Holger Rootzén Jan Sternby

Consistency in Least Squares Estimation: A Bayesian Approach

Preprint
January
1984

1

Institute of Mathematical Statistics
University of Copenhagen

CONSISTENCY IN LEAST SQUARES ESTIMATION: A BAYESIAN APPROACH

Holger Rootzén
Institute of Mathematical Statistics
University of Copenhagen
5, Universitetsparken
DK-2100 Copenhagen Ø
Denmark

Jan Sternby
Kockumation AB
Box 1044
S-212 10 Malmö
Sweden

Abstract

In a previous paper, Sternby (1977), the convenience of using martingale theory in the analysis of Bayesian Least Squares estimation was demonstrated. However, certain restrictions had to be imposed on either the feedback structure or on the initial values for the estimation. In the present paper these restrictions are removed, and necessary and sufficient conditions for strong consistency (in a Bayesian sense) are given for the Gaussian white noise case without any assumptions on closed loop stability or on the feedback structure.

In the open loop case the poles are shown to be consistently estimated, a.e., and in the closed loop case certain choices of control law are shown to assure consistency. Finally adaptive control laws are treated, and implicit self-tuning regulators are shown to converge to the desired control laws.

1. Introduction

The simplicity of the Least Squares (LS) method for parameter identification in dynamical systems has made it widely used. As compared to ordinary regression analysis, this leads to additional problems in the study of the asymptotic properties of the estimates. Natural questions to ask are the following: What happens if there is feedback in the system? What happens if parameters of the feedback temporarily take values which in stationarity would lead to an unstable closed loop system? What happens if the noise is coloured? Some of these questions have been answered in recent years, but a complete solution is not yet found. E.g. one difficult problem concerns the case of (possibly adaptive) feedback without stability assumptions on the closed loop system.

In a previous paper, Sternby (1977), a Bayesian approach was shown to be convenient in proving consistency. Using martingale theory, consistency was established under a certain condition, which did not include any explicit stability assumptions. However, for a technical reason, restrictions had to be imposed on either the feedback structure or on the initial values for the estimates. Here we remove these restrictions and give a (necessary and) sufficient condition for consistency in linear discrete time systems with Gaussian equation noise. The notation and general theory needed for this are given in Sections 2 and 3, while the result itself is stated and proved in Section 4.

In Section 4, the consistency condition is moreover interpreted in terms of conditions on the feedback, and some different ways to ensure consistency are indicated. Finally, in Section 5 we discuss adaptive control laws, and in particular take self-tuning regulators as examples of feedbacks that allow the "true" control law to be obtained asymptotically.

The single-input, single-output (SISO) and the multivariable cases can be treated simultaneously as was done in Sternby (1977), but in order to emphasize the basic ideas this paper is written out only for the SISO case.

2. The least squares identification method

Let the true system be described by

$$y(t) = \phi(t)^T x + e(t) \tag{1}$$

with $\phi(t)$ a n-dimensional column vector of variables known at time $t-1$, i.e. $\phi(t)$ is a function of previous values of inputs and outputs. E.g. for a linear kth order transfer function n would equal 2k and $\phi(t)$ would be $\phi(t)^T = [y(t-1),\ldots, y(t-k), u(t-1),\ldots, u(t-k)]$, where u are the system inputs. The sequence $\{e(t)\}$ is some unobservable one-dimensional disturbance and x is a *random* n- dimensional column vector of parameters which we want to estimate from observations of the y(t)'s.

Usually x is considered as a constant, but unknown, vector that is to be estimated. By taking x random we thus consider a whole class of systems with the structure (1), and picking a certain member of that class is the first part of the identification experiment. The second part of the experiment then generates the e(t)'s and thus the observables $\{y(t)\}$.

The underlying probability space then supports both the parameter x and the noise sequence $\{e(t)\}$, and will be thought of as a product of a space supporting x and a space supporting $\{e(t)\}$. Primarily we will be interested in a.e. convergence of estimates, $\hat{x}_t \to \hat{x}_\infty$ say, and "a.e." then means almost everywhere with respect to the joint distribution of x and $\{e(t)\}$. By standard arguments this convergence is equivalent to the existence of a set $A \subset R^n$ with $P(x \in A) = 1$, and such that <u>for each $x \in A$</u>, $\hat{x}_t \to \hat{x}_\infty$ a.e. with respect to the conditional distribution given the value of x. (Formally, if we write $h(x_0) = P(\hat{x}_t \not\to \hat{x}_\infty | x = x_0)$ for the conditional probability that the estimates do not converge to $\hat{x}_\infty$ if the true parameter value is x_0, and use that probabilities can be obtained as averages of conditional probabilities, we have that if $\hat{x}_t$ converges a.e. then

