ON MAXIMUM LIKELIHOOD IDENTIFICATION
OF LINEAR STATE SPACE MODELS

by

KHALED I. YARED

B.E., American University of Beirut (1974)
S.M., Massachusetts Institute of Technology (1976)
E.E., Massachusetts Institute of Technology (1977)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREE OF
DOCTOR OF PHILOSOPHY
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September, 1979

Signature of Author

Department of Electrical Engineering and Computer Science,
June 29, 1979

Certified by

Thesis Supervisor

Accepted by

Chairman, Department Committee
DISCLAIMER OF QUALITY

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available. If you are dissatisfied with this product and find it unusable, please contact Document Services as soon as possible.

Thank you.

Some pages in the original document contain color pictures or graphics that will not scan or reproduce well.
ON MAXIMUM LIKELIHOOD IDENTIFICATION OF LINEAR STATE SPACE MODELS

by

KHALED I. YARED

Submitted to the Department of Electrical Engineering and Computer Science on June 29, 1979, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

The maximum likelihood method, as applied to the identification of linear state space models, is studied. The asymptotic properties of ML parameter estimates in this context are first analyzed and some recent general theoretical results are interpreted in the state space framework. In particular, information theoretic concepts are used to obtain insight into the nature of the asymptotic state space model selected by the ML method when the model set considered does not include the "true" system. The special problem of identification under feedback is also reviewed.

Then, attention is turned to some practical aspects of computing ML parameter estimates for state space models. In particular, the use of adjoint computations in evaluating the gradient and Hessian of the likelihood function is investigated. Simple examples with computer simulated data show that significant computational savings can be thus achieved.

THESIS SUPERVISOR:

Nils R. Sandell, Jr.

Title: Associate Professor of Electrical Engineering and Computer Science
ACKNOWLEDGEMENT

I wish to express my gratitude to Professor Nils Sandell, my thesis supervisor, for his patient guidance, friendly encouragement and constant support. His insight and ideas were essential to the development of this work.

I was fortunate to have Professors Peter Caines, Fred Schweppe and Gunter Stein as my thesis readers. I would like to express to them my deepest appreciation for their invaluable comments and suggestions.

I thank James Carrig for his help in the ungratifying task of writing and debugging computer programs.

I am grateful to my friends and fellow students Paul Toldalagi, Gregory Lauer, David Rossi and Demosthenis Teneketzis for their unreserved help, for many stimulating discussions and especially for creating a friendly atmosphere.

I also want to thank Fifa Monserrate for her excellent work in typing this thesis and Arthur Giordani of the drafting department for his help with the figures.

Finally, I wish to thank my wife Dalal for her invaluable moral support.

