A PARTITIONED RECURSIVE ALGORITHM FOR THE ESTIMATION OF DYNAMICAL AND INITIAL-CONDITION PARAMETERS FROM CROSS-SECTIONAL DATA

David W. Porter, member, IEEE
Malcolm D. Shuster, senior member, IEEE
Bruce P. Gibbs, member, IEEE
Business and Technological Systems, Inc.
Aerospace Building, Suite 440
10210 Greenbelt Road
Seabrook, Maryland 20706

William S. Levine, senior member, IEEE
Department of Electrical Engineering
University of Maryland
College Park, Maryland 20742

Twenty-Second Conference on Decision and Control
San Antonio, Texas - December 14-16, 1983

A PARTITIONED RECURSIVE ALGORITHM FOR THE ESTIMATION OF DYNAMICAL AND INITIAL-CONDITION PARAMETERS FROM CROSS-SECTIONAL DATA

David W. Porter, member, IEEE
Malcolm D. Shuster, senior member, IEEE
Bruce P. Gibbs, member, IEEE
Business and Technological Systems, Inc.
Aerospace Building, Suite 440
10210 Greenbelt Road
Seabrook, Maryland 20706

William S. Levine, senior member, IEEE
Department of Electrical Engineering
University of Maryland
College Park, Maryland 20742

Abstract

Many practical applications require the simultaneous estimation of unknown dynamical parameters and unknown initial means and covariances from an ensemble of tests. A recursive algorithm which asymptotically obtains the maximum likelihood estimate of both sets of unknown parameters is presented. The computational requirements of the algorithm are greatly reduced by partitioning the parameter vector into initial and dynamical parameters and making use of a sufficient statistic as an intermediate variable for the estimation of initial condition parameters. This partitioning leads to a two-tier filter for calculating some of the required parameter sensitivities. The results are illustrated by an application to a simplified robotic system.

1. Introduction

The majority of work on system identification has been concerned with system parameter identification from single sample longitudinal data. Goodrich and Caines [1] point out that many identification problems require system parameters to be identified from cross-sectional non-stationary data. For example, in areas as diverse as robotics and inertial navigation it is often necessary to estimate both initial condition and dynamical parameters. The point is that it is frequently impossible to directly measure the complete initial state. Since dynamical parameters are unknown it is impossible to use standard state estimation techniques to obtain the initial conditions. Furthermore, if the same instrumentation is used to test an ensemble of systems, then troublesome correlations can arise between tests due to common instrumentation errors.

An obvious way to compute the parameter estimates from cross-sectional non-stationary data is via maximum likelihood or, more generally, some other prediction-error scheme [2]. Such schemes require batch processing of the data. A recursive approach offers many well known advantages relative to batch. An advantage particularly important for cross-sectional analysis is the ability to change the testing procedure and the system design in response to results from previous tests and then continue testing without having to reprocess the previous tests.

For many problems with realistic state sizes the computational burden of batch processing is so large that batch processing is impractical. One reason for this is that it is often necessary to estimate initial covariances as well as initial means and variances. Straightforward batch maximum likelihood methods [2] require differentiating a Kalman filter for every parameter being estimated. The result is an impractically large computational burden.

The main result in this paper is a recursive algorithm motivated by [7] that processes each data point exactly once and converges, as the number of data points goes to infinity, to the maximum likelihood estimates of both the dynamical parameters and the initial mean and covariance. The computational requirements of the algorithm are greatly reduced by partitioning the parameter vector into initial condition and dynamical parameters and making use of a sufficient statistic as an intermediate variable for the estimation of the initial-condition parameters. See [3,4] for earlier uses of a special case of this idea. This partitioning leads to a two-tier filter, as in the treatment of bias in recursive estimation [5,6], for calculating some of the required parameter sensitivities, which further reduces computation.

The paper is organized as follows. The system model is described in Section 2. Section 3 describes an algorithm which is recursive from test to test but which requires a batch computation for each test. In Section 4, this batch computation is replaced by a partitioned recursion within each test that greatly reduces the necessary computations. Section 5 contains a numerical example that is a simplified version of an application to manipulators. Finally, Section 6 gives some conclusions and two suggestions for further research.

2. System Model

Consider a system model of the form

$$\underline{x}^i(t+1) = \underline{A}^i_\theta(t)\,\underline{x}^i(t) + \underline{w}^i(t) \quad ; \; t=0,1,2,\ldots,n_i \quad i=1,2,\ldots,M \qquad (1)$$

$$\underline{y}^i(t) = \underline{C}^i_\theta(t)\,\underline{x}^i(t) + \underline{v}^i(t) \quad ; \quad t=1,2,\ldots,n_i \qquad (2)$$

where

$$\underline{w}^i(t) \sim N(\underline{0},\underline{W}^i_\theta(t))$$

$$\underline{v}^i(t) \sim N(\underline{0},\underline{V}^i_\theta(t))$$

$$\underline{x}^i(0) = \underline{a}^i + \underline{T}_i\,\underline{b}^i$$

$$\underline{a}^i \sim N(\underline{0},\underline{\Sigma}^i_{a\theta}) \; ; \; \underline{b}^i \sim N(\underline{\mu}_{b\theta},\underline{\Sigma}_{b\theta})$$

$\underline{w}^i(t)$, $\underline{v}^i(t)$ are "white" and are independent of each other and of $\underline{x}^i(0)$.

The subscript θ indicates dependence on an unknown vector of parameters $\underline{\theta} \in \Theta$.

The superscript i denotes the i^{th} test.

