April 1978 Report ESL-R-814

MAXIMUM LIKELIHOOD IDENTIFICATION OF STATE SPACE MODELS FOR LINEAR DYNAMIC SYSTEMS

by

Nils R. Sandell, Jr. and Khaled I. Yared

Abstract

Maximum likelihood (ML) identification of state space models for linear dynamic systems is presented in a unified tutorial form. First linear filtering theory and classical maximum likelihood theory are reviewed. Then ML identification of linear state space models is discussed. A compact user-oriented presentation of results scattered in the literature is given for computing the likelihood function, maximizing it, evaluating the Fisher information matrix and finding the asymptotic properties of ML parameter estimates. The practically important case where a system is described by a simpler model is also briefly discussed.

Electronic Systems Laboratory
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

TABLE OF CONTENTS

SECTION 1

INTRODUCTION

One of the most important problems in the study of systems (either physical or socio-economic) is that of obtaining adequate mathematical models to describe in some way the behavior of those systems. This is indeed an important issue since critical decisions about what to measure, what to control and in what manner, are usually explicitly or implicitly based on such models. For example, modern control theory results pertaining to the design of optimal observers and regulators assume the knowledge of dynamic models, usually in state space form.

The purpose of this report is to give, in tutorial form, a unified presentation of one approach to the problem of determining mathematical models from measurements made on the actual system. Attention is restricted to one important class of models (namely linear state space models) and one method (namely the maximum likelihood identification method). The report is mainly addressed to readers familiar with Kalman filtering techniques who would like to obtain a working knowledge of the maximum likelihood approach to system identification. It is not a survey of the subject and so our references are confined to those works that were directly used in preparing the report. Finally, we make no claim to originality except for our declared tutorial purpose.

1.1 Overview of Issues in System Identification

It is important to distinguish between modeling and identification. The former is a deductive process, using physical laws or even intuitive socio-economic relationships in establishing a model, whereas the latter is an inductive process, using data obtained from observations of the actual system (see Figure 1.1) for that purpose. But these are complementary in

the sense that from physical laws one can deduce model structures and possibly parameter ranges, while identification techniques give parameter values and verify proposed model structures (see e.g. Schweppe [1]). Those identified parameter values will of course depend on the system producing the data, but also on the experimental conditions, the hypothesized model structure (or model set), and the identification method used.

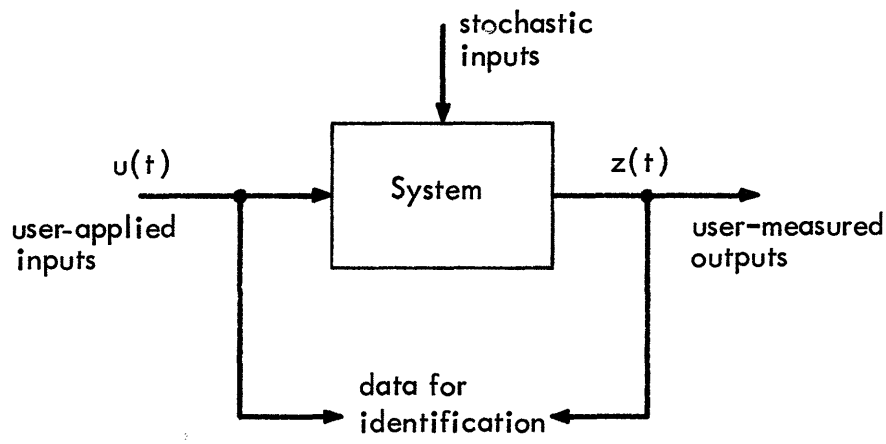


Figure 1.1 Data Gathering for Identification Experiment

Experimental conditions include the choice of the applied inputs (e.g. none, chosen in an open loop fashion, determined partly by some output feedback process, etc.). The case where no inputs are applied (and the system is driven by stochastic inputs only) corresponds to the case traditionally considered in time series analysis. Whenever it is possible to do so, choosing a good set of inputs to excite the system adequately is an impor-

tant issue which has been studied by, e.g., Mehra [2]. In some instances one might have a feedback loop around the system to be identified. The identification of systems under closed loop control may present significant problems. One example, pointed out by, e.g., Åström and Eykhoff [3], is illustrated in Figure 1.2: An attempt to identify H_0 from measurements of u and z by correlation analysis, say, will give the estimate of H_0 as being $\frac{1}{H_F}$ (i.e. the inverse of the transfer function of the feedback system). Note also that one might not be able or not even willing (e.g. for stability reasons) to open the loop in such cases. A good survey of identification of processes in closed loop has been given by Gustavsson, Ljung and Söderström [4]; we will not discuss this problem further in this report.

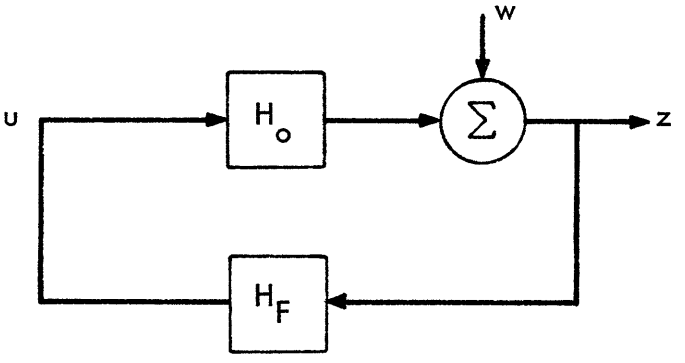


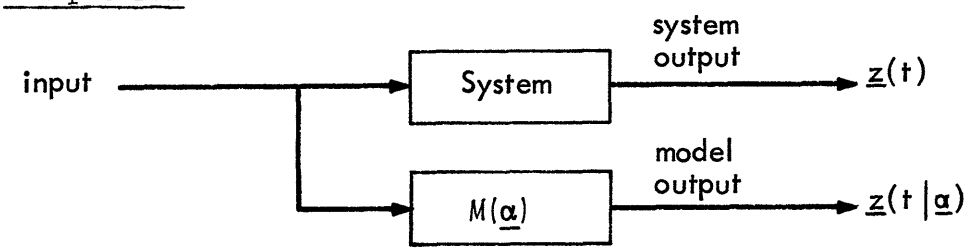
Figure 1.2 Simple Example of Closed Loop System

A model set M is a class of models describing a behavior of the system under study and parametrized in a certain manner by a parameter vector $\underline{\alpha}$. The parametrization of general multivariable systems is not an easy task and the structural aspects of linear systems identification are still an area of active study. However, the model set and its parametrization are often given from physical considerations. There are also so-called nonparametric techniques for establishing a model (e.g., cor-

relation and spectral analysis of time series when the system is driven by a stochastic input only; step and frequency response analysis when a user applied input is possible), but these will not be discussed in this report.

As for the various identification methods used to estimate values for the different parameters, once a model set has been chosen, they generally consist of minimizing some model-dependent function of the data (Example 1.1). The methods vary according to the choice of that function and the criteria for this choice range from ad-hoc considerations to tenets of statistical optimality. An excellent survey of identification methods with an extensive bibliography has been given by Åström and Eykhoff [3].

Example 1.1



choose $\underline{\alpha}$ to minimize

$$\ell(\underline{z}(0),\underline{z}(1),\ldots,\underline{z}(t);\underline{\alpha}) = \sum_{\tau=0}^{t} [\underline{z}(t) - \underline{z}(t|\underline{\alpha})]^2 \quad \blacksquare$$

One can also distinguish between on-line and off-line algorithms for system identification. In general, off-line algorithms estimate system parameters from a given, fixed set of input-output data whereas on-line algorithms, used when a model has to be identified in real time, update their parameter estimate as they receive new input-output data pairs. The boundary between on- and off-line methods is indistinct as it depends on process speed, requirements of the method and computational resources.

Finally, we remark that the accuracy required of an identification method depends on its particular use. For general purpose models or models to be used in filtering applications (Figure 1.3) one may need accurate parameter esti-

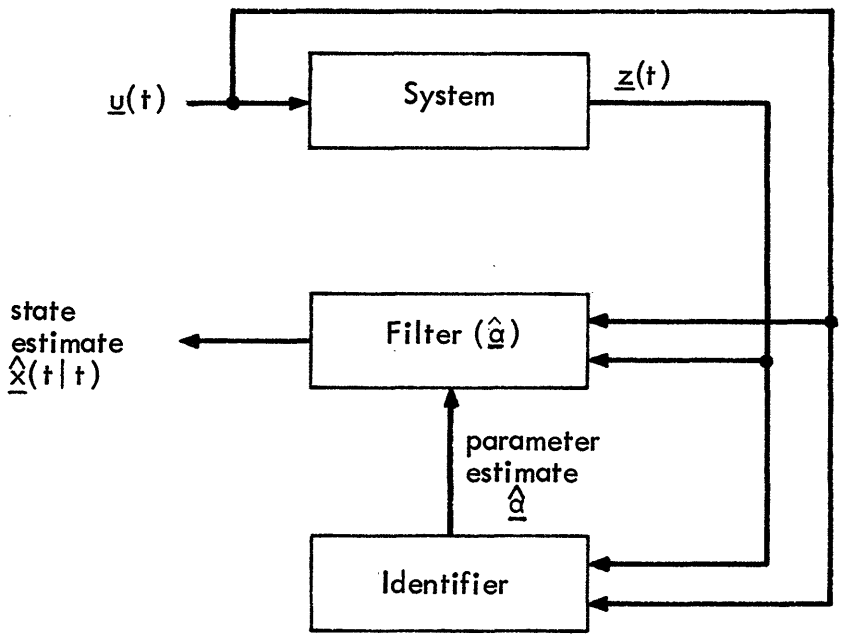


Figure 1.3 Adaptive Estimation

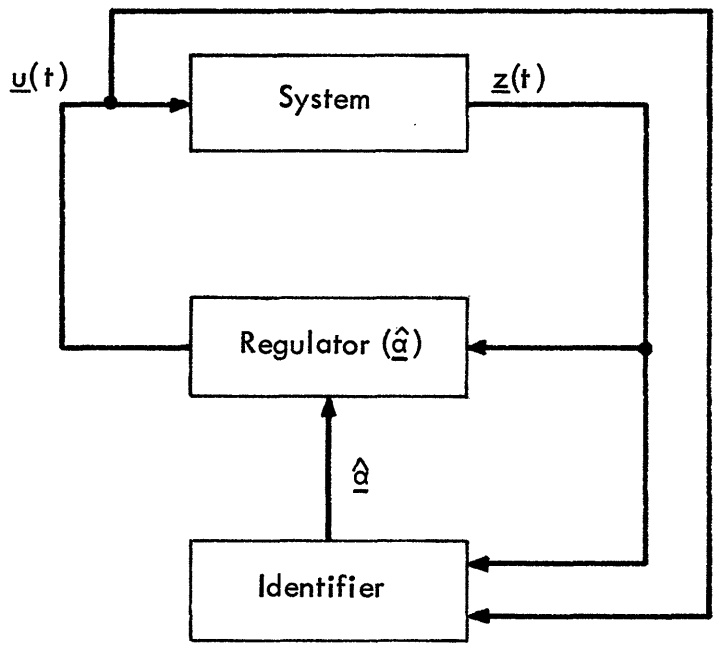


Figure 1.4 Adaptive Control

mates whereas for adaptive control (Figure 1.4) one may just need a good model of input-output behavior.

Figure 1.5 summarizes the interrelationship between the different concepts discussed above. The objectives of the present report are now discussed in the context of those issues.

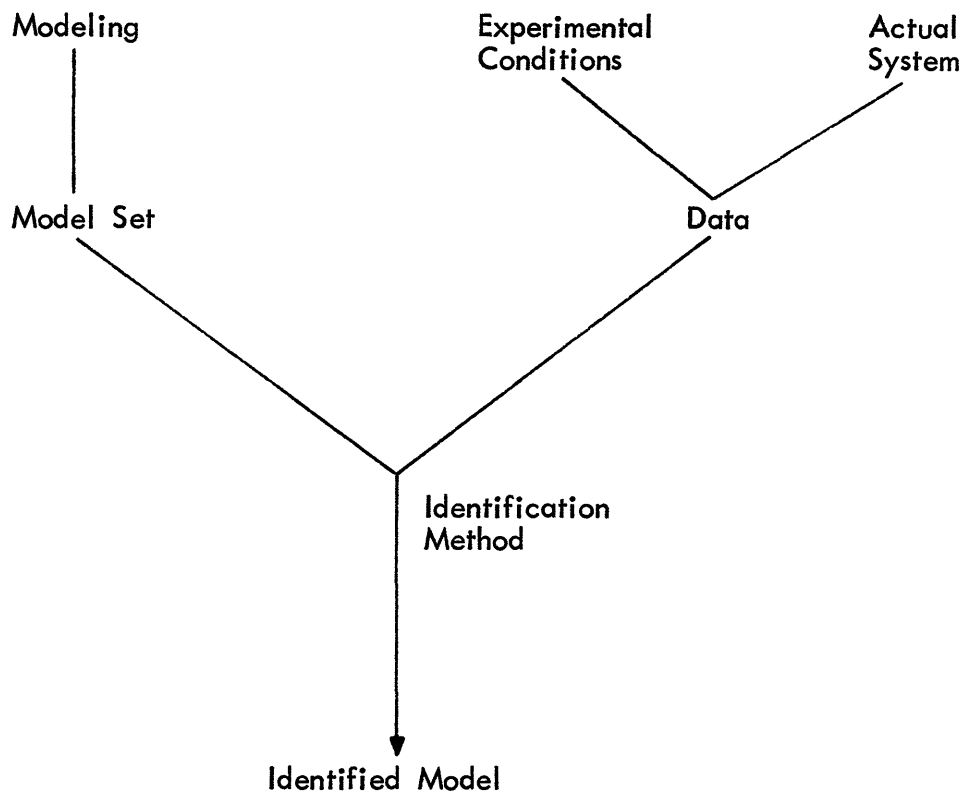


Figure 1.5

1.2 Scope and Objectives of Report

As already mentioned, this report will be limited to linear state space models and the maximum likelihood identification method. On the one

hand, this provides for a unified presentation and on the other hand it is felt to be an important case. Indeed, the maximum likelihood method is generally accepted by serious practitioners as the best method if one is not computationally constrained. It also has statistical properties which essentially say that the method is optimal for long measurement sequences. Furthermore, many control and estimation techniques are based on linear state space models. So, with the goal of summarizing the present state of maximum likelihood identification of linear state space models, and of collecting in one place and in a unified fashion results scattered in the literature, the report is based on the following outline.

Section 2 establishes the notation which will be used for linear state space models and presents a specific formulation of the parameter identification problem in that context. Then a brief review of linear state estimation theory is given in Section 3 with an emphasis on sensitivity analysis (covariance analysis) and on reduced order filtering. Our motivation for this review is the fact that Kalman filtering equations are basic to the maximum likelihood equations given in Section 5. Another reason is our desire to use the results of the state estimation problem to provide goals for the parameter estimation problem. Section 4 presents a brief review of the classical maximum likelihood theory in preparation for Section 5. Section 5 contains the development of the maximum likelihood identification method for linear state space models that is the objective of this report. We emphasize that the contents of this section have been deliberately limited in scope in order to give a compact and user-oriented presentation. Thus there are many topics in identification theory that will not be treated, but only some of the most important concepts emphasized.

SECTION 2

FORMULATION

As was mentioned above, this report will deal with linear state space models only. We assume that the reader is familiar with such models and appreciates their usefulness. Our notation is the following:

State Dynamics:

$$\underline{x}(t+1) = \underline{A}(t)\underline{x}(t) + \underline{B}(t)\underline{u}(t) + \underline{L}(t)\underline{\xi}(t) \tag{2.1}$$

Measurement Equation:

$$\underline{z}(t+1) = \underline{C}(t+1)\underline{x}(t+1) + \underline{\theta}(t+1) \tag{2.2}$$

where $t = 0,1,2,\ldots$ is the time index

$\underline{x}(t) \in R^n$ the state vector (non-white stochastic sequence)

$\underline{u}(t) \in R^m$ the deterministic input sequence

$\underline{\xi}(t) \in R^p$ the white plant noise

$\underline{\theta}(t) \in R^r$ the white measurement noise

$\underline{z}(t) \in R^r$ the measurement vector

Probabilistic Information:

The initial state $\underline{x}(0)$ is Gaussian with

$$E\{\underline{x}(0)\} = \overline{\underline{x}}(0) \tag{2.3}$$

$$\text{cov}[\underline{x}(0);\underline{x}(0)] = \underline{\Sigma}_0 = \underline{\Sigma}_0' \geq 0 \tag{2.4}$$

The plant noise $\underline{\xi}(t)$ is Gaussian discrete white noise with:

$$E\{\underline{\xi}(t)\} = \underline{0} \tag{2.5}$$

$$\text{cov}[\underline{\xi}(t), \underline{\xi}(\tau)] = \underline{\Xi}(t)\delta_{t\tau} \tag{2.6}$$

$$\underline{\Xi}(t) = \underline{\Xi}'(t) \geq \underline{0} \tag{2.7}$$

The measurement noise $\underline{\theta}(t)$ is Gaussian discrete white noise with:

$$E\{\underline{\theta}(t)\} = \underline{0} \tag{2.8}$$

$$\text{cov}[\underline{\theta}(t);\underline{\theta}(\tau)] = \underline{\Theta}(t)\ \delta_{t\tau} \tag{2.9}$$

$$\underline{\Theta}(t) = \underline{\Theta}'(t) > \underline{0} \tag{2.10}$$

(i.e., every measurement is corrupted by white noise)

$$\underline{x}(0),\ \underline{\xi}(t),\ \underline{\theta}(\tau) \tag{2.11}$$

are independent for all t,τ.

Note the use of a discrete time format which is compatible with the way data is collected using modern digital technology.

It is assumed that the system under study has been modeled in the above form, but that some parameters still need to be determined. Typically those would be coefficients in the entries of the model matrices, as illustrated by Examples 2.1 and 2.2. Note first that if we denote those unknown parameters by the vector $\underline{\alpha}$, the dependancy of the model matrices can be made explicit by the notation $\underline{A}(t;\underline{\alpha})$, $\underline{B}(t;\underline{\alpha})$, etc. Note also that the system can be time varying but that $\underline{\alpha}$ must be time invariant (at least according to the time scale of the identification experiment).

Example 2.1

If a system is of unknown structure but we assume it is time invariant and has a third order transfer function, we can write a general third order model with the following matrices

$$\underline{A}(\underline{\alpha}) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \qquad \underline{B}(\underline{\alpha}) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad \underline{C}(\underline{\alpha}) = \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix}$$

In this case, assuming that the other matrices in the model are known, the unknown parameter vector $\underline{\alpha}$ is

$$\underline{\alpha} = \begin{bmatrix} a_{31} & a_{32} & a_{33} & c_1 & c_2 & c_3 \end{bmatrix}'$$

Note that in this case $\underline{\alpha}$ does not necessarily correspond to any

physical quantities, since the model chosen is in a canonical form with no a priori structural knowledge being used. ■

Example 2.2

This example illustrates how an unknown physical parameter does not usually enter the model as a single matrix element. It also illustrates the discretization of a continuous time model in a sampled data environment. Consider the continuous time scalar model:

$$\dot{x} = -\alpha x(t) + u(t)$$

where α^{-1} is an unknown time constant. It can be discretized as follows:

$$x(t+\Delta t) = e^{-\alpha\Delta t}\, x(t) + \int_t^{t+\Delta t} e^{-\alpha(t+\Delta t-\tau)} u(\tau)\, d\tau$$

Assume $u(\tau)$ = constant, $t \leq \tau < t+\Delta t$, we then get the discrete time model

$$x(t+\Delta t) = A(\alpha)x(t) + B(\alpha)u(t)$$

where

$$A(\alpha) = e^{-\alpha\Delta t}$$

$$B(\alpha) = \int_t^{t+\Delta t} e^{-\alpha(t+\Delta t-\tau)}\, d\tau$$ ■

The problem now is to estimate the parameter vector $\underline{\alpha}$ using measured values of $\underline{u}(t)$ and $\underline{z}(t)$. This can be viewed as a nonlinear filtering problem by defining the augmented state vector

$$\begin{bmatrix} \underline{x}(t) \\ \underline{\alpha}(t) \end{bmatrix}, \text{ where } \underline{\alpha}(t) \equiv \underline{\alpha}$$

and considering the augmented system:

$$\begin{bmatrix} \underline{x}(t+1) \\ \underline{\alpha}(t+1) \end{bmatrix} = \begin{bmatrix} \underline{A}(t;\underline{\alpha}) & 0 \\ 0 & \underline{I} \end{bmatrix} \begin{bmatrix} \underline{x}(t) \\ \underline{\alpha}(t) \end{bmatrix} + \begin{bmatrix} \underline{B}(t;\underline{\alpha}) \\ \underline{0} \end{bmatrix} \underline{u}(t) + \begin{bmatrix} \underline{L}(t;\underline{\alpha}) \\ 0 \end{bmatrix} \underline{\xi}(t)$$

$$\underline{z}(t) = [\underline{C}(t;\underline{\alpha}) \quad 0] \begin{bmatrix} \underline{x}(t) \\ \underline{\alpha}(t) \end{bmatrix} + \underline{\Theta}(t)$$

The problem becomes that of determining state estimates $\hat{\underline{x}}(t|t)$ and $\hat{\underline{\alpha}}(t|t)$ of $\underline{x}(t)$ and $\underline{\alpha}$ respectively, given $\{\underline{z}(0),\underline{z}(1),\ldots,\underline{z}(t)\}$, and this is a nonlinear filtering problem since $\underline{x}(t+1)$ depends nonlinearly on $\underline{x}(t)$ and $\underline{\alpha}$. Therefore, one possible and general approach, appealing to Kalman filter designers because of its simplicity, is that of the extended Kalman filter [5]. This consists of repeatedly relinearizing the equations about current estimate values and using linear Kalman filtering to obtain new estimates. But difficulties connected with bias and divergence of estimates often arise in practice (see [18] for a recent analysis), so a more reliable approach is desirable.

Fundamentally, identification is a nonlinear estimation problem with a very special structure. Therefore algorithms that exploit that structure might be expected to yield more useful results. One such technique, the maximum likelihood identification method, exploits the structure of the identification problem very effectively. Indeed note that in the problem as defined above, if $\underline{\alpha}$ were fixed, we would have the linear state estimation problem solved by the standard Kalman filter. But as we shall see later in Section 5, the maximum likelihood method for estimating the parameters of a linear dynamic system will consist of minimizing a function of quantities which are generated by that Kalman filter. Thus we see that the Kalman filter equations play an important role in ML identification of linear dynamic systems, and so we give a brief summary of Kalman filtering theory in the next section.

SECTION 3

LINEAR ESTIMATION REVIEW

The purpose of this section is to collect needed results from linear estimation theory. These results will be needed directly in Section 5 as well as to provide goals for the parameter identification problem.

The standard estimation problem for linear state space models described by (2.1) through (2.11) is that of determining $\hat{\underline{x}}(t|t)$, the best estimate of $\underline{x}(t)$ given $\{\underline{z}(0),\underline{z}(1),\ldots,\underline{z}(t)\}$ in the sense

$$E\{[x_i(t) - \hat{x}_i(t|t]^2\} \leq E\{[x_i(t) - \tilde{x}_i(t|t)]^2\}$$

where $\tilde{x}_i(t|t)$ is any other causal estimate. Assuming knowledge of all matrices in the model, the solution to this problem is well known (e.g. [1],[5],[6]) and is given by a discrete time Kalman filter. This involves the following:

Off-line Calculations

- Initialization (t=0):

$$\underline{\Sigma}(0|0) = \text{cov}[\underline{x}(0);\underline{x}(0)] \tag{3.1}$$

- Predict Cycle:

$$\underline{\Sigma}(t+1|t) = \underline{A}(t)\underline{\Sigma}(t|t)\underline{A}'(t) + \underline{L}(t)\underline{\Xi}(t)\underline{L}'(t) \tag{3.2}$$

- Update Cycle:

$$\underline{\Sigma}(t+1|t+1) = \underline{\Sigma}(t+1|t) - \underline{\Sigma}(t+1|t)\,\underline{C}'(t+1)\,[\underline{C}(t+1)\underline{\Sigma}(t+1|t)\,\underline{C}'(t+1) + \underline{\Theta}(t+1)]^{-1}\underline{C}(t+1)\underline{\Sigma}(t+1|t) \tag{3.3}$$

- Filter Gain Matrix:

$$\underline{H}(t+1) = \underline{\Sigma}(t+1|t+1)\,\underline{C}'(t+1)\underline{\Theta}^{-1}(t+1) \tag{3.4}$$

On-line Calculations

- Initialization (t=0):

$$\hat{\underline{x}}(0|0) = E\{\underline{x}(0)\} \tag{3.5}$$

- Predict Cycle:

$$\hat{\underline{x}}(t+1|t) = \underline{A}(t)\hat{\underline{x}}(t|t) + \underline{B}(t)\underline{u}(t) \tag{3.6}$$

- Update Cycle:

$$\underbrace{\hat{\underline{x}}(t+1|t+1)}_{\text{updated estimate}} = \underbrace{\hat{\underline{x}}(t+1|t)}_{\text{predicted estimate}} + \underline{H}(t+1)\{\underbrace{\underline{z}(t+1) - \underline{C}(t+1)\hat{\underline{x}}(t+1|t)}_{\text{residual } \underline{r}(t+1)}\} \tag{3.7}$$

The structure of the system dynamics and measurements and that of the corresponding discrete time Kalman filterare shown in Figures 3.1 and 3.2 respectively. As presented above, the filter algorithm has two kinds of equations:

- Off-line equations with which the filter gain matrix and the error covariance matrix can be evaluated before the gathering of data.
- On-line equations which generate the state estimate.