$$0 = P(\hat{x}_t \not\to \hat{x}_\infty) = E(h(x)).$$

Since $h(x) \geq 0$ for all x it follows that there is a set $A \in \mathbb{R}^n$ with $P(x \in A) = 1$

such that $h(x) = 0$ for $x \in A$ (since otherwise $E(h(x))$ would be strictly positive). Hence,

$$P(\hat{x}_t \to \hat{x}_\infty \quad | x = x_0) = 1 - h(x_0) = 1 \ , \text{ for } x_0 \in A \ ,$$

as claimed.) Thus, in the present context, convergence a.e. means "convergence for almost every realization of almost every system," and nothing can be directly deduced about a specific given system with fixed true parameters. For a further discussion of this point of view, see Sternby (1977).

Let F_t be the σ-algebra containing all the information available at time t so that in particular $F_t \subset F_{t+1}$ for all t, and let $F_\infty = \lim_{t\to\infty} F_t$ be the σ-algebra generated by all the F_t's. E.g. for the linear kth order transfer function considered above, F_t would be the σ-algebra generated by the outputs $y(1),\ldots, y(t)$ and the inputs $u(1),\ldots, u(t)$. Furthermore, we will write $\zeta \in F_t$ if ζ is F_t-measurable, i.e. if "ζ is a function of the variables generating F_t," and similarly for $\zeta \in F_\infty$. In the Bayesian approach to least squares estimation, the estimate at time t, $\hat{x}_t$, say, is defined to be the function of the observations which minimizes the conditional mean square error. Thus, more precisely, $\hat{x}_t$ is uniquely determined by the requirements that $\hat{x}_t \in F_t$ and that

$$E[(x - \zeta)(x - \zeta)^T | F_t] - E[(x - \hat{x}_t)(x - \hat{x}_t) | F_t]$$

is positive definite for any random vector $\zeta \in F_t$. It is well known (see e.g. Kalman (1960) or Jaswinski (1970), ch. 5.2) that $\hat{x}_t = E[x|F_t]$.

This should be compared to the ordinary LS estimate ξ_t of x which is obtained by minimizing V_t with respect to t, where $V_t = V_t(\xi)$ is defined by

$$V_t(\xi) = \frac{1}{t-t_0} \sum_{s=t_0+1}^{t} [y(s) - \phi(s)^T \xi]^2 \ . \qquad (1)$$

Here the lower limit of summation, $t_0 + 1$ has to be chosen such that $\phi(s)$, $s = t_0 + 1,\ldots,t$ have been actually observed at time t. In Section 4 we will see that in an important special case $\hat{\xi}_t = \hat{x}_t$, and use this to prove consistency for $\hat{\xi}_t$ from the corresponding result for $\hat{x}_t$.

Finally, for later use, we introduce the notation $\mathcal{N}(S)$ for the null space of a matrix S, and P_t for the conditional covariance matrix of x at time t, i.e., $P_t = \mathcal{E}[(x - \hat{x}_t)(x - \hat{x}_t)^T | \mathcal{F}_t]$.

3. General results

Convergence of Bayes estimates have been studied rather extensively in the statistical literature, see e.g. Schwartz (1965) and the references therein. In the present context, the basic result is that $\hat{x}_t$ and P_t converge a.e. Here we will use a variant of the formulation in Sternby (1977).

Theorem 1. Suppose that the distribution of the true parameter x has finite second moments. Then $\hat{x}_t$ and P_t converge a.e. as $t \to \infty$, and, denoting the limits by $\hat{x}_\infty$ and P_∞,

$$\hat{x}_\infty = E[x \mid F_\infty],$$

and

$$P_\infty = E[(x - \hat{x}_\infty)(x - \hat{x}_\infty)^T \mid F_\infty]. \qquad (2)$$

Proof. All the assertions except (2) are established in Sternby (1977) (Theorem 1 and the subsequent remark). In the cited reference it is also noted that

$$E[(x - \hat{x}_t)(x - \hat{x}_t)^T \mid F_t] = E[xx^T \mid F_t] - \hat{x}_t\hat{x}_t^T . \qquad (3)$$