The initial condition of the i^{th} test, $\underline{x}^i(0)$, has been decomposed into two components. One component $\underline{a}^i$ is assumed to have known (zero) mean and known covariance or a covariance which is a known function of possibly unknown dynamical parameters so that the statistics of $\underline{a}^i$ will not be estimated as separate parameters. Such a component in the per test initial condition arises, for example, from Markov processes which achieve steady state before the inception of the test. The component $T_i \underline{b}^i$ is that part of the the initial condition which is not a known function of the system dynamical parameters. The matrix T_i is assumed known and is included to allow for the possibility that the initial-condition parameters of interest ($\underline{\mu}_{b\theta}$, $\underline{\Sigma}_{b\theta}$) need not enter each test in identical fashion.

There are several points that should be emphasized in connection with the model. First, the time dependence of $\underline{A}^i_\theta(t)$ can be the result of a feedback control system, possibly including a Kalman filter, being applied to the basic dynamical system. Since one might change the feedback control gains as a result of earlier tests, $A^i_\theta(t)$ can vary with i (as denoted by the superscript i). Similarly, we can allow the conditions of the i^{th} test to change based on information from previous tests. Of course, the unknown parameters do not change. Although more general results are possible, we will assume that all of the dependence on previous tests is known and linear. For example, correlated tests can be described this way. This preserves the Gaussian distribution of the complete set of data. Second, the unknown $\mu_{b\theta}$ and $\Sigma^i_{b\theta}$ guarantee the need for multiple tests.

3. Test-to-Test Recursive Algorithm

The basic problem is to find $\hat{\underline{\theta}}_{ML}$, the maximum likelihood estimate of $\underline{\theta}$, given the data $\underline{y}^i(t)$ for $t = 0,1,\ldots,n_i$ and $i = 1,2,\ldots,M$. By definition,

$$\hat{\underline{\theta}}_{ML} = \arg \min_{\underline{\theta}} L(\underline{\theta},\underline{Y}) \tag{3}$$

where $L(\underline{\theta},\underline{Y})$ denotes the negative log likelihood and $\underline{Y}$ denotes the collection of $\underline{y}^i(t)$ for $t = 1,2,\ldots,n_i$ and $i = 1,2,\ldots,M$. Of course, under the assumptions we have made $\underline{Y}$ is Gaussian distributed with some mean and variance which depend on $\underline{\theta}$. Because of our assumption that the details of the i^{th} test can depend on the results of previous tests, the individual tests are not independent. Thus, some labor is required in order to obtain a recursive calculation of $\hat{\underline{\theta}}_{ML}$.

The first step is to concatenate the data vectors $\underline{y}^i(t)$ for the i^{th} test to form a vector $\underline{y}(i) = [\underline{y}^i(1)^T , \underline{y}^i(2)^T \ldots, \underline{y}^i(n_i)^T]^T$.

Then

$$\underline{y}(i) = \underline{H}^i_\theta \underline{x}^i(0) + \underline{v}(i) \qquad i = 1,2,\ldots,M \tag{4}$$

where $\underline{H}^i_\theta$ is computed from the $\underline{A}^i_\theta(t)$ and $\underline{C}^i_\theta(t)$, $\underline{v}(i) \sim N(\underline{0},R^i_\theta)$, and $\underline{R}^i_\theta$ is computed from $A^i_\theta(t)$, $C^i_\theta(t)$, $\underline{V}^i_\theta(t)$, and $\underline{W}^i_\theta(t)$.

If the tests were all independent we could now rewrite Eq. (3) in the more convenient form

$$\hat{\underline{\theta}}_{ML} = \arg \min_{\underline{\theta}} \frac{1}{M} \sum_{i=1}^{M} \ell(\underline{\theta},\underline{y}(i)) \tag{5}$$

where

$$\ell(\underline{\theta},\underline{y}(i)) = \frac{1}{2} \operatorname{tr} \left[\underline{S}^{i\,-1}_\theta (\underline{y}(i)-\hat{\underline{y}}(i)) (\underline{y}(i)-\hat{\underline{y}}(i))^T \right] + \frac{1}{2} \log \det \underline{S}^i_\theta \tag{6}$$

and

$$\underline{S}^i_\theta = \underline{H}^i_\theta \underline{\Sigma}_{a\theta} \underline{H}^{i\,T}_\theta + \underline{H}^i_\theta \underline{T}_i \underline{\Sigma}_{b\theta} \underline{T}^T_i \underline{H}^{i\,T}_\theta + \underline{R}^i_\theta = \operatorname{cov}(\underline{y}(i)) \tag{7}$$

$$\hat{\underline{y}}(i) = E(\underline{y}(i)) = \underline{H}^i_\theta \underline{T}_i \underline{\mu}_{b\theta} \tag{8}$$

Even though the tests are not independent we can obtain essentially the same form for $\hat{\underline{\theta}}_{ML}$ as is given in Eq. (5). The idea, explained in detail in [2], is to write L in terms of conditional probabilities. In detail, let $\underline{Y}_i$ denote the data set consisting of all $\underline{y}^i(t)$ such that $t = 1,2,\ldots,n_i$ and $i = 1,2,\ldots,J$. We can then rewrite Eq. (3) as

$$\hat{\underline{\theta}}_{ML} = \arg \min_{\underline{\theta}} \frac{1}{M} \sum_{i=1}^{M} \hat{\ell}(\underline{\theta},\underline{y}(i)) \tag{9}$$

where

$$\hat{\ell}(\theta,y(i)) = \frac{1}{2} \operatorname{tr} \left[\hat{\underline{S}}^{i\,-1}_\theta (\underline{y}(i) - \hat{\underline{y}}(i|\underline{Y}_{i-1})) (\underline{y}(i) - \hat{\underline{y}}(i|\underline{Y}_{i-1}))^T \right] + \frac{1}{2} \log \det \hat{\underline{S}}^i_\theta \tag{10}$$

and

$$\hat{\underline{y}}(i|\underline{Y}_{i-1}) = E(\underline{y}(i)|\underline{Y}_{i-1}) \tag{11}$$

$$\hat{\underline{S}}^i_\theta = \operatorname{cov} (\underline{y}(i) - \hat{\underline{y}}(i|\underline{Y}_{i-1})) \tag{12}$$