This research was conducted at the M.I.T. Laboratory for Information and Decision Systems with partial support provided by the Department of Energy under contract ERDA-E(49-18)-2087.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>2</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>3</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>4</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>7</td>
</tr>
<tr>
<td>CHAPTER 1: INTRODUCTION</td>
<td>8</td>
</tr>
<tr>
<td>1.1 General Discussion and Motivation</td>
<td>8</td>
</tr>
<tr>
<td>1.2 Background</td>
<td>11</td>
</tr>
<tr>
<td>1.3 Contents and Contribution</td>
<td>16</td>
</tr>
<tr>
<td>CHAPTER 2: ASYMPTOTIC PROPERTIES OF ML IDENTIFICATION</td>
<td>18</td>
</tr>
<tr>
<td>2.1 The Information Distance Between Models</td>
<td>18</td>
</tr>
<tr>
<td>2.2 Problem Formulation</td>
<td>22</td>
</tr>
<tr>
<td>2.3 Convergence Analysis for Stationary Systems</td>
<td>29</td>
</tr>
<tr>
<td>2.4 Identifiability Analysis for Stationary Systems</td>
<td>37</td>
</tr>
<tr>
<td>2.4.1 Global identifiability</td>
<td>37</td>
</tr>
<tr>
<td>2.4.2 Local identifiability</td>
<td>47</td>
</tr>
<tr>
<td>2.5 Convergence Analysis for Time Invariant Systems with Deterministic Inputs</td>
<td>48</td>
</tr>
<tr>
<td>2.6 Identifiability Analysis for Time Invariant Systems with Deterministic Inputs</td>
<td>54</td>
</tr>
<tr>
<td>2.7 Asymptotic Normality</td>
<td>58</td>
</tr>
<tr>
<td>2.8 Summary</td>
<td>63</td>
</tr>
<tr>
<td>CHAPTER 3: IDENTIFICATION OF SYSTEMS UNDER FEEDBACK</td>
<td>64</td>
</tr>
<tr>
<td>3.1 Review of Some Examples</td>
<td>65</td>
</tr>
<tr>
<td>3.2 Overview of Standard Treatments</td>
<td>69</td>
</tr>
<tr>
<td>Table of Contents (continued)</td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td></td>
</tr>
<tr>
<td>3.3 LTI State Space Models Under Feedback</td>
<td>Page 75</td>
</tr>
<tr>
<td>3.4 Summary</td>
<td>Page 79</td>
</tr>
<tr>
<td>CHAPTER 4: EVALUATION OF THE ML ESTIMATE FOR LINEAR STATE SPACE MODELS</td>
<td></td>
</tr>
<tr>
<td>4.1 Gradient Evaluation</td>
<td></td>
</tr>
<tr>
<td>4.1.1 The forward filter</td>
<td>Page 80</td>
</tr>
<tr>
<td>4.1.2 The mixed filter</td>
<td>Page 86</td>
</tr>
<tr>
<td>4.1.3 The backward filter</td>
<td>Page 89</td>
</tr>
<tr>
<td>4.1.4 Gradient evaluation for time invariant systems</td>
<td>Page 90</td>
</tr>
<tr>
<td>4.2 Hessian Evaluation</td>
<td>Page 93</td>
</tr>
<tr>
<td>4.3 Information Matrix Evaluation</td>
<td>Page 98</td>
</tr>
<tr>
<td>CHAPTER 5: NUMERICAL EXAMPLES AND RESULTS</td>
<td></td>
</tr>
<tr>
<td>5.1 Description of Programs</td>
<td></td>
</tr>
<tr>
<td>5.1.1 ML parameter estimates</td>
<td>Page 104</td>
</tr>
<tr>
<td>5.1.2 Minimum information measure parameters</td>
<td>Page 112</td>
</tr>
<tr>
<td>5.2 Comparison of Forward and Mixed Filter Approaches</td>
<td>Page 113</td>
</tr>
<tr>
<td>5.3 Minimum Information Measure Parameters</td>
<td>Page 121</td>
</tr>
<tr>
<td>5.4 Convergence of the Scaled Likelihood Function</td>
<td>Page 124</td>
</tr>
<tr>
<td>5.5 Summary</td>
<td>Page 124</td>
</tr>
<tr>
<td>CHAPTER 6: SUMMARY AND SUGGESTIONS</td>
<td>Page 128</td>
</tr>
<tr>
<td>APPENDIX A</td>
<td>Page 130</td>
</tr>
<tr>
<td>APPENDIX B</td>
<td>Page 133</td>
</tr>
</tbody>
</table>
Table of Contents (continued)  

