|\mathcal F_{n-1}] = \int g(\widehat\epsilon_n - \epsilon_n + u) f(u) du$$
$$= \left\{ \int g'(t^*(\widehat\epsilon_n - \epsilon_n) + u) f(u) du \right\} (\widehat\epsilon_n - \epsilon_n),$$

where $t^* \in (0,1)$. Because $C(q^{-1})(\widehat\epsilon_n - \epsilon_n) = -(\widehat{\boldsymbol\beta}_{n-1} - \boldsymbol\beta)^T \boldsymbol\phi_n$, $C(q^{-1})$ is stable and $\widehat{\boldsymbol\beta}_n$ is consistent, we have

$$\sum_{i=1}^n \boldsymbol\phi_i^T(\widehat{\boldsymbol\beta}_{i-1} - \boldsymbol\beta) g(\widehat\epsilon_i)$$
$$= \sum_{i=1}^n \boldsymbol\phi_i^T(\widehat{\boldsymbol\beta}_{i-1} - \boldsymbol\beta)(\widehat\epsilon_i - \epsilon_i) E g'(\epsilon_1) + o(1) \sum_{i=1}^n [\boldsymbol\phi_i^T(\widehat{\boldsymbol\beta}_{i-1} - \boldsymbol\beta)]^2$$
$$= -(1+o(1)) \sum_{i=1}^n [\boldsymbol\phi_i^T(\widehat{\boldsymbol\beta}_{i-1} - \boldsymbol\beta)]^2 E\widetilde g(\epsilon_1) + o(1) \sum_{i=1}^n \boldsymbol\phi_i^T \widetilde{\mathbf P}_i \boldsymbol\phi_i^T \quad \text{a.s.} \quad (4.12)$$

Combining (4.11) and (4.12) yields

$$V_n + \sum_{i=1}^n [\boldsymbol\phi_i^T(\widehat{\boldsymbol\beta}_{i-1} - \boldsymbol\beta)]^2 E\widetilde g(\epsilon_1) \le (1+o(1)) \sum_{i=1}^n \boldsymbol\phi_i^T \widetilde{\mathbf P}_i \boldsymbol\phi_i E g^2(\epsilon_1) \quad \text{a.s.} \quad (4.13)$$

Since $\widetilde{\mathbf P}_n^{-1} = (E\widetilde g(\epsilon_1) + o(1)) \mathbf P_n^{-1}$ a.s., it follows that with probability 1,

$$\sum_{i=1}^n \boldsymbol\phi_i^T \widetilde{\mathbf P}_i \boldsymbol\phi_i = (1+o(1))[E\widetilde g(\epsilon_1)]^{-1} \sum_{i=1}^n \boldsymbol\phi_i^T \mathbf P_i \boldsymbol\phi_i$$
$$= (1+o(1))[E\widetilde g(\epsilon_1)]^{-1} \log \mathbf P_n^{-1}$$
$$= (1+o(1))(p+k+h)[E\widetilde g(\epsilon_1)]^{-1} \log n.$$

Putting this in (4.13), we obtain

$$V_n + E\widetilde g(\epsilon_1) \sum_{i=1}^n (\widehat\epsilon_n - \epsilon_n)^2 \le (1+o(1)) E g^2(\epsilon_1) [E\widetilde g(\epsilon_1)]^{-1} (p+k+h) \log n \quad \text{a.s.} \quad (4.14)$$

From (4.14), it follows that with probability 1, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|^2 = O(\log n/n)$, which implies that the monitoring is not used for all large n, provided that $\delta_n = o(n^{1/2-\alpha})$ for some $\alpha > 0$. In other words, $\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\beta}}_{n-1} + \widetilde{\mathbf{P}}_n \boldsymbol{\phi}_n g(\epsilon_n)$ for all large n. This recursion leads to

$$\begin{aligned}
\widetilde{\mathbf{P}}_n^{-1}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) &= \widetilde{\mathbf{P}}_n^{-1}(\widehat{\boldsymbol{\beta}}_{n-1} - \boldsymbol{\beta} + \widetilde{\mathbf{P}}_n \boldsymbol{\phi}_n g(\widehat{\epsilon}_n)) \\
&= \widetilde{\mathbf{P}}_{n-1}^{-1}(\widehat{\boldsymbol{\beta}}_{n-1} - \boldsymbol{\beta}) + \boldsymbol{\phi}_n[\boldsymbol{\phi}_n^T(\widehat{\boldsymbol{\beta}}_{n-1} - \boldsymbol{\beta})\widetilde{g}(\widehat{\epsilon}_n) + g(\widehat{\epsilon}_n)] \\
&= \sum_{i=1}^n \boldsymbol{\phi}_i[\boldsymbol{\phi}_i^T(\widehat{\boldsymbol{\beta}}_{i-1} - \boldsymbol{\beta})\widetilde{g}(\widehat{\epsilon}_i) + g(\widehat{\epsilon}_i) - g(\epsilon_i)] + \sum_{i=1}^n \boldsymbol{\phi}_i g(\epsilon_i) + O(1) \quad \text{a.s.}
\end{aligned}$$

Using $g(\widehat{\epsilon}_i) - g(\epsilon_i) \cong g'(\epsilon)(\widehat{\epsilon}_i - \epsilon_i)$, it can be shown that the first term on right-hand side of the above equality is of order $o(\sqrt{n})$.