Since $\hat{x}_t \to \hat{x}_\infty$ a.e., $\hat{x}_t\hat{x}_t^T \to \hat{x}_\infty\hat{x}_\infty^T$, and it follows from the martingale convergence theorem (see e.g. Chung 1968, p. 313) that $E[xx^T \mid F_t] \to E[xx^T \mid F_\infty]$, a.e. Since moreover the lefthand side of (3) converges to P_∞ it follows that a.e.

$$P_\infty = E[xx^T \mid F_\infty] - \hat{x}_\infty\hat{x}_\infty^T ,$$

and (2) then is obtained by applying (3), with t replaced by ∞, to the righthand side. #

In the next result , which is a minor generalization from Sternby (1977), the limit $\hat{x}_\infty$ is examined. The statement is slightly complicated by the need to handle cases where the probability for P_t to tend to zero is less than one, or where only some modes, but not all, may be consistently estimated.

Theorem 2. Let $a \in F_\infty$ be a n-dimensional random vector. If the assumptions of Theorem 1 are satisfied, then

$$a^T\hat{x}_t \to a^T x \quad \text{a.e. on } \{\omega | P_\infty a = 0\}.$$

In particular, if $P(P_\infty a = 0) = 1$ then $a^T\hat{x}_t \to a^T x$ a.e., and if furthermore a is nonrandom, then $a^T\hat{x}_t \to a^T x$ also in mean square.

Proof. Let $1(a)$ equal one if $P_\infty a = 0$ and zero if $P_\infty a \neq 0$. By assumption $a \in F_\infty$ and hence also $1(a) \in F_\infty$, since $1(a)$ is a function of $P_\infty \in F_\infty$ and a. Thus a and $1(a)$ can be treated as constants when taking expectations conditioned on F_∞, and hence, using (2) for the third equality, we get that

$$E[1(a)(a^T(x-\hat{x}_\infty))^2 | F_\infty] = 1(a)a^T E[(x-\hat{x}_\infty)(x-\hat{x}_\infty)^T | F_\infty]a = 1(a)a^T P_\infty a = 0,$$

since $1(a)$ is zero on the set where $a^T P_\infty a \neq 0$. Thus $1(a)(a^T(x-\hat{x}_\infty))^2$ has expectation zero and is nonnegative, and hence has to be zero a.e., or equivalently $a^T x = a^T x_\infty$ a.e. on $\{\omega | 1(a) = 1\} = \{\omega | P_\infty a = 0\}$. Since $a^T\hat{x}_t \to a^T\hat{x}_\infty$ a.e., this proves the first part of the theorem.

Further, to prove the last assertion, about mean square convergence, it is sufficient to prove that $(a^T\hat{x}_t)^2$ is uniformly integrable (uniform integrability is defined in Chung (1968), p. 89, and the result is stated on p. 90 therein). However, by (3)

$$0 \le (a^T\hat{x}_t)^2 \le E[(a^T x)^2 | F_t] \ ,$$

and since $E[(a^T x)^2 | F_t]$ is uniformly integrable by the martingale convergence theorem, this proves uniform integrability of $(a^T\hat{x}_t)^2$ (Chung (1968), p. 313 and p. 93). #

For systems picked randomly as our model prescribes, this theorem answers the consistency question as long as the conditional mean is used as estimate. But also in a practical situation, when we are faced with a particular system, this theorem should be encouraging. If there is no reason to suspect that the true system belongs to a null set (e.g. through overparameterization), then the estimate can be assumed to converge to the true value if the conditional variance tends to zero.

It is thus desirable to use the conditional mean as an estimate. But unfortunately this is in general difficult to calculate.

4. Main results--the Gaussian case

We now focus interest on the special--but important--Gaussian white noise case. Specifically, we will assume that

(a) $e(1), e(2), \ldots$ are independent Gaussian random variables with mean zero and variances $\sigma^2 > 0$, and

(b) $\{e(t)\}_1^\infty$ is independent of x and F_0, $\phi(t) \in F_{t-1}$, and F_t is the σ-algebra generated by F_0 and $y(1), \ldots, y(t)$, for $t \geq 1$.

Initially it will further be assumed that

(c) the conditional distribution of the parameter vector x given F_0 is n-dimensional normal, with (possibly random) mean m and covariance R.