Furthermore, one can distinguish within each kind of equation two different cycles which can be justified in the following heuristic manner:

- A predict cycle where knowledge of the dynamics of the system is used to obtain a predicted estimate (see equation (3.6)); and at this point uncertainty is increased due to the plant noise input and increased or decreased from propagation by the dynamics of the system (see equation (3.2)).
- An update cycle where the residual between the latest observation and the corresponding predicted observation is used to obtain an updated estimate (see equation (3.7)); and uncertainty is decreased due to the processing of the new observation (see equation (3.3)). Note that the predicted observation is obtained from the predicted state estimate by

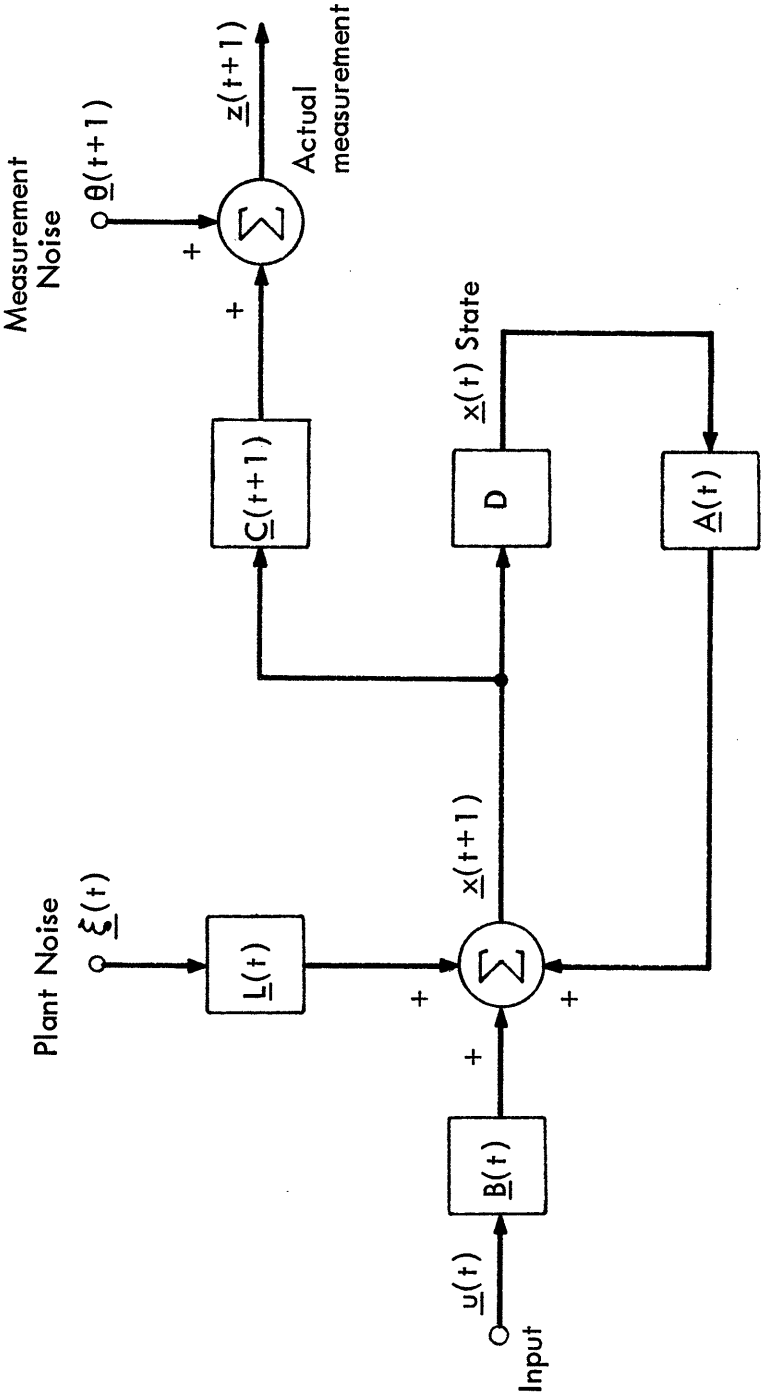


Figure 3.1 Structure of System Dynamics and Measurements

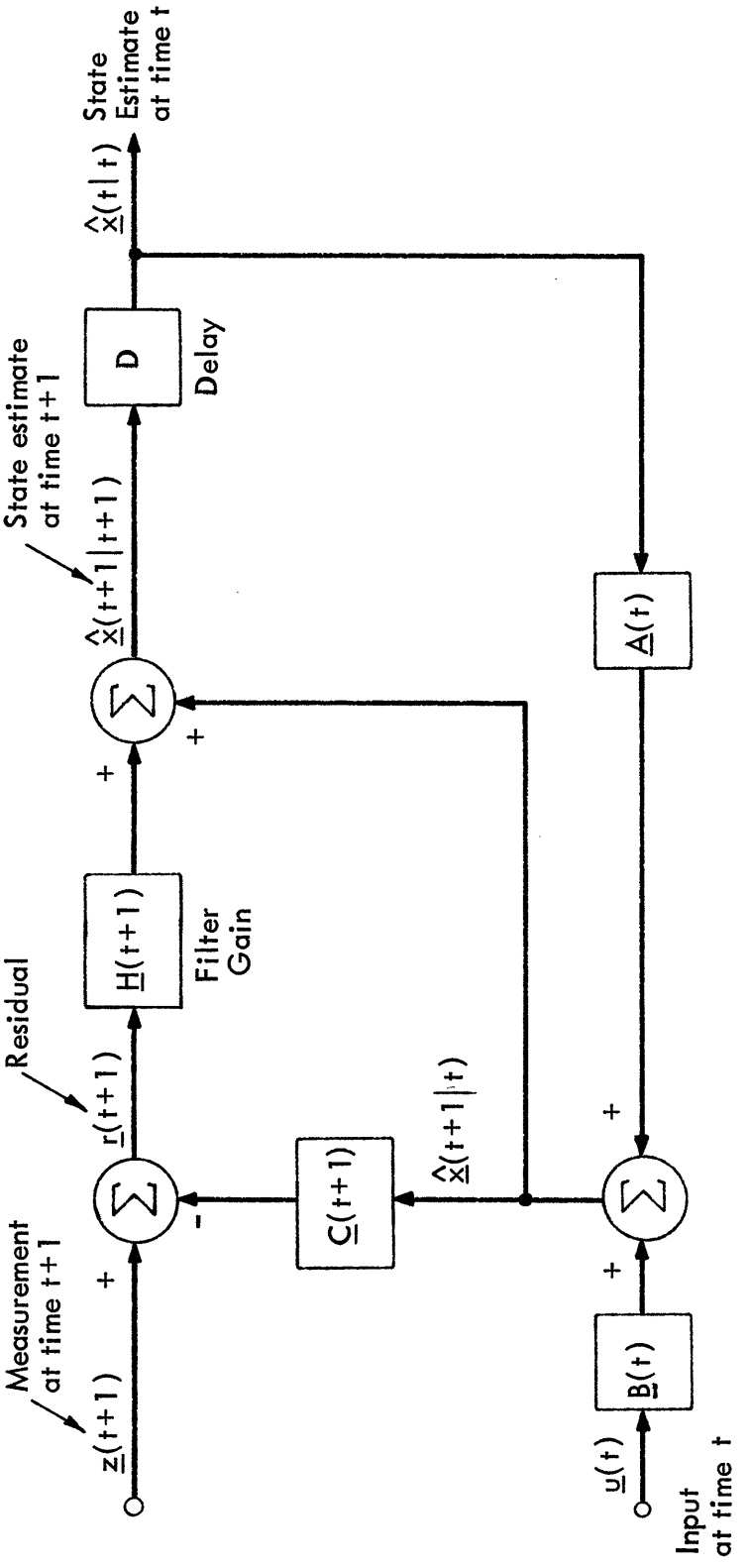


Figure 3.2 Structure of Discrete-Time Kalman Filter

$$\hat{\underline{z}}(t|t-1) = \underline{C}(t)\hat{\underline{x}}(t|t-1) \tag{3.8}$$

and that its error covariance matrix is

$$\underline{S}(t) \equiv E\{[\underline{z}(t) - \hat{\underline{z}}(t|t-1)]\,[z(t) - \hat{z}(t|t-1)]'\}$$

$$= \underline{C}(t)\underline{\Sigma}(t|t-1)\,\underline{C}'(t) + \underline{\Theta}(t) \tag{3.9}$$

From (3.1) through (3.9) one can also get the following form which will be useful in the sequel

$$\hat{\underline{x}}(t+1|t) = \underline{A}(t)\hat{\underline{x}}(t|t-1) + \underline{B}(t)\underline{u}(t) +$$

$$+ \underline{A}(t)\underline{H}(t)\,[\underline{z}(t) - \underline{C}(t)\hat{\underline{x}}(t|t-1)] \tag{3.10}$$

$$\underline{H}(t) = \underline{\Sigma}(t|t-1)\,\underline{C}'(t)\,[\underline{C}(t)\underline{\Sigma}(t|t-1)\,\underline{C}'(t) + \underline{\Theta}(t)]^{-1}$$

$$= \underline{\Sigma}(t|t-1)\,\underline{C}'(t)\underline{S}^{-1}(t) \tag{3.11}$$

$$\underline{\Sigma}(t+1|t) = \underline{A}(t)\underline{\Sigma}(t|t-1)\,\underline{A}'(t) + \underline{L}(t)\underline{\Xi}(t)\underline{L}(t) -$$

$$- \underline{A}(t)\underline{H}(t)\underline{S}(t)\underline{H}'(t)\underline{A}'(t) \tag{3.12}$$

Finally, let $\underline{z}^{t-1}$ denote the set of past measurements $\{\underline{z}(0),\underline{z}(1),\ldots,\underline{z}(t-1)\}$. In Section 5 we will need the conditional probability density $p(\underline{z}(t)\,|\,\underline{z}^{t-1})$ of measurement $\underline{z}(t)$ given $\underline{z}^{t-1}$. But it also follows from the above solution to the linear estimation problem that, under the Gaussian assumption, the residuals

$$\underline{r}(t) \equiv \underline{z}(t) - \hat{\underline{z}}(t|t-1)$$

form an independent Gaussian sequence and that

$$p(\underline{z}(t)\,|\,\underline{z}^{t-1}) = (2\pi)^{-\frac{r}{2}}\,(\det[\underline{S}(t)]^{-\frac{1}{2}}e^{-\frac{1}{2}\underline{r}'(t)\underline{S}^{-1}(t)\underline{r}(t)} \tag{3.13}$$

The above equations are valid for the general case of a time-varying linear system. In the special time invariant case, the system matrices will be constant, but the Kalman filter will still in general have a time varying gain $\underline{H}(t)$. However, under reasonable assumptions (see, e.g., [1]), it can be shown that the Riccati equation (3.12) has a limiting solution as $t\to\infty$, which in turn implies that the Kalman filter gain becomes con-

stant. A common procedure is to use this steady state gain instead of the optimal time-varying gain. The resulting filter is termed the steady state Kalman filter. Clearly, this procedure makes sense when the time horizon of the filtering problem is large relative to the effective convergence time of the Riccati equation (3.12).

The procedure of using the steady-state Kalman filter in place of the time-varying filter has two important practical ramifications. First, equation (3.12) can be replaced by the algebraic Riccati equation

$$\underline{\Sigma} = \underline{A}\underline{\Sigma}\underline{A}' + \underline{L}\underline{\Xi}\underline{L}' - \underline{A}\underline{\Sigma}\underline{C}'(\underline{C}\underline{\Sigma}\underline{C}' + \underline{\Theta})^{-1}\underline{C}\underline{\Sigma}\underline{A}' \ . \tag{3.14}$$

Equation (3.14) can be solved directly by methods more efficient than the iteration of (3.12) to steady state. Second, since the filter gain is constant,

$$\underline{H} = \underline{\Sigma}\underline{C}'(\underline{C}\underline{\Sigma}\underline{C}' + \underline{\Theta})^{-1} \tag{3.15}$$

implementation of the filter is greatly simplified.

Before concluding this section, note that an important assumption underlying the theory presented above is that the actual system and the model used for the filter are identical. But in practice, due to neglected dynamic effects, uncertain parameter values, etc., this is never the case. Moreover, a reduced order filter is often deliberately employed to reduce on-line computational requirements or sensitivity to poorly known parameters. For these reasons, it is necessary to be able to evaluate the performance of a mismatched Kalman filter, i.e., a Kalman filter based on a model that differs from the actual system.

Suppose therefore that the true system is as previously specified (but assuming for simplicity $\underline{u}(t) \equiv 0$) and that the filter used is:

$$\hat{\underline{x}}(t+1|t) = \underline{A}^r(t)\hat{\underline{x}}(t|t-1) + \underline{A}^r(t)\underline{H}^r(t)\{\underline{z}(t) - \underline{C}^r(t)\hat{\underline{x}}(t|t-1)\} \tag{3.16}$$

where $\hat{\underline{x}}$ is of dimension $n^r \leq n$. Suppose also that $\hat{\underline{x}}$ is an estimate of $\underline{W}\underline{x}$

where $\underline{W}$ is some $\check{n}^r \times n$ matrix.

Example 3.1

If $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ and filter omits x_3 then $\underline{W} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$. ■

Then, one can compute $\underline{\Sigma}(t|t-1) \equiv E\{[\underline{W}\underline{x}(t) - \hat{\underline{x}}(t|t-1)][\underline{W}\underline{x}(t) - \hat{\underline{x}}(t|t-1)]'\}$ as follows. Define

$$\overline{\underline{A}^r}(t) \equiv \underline{A}^r(t)(\underline{I} - \underline{H}^r(t)\underline{C}^r(t))$$

$$\underline{A}_{aug}(t) \equiv \left[\begin{array}{c|c} \underline{A}(t) & \underline{0} \\ \hline \underline{A}^r(t)\underline{H}^r(t)\underline{C}(t) & \overline{\underline{A}^r}(t) \end{array}\right]$$

$$\underline{L}_{aug}(t) \equiv \left[\begin{array}{c|c} \underline{L}(t) & \underline{0} \\ \hline \underline{0} & \underline{A}^r(t)\underline{H}^r(t) \end{array}\right]$$

$$\underline{\Xi}_{aug}(t) \equiv \left[\begin{array}{c|c} \underline{\Xi}(t) & \underline{0} \\ \hline \underline{0} & \underline{\Theta}(t) \end{array}\right]$$

$$\underline{x}_{aug}(t) \equiv \left[\begin{array}{c} \underline{x}(t) \\ \hline \hat{\underline{x}}(t|t-1) \end{array}\right]$$

Then, from (2.1) and (3.14)

$$\underline{x}_{aug}(t+1) = \underline{A}_{aug}(t)\underline{x}_{aug}(t) + \underline{L}_{aug}(t)\begin{bmatrix} \underline{\xi}(t) \\ \underline{\theta}(t) \end{bmatrix} \tag{3.17}$$

If we now let $\underline{\Gamma}(t)$ be the $(n+n^r)\times(n+n^r)$ solution of the covariance equation for the above augmented system, we get:

$$\underline{\Gamma}(t+1) = \underline{A}_{aug}(t)\underline{\Gamma}(t)\underline{A}'_{aug}(t) + \underline{L}_{aug}(t)\underline{\Xi}_{aug}(t)\underline{L}'_{aug}(t) \tag{3.18}$$

and

$$\underline{\Sigma}(t|t-1) = [\underline{W} \mid -\underline{I}]\,\underline{\Gamma}(t)\,[\underline{W} \mid -\underline{I}]' \tag{3.19}$$

This covariance analysis is routinely used to study the loss in accuracy when the filter order is reduced for computational savings. The fact that this analysis can be carried out off-line makes it an important tool in reduced order filter design.

The preceeding equations of Kalman filtering theory (particularly (3.13)) will prove basic to the maximum likelihood identification problem considered in Section 5. Moreover, there are some properties of Kalman filtering theory that are important for setting goals for maximum likelihood identification theory. First, note that the Kalman filtering equations provide an optimal data processing algorithm in the sense that no other set of equations can give state estimates with less mean square error. Second, note that as well as providing state estimates, a quantitative measure $\underline{\Sigma}(t|t)$ of the accuracy of those estimates is obtained. Third, note that $\underline{\Sigma}(t|t)$ can be evaluated off-line before any measurements are made, so that system performance with various alternative hardware components and operating conditions can be evaluated even before the system is built. Fourth, the filtering equations are general purpose, applying equally well to navigation systems or power systems or to any system that can be modeled by (2.1) - (2.11). Fifth, the sensitivity of filter performance to modeling errors, either inadvertant or deliberate, can be readily evaluated. These properties of Kalman filtering theory provide desirable goals for any theory of system identification. We will see that the maximum likelihood identification theory does attain these goals, but only in an asymptotic sense for very long sequences of measurements.

SECTION 4

CLASSICAL MAXIMUM LIKELIHOOD THEORY

The purpose of this section is to briefly review classical maximum likelihood theory. By this we mean those issues of the maximum likelihood method which were treated in the statistics literature before this method was introduced as a tool for system identification. The problem is that of identifying the unknown values $\underline{\alpha}$ parametrizing the probability density[1] $p(\underline{z}^t;\underline{\alpha})$ of all past observations $\{\underline{z}(0),\underline{z}(1),\ldots,\underline{z}(t)\}$. Since $\underline{\alpha}$ is not a random variable this is not, strictly speaking, a conditional density but rather a family of density functions, one for each value of $\underline{\alpha}$. So, for a fixed set of past observations $\underline{z}^t$, $p(\underline{z}^t;\underline{\alpha})$ can be viewed as a function of $\underline{\alpha}$, called the likelihood function. The maximum likelihood estimate $\hat{\underline{\alpha}}$ of $\underline{\alpha}$ is then defined to be the maximum of this function, i.e. the value of $\underline{\alpha}$ that is most likely to have caused the particular set of observations $\underline{z}^t$ (see Figure 4.1). It is remarkable that this simple idea leads to an

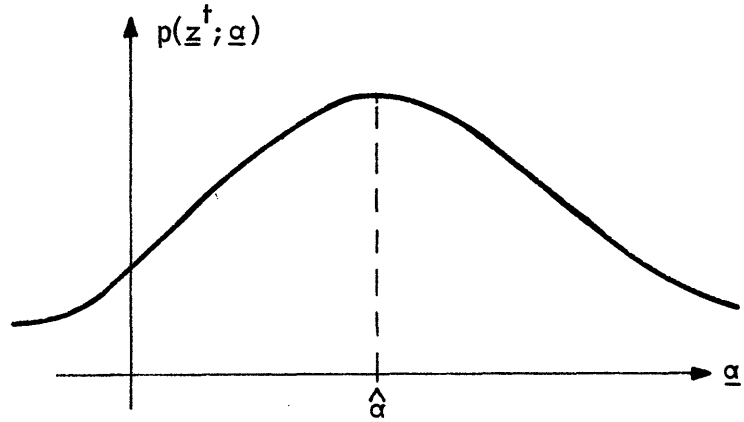


Figure 4.1

[1] Note that such a probability density is induced by the model (2.1)-(2.11).

estimate with a number of desirable properties.

To discuss properties of the maximum likelihood method, it is necessary to introduce several general concepts. To this end, consider some arbitrary estimate $\underline{\hat{\alpha}}(\underline{z}^t)$ of $\underline{\alpha}$ given the observations $\underline{z}^t$. Define the bias $\underline{b}(\underline{\alpha})$ in the estimator $\underline{\hat{\alpha}}(\cdot)$ by the equation

$$\begin{aligned}\underline{b}(\underline{\alpha}) &= \underline{\alpha} - E\{\underline{\hat{\alpha}}(\underline{z}^t) \,|\, \underline{\alpha}\} \\ &= \underline{\alpha} - \int \underline{\hat{\alpha}}(\underline{z}^t) p(\underline{z}^t;\underline{\alpha})\, d\underline{z}t \end{aligned} \tag{4.1}$$

and the error covariance matrix $\underline{\Sigma}(\underline{\alpha})$ of the estimator $\underline{\hat{\alpha}}(\cdot)$ by

$$\begin{aligned}\underline{\Sigma}(\underline{\alpha}) &= E\{(\underline{\alpha} - \underline{\hat{\alpha}}(\underline{z}^t) - \underline{b}(\underline{\alpha}))(\underline{\alpha} - \underline{\hat{\alpha}}(\underline{z}^t) - \underline{b}(\underline{\alpha}))' \,|\, \underline{\alpha}\} \\ &= \int (\underline{\alpha} - \underline{\hat{\alpha}}(\underline{z}^t) - \underline{b}(\underline{\alpha}))(\underline{\alpha} - \underline{\hat{\alpha}}(\underline{z}^t) - \underline{b}(\underline{\alpha}))' p(\underline{z}^t;\underline{\alpha})\, d\underline{z}^t \,. \end{aligned} \tag{4.2}$$

Clearly, desirable properties of an estimator are that it be unbiased ($\underline{b}(\underline{\alpha}) = 0$) and that it have minimum error covariance matrix, i.e., that the diagonal elements of $\underline{\Sigma}(\underline{\alpha})$ should be as small as possible. (Note that the ith diagonal element of $\underline{\Sigma}(\underline{\alpha})$ is the mean square error in the estimate of α_i.)

In general, it is very difficult to compute either $\underline{b}(\underline{\alpha})$ or $\underline{\Sigma}(\underline{\alpha})$. However, for any unbiased estimator we have the following Cramer-Rao lower bound [7].

$$\underline{\Sigma}(\underline{\alpha}) \geq \underline{I}_{\underline{z}^t}^{-1}(\underline{\alpha}) \tag{4.3}$$

Here $\underline{I}_{\underline{z}^t}(\underline{\alpha})$ is the Fisher information matrix defined by

$$\underline{I}_{\underline{z}^t}(\underline{\alpha}) = -E\left\{\frac{\partial^2}{\partial \underline{\alpha}^2} \ln p(\underline{z}^t;\underline{\alpha}) \,|\, \underline{\alpha}\right\} \tag{4.4}$$

or equivalently

$$\underline{I}_{\underline{z}^t}(\underline{\alpha}) = E\left\{\left[\frac{\partial}{\partial \underline{\alpha}} \ln p(\underline{z}^t;\underline{\alpha})\right]\left[\frac{\partial}{\partial \underline{\alpha}} \ln p(\underline{z}^t;\underline{\alpha})\right]' \,|\, \underline{\alpha}\right\} . \tag{4.5}$$

There are several points that need to be emphasized concerning the Cramer-Rao Lower Bound. First, note that (4.3) is equivalent to

$$\underline{\Sigma}(\underline{\alpha}) - \underline{I}_{\underline{z}t}^{-1}(\underline{\alpha}) \geq 0$$

so that in particular every diagonal element of $\underline{\Sigma}(\underline{\alpha})$ must be no smaller than the corresponding element of $\underline{I}_{\underline{z}t}^{-1}(\underline{\alpha})$. Thus the Cramer-Rao lower bound provides a lower bound on the accuracy to which any component of $\underline{\alpha}$ can be estimated. A second point to be made is that the technical assumptions required in the derivation of the lower bound do not include any assumptions of linearity or Gaussianess. Thus the bound is well-suited for nonlinear problems such as the parameter identification problem for dynamical systems. Finally, note the dependence of $\underline{b}(\underline{\alpha})$, $\underline{\Sigma}(\underline{\alpha})$, and $\underline{I}_{\underline{z}t}(\underline{\alpha})$ on $\underline{\alpha}$. It is generally true that the performance of an estimator in a nonlinear estimation problem is dependent on the quantity being estimated.

The preceeding discussion is illustrated in the following example which involves a simple version of the dynamic problems to be treated later.