Because of our assumptions that the conditions of the i^{th} test depend linearly on the results of previous tests and that this dependence is known, all of the terms in Eqs. (9-12) can be computed explicitly. Note that all of the expectations and covariances in Eqs. (7), (8), (11) and (12) are evaluated as func-

tions of $\underline{\theta}$, as if $\underline{\theta}$ was known. Thus, the actual calculation of $\hat{\underline{\theta}}_{ML}$ via Eq. (3) or (5) or (9) is still fundamentally a batch calculation. However, Eqs. (9) through (12) express $\hat{\underline{\theta}}_{ML}$ in terms of the so-called prediction error and Ljung [7] has given a technique for converting a batch prediction error estimator to a recursive estimator whose estimate converges to the batch estimate as the amount of data goes to infinity.

Ljung's procedure amounts to writing

$$\hat{\underline{\theta}}(i) = \hat{\underline{\theta}}(i-1) + \gamma(i)\, \underline{R}^{-1}(i) \left[-\frac{d\hat{\ell}}{d\underline{\theta}} \Big|_{\hat{\underline{\theta}}(i-1),\underline{\varepsilon}(i)} \right] \quad (13)$$

$$\underline{R}(i) = \underline{R}(i-1) + \gamma(i) \left\{ \frac{d^2\hat{\ell}}{d\underline{\theta}^2} \Big|_{\hat{\underline{\theta}}(i-1),\underline{\varepsilon}(i)} + \delta \underline{I} - \underline{R}(i-1) \right\} \quad (14)$$

where

$\hat{\underline{\theta}}(i)$ is the estimate of $\underline{\theta}$ based on data $\underline{Y}_i$

$\gamma(i) = 1/i$ although more general forms are possible

$$\underline{\varepsilon}(i) = \underline{y}(i) - E(\underline{y}(i) | \underline{Y}_{i-1}, \hat{\underline{\theta}}(i-1)) \quad (15)$$

δ is a small positive number that is large enough to ensure $\underline{R}(i) > 0$ for all i

and

$$\hat{\ell}(\underline{\theta},\underline{\varepsilon}(i)) = \frac{1}{2} \operatorname{tr} [\hat{\underline{S}}_{\theta}^{i\,-1} \underline{\varepsilon}(i)\, \underline{\varepsilon}^T(i)] + \frac{1}{2} \log \det \hat{\underline{S}}_{\theta}^{i} \quad (16)$$

Ljung [7] proves, under some assumptions we will discuss below, that $\hat{\underline{\theta}}(i)$ converges w.p.1 either to the set

$$D_C = \{\underline{\theta} \mid \frac{d}{d\underline{\theta}} \bar{V}(\underline{\theta}) = 0\} \quad (17)$$

where, it can be shown that

$$\bar{V}(\underline{\theta}) = \lim_{M\to\infty} L(\underline{\theta},\underline{Y}_M) \quad (18)$$

or to the boundary of the model set as $i \to \infty$. Actually convergence is proved for a class of positive semi-definite approximations to

$$\frac{d^2\hat{\ell}}{d\underline{\theta}^2}.$$

Furthermore, among isolated points of D_C, only local minima of $\bar{V}(\underline{\theta})$ are possible convergence points. Note that $\hat{\underline{\theta}}(i)$ really converges to a local minimum of the negative log likelihood or to a value on the boundary of the admissible parameter set. However, this is all any batch algorithm achieves.

This convergence result is based on three assumptions:

A1: $$\bar{V}(\underline{\theta}) = \lim_{M\to\infty} E(L(\underline{\theta},\underline{Y}_M)) \quad (19)$$

A2: $\hat{y}(i|\underline{Y}_{i-1})$ is computed via equations of the form

$$\underline{\phi}(i+1) = \underline{F}(\underline{\theta})\, \underline{\phi}(i) + \underline{G}(\underline{\theta})\, \underline{y}(i) \quad (20)$$

$$\hat{\underline{y}}(i|\underline{Y}_{i-1}) = \underline{H}(\underline{\theta})\, \underline{\phi}(i) + \underline{H}(\underline{\theta})\, \underline{\mu}_{b\theta} \quad (21)$$

$\underline{F}(\underline{\theta})$ has all its eigenvalues strictly inside the unit circle and $\underline{F}(\underline{\theta})$, $\underline{G}(\underline{\theta})$ and $\underline{H}(\underline{\theta})$ are twice differentiable for all $\underline{\theta}$ in the compact set of possible parameter values. Note that the second term on the right hand side of Eq. (21) is not included in Ljung's version. The extra term is needed to handle unknown initial conditions. The proof of convergence is straightforward.

A3: The test procedure that actually generates the data is "exponentially stable". That is, the influence of any test on future tests decreases exponentially.

Of course, it is not obvious that Ljung's result can be applied to the present problem. Thus, we must show that assumptions A1-A3 hold true here.