The important case covered by (a)-(c) is the kth order transfer function briefly discussed above. In fact, suppose F_0 is the trivial σ-algebra (which contains no information), suppose $\phi(t)^T = [y(t-1), \ldots, y(t-k), u(t-1), \ldots, u(t-k)]$, where the input u(t) only depends on $y(1), \ldots, y(t)$ and suppose that $y(-k+1), \ldots, y(0)$ are nonrandom. Then, if (a) and (b) hold and the parameter vector x is independent of the noise variables $\{e(t)\}$ and if the (unconditional) distribution of x is normal with (constant) mean m and covariance R, it follows readily that also (c) holds. In this, the condition that $y(-k+1), \ldots, y(0)$ are nonrandom is rather awkward--this problem with initial values is the reason for the slightly involved formulation of (c). As will be seen this initialization problem can be circumvented in the study of consistency of the (ordinary) LS estimates $\hat{\xi}_t$, but it is necessary for obtaining an explicit expression for $\hat{x}_t$.

If (a)-(c) are satisfied, then $\hat{x}_t, P_t$ can be computed recursively from the equations

$$\hat{x}_t = \hat{x}_{t-1} + K(t)[y(t) - \phi(t)^T \hat{x}_{t-1}],$$
$$P_t = P_{t-1} - P_{t-1}\phi(t)[\sigma^2 + \phi(t)^T P_{t-1}\phi(t)]^{-1}\phi(t)^T P_{t-1} \tag{4}$$

with

$$K(t) = P_{t-1}\phi(t)[\sigma^2 + \phi(t)^T P_{t-1}\phi(t)]^{-1},$$

using the initial conditions $\hat{x}_0 = m$, $P_0 = R$. (Here and in the sequel we arbitrarily assume that the recursion starts at $t = 0$. This is only for convenience of notation, and we might of course as well replace 0 by any other time point t_0.) The recursion (4) was obtained by Åström and Wittenmark (1970), with a further proof given in Åström (1978)--the conditions (a)-(c) are stated in a slightly different way in their derivations. (In fact the proof is quick: Using (a)-(c) and the standard formula for conditioning in multivariate Gaussian distributions it follows that the conditional distribution of x given F_1 is Gaussian with mean $\hat{x}_1$ and covariance matrix P_1, and the general result then follows by repeating this argument for $t=2,3,\ldots$)

A basic fact about the (ordinary) LS-estimate $\hat{\xi}_t$, which we will use to carry over the consistency results for $\hat{x}_t$ to $\hat{\xi}_t$, is that $\hat{\xi}_t$ also can be computed recursively, by very similar equations as for $\hat{x}_t$, viz.

$$\begin{aligned} \hat{\xi}_t &= \hat{\xi}_{t-1} + K(t)[y(t) - \phi(t)^T\hat{\xi}_{t-1}], \\ S_t &= S_{t-1} - S_{t-1}\phi(t)[1 + \phi(t)^T S_{t-1}\phi(t)]^{-1}\phi(t) S_{t-1}, \end{aligned} \tag{5}$$

with

$$K(t) = S_{t-1}\phi(t)[1 + \phi(t)^T S_{t-1}\phi(t)]^{-1},$$

and some suitable initial values, $\hat{\xi}_0, S_0$, with S_0 a nonnegative definite matrix. Here, to obtain the exact LS estimate which minimizes (1), one should start the recursion at some t-value, say t_1 such that (2) has a unique minimum for $\xi = \hat{\xi}_{t_1}$, and use $\hat{\xi}_{t_1}$ and $S_{t_1} = [\sum_1^{t_1}\phi(s)\phi(s)^T]^{-1}$ as initial values. However, in practice one would often use some conventional (incorrect) starting values, and this is the situation we analyze here, and as before we will use $t_1 = 0$. In passing it may also be remarked that it is obvious that S_t always converges, since it is nonincreasing and bounded from below.