Example 4.1

Consider the scalar model

$$\begin{cases} x(t+1) = ax(t) & , \ t=0 \\ z(t) = x(t) + \theta(t) & , \ t=0,1 \end{cases}$$

where $\theta(0)$ and $\theta(1)$ are independent random variables with probability density

$$p(\theta(t)) = \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}\theta^2(t)}$$

and where a and x(0) are unknown. In terms of the notation previously defined

$$\underline{\alpha} = \begin{bmatrix} x(0) \\ a \end{bmatrix}$$

and $\underline{z}^t$ can be conveniently arranged as

$$\underline{z}^t = \begin{bmatrix} z(0) \\ z(1) \end{bmatrix} = \begin{bmatrix} x(0) + \theta(0) \\ ax(0) + \theta(1) \end{bmatrix}.$$

Then,

$$p(\underline{z}^t;\underline{\alpha}) = \frac{1}{2\pi} e^{-\frac{1}{2}[(z(0)-x(0))^2 + (z(1)-ax(0))^2]}$$

or

$$\ln p(\underline{z}^t;\underline{\alpha}) = -\ln(2\pi) - \tfrac{1}{2}[(z(0) - x(0))^2 + (z(1) - ax(0))^2]$$

and

$$\frac{\partial}{\partial \underline{\alpha}} \ln p(\underline{z}^t;\underline{\alpha}) = \begin{bmatrix} z(0) - x(0) + a[z(1) - ax(0)] \\ x(0)[z(1) - ax(0)] \end{bmatrix}$$

$$= \begin{bmatrix} \theta(0) + a\theta(1) \\ x(0)\theta(1) \end{bmatrix}$$

Now the Fisher information matrix can be evaluated, using (4.5)

$$\underline{I}_{\underline{z}^t}(\underline{\alpha}) = E\left\{ \begin{bmatrix} \theta(0) + a\theta(1) \\ x(0)\theta(1) \end{bmatrix} \left[\theta(0) + a\theta(1) \,\vdots\, x(0)\theta(1)\right] \middle| \underline{\alpha} \right\}$$

$$= \begin{bmatrix} 1+a^2 & ax(0) \\ ax(0) & x^2(0) \end{bmatrix}$$

and

$$\underline{I}^{-1}_{\underline{z}^t}(\underline{\alpha}) = \begin{bmatrix} 1 & -\frac{a}{x(0)} \\ -\frac{a}{x(0)} & \frac{1}{x^2(0)} + \frac{a^2}{x^2(0)} \end{bmatrix}$$

From the Cramer-Rao lower bound, any unbiased estimator will satisfy:

$$E\{(\underline{\alpha}-\hat{\underline{\alpha}})(\underline{\alpha}-\hat{\underline{\alpha}})' \,|\, \underline{\alpha}\} \geq \underline{I}^{-1}_{\underline{z}^t}(\underline{\alpha})$$

implying

$$E\{(x(0) - \hat{x}(0))^2\} \geq 1$$

and

$$E\{(a - \hat{a})^2\} \geq \frac{1}{x^2(0)} + \frac{a^2}{x^2(0)} .$$

As mentioned above, the bound depends on $\underline{\alpha}$ which is unknown. In particular note the dependence on $x(0)$. If $x(0) = 0$ the information matrix is singular; which is consistent with the fact that in this case, $z(1)$ contains no information about a anymore. On the other hand, the larger $|x(0)|$ (i.e. the more the system is originally excited), the better it can be identified.

Finally, the maximum likelihood estimate of $\underline{\alpha}$ is the one which maximizes $\ln p(\underline{z}^t;\underline{\alpha})$ above:

$$\hat{x}(0) = z(0) \quad ; \quad \hat{a} = \frac{z(1)}{\hat{x}(0)}$$

and one can verify that

$$E\{(x(0) - \hat{x}(0))^2\} = 1$$

$$E\{(a - \hat{a})^2\} = E\left\{\left[a - \frac{ax(0) + \theta(1)}{x(0) + \theta(0)}\right]^2\right\} \approx \frac{1}{x^2(0)} + \frac{a^2}{x^2(0)}$$

since $$\frac{ax(0) + \theta(1)}{x(0) + \theta(0)} \approx a + \frac{\theta(1)}{x(0)} + \frac{a}{x(0)}\ \theta(0)$$

for $\theta(0)$ small relative to $x(0)$. Thus we see that the Cramer-Rao lower bound becomes tight as the signal-to-noise ratio increases. ■

We now return to the statement of the classical properties of the maximum likelihood estimate. First consider an arbitrary estimator, and suppose that this estimator is unbiased and efficient, i.e., it satisfies the Cramer-Rao lower bound with equality. It can be shown [7], that if any such estimator exists, it is necessarily a maximum likelihood estimator. Since an unbiased, efficient estimator is clearly optimal in a mean square estimation error sense with respect to the class of unbiased estimators, this property does provide some motivation for using maximum likelihood estimates. Note however that there is no guarantee that an unbiased, efficient estimate will

exist.

Perhaps more interesting are the asymptotic properties of the maximum likelihood estimate. In the classical theory, statisticians usually assume independent observations[1] so that

$$p(\underline{z}^t;\underline{\alpha}) = \prod_{i=0}^{t} p(\underline{z}(i);\underline{\alpha}) \ . \tag{4.6}$$

The idea is that the observations are presumed to be the result of a sequence of independent experiments. The asymptotic properties of the maximum likelihood estimate are concerned with the limiting behavior as the number of observations becomes infinite.

A crucial assumption for the asymptotic properties given below is the identifiability condition

$$p(\underline{z}(t);\underline{\alpha}_1) \not\equiv p(\underline{z}(t);\underline{\alpha}_2) \quad \text{for all } \underline{\alpha}_1 \neq \underline{\alpha}_2 \ . \tag{4.7}$$

This assumption simply means that no two parameters lead to observations with identical probabilistic behavior. Clearly, if the identifiability condition is violated for some pair $\underline{\alpha}_1, \underline{\alpha}_2$ of parameters, than $\underline{\alpha}_1$ and $\underline{\alpha}_2$ cannot be distinguished no matter how many observations are made. (Compare with example 5.3, page 54.)

Now let $\hat{\underline{\alpha}}_t$ denote the maximum likelihood estimate of $\underline{\alpha}$ given $\underline{z}^t$. Under the above conditions of independent observations, identifiability, and additional technical assumptions (see [7]) we have the following results.

Consistency

$\hat{\underline{\alpha}}_t \to \underline{\alpha}$ with probability 1 as $t \to \infty$.

---

[1] This is not the case for $\underline{z}(t)$ generated by (2.1) - (2.11).

Asymptotic Unbiasedness

$$E\{\hat{\underline{\alpha}}_t | \underline{\alpha}\} \to \underline{\alpha} \text{ as } t \to \infty.$$

Asymptotic Normality

$\hat{\underline{\alpha}}_t$ tends towards a Gaussian random variable as $t \to \infty$.

Asymptotic Efficiency

$$E\{(\underline{\alpha} - \hat{\underline{\alpha}}_t)(\underline{\alpha} - \hat{\underline{\alpha}}_t)' | \underline{\alpha}\} \to \underline{I}_{\underline{z}^t}^{-1}(\underline{\alpha}) \text{ as } t \to \infty.$$

In other words, as the number of processed observations becomes infinite, the maximum likelihood estimate $\hat{\underline{\alpha}}_t$ converges to the true value of $\underline{\alpha}$, and the parameter estimate error $\hat{\underline{\alpha}}_t - \underline{\alpha}$ is asymptotically normally distributed with covariance matrix $\underline{I}_{\underline{z}^t}^{-1}(\underline{\alpha})$ so that the Cramer-Rao lower bound is asymptotically tight.

Notice that the independence assumption (4.6) implies an additive form for the information matrix $\underline{I}_{\underline{z}^t}(\underline{\alpha})$. Specifically, we have

$$\begin{aligned}
\underline{I}_{\underline{z}^t}(\underline{\alpha}) &= -E\left\{\frac{\partial^2}{\partial \underline{\alpha}^2} \ln p(\underline{z}^t;\underline{\alpha})\right\} \\
&= -E\left\{\frac{\partial^2}{\partial \underline{\alpha}^2} \ln \prod_{i=0}^{t} p(\underline{z}(i);\underline{\alpha})\right\} \\
&= \sum_{i=0}^{t} -E\left\{\frac{\partial^2}{\partial \underline{\alpha}^2} \ln p(\underline{z}(i);\underline{\alpha})\right\} \\
&= (t+1)\underline{I}_{\underline{z}}(\underline{\alpha}) \qquad (4.8)
\end{aligned}$$

where $\underline{I}_{\underline{z}}(\underline{\alpha})$ is the information matrix for a single observation. In terms of the asymptotic covariance matrix, we see that

$$E\{(\hat{\underline{\alpha}}_t - \underline{\alpha})(\hat{\underline{\alpha}}_t - \underline{\alpha})'\} \approx \frac{1}{t} \underline{I}_{\underline{z}}^{-1}(\underline{\alpha}) \qquad (4.9)$$

for large t.

Equation (4.9) is extremely important from an applications point of view. By asymptotic unbiasedness, we know that $\hat{\underline{\alpha}}_t$ has expected value $\underline{\alpha}$

for large t. From (4.9), we can compute the standard deviation of the error in the estimate of each component of α_i. Since $\hat{\underline{\alpha}}_t$ is known to be asymptotically Gaussian, we can even compute confidence intervals, i.e., intervals about the estimates $(\hat{\underline{\alpha}}_t)_i$ in which $\underline{\alpha}_i$ is known to lie with a specified probability. Moreover, if we turn the problem around and consider the issue of experimental design, we can choose the number of observations t so that any desired level of asymptotic estimation accuracy is achieved.[1] Note that the characteristic $\frac{1}{\sqrt{t}}$ behavior of the estimate error standard deviation implies that a ten-fold increase in accuracy requires a hundred-fold increase in the number of observations.

To conclude this section, it is useful to compare the results provided by maximum likelihood estimation theory for the nonlinear parameter estimation problem with those provided by Kalman filtering theory for the linear state estimation problem. First, recall that the Kalman filtering equations provide optimal estimates. The maximum likelihood estimates are also optimal, in a slightly different sense, but only asymptotically in general. Second, the Kalman filter provides a state covariance matrix that provides an indication of estimate accuracy. In the maximum likelihood theory, the inverse information matrix plays this role, but in general provides a lower bound which is only asymptotically tight. Third, recall that the Kalman filter error covariance matrix is precomputable so that the performance of various alternative hardware configurations can be evaluated before components are procured and measurements are made. The information matrix $\underline{I}_{\underline{z}t}(\underline{\alpha})$ is likewise precomputable without measurements, but it depends on the unknown parameter $\underline{\alpha}$. Thus one has to assume some plausible

[1] Note that $\underline{I}_{\underline{z}}(\underline{\alpha})$ depends on $\underline{\alpha}$, so the number of observations must actually be computed for some plausible range of values of $\underline{\alpha}$.

SECTION 5

MAXIMUM LIKELIHOOD ESTIMATION OF LINEAR DYNAMIC SYSTEMS

In this section we finally take up the main topic of this report, the maximum likelihood identification method for determining the parameters of a linear-Gaussian state space model of a dynamic system. This class of models was described in Section 2, and the general maximum likelihood method was described in Section 4. We will see that the Kalman filtering theory reviewed in Section 3 plays a key role in the subsequent development.

The basic difficulty in applying the maximum likelihood method to the identification problem is that the observation process is not independent, i.e.,

$$p(\underline{z}^t;\underline{\alpha}) \neq p(\underline{z}(0);\underline{\alpha}) \cdots p(\underline{z}(t);\underline{\alpha}) \ .$$

This leads to practical difficulties in the computation of the likelihood function and the information matrix since $p(\underline{z}^t;\underline{\alpha})$ is a density defined over a high dimensional space and is cumbersome to deal with. Moreover, recall from the previous section that the asymptotic results concerning the maximum likelihood method were all predicated on the assumption of independent observations, so that there are theoretical difficulties as well.

The key idea in the extension of the maximum likelihood method to the identification problem is to write the more general factorization

$$p(\underline{z}^t;\underline{\alpha}) = p(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha}) \cdots p(\underline{z}(1)\,|\,\underline{z}(0);\underline{\alpha})\ p(\underline{z}(0);\underline{\alpha})$$

and to recall that (in the linear-Gaussian case) $p(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha})$ is characterized completely by quantities computed by the Kalman filter corresponding to $\underline{\alpha}$. We will see in the sequel that this simple idea will lead to methods for computing and maximizing the likelihood function and for computing the information matrix, as well as an extension of all the classical asymptotic properties of the maximum likelihood estimate. We will even be able to

examine the behavior of the maximum likelihood estimate under modeling errors (deliberate or inadvertent), an issue that does not seem to arise in the classical case.

5.1 Computation of the Likelihood Function

The problem is to evaluate the maximum likelihood estimate $\hat{\underline{\alpha}}_t$ of $\underline{\alpha}$ which, by definition, maximizes $p(\underline{z}^t;\underline{\alpha})$. As mentioned above, this probability density function factors as follows:

$$p(\underline{z}^t;\underline{\alpha}) = p(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha}) \ \ldots\ p(\underline{z}(0);\underline{\alpha})$$

But recall from Section 3 (equation (3.13)) that the conditional probability density $p(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha})$ of the current observations $\underline{z}(\tau)$ given past observations $\underline{z}^{\tau-1}$ is (in the case of linear-Gaussian state space models):

$$p(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha}) = (2\pi)^{-r/2}(\det[\underline{S}(\tau;\underline{\alpha})])^{-\frac{1}{2}}\, e^{-\frac{1}{2}\underline{r}'(\tau;\underline{\alpha})\underline{S}^{-1}(\tau;\underline{\alpha})\underline{r}(\tau;\underline{\alpha})} \tag{5.1}$$

Since $\underline{r}(\tau;\underline{\alpha}) = \underline{z}(\tau) - \hat{\underline{z}}(\tau\,|\,\tau\text{-}1;\underline{\alpha})$, where $\hat{\underline{z}}(\tau\,|\,\tau\text{-}1;\underline{\alpha})$ and $\underline{S}(\tau;\underline{\alpha})$ are generated by the Kalman filter corresponding to $\underline{\alpha}$, the likelihood function is readily computable for every $\underline{\alpha}$ and set of data $\underline{z}^t$.

Now to simplify the manipulation of the above quantities, it is customary to equivalently maximize $\ln p(\underline{z}^t;\underline{\alpha})$ instead of $p(\underline{z}^t;\underline{\alpha})$ itself. This has the advantage of transforming the products into sums, which will be useful when we will need to compute derivatives. It also replaces the exponential term in (5.1) by a more amenable quadratic term. Indeed,

$$\ln p(\underline{z}^t;\underline{\alpha}) = \sum_{\tau=0}^{t} \ln p(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha})$$

where $p(\underline{z}(0)\,|\,\underline{z}^{-1};\underline{\alpha}) \equiv p(\underline{z}(0);\underline{\alpha})$; and

$$\ln p(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha}) = -\frac{r}{2}\ln(2\pi) - \tfrac{1}{2}\ln(\det[\underline{S}(\tau;\underline{\alpha})]) - \tfrac{1}{2}\underline{r}'(\tau;\underline{\alpha})\underline{S}^{-1}(\tau;\underline{\alpha})\underline{r}(\tau;\underline{\alpha}) \tag{5.2}$$

Furthermore, note that $-\frac{r}{2}\ln(2\pi)$ is a constant term independent of $\underline{\alpha}$.

range of values of $\underline{\alpha}$ for pre-experimental studies. Fourth, the Kalman filtering concept is general in the sense that it is not restricted to any one application area. The maximum likelihood concept is even more general as it applies to nonlinear as well as linear parameter estimation problems. Fifth, the Kalman filtering theory permits the sensitivity of filter performance to modeling errors to be readily assessed. However, such a sensitivity theory does not appear to have been developed in the classical maximum likelihood theory.

Viewed as a possible concept for application to the parameter identification problem of linear dynamic systems, maximum likelihood estimation theory can be seen from the discussion of this section to offer great potential. However, some questions remain. Can maximum likelihood estimates and related quantities such as the information matrix be readily computed for the parameters of linear dynamic systems? Are the asymptotic properties that motivate the use of the maximum likelihood estimate valid if the independence assumption (4.6) is relaxed? Can we determine the asymptotic behavior of the maximum likelihood estimate under modeling errors? We will see in the next section that the answer to all these questions is in the affirmative.

Therefore the maximum likelihood estimate $\hat{\underline{\alpha}}_t$ can be more simply evaluated by minimizing the "interesting part" of the negative log likelihood function:

$$\zeta(\underline{z}^t;\underline{\alpha}) \equiv -\left[\ln p(\underline{z}^t;\underline{\alpha}) + \ln(2\pi)^{\frac{(t+1)r}{2}}\right]$$

$$= \sum_{\tau=0}^{t} \zeta(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha}) \tag{5.3}$$

where

$$\zeta(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha}) \equiv \tfrac{1}{2}\ln(\det[\underline{S}(\tau;\underline{\alpha})]) + \tfrac{1}{2}\,\underline{r}'(\tau;\underline{\alpha})\underline{S}^{-1}(\tau;\underline{\alpha})\underline{r}(\tau;\underline{\alpha}) \tag{5.4}$$

Note that the new log likelihood function $\zeta(\underline{z}^t;\underline{\alpha})$ has two parts: a deterministic part which depends only on $\underline{S}(\tau;\underline{\alpha})$, $\tau=0,\ldots,t$ and which is therefore precomputable (see Section 3); and a quadratic in the residuals part, which therefore depends on the data. The steps involved in computing $\zeta(\underline{z}^t;\underline{\alpha})$ are summarized in Figure 5.1.

The above equations are valid for the general case of a time varying system. Note that each evaluation of the likelihood function requires the processing of the observations by a time-varying Kalman filter. A very common practice in the case of time invariant systems is to instead use the steady state Kalman filter. This introduces an approximation into the computation of the likelihood function, but this approximation will be good if the optimal time-varying Kalman filter reaches steady state in a time that is short relative to the time interval of the observations. Of course, use of the steady state Kalman filter greatly simplifies the calculation of the likelihood function. As we will see in the next two sections, there is also a great simplification in the computation of the gradient of the likelihood function and of the information matrix.

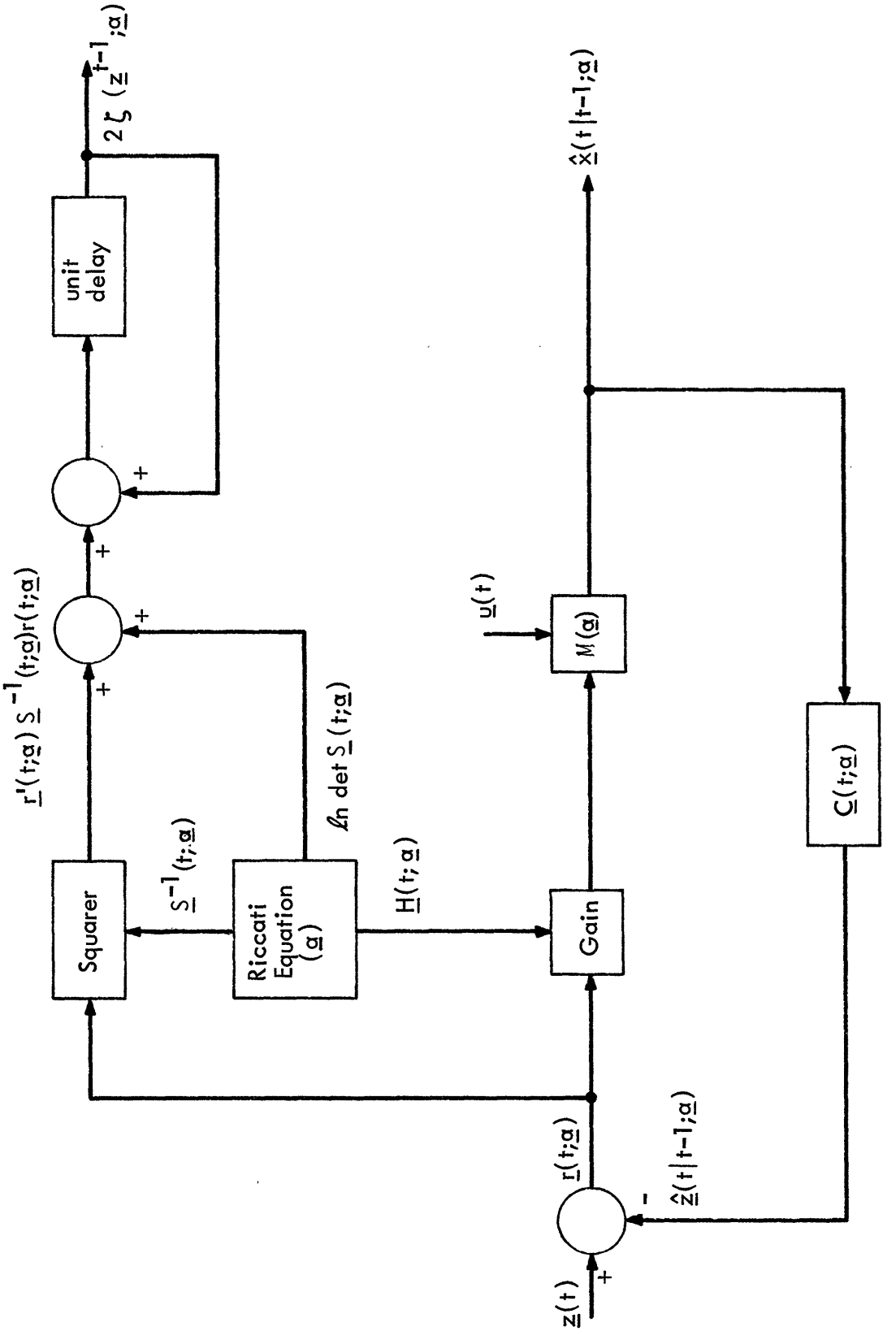


Figure 5.1 Computation of $\zeta(\underline{z}^t;\underline{\alpha})$

5.2 Maximization of the Likelihood Function

As was mentioned above, maximizing the likelihood function is equivalent to minimizing $\zeta(\underline{z}^t;\underline{\alpha})$ defined in (5.3). The cases where the parameter space in which $\underline{\alpha}$ lies is discrete or continuous will be now discussed separately.

5.2.1 Case 1. Discrete parameter space: $\underline{\alpha} \in \{\underline{\alpha}_1, \underline{\alpha}_2, \ldots, \underline{\alpha}_N\}$

In this case one can run N parallel Kalman filters to generate $\zeta(\underline{z}^t;\underline{\alpha}_k)$, $k=1,\ldots,N$. The choice of $\hat{\underline{\alpha}}_t$ is then trivial.

As an aside, we remark that if a priori probabilities $\Pr\{\underline{\alpha} = \underline{\alpha}_k\}$ are available, then one can also get:

$$P_k(t) \equiv \Pr\{\underline{\alpha} = \underline{\alpha}_k \mid \underline{z}^t\} \tag{5.5}$$

$$= \frac{\Pr\{\underline{\alpha} = \underline{\alpha}_k\}\ e^{-\zeta(\underline{z}^t;\underline{\alpha}_k)}}{\sum_{\ell=1}^{N} \Pr\{\underline{\alpha} = \underline{\alpha}_\ell\}\ e^{-\zeta(\underline{z}^t;\underline{\alpha}_\ell)}} \tag{5.6}$$

These a posteriori probabilities can then be used to weight the corresponding estimates $\hat{\underline{x}}(t|t-1;\underline{\alpha}_k)$ or corresponding controls $\underline{u}_k(t)$, and therefore generate an on-line adaptive estimate or control law. This is illustrated in Figures 5.2 and 5.3 and is an example of the "multiple model" techniques which can be used in many adaptive estimation and control applications (see e.g. [8]). Notice the parallel structure of the computations which can be exploited in advanced digital controller architectures.