Assumption (A1) is proven by expanding (19) to

$$\bar{V}(\underline{\theta}) = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} E\{\hat{\ell}(\underline{\theta},\underline{\varepsilon}(i))\} \quad (22)$$

where expectation is with respect to the true $\underline{\theta}$ and $\underline{\varepsilon}(i)$ is calculated using some other $\underline{\theta}$. So,

$$\bar{V}(\underline{\theta}) = \lim_{M\to\infty} \frac{1}{2M} \sum_{i=1}^{M} [\log \det \hat{\underline{S}}_{\theta}^{i} + \operatorname{tr} [(\hat{\underline{S}}_{\theta}^{i})^{-1}\, P_{\varepsilon}^{i}(\underline{\theta})]] \quad (23)$$

where $P_{\varepsilon}^{i}(\underline{\theta})$ = true covariance of $\underline{\varepsilon}(i)$. In order for the limit on the right-hand side of the above equation to exist we have to impose some condition on the way tests can vary. A sufficient condition is for the change in conditions from test to test to go to zero as $i \to \infty$. If this condition is satisfied then

$$\bar{V}(\underline{\theta}) = \frac{1}{2} \log \det \hat{\underline{S}}_{\theta} + \frac{1}{2} \operatorname{tr} [\hat{\underline{S}}_{\theta}^{-1}\, P_{\underline{\varepsilon}}(\underline{\theta})] \quad (24)$$

It is trivial to show that the remaining assumptions are satisfied when the tests are independent ($\underline{F}(\underline{\theta}) = \underline{0}$, $\underline{H}(\underline{\theta}) = \underline{H}_{\theta}^{i} T_i$, $\underline{G}(\underline{\theta}) = \underline{0}$). When the tests are coupled via some dependence of test conditions on previous tests then the assumptions impose conditions on the form of coupling. The key is Eqs. (20) and (21).

Note that Ljung's result also can be used to show that

$$\lim_{i\to\infty} \underline{R}(i) - \delta \underline{I}$$

$$= \lim_{M\to\infty} \frac{1}{M} \{\text{Fisher information matrix at truth}\} \qquad (25)$$

4. Calculation of Quantities for Individual Tests

In general, even for independent tests, the measurements within a test will be highly correlated. Hence, the computation of

$$\frac{d\hat{\ell}(\underline{\theta},\underline{\varepsilon}(i))}{d\underline{\theta}} \quad \text{and of} \quad \frac{d^2\hat{\ell}(\underline{\theta},\underline{\varepsilon}(i))}{d\underline{\theta}^2}$$

is potentially very burdensome. The balance of this section describes, in four steps, the reduction of the computational requirements to a reasonable level.

Step 1: There are two well known simplifications. First, write $\hat{\ell}$ in terms of the innovations process within a single test. That is

$$\hat{\ell}(\underline{\theta},\underline{\varepsilon}(i)) \qquad (26)$$

$$= \frac{1}{2} \sum_{k=1}^{n_i} \{\underline{\nu}_{\underline{\theta}}^{i\,T}(k)\, \underline{B}_{\underline{\theta}}^{i\,-1}(k)\, \underline{\nu}_{\underline{\theta}}^{i}(k) + \log \det \underline{B}_{\underline{\theta}}^{i}(k)\}$$

where

$$\underline{\nu}_{\underline{\theta}}^{i}(t) \qquad (27)$$

$$= \underline{y}^i(t=k) - E_{\underline{\theta}}(\underline{y}^i(t=k) \,|\, \underline{Y}_{t-1}, \underline{y}^i(1), \ldots \underline{y}^i(t=k-1))$$

and

$$\underline{B}_{\underline{\theta}}^{i}(k) = E_{\underline{\theta}}\{\underline{\nu}_{\underline{\theta}}^{i}(k)\, \underline{\nu}_{\underline{\theta}}^{i\,T}(k)\} \qquad (28)$$

Then, the calculation of $\frac{d\hat{\ell}}{d\underline{\theta}}$ is straightforward and gives (suppressing the test index i and the dependence on $\underline{\theta}$)

$$\frac{\partial \hat{\ell}(\underline{\theta},\underline{\varepsilon}(i))}{\partial \theta_m}$$

$$= \sum_{k=1}^{n} \{\underline{\nu}^T(k)\, \underline{B}^{-1}(k)\, \frac{\partial \underline{\nu}}{\partial \theta_m}(k)$$

$$- \frac{1}{2}\, \underline{\nu}^T(k)\, \underline{B}^{-1}(k)\, \frac{\partial \underline{B}(k)}{\partial \theta_m}\, \underline{B}^{-1}(k)\, \underline{\nu}(k)$$

$$+ \frac{1}{2} \operatorname{tr}\, [\underline{B}^{-1}(k)\, \frac{\partial \underline{B}(k)}{\partial \theta_m}]\} \qquad (29)$$

Second, approximate $\frac{d^2\hat{\ell}}{d\underline{\theta}^2}$ by

$$\frac{\partial^2 \hat{\ell}}{\partial \theta_m \partial \theta_n} = \sum_{k=1}^{n} \{\frac{1}{2} \operatorname{tr}\, [\underline{B}^{-1}(k)\, \frac{\partial \underline{B}(k)}{\partial \theta_m}\, \underline{B}^{-1}(k)\, \frac{\partial \underline{B}(k)}{\partial \theta_n}]$$

$$+ \operatorname{tr}\, [\underline{B}^{-1}(k)\, \frac{\partial \underline{\nu}(k)}{\partial \theta_m}\, \frac{\partial \underline{\nu}^T(k)}{\partial \theta_n}]\} \qquad (30)$$

The fact that the expectation of the right hand side of (30) is exactly the per-test Fisher information matrix is sufficient to guarantee that the results of the previous section are unaffected by the approximation.

Note that the calculations required in Eq. (29) and (30) require that one construct and run a Kalman filter to give $\underline{B}(k)$ and $\underline{\nu}(k)$. Then, these equations must be differentiated with respect to $\underline{\theta}$ to give difference equations which must be sequentially solved to give

$$\frac{\partial \underline{B}(k)}{\partial \theta_m} \quad \text{and} \quad \frac{\partial \underline{\nu}(k)}{\partial \theta_m} \quad \text{for each } m .$$

The computational burden of these operations can be considerable.