In the main result we make use of the concept of absolute continuity. A probability distribution P is said to be absolutely continuous with respect to another probability distribution Q if events which have zero Q-probability also have zero P-probability, i.e. if $Q(A) = 0$ implies that $P(A) = 0$. It can be shown that equivalently P is absolutely continuous with respect to Q if it has a density with respect to Q, i.e. if P-expectations can be computed by first multiplying with a density function and then taking expectation with respect to Q. Absolute continuity of general measures is defined in the same way. E.g. if a Gaussian distribution in $\mathbb{R}^n$ has a strictly positive definite covariance matrix, then it has a strictly positive density function and hence is absolutely continuous with respect to Lebesgue measure in $\mathbb{R}^n$ (i.e. volume measure in $\mathbb{R}^n$), and vice versa. An important property of absolute continuity is that it preserves a.e. convergence, i.e. if a sequence converges a.e., $\hat{x}_t \to \hat{x}_\infty$, say, with respect to Q and P is absolutely continuous with respect to Q, then $\hat{x}_t \to \hat{x}_\infty$ a.e. with respect to P. (The proof of this assertion is immediate: That $\hat{x}_t \to \hat{x}_\infty$ a.e. under Q is the same as Q ($\hat{x}_t$ does not converge to $\hat{x}_\infty) = 0$, which by absolute continuity implies that $P(\hat{x}_t$ does not converge to $\hat{x}_\infty) = 0$, so that $\hat{x}_t \to \hat{x}_\infty$ a.e. under P.)

Theorem 3. Let $\{y(t)\}$ be the output of a system (1) which satisfies (a) and (b) above, and let $\hat{\xi}_t$ be the LS-estimate generated by (5), with initial values $\hat{\xi}_0, S_0$.

(i) If the true conditional distribution of the parameter vector x given F_0 a.e. is absolutely continuous with respect to the Gaussian distribution with mean $\hat{\xi}_0$ and covariance matrix S_0, then $\hat{\xi}_t$ converges a.e., and if the random vector $a \in F_\infty$, then

$$a^T\hat{\xi}_t \to a^T x, \text{ a.e. on } \{\omega | S_\infty a = 0\}.$$

(ii) If S_0 is strictly positive definite and $S_t \to 0$ a.e., then there is a set $A \in \mathbb{R}^n$ such that the complement of A has measure (or "volume") zero and such that

$$\hat{\xi}_t \to x,$$

a.e. for each fixed $x \in A$ (i.e. with respect to the conditional distribution given x).

Proof. Suppose first that in addition (c) holds, with $m = \hat{\xi}_0$ and $R = \sigma^2 S_0$, so that $\hat{x}_t$ and P_t satisfy (4) with these initial conditions. It then follows at once from the forms of (4) and (5) that $\hat{x}_t$ and P_t/σ^2 satisfy (5) (with $\hat{\xi}_t$ replaced by $\hat{x}_t$ and S_t by P_t/σ^2). Thus $\hat{\xi}_t = \hat{x}_t$ and $S_t = P_t/\sigma^2$ also for $t=1,2,\ldots$, and it follows from Theorems 1 and 2 that $\hat{\xi}_t \to \hat{x}_\infty$ and $S_t \to P_\infty/\sigma^2$ a.e. and that

$$a^T \xi_t \to a^T x \quad \text{a.e. on } \{\omega \mid P_\infty a = 0\},$$

where $\{\omega \mid P_\infty a = 0\} = \{\omega \mid \sigma^2 S_\infty a = 0\} = \{\omega \mid S_\infty a = 0\}$, a.e.

Now, if the true distribution of x given F_0 a.e. is absolutely continuous with respect to the Gaussian distribution with mean $\hat{\xi}_0$ and covariance matrix S_0, it follows straightforwardly that the entire true distribution is absolutely continuous with respect to the distribution where the conditional distribution of x was assumed to be Gaussian with mean $\hat{\xi}_0$ and covariance matrix $\sigma^2 S_0$. That $\hat{\xi}_t$ converges a.e. under the true distribution then follows at once, since convergence a.e. is preserved under absolute continuity, as discussed immediately before the theorem. Furthermore, above it is shown that $\{\omega \mid S_\infty a = 0$ and $a^T\hat{\xi}_t \not\to a^T x\}$ has probability 0 if (c) holds with $m = \hat{\xi}_0$, $R = \sigma^2 S_0$. By absolute continuity it then also has probability zero in the correct distribution, which completes the proof of part (i).