5.2.2 Case 2. Continuous parameter space: $\underline{\alpha} \in R^{\ell}$

In this case, a numerical optimization technique is required. These generally require $\frac{\partial \zeta(\underline{z}^t;\underline{\alpha})}{\partial \underline{\alpha}}$ and sometimes $\frac{\partial^2 \zeta(\underline{z}^t;\underline{\alpha})}{\partial \underline{\alpha}^2}$, respectively the gradient and

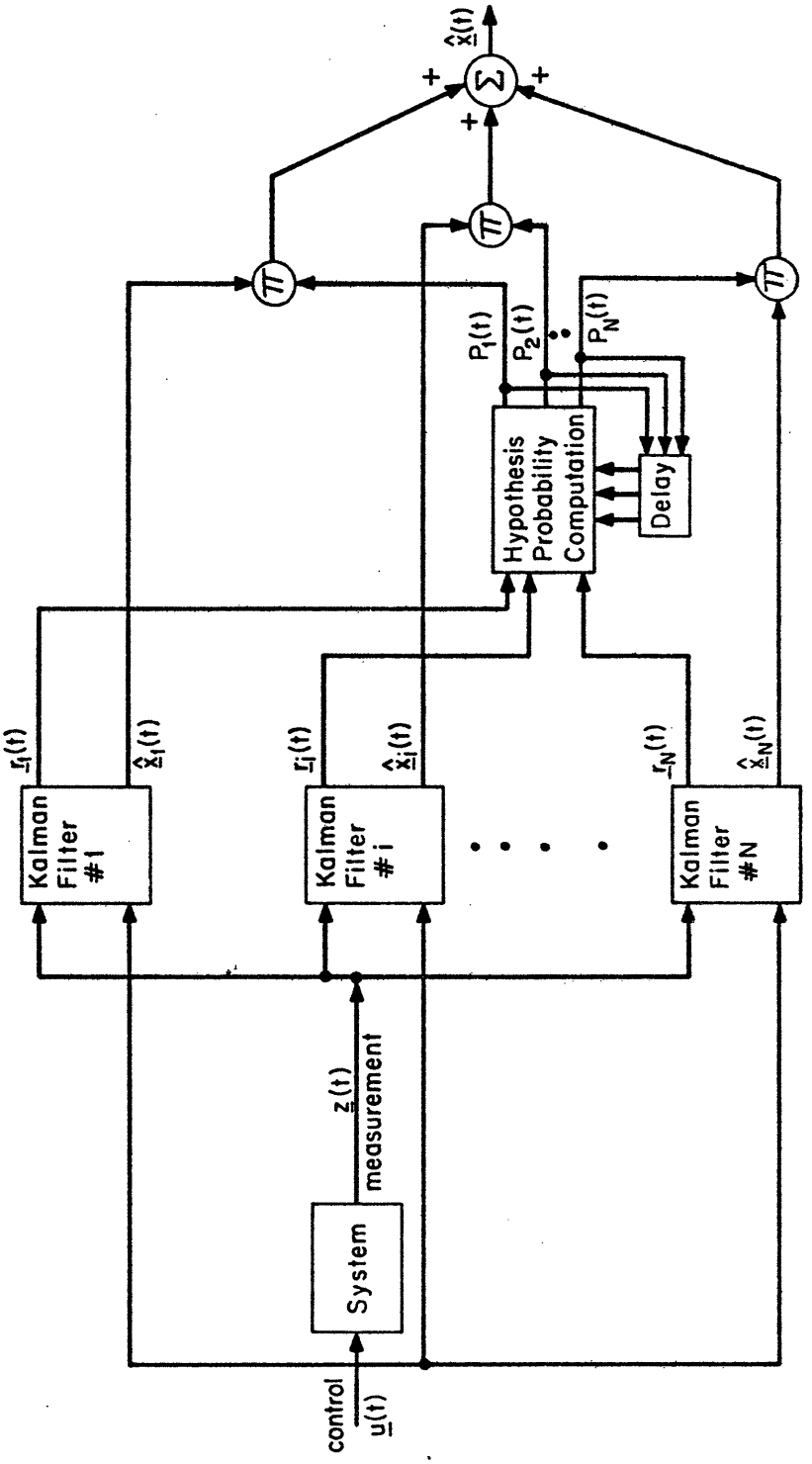


Figure 5.2: Multiple Model Identification and Estimation

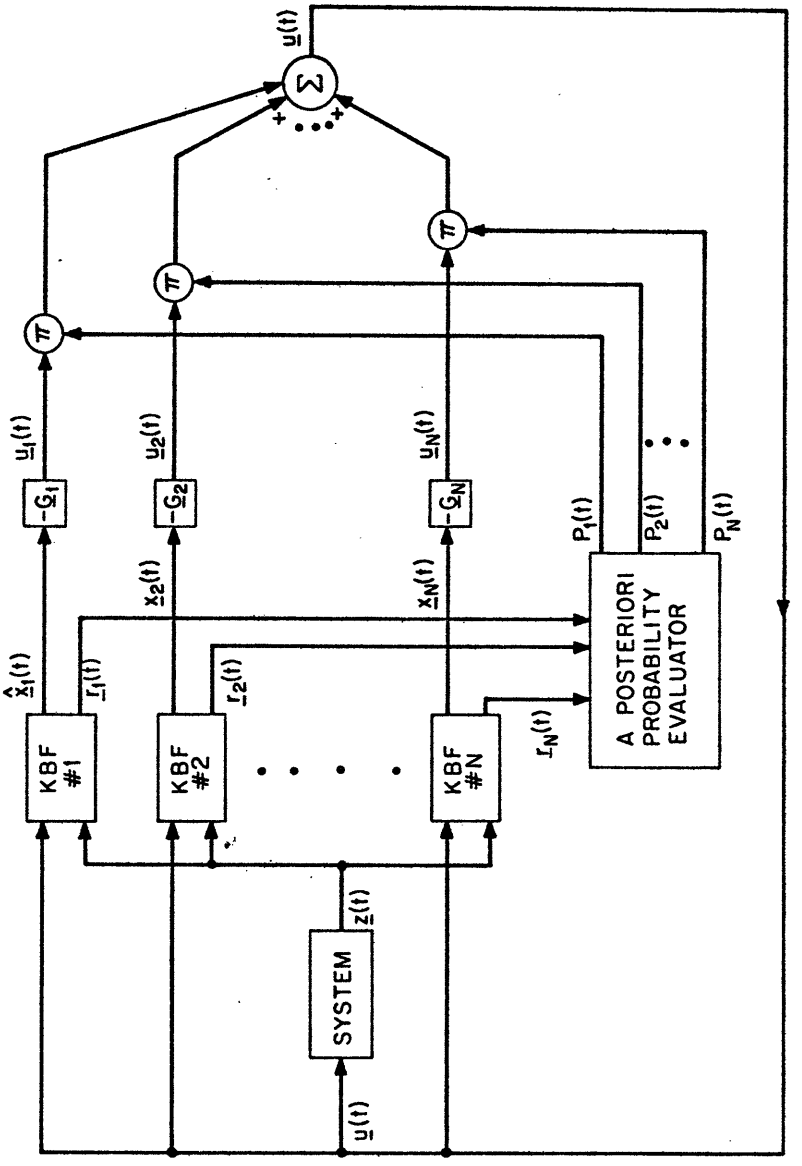


Figure 5.3: Multiple Model Adaptive Control

the Hessian of $\zeta(\underline{z}^t;\underline{\alpha})$. Thus our problem is to derive expressions for these quantities.

(i) Gradient Evaluation

Forward Filter:

We proceed by straightforward differentiation (see, e.g., [9] for a similar treatment). From (5.3)

$$\frac{\partial \zeta(\underline{z}^t;\underline{\alpha})}{\partial \alpha_i} = \sum_{\tau=0}^{t} \frac{\partial \zeta(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha})}{\partial \alpha_i} \tag{5.7}$$

where α_i denotes the ith component of $\underline{\alpha}$. From (5.4)

$$\begin{aligned}\frac{\partial \zeta(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha})}{\partial \alpha_i} &= \underline{r}'(\tau;\underline{\alpha})\underline{S}^{-1}(\tau;\underline{\alpha})\,\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i} \\ &\quad - \tfrac{1}{2}\underline{r}'(\tau;\underline{\alpha})\underline{S}^{-1}(\tau;\underline{\alpha})\,\frac{\partial \underline{S}(\tau;\underline{\alpha})}{\partial \alpha_i}\,\underline{S}^{-1}(\tau;\underline{\alpha})\underline{r}(\tau;\underline{\alpha}) \\ &\quad + \tfrac{1}{2}\,\mathrm{tr}\left[\underline{S}^{-1}(\tau;\underline{\alpha})\,\frac{\partial \underline{S}(\tau;\underline{\alpha})}{\partial \alpha_i}\right]\end{aligned} \tag{5.8}$$

so that one needs to evaluate $\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}$ and $\frac{\partial \underline{S}(\tau;\underline{\alpha})}{\partial \alpha_i}$. The former is obtained by differentiating (3.8),

$$\frac{\partial \underline{r}(t;\underline{\alpha})}{\partial \alpha_i} = -\frac{\partial \underline{C}(t;\underline{\alpha})}{\partial \alpha_i}\,\hat{\underline{x}}(t\,|\,t-1;\underline{\alpha}) - \underline{C}(t;\underline{\alpha})\,\frac{\partial \hat{\underline{x}}(t\,|\,t-1;\underline{\alpha})}{\partial \alpha_i} \tag{5.9}$$

From (3.10), we can obtain the filter sensitivity equations

$$\frac{\partial \hat{\underline{x}}(t+1\,|\,t;\underline{\alpha})}{\partial \alpha_i} = \bar{\underline{A}}(t;\underline{\alpha})\,\frac{\partial \hat{\underline{x}}(t\,|\,t-1;\underline{\alpha})}{\partial \alpha_i} + \underline{\omega}_i(t;\underline{\alpha}) \tag{5.10}$$

where

$$\begin{aligned}\underline{\omega}_i(t;\underline{\alpha}) &\equiv \frac{\partial \bar{\underline{A}}(t;\underline{\alpha})}{\partial \alpha_i}\,\hat{\underline{x}}(t\,|\,t-1;\underline{\alpha}) + \frac{\partial \underline{B}(t;\underline{\alpha})}{\partial \alpha_i}\,\underline{u}(t) \\ &\quad + \left\{\frac{\partial}{\partial \alpha_i}\,[\underline{A}(t;\underline{\alpha})\underline{H}(t;\underline{\alpha})]\right\}\underline{z}(t)\end{aligned} \tag{5.11}$$

$$\bar{\underline{A}}(t;\underline{\alpha}) \equiv \underline{A}(t;\underline{\alpha})\,[\underline{I} - \underline{H}(t;\underline{\alpha})\underline{C}(t;\underline{\alpha})] \tag{5.12}$$

and where, from (3.11),

$$\begin{aligned}\frac{\partial \underline{H}(t;\underline{\alpha})}{\partial \alpha_i} &= \left\{\frac{\partial}{\partial \alpha_i}\,[\underline{\Sigma}(t\,|\,t-1;\underline{\alpha})\underline{C}'(t;\underline{\alpha})]\right\}\underline{S}^{-1}(t;\underline{\alpha}) \\ &\quad - \underline{\Sigma}(t\,|\,t-1;\underline{\alpha})\underline{C}'(t;\underline{\alpha})\underline{S}^{-1}(t;\underline{\alpha})\,\frac{\partial \underline{S}(t;\underline{\alpha})}{\partial \alpha_i}\,\underline{S}^{-1}(t;\underline{\alpha})\end{aligned} \tag{5.13}$$

From (3.9), we have

$$\frac{\partial \underline{S}(t;\underline{\alpha})}{\partial \alpha_i} = \underline{C}(t;\underline{\alpha}) \frac{\partial \underline{\Sigma}(t|\,t-1;\underline{\alpha})}{\partial \alpha_i} \underline{C}'(t;\underline{\alpha}) + \frac{\partial \underline{\Theta}(t;\underline{\alpha})}{\partial \alpha_i}$$

$$+ \frac{\partial \underline{C}(t;\underline{\alpha})}{\partial \alpha_i} \underline{\Sigma}(t|t-1;\underline{\alpha}) \underline{C}'(t;\underline{\alpha}) + \underline{C}(t;\underline{\alpha}) \underline{\Sigma}(t|\,t-1;\underline{\alpha}) \frac{\partial \underline{C}'(t;\underline{\alpha})}{\partial \alpha_i} \tag{5.14}$$

and, from (3.12), we obtain the Riccati sensitivity equations

$$\frac{\partial \underline{\Sigma}(t+1|\,t;\underline{\alpha})}{\partial \alpha_i} = \bar{\underline{A}}(t;\underline{\alpha}) \frac{\partial \underline{\Sigma}(t|\,t-1;\underline{\alpha})}{\partial \alpha_i} \bar{\underline{A}}'(t;\underline{\alpha}) + \underline{\Omega}_i(t;\underline{\alpha}) + \underline{\Omega}_i'(t;\underline{\alpha}) \tag{5.15}$$

where

$$\underline{\Omega}_i(t;\underline{\alpha}) \equiv \frac{\partial \underline{A}}{\partial \alpha_i} \underline{\Sigma}(t|\,t-1;\underline{\alpha}) \bar{\underline{A}}' - \underline{A}\underline{H} \frac{\partial \underline{C}}{\partial \alpha_i} \underline{\Sigma}(t|\,t-1;\underline{\alpha}) \bar{\underline{A}}'$$

$$+ \tfrac{1}{2} \frac{\partial}{\partial \alpha_i} [\underline{L}\underline{\Xi}\underline{L}'] + \tfrac{1}{2} \underline{A}\underline{H} \frac{\partial \underline{\epsilon}}{\partial \alpha_i} \underline{H}' \underline{A}' \tag{5.16}$$

The derivation of (5.15) and (5.16) may not be as clear as that of (5.10) and (5.11) and is therefore given, in more detail, in Appendix A.

Figure 5.4 summarizes the steps described above in the evaluation of $\frac{\partial \zeta(\underline{z}(t)\,|\,\underline{z}^t;\underline{\alpha})}{\partial \alpha_i}$. Since the recursive equations (5.10) and (5.15) run forward in time, this approach can be referred to as the forward filter evaluation of $\frac{\partial \zeta}{\partial \underline{\alpha}}$. Recalling that $\underline{\alpha}$ is ℓ-dimensional, this evaluation then requires:

1 Riccati equation (3.12) or n^2 equations

1 filter equation (3.10) or n equations

ℓ Riccati sensitivity equations (5.15) or $n^2\ell$ equations

ℓ filter sensitivity equations (5.10) or $n\ell$ equations

or roughly the equivalent of $(\ell+1)$ Kalman filters. While this is quite expensive computationally, it can be carried out in a straightforward fashion.

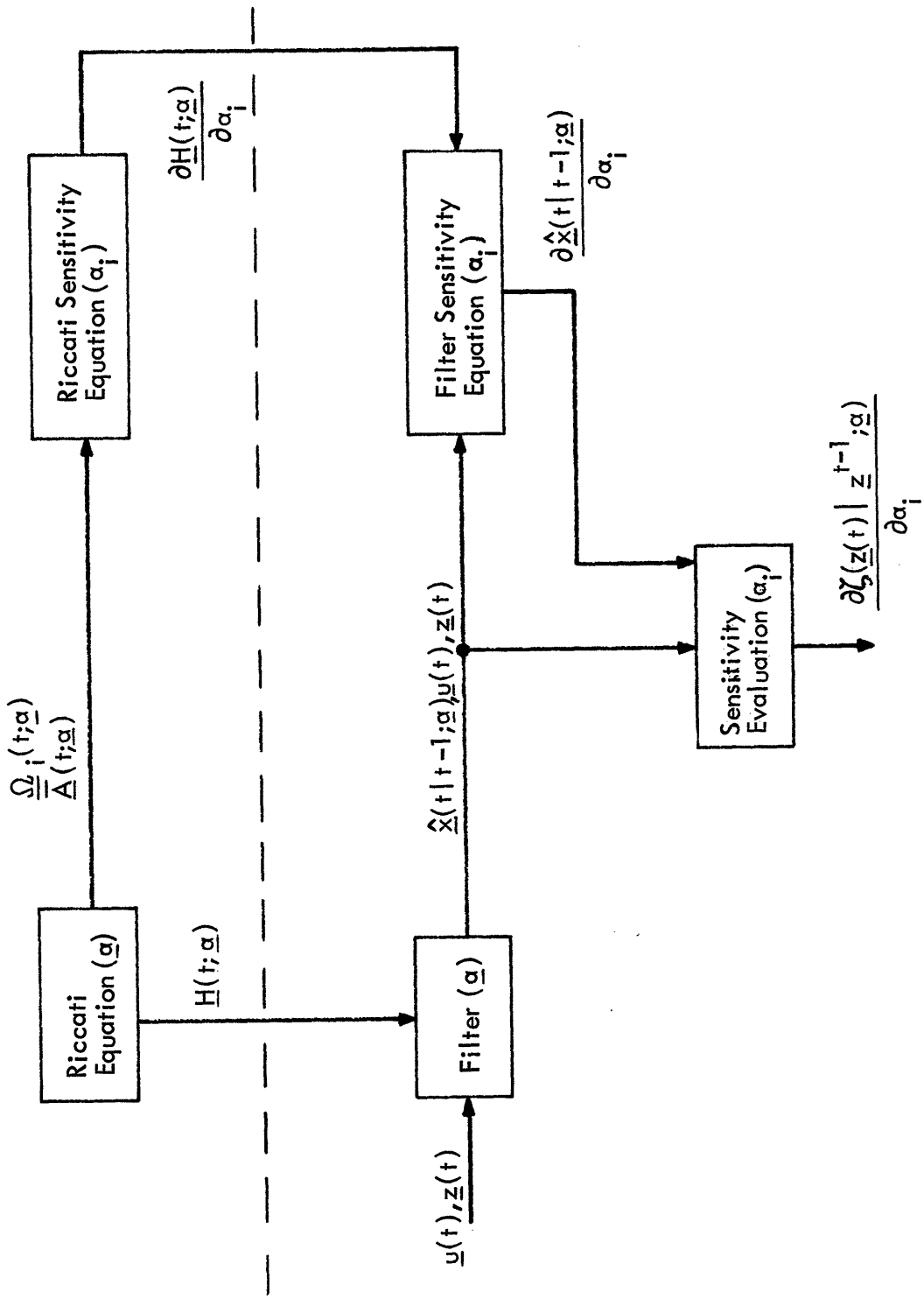


Figure 5.4: Forward Filter Evaluation of $\frac{\partial \zeta}{\partial \underline{\alpha}}$

For the case of a time invariant system and observations processed by a steady state Kalman filter, the equations above simplify. In this case, $\underline{S}(\underline{\alpha})$ and $\underline{H}(\underline{\alpha})$ are independent of t, and the time varying matrix equation (5.15) reduces to a steady state version

$$\frac{\partial \underline{\Sigma}}{\partial \alpha_i} = \bar{\underline{A}} \frac{\partial \underline{\Sigma}}{\partial \alpha_i} \bar{\underline{A}}' + \underline{\Omega}_i + \underline{\Omega}_i' \tag{5.17}$$

This equation is termed an algebraic Lyapunov equation, and an efficient method for its direct solution, called the Bartels-Stewart method, is available [17]. It is noteworthy that a major part of the computations in this algorithm involve only operations on $\bar{\underline{A}}$ which do not have to be repeated when the equation is resolved for the different $\underline{\Omega}_i$.

There is also an alternate backward filter approach, with the possibility of reduced computation, to the problem of evaluating the gradient of the likelihood function. Reexpress the forward filter equations in the following form:

$$\frac{\partial \zeta(\underline{z}^t;\underline{\alpha})}{\partial \alpha_i} = \sum_{\tau=0}^{t} \beta_i(\underline{z}^\tau;\underline{\alpha}) + \sum_{\tau=0}^{t} \underline{\gamma}'(\underline{z}^\tau;\underline{\alpha}) \frac{\partial \hat{\underline{x}}(\tau|\,\tau-1;\underline{\alpha})}{\partial \alpha_i} + \sum_{\tau=0}^{t} \operatorname{tr}\left[\underline{\Gamma}(\underline{z}^\tau;\underline{\alpha}) \frac{\partial \underline{\Sigma}(\tau|\,\tau-1;\underline{\alpha})}{\partial \alpha_i}\right] \tag{5.18}$$

$$\frac{\partial \hat{\underline{x}}(t+1|\,t;\underline{\alpha})}{\partial \alpha_i} = \bar{\underline{A}}(t;\underline{\alpha}) \frac{\partial \hat{\underline{x}}(t|\,t-1;\underline{\alpha})}{\partial \alpha_i} - \bar{\underline{A}}(t;\underline{\alpha}) \frac{\partial \underline{\Sigma}(t|\,t-1;\underline{\alpha})}{\partial \alpha_i} \underline{\gamma}(\underline{z}^t;\underline{\alpha}) + \bar{\underline{\omega}}_i(t;\underline{\alpha}) \tag{5.19}$$

$$\frac{\partial \underline{\Sigma}(t+1|\,t;\underline{\alpha})}{\partial \alpha_i} = \bar{\underline{A}}(t;\underline{\alpha}) \frac{\partial \underline{\Sigma}(t|\,t-1;\underline{\alpha})}{\partial \alpha_i} \bar{\underline{A}}'(t;\underline{\alpha}) + \underline{\Omega}_i(t;\underline{\alpha}) + \underline{\Omega}_i'(t;\underline{\alpha}) \tag{5.15}$$

where

$$\beta_i(\underline{z}^t;\underline{\alpha}) \equiv \tfrac{1}{2} \operatorname{tr}\left\{\underline{S}^{-1}(t;\underline{\alpha})\left[\underline{I}-\underline{r}(t;\underline{\alpha})\underline{r}'(t;\underline{\alpha})\underline{S}^{-1}(t;\underline{\alpha})\right] \times \left[2\underline{C}(t;\underline{\alpha})\underline{\Sigma}(t|\,t-1;\underline{\alpha}) \frac{\partial \underline{C}'(t;\underline{\alpha})}{\partial \alpha_i} + \frac{\partial \underline{\Theta}(t;\underline{\alpha})}{\partial \alpha_i}\right]\right\} - \underline{r}'(t;\underline{\alpha})\underline{S}^{-1}(t;\underline{\alpha}) \frac{\partial \underline{C}(t;\underline{\alpha})}{\partial \alpha_i} \hat{\underline{x}}(t|\,t-1;\underline{\alpha}) \tag{5.20}$$

$$\underline{\gamma}(\underline{z}^t;\underline{\alpha}) \equiv -\underline{C}'(t;\underline{\alpha})\underline{S}^{-1}(t;\underline{\alpha})\underline{r}(t;\underline{\alpha}) \tag{5.21}$$

$$\underline{\Gamma}(\underline{z}^t;\underline{\alpha}) \equiv \tfrac{1}{2}\underline{C}'(t;\underline{\alpha})\underline{S}^{-1}(t;\underline{\alpha})[\underline{I} - \underline{r}(t;\underline{\alpha})\underline{r}'(t;\underline{\alpha})\underline{S}^{-1}(t;\underline{\alpha})]\underline{C}(t;\underline{\alpha}) \tag{5.22}$$

and where (dropping the t and $\underline{\alpha}$ arguments from matrix notations for clarify)

$$\begin{aligned}\bar{\underline{\omega}}_i(t;\underline{\alpha}) = &\left[\frac{\partial \underline{A}}{\partial \alpha_i} - \underline{A}\underline{H}\frac{\partial \underline{C}}{\partial \alpha_i}\right]\hat{\underline{x}}(t|t-1;\underline{\alpha}) \\ &+\left\{\frac{\partial \underline{A}}{\partial \alpha_i}\underline{H} + \underline{A}\underline{\Sigma}(t|t-1;\underline{\alpha})\left[\frac{\partial \underline{C}'}{\partial \alpha_i} - \underline{C}'\underline{S}^{-1}\frac{\partial \underline{\Theta}}{\partial \alpha_i}\right.\right. \\ &\left.\left. - \underline{C}'\underline{S}^{-1}\frac{\partial \underline{C}}{\partial \alpha_i}\underline{\Sigma}(t|t-1;\underline{\alpha})\underline{C}' - \underline{C}'\underline{S}^{-1}\underline{C}\underline{\Sigma}(t|t-1;\underline{\alpha})\frac{\partial \underline{C}'}{\partial \alpha_i}\right]\underline{S}^{-1}\right\}\underline{r}(t;\underline{\alpha}) \\ &+\frac{\partial \underline{B}}{\partial \alpha_i}\underline{u}(t)\end{aligned} \tag{5.23}$$

(This form is derived in Appendix B.) The theory of adjoint equations, briefly summarized in Appendix C, suggests the possibility of replacing the $n\ell + n^2\ell$ forward filter equations (5.19) and (5.15) with $n + n^2$ adjoint equations running backward in time, and using the adjoint variables to evaluate the second and third terms of (5.18). Indeed, by direct application of Corollary 3 of Appendix C we have the following