Step 2: A major reduction in computations is achieved by avoiding differentiation of Kalman filters with respect to the elements of $\underline{\mu}_{b\theta}$ and $\underline{\Sigma}_{b\theta}$. We demonstrate this for the case where the tests are independent. For convenience of presentation, it is assumed here that the parameters to be estimated are the elements of $\underline{\mu}_{\theta} = \underline{\mu}$ and $\underline{\Sigma}_{b\theta} = \underline{\Sigma}$ along with the parameters in $\underline{A}_{\underline{\theta}}^{i}(t)$, $\underline{C}_{\underline{\theta}}^{i}(t)$, $\underline{W}_{\underline{\theta}}^{i}(t)$, $R_{\underline{\theta}}^{i}(t)$ and $\underline{\Sigma}_{\underline{\theta}}^{i}$. The $\underline{\mu},\underline{\Sigma}$ parameters are contained in the vector $\underline{\beta}$ and the remaining parameters are denoted by $\underline{\alpha}$ so $\underline{\theta} = [\underline{\beta}^T, \underline{\alpha}^T]^T$.

The $\underline{\mu},\underline{\Sigma}$ derivatives are avoided by expressing the likelihood in terms of the per-test maximum-likelihood estimate of $\underline{b}^i$ denoted by $\hat{\underline{b}}^i$. This can be done because $\hat{\underline{b}}^i$, $i = 1,\ldots,M$ is a sufficient statistic for $\underline{\beta}$ as shown below through an argument relying on and generalizing the results in [3,4].

Since $\hat{\underline{b}}^i$ is a sufficient statistic for $\underline{b}^i$, the factorization criterion [8] provides that

$$p(\underline{y}(i) \,|\, \underline{b}^i) = h_i(\underline{y}(i))\; p(\hat{\underline{b}}^i \,|\, \underline{b}^i) \qquad (31)$$

where h_i is a function only of the data multiplying the density of $\hat{\underline{b}}^i$ to get the likelihood function relative to $\underline{b}^i$, namely $p(\underline{y}(i) \,|\, \underline{b}^i)$. Now multiplying by the density of $\underline{b}^i$ given $\underline{\mu}$, Σ and integrating out $\underline{b}^i$ gives the likelihood relative to $\underline{\beta}$ as

$$p(\underline{y}(i)) \,|\, \underline{\mu},\underline{\Sigma}) = h_i(\underline{y}(i))\; p(\underline{\mu},\underline{\Sigma}) \qquad (32)$$

But h_i can be determined [illegible] $\underline{b}^i = \underline{0}$ and from (32)

$$p(\underline{y}(i)|\underline{\mu},\underline{\Sigma}) = \frac{p(\underline{y}(i)|\underline{b}^i = 0)}{p(\hat{\underline{b}}^i|\underline{b}^i = 0)} \cdot p(\hat{\underline{b}}^i|\underline{\mu},\underline{\Sigma}) \qquad (33)$$

Now substituting the respective densities for $\hat{\underline{b}}^i|\underline{\mu},\underline{\Sigma}$ and $\hat{\underline{b}}^i|(\underline{b}^i = 0)$ and using the Kalman filter representation for the density of $\underline{y}(i)$ gives within additive constants

$$- \log p(\underline{y}(i)|\underline{\mu},\underline{\Sigma})$$

$$= \frac{1}{2}\Big[\log \det (\underline{\Sigma} + \underline{P}(i))$$

$$+ (\hat{\underline{b}}^i - \underline{\mu})^T (\underline{\Sigma} + \underline{P}(i))^{-1}(\hat{\underline{b}}^i - \underline{\mu})$$

$$- \log \det \underline{P}(i) - \hat{\underline{b}}^{i\,T} \underline{P}(i)^{-1} \hat{\underline{b}}^i$$

$$+ \sum_{k=1}^{n_i} (\log \det \underline{B}_I^i(k) + \underline{\nu}_I^{i\,T}(k)\, \underline{B}_I^{i\,-1}(k)\, \underline{\nu}_I^i(k))\Big] \qquad (34)$$

where $\underline{P}(i)$ is the estimate error covariance of $\hat{\underline{b}}^i$. The I subscript on the Kalman filter quantities refers to the fact that the filter is ignorant of $\underline{b}^i$ since $\underline{b}^i = 0$. Note that the filter in (26) models $\underline{b}^i \sim N(\underline{\mu},\underline{\Sigma})$ and consequently may have a much higher state dimension. Equation (34), if summed over the tests, proves sufficiency of $\hat{\underline{b}}^i$, $i=1,\ldots,M$, for $\underline{\mu},\underline{\Sigma}$. Further, (34) is a replacement of (26) for $\hat{\ell}(\underline{\theta},\underline{\varepsilon}(i))$ where $\hat{\underline{b}}^i$, $\underline{P}(i)$, $\underline{\nu}_I^i(k)$, and $\underline{B}_I^i(k)$ are functions of $\underline{\alpha}$. Thus, $\hat{\ell}(\underline{\theta},\underline{\varepsilon}(i))$ is an explicit function of $\underline{\mu},\underline{\Sigma}$ that can be differentiated without differentiating Kalman filters.

Step 3:

The maximum-likelihood estimate of the initial condition $\hat{\underline{b}}^i$ can be computed from the residuals of the Kalman filter in which $\underline{b}^i$ has been arbitarily set to zero (ignorant filter) [5,6], again assuming independence. Note that

$$\underline{\nu}_I^i(k) = \underline{T}_k^i\, \underline{b}^i + \underline{\nu}_{TI}^i(k) \qquad (35)$$

where $\underline{T}_k^i$ is a transformation relating the true $\underline{b}^i$ to ignorant filter residuals and $\underline{\nu}_{TI}^i(k)$ is the filter residual that would have been obtained if $\underline{b}^i$ had truly been zero. $\underline{T}_k^i$ is easily expressed as a simple recursion using transition matrices and Kalman gains from the ignorant filter. Further, $\underline{\nu}_{TI}^i(k)$ is a white sequence with covariance given exactly by $\underline{B}_I^i(k)$. Consequently, $\hat{\underline{b}}^i$, $\underline{P}(i)$ can be computed by treating (35) as a measurement on $\underline{b}^i$ and performing numerically efficient and well-conditioned estimation of $\underline{b}^i$ based on stacking the measurements over time.