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = n\widetilde{\mathbf{P}}_n \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\phi}_i g(\epsilon_i) + o(1) \Rightarrow N\left(0, \ \ \boldsymbol{A}^{-1}[E\widetilde{g}(\epsilon_1)]^{-2} E g^2(\epsilon_1)\right), \tag{4.15}$$

establishing the asymptotic normality of the monitored recursive estimator.

4.2. Adaptive semiparametric estimation

The subject of adaptive estimation began with Stein's (1956) work on estimating and testing hypotheses about a Euclidean parameter $\boldsymbol{\beta}$ or, more generally, a function $h(\boldsymbol{\beta})$ in the presence of an infinite-dimensional nuisance parameter G. He considered the problem of when and how one can estimate $\boldsymbol{\beta}$ when G is unknown as well asymptotically as when G is known. In the case of i.i.d. observations, Bickel (1982) and Fabian and Hannan (1982) provided general solutions to this problem. Their theory was subsequently extended by Kreiss (1987) to study adaptive estimation of the parameters of a stationary ARMA model. We consider the somewhat more general case of the ARMAX model (2.14).

For the ARMA model (i.e., $B(q^{-1}) = 0$), Kreiss (1987) made use of the stability of $A(z)$ and $C(z)$ to derive the following power series representations for some $\eta > 1$:

$$\frac{1}{A(z)} = \sum_{j=0}^\infty \alpha_j(\boldsymbol{\beta}) z^j, \quad \frac{1}{C(z)} = \sum_{j=0}^\infty \gamma_j(\boldsymbol{\beta}) z^j, \quad \text{for all } |z| \le \eta. \tag{4.16}$$

With $\gamma_j(\boldsymbol{\beta})$ defined from the invertibility of $C(z)$, let

$$\mathbf{Z}_{i-1}(\boldsymbol{\beta}) = \sum_{i=0}^{t-1} \gamma_i(\boldsymbol{\beta}) \boldsymbol{\psi}_{t-i}(\boldsymbol{\beta}), \tag{4.17}$$

which plays a basic role in Kreiss' construction of adaptive estimators. Under the identifiability assumption that $a_p \neq 0, c_h \neq 0$ and $A(z), C(z)$ have no common zero and under some additional assumptions, Kreiss (1987) showed that when f is unknown, it is possible to construct adaptive estimators which have the same asymptotic optimality properties as the maximum likelihood estimator that assumes f to be known and that is constrained to lie in a $O(n^{-1/2})$-neighborhood of the true parameter $\boldsymbol{\beta}$. Starting with a $\sqrt{n}$-consistent preliminary estimator $\boldsymbol{\beta}_n^*$ of $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}_n^*$ can assume values in a prescribed discrete set, he defined an adaptive estimator $\widehat{\boldsymbol{\beta}}_n$ by a linear approximation around $\boldsymbol{\beta}_n^*$ to the estimating equation (4.5) in which the unknown g is replaced by an estimate $\widehat{g}_{n,i}$:

$$\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_n^* + \left\{ \widehat{I}_n \sum_{t=1}^{n} \boldsymbol{Z}_{t-1}(\boldsymbol{\beta}_n^*) \boldsymbol{Z}_{t-1}^T(\boldsymbol{\beta}_n^*) \right\}^{-1} \sum_{i=1}^{n} \widehat{g}_{n,i}(\epsilon_i(\boldsymbol{\beta}_n^*)) \mathbf{Z}_{i-1}^T(\boldsymbol{\beta}_n^*), \tag{4.18}$$

where $\widehat{I}_n$ is an estimate of $I(f) = \int (f'/f)^2 f dx$. The estimator $\widehat{g}_{n,i}(x)$ of $g(x)$ in (4.18) depends on the $n-1$ quantities $\epsilon_1(\boldsymbol{\beta}_n^*), \ldots, \epsilon_{i-1}(\boldsymbol{\beta}_n^*)$, $\epsilon_{i+1}(\boldsymbol{\beta}_n^*), \ldots, \epsilon_n(\boldsymbol{\beta}_n^*)$.

The adaptive estimator (4.18) uses ideas similar to those of Bickel (1982) and Fabian and Hannan (1982) in the i.i.d. case, but is much more complicated because it involves the inversion (4.16) of the operator $C(q^{-1})$ for evaluating the $\mathbf{Z}_{t-1}(\boldsymbol{\beta}_n^*)$ in (4.18). This is particularly inconvenient when one needs to update (4.18) sequentially whenever new data become available. The heavy computational burden for (4.18) is in sharp contrast to the recursive "on-line" estimators emphasized in the engineering literature, where the primary purpose of parameter estimation for the ARMAX system is to support decisions that have to be taken during the operation of the system.