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPENDIX C</td>
<td>135</td>
</tr>
<tr>
<td>APPENDIX D</td>
<td>137</td>
</tr>
<tr>
<td>APPENDIX E</td>
<td>143</td>
</tr>
<tr>
<td>APPENDIX F</td>
<td>146</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>149</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>A Simple Closed Loop Configuration</td>
<td>66</td>
</tr>
<tr>
<td>3.2</td>
<td>A Simple Closed Loop Configuration</td>
<td>66</td>
</tr>
<tr>
<td>3.3</td>
<td>A General Closed Loop Configuration</td>
<td>71</td>
</tr>
<tr>
<td>4.1</td>
<td>Forward Filter Evaluation of $\frac{\partial x}{\partial \alpha}$</td>
<td>84</td>
</tr>
<tr>
<td>4.2</td>
<td>Mixed Filter Evaluation of $\frac{\partial x}{\partial \alpha}$</td>
<td>88</td>
</tr>
<tr>
<td>4.3</td>
<td>Backward Filter Evaluation of $\frac{\partial x}{\partial \alpha}$</td>
<td>91</td>
</tr>
<tr>
<td>4.4</td>
<td>Linear System for Information Matrix Element Computation</td>
<td>101</td>
</tr>
<tr>
<td>5.1</td>
<td>Flowchart Describing Subroutine CALCF</td>
<td>107</td>
</tr>
<tr>
<td>5.2</td>
<td>Flowchart Describing Subroutine CALCGF</td>
<td>108</td>
</tr>
<tr>
<td>5.3</td>
<td>Flowchart Describing Subroutine CALCGM</td>
<td>108</td>
</tr>
<tr>
<td>5.4</td>
<td>One Unknown Parameter in Second Order Model</td>
<td>116</td>
</tr>
<tr>
<td>5.5</td>
<td>Two Unknown Parameters in Second Order</td>
<td>117</td>
</tr>
<tr>
<td>5.6</td>
<td>Three Unknown Parameters in Second Order Model</td>
<td>118</td>
</tr>
<tr>
<td>5.7</td>
<td>Higher Order Models</td>
<td>120</td>
</tr>
<tr>
<td>5.8</td>
<td>Minimum Information Measure Parameters</td>
<td>123</td>
</tr>
<tr>
<td>5.9</td>
<td>Convergence of the Scaled Likelihood Function (*EP)</td>
<td>125</td>
</tr>
<tr>
<td>5.10</td>
<td>Convergence of the Scaled Likelihood Function (*EP)</td>
<td>126</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 General Discussion and Motivation

One of the most important problems in the study of systems (may they be physical or socio-economical) is that of obtaining an adequate mathematical model to describe in some way the behavior of the actual system. This is the goal of system identification which usually involves (see e.g. Schewpe [1]):

- Choosing a model structure with the help of known physical laws, a priori knowledge or sometimes intuitive relationships;
- Running identification experiments and collecting data from observations of the actual system. The data is used to find parameter values and select, from the specified set of models, the best model in some sense.

The outcome of such experiments will therefore depend on the system itself, the experimental conditions, the hypothesized model structure (or model set) and the identification method used.

Experimental conditions include the choice of the applied inputs (e.g. none, chosen in an open loop, determined partly by some output feedback process, etc...). Whenever it is possible to do so, choosing a good set of inputs to excite the system adequately is an important issue which has been studied by many authors (see e.g. Mehra [2] for
a survey, also Goodwin and Payne [3]). On the other hand, as was ob-
served by Åström and Eykhoff [4], the identification of systems under
closed loop control may present significant problems.

A model set $M$ can be viewed as a class of functions describing a
behavior of the system under study and parametrized in a certain manner
by a parameter vector $\alpha$. If to each parameter $\alpha$ in a set $\mathcal{P}$ there
corresponds a unique model in $M$ and vice versa, the terms model set
and parameter set can be used interchangeably and the model selection
problem can be viewed as a parameter estimation problem. This para-
metrization of multivariable systems is not, in general, an easy task
and the structural aspects of linear systems identification are still
an area of active study (see e.g. Clark [5]). There are also so-
called nonparametric techniques for the model selection problem (e.g.
correlation 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).

As for the various identification methods used to estimate para-
meter values, once a model set has been chosen, they generally consist
of minimizing some model dependent function of the data. 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 with an extensive
bibliography has been given by Åström and Eykhoff [4].
One main subject of this thesis is the maximum likelihood identification method in an environment of linear state space models. In particular, we will be assuming that the system under study can be adequately represented by a linear state space model and the selection of the "best" model will be done from linear state space model sets which do not necessarily contain the "true" model. This is felt to be an important subject for several reasons. On one hand, the maximum likelihood method is widely accepted as the best method if one is not computationally constrained, and turns out to have many desirable statistical properties. On the other hand, many control and estimation techniques are based on linear state space models. Furthermore, the special interest in model sets not necessarily containing the true model stems from the often encountered situation where lower order models are used. 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 it turns out, 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.