Finally, if the hypothesis of (ii) is satisfied, it also follows, using the argument in the beginning of Section 2, that there is a set $A \in \mathbb{R}^n$ with probability one in the Gaussian distribution with mean $\hat{\xi}_0$ and covariance matrix $\sigma^2 S_0$ such that $\hat{\xi}_t \to x$ a.e. with respect to the conditional distribution given x, for all $x \in A$. Since S_0 is assumed to be strictly positive definite this implies that the complement of A, which has probability zero in the Gaussian distribution, also has Lebesgue measure (or volume) zero, which concludes the proof of part (ii). #

Remark 1: Note that no assumptions are made in Theorem 3 on stability or feedback. Consistency is only coupled to the condition $S_\infty = 0$, with $S_\infty = 0$ a.e. $\Rightarrow \hat{\xi}_t \to x$ a.e. Theorem 3 thus replaces and supersedes Theorem 3 and corollaries 1 and 2 of Sternby (1977). The condition $S_\infty = 0$ will be further examined below. Note also that the awkward condition (c) (which for the kth order transfer function translates to $y(-k+1),\ldots, y(0)$ being non-random) is not needed for the result, but only the very weak requirement of absolute continuity of the conditional distributions of x given F_0. The condition (b) may be slightly weakened, in that the result holds also if F_t contains some extra information in addition to knowledge of $y(1),\ldots, y(t)$ and F_0, as long as this extra information is independent of x and $\{e(t)\}$.

Remark 2: Using the same technique with absolute continuity of measures, the counterpart to Theorem 5 of Sternby (1977) can also be shown so that, for $\hat{\xi}_\infty = \lim_t \hat{\xi}_t$, the sets where $S_\infty = 0$ and $\hat{\xi}_\infty = x$ can differ only by a null set.

Remark 3: It would be desirable to be able to drop the Gaussian assumption on the noise sequence as well. But this is more involved, since it requires the absolute continuity of a product measure of infinitely many components.

Remark 4: In Sternby (1977) also mean square convergence was shown. This part cannot be extended in the same way, since the distribution of the estimates is involved, and not only null sets. In fact, the second moments need not even exist.

Remark 5: In general nothing can be directly deduced from the theorem about a specific system, since the notion a.e. always excludes parameter sets with zero probability. However, if $\phi(t)$ does not depend on the parameters, then consistency a.e. in the product measure implies consistency a.e. for every parameter vector x. We thus have a non-Bayesian consistency result, e.g. for ordinary regression analysis, where $\phi(t)$ does not contain old inputs or outputs of a dynamical system. On the other hand, for this case stronger results can be shown by other methods, see e.g. Solo (1981).

The most straightforward application of Theorem 3 is to cases where the whole S_t-matrix tends to zero. Then all components of the estimate will be consistent. The next theorem gives a condition for this to happen.

Theorem 4: With notation as in Theorem 3 and $S_o > 0$

$$\{\omega | S_t \to 0\} = \{\omega | \sum_{s=0}^{t} [a^T \phi(s)]^2 \text{ diverges for every constant column vector } a \neq 0\}$$

Proof. The corresponding proof for P_t in Sternby (1977) goes through without changes, since only the structure of equation (5) is used. #

Using Theorems 3 and 4 we see that there are only two possibilities:

I) $S_t \to 0 \Rightarrow \hat{\xi}_t \to x$,

or II) $S_t \not\to 0 \Rightarrow \sum [a^T \phi(t)]^2$ converges for some $a \neq 0 \Rightarrow a^T \phi(t) \to 0$.

The second case imposes certain restrictions on the control law. For the analysis the following lemma is useful.

Lemma 1: Let $\{e(t)\}$ be a sequence of independent, zero mean normal random variables with a variance $\sigma^2(t)$ bounded away from zero, $0 < \varepsilon \le \sigma^2(t)$, for all t. Let $\{v(t)\}$ be another sequence with $e(t)$ independent of $\{v(k);\ k \le t\}$. Then $\sum^t [e(k) + v(k)]^2$ diverges a.e.

This result is intuitively clear, and the proof (using e.g. the extended Borel-Cantelli lemma, see Breiman (1968)) is omitted, being a standard exercise in probability theory.

Using the lemma, the a in case II above can be further specified. If u(k) is the latest input present in $a^T \phi$ then all components of a, corresponding to y(t) for $t > k$, must be zero. This is because y(t) always contains e(t), and $\sum [a^T \phi(t)]^2$ would otherwise diverge according to the lemma.

It is thus clear that either the parameters converge to their true values or the control law will converge to a linear one satisfying $a^T \cdot \phi(t) = 0$.

As a special case, if $\phi(t)$ does not contain any inputs, (pure AR model or all b-parameters known) then II) cannot happen, and the estimates are consistent.