Now the derivatives of $\hat{\ell}(\underline{\theta},\underline{\varepsilon}(i))$ with respect to $\underline{\mu}$, $\underline{\Sigma}$, $\underline{\alpha}$ are straightforward to compute from (34) recognizing that $\hat{\underline{b}}^i$, $\underline{P}(i)$, $\underline{\nu}_I^i(k)$ and $\underline{B}_I^i(k)$ are all functions of $\underline{\alpha}$. The derivatives of the ignorant filter quantities plus $\hat{\underline{b}}^i$ and $\underline{P}(i)$ are in turn obtained as recursions by differentiating the ignorant filter recursions.

If there are many $\underline{\mu},\underline{\Sigma}$ parameters then the above provides an enormous computational savings relative to equation (29). The Fisher information relative to $\underline{\mu},\underline{\Sigma}$ can be used as an approximation to the Hessian of $\hat{\ell}(\underline{\theta},\underline{\varepsilon}(i))$

$$\frac{\partial^2 \hat{\ell}(\underline{\theta},\underline{\varepsilon}(i))}{\partial \underline{\mu}\, \partial \underline{\mu}^T} = (\underline{\Sigma} + \underline{P}(i))^{-1} \qquad (36)$$

$$\frac{\partial^2 \hat{\ell}(\underline{\theta},\underline{\varepsilon}(i))}{\partial \mu_j\, \partial \Sigma_{mn}} = 0 \quad \text{for all } j, m, n \qquad (37)$$

$$\frac{\partial^2 \hat{\ell}(\underline{\theta},\underline{\varepsilon}(i))}{\partial \Sigma_{mn}\, \partial \Sigma_{pq}}$$

$$= C_{mn}\, C_{pq}\, ([(\underline{\Sigma} + \underline{P}(i))^{-1}]_{mp}\, [(\underline{\Sigma} + \underline{P}(i))^{-1}]_{nq}$$

$$+ [(\underline{\Sigma} + \underline{P}(i)^{-1}]_{mq}\, [(\underline{\Sigma} + P(i))^{-1}]_{np}) \qquad (38)$$

where C_{mn} is $\frac{1}{2}$ for $m = n$ and 1 for $m \neq n$.

The approximate Hessian relative to $\underline{\alpha}$ is still provided by equation (30), which requires a standard Kalman filter and its derivatives relative to $\underline{\alpha}$. If the cross terms in the Hessian between $\underline{\mu},\underline{\Sigma}$ and $\underline{\alpha}$ are approximated by zero, then the convergence of $\hat{\underline{\theta}}(i)$ follows from the results given in Section 3. Thus, the algorithm given in step 2 up to this point is a completely recursive algorithm guaranteed asymptotically to perform as well as batch maximum likelihood. Further, derivatives of the Kalman filter with respect to $\underline{\mu},\underline{\Sigma}$ have been completely eliminated.

The result of Section 3 relating $\underline{R}(i)$ to the Fisher information does not follow because of the approximation for the cross terms in the Hessian. However, in many practical problems, such as inertial navigation or orbit determination, a core set of dynamics known from physical principles is driven by random biases $\underline{b}$ and measured through known dynamics. In such a situation the $\underline{H}_2^i$ of (4) is not a function of $\underline{\theta}$. Based solely on this fact the cross Fisher information between $\underline{\mu},\underline{\Sigma}$ and $\underline{\alpha}$ can be obtained after some manipulation and the cross-Hessian terms can be approximated by the Fisher information giving

$$\frac{\partial^2 \hat{\ell}(\underline{\theta},\underline{y}(i))}{\partial \mu_j\, \partial \alpha_m} = 0 \quad \text{for all } j \text{ and } m \qquad (39)$$

$$\frac{\partial^2 \hat{\ell}(\underline{\theta},\underline{y}(i))}{\partial \Sigma_{mn}\, \partial \alpha_j} = - C_{mn} \frac{\partial}{\partial \alpha_j} [(\underline{\Sigma} + \underline{P}(i))^{-1}]_{mn} \qquad (40)$$

The results of Section 3 relating $\underline{R}(i)$ to the Fisher information are unaffected by the above approximation to the Hessian.

Step 4: The above steps still require the computation of standard Kalman filter residual covariances and derivatives of the residuals and residual covariances with respect to $\underline{\alpha}$. However, the standard Kalman filter can be expressed as a two-tier combination of the ignorant Kalman filter already being used

and a bias-restoring filter using the results of [5], [6]. Further, the filter derivatives can be similarly expressed in terms of the existing ignorant filter derivatives and derivatives of the bias-restoring filter. Thus, the standard filter and its derivatives can be eliminated entirely with the addition of a bias-restoring filter and its derivatives. Note that the bias-restoring filter described here differs from the estimator giving $\underline{b}^i$ in that the bias-restoring filter must operate recursively and model $\underline{b} \sim N(\underline{\mu},\underline{\Sigma})$ to provide quantities for (30). As a final comment, even further computational advantages can be realized by making the ignorant filter ignorant also of bias states of known a priori mean and covariance where a reduction in state dimension occurs. All the quantities needed for parameter estimation can be obtained by again using the ideas in [5] and [6].