We next construct an "on-line" adaptive procedure that achieves the semiparametric efficiency. Following Kreiss (1987), we assume that f is symmetric about the origin. We also assume that f is differentiable as many times, and has as many moments, as needed. We again, as in Section 4.1, require the availability of the auxiliary estimate $\widetilde{\boldsymbol{\beta}}_n$ to monitor the recursion. Define the normalized efficient score

$$g_e(u) = \frac{-\frac{f'(u)}{f(u)}}{E\left[\frac{f'(\epsilon_1)}{f(\epsilon_1)}\right]^2}, \tag{4.19}$$

which is "normalized" because $E[g_e'(\epsilon_1)] = 1$. Suppose that we have a consistent estimate $\widehat{g}_n \in \mathcal{F}_{n-1}$ of g_e and that $\widehat{g}_n$ is an odd function with derivative $\widehat{g}_n'$ converging to g_e'. Such an estimate may be constructed using residuals from either the auxiliary estimate $\widetilde{\beta}_i$ or simply $\widehat{\beta}_i$, $i \leq n-1$. Define, recursively,

$$\mathbf{P}_n^{-1} = \mathbf{P}_{n-1}^{-1} + \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T, \tag{4.20a}$$

$$\widehat{\boldsymbol{\beta}}_n = \begin{cases} \widehat{\boldsymbol{\beta}}_{n-1} + \mathbf{P}_n\boldsymbol{\phi}_n\widehat{g}_n(\widehat{\epsilon}_n) \text{ if } \|\widehat{\boldsymbol{\beta}}_{n-1} + \mathbf{P}_n\boldsymbol{\phi}_n\widehat{g}_n(\widehat{\epsilon}_n) - \widetilde{\boldsymbol{\beta}}_n\| \le \delta_n, \\ \widetilde{\boldsymbol{\beta}}_n \text{ otherwise.} \end{cases} \tag{4.20b}$$

To derive the asymptotic properties of this recursive adaptive estimator, we again use the extended stochastic Liapounov function $V_n = (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^T\mathbf{P}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. Similar to the derivation in Section 4.1, with probability 1,

$$V_n \le V_{n-1} + [\boldsymbol{\phi}_n^T(\widehat{\boldsymbol{\beta}}_{n-1} - \boldsymbol{\beta})]^2 + 2\boldsymbol{\phi}_n^T(\widehat{\boldsymbol{\beta}}_{n-1} - \boldsymbol{\beta})\widehat{g}_n(\widehat{\epsilon}_n) + \boldsymbol{\phi}_n^T\mathbf{P}_n\boldsymbol{\phi}_n\widehat{g}_n^2(\widehat{\epsilon}_n) \tag{4.21}$$

for all large n. Moreover, it can be shown that

$$\sum_{i=1}^n \boldsymbol{\phi}_i^T(\widehat{\boldsymbol{\beta}}_{i-1} - \boldsymbol{\beta})\{\widehat{g}_i(\widehat{\epsilon}_i) - E[\widehat{g}_i(\widehat{\epsilon}_i)|\mathcal{F}_{i-1}]\} = o(\sum_{i=1}^n [\boldsymbol{\phi}_i^T(\widehat{\boldsymbol{\beta}}_{i-1} - \boldsymbol{\beta})]^2) \quad \text{a.s.}. \tag{4.22}$$

In addition, since $\widehat{g}_i \in \mathcal{F}_{i-1}$, $\widehat{\epsilon}_i - \epsilon_i \in \mathcal{F}_{i-1}$ and $\int \widehat{g}_i(u)f(u)du = 0$,

$$\begin{aligned} &\sum_{i=1}^n \boldsymbol{\phi}_i^T(\widehat{\boldsymbol{\beta}}_{i-1} - \boldsymbol{\beta})E[\widehat{g}_i(\widehat{\epsilon}_i)|\mathcal{F}_{i-1}] \\ &= \sum_{i=1}^n \boldsymbol{\phi}_i^T(\widehat{\boldsymbol{\beta}}_{i-1} - \boldsymbol{\beta})\int \widehat{g}_i(\widehat{\epsilon}_i - \epsilon_i + u)f(u)du \\ &= (1+o(1))\sum_{i=1}^n \int \widehat{g}_i'(u)f(u)du\boldsymbol{\phi}_i^T(\widehat{\boldsymbol{\beta}}_{i-1} - \boldsymbol{\beta})(\widehat{\epsilon}_i - \epsilon_i) \\ &= (1+o(1))\sum_{i=1}^n \boldsymbol{\phi}_i^T(\widehat{\boldsymbol{\beta}}_{i-1} - \boldsymbol{\beta})(\widehat{\epsilon}_i - \epsilon_i) \\ &= -(1+o(1))\sum_{i=1}^n (\widehat{\epsilon}_i - \epsilon_i)^2 \quad \text{a.s.}. \end{aligned}$$

Combining this with (4.21) and (4.22) yields

$$\begin{aligned} V_n + \sum_{i=1}^n (\widehat{\epsilon}_i - \epsilon_i)^2 &\le (1+o(1))\sum_{i=1}^n \boldsymbol{\phi}_i^T\mathbf{P}_i\boldsymbol{\phi}_i\widehat{g}_i^2(\widehat{\epsilon}_i) \\ &= (1+o(1))\Big\{E\Big[\frac{f'(\epsilon_1)}{f(\epsilon_1)}\Big]^2\Big\}^{-1}(p+k+h)\log n \quad \text{a.s.}. \end{aligned} \tag{4.23}$$

It follows from (4.23) that $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|^2 = O(\log n/n)$ a.s., so that the monitored modification $\widehat{\boldsymbol{\beta}}_n = \widetilde{\boldsymbol{\beta}}_n$ is in fact used only finitely many times with probability 1.