One objective of this thesis is to help, by a unified presentation, bridge a gap that is felt to exist between practical state space,
maximum likelihood identification techniques and some recent theoretical results in parameter identification.

1.2 Background

The study of parameter estimation techniques started with the introduction of the least squares criterion by Gauss [6] in 1809. The maximum likelihood method (ML) was introduced by Fisher [7] in 1922 and has since been widely used for parameter estimation due to its intuitive appeal and its good asymptotic properties mentioned above. For a long time after its introduction, the ML method had been mainly a tool of traditional statistics. State variable techniques were first used to compute likelihood functions by Schweppe [8] in the context of signal detection; and the first application of the ML method to system identification is due to Åström and Bohlin [9] who considered, in 1965, a single input single output system application.

Since then, the problem of computing ML parameter estimates in the context of system identification has been studied by many authors. Among those are e.g. Box and Jenkins [10], Kashyap [11] who discussed ML identification of stochastic linear systems, but not considering state space models. The latter also presented a method for computing the gradients of the likelihood function using the theory of adjoint equations in order to save on computations. Mehra [12] derived a set of equations for the ML identification of linear dynamic systems using a Kalman filter representation; computational aspects of ML
identification in connection with state space models were further considered by e.g. Schweppe [1,13], Maybeck [14], Gupta and Mehra [15], Gupta [16], Sandell and Yared [17], Best [18], and Goodrich and Caines [19]. An interesting approach to the likelihood function minimization, the parallel ML identification algorithm, has also been used for aeronautical applications by Stein et al. [20].

However, the success of the various algorithms used depended on, among many factors, the consistency property of ML estimates. Wald [21] established this property for the case of independent identically distributed observations by using the strong law of large numbers. His result was extended by Roussas [22] to the case of ergodic Markov observation sequences by using the ergodic theorem. The consistency of various parameter estimation techniques has also been actively studied in the context of system identification but many proofs of strong consistency (i.e. convergence w.p.1) were not complete. Whereas pointwise convergence only of the associated probability densities on the parameter set was sufficient for consistent estimates on finite parameter sets, the need for uniform convergence in the case of infinite parameter sets was missed. It seems that the first complete strong consistency proofs for application to system identification were given by Rissanen and Caines [23] and Ljung [26].

Using a version of the law of large numbers, Ljung [26,27,28,29] proved strong consistency for a general class of prediction error
techniques in a general system's innovations representation. No assumptions of ergodicity or stationarity of the stochastic processes were made, thus allowing quite general models including time varying feedback. In [28] and [29] Ljung discussed in broad terms the case of the model set not containing the true system. In [23] and [24] Caines and Rissanen proved strong consistency of the ML identification method for stationary Gaussian processes using the ergodic theorem. This approach was then generalized by Caines [25] to the consistency proof for stationary processes of the more general class of prediction error techniques (which he shows to include the ML technique for the case of stationary Gaussian observation sequences). Asymptotic normality and efficiency of general prediction error estimates were also discussed by Caines and Ljung [30,31]. The work of Hannan and Dunsmuir [32,33,34,35] on the identification of ARMA models is also of particular relevance here.

Finally, the convergence of maximum likelihood parameter estimates was studied in an explicit state space framework and for the case of the model set not containing the true system by Baram and Sandell [36]. Here, information theoretic concepts were used to help formulating the results but the convergence analysis was restricted to finite model sets, and models driven by plant noise only.