Backward Filter

$$\begin{aligned}\frac{\partial \zeta(\underline{z}^t;\underline{\alpha})}{\partial \alpha_i} = &\sum_{\tau=0}^{t}\beta_i(\underline{z}^\tau;\underline{\alpha}) \\ &+\underline{\lambda}'(0;\underline{\alpha})\frac{\partial \hat{\underline{x}}(0;\underline{\alpha})}{\partial \alpha_i} + \sum_{\tau=1}^{t}\underline{\lambda}'(\tau;\underline{\alpha})\bar{\underline{\omega}}_i(\tau-1;\underline{\alpha}) \\ &+\operatorname{tr}\left[\underline{\Lambda}(0;\underline{\alpha})\frac{\partial \underline{\Sigma}(0;\underline{\alpha})}{\partial \alpha_i}\right] + \sum_{\tau=1}^{t}\operatorname{tr}\left[\underline{\Lambda}(t;\underline{\alpha})\left(\underline{\Omega}_i(\tau-1;\underline{\alpha}) + \underline{\Omega}_i'(\tau-1;\underline{\alpha})\right)\right]\end{aligned} \tag{5.24}$$

$$\underline{\lambda}(\tau;\underline{\alpha}) = \bar{\underline{A}}'(\tau;\underline{\alpha})\underline{\lambda}(\tau+1;\underline{\alpha}) + \underline{\gamma}(\underline{z}^\tau;\underline{\alpha})$$

$$\underline{\lambda}(t;\underline{\alpha}) = \underline{\gamma}(\underline{z}^t;\underline{\alpha}) \tag{5.25}$$

$$\underline{\Lambda}(\tau;\underline{\alpha}) = \bar{\underline{A}}'(\tau;\underline{\alpha})\underline{\Lambda}(\tau+1;\underline{\alpha})\bar{\underline{A}}(\tau;\underline{\alpha}) - \bar{\underline{A}}'(\tau;\underline{\alpha})\underline{\lambda}(\tau+1;\underline{\alpha})\underline{\gamma}'(\underline{z}^\tau;\underline{\alpha}) + \underline{\Gamma}(\underline{z}^\tau;\underline{\alpha})$$

$$\underline{\Lambda}(t;\underline{\alpha}) = \underline{\Gamma}(\underline{z}^t;\underline{\alpha}) \tag{5.26}$$

These equations run backward in time and can be referred to as the backward filter evaluation of $\frac{\partial \zeta}{\partial \alpha_i}$ (summarized in Figure 5.5). Furthermore, in com-

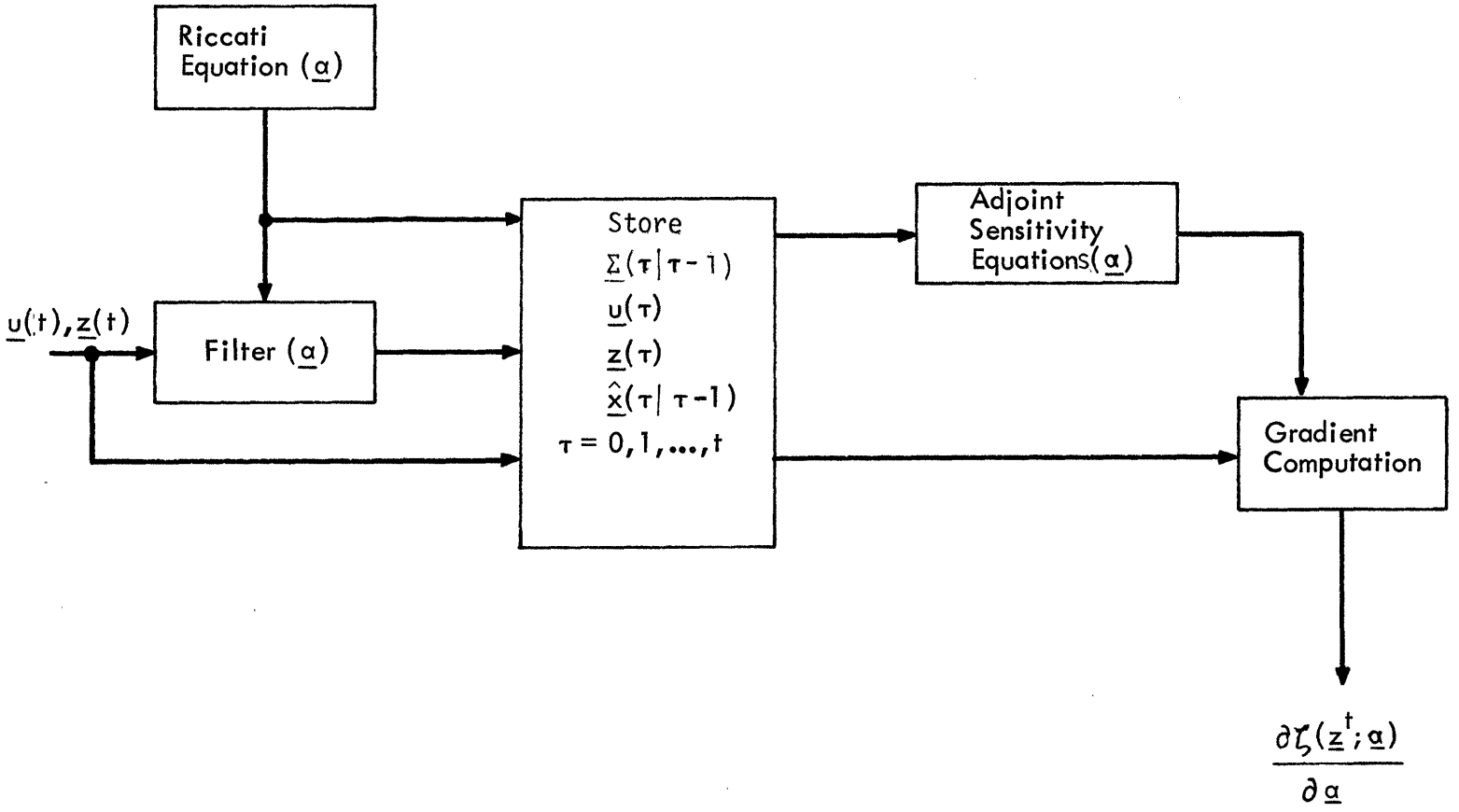


Figure 5.5: Backward Filter Evaluation of $\frac{\partial \zeta}{\partial \underline{\alpha}}$

parison with the forward filter approach, the backward filter approach requires:

1 Riccati equation (3.12) or n^2 equations

1 filter equation (3.10) or n equations

1 adjoint Riccati sensitivity equation (5.26) or n^2 equations

1 adjoint filter sensitivity equation (5.25) or n equations

or roughly the equivalent of 2 (1 forward and 1 backward) Kalman filters. This apparently represents less computation but the storage burden is now increased. Therefore, the relative advantages of either the forward or backward filter approaches depend heavily on the particular application.

As a final comment, we note that $\underline{\lambda}(t)$ and $\underline{\Lambda}(t)$ in the adjoint equations have interpretations as Lagrange multipliers or costates of an optimal control problem (see e.g. [10]).

(ii) Hessian Evaluation

The evaluation of the Hessian of $\zeta(\underline{z}^t;\underline{\alpha})$ proceeds, in principle, as for the gradient. However it requires roughly the equivalent of ℓ^2 Kalman filters which is a very heavy computational constraint and so is not attempted in practice. An alternate approach is to use the information matrix

$$\underline{I}_{\underline{z}^t}(\underline{\alpha}) = E\left\{\frac{\partial^2 \zeta(\underline{z}^t;\underline{\alpha})}{\partial \underline{\alpha}^2}\middle|\underline{\alpha}\right\}$$

instead, or, as is usually done, an approximation thereof (see e.g. [9] and Section 5.3 below).

(iii) Numerical Minimization

At this point one can use a gradient algorithm of the form:

$$\hat{\underline{\alpha}}^{k+1} = \hat{\underline{\alpha}}^k - \sigma^k \underline{W}^k \frac{\partial \zeta(\underline{z}^t;\hat{\underline{\alpha}}^k)}{\partial \underline{\alpha}}$$

σ^k being determined by a one dimensional search and the choice of $\underline{W}^k$ determined by the choice of one of the following methods [20].

Steepest Descent Method: $\underline{W}^k = \underline{I}$

This method is simple but has a slow convergence.

Newton-Raphson Method: $\underline{W}^k = \left[\frac{\partial^2 \zeta(\underline{z}^t;\hat{\underline{\alpha}}^k)}{\partial \underline{\alpha}^2}\right]^{-1}$

This has the fastest convergence but is complex and expensive.

Approximate Newton-Raphson Method:

The idea here is, as mentioned above, to approximate $\frac{\partial^2 \zeta(\underline{z}^t;\hat{\underline{\alpha}}^k)}{\partial \underline{\alpha}^2}$ by its expected value $\underline{I}_{\underline{z}^t}(\hat{\underline{\alpha}}^k)$ and to further approximate $\underline{I}_{\underline{z}^t}(\hat{\underline{\alpha}}^k)$ by the expression given in Section 5.3. This is the most common method.

Quasi-Newton Method:

Here, $\underline{W}^k$ is again an approximation to $\left[\frac{\partial^2 \zeta(\underline{z}^t;\hat{\underline{\alpha}}^k)}{\partial \underline{\alpha}^2}\right]^{-1}$ but is built up during the minimization process which starts out like the steepest descent and then switches over to become like the Newton-Raphson.

Note that the approximate Newton-Raphson method requires that the information matrix be nonsingular. We will see below that singularity of the information matrix implies that our model set is overparametrized.

5.3 Information Matrix

Recall that the information matrix is defined by the equivalent expressions (4.4) and (4.5). Therefore, from (5.3), we can write the information matrix as

$$\underline{I}_{\underline{z}^t}(\underline{\alpha}) = E\left\{\sum_{\tau=0}^{t} \frac{\partial^2 \zeta(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha})}{\partial \underline{\alpha}^2}\,\Big|\,\underline{\alpha}\right\}. \tag{5.27}$$

It is useful to define the conditional information matrix $\underline{I}_{\underline{z}(t)|\underline{z}^{t-1}}(\underline{\alpha})$ by the equivalent expressions

$$\underline{I}_{\underline{z}(t)|\underline{z}^{t-1}}(\underline{\alpha}) = E\left\{\frac{\partial^2 \zeta(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha})}{\partial \underline{\alpha}^2}\,\Big|\,\underline{z}^{t-1};\underline{\alpha}\right\}$$

$$= E\left\{\frac{\partial \zeta(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha})}{\partial \underline{\alpha}}\,\frac{\partial \zeta'(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha})}{\partial \underline{\alpha}}\,\Big|\,\underline{z}^{t-1};\underline{\alpha}\right\} \tag{5.28}$$

so that we can write

$$\underline{I}_{\underline{z}^t}(\underline{\alpha}) = E\left\{\sum_{\tau=0}^{t} \underline{I}_{\underline{z}(\tau)|\underline{z}^{\tau-1}}(\underline{\alpha}) \middle| \underline{\alpha}\right\}$$

$$= \sum_{\tau=0}^{t} \bar{\underline{I}}_{\underline{z}(\tau)|\underline{z}^{\tau-1}}(\underline{\alpha}) \tag{5.29}$$

From (5.29), we can see that to derive an expression for $\underline{I}_{\underline{z}^t}(\underline{\alpha})$, we need to first obtain an expression for $\underline{I}_{\underline{z}(\tau)|\underline{z}^{\tau-1}}(\underline{\alpha})$ and then to compute its expected value $\bar{\underline{I}}_{\underline{z}(\tau)|\underline{z}^{\tau-1}}(\underline{\alpha})$.

The calculation of $\underline{I}_{\underline{z}(\tau)|\underline{z}^{\tau-1}}(\underline{\alpha})$ is carried out in Appendix D, where the equation

$$\left[\underline{I}_{\underline{z}(\tau)|\underline{z}^{\tau-1}}(\underline{\alpha})\right]_{ij} = \operatorname{tr}\left[\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i} \frac{\partial \underline{r}'(\tau;\underline{\alpha})}{\partial \alpha_j} \underline{S}^{-1}(\tau;\underline{\alpha})\right]$$

$$+ \tfrac{1}{2} \operatorname{tr}\left[\frac{\partial \underline{S}(\tau;\underline{\alpha})}{\partial \alpha_i} \underline{S}^{-1}(\tau;\underline{\alpha}) \frac{\partial \underline{S}(\tau;\underline{\alpha})}{\partial \alpha_j} \underline{S}^{-1}(\tau;\underline{\alpha})\right] \tag{5.30}$$

for its ij^{th} element is derived. Using (5.30) and the equation

$$E\left\{\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i} \frac{\partial \underline{r}'(\tau;\underline{\alpha})}{\partial \alpha_j}\right\} = \underline{S}_{ij}(\tau;\underline{\alpha}) + \overline{\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}} \overline{\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_j}}, \tag{5.31}$$

where $\overline{\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}}$ denotes the mean of $\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}$ and $\underline{S}_{ij}(\tau;\underline{\alpha})$ denotes the covariance matrix between $\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}$ and $\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_j}$, we can obtain the equation

$$\left[\bar{\underline{I}}_{\underline{z}(\tau)|\underline{z}^{\tau-1}}(\underline{\alpha})\right]_{ij} = \operatorname{tr}\left[\overline{\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}} \overline{\frac{\partial \underline{r}'(\tau;\underline{\alpha})}{\partial \alpha_j}} \underline{S}^{-1}(\tau;\underline{\alpha})\right] +$$

$$+ \operatorname{tr}\left[\underline{S}_{ij}(\tau;\underline{\alpha})\underline{S}^{-1}(\tau;\underline{\alpha}) + \tfrac{1}{2} \frac{\partial \underline{S}(\tau;\underline{\alpha})}{\partial \alpha_i} \underline{S}^{-1}(\tau;\underline{\alpha}) \frac{\partial \underline{S}(\tau;\underline{\alpha})}{\partial \alpha_j} \underline{S}^{-1}(\tau;\underline{\alpha})\right]. \tag{5.32}$$

The corresponding expression for the information matrix is

$$\left[\underline{I}_{\underline{z}^t}(\underline{\alpha})\right]_{ij} = \sum_{\tau=0}^{t} \operatorname{tr}\left[\overline{\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}} \overline{\frac{\partial \underline{r}'(\tau;\underline{\alpha})}{\partial \alpha_j}} \underline{S}^{-1}(\tau;\underline{\alpha})\right] +$$

$$+ \sum_{\tau=0}^{t} \operatorname{tr}\left[\underline{S}_{ij}(\tau;\underline{\alpha})\underline{S}^{-1}(\tau;\underline{\alpha}) + \tfrac{1}{2} \frac{\partial \underline{S}(\tau;\underline{\alpha})}{\partial \alpha_i} \underline{S}^{-1}(\tau;\underline{\alpha}) \frac{\partial \underline{S}(\tau;\underline{\alpha})}{\partial \alpha_j} S^{-1}(\tau;\underline{\alpha})\right] \tag{5.33}$$

Evaluating $\underline{\underline{S}}_{ij}(\tau;\underline{\alpha})$ and $\overline{\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}}$ is conceptually straightforward although computationally expensive. From equation (5.9), it follows that $\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}$ and $\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_j}$ are the outputs of a 4n-dimensional linear system with state vector consisting of $\underline{x}(t)$, $\hat{\underline{x}}(t|t-1;\underline{\alpha})$, $\frac{\partial \hat{\underline{x}}(t|t-1;\underline{\alpha})}{\partial \alpha_i}$, and $\frac{\partial \hat{\underline{x}}(t|t-1;\underline{\alpha})}{\partial \alpha_j}$ with white noise inputs $\underline{\xi}(t)$, $\underline{\theta}(t)$ and with input $\underline{u}(t)$ (Figure 5.6). Therefore, we can solve the usual equations for the mean and covariance of this system (see, e.g., Chapter 4 of [1]) to generate $\underline{\underline{S}}_{ij}(\tau;\underline{\alpha})$ and $\overline{\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}}$. We omit the details of this computation; some of the ideas are illustrated in the following simple example.

Example 5.1

Our system model is

$$x(t+1) = \alpha u(t) + \xi(t)$$

$$z(t) = x(t) + \theta(t)$$

where $x(0)$, $\xi(t)$, $\theta(s)$ are all independent, zero mean, and have mean square value equal to one. The true system has the same form, with $\alpha = 1$. Clearly, we have

$$\hat{x}(t|t-1;\alpha) = \hat{z}(t|t-1;\alpha) = \alpha u(t-1)$$

and

$$S(t) = E\{[z(t) - \hat{z}(t|t-1;\alpha)]^2|\alpha\} = 2$$

independent of α. Therefore,

$$\frac{\partial r(t;\alpha)}{\partial \alpha} = u(t)$$

which (trivially) has mean $u(t)$ and covariance 0. Therefore, from (5.30),

$$I_{z(\tau)|z^{\tau-1}}(\alpha) = \bar{I}_{z(\tau)|z^{\tau-1}}(\alpha) = \tfrac{1}{2} u^2(\tau)$$

so that

$$I_{z^t}(\alpha) = \sum_{\tau=0}^{t} \tfrac{1}{2} u^2(\tau) \ . \qquad \blacksquare$$

Although the information matrix computation in the above example is extremely simple, it is in general quite expensive to calculate the information matrix. The mean and covariance equations of a 4n dimensional linear dynamic system must be propagated and the sum (5.33) accumulated to determine

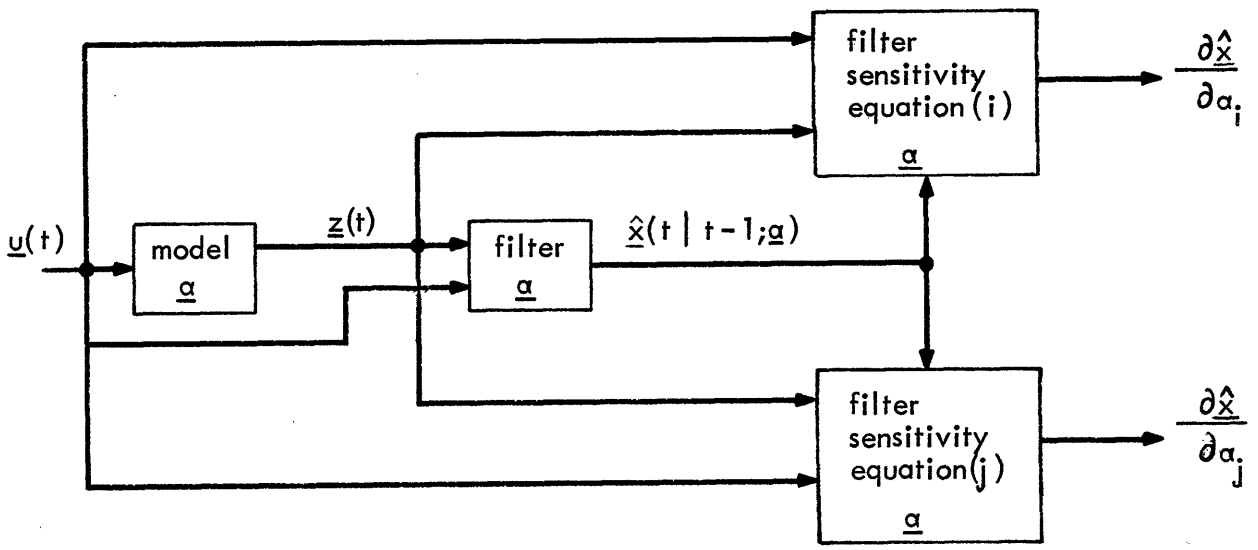


Figure 5.6. Linear System for Information Matrix Element Computation

one element of the information matrix. This calculation must be repeated $\ell(\ell-1)/2$ times since there are that many distinct elements of the symmetric information matrix.

As was the case for the evaluation of the likelihood function and its gradient, the evaluation of the information matrix simplifies somewhat if the system is stationary and it is assumed that the likelihood function can be evaluated by a steady state Kalman filter. As discussed previously, this assumption introduces an approximation that is valid if the observation time interval is long compared to the time required for the optimal Kalman filter to converge to its steady state. Under this approximation, we have $\underline{S}(\tau;\underline{\alpha}) \equiv \underline{S}(\underline{\alpha})$ and $\underline{S}_{ij}(\tau;\underline{\alpha}) \equiv \underline{S}_{ij}(\underline{\alpha})$ constant, and these matrices can be evaluated by solving the steady state covariance equations for the system of Figure 5.6. (The special form of this system can be exploited in the computations.) The resulting expression for the information matrix is

$$\underline{I}_{\underline{z}^t}(\underline{\alpha}) = \sum_{\tau=0}^{t} \operatorname{tr}\left[\frac{\overline{\partial \underline{r}(\tau;\underline{\alpha})}}{\partial \alpha_i} \frac{\overline{\partial \underline{r}'(\tau;\underline{\alpha})}}{\partial \alpha_j} \underline{S}^{-1}(\underline{\alpha})\right]$$

$$+ (t+1) \operatorname{tr}\left[\underline{S}_{ij}(\underline{\alpha})\underline{S}^{-1}(\underline{\alpha}) + \tfrac{1}{2} \frac{\partial \underline{S}(\underline{\alpha})}{\partial \alpha_i} \underline{S}^{-1}(\underline{\alpha}) \frac{\partial \underline{S}(\underline{\alpha})}{\partial \alpha_j} \underline{S}^{-1}(\underline{\alpha})\right]. \tag{5.34}$$

The first term in (5.34) cannot be simplified without further assumptions on $\underline{u}(t)$ (e.g., $\underline{u}(t)$ periodic).

An alternative to the solution of mean and covariance equations for determining the information matrix is to use the stochastic approximation

$$\underline{I}_{\underline{z}^t}(\underline{\alpha}) \approx \sum_{\tau=0}^{t} \operatorname{tr}\left[\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i} \frac{\partial \underline{r}'(\tau;\underline{\alpha})}{\partial \alpha_j} \underline{S}^{-1}(\tau;\underline{\alpha})\right.$$

$$\left. + \tfrac{1}{2} \frac{\partial \underline{S}(\tau;\underline{\alpha})}{\partial \alpha_i} \underline{S}^{-1}(\tau;\underline{\alpha}) \frac{\partial \underline{S}(\tau;\underline{\alpha})}{\partial \alpha_j} \underline{S}^{-1}(\tau;\underline{\alpha})\right]. \tag{5.35}$$

Note that the right hand side of (5.35) is a random variable that has expected value equal to the left hand side (from (5.29) and (5.30)). Therefore, equation (5.35) makes sense if the standard deviation of its right hand side is much smaller than its expected value. This will be the case, for example, if the dominant system excitation is the known input $\underline{u}(t)$ rather than the stochastic input $\underline{\xi}(t)$ so that the first term of (5.33) dominates the second. Alternatively, in the absense of deterministic inputs but with the assumption of a stationary system, it can be shown that the approximation (5.35) is good if the observation interval $[0,t]$ is much longer than the correlation times of $\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}$ and $\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_j}$.

Notice that all the quantities in (5.35) are evaluated during the computation of the gradient $\frac{\partial \zeta(\underline{z}^t;\underline{\alpha})}{\partial \underline{\alpha}}$. Thus the approximation (5.35) is readily computed during a gradient search for the maximum likelihood estimate $\hat{\underline{\alpha}}_t$. In fact, (5.35) is the approximation to the information matrix used in the approximate Newton-Raphson method (Section 5.2 above).

Another use of (5.35) is in on-line identification for adaptive estimation and control. The idea is to extend the multiple model adaptive algorithms briefly mentioned in Section 5.2.1 by evaluating not only $\zeta(\underline{z}^t;\underline{\alpha}_K)$ but also $\frac{\partial \zeta(\underline{z}^t;\underline{\alpha}_K)}{\partial \underline{\alpha}}$ and the approximate expression for $\underline{I}_{\underline{z}^t}(\underline{\alpha}_K)$ for a fixed number of parameter values $\underline{\alpha}_K$, $K = 0,1,\ldots,N$. One can then pick the smallest $\zeta(\underline{z}^t;\underline{\alpha}_K)$ and take one approximate Newton-Raphson step away from $\underline{\alpha}_K$ to interpolate between models in the parameter space. This method, termed the parallel channel maximum likelihood adaptive algorithm, has been applied in a number of practical problems [11]. It has two very significant potential advantages over other modifications of the maximum likelihood identification method for on-line applications. First, the practice of anchoring the Kalman filters at a fixed number of points in parameter space leads to an algorithm with a stable and predictable

behavior; divergence problems cannot occur. Second, by careful selection of the $\underline{\alpha}_K$ one can eliminate the problem of convergence to local minima of the likelihood function associated with on-line, recursive implementations of the maximum likelihood method.

So far we have emphasized the role of the information matrix (or an approximation thereof) in algorithms for numerical minimization of the negative log likelihood function. However, the information matrix also plays a crucial role in analyses associated with identification problems. Indeed, we have argued that a parameter estimate is of little value without an indication of its accuracy.