To derive the asymptotic normality of the recursive adaptive estimator, we write

$$
\begin{aligned}
\mathbf{P}_n^{-1}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}) &= \mathbf{P}_{n-1}^{-1}(\widehat{\boldsymbol{\beta}}_{n-1}-\boldsymbol{\beta})+\boldsymbol{\phi}_n[\boldsymbol{\phi}_n^T(\widehat{\boldsymbol{\beta}}_{n-1}-\boldsymbol{\beta})+\widehat{g}_n(\widehat{\epsilon}_n)]\\
&=\sum_{i=1}^n \boldsymbol{\phi}_i[\boldsymbol{\phi}_i^T(\widehat{\boldsymbol{\beta}}_{i-1}-\boldsymbol{\beta})+\widehat{g}_i(\widehat{\epsilon}_i)-\widehat{g}_i(\epsilon_i)]+\sum_{i=1}^n \boldsymbol{\phi}_i\widehat{g}_i(\epsilon_i)+O(1) \quad \text{a.s..}
\end{aligned}
$$

The first term on the right-hand side of the last equality can be shown to be of order $o(\sqrt{n})$. The second term, when properly normalized, has a limiting normal distribution by the Martingale Central Limit Theorem. Therefore, by (4.3),

$$
\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}) &= (n\mathbf{P}_n^{-1})\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\phi}_i\widehat{g}_i(\epsilon_i)+o(1)\\
&\Rightarrow N\left(0,\ \left\{\boldsymbol{A}E\Big[\frac{f'(\epsilon_1)}{f(\epsilon_1)}\Big]^2\right\}^{-1}\right). \qquad (4.24)
\end{aligned}
$$

4.3. Recursive GMM estimators

The auxiliary consistent estimates in Sections 4.1 and 4.2 that are used to monitor the recursive M-estimators or adaptive estimators can be constructed by recursive modifications of Hansen's (1982) generalized method of moments (GMM), which takes the form of spectral factorization and Yule-Walker-type equations when applied to ARMAX models. Let $c_0=1$ and $c_i'=\sigma c_i$ for $i=0,\ldots,h$. Let $v_n=C(v^{-1})\epsilon_n$. Then $\{v_n\}_{n>h}$ is a covariance-stationary sequence whose autocovariances $\gamma_j=E(v_nv_{n+j})$ are given by

$$
\gamma_j=\sum_{i=0}^{h-j} c_i'c_{i+j}' \ (j=0,\ldots,h), \qquad (4.25)
$$

and $\gamma_j=0$ if $j>h$. Letting $\Gamma(z)=\Sigma_{j=-h}^h\gamma_jz^j$, note that (4.25) is equivalent to the spectral factorization $\Gamma(z)=\sigma^2C(z)C(z^{-1})$. Given $\gamma_0,\ldots,\gamma_h$, one can solve the system of nonlinear equations (4.25) by standard Newton-Raphson iterations, which Tunnicliffe-Wilson (1979) showed to be quadratically convergent by making use of spectral factorization and the theory of harmonic functions. In practice, the γ_j in (4.25) are unknown and have to be replaced by estimates based on observed data. Suppose that at time n one has an estimate $\widehat{\gamma}_{n,j}$ of γ_j for which

$$
\sum_{n=1}^{\infty}|\widehat{\gamma}_{n+1,j}-\widehat{\gamma}_{n,j}|<\infty \text{ a.s. for } j=0,\ldots,h, \qquad (4.26a)
$$

that there exists δ_n, depending on the input-output data $y_1, u_1, \ldots, y_n, u_n$ up to stage n, such that $\delta_n \to 0$ a.s., and that

$$\max_{0\le j\le h} |\widehat{\gamma}_{n,j} - \gamma_j| = o(\delta_n) \text{ a.s.}, \ \delta_{n-1} + \delta_{n+1} + n^{-a} = O(\delta_n) \text{ a.s.} \tag{4.26b}$$

for some $a > 0$. Note that (4.26b) implies that the $\widehat{\gamma}_{n,j}$ are strongly consistent. Making use of (4.26) and a modification of Tunnicliffe-Wilson's argument, Lai and Ying (1992) established the strong consistency of the following recursive estimate $\widehat{\mathbf{c}}_n = (\widehat{c}_{n,0}, \ldots, \widehat{c}_{n,h})^T$ of $\mathbf{c}' = (c'_0, \ldots, c'_h)^T$. Let $\mathbf{\Delta}_{n-1}$ be a $(h+1)\times(h+1)$ matrix whose (r,s)th element is $\widehat{c}_{n-1,r+s} + \widehat{c}_{n-1,r-s}$, setting $\widehat{c}_{n-1,j} = 0$ if $j > h$ or $j < 0$. Let $\widehat{c}_{0,0} > 0 = \widehat{c}_{0,1} = \cdots = \widehat{c}_{0,h}$ and

$$\widehat{\mathbf{c}}_n = \widehat{\mathbf{c}}_{n-1} - \mathbf{\Delta}_{n-1}^{-1}\Big(\sum_{i=0}^{h-j} \widehat{c}_{n-1,i}\widehat{c}_{n-1,i+j} - \widehat{\gamma}_{n,j}\Big)^T_{0\le j\le h}, \ j \ge 1. \tag{4.27}$$