As it turns out, one important issue connected with the consistency of parameter estimates is that of identifiability of parameters, and
as mentioned above this is still an area of active study. Identifiability has been defined (see e.g. Åström and Bohlin [9], Ljung et al. [46], Tse and Anton [37]) in terms of conditions related to consistency of parameter estimates and the maxima of the likelihood functions. The latter have been investigated by e.g. Åström and Söderström [38] and Söderström [39] for autoregressive moving average type models and Bar-Shalom [40] in the context of simultaneous state and parameter estimation. Now it is well known that, given any state space model in general, there are many equivalent realizations of a particular input output response; and a standard approach to overcome this nonuniqueness problem is the use of canonical forms (e.g. Glover [41], Mehra [42] among many others). However canonical forms have the disadvantage of losing the physical interpretation of the original parameters. Furthermore, to obtain canonical forms, certain structural indices (such as the Kronecker invariants) must also be determined (see e.g. Caines [43], Glover and Willems [44], Tse and Wienert [45]). Therefore, attention will be directed to situations where one can incorporate one's a priori knowledge in forming models with fewer unknown parameters. Glover [41] has also addressed the identifiability question in this context and has derived conditions for local, global and partial identifiability assuming that the transfer functions can be asymptotically identified.
As mentioned previously, systems operating in closed loop might present special identifiability problems. An example given by Åström and Eykhoff [4] shows how identification methods based on correlation analysis can give the estimate of the open loop transfer function as being the inverse of the transfer function of the feedback. Another example given by e.g. Ljung et al. [46] shows how certain experimental conditions do not guarantee identifiability for certain model structures. Many such problems have been discussed in the literature, mostly concerning linear models, and the solutions proposed include the use of nonlinear feedback (e.g. Fisher [47]); shifting between an appropriate number of linear regulators (e.g. Ljung et al. [46], Söderström et al. [48]); addition of a small perturbation in the input (e.g. Saridis and Lobbia [49]); addition of noise in the regulator (e.g. Vorchik et al. [50]). But most studies do not consider multiple input multiple output state space models. Glover [41] applies the identifiability tests mentioned above to such models by expressing the closed loop model in terms of the open loop parameters and the parameters of the feedback law. An extensive survey of identifiability issues concerning processes in closed loop has been given by Gustavsson et al. [51].

Finally, other problems, connected with the identification of systems under feedback, arise when the identifier is part of an online adaptive control scheme. Some of those problems are overall.
stability and convergence of the regulator, and have been discussed by various authors in the context of specific applications (e.g. aeronautical applications with Stein et al. [20], Athans et al. [52] and Greene [53]; industrial processes and self-tuning regulators with Aström et al. [54]).

1.3 Contents and Contribution

The first part of this thesis, containing chapters 2 and 3, discusses properties of the maximum likelihood method as applied to state space models. In Chapter 2, a formulation of the identification problem is first given. Then asymptotic properties (convergence, normality) of the ML method are analyzed using the methods of Caines [25] and Ljung [31]. The contributions here are a generalization of Baram and Sandell's [36] convergence analysis for stationary models to continuous model sets, and the interpretation of other theoretical results from the viewpoint of state space models. In particular some additional insights into the nature of the asymptotic models selected by the maximum likelihood method are given. Chapter 3 reviews the particular problem of identifying open loop dynamics from data collected under closed loop operation and reexamines that problem in the light of the results of Chapter 2.

The second part of the thesis turns to the practical aspects of computing the maximum likelihood estimates of parameters of linear Gaussian state space models. Chapter 4 discusses different ways of
evaluating the gradient and Hessian of the likelihood function as well as the information matrix. The contributions here are the evaluation of the gradient and Hessian of the likelihood function using adjoint computations which present significant computational savings. In Chapter 5, some examples are run on computer simulated data to illustrate two main points: the computational savings obtained from the use of adjoint computations and the nature of asymptotic models selected by the ML method from a set of lower order models.

Chapter 6 contains a summary of the thesis as well as suggestions for further research.
CHAPTER 2
ASYMPTOTIC PROPERTIES OF ML IDENTIFICATION FOR LINEAR STATE SPACE MODELS

Asymptotic properties are an important factor in the study of parametric identification methods. They might affect the choice of a particular method as well as the amount of data to be collected when setting up an identification experiment. In this chapter, convergence and other asymptotic properties of ML identification will be studied in the context of linear state space model sets.