For the maximum likelihood parameter identification method, the Cramer-Rao lower bound

$$E\{(\underline{\alpha} - \underline{\hat{\alpha}}_t)(\underline{\alpha} - \underline{\hat{\alpha}}_t)' | \underline{\alpha}\} \geq \underline{I}_{\underline{z}^t}^{-1}(\underline{\alpha})$$

provides a lower bound on the accuracy of parameter estimates.[1] As we have shown above, the information matrix can be precomputed without actual observations, and thus the Cramer-Rao lower bound can serve as a tool for experimental design. The maximum accuracy of parameter estimation can be evaluated as a function of such experimental conditions as sensor quality, system excitation, number of observations, etc. before an identification experiment is performed.

5.4 Asymptotic Properties

In this subsection the following assumptions are added to the linear-Gaussian model considered so far.

- The system is time invariant.
- The noise processes are stationary.

---

[1] We will see in the next section that under certain conditions this bound is asymptotically tight.

• A steady state filter is used to compute the likelihood function.

Furthermore, let $\underline{\alpha}_0$ denote the true parameter which is assumed to belong to M. Then, the asymptotic properties of maximum likelihood identification mentioned in Section 4 apply as follows.

5.4.1 Consistency ($\lim_{t\to\infty} \hat{\underline{\alpha}}_t = \underline{\alpha}_0$)

In Section 4 we discussed the consistency of maximum likelihood estimates for independent, identically distributed observations. Recall that the basic condition for consistency was that no parameter has the same single observation likelihood function as the true observation. Since the conditional likelihood function $p(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha})$ is the analog for dependent observations of the single observation likelihood function, it is reasonable to conjecture that an identifiability condition of the form

$$p(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha}) \neq p(z(t)|z^{t-1};\alpha_0)$$

or equivalently

$$\zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha}) \neq \zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha})$$

for all $\underline{\alpha} \neq \underline{\alpha}_0 \in M$ would be sufficient for consistency. This is essentially the situation for the case we are considering, except for some difficulties associated with the presence of the inputs $\underline{u}(t)$. We will see that the above inequalities can be checked in terms of quantities associated with the steady state Kalman filters corresponding to $\underline{\alpha}$ and $\underline{\alpha}_0$.

Recall that $\hat{\underline{\alpha}}_t$ minimizes $\zeta(\underline{z}^t;\underline{\alpha})$ which depends on $\underline{\alpha}$ through the Kalman filter residuals $\underline{r}(t;\underline{\alpha})$ as well as their covariance $\underline{S}(\underline{\alpha})$.[1] Rewrite (5.8) as

$$\frac{\partial \zeta(\underline{z}(\tau)|\underline{z}^{\tau-1};\underline{\alpha})}{\partial \alpha_i} = \underline{r}'(\tau;\underline{\alpha})\underline{S}^{-1}(\underline{\alpha})\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i} + \tfrac{1}{2}\,\mathrm{tr}\left[(\underline{I} - \underline{S}^{-1}(\underline{\alpha})\underline{r}(\tau;\underline{\alpha})\underline{r}'(\tau;\underline{\alpha}))\underline{S}^{-1}(\underline{\alpha})\frac{\partial \underline{S}(\underline{\alpha})}{\partial \alpha_i}\right]$$

where the first term is the i^{th} component of the gradient of

[1] Recall the stationarity and steady state assumptions made above.

$\zeta(\underline{z}(\tau)|\underline{z}^{\tau-1};\underline{\alpha})$ holding $\underline{S}(\underline{\alpha})$ fixed and the second term is the i^{th} component of the gradient holding $\underline{r}(\tau;\underline{\alpha})$ fixed. We then see that the maximum likelihood technique involves minimizing a weighted quadratic criterion in the residuals as well as fitting the residuals second moment to their presumed covariance $\underline{S}(\underline{\alpha})$, precomputed through the algebraic matrix Riccati equation corresponding to $\underline{\alpha}$. One would therefore expect that if, for some $\underline{\alpha} \neq \underline{\alpha}_0$, $\underline{r}(t;\underline{\alpha})$ and $\underline{S}(\underline{\alpha})$ are respectively identical to $\underline{r}(t;\underline{\alpha}_0)$ and $\underline{S}(\underline{\alpha}_0)$, $\underline{\alpha}_0$ will not be identifiable.[1]

Now by virtue of the stationarity assumptions made above we can consider the frequency domain description of the steady state Kalman filter as shown in Figure 5.7. Let

$$\underline{G}(\mathcal{z};\underline{\alpha}) \equiv \underline{C}(\underline{\alpha})(\mathcal{z}\underline{I} - \underline{A}(\underline{\alpha}))^{-1}\underline{B}(\underline{\alpha})$$

and

$$\underline{H}(\mathcal{z};\underline{\alpha}) \equiv \underline{C}(\underline{\alpha})(\mathcal{z}\underline{I} - \underline{A}(\underline{\alpha}))^{-1}\underline{A}(\underline{\alpha})\underline{H}(\underline{\alpha}) + \underline{I} .$$

Then

$$\underline{r}(\mathcal{z};\underline{\alpha}) = -\underline{H}(\mathcal{z};\underline{\alpha})^{-1}\underline{G}(\mathcal{z};\underline{\alpha})\underline{u}(\mathcal{z}) + \underline{H}(\mathcal{z};\underline{\alpha})^{-1}\underline{z}(\mathcal{z}) \tag{5.36}$$

and, in view of our discussion above, it comes as no surprise that it is necessary for consistency to have

$$\underline{G}(\mathcal{z};\underline{\alpha}) \not\equiv G(\mathcal{z};\underline{\alpha}_0) \tag{5.37}$$

or

$$\underline{H}(\mathcal{z};\underline{\alpha}) \not\equiv (H(\mathcal{z};\underline{\alpha}) \tag{5.38}$$

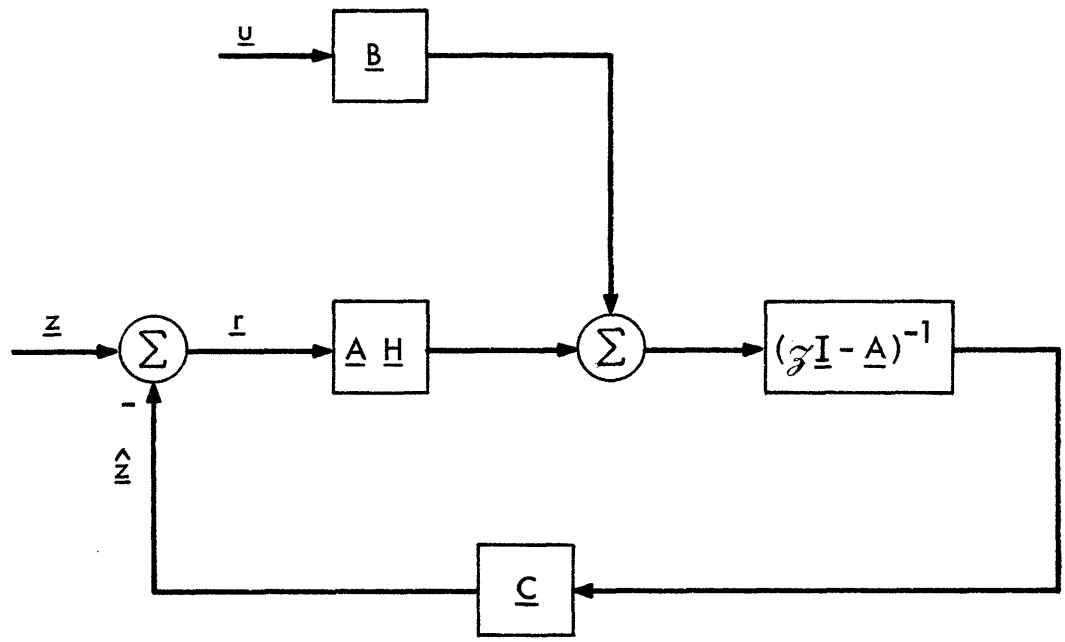
or

$$\underline{S}(\underline{\alpha}) \neq \underline{S}(\underline{\alpha}_0) \tag{5.39}$$

for all $\underline{\alpha} \neq \underline{\alpha}_0$, $\underline{\alpha} \in M$ since otherwise some other parameter would have the same likelihood function as the true parameter.

---

[1] We use the phrases "$\underline{\alpha}_0$ is identifiable" and "the maximum likelihood estimate is consistent" interchangeably.



$\mathcal{Z}$ = transform variable

Figure 5.7 Frequency Domain Description of the Steady State Kalman Filter

Let us now raise the question of the sufficiency of those conditions and let us restrict our attention to the case where $\underline{u}(t) \equiv \underline{0}$[1] (or equivalently, $\underline{G}(\mathcal{J};\underline{\alpha}) \equiv 0 \quad \forall \alpha \in M$). In [13], Caines shows that, under assumptions of stationarity of the inputs and outputs of the system (which in our present case follows from stability), $\underline{\alpha}_t$ converges into the set M_0 of parameters minimizing $E\{\zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha})|\underline{\alpha}_0\}$. Note that under our present assumptions this is a time invariant function of $\underline{\alpha}$. It is also implicitly shown in [13] that if condition (5.38) is satisfied, then $\underline{\alpha}_0$ is the only element of M_0 and consistency of $\hat{\underline{\alpha}}_t$ follows. This says that if the steady state Kalman filter transfer function $\underline{H}(\mathcal{J};\underline{\alpha}_0)$ corresponding to $\underline{\alpha}_0$ is different from that corresponding to $\underline{\alpha} \neq \underline{\alpha}_0$, then $\hat{\underline{\alpha}}_t$ is consistent. If, however, these transfer functions are the same and condition (5.38) does not hold, then condition (5.39) is necessary and sufficient for consistency. This result is shown in Appendix F and is illustrated by the following example.

Example 5.2

Consider the system

$$\underline{x}(t+1) = \underline{\xi}(t) \quad ; \quad \underline{\xi}(t) \sim N(0,\underline{\Xi})$$

$$\underline{z}(t) = \underline{x}(t) + \underline{\theta}(t) \quad ; \quad \underline{\theta}(t) \sim N(0,\underline{\Theta})$$

where $\underline{\alpha}$ contains unknown parameters in $\underline{\Xi}$ and $\underline{\Theta}$. Clearly

$$\hat{\underline{z}}(t|t-1;\underline{\alpha}) = (\hat{\underline{x}}t|t-1;\underline{\alpha}) = 0 \qquad \forall \alpha \in M$$

so that

$$\underline{H}(\mathcal{J};\underline{\alpha}) = \underline{I} \text{ and } \underline{G}(\mathcal{J};\underline{\alpha}) = 0 \qquad \forall \alpha \in M$$

However,

$$\underline{S}(\underline{\alpha}) = \underline{\Xi}(\underline{\alpha}) + \underline{\Theta}(\underline{\alpha})$$

so that $\hat{\underline{\alpha}}_t$ will converge to $\underline{\alpha}_0$ if and only if

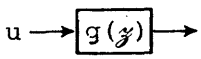
$$\underline{S}(\alpha) \neq \underline{S}(\underline{\alpha}_0) \qquad \forall \underline{\alpha} \neq \underline{\alpha}_0 \ , \ \underline{\alpha} \in M \qquad \blacksquare$$

---

[1] The case where general deterministic inputs $\underline{u}(t)$ are present requires a more elaborate analysis which is briefly sketched out in Appendix E.

The above discussion has considered global identifiability, but one can also define a notion of local identifiability. The following examples will clarify this point.

Example 5.3



$$x(t+1) = ax(t) + bu(t)$$

$$z(t) = cx(t)$$

$$\underline{\alpha}' = [\, a \quad b \quad c \,]$$

$$g(\mathscr{y};\underline{\alpha}) = \frac{cb}{\mathscr{y}-a}$$

and one cannot identify c and b, only their product cb. ■

Example 5.4

$$x_1(t+1) = a_1 x_1(t) + \xi(t)$$

$$x_2(t+1) = a_2 x_2(t) + \xi(t) \quad ; \quad a_1 \neq a_2$$

$$z(t) = x_1(t) + x_2(t) + \theta(t)$$

and if a_1 and a_2 have their true values exchanged they give rise to the same $E\{\zeta(z(t)\,|\,z^{t-1};\alpha)\}$ thus making the two sets of values indistinguishable. ■

Note in example 5.4 that there exist neighborhoods about the true values of a_1 and a_2 such that the maximum likelihood method will be consistent if restricted to these neighborhoods. On the other hand, in example 5.3 no such neighborhoods about the true values of c and b can be found. The situation in example 5.4 is termed local identifiability and often suffices for practical purposes.

Local identifiability can be investigated by determining the rank of the information matrix.[1] Indeed, under the above cited assumptions of

[1] Determination of rank is a difficult problem for which a sophisticated numerical analytic technique is required. The singular value decomposition approach is recommended [17].

Caines [13] and $\underline{u}(t) \equiv \underline{0}$,[1] we formally have, for some neighborhood of $\underline{\alpha}_0$

$$E\{\zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha})\} - E\{\zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha}_0)$$

$$= (\underline{\alpha} - \underline{\alpha}_0)' \, E\left\{\frac{\partial \zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha})}{\partial \underline{\alpha}}\right\}_{\underline{\alpha}=\underline{\alpha}_0}$$

$$+ \tfrac{1}{2}(\underline{\alpha} - \underline{\alpha}_0)' \, E\left\{\frac{\partial^2 \zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha})}{\partial \underline{\alpha}^2}\right\}_{\underline{\alpha}=\underline{\alpha}_0} (\underline{\alpha} - \underline{\alpha}_0)$$

$$+ \text{h.o.t.} \qquad (5.40)$$

But the first term after the equality sign vanishes for $\underline{\alpha} = \underline{\alpha}_0$ and the matrix of the second term is $\bar{\underline{I}}_{\underline{z}(t)|\underline{z}^{t-1}}(\underline{\alpha}_0)$ by definition.[2] Therefore, in view of the above consistency results, the positive definiteness of $\bar{\underline{I}}_{\underline{z}(t)|\underline{z}^{t-1}}(\underline{\alpha}_0)$ is a sufficient condition ([20]) for $\underline{\alpha}_0$ to be a unique local minimum of $E\{\zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha})\}$ and for $\hat{\underline{\alpha}}_t$ to be locally consistent.

So far we have discussed identifiability conditions for the consistency (global and local) of maximum likelihood identification under assumptions of stationarity and steady state Kalman filters. We now briefly address the other asumptotic properties mentioned in Section 4 under the same assumptions and conditions.

5.4.2 Asymptotic unbiasedness

As indicated in Section 4, this follows from consistency since, under very general technical conditions,

$$\lim_{t\to\infty} E\{\hat{\underline{\alpha}}_t|\underline{\alpha}_0\} = E\{\lim_{t\to\infty} \hat{\underline{\alpha}}_t|\underline{\alpha}_0\} = \underline{\alpha}_0$$

5.4.3 Asymptotic normality

As $t\to\infty$, $\hat{\underline{\alpha}}_t$ tends to a Gaussian random vector with mean $\underline{\alpha}_0$ and covari-

---

[1] See footnote p. 53.

[2] Under the present assumptions $\underline{I}_{\underline{z}(t)|\underline{z}^{t-1}}(\underline{\alpha}_0)$ is a time invariant quantity (see also Section 5.3).

ance $\underline{I}_{\underline{z}^t}^{-1}(\underline{\alpha}_0)$. This was shown in [14] under general conditions of smoothness and boundedness of the innovations. As argued in Section 4, this also allows us to compute confidence intervals for $\underline{\alpha}_0$ about the estimate $\hat{\underline{\alpha}}_t$ which is a useful tool in applications.

5.4.4 Asymptotic efficiency

$$\lim_{t\to\infty} E\{(\hat{\underline{\alpha}}_t - \underline{\alpha}_0)(\hat{\underline{\alpha}}_t - \underline{\alpha}_0)' \,|\, \underline{\alpha}_0\} = \underline{I}_{\underline{z}^t}^{-1}(\underline{\alpha}_0)$$

follows from asymptotic normality. In conjunction with asymptotic unbiasness, this means that the Cramer-Rao lower bound is achieved for a large number of observations.

Note that under the assumptions of this section, we have from (5.34), for $\underline{u}(t) \equiv 0$

$$\underline{I}_{\underline{z}^t}(\underline{\alpha}_0) = (t+1)\,\mathrm{tr}\left[\underline{S}_{ij}(\underline{\alpha}_0)\underline{S}^{-1}(\underline{\alpha}_0) + \tfrac{1}{2}\frac{\partial \underline{S}(\underline{\alpha}_0)}{\partial \alpha_i}\underline{S}^{-1}(\underline{\alpha}_0)\frac{\partial \underline{S}(\underline{\alpha}_0)}{\partial \alpha_j}\underline{S}^{-1}(\underline{\alpha}_0)\right]$$

$$= (t+1)\,\bar{\underline{I}}_{\underline{z}(t)\,|\,\underline{z}^{t-1}}(\underline{\alpha}_0)$$

and here again, in terms of the asymptotic covariance matrix

$$E\{(\hat{\underline{\alpha}}_t - \underline{\alpha}_0)(\hat{\underline{\alpha}}_t - \underline{\alpha}_0)'\} \approx \frac{1}{t}\bar{\underline{I}}^{-1}_{\underline{z}(t)\,|\,\underline{z}^{t-1}}(\underline{\alpha}_0)$$

This generalizes equation (4.9) and plays a similar role from an applications point of view.

The asymptotic efficiency property of ML estimates is quite important. It says, for the special case of very long observation sequences, that the maximum likelihood identification method is an optimal method in the sense that it gives unbiased parameter estimates with minimum error covariance matrix. Note that the asymptotic efficiency property agrees with our earlier statement that the Cramer-Rao lower bound tends to be tight when the signal-to-noise ratio is high; the large number of observations effectively permits us to average the noise down to a low level.

Finally, we recall our discussion of Section 5.3, where we pointed

out that since the information matrix is precomputable it can play an important role in experimental design. The fact that the inverse information matrix is an asymptotically tight lower bound to the error covariance matrix provides additional support for that discussion.

5.5 Maximum Likelihood Estimation Under Modeling Errors

One is often in a situation where the true system is not a member of the model set in use. This might happen inadvertently, as in the case where one is ignorant of the true system's order, or deliberately, as in the case where one uses lower order models to reduce the computational burden. As we have seen, maximum likelihood identification requires repeated solution of the Kalman filtering problem which is a significant computational burden if the dimension of the models considered is large; so there is indeed a great incentive to work with reduced order models.

The previous analysis does not apply to this situation and the question of convergence of the ML estimate has to be reanalyzed in the present context. One possible approach to this problem has been suggested by Baram and Sandell ([15]) and relies on some information theoretic concepts which will now be summarized.

5.5.1 Information definition and properties

The analysis presented in this part is general, applying to any model set M, finite or nonfinite, and requires no assumptions of Gaussianess or stationarity.

Recall the probability density of past to present observations

$$p(\underline{z}^t;\underline{\alpha}) = p(\underline{z}(t)|\,\underline{z}^{t-1};\underline{\alpha}) \ \ldots\ p(z(0);\underline{\alpha})$$

where $\underline{\alpha}$ can now be any element of $T \equiv M \cup \{*\}$ and where * denotes the "true" parameter. If for some pair of parameters $\underline{\alpha}_1$, $\underline{\alpha}_2$ we have

$$p(\underline{z}^t;\underline{\alpha}_1) > p(\underline{z}^t;\underline{\alpha}_2)$$

then it is natural to say that the information in the observations contained

in z^t favors $\underline{\alpha}_1$ over $\underline{\alpha}_2$. The above equation is equivalent to

$$\ln p(\underline{z}^t;\underline{\alpha}_1) > \ln p(\underline{z}^t;\underline{\alpha}_2)$$

or

$$\ln \frac{p(\underline{z}^t;\underline{\alpha}_1)}{p(\underline{z}^t;\underline{\alpha}_2)} > 0 .$$

Therefore, $\ln \frac{p(\underline{z}^t;\underline{\alpha}_1)}{p(\underline{z}^t;\underline{\alpha}_2)}$ can be regarded as a measure of information in $\underline{z}^t$ favoring $\underline{\alpha}_1$ over $\underline{\alpha}_2$. Similarly,

$$\ln \frac{p(\underline{z}^t;\underline{\alpha}_1)}{p(\underline{z}^t;\underline{\alpha}_2)} - \ln \frac{p(\underline{z}^{t-1};\underline{\alpha}_1)}{p(\underline{z}^{t-1};\underline{\alpha}_2)} = \ln \frac{p(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha}_1)}{p(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha}_2)}$$

can be regarded as a measure of the new information in $\underline{z}(t)$ favoring $\underline{\alpha}_1$ over $\underline{\alpha}_2$. Finally define,

$$J_t(\underline{\alpha}_1;\underline{\alpha}_2) \equiv E_* \left\{ \ln \frac{p(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha}_1)}{p(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha}_2)} \right\} \tag{5.41}$$

the expected new information in $\underline{z}(t)$ favoring $\underline{\alpha}_1$ over $\underline{\alpha}_2$.

Note that the expected value in (5.41) is taken with respect to the true probability measure. So $J_t(\underline{\alpha}_1;\underline{\alpha}_2)$ can only be computed if the true probability is known. But it is still useful as an analytical tool as will be shown later.

Now some of the properties of $J_t(\underline{\alpha}_1;\underline{\alpha}_2)$ are presented. The proofs can be found in [15].

i) For any $\underline{\alpha} \in M$,

$$J_t(*;\underline{\alpha}) \geq 0 \tag{5.42}$$

with the equality holding if and only if

$$p(\underline{z}^t;*) = p(\underline{z}^t;\underline{\alpha}) \quad \text{a.s.}$$

I.e., on the average, the true model is always favored by the observations.

ii) $$|J_t(\underline{\alpha}_i;\underline{\alpha}_i)| = 0$$

$$|J_t(\underline{\alpha}_i;\underline{\alpha}_j)| = |J_t(\underline{\alpha}_j;\underline{\alpha}_i)|$$

$$|J_t(\underline{\alpha}_i;\underline{\alpha}_k)| \leq |J_t(\underline{\alpha}_i;\underline{\alpha}_j)| + |J_t(\underline{\alpha}_j;\underline{\alpha}_k)|$$

I.e.,

$$d_t(\underline{\alpha}_i;\underline{\alpha}_j) \equiv |J_t(\underline{\alpha}_i;\underline{\alpha}_j)| \tag{5.43}$$

constitutes a pseudo metric on T or an information distance between $\underline{\alpha}_i$ and $\underline{\alpha}_j$.

5.5.2 Application to linear systems

Returning now to the linear-Gaussian case with stationary system and model, consider the true system described by equations of the form (2.1) through (2.11) assuming $\underline{u}(t) \equiv 0$ (i.e. only noise inputs). This true system can then be specified by the n^* dimensional time invariant matrices:

$$\{\underline{A}(*),\underline{L}(*),\underline{C}(*),\underline{\Xi}(*),\underline{\Theta}(*)\} \tag{5.44}$$

and the n^{α} dimensional model set by

$$M(\underline{\alpha}) = \{(A(\underline{\alpha}),L(\underline{\alpha}),C(\underline{\alpha}),\underline{\Xi}(\underline{\alpha}),\Theta(\underline{\alpha}));\ \underline{\alpha} \in M\} \tag{5.45}$$

As before

$$\begin{aligned}\underline{S}(t;\underline{\alpha}) &\equiv E_{\alpha}\{(\underline{z}(t) - \hat{\underline{z}}(t;\underline{\alpha}))(\underline{z}(t) - \hat{\underline{z}}(t;\underline{\alpha}))'\} \\ &= E_{\alpha}\{\underline{r}(t;\underline{\alpha})r'(t;\underline{\alpha})\}\end{aligned} \tag{5.46}$$

denotes the predicted observation error covariance assuming that $\underline{\alpha}$ is the true parameter. If each model in (5.45) is detectable and controllable (see [6]) the steady state limit

$$\underline{S}(\underline{\alpha}) = \lim_{t\to\infty} \underline{S}(t;\underline{\alpha}) \tag{5.47}$$

exists and has a finite positive definite value.

Furthermore, let

$$\underline{S}_*(t;\underline{\alpha}) \equiv E_*\{\underline{r}(t;\underline{\alpha})\underline{r}'(t;\underline{\alpha})\} \tag{5.48}$$

denote the observation error covariance of the predictor corresponding to $\underline{\alpha}$ when in fact the model corresponding to * is the correct one. Here again, let

$$\underline{S}_*(\underline{\alpha}) = \lim_{t\to\infty} \underline{S}_*(t;\underline{\alpha}) \tag{5.49}$$

be the steady state limit if it exists. $\underline{S}_*(\underline{\alpha})$ is generated by solving the covariance equation for the $(n^* + n^{\alpha})$ linear system described in Appendix G. (One will note the similarity to the reduced order filter computations in equations (3.16) through (3.18)).