Perform a stability test for the polynomial $\Sigma_{j=0}^h \widehat{c}_{n,j} z^j$, redefining (4.27) by $\widehat{\mathbf{c}}_n = \widehat{\mathbf{c}}_{n-1}$ if and only if the stability test fails. Note that (4.27) is essentially a one-step Newton-Raphson approximation, initialized at $\widehat{\mathbf{c}}_{n-1}$, to the solution $(c'_{n,0}, \ldots, c'_{n,h})^T$ of the system of equations $\widehat{\gamma}_{n,j} = \Sigma_{i=0}^{h-j} c'_{n,i} c'_{n,i+j} (j = 0, \ldots, h)$. From (4.25) and (4.26), it follows that $\widehat{\mathbf{c}}_n = \mathbf{c}' + o(\delta_n)$ a.s., and therefore

$$\frac{\widehat{c}_{n,j}}{\widehat{c}_{n,0}} = c_j + o(\delta_n) \text{ a.s. for } j = 1, \ldots, h. \tag{4.28}$$

To construct recursive GMM estimates $\widehat{\gamma}_{n,j}$ of γ_j so that (4.26) holds, we assume the identifiability condition

$$z^p A(z^{-1}), z^{k-1} B(z^{-1}) \text{ and } z^h C(z^{-1}) \text{ are relatively prime polynomials.} \tag{4.29}$$

Let $\boldsymbol{\lambda} = (-a_1, \ldots, -a_p, b_1, \ldots, b_k)^T$ be the vector of ARX parameters and let $\mathbf{v}_n = (y_{n-1}, \ldots, y_{n-p}, u_{n-d}, \ldots, u_{n-d-k+1})^T$. Note that (2.14) can be written in the form

$$y_n = \boldsymbol{\lambda}^T \mathbf{v}_n + \epsilon_n + c_1\epsilon_{n-1} + \cdots + c_h\epsilon_{n-h}. \tag{4.30}$$

Letting $\mathcal{G}_n$ be the σ-field generated by $\{\epsilon_i : i \le n\}$, we introduce instrumental variables $\mathbf{z}_n$, which are $\mathcal{G}_{n-h-1}$-measurable $\nu \times 1$ random vectors with $\nu \ge p+k$, and estimate $\boldsymbol{\lambda}$ by using the sample covariances between $\mathbf{v}_n$ and $\mathbf{z}_n$ and between y_n and $\mathbf{z}_n$, i.e.,

$$\widehat{\boldsymbol{\lambda}}_n = (\mathbf{V}_n^T\mathbf{V}_n)^{-1}\mathbf{V}_n^T\mathbf{Z}_n, \text{ where } \mathbf{V}_n = \sum_{i=1}^n \mathbf{z}_i\mathbf{v}_i^T, \mathbf{Z}_n = \sum_{i=1}^n \mathbf{z}_i y_i. \tag{4.31}$$

From (4.30) it follows that $\widehat{\boldsymbol{\lambda}}_n = \boldsymbol{\lambda} + (\mathbf{V}_n^T\mathbf{V}_n)^{-1}\mathbf{V}_n^T\Sigma_{i=1}^n \mathbf{z}_i(\epsilon_i + c_1\epsilon_{i-1} + \cdots + c_h\epsilon_{i-h})$. We next describe two typical classes of input-output data for which $\mathbf{z}_n$ can be defined so that the corresponding $\boldsymbol{\lambda}_n$ satisfy

$$\widehat{\boldsymbol{\lambda}}_n = \boldsymbol{\lambda} + o(\delta_n) \text{ a.s.}, \tag{4.32}$$

and then show how such $\widehat{\boldsymbol{\lambda}}_n$ can be used to define estimates $\widehat{\gamma}_{n,j}$ of γ_j so that (4.26) is satisfied.

(A) Suppose that $\{u_n\}$ is a stationary sequence independent of $\{\epsilon_n\}$ such that $(Eu_iu_j)_{1\le i,j\le m}$ is positive definite for every $m \ge 1$ and $E|u_1|^\alpha < \infty$ for some $\alpha > 4$. Define $\mathbf{z}_n = (y_{n-h-1}, \ldots, y_{n-h-p}, u_{n-d+p}, \ldots, u_{n-d-p-k+1})^T$. Then $n^{-1}\Sigma_{i=1}^n \mathbf{z}_i\mathbf{v}_i^T$ converges a.s. to a nonrandom $(3p+k)\times(p+k)$ matrix $\boldsymbol{H}$ that has full rank $p+k$. Moreover, since $(y_{n-h-1}, \ldots, y_{n-h-p})^T$ is $\mathcal{F}_{n-h-1}$-measurable and $\{u_n\}$ is independent of $\{\epsilon_n\}$, it follows from martingale limit theorems that

$$\sum_{i=1}^n \mathbf{z}_i(\epsilon_i + c_1\epsilon_{i-1} + \cdots + c_h\epsilon_{i-h}) = O((n\log\log n)^{1/2}) \text{ a.s.}$$