5. Numerical Example

Typical robot manipulators have dynamics of the form

$$\ddot{\underline{x}}(t) = \underline{f}(\underline{x}(t), \dot{\underline{x}}(t)) + \underline{G}\,(\underline{x}(t), \dot{\underline{x}}(t))\, \underline{u}(t) \qquad (41)$$

where

$\underline{x}(t)$ is a vector of generalized position coordinates

$\underline{u}(t)$ is the control

$\underline{G}(\underline{x},\dot{\underline{x}})$ is an invertible matrix for all $\underline{x}$, $\dot{\underline{x}}$

A possible control scheme for this system is to let

$$\underline{u}(t) = -\underline{G}^{-1}(\underline{x}(t),\dot{\underline{x}}(t))\,[\underline{f}(\underline{x}(t),\dot{\underline{x}}(t)) + \underline{G}(\underline{x}(t),\dot{\underline{x}}(t))\,\underline{\lambda}(t)] \qquad (42)$$

where

$\underline{x}(t)$ is a nominal path

$\underline{\lambda}(t)$ is a new control to be described below.

Under the extremely optimistic assumption that both of the equations above are exactly satisfied

$$\ddot{\underline{x}}(t) = \underline{\lambda}(t) \qquad (43)$$

If $\underline{\lambda}^i(t)$ is chosen to be a very tight linear feedback control (based on observations of $x^i(t)$) then, we can regard the errors which cause our system to deviate from Eq. (43) approximately as "noise". Since the sole purpose of our model is to enable us to design a good feedback controller for $\underline{x}^i(t)$, such a simple model may well be adequate.

Note that our example is a discretely sampled continuous time system rather than discrete. Hence the continuous time dynamics must be first written as equivalent discrete dynamics in order for the above results to be applied without modification (although these modifications are trivial).

The problem is then to implement the above control scheme on a manipulator, measure the performance of the manipulator in a given task and use this data to specify the unknown parameters of the model (thereby specifying the manipulator accuracy). The instrumentation that is normally used for these tests measures $\underline{x}(t)$ but not $\dot{\underline{x}}(t)$.

A simple example of the system is shown in Figure 1. The second order system is shown inside the dashed block. In addition to the control signal λ the system is driven also by zero-mean colored noise which has been modeled as a first-order Markov process, x_3

$$\dot{x}_3^i(t) = -\frac{1}{\tau}\, x_3^i(t) + w^i(t) \qquad , \qquad (44)$$

which is assumed to be in steady state, hence,

$$w^i(t) \sim \text{zero mean and white with power spectral density } q \qquad (45)$$

and

$$x_3^i(0) \sim N(0, q\tau/2) \qquad (46)$$

The state vector is $\underline{x} = [x_1 , x_2 , x_3]^T$.

The control signal is computed in our example by a Kalman filter which determines λ from noisy measurements z of x_1. Thus,

$$z^i(k) = \underline{C}^i(k)\, \underline{x}^i(k) + v^i(k) \qquad k = 1,2,\ldots,n_f \qquad (47)$$

where

$$C^i(k) = [1,0,0] , \quad v^i(k) \sim N(0, .0025) \qquad (48)$$

Measurements of the control signal are taken simultaneously with the measurements z. Hence, there is no need to specify the control law for the present example.

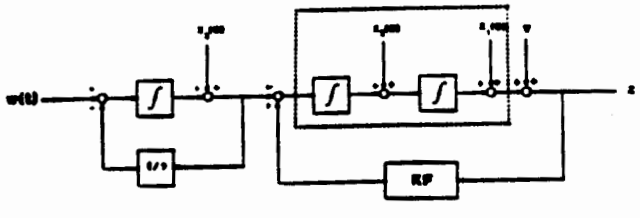

Fig. 1 System Block Diagram

It is assumed for our example that the tests are independent and that x_1 and x_2 are initially uncorrelated. The initial conditions of the system as given by Eqs. (1) and (2) above are

$$\underline{\Sigma}_{a\theta} = \text{Diag}\,(0,0,\, q\tau/2) \qquad , \qquad (49)$$

$$\underline{\Sigma}_{b\theta} = \text{Diag}\,(\Sigma_{11}, \Sigma_{22}) \quad , \tag{50}$$

$$\underline{\Gamma}_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{51}$$

The initial-condition and dynamic parameter vectors are, therefore $\underline{\beta} = [\mu_1, \mu_2, \Sigma_{11}, \Sigma_{22}]^T$, $\underline{\alpha} = [q, \tau]^T$. The measurement model and measurement error variance are assumed to be known.

Twenty tests were performed each containing 200 measurements taken at intervals of 1 sec. The estimates are compared with the true values of the parameters used in simulating the data in Table 1. The estimation errors (± one standard deviation) given in the table are computed as the square roots of the diagonal elements of the inverse of the Fisher information matrix, as given by Eqs. (36) and (38) for initial-condition parameters, or by the approximation of the Fisher information matrix given by Eq. (30) for the dynamical parameters. The batch estimate processed all twenty tests recursively iterating the scoring algorithm until convergence was reached. The recursive algorithm was begun by processing five tests in batch. The agreement is observed to be very satisfactory. Four of the six parameters fall within the predicted 1σ error levels as expected. The deviation of $\hat{\Sigma}_{22}$ from the true value might seem large. Half of this error, in fact, is due to the particular realization of the initial condition, which is known since the data is simulated. The sampled variance of the actual initial conditions on x_2 was found to be 0.38.

Parameter	True Value	Stating Value	Batch Estimate	Recursive Estimate
μ_1	1.0	0.	1.27 ± .20	1.35 ± .18
Σ_{11}	1.0	0.5	.77 ± .26	.84 ± .19
μ_2	.5	0.	.44 ± .16	.50 ± .14
Σ_{22}	.25	0.5	.43 ± .15	.49 ± .11
$1/\tau$	.05	.04	.054 ± .005	.052 ± .005
q	.025	.04	.0252 ± .001	.0252 ± .001

Table 1
Parameter Estimates for Twenty Tests

Figures 2 and 3 show the detailed behavior of the estimates of μ_1 and Σ_{11} (with their one standard deviation error bars) as a function of the numbers of tests. The first point in each case is obtained by batch ML estimation of the initial mean and variance using the first five tests. The succeeding points show the result of using Ljung's recursive algorithm, Eqs. (13) and (14) for tests 6 through 20. The large point on the far right is the result for the batch ML estimation using all twenty tests. The solid line shows the true sampled mean and variance of the actual realizations of the initial conditions. The estimated values are seen to follow the true sampled values with remarkable fidelity. Figures 4 shows similar results for $1/\tau$. Again the agreement is seen to be very satisfactory with the recursive estimate marching systematically toward the batch result. The estimation errors for q were so small even for five tests that these need not be shown.