Finally, assume that the residuals sequence $\underline{r}(t;\underline{\alpha})$ is ergodic (a sufficient condition would be the stability and observability of the corresponding stationary model $M(\underline{\alpha})$). Then, the conditional probability density of $\underline{z}(t)$ given the past observations $\underline{z}^{t-1}$ corresponding to a model $M(\underline{\alpha})$ is given by (5.1) and the information distance between two models $M(\underline{\alpha}_1)$ and $M(\underline{\alpha}_2)$ can be derived as follows. From (5.41) and (5.3)

$$J_t(\underline{\alpha}_1;\underline{\alpha}_2) = E_*\{\zeta(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha}_2)\} - E_*\{\zeta(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha}_1)\} \tag{5.50}$$

where, under the additional steady state assumptions used in Section 5.4,

$$E_*\{\zeta(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha})\} = \tfrac{1}{2} \ln \det[\underline{S}(\underline{\alpha})] + \tfrac{1}{2} \operatorname{tr}[\underline{S}^{-1}(\underline{\alpha})\underline{S}_*(\underline{\alpha})] \tag{5.51}$$

is a time invariant function of $\underline{\alpha}$, as argued in Section 5.4.1. We also have from (5.41) through (5.43)

$$\begin{aligned} J(\underline{\alpha}_1;\underline{\alpha}_2) &= J(*;\underline{\alpha}_2) - J(*;\underline{\alpha}_1) \\ &= d(*;\underline{\alpha}_2) - d(*;\underline{\alpha}_1) \end{aligned} \tag{5.52}$$

where $d(*;\underline{\alpha})$ is the information distance between * and $\underline{\alpha}$.

It can then be shown that, under the above assumptions of stationarity and ergodicity (see [16]), maximum likelihood estimates on the compact parameter set M converge almost surely to $\underline{\alpha}_0$ where

$$d(*;\underline{\alpha}_0) \leq d(*;\underline{\alpha}) \tag{5.53}$$

for all $\underline{\alpha} \in M$.

This means that maximum likelihood estimates converge to the parameter

in M closest to the true model. This also represents a generalization of the consistency result of Section 5.4.1. Indeed, in view of (5.50) - (5.52), condition (5.53) holds if and only if

$$E_*\{\zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha}_0)\} \leq E_*\{\zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha})\} \tag{5.54}$$

and the same identifiability issues as those discussed in Section 5.4.1 are relevant here to $\underline{\alpha}_0$ which satisfies (5.53). In other words, the consistency and identifiability properties connected with the true parameter in Section 5.4.1 generalize here to the parameter which minimizes the information distance to the true parameter.

Recall now that this section was concerned with linear stationary Gaussian models with only noise inputs. Here again, as mentioned in Section 5.4, the presence of deterministic inputs complicates the analysis. Indeed, the information distance defined above, depends in this case on $\underline{u}^t$ as illustrated by the following simple example.

Example 5.5

Consider the true model:

$$\begin{cases} \begin{bmatrix} x_1(t+1) \\ x_2(t+2) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} \\ y(t) = x_1(t) + x_2(t) \end{cases}$$

and the lower order models

$$M(\alpha_1): \begin{cases} x(t+1) = x(t) + u_1(t) + u_2(t) \\ y(t) = x(t) \end{cases}$$

and

$$M(\alpha_2): \begin{cases} x(t+1) = -x(t) + u_1(t) + u_2(t) \\ y(t) = x(t) \end{cases}$$

Then if $u_1(t) \equiv 1$ and $u_2(t) \equiv 0$

Then if $u_1(t) \equiv 1$ and $u_2(t) \equiv 0$

$$J(\alpha_1;\alpha_2|\underline{u}^t) > 0$$

forcing the choice of $M(\alpha_1)$; and if $u_1(t) \equiv 0$ and $u_2(t) \equiv 1$

$$J(\alpha_2;\alpha_1|\underline{u}^t) > 0$$

forcing the choice of $M(\alpha_2)$. This makes sense since in each case we are exciting only one of the two modes of the true system. ■

The convergence analysis in the presence of deterministic inputs as well as the other asymptotic properties of maximum likelihood estimates mentioned in Section 5.4 will not be discussed in this report.

SECTION 6

SUMMARY AND CONCLUSIONS

In this report, we have given a brief introduction to the maximum likelihood method for identifying the parameters of a linear-Gaussian state space model. We have ignored or only mentioned briefly a number of important issues including approximate maximum likelihood identification of nonlinear systems, special cases of the basic formulation that can be implemented with less computation, determining the best model order, and many others. Rather, we have concentrated on the issues of computing and interpreting the maximum likelihood estimate of the unknown parameters in a general linear-Gaussian state space model of fixed order. Our most basic conclusions were the following

- Maximum likelihood theory provides asymptotically optimal estimates in the sense that they are asymptotically unbiased and achieve the Cramer-Rao lower bound.
- A quantitative measure of estimation accuracy is provided by the Cramer-Rao lower bound which is asymptotically tight.
- Asymptotic accuracy of parameter estimates can be determined off-line by computation of the Cramer-Rao lower bound so that various alternative experimental conditions can be evaluated before data is gathered.
- The maximum likelihood equations are general-purpose, valid for any linear state space model and involving compuations familiar to Kalman filter designers.
- The asymptotic sensitivity of the maximum likelihood estimates to modeling errors, either inadvertant or deliberate, can be readily assessed.

It seems to us that the above properties are essential for any adequate theory of system identification. It is notable that these properties of the maximum likelihood method for the nonlinear parameter identification problem are similar to the properties of the Kalman filtering method for the linear state estimation problem. Kalman filtering theory has become a basic tool for off-line studies of system performance during preliminary design studies by covariance simulation, and for on-line integration of multisensor systems. We feel that maximum likelihood theory will become a basic tool for off-line problems of identification experiment design and for processing of experimental data to extract estimates of system parameters.

APPENDIX A

DERIVATION OF THE RICCATI SENSITIVITY EQUATION (5.15)

Consider the Riccati Equation (3.12)

$$\underline{\Sigma}(t+1|\,t;\underline{\alpha}) = \underline{A}(t;\underline{\alpha})\underline{\Sigma}(t|\,t-1;\underline{\alpha})\underline{A}'(t;\underline{\alpha}) + \underline{L}(t;\underline{\alpha})\underline{\Xi}(t;\underline{\alpha})\underline{L}'(t;\underline{\alpha})$$
$$- \underline{A}(t;\underline{\alpha})\underline{H}(t;\underline{\alpha})\underline{S}(t;\underline{\alpha})\underline{H}'(t;\underline{\alpha})\underline{A}'(t;\underline{\alpha})$$

Dropping the arguments and underscores from the matrix notations for clarity, differentiate with respect to α_i and use (5.13) for $\frac{\partial \underline{H}}{\partial \alpha_i}$

$$\frac{\partial \Sigma(t+1|\,t;\underline{\alpha})}{\partial \alpha_i} = A \frac{\partial \Sigma(t|\,t-1;\underline{\alpha})}{\partial \alpha_i} A' + \frac{\partial}{\partial \alpha_i}[L \Xi L']$$
$$- AHC \frac{\partial \Sigma(t|\,t-1;\underline{\alpha})}{\partial \alpha_i} A'$$
$$- A \frac{\partial \Sigma(t|\,t-1;\underline{\alpha})}{\partial \alpha_i} C'H'A'$$
$$- AH \frac{\partial C}{\partial \alpha_i} \Sigma(t|\,t-1;\underline{\alpha})A'$$
$$- A\Sigma(t|\,t-1;\underline{\alpha}) \frac{\partial C'}{\partial \alpha_i} H'A'$$
$$+ AH \frac{\partial S}{\partial \alpha_i} H'A'$$
$$+ \frac{\partial A}{\partial \alpha_i} \Sigma(t|\,t-1;\underline{\alpha})[I - C'H']A'$$
$$+ A[I - HC]\Sigma(t|t-1;\underline{\alpha}) \frac{\partial A'}{\partial \alpha_i} \qquad (A.1)$$

Substituting for $\frac{\partial S}{\partial \alpha_i}$ from (5.14) and using the definition of $\bar{A}$ in (5.12),

The first, third, fourth terms and the first substituted term in (A.1) group into:

$$\bar{A} \frac{\partial \Sigma(t|\,t-1;\underline{\alpha})}{\partial \alpha_i} \bar{A}'$$

The eighth and ninth terms into:

$$\frac{\partial A}{\partial \alpha_i} \Sigma(t|\,t-1;\underline{\alpha})\,\bar{A}' + \bar{A}\Sigma(t|\,t-1;\underline{\alpha}) \frac{\partial A'}{\partial \alpha_i}$$

The fifth, sixth terms and the third and fourth substituted terms into:

APPENDIX B

DERIVATION OF THE FORWARD FILTER EQUATIONS (5.18) TO (5.23)

As in Appendix A, the arguments and underscores are dropped from matrix notations for clarity.

Consider equation (5.8)

$$\frac{\partial \zeta(\underline{z}(t)|\,\underline{z}^{t};\underline{\alpha})}{\partial \alpha_i} = r'S^{-1}\frac{\partial r}{\partial \alpha_i} - \tfrac{1}{2}r'S^{-1}\frac{\partial S}{\partial \alpha_i}S^{-1}r + \operatorname{tr}\left[S^{-1}\frac{\partial S}{\partial \alpha_i}\right]$$

$$= r'S^{-1}\frac{\partial r}{\partial \alpha_i} + \tfrac{1}{2}\operatorname{tr}\left[S^{-1}(I - rr'S^{-1})\frac{\partial S}{\partial \alpha_i}\right]$$

And substitute for $\frac{\partial r}{\partial \alpha_i}$ from (5.9) and $\frac{\partial S}{\partial \alpha_i}$ from (5.14).

$$\frac{\partial \zeta(\underline{z}(t)|\,\underline{z}^{t};\underline{\alpha})}{\partial \alpha_i} = r'S^{-1}\left[-\frac{\partial C}{\partial \alpha_i}\hat{x}(t|\,t-1;\underline{\alpha}) - C\frac{\partial \hat{x}(t|\,t-1;\underline{\alpha})}{\partial \alpha_i}\right]$$

$$+ \tfrac{1}{2}\operatorname{tr}\left[S^{-1}(I - rr'S^{-1})\left(C\frac{\partial \Sigma(t|\,t-1;\underline{\alpha})}{\partial \alpha_i}C' + \frac{\partial \Theta}{\partial \alpha_i}\right.\right.$$

$$\left.\left. + \frac{\partial C}{\partial \alpha}\Sigma(t|\,t-1;\underline{\alpha})C' + C\Sigma(t|\,t-1;\underline{\alpha})\frac{\partial C'}{\partial \alpha}\right)\right]$$

$$= \tfrac{1}{2}\operatorname{tr}\left[(S^{-1} - S^{-1}rr'S^{-1})\left(2C\Sigma(t|\,t-1;\underline{\alpha})\frac{\partial C'}{\partial \alpha} + \frac{\partial \Theta}{\partial \alpha_i}\right)\right]$$

$$- r'S^{-1}\frac{\partial C}{\partial \alpha_i}\hat{x}(t|\,t-1;\underline{\alpha})$$

$$- r'S^{-1}C\frac{\partial \hat{x}(t|\,t-1;\underline{\alpha})}{\partial \alpha_i}$$

$$+ \tfrac{1}{2}\operatorname{tr}\left[C'(S^{-1} - S^{-1}rr'S^{-1})C\frac{\partial \Sigma(t|\,t-1;\underline{\alpha})}{\partial \alpha_i}\right]$$

where for the first term the identity

$$\operatorname{tr}(AB) = \operatorname{tr}(A'B')$$

was used, and for the last term the identity

$$\operatorname{tr}(AB) = \operatorname{tr}(BA)$$

was used.

Equation (5.18) with (5.20), (5.21) and (5.22) now follow.

$$- AH \frac{\partial C}{\partial \alpha_i} \Sigma(t|\, t-1;\underline{\alpha})\bar{A}' - \bar{A}\Sigma(t|\, t-1;\underline{\alpha}) \frac{\partial C'}{\partial \alpha_i} H'A'$$

Finally the second term and the second substituted term remain as:

$$\frac{\partial}{\partial \alpha_i} [L\Xi L'] + AH \frac{\partial \epsilon}{\partial \alpha_i} H'A'$$

Equations (5.15) and (5.16) now follow.

Consider now equations (5.10) and (5.11). Using (5.12) we can reexpress (5.11) as

$$\underline{\omega}_i(t;\underline{\alpha}) = \frac{\partial}{\partial\alpha_i}[A(I-HC)]\hat{x}(t|t-1;\underline{\alpha}) + \frac{\partial}{\partial\alpha_i}[AH]\,z + \frac{\partial B}{\partial\alpha_i}u$$

$$= \left(\frac{\partial A}{\partial\alpha_i} - AH\frac{\partial C}{\partial\alpha_i}\right)\hat{x}(t|t-1;\underline{\alpha}) + \frac{\partial}{\partial\alpha_i}[AH]r + \frac{\partial B}{\partial\alpha_i}u$$

But, from (5.13) and (5.14) we have

$$\frac{\partial}{\partial\alpha_i}[AH] = \frac{\partial A}{\partial\alpha_i}H + A\frac{\partial\Sigma}{\partial\alpha_i}C'S^{-1} + A\Sigma\frac{\partial C'}{\partial\alpha_i}S^{-1} - A\Sigma C'S^{-1}C\frac{\partial\Sigma}{\partial\alpha_i}C'S^{-1}$$

$$- A\Sigma C'S^{-1}\frac{\partial\Theta}{\partial\alpha_i}S^{-1} - A\Sigma C'S^{-1}\frac{\partial C}{\partial\alpha_i}\Sigma C'S^{-1}$$

$$- A\Sigma C'S^{-1}C\,\Sigma\frac{\partial C'}{\partial\alpha_i}S^{-1}\ .$$

The second and fourth term group into

$$-\bar{A}\frac{\partial\Sigma}{\partial\alpha_i}C'S^{-1}$$

and equation (5.19) with (5.23) now follow.

APPENDIX C

ADJOINT COMPUTATIONS

Some properties of adjoint equations are discussed.

Lemma:

Suppose $J = \langle \underline{c}, \underline{x} \rangle$ where $\underline{A}\underline{x} = \underline{b}$, $\underline{A}$ invertible.

Then $J = \langle \underline{\lambda}, \underline{b} \rangle$ where $\underline{A}^*\underline{\lambda} = \underline{c}$.

Proof:

$$J = \langle \underline{c}, \underline{A}^{-1}\underline{b} \rangle = \langle (\underline{A}^{-1})^* \underline{c}, \underline{b} \rangle = \langle \underline{\lambda}, \underline{b} \rangle$$

where $(\underline{A}^{-1})^* \underline{c} = (A^*)^{-1} \underline{c} = \underline{\lambda}$

or $\underline{A}^* \underline{\lambda} = c.$ ■

Corollary 1

Suppose $J = \sum_{t=0}^{T} \underline{c}_t' \underline{x}_t$

where $\underline{x}_{t+1} = \underline{A}_t \underline{x}_t + \underline{u}_t$, $\underline{x}_0$ given.

Then $J = \underline{\lambda}_0' \underline{x}_0 + \sum_{t=1}^{T} \underline{\lambda}_t' \underline{u}_{t-1}$

where $\underline{\lambda}_t = \underline{A}_t' \underline{\lambda}_{t+1} + \underline{c}_t$, $\underline{\lambda}_T = \underline{c}_T$

Proof:

$$\begin{bmatrix} I & 0 & 0 & \cdots & 0 & 0 & 0 \\ -A_0 & I & 0 & \cdots & 0 & 0 & 0 \\ \cdot & \cdot & & & & & \\ & \cdot & \cdot & & & & \\ & & \cdot & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & -A_{T-2} & I & 0 \\ & & & & 0 & -A_{T-1} & I \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \cdot \\ \cdot \\ \cdot \\ \\ x_T \end{bmatrix} = \begin{bmatrix} x_0 \\ u_0 \\ \cdot \\ \cdot \\ \cdot \\ \\ u_{T-1} \end{bmatrix}$$

From the lemma

$$J = \begin{bmatrix} \lambda_0' & \lambda_1' & \cdots & \lambda_T' \end{bmatrix} \begin{bmatrix} x_0 \\ u_0 \\ \cdot \\ \cdot \\ \cdot \\ u_{T-1} \end{bmatrix}$$

where

$$\begin{bmatrix} I & -A_0' & 0 & \cdots & 0 & 0 \\ 0 & I & -A_1' & \cdots & 0 & 0 \\ & & & & & \\ & & & & I & -A_{T-1}' \\ & & & & 0 & I \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \cdot \\ \cdot \\ \cdot \\ \lambda_T \end{bmatrix} = \begin{bmatrix} c_0 \\ c_1 \\ \cdot \\ \cdot \\ \cdot \\ c_T \end{bmatrix}$$

i.e. $\lambda_T = c_T$

$$\lambda_t = A_t' \lambda_{t+1} + c_t$$

■

Corollary 2

Suppose $J = \sum_{t=0}^{T} \operatorname{tr}[\underline{C}_t \underline{X}_t]$, $\underline{C}_t$ symmetric

where $\underline{X}_{t+1} = \underline{A}_t \underline{X}_t \underline{A}_t' + \underline{U}_t$, $\underline{X}_0$ given and symmetric, $\underline{U}_t$ symmetric.

Then $J = \operatorname{tr}(\underline{\Lambda}_0 \underline{X}_0) + \sum_{t=1}^{T} \operatorname{tr}[\underline{\Lambda}_t \underline{U}_{t-1}]$

where $\underline{\Lambda}_{t-1} = \underline{A}_{t-1}' \underline{\Lambda}_t \underline{A}_{t-1} + \underline{C}_{t-1}$, $\underline{\Lambda}_T = \underline{C}_T$

Proof:

Define formally the operator

$$\mathscr{L}(X_0, X_1, \ldots, X_T) \equiv (X_0, -A_0 X_0 A_0' + X_1, \ldots, -A_{T-1} X_{T-1} A_{T-1}' + X_T)$$

and the inner product

$$<(C_0,C_1,\ldots,C_T),(X_0,X_1,\ldots,X_T)> \equiv \sum_{t=0}^{T} \mathrm{tr}[C_t X_t]$$

Evaluate $J = <(C_0,C_1,\ldots,C_T),\ (X_0,X_1,\ldots,X_T)>$

subject to $\mathscr{L}(X_0,\ldots,X_T) = (X_0,U_0,\ldots,U_{T-1})$

From the lemma,

$$J = <(\Lambda_0,\Lambda_1,\ldots,\Lambda_T),(X_0,U_0,\ldots,U_{T-1})>$$

where $\mathscr{L}^*(\Lambda_0,\Lambda_1,\ldots,\Lambda_T) = (C_0,C_1,\ldots,C_T)$

But $\mathscr{L}^*(\Lambda_0,\Lambda_1,\ldots,\Lambda_T) = (\Lambda_0 - A_0'\Lambda_1 A_0,\ldots,\Lambda_{T-1} - A_{T-1}'\Lambda_T A_{T-1},\Lambda_T)$

indeed: $<\mathscr{L}(X_0,X_1,\ldots,X_T),(\Lambda_0,\Lambda_1,\ldots,\Lambda_T)>$

$$= \mathrm{tr}[X_0\Lambda_0] + \mathrm{tr}\sum_{t=1}^{T} \mathrm{tr}[(-A_{t-1}X_{t-1}A'_{t-1} + X_t)\Lambda_t]$$

$$= \mathrm{tr}[X_0\Lambda_0] + \sum_{t=1}^{T} \mathrm{tr}[-X_{t-1}A'_{t-1}\Lambda_t A_{t-1} + X_t\Lambda_t]$$

$$= \mathrm{tr}[X_0\Lambda_0 - X_0A'_0\Lambda_1A_0] + \mathrm{tr}[X_1\Lambda_1 - X_1A'_1\Lambda_2A_1] + \ldots$$

$$\ldots + \mathrm{tr}[X_{T-1}\Lambda_{T-1} - X_{T-1}A'_{T-1}\Lambda_T A_{T-1}] + \mathrm{tr}[X_T\Lambda_T]$$

$$= \sum_{t=0}^{T-1} \mathrm{tr}[X_t(\Lambda_t - A_t'\Lambda_{t+1}A_t)] + \mathrm{tr}[X_T\Lambda_T]$$

$$= <(X_0,X_1,\ldots,X_T),\mathscr{L}^*(\Lambda_0,\Lambda_1,\ldots,\Lambda_T)> \quad \blacksquare$$

Corollary 3

Suppose $J = \sum_{t=0}^{T} \{\underline{c}_t'\underline{x}_t + \mathrm{tr}[\underline{C}_t\underline{X}_t]\}$, $\underline{C}_t$ symmetric

where $\underline{x}_{t+1} = \underline{A}_t\underline{x}_t - \underline{A}_t\underline{X}_t\underline{c}_t + \underline{u}_t$; $\underline{x}_0$ given

and $\underline{X}_{t+1} = \underline{A}_t\underline{X}_t\underline{A}_t' + \underline{U}_t$; $\underline{X}_0$ given and symmetric, $\underline{U}_t$ symmetric.

Then $J = \underline{\lambda}_0'\underline{x}_0 + \mathrm{tr}[\underline{\Lambda}_0\underline{X}_0] + \sum_{t=1}^{T} \{\underline{\lambda}_t'\underline{u}_{t-1} + \mathrm{tr}[\underline{\Lambda}_t\underline{U}_{t-1}]\}$

where $\underline{\lambda}_t = \underline{A}_t'\underline{\lambda}_{t+1} + \underline{c}_t$, $\underline{\lambda}_T = \underline{c}_T$

and $\underline{\Lambda}_t = \underline{A}_t'\underline{\Lambda}_{t+1}\underline{A}_t - \underline{A}_t'\underline{\lambda}_{t+1}\underline{c}_t' + \underline{C}_t$, $\underline{\Lambda}_T = \underline{C}_T$.

Proof:

Define formally the operator

$$\mathscr{L}(x_0,x_1,\ldots,x_T,X_0,X_1,\ldots,X_T) \equiv (x_0, -A_0x_0 + A_0X_0c_0 + x_1,\ldots, -A_{T-1}x_{T-1}$$

$$+ A_{T-1}X_{T-1}c_{T-1} + x_T, X_0, -A_0X_0A_0' + X_1,\ldots, -A_{T-1}X_{T-1}A_{T-1}' + X_T)$$

and the inner product on $\prod_{i=0}^{T} R^n \times \prod_{i=0}^{T} R^{n\times n}$

$$<(c_0,c_1,\ldots,c_T,C_0,C_1,\ldots,C_T),\ (x_0,x_1,\ldots,x_T,X_0,X_1,\ldots,X_T)>$$

$$\equiv \sum_{t=0}^{T} \{c_t' x_t + tr[C_tX_t]\}$$

Evaluate $J = <(c_0,c_1,\ldots,c_T,C_0,C_1,\ldots,C_T),\ (x_0,x_1,\ldots,x_T,X_0,\ldots,X_T)>$

subject to

$$\mathscr{L}(x_0,\ldots,x_T,X_0,\ldots,X_T) = (x_0,u_0,\ldots,u_{T-1},X_0,U_0,\ldots,U_{T-1}) .$$

From the lemma,

$$J = <(\lambda_0,\lambda_0,\ldots,\lambda_T,\Lambda_0,\Lambda_1,\ldots,\Lambda_T),\ (x_0,u_0,\ldots,u_{T-1},X_0,U_0,\ldots,U_{T-1})>$$

where $\mathscr{L}^*(\lambda_0,\ldots,\lambda_T,\Lambda_0,\ldots,\Lambda_T) = (c_0,\ldots,c_T,C_0,\ldots,C_T)$.