For closed loop systems there are several ways to guarantee consistency by avoiding $a^T\phi(t) \to 0$:

- use a nonlinear control law
- use a time-varying control law, e.g. an ever-lasting shift between two constant and linear control laws
- use a more complex linear control law than can be described by $a^T\phi = 0$.

These results agree with those of Gustavson et. al (1977) for more general identification schemes.

In the open loop case, those components of a which correspond to outputs must be zero. This can be seen from Lemma 1 and the fact that with open loop control the noise and input sequences are independent. In the next section (Theorem 5) the results are shown to imply that all estimates corresponding to outputs will be consistent. As discussed in Sternby (1977), a condition similar to, but weaker than persistently exciting input is required to assure consistency also for the input parameters.

5. Adaptive control laws

In order to handle general control laws, where possibly $a^T\phi(t) \nrightarrow 0$, we need more detailed information on which parameters will converge. The following lemma concerning the null space of S_∞, is then useful.

Lemma 2: If $S_0 > 0$ then,

$$a \perp N(S_\infty) \Rightarrow \sum_{s=0}^{\infty} [a^T\phi(s)]^2 < \infty \text{ , a.e.}$$

Proof. The same technique is used as in the appendix of Sternby (1977) proving a similar lemma. In particular we may assume that S_∞ is diagonal, with entries $\lambda_1, \lambda_2, \ldots, \lambda_n$ in the diagonal, so that λ_j is the jth eigenvalue of S_∞. Then

$$(S_t^{-1})_{jj} \leq 1/\lambda_j \quad \text{for every } \lambda_j \neq 0.$$

This is shown as in Sternby (1977) for one element at a time by setting all eigenvalues in S_∞ to zero except λ_j. But a has no components in $N(S_\infty)$ so that for any t,

$$a^T S_t^{-1} a \leq \sum_{\lambda_j \neq 0} (\frac{1}{\lambda_j}) \cdot a^T a < \infty.$$

Now, as in the proof of Theorem 4 of Sternby (1977), $S_t^{-1} = S_0^{-1} + \sum_0^t \phi(s)\phi(s)^T$, so that $a^T S_t^{-1} a = a^T S_0^{-1} a + \sum_0^t [a^T\phi(s)]^2 > \sum_0^t [a^T\phi(s)]^2$, which concludes the proof of the theorem. #

The next theorem shows what will happen with the LS method for general control laws. Notice that no stability condition is required, and that no control law is specified.

Theorem 5. If the hypothesis (i) of Theorem 3 is satisfied and $S_0 > 0$ then

$$\sum_{t=0}^{\infty} [(\hat{\xi}_\infty - x)^T\phi(t)]^2 < \infty \text{ , a.e.}$$

Proof. Theorem 3 shows that with probability one, $\hat{\xi}_\infty - x$ is orthogonal to any given vector in the null space of S_∞. To be able to use Lemma 2, however, we must show that w.p. 1, $\hat{\xi}_\infty - x$ is simultaneously orthogonal to all the vectors in the null space of S_∞. Since $N(S_\infty)$ is finite dimensional, it is spanned by a set of random vectors $v_1, \ldots, v_p$, where $p = \dim(N(S_\infty))$ also is random. Further let $v_j = 0$ for $p < j \leq n$ and define $\Omega_j = \{\omega | v_j^T(\hat{\xi}_\infty - x) = 0\}$ so that $P(\Omega_j) = 1$ for each j by Theorem 3, since clearly $S_\infty v_j = 0$ and v_j can be chosen such that $v_j \in F_\infty$, and then also

$$P(\hat{\xi}_\infty - x \perp v_j \text{ for } j=1,\ldots, p) = P(\bigcap_{j=1}^{p} \Omega_j) = 1.$$

Thus, with probability one $\hat{\xi}_\infty - x$ is orthogonal to $v_1, \ldots, v_p$ and hence to $N(S_\infty)$, and the result follows from Lemma 2. #

This result has also been obtained by Ljung (1974) in a non-Bayesian version, but under the assumption of closed loop stability.

With the assumptions of Theorem 3, any adaptive control law based on Least Squares estimation, where the estimates enter the control law only as $\hat{\xi}_t \phi(t)$ will thus converge to the corresponding control law for known parameters. Under these circumstances, closed loop stability for the adaptive system is closely coupled to the choice of control law for the corresponding problem with known parameters. To actually prove stability one has to examine the different algorithms separately. This is often straightforward, but will not be further discussed in this paper.