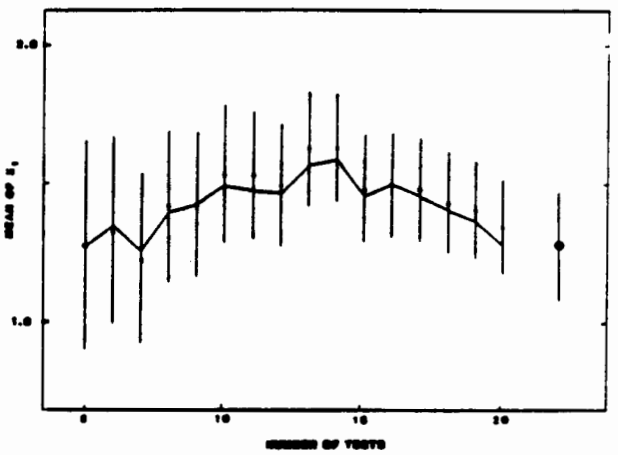

Fig. 2 Recursive and batch estimates of mean with 1σ error bars and samle mean

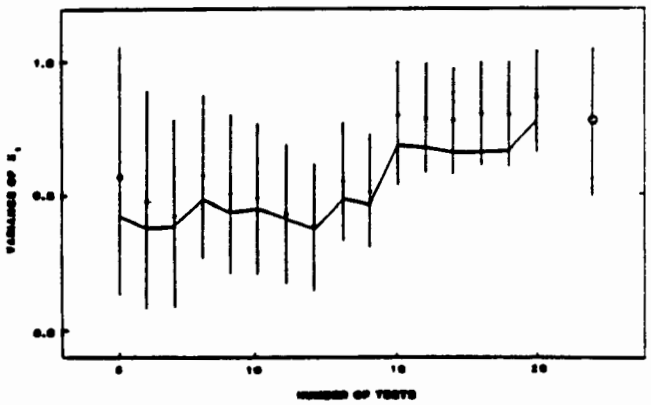

Fig. 3 Recursive and batch estimates of variance with 1σ error bars and true sample variance

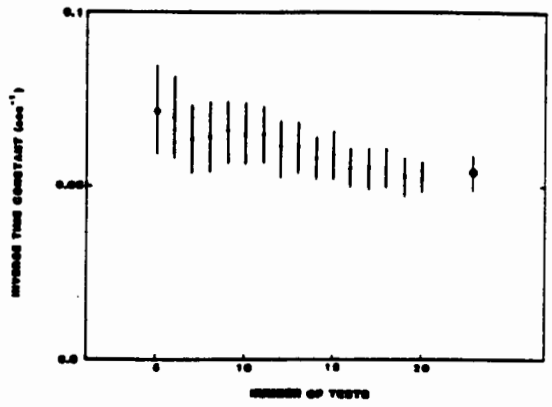

Fig. 4 Recursive and batch estimates of inverse time constant with 1σ error bars

6. Conclusions

We have exhibited a partitioned recursive algorithm for calculating the maximum likelihood estimate of the unknowns $\underline{\theta}$. We have shown that the recursion converges to the same estimates as the batch maximum likelihood approach when the number of tests tends to ∞. Note that these results hold when the tests are of unequal length and are carried out under different conditions. The partitioning of the parameters into initial condition and dynamic parameters provides a crucial improvement in numerical properties.

Two areas of further research are believed to be important. First, many of the potential applications of these ideas involve nonlinear systems. Since the results presented here rely primarily on derivatives of the log likelihood there is hope that they could be extended to apply to reasonable classes of nonlinear systems. Second, it would be very useful to have estimates of how the rate of convergence depends on the scheme for finding $\underline{R}(i)$.

Acknowledgement

The authors are grateful to Mr. Mark S. Asher for providing the numerical estimates appearing in Section 5.

References

[1] R. L. Goodrich and P. E. Caines, "Linear System Identification from Nonstationary Cross-Sectional Data," IEEE Trans. Automat. Contr., Vol. AC-24, no. 3, pp. 403-411, June 1979.

[2] K. J. Åström, "Maximum Likelihood and Prediction-Error Methods", Automatica, Vol. 16, no. 5, pp. 551-574, Sept. 1980.

[3] L. J. Levy, R. H. Shumway, D. E. Olsen, and F. C. Deal, Jr., "Model Validation from an Ensemble of Kalman Filter Tests", Proc. 21st Midwestern Symposium on Circuits and Systems, Ames, IA, 1978.

[4] R. H. Shumway, D. E. Olsen, and L. J. Levy, "Estimation and Test of Hypothesis for the Initial Mean and Covariance in the Kalman Filter", American Statistical Association Conference, San Diego, CA, August 1978.

[5] B. Friedland, "Treatment of Bias in Recursive Filtering", IEEE Trans. Automat. Contr., Vol. AC-17, no. 1, pp. 359-367, August 1969.

[6] G. Bierman, "The Treatment of Bias in the Square Root Information Filter Smoother", Conference on Decision and Control, 1973.

[7] L. Ljung, "Analysis of a General Recursive Prediction Error Identification Algorithm", Automatica, Vol. 17, no. 1, pp. 89-99, January 1981.

[8] C. R. Rao, Linear Statistical Inference and its Applications, John Wiley & Sons, New York, 1973.