Hence, defining $\widehat{\boldsymbol{\lambda}}_n$ as in (4.31), we have $\widehat{\boldsymbol{\lambda}}_n = \boldsymbol{\lambda} + o(n^{-1/2}(\log n)^\rho)$ a.s. for every $\rho > 0$, and $\Sigma_{n=1}^\infty \|\widehat{\boldsymbol{\lambda}}_n - \widehat{\boldsymbol{\lambda}}_{n-1}\| < \infty$ a.s. Letting $e_{n,t} = y_t - \widehat{\boldsymbol{\lambda}}_n^T\mathbf{x}_t$, define $\widehat{\gamma}_{n,j} = (n-j)^{-1}\Sigma_{t=1}^{n-j} e_{n,t}e_{n,t+j}$ for $j = 0, \ldots, h$, which can also be expressed in the form

$$(n-j)\widehat{\gamma}_{n,j} = \sum_{t=1}^{n-j} y_t y_{t+j} - \widehat{\boldsymbol{\lambda}}_n^T\Big(\sum_{t=1}^{n-j} y_{t+j}\mathbf{v}_t\Big) - \widehat{\boldsymbol{\lambda}}_n^T\Big(\sum_{t=1}^{n-j} y_t\mathbf{v}_{t+j}\Big) + \widehat{\boldsymbol{\lambda}}_n^T\Big(\sum_{t=1}^{n-j}\mathbf{v}_t\mathbf{v}_{t+j}^T\Big)\widehat{\boldsymbol{\lambda}}_n, \tag{4.33}$$

thereby providing a convenient formula for updating the estimates $\widehat{\gamma}_{n,j}$ without having to calculate the residuals $e_{n,t}$ for all $t \le n$ at every stage n. The $\widehat{\gamma}_{n,j}$ thus defined satisfy (4.26a, b) with $\delta_n = n^{-1/2}(\log n)^\rho$, for any choice of $\rho > 0$.

(B) Suppose $\sup_n|\epsilon_n| < \infty$ a.s. We now weaken the assumption of the preceding paragraph that the entire sequence $\{u_n\}$ be stationary and independent of $\{\epsilon_n\}$, in order to extend the above ideas to adaptive control applications. Suppose that $\Sigma_{i=1}^n u_i^2 = O(n)$ a.s., and that there exist integer-valued random variables $(1 \le)n_1 < m_1 \le n_2 < m_2 \le \cdots$ satisfying the following assumptions:

(a) n_i is a stopping time (i.e., $\{n_i = t\} \in \mathcal{F}_t$ for all t) and $m_i - n_i$ is $\mathcal{F}_{n_i}$-measurable;

(b) $m_{i+1} - n_{i+1} = O(\#_i)$ and $\liminf(m_{i+1} - n_{i+1})/\{\#_i^{1/2}(\log\#_i)^c\} > 0$ a.s., for some $c > 3/2$, where $\#_n = \sum_{r=1}^n (m_r - n_r)$;

(c) $\max\left\{\sum_{\nu=0}^{p-1} y_{n_i-\nu}^2, \sum_{\nu=1}^{k+d-2} u_{n_i-\nu}^2\right\} = o(m_i - n_i)$ a.s.;

(d) $u_n = w_n$ for $n_i \le n < m_i$, where w_n are independent with $\sup_n |w_n| < \infty$ a.s. and are independent of $\{\epsilon_n\}$.

Letting $J = \cup_{i=1}^{\infty}\{n : n_i + p + k + h + d \le n \le m_i\}$, define the $(3p+k)\times 1$ vectors

$$\mathbf{z}_n = \begin{cases} (y_{n-h-1}, \ldots, y_{n-h-p}, u_{n-d+p}, \ldots, u_{n-d-p-k+1})^T, & \text{if } n \in J, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{4.34}$$

This means that the $\widehat{\boldsymbol{\lambda}}_n$ defined by (4.31) need only be updated at $n \in J$ since $\mathbf{V}_n = \mathbf{V}_{m_i}$ and $\mathbf{Z}_n = \mathbf{Z}_{m_i}$ for $m_i < n < n_{i+1} + p + k + h + d$. In fact, we need update the $\widehat{\boldsymbol{\lambda}}_n$ only at $n = m_i$ as we define $\widehat{\gamma}_{n,j}$ by (4.33) for $n \in \{m_1, m_2, \ldots\}$, and set

$$\widehat{\gamma}_{n,j} = \widehat{\gamma}_{m_i,j} \text{ for } m_i < n < m_{i+1}. \tag{4.35}$$

Since $A(z)$ is stable and since $\Sigma_{i=1}^n \epsilon_i^2 + \Sigma_{i=1}^n u_i^2 = O(n)$ a.s., it follows that $\Sigma_{i=1}^n y_i^2 = O(n)$ a.s. Assumption (d) is tantamount to using exogenous probing inputs w_n (that are independent of the ϵ_n) only *occasionally* at stages $n \in J$, so that at other times feedback control laws can be used to generate inputs. Assumption (b) says that these probing inputs are introduced in blocks of at least certain prescribed lengths that eventually tend to ∞, while assumption (c) requires that each of these blocks begins only when the input-output data for several past observations do not exceed a certain magnitude, which can be done since $\Sigma_{i=1}^n u_i^2 + \Sigma_{i=1}^n y_i^2 = O(n)$ a.s. Since $\{w_t\}$ and $\{\epsilon_t\}$ are independent, we can modify the proof of Theorem 4 and Corollary 2 of Lai and Ying (1991a) to conclude that

$$\frac{\lim_{r\to\infty}\left(\sum_{i=1}^{r}\sum_{t=n_i+p+k+h+d}^{m_i} \mathbf{z}_i\mathbf{v}_i^T\right)}{\#_r} = \boldsymbol{H}, \widehat{\gamma}_{m_i,j} = \gamma_j + O(\#_i^{-1/2}(\log \#_i)^{\rho}) \text{ a.s.} \tag{4.36}$$

for any $\rho > \frac{1}{2}$ and $j = 0, \ldots, h$, where H is a nonrandom matrix of full rank $p + k$. Making use of (4.36) together with assumption (b), it can be shown that (4.26a, b) again holds for the $\widehat{\gamma}_{n,j}$ defined by (4.35).