Several self-tuning regulators fall within the category discussed above. The common minimum variance type implicit self-tuners (regulator parameters directly estimated) are of this type, whether the leading b parameter is estimated or not. Also the implicit pole placement self-tuner fits into this framework.

For a more general adaptive control law, Theorem 3 shows that the parameters will almost always converge. Theorem 5 and Lemma 1 show that, unless $\hat{\xi}_\infty = x$ the final control law can be written $(\hat{\xi}_\infty - x)^T \phi(t) = 0$. This expression for the control law can be compared to its original definition, and it might then in some cases be possible to show convergence of the control law to the desired one. It is however also easy to give examples when this is not possible.

6. Conclusion

It has been shown that in the Gaussian white noise case, consistency for the LS estimate follows if the "conditional variance" S_t tends to zero. This requires no assumptions on stability or on the control law.

Furthermore, a condition is given for the "conditional variance" to tend to zero, which imposes certain restrictions on the control law. Finally, a general condition is given for the possible convergence points of the estimates. The last condition can in some cases be used to show convergence of adaptive control laws to the desired ones. The assumption of Gaussian noise is essential to our method of analysis. However, rather similar results to Theorem 3 hold also without this assumption, but with some conditions on the rate of convergence to zero by S_t. This is proved in an important paper by Lai and Wei (1982), which appeared after the submission of the present paper.

Using Theorem 5, and maybe extensions thereof, it is possible to examine the asymptotic properties of adaptive control laws. It might be difficult to get general results, but specific algorithms are easily analyzed. Implicit versions of self-tuning regulators have been discussed in this paper. In the corresponding explicit self-tuners the parameters of the control law are calculated from a linear transformation of the estimates and should be possible to handle with Theorem 5. The examination of other adaptive control laws with this tool is left as a topic for future work.

Acknowledgement

We want to thank an anonymous referee for a careful reading which led to substantial improvements of the presentation of this paper.

7. References

Åström, K.J. and Wittenmark, B. (1971): Problems of identification and control. J. Math. Anal. Appl. 34, 90.

Åström, K.J. (1978): Stochastic control problems. In W.A. Coppel (Ed.): Mathematical Control Theory. Springer Lecture Notes in Mathematics, 680, 1-67. Berlin.

Breiman, L. (1968): Probability. Addison-Wesley. New York.

Chung, K.L. (1968): A Course in Probability Theory. Harcourt, Brace & World, Inc., New York.

Gustavsson, I. et. al (1977): Identification of processes in closed loop-identifiability and accuracy aspects. Automatica 13, 59.

Jazwinski, A.H. (1970): Stochastic Processes and Filtering Theory. Academic Press.

Kalman, R.E. (1960): A new approach to linear filtering and prediction problems. Trans. ASME, Ser D. J. Basic Eng. 82, 35.

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

Ljung, L. and Wittenmark, B. (1974): Asymptotic Properties of Self-tuning Regulators. Rept. 7404, Dept. of Automatic Control, Lund Institute of Technology, Sweden.

Schwartz, L. (1965): On Bayes Procedures. Z. Wahrsch. verw. Geb. 4, 10.

Solo, V. (1981): Strong consistency of least squares estimators in regression with correlated disturbances. Ann. Statist. 9, 689.

Sternby, J. (1977): On consistency for the method of least squares using martingale theory. IEEE Trans. Aut. Control 22, 346.

PREPRINTS 1983

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Jacobsen, Martin: Two Operational Characterizations of Cooptional Times.

No. 2 Hald, Anders: Nicholas Bernoulli's Theorem.

No. 3 Jensen, Ernst Lykke and Rootzén, Holger: A Note on De Moivre's Limit Theorems: Easy Proofs.

No. 4 Asmussen, Søren: Conjugate Distributions and Variance Reduction in Ruin Probability Simulation.

No. 5 Rootzén, Holger: Central Limit Theory for Martingales via Random Change of Time.

No. 6 Rootzén, Holger: Extreme Value Theory for Moving Average Processes.

No. 7 Jacobsen, Martin: Birth Times, Death Times and Time Substitutions in Markov Chains.

No. 8 Hougaard, Philip: Convex Functions in Exponential Families.

PREPRINTS 1984

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Rootzén, Holger and Sternby, Jan: Consistency in Least Squares Estimation: A Bayesian Approach.