As mentioned previously, one is usually in a situation where the true system is not a member of the model set in use. Therefore, Baram and Sandell's [36] information measure is first briefly reviewed as a tool for comparing different members of a model set. Then, the ML identification problem is formulated in the context of linear state space models. The convergence analysis of [36] for stationary Gaussian linear systems is then generalized to continuous model sets, and the identifiability of unknown parameters is discussed. The same analysis is also extended to linear time invariant systems driven by deterministic inputs. Finally, asymptotic normality of the ML estimates is discussed.

2.1 The Information Distance Between Models

This part is a brief recapitulation of some of the results of
In general, one does not have independent observations when performing a system identification experiment. A standard general model for such a sequence of observations \{z(t)\} is therefore their joint probability distribution function. Let \( Z^t \overset{\Delta}{=} (z(1), \ldots, z(t)) \) and \( P \) be a parameter set such that to each \( \alpha \in P \) there corresponds a conditional probability density function \( f_\alpha(z(t)|Z^{t-1}) \) for each \( t \geq 1 \) with \( f_\alpha(z(1)|Z^{0}) = f_\alpha(z(1)) \). Let \( f_*(z(t)|Z^{t-1}) \) be the correct sequence of conditional probability densities of \{z(t)\}. We call \(*\) the true parameter. Note that \(*\) does not necessarily belong to \( P \).

Let \( Q = \{*\} \cup P \). For each \( \alpha \in Q \) we have \( f_\alpha(Z^t) = \prod_{t=1}^{t} f_\alpha(z(t)|Z^{t-1}) \).

For each pair \( \alpha_1, \alpha_2 \in P \) we call

\[
\frac{f_{\alpha_1}(z^t)}{f_{\alpha_2}(z^t)} \quad \text{and} \quad \frac{f_{\alpha_1}(z(t)|Z^{t-1})}{f_{\alpha_2}(z(t)|Z^{t-1})}
\]

the likelihood ratio and the conditional likelihood ratio, respectively, between \( \alpha_1 \) and \( \alpha_2 \).

If for some pair of parameters \( \alpha_1, \alpha_2 \in Q \)

\[
f_{\alpha_1}(Z^t) > f_{\alpha_2}(Z^t)
\]

or, equivalently,

\[
\log f_{\alpha_1}(Z^t) > \log f_{\alpha_2}(Z^t)
\]
we say that the parameter $\alpha_1$ is favored over the parameter $\alpha_2$ by the observations $Z^t$. Then $\log f_{\alpha_1}(Z^t)$ may be regarded as a measure of the information in $Z^t$ for selecting a parameter from the set $\mathbb{Q}$. The difference

$$
\log f_{\alpha_1}(Z^t) - \log f_{\alpha_2}(Z^t) = \log h_{\alpha_2}(Z^t)
$$

(2.1)

is then a measure of the information in $Z^t$ for selecting between $\alpha_1$ and $\alpha_2$. If (2.1) is positive then $\alpha_1$ is favored and if it is negative then $\alpha_2$ is favored. The difference

$$
\log h_{\alpha_2}(Z^t) - \log h_{\alpha_2}(Z^{t-1}) = \log h_{\alpha_2}(z(t) | Z^{t-1})
$$

(2.2)

is then a measure of the difference between the information favoring $\alpha_1$ against $\alpha_2$ at instant $t$ and the information favoring $\alpha_1$ against $\alpha_2$ at instant $t-1$. It can then be regarded as a measure of the information favoring $\alpha_1$ against $\alpha_2$ in the observation $z(t)$. We define

$$
I_t(\alpha_1; \alpha_2) = \Delta \{ \log h_{\alpha_2}(z(t) | Z^{t-1}) \}
$$

(2.3)

as the mean information in $z(t)$ favoring $\alpha_1$ against $\alpha_2$.

We now state some properties of the information measure defined above that will prove to be useful in the following analysis. These properties follow almost immediately from the definition.
For any $a \in \mathcal{P}$ and for each $t \geq 0$ we have

$$I_t(a; x) \geq 0$$

with equality if and only if $f_a(z(t)|z_{t-1}^t) = f_a(z(t)|z_{t-1}^t)$ a.s.