Without assuming $A(z)$ to be stable, Lai and Ying (1991a) modified the preceding approach that uses directly occasional blocks of white-noise probing inputs w_t by running SG in parallel and introducing w_t via (3.15). Their recursive

GMM estimators used to monitor RML2 are still defined by (4.27) and (4.35) from these occasional blocks of well excited input-output data.

5. Conclusion

In his article in this issue, Chan says, "By reviewing (Wei's) publications, one often comes up with more than the stated results. In fact, there are numerous possible areas of extension due to Wei's work." He then mentions several examples in time series analysis and concludes that Wei's work in asymptotic inference "has greatly broadened the scope of time series and stochastic regression." In this article we have focused on another area of Wei's research that reaffirms Chan's conclusions. In particular, in Section 4 we have given one such area of extension by using the theory of recursive estimation and extended stochastic Liapounov functions to study efficient semiparametric estimation in ARMAX models.

Acknowledgement

This research was supported by the National Science Foundation under DMS-0305749 at Stanford University and DMS-0203798 and DMS-0504871 at Columbia University.

References

Anderson, T. W. and Taylor, J. (1976). Some experimental results on the statistical problem of least squares estimates in control problems. *Econometrica* **44**, 1289-1302.

Aokoi, M. (1974). On some price adjustment schemes. *Ann. Econ. Soc. Measurements* **3**, 95-116.

Åström, K. and Eykhof, P. (1971). System identification – a survey. *Automatica* **7**, 123-167.

Åström, K. and Wittenmark, B. (1973). On self-tuning regulators. *Automatica* **9**, 195-199.

Banach, S. (1930). Über einige Eigenschaften der lacunaren trigonometrishe Reihen. *Studia Math.* **2**, 207-220.

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647-671.

Becker, A., Kumar, P. R. and Wei, C. Z. (1985). Adaptive control with stochastic approximation algorithm: geometry and convergence. *IEEE Trans. Automat. Control* **30**, 330-338.

Caines, P. E. and Lafortune, S. (1984). Adaptive control with recursive identification for stochastic linear systems. *IEEE Trans. Automat. Control* **AC-29**, 312-321.

Chan, N. H. (2006). Inference in time series and stochastic processes. *Statist. Sinica* **16**, 683-696.

Chen, G. J., Lai, T. L. and Wei, C. Z. (1981). Convergence systems and strong consistency of least squares estimates in regression model. *J. Multivariate Anal.* **11**, 319-333.

Fabian, V. and Hannan, J. (1982). On estimation and adaptive estimation for locally asymptotically normal families. *Z. Wahrsch. Verw. Gebiete* **59**, 459-478.

Feldbaum, A. A. (1961). The theory of dual control I-II. *Autom. Remote Control* **21**, 874-883, 1022-1039.

Gaposhkin, V. F. (1966). Lacunary series and independent functions. *Russian Math. Surveys* 216, 1-82.

Gladyshev, E. G. (1965). On stochastic approximation. *Theory Probab. Appl.* **10**, 275-278.

Goodwin, G. C., Ramage, P. J. and Cainse, P. E. (1981). Discrete time stochastic adaptive control. *SIAM J. Control Optim.* **19**, 829-853.

Guo, L. (1995). Convergence and logarithmic laws of self-tuning regulators. *Automatica* **31**, 435-450.

Guo, L. and Chen, H. F. (1988). Convergence rate of ELS based adaptive tracker. *Syst. Sci & Math. Sci.* **1**, 131-138.

Guo, L. and Chen, H. F. (1991). The Åström-Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers. *IEEE Trans. Automat. Control* **36**, 802-812.

Han, J., Lai, T. L. and Spivakovsky, V. (2005). Approximate policy optimization and adaptive control in regression models. To appear in *Computational Economics*.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029-1054.

Kreiss, J-P. (1987). On adaptive estimation in stationary ARMA processes. *Ann. Statist.* **15**, 112-133.

Lai, T. L. (1986). Asymptotically efficient adaptive control in stochastic regression models. *Adv. Appl. Math.* **7**, 23-45.

Lai, T. L. (2003). Stochastic approximation. *Ann. Statist.* **31**, 391-406.

Lai, T.L. and Robbins, H. (1977). Strong consistency of least-squares estimates in regression models. *Proc. Natl. Acad. Sci. USA* **74**, 2667-2669.

Lai, T. L. and Robbins, H. (1979). Adaptive design and stochastic approximation. *Ann. Statis.* **7**, 1196-1221.

Lai, T. L. and Robbins, H. (1981). Consistency and asymptotic efficiency of slope in stochastic approximation schemes. *Z. Wahrsch. Verw. Gebiete* **56**, 329-360.