But $\mathscr{L}^*(\lambda_0,\ldots,\lambda_T,\Lambda_0,\ldots,\Lambda_T) = (\lambda_0 - A_0'\lambda_1,\ldots,\lambda_{T-1} - A_{T-1}'\lambda_T,\lambda_T,\Lambda_0 - A_0'\Lambda_1A_0 +$

$$+ A_0'\lambda_1c_0',\ldots,\Lambda_{T-1} - A_{T-1}'\Lambda_TA_{T-1} + A_{T-1}'\lambda_Tc_{T-1}',\Lambda_T) .$$

Indeed $<\mathscr{L}(x_0,\ldots,x_T,X_0,\ldots,X_T),\ (\lambda_0,\ldots,\lambda_T,\Lambda_0,\ldots,\Lambda_T)>$

$$= x_0'\lambda_0 + \sum_{t=1}^{T} (-A_{t-1}x_{t-1} + A_{t-1}X_{t-1}c_{t-1} + x_t)'\lambda_t$$

$$+ tr[X_0\Lambda_0] + \sum_{t=1}^{T} tr[(-A_{t-1}X_{t-1}A_{t-1}' + X_t)\Lambda_t]$$

$$= x_0'\lambda_0 + \sum_{t=1}^{T} \{-x_{t-1}'A_{t-1}'\lambda_t + tr[X_{t-1}A_{t-1}'\lambda_tc_{t-1}'] + x_t'\lambda_t\}$$

$$+ tr[X_0\Lambda_0] + \sum_{t=1}^{T} tr[-X_{t-1}A_{t-1}'\Lambda_tA_{t-1} + X_t\Lambda_t]$$

$$= x_0' (\lambda_0 - A_0'\lambda_1) + x_1' (\lambda_1 - A_1'\lambda_2) + \cdots + x_T'\lambda_T$$

$$+ \operatorname{tr}[X_0(\Lambda_0 - A_0'\Lambda_1 A_0 + A_0'\lambda_1 c_0')] + \cdots + \operatorname{tr}[X_T\Lambda_T]$$

$$= \sum_{t=0}^{T-1} x'_t (\lambda_t - A_t'\lambda_{t+1}) + x_T'\lambda_T$$

$$+ \sum_{t=0}^{T-1} \operatorname{tr}[X_t(\Lambda_t - A_t'\Lambda_{t+1}A_t + A_t'\lambda_{t+1}c'_t)] + \operatorname{tr}[X_T\Lambda_T]$$

$$= \langle (x_0,\ldots,x_T,X_1,\ldots,X_T), \mathscr{L}^*(\lambda_0,\ldots,\lambda_T,\Lambda_0,\ldots,\Lambda_T)\rangle \quad \blacksquare$$

APPENDIX D

DERIVATION OF THE INFORMATION MATRIX EQUATION (5.30)

Here again, the arguments and underscores are dropped from matrix notations for clarity. From (5.28)

$$\left[\underline{I}_{\underline{z}(\tau)\,|\,\underline{z}^{\tau-1}}(\underline{\alpha})\right]_{ij} = E\left\{\frac{\partial \zeta(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha})}{\partial \alpha_i}\,\frac{\partial \zeta'(\underline{z}(\tau)\,|\,\underline{z}^{\tau-1};\underline{\alpha})}{\partial \alpha_j}\,\middle|\,\underline{z}^{\tau-1};\underline{\alpha}\right\} \quad (D.1)$$

and from (5.8)

$$= E\left\{\left(r'S^{-1}\frac{\partial r}{\partial \alpha_i} - \tfrac{1}{2}r'S^{-1}\frac{\partial S}{\partial \alpha_i}S^{-1}r + \tfrac{1}{2}\mathrm{tr}\left[S^{-1}\frac{\partial S}{\partial \alpha_i}\right]\right)\right.$$
$$\left.\left(r'S^{-1}\frac{\partial r}{\partial \alpha_j} - \tfrac{1}{2}r'S^{-1}\frac{\partial S}{\partial \alpha_j}S^{-1}r + \tfrac{1}{2}\mathrm{tr}\left[S^{-1}\frac{\partial S}{\partial \alpha_j}\right]\right)\middle|\,\underline{z}^{\tau-1};\underline{\alpha}\right\} \quad (D.2)$$

$$= E\left\{\left(r'S^{-1}\frac{\partial r}{\partial \alpha_i}\right)\left(r'S^{-1}\frac{\partial r}{\partial \alpha_j}\right)\right.$$
$$+ \tfrac{1}{4}\left(r'S^{-1}\frac{\partial S}{\partial \alpha_i}S^{-1}r\right)\left(r'S^{-1}\frac{\partial S}{\partial \alpha_j}S^{-1}r\right)$$
$$+ \tfrac{1}{4}\,\mathrm{tr}\left[S^{-1}\frac{\partial S}{\partial \alpha_i}\right]\mathrm{tr}\left[S^{-1}\frac{\partial S}{\partial \alpha_j}\right]$$
$$+ \left(r'S^{-1}\frac{\partial r}{\partial \alpha_i}\right)\left(-\tfrac{1}{2}r'S^{-1}\frac{\partial S}{\partial \alpha_j}S^{-1}r\right)$$
$$+ \left(r'S^{-1}\frac{\partial r}{\partial \alpha_i}\right)\left(\tfrac{1}{2}\mathrm{tr}\left[S^{-1}\frac{\partial S}{\partial \alpha_j}\right]\right)$$
$$+ \left(-\tfrac{1}{2}r'S^{-1}\frac{\partial S}{\partial \alpha_i}S^{-1}r\right)\left(r'S^{-1}\frac{\partial r}{\partial \alpha_j}\right)$$
$$+ \left(-\tfrac{1}{2}r'S^{-1}\frac{\partial S}{\partial \alpha_i}S^{-1}r\right)\left(\tfrac{1}{2}\mathrm{tr}\left[S^{-1}\frac{\partial S}{\partial \alpha_j}\right]\right)$$
$$+ \left(\tfrac{1}{2}\mathrm{tr}\left[S^{-1}\frac{\partial S}{\partial \alpha_i}\right]\right)\left(r'S^{-1}\frac{\partial r}{\partial \alpha_j}\right)$$
$$\left.+ \left(\tfrac{1}{2}\mathrm{tr}\left[S^{-1}\frac{\partial S}{\partial \alpha_i}\right]\right)\left(-\tfrac{1}{2}r'S^{-1}\frac{\partial S}{\partial \alpha_i}S^{-1}r\right)\middle|\,\underline{z}^{\tau-1};\underline{\alpha}\right\} \quad (D.3)$$

Recall now from Section 3 that, conditioned on $\underline{z}^{\tau-1}$, $\underline{r}(\tau;\underline{\alpha})$ has zero mean and covariance $\underline{S}(\tau;\underline{\alpha})$ and that $\frac{\partial \underline{r}(\tau;\underline{\alpha})}{\partial \alpha_i}$ is deterministic.

Therefore, using the identities

$$E\{(\underline{c}'\underline{x})(\underline{x}'\underline{A}\underline{x})\} = 0$$

$$E\{(\underline{x}'\underline{A}\underline{x})(\underline{x}'\underline{A}\underline{x})\} = (\mathrm{tr}[\underline{\Sigma}\underline{A}])^2 + 2\ \mathrm{tr}[\underline{\Sigma}\underline{A}\underline{\Sigma}\underline{A}]$$

for $\underline{x} \sim N(\underline{0},\underline{\Sigma})$, $\underline{A} = \underline{A}'$

the fourth, fifth, sixth and eighth terms of (D.3) are zero and (D.3) reduces to:

$$\begin{aligned}\left[\underline{I}_{\underline{z}(\tau)|\underline{z}^{\tau-1}}(\underline{\alpha})\right]_{ij} &= \mathrm{tr}\left[\frac{\partial r}{\partial \alpha_i}\frac{\partial r'}{\partial \alpha_j}S^{-1}\right] \\ &\quad + \tfrac{1}{4}\left(\mathrm{tr}\left[\frac{\partial S}{\partial \alpha_i}S^{-1}\right]\mathrm{tr}\left[\frac{\partial S}{\partial \alpha_j}S^{-1}\right] + 2\ \mathrm{tr}\left[\frac{\partial S}{\partial \alpha_i}S^{-1}\frac{\partial S}{\partial \alpha_j}S^{-1}\right]\right) \\ &\quad + \tfrac{1}{4}\ \mathrm{tr}\left[S^{-1}\frac{\partial S}{\partial \alpha_i}\right]\ \mathrm{tr}\left[S^{-1}\frac{\partial S}{\partial \alpha_j}\right] \\ &\quad - \tfrac{1}{4}\ \mathrm{tr}\left[\frac{\partial S}{\partial \alpha_i}S^{-1}\right]\ \mathrm{tr}\left[\frac{\partial S}{\partial \alpha_j}S^{-1}\right] \\ &\quad - \tfrac{1}{4}\ \mathrm{tr}\left[\frac{\partial S}{\partial \alpha_i}S^{-1}\right]\ \mathrm{tr}\left[\frac{\partial S}{\partial \alpha_j}S^{-1}\right] \qquad \text{(D.4)}\end{aligned}$$

$$\left[\underline{I}_{\underline{z}(\tau)|\underline{z}^{\tau-1}}(\underline{\alpha})\right]_{ij} = \mathrm{tr}\left[\frac{\partial r}{\partial \alpha_i}\ \frac{\partial r'}{\partial \alpha_j}S^{-1}\right] + \tfrac{1}{2}\ \mathrm{tr}\left[\frac{\partial S}{\partial \alpha_i}S^{-1}\frac{\partial S}{\partial \alpha_j}S^{-1}\right] \qquad \text{(D.5)}$$

which is (5.30).

APPENDIX E

CONSISTENCY IN THE PRESENCE OF DETERMINISTIC INPUTS

We have already argued in Section 5.4.1 that condition (5.37) could be necessary for $\hat{\underline{\alpha}}_t$ to be consistent. However, we shall now see that its sufficiency depends on additional assumptions on the deterministic inputs $\underline{u}(t)$. Note that in the presence of such inputs, $E\{\zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha})\}$ is not, in general, time invariant and the global convergence result generalizes as follows. In [21], Ljung shows that $\hat{\underline{\alpha}}_t$ converges into the set $M_0(t)$ of parameters minimizing

$$\lim_{t\to\infty} \frac{1}{t+1} \sum_{\tau=0}^{t} E\{\zeta(\underline{z}(t)|\underline{z}^{t-1};\underline{\alpha})$$

and this set depends in general on the input signal. In view of (5.36) it is therefore reasonable to expect condition (5.37) to be sufficient if the input $\underline{u}(t)$ is general enough to excite all modes of the system. No rigorous proof of this will be given in this report. However, we shall take a closer look at the following special case.

In [12], Ljung uses the prediction error parameter estimate obtained by minimizing a scalar function of the matrix

$$\sum_{\tau=0}^{t} [\underline{R}^{\frac{1}{2}}(\tau)\underline{r}(\tau;\underline{\alpha})][\underline{R}^{\frac{1}{2}}(\tau)\underline{r}(\tau;\underline{\alpha})]'$$

where $\underline{R}(\tau)$ is some positive definite weighting matrix, and shows that under general conditions of bounded fourth moments of the residuals $\underline{r}(t;\underline{\alpha})$, search over models leading to stable Kalman filters and overall system stability, this prediction error estimate converges into the set of models that give the same output prediction as the true system in the sense:

$$\lim_{t\to\infty} \inf \frac{1}{t+1} \sum_{\tau=0}^{t} \left|\hat{\underline{z}}(\tau;\underline{\alpha}_0) - \hat{\underline{z}}(\tau;\underline{\alpha})\right|^2 = 0 .$$

Here again this set depends in general on the input signal and will be contained in the set of all models with same input-output relation as true

model, if the input is general enough to excite all modes of the system. It is shown in [12] that a sufficient condition would be for $\underline{u}(t)$ to be independent of the process noise and be persistently exciting.[1] It is also shown in [12] that the above prediction error identification method includes the maximum likelihood method only in the cases where $\underline{S}$ is completely known (i.e. independent of $\underline{\alpha}$) or $\underline{S}$ is completely unknown (i.e. all its elements are free and part of $\underline{\alpha}$). In those cases $\underline{S}(\underline{\alpha})$ is not computed through a Riccati equation and consistency of $\hat{\underline{\alpha}}_t$ follows if (5.37) or (5.38) hold.

As for local consistency, by the same argument as above, equation (5.40) is not in general time invariant and so the local identifiability result of Section 5.4.1 must be generalized as follows. In [19], Tsé shows that local identifiability of the true parameter $\underline{\alpha}_0$ follows from positive definiteness (non-singularity) of the average information matrix

$$\lim_{t\to\infty} \frac{1}{t+1} \sum_{\tau=0}^{t} \bar{\underline{I}}_{\underline{z}(t)|\underline{z}^{t-1}}(\underline{\alpha}_0)$$

and here again we will note that this result depends in general on the input signal $\underline{u}(t)$.

---

[1] $\underline{u}(t)$ is persistently exciting if, for all M, there exists $\delta(M)$ and $N_0(M)$ such that

$$\delta \underline{I} < \frac{1}{N} \sum_{1}^{N} \underline{u}_M(t)\, \underline{u}'_M(t) < \frac{1}{\delta}\, \underline{I}$$

for $N > N_0$ and where

$$\underline{u}'_M(t) \equiv [\underline{u}'(t) \ \ldots \ \underline{u}'(t-M)] \ .$$

APPENDIX F

ON CONDITION (5.39)

We first prove the following lemma

Lemma:

Let $\underline{A}$ and $\underline{B}$ be two real symmetric $r \times r$ positive definite matrices, then

$$\ln(\det[\underline{A}]) + \mathrm{tr}[\underline{A}^{-1}\underline{B}] \geq \ln(\det[\underline{B}]) + r \qquad \text{(E.1)}$$

with equality holding if and only if $\underline{A} = \underline{B}$.

Proof:

(E.1) is equivalent to

$$\mathrm{tr}[\underline{A}^{-1}\underline{B}] - \ln(\det[\underline{B}]) + \ln(\det[\underline{A}]) \geq r$$

$$\mathrm{tr}[\underline{A}^{-1}\underline{B}] - \ln\left(\frac{\det[\underline{B}]}{\det[\underline{A}]}\right) \geq r$$

$$\mathrm{tr}[\underline{A}^{-1}\underline{B}] - \ln(\det[\underline{A}^{-1}\underline{B}]) \geq r$$

Let $\underline{A}^{\frac{1}{2}}$ denote the real symmetric positive square root of $\underline{A}$ and $\underline{X} \equiv \underline{A}^{-\frac{1}{2}}\underline{B}\underline{A}^{-\frac{1}{2}}$ then $\underline{X}$ is a real symmetric positive definite matrix with eigenvalues $\lambda_i(\underline{X}) > 0$, $i = 1,\ldots,r$, $\mathrm{tr}[\underline{X}] = \sum_{i=1}^{r} \lambda_i(\underline{X})$ and $\det[\underline{X}] = \prod_{i=1}^{r} \lambda_i(\underline{X})$ so that (E.1) is equivalent to

$$\mathrm{tr}[\underline{X}] - \ln(\det[\underline{X}] \geq r$$

or

$$\sum_{i=1}^{r} [\lambda_i(\underline{X}) - \ln \lambda_i(\underline{X})] \geq r \ .$$

Since for any scalar $x > 0$

$$x - \ln x \geq 1$$

with equality holding if and only if $x = 1$ (E.1) follows, with equality holding if and only if

$$\lambda_i(\underline{X}) = 1 \qquad \forall i = 1,\ldots,r \ .$$

Since $\underline{X}$ is real symmetric equality holds if and only if

$$\underline{X} \equiv \underline{A}^{-\frac{1}{2}}\underline{B}\underline{A}^{-\frac{1}{2}} = \underline{I}$$

or

$$\underline{A} = \underline{B} \,. \qquad \blacksquare$$

Now recalling our discussion in Section 5.4.1, if (5.38) does not hold for some $\underline{\alpha} \neq \underline{\alpha}_0$ then

$$E\{\underline{r}(t;\underline{\alpha})\underline{r}'(t;\underline{\alpha})\,|\,\underline{\alpha}_0\} = E\{\underline{r}(t;\underline{\alpha}_0)\underline{r}'(t;\underline{\alpha}_0)\,|\,\underline{\alpha}_0\} \equiv \underline{S}(\underline{\alpha}_0)$$

and

$$E\{\zeta(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha})\,|\,\underline{\alpha}_0\} = \tfrac{1}{2}\ln(\det[\underline{S}(\underline{\alpha})]) + \tfrac{1}{2}\,\mathrm{tr}[\underline{S}^{-1}(\underline{\alpha})\underline{S}(\underline{\alpha}_0)]$$

But

$$E\{\zeta(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha}_0)\,|\,\underline{\alpha}_0\} = \tfrac{1}{2}\ln(\det[\underline{S}(\underline{\alpha}_0)]) + \tfrac{1}{2}\,\mathrm{tr}[\underline{S}^{-1}(\underline{\alpha}_0)\underline{S}(\underline{\alpha}_0)]$$

$$= \tfrac{1}{2}\ln(\det[\underline{S}(\underline{\alpha}_0)]) + \frac{r}{2}$$

and, under the assumptions of Section 5.4.1, $\underline{S}(\underline{\alpha}_0)$ and $\underline{S}(\underline{\alpha})$ are real symmetric positive definite matrices. Therefore, from the lemma above,

$$E\{\zeta(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha})\,|\,\underline{\alpha}_0\} \geq E\{\zeta(\underline{z}(t)\,|\,\underline{z}^{t-1};\underline{\alpha}_0)\,|\,\underline{\alpha}_0\}$$

with equality holding if and only if $S(\underline{\alpha}) = S(\underline{\alpha}_0)$.

This means that if (5.38) does not hold, condition (5.39) ($\underline{S}(\underline{\alpha}) \neq \underline{S}(\underline{\alpha}_0)$) is necessary and sufficient for $\underline{\alpha}_0$ to be the only element of M_0 from which consistency of $\hat{\underline{\alpha}}_t$ follows.

APPENDIX G

DERIVATION OF $\underline{S}_*(\underline{\alpha})$ (EQUATION 5.49)

Consider the $(n^* + n^{\alpha})$ dimensional dynamic equation generating simultaneously the true state $\underline{x}(t;*)$ and the estimate $\hat{\underline{x}}(t|\, t-1;\underline{\alpha})$ corresponding to the reduced order model $M(\underline{\alpha})$:

$$\begin{bmatrix} \underline{x}(t+1;*) \\ \hat{\underline{x}}(t+1|\, t;\underline{\alpha}) \end{bmatrix} = \underline{A}^*_{\alpha} \begin{bmatrix} \underline{x}(t;*) \\ \hat{\underline{x}}(t|\, t-1;\underline{\alpha}) \end{bmatrix} + \underline{L}^*_{\alpha} \begin{bmatrix} \underline{\xi}(t) \\ \underline{\theta}(t) \end{bmatrix} \tag{E.1}$$

where

$$\underline{A}^*_{\alpha} \equiv \begin{bmatrix} \underline{A}(*) & \underline{0} \\ \underline{A}(\underline{\alpha})\underline{H}(\underline{\alpha})\underline{C}(*) & \underline{A}(\underline{\alpha})[\underline{I} - \underline{H}(\underline{\alpha})\underline{C}(\underline{\alpha})] \end{bmatrix} \tag{E.2}$$

$$\underline{L}^*_{\alpha} \equiv \begin{bmatrix} \underline{L}(*) & \underline{0} \\ \underline{0} & \underline{A}(\underline{\alpha})\underline{H}(\underline{\alpha}) \end{bmatrix} \tag{E.3}$$

$$\underline{H}(\underline{\alpha}) = \underline{\Sigma}(t|\, t-1;\underline{\alpha})\ \underline{C}'(\underline{\alpha})\,[\underline{C}(\underline{\alpha})\underline{\Sigma}(t|\, t-1;\underline{\alpha})\underline{C}'(\underline{\alpha}) + \underline{\Theta}(\underline{\alpha})]^{-1} \tag{E.4}$$

(obtained directly from equations (2.1), (2.2), (3.10) and (3.11)).

Also, let

$$\underline{\Xi}^* = \begin{bmatrix} \underline{\Xi}(*) & \underline{0} \\ \underline{0} & \underline{\Theta}(*) \end{bmatrix} \quad \text{and} \quad \underline{C}^*_{\alpha} = [\underline{C}(*) \quad -\underline{C}(\underline{\alpha})] \tag{E.5}$$

Then

$$\underline{\Sigma}^*_{\alpha}(t) \equiv E\left\{ \begin{bmatrix} \underline{x}(t;*) \\ \hat{\underline{x}}(t|\, t-1;\underline{\alpha}) \end{bmatrix} [\underline{x}'(t;*) \quad \hat{\underline{x}}'(t|\, t-1;\underline{\alpha})] \right\} \tag{E.6}$$

is generated by the Lyapunov equation

$$\underline{\Sigma}^*_{\alpha}(t+1) = \underline{A}^*_{\alpha}\underline{\Sigma}^*_{\alpha}(t)\,\underline{A}^{*\prime}_{\alpha} + \underline{L}^*_{\alpha}\underline{\Xi}^*\underline{L}^{*\prime}_{\alpha} \tag{E.7}$$

Let $$\underline{\Sigma}^*_{\alpha} = \lim_{t\to\infty} \underline{\Sigma}^*_{\alpha}(t) \tag{E.8}$$

denote its steady state value. This limit exists if $\underline{A}^*_{\alpha}$ has all its eigenvalues inside the unit circle.

Finally

$$\underline{S}_*(\underline{\alpha}) = \underline{C}^*_{\alpha}\,\underline{\Sigma}^*_{\alpha}\,\underline{C}^{*\prime}_{\alpha} + \underline{\Theta}(*) \qquad \text{(E.9)}$$

REFERENCES

[1] F. C. Schweppe, Uncertain Dynamic Systems, Prentice Hall, Englewood Cliffs, New Jersey, 1973.

[2] R. K. Mehra, "Synthesis of Optimal Inputs for Multiinput-Multioutput Systems with Process Noise", System Identification: Advances and Case Studies, R. K. Mehra and D. G. Lainiotis, Eds., Academic Press, New York, 1976, pp. 211-249.

[3] K. J. Åström and P. Eykhoff, "System Identification-A Survey", Automatica, Vol. 7, pp. 123-162, 1971.

[4] I. Gustavsson, L. Ljung and T. Söderström, "Survey Paper-Identification of Processes in Closed Loop-Identifiability and Accuracy Aspects", Automatica, Vol. 13, pp. 59-75, 1977.

[5] A. H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press, New York, 1970.

[6] A. Gelb, Ed., Applied Optimal Estimation, M.I.T. Press, Cambridge, Mass., 1974.

[7] S. S. Wilks, Mathematical Statistics, John Wiley & Sons, Inc., New York, 1962.

[8] M. Athans et al., "The Stochastic Control of the F-8C Aircraft Using a Multiple Model Adaptive Control Method-Part I: Equilibrium Flight", IEEE Trans. on Automatic Control, Vol. AC-22, No. 5, Oct. 1977, pp. 768-780.

[9] N. K. Gupta and R. K. Mehra, "Computational Aspects of Maximum Likelihood Estimation and Reduction in Sensitivity Function Calculations", IEEE Trans. on Automatic Control, Vol. AC-19, No. 6, Dec. 1974, pp. 774-783.

[10] R. L. Kashyap, "Maximum Likelihood Identification of Stochastic Linear Systems," IEEE Trans. on Automatic Control, Vol. AC-15, No. 1, Feb. 1970, pp. 25-34.

[11] G. Stein, G. L. Hartmann and R. C. Hendrick, "Adaptive Control Laws for F-8 Flight Test", IEEE Trans. on Automatic Control, Vol. AC-22, No. 5, Oct. 1977, pp. 758-767.

[12] L. Ljung, "On the Consistency of Prediction Error Identification Methods", System Identification: Advances and Case Studies, R. K. Mehra and D. G. Lainiotis, Eds., Academic Press, New York, 1976, pp. 121-164.

[13] P. E. Caines, "Prediction Error Identification Methods for Stationary Stochastic Processes", IEEE Trans. on Automatic Control, Vol. AC-21, No. 4, Aug. 1976, pp. 500-505.

[14] P. E. Caines and L. Ljung, "Asymptotic Accuracy and Normality of Prediction Error Estimators", Research Report 7602, 1976, Dept. of Electrical Engineering, Univ. of Toronto, Canada

[15] Y. Baram and N. R. Sandell, Jr., "An Information Theoretic Approach to Dynamical System Modeling and Identification", IEEE Trans. on Automatic Control, Vol. AC-23, No. 1, Feb. 1978.

[16] K. I. Yared, Ph.D. Thesis, to appear, M.I.T.

[17] A. J. Laub, personal communication.

[18] L. Ljung, "The Extended Kalman Filter as a Parameter Estimator for Linear Systems," Report LiTH-ISY-I-0154, Dept. of Electrical Engineering, Linköping University, Sweden, 1977.

[19] E. Tse, "Information Matrix and Local Identifiability of Parameters," presented at the JACC 1973, Columbus, Ohio.

[20] D. Luenberger, Introduction to Linear and Nonlinear Programming, Addison-Wesley Publishing Co., Inc., 1973.

[21] L. Ljung, "Convergence Analysis of Identification Methods," to appear.