Lai, T.L. and Robbins, H. (1982). Iterated least squares in multiperiod control. *Adv. Appl. Math.* **3**, 50-73.

Lai, T. L., Robbins, H. and Wei, C. Z. (1978). Strong consistency of least squares estimates in multiple regression. *Proc. Natl. Acad. Sci. USA* **75**, 3034-3036.

Lai, T. L., Robbins, H. and Wei, C. Z. (1979). Strong consistency of least squares estimates in multiple regression II. *J. Multivariate Anal.* **9**, 343-361.

Lai, T. L. and Wei, C. Z. (1982a). A law of the iterated logarithm for double arrays of independent random variables with applications to regression and time series models. *Ann. Probab.* **10**, 320-335.

Lai, T. L. and Wei, C. Z. (1982b). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10**, 154-166.

Lai, T. L. and Wei, C. Z. (1982c). Asymptotic properties of projections with applications to stochastic regression problems. *J. Multivariate Anal.* **12**, 346-370.

Lai, T. L. and Wei, C. Z. (1983a). Lacunary systems and generalized linear processes. *Stoch. Proc. Appl.* **14**, 187-199.

Lai, T. L. and Wei, C. Z. (1983b). Asymptotic properties of general autoregressive models and strong consistency of least squares estimates of their parameters. *J. Multivariate Anal.* **13**, 1-23.

Lai, T. L. and Wei, C. Z. (1984). Moment inequalities with applications to regression and time series models. In *Inequalities in Statistics and Probability* (Edited by Y. L. Tong), 165-172. Monograph Series, Institute of Mathematical Statistics.

Lai, T. L. and Wei, C. Z. (1985). Orthonormal Banach systems with applications to linear processes. *Z. Wahrsch. Verw. Gebiete* **70**, 381-393.

Lai, T. L. and Wei, C. Z. (1986a). Extended least squares and their applications to adaptive control and prediction in linear systems. *IEEE Trans. Automat. Control* **31**, 898-906.

Lai, T. L. and Wei, C. Z. (1986b). On the concept of excitation in least squares identication and adaptive control. *Stochastics* **16**, 227-254.

Lai, T. L. and Wei, C. Z. (1987). Asymptotically efficient self-tuning regulators. *SIAM J. Contr. Optim.* **25.**

Lai, T. L. and Ying, Z. (1991a). Parallel recursive algorithms in asymptotically efficient adaptive control of linear stochastic systems. *SIAM J. Control Optim.* **29**, 1091-1127.

Lai, T. L. and Ying, Z. (1991b). Recursive identification and adaptive prediction in linear stochastic systems. *SIAM J. Control Optim.* **29**, 1061-1190.

Lai, T. L. and Ying, Z. (1992). Recursive solutions of estimating equations and adaptive spectral factorization. *IEEE Trans. Automat. Control* **37**, 240-243.

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Automat. Control* **22**, 551-575.

Ljung, L. and Söderström, T. (1983). *Theory and Practice of Recursive Estimation*. MIT Press.

Martin, R. D. and Yohai, V. J. (1985). Robustness in time series and estimating ARMA models. In *Handbook of Statistics* **5** (Edited by E. J. Hannan, P. R. Krishnaiah and M. M. Rao), 119-155. North Holland, Amsterdam.

Móricz, F. (1976). Moment inequalities and strong laws of large numbers. *Z. Wahrsch. Verw. Gebiete* **35**, 299-314.

Nevel'son, M. B. and Has'minskii, R. Z. (1973). *Stochastic Approximation and Recursive Estimation*. Amer. Math. Soc., Providence, RI.

Prescott, E. C. (1972). The multiperiod control problem under uncertainty. *Econometrica* **40**, 1043-1058.

Robbins, H. and Siegmund, D. (1970). A convergence theorem for nonnegative lamost supermartingales and some applications. In *Optimizing Methods in Statistics* (Edited by J. Rustagi), 233-257. Academic Press, New York.

Solo, V. (1979). The convergence of AML. *IEEE Trans. Automat. Control* **24**, 958-962.

Stein, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1**, 187-196. Univ. California Press.

Tunnicliffe-Wilson, G. (1979). Factorization of the covariance generating function of a pure moving average process. *SIAM J. Numer. Anal.* **6**, 1-7.

Wei, C. Z. (1980). Limit theorems for weighted sums with applications to regression and time series models. Ph.D. dissertation, Columbia University.

Wei, C. Z. (1985). Asymptotic properties of least squares estimates in stochastic regression models. *Ann. Statist.* **13**, 1498-1508.

Wei, C. Z. (1987a). Multivariate adaptive stochastic approximation. *Ann. Statist.* **15**, 1115-1130.

Wei, C. Z. (1987b). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *Ann. Statist.* **15**, 1667-1682.

Wieland, V. (2000). Learning by doing and the value of experimentation. *J. Econom. Dynam. Control* **24**, 501-534.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.

Department of Statistics, Stanford University, Sequoia Hall, Stanford, CA 94305, U.S.A.

E-mail: lait@stat.stanford.edu

Department of Statistics, Columbia University, New York, NY 10027, U.S.A.

E-mail: zying@stat.columbia.edu

(Received September 2005; accepted January 2006)