Consequently, if $a_\ast \in \mathcal{P}$ is the true parameter, then for any $a' \in \mathcal{P}$

$I_t(a; a')$ is maximized on $\mathcal{P}$ at $a = a_\ast$. This maximum is unique unless

for some $a \in \mathcal{P}$ $f_a(z(t)|z_{t-1}^t) = f_{a_\ast}(z(t)|z_{t-1}^t)$ a.s. We also have

Theorem 2.1

Let $\mathcal{Q} = \{\ast\} \cup \mathcal{P}$. The sequence $|I_t(a_1; a_2)|; a_1, a_2 \in \mathcal{P}$ is a

sequence of pseudo metrics on $\mathcal{Q}$. It is a sequence of metrics if and

only if $I_t(a_1; a_2) = 0$ implies $a_1 = a_2$.

Proof: See [36].

A direct consequence of Theorem 2.1 is that if $a_1$ and $a_2$ are

any pair of parameters in the parameter space $\mathcal{P}$ then $a_1$ is closer to

the true parameter $\ast$ than $a_2$ in the metric $|I_t(\cdot; \cdot)|$ if and only if

$I_t(a_1; a_2) > 0$.

Note that the information measure defined by (2.3) can only

be computed if the true parameter (or equivalently the true probabil-

ity density) is known. It is nevertheless useful in the analysis

of the asymptotic behavior of identification procedures as we shall

see in the sequel.
2.2 Problem Formulation

Consider the linear discrete time system of the form

\[
x(t+1) = A_x(t)x(t) + B_x(t)u(t) + I_x(t)\xi(t) \tag{2.5}
\]

\[
z(t) = C_x(t)x(t) + \theta(t) \tag{2.6}
\]

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

\( x(t) \in \mathbb{R}^n \) the state vector (nonwhite stochastic sequence)

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

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

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

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

with

\[
E\{x(0)\} = 0 \tag{2.7}
\]

\[
E\{x(0)x'(0)\} = \Sigma_x(0); \quad \Sigma_x(0) = \Sigma_x'(0) \geq 0 \tag{2.8}
\]

and where \( \{\xi(t)\} \) and \( \{\theta(t)\} \) are uncorrelated and mutually uncorrelated Gaussian sequences, independent of \( x(0) \), with

\[
E\{\xi(0)\} = E\{\theta(0)\} = 0 \tag{2.9}
\]

\[
E\{\xi(t)\xi'(t)\} = \Sigma_x(t); \quad \Sigma_x(t) = \Sigma_x'(t) \geq 0 \tag{2.10}
\]

\[
E\{\theta(t)\theta'(t)\} = \Theta_x(t); \quad \Theta_x(t) = \Theta_x'(t) \geq 0 \tag{2.11}
\]
Let

\[ M = \{(A_\alpha, B_\alpha, C_\alpha, L_\alpha, \Xi_\alpha, \Theta_\alpha, \Sigma(0)); \alpha \in \mathcal{P}\} \]  

(2.12)

be a model set of the same structure as (2.5)-(2.11) but possibly of different dimension (e.g. \( n_\alpha \neq n_{\alpha^*} \)). The matrices in \( M \) depend on the unknown parameter \( \alpha \) (e.g. the vector of unknown entries in the matrices) which ranges over a compact subset \( \mathcal{P} \) of \( \mathbb{R}^d \).

Let

\[ \hat{z}(t;\alpha) = E_{\alpha}\{z(t) \mid Z_{t-1}\} ; \quad \alpha \in \{\} \cup \mathcal{P} \]  

(2.13)

denote the one step least square prediction of \( z(t) \) given all the past measurements \( Z_{t-1} \triangleq \{z(0), z(1),...,z(t-1)\} \) assuming that \( \alpha \) is the true parameter; and let

\[ S_{\alpha}(t) = E_{\alpha} \{[z(t) - \hat{z}(t;\alpha)][z(t) - \hat{z}(t;\alpha)]'\} \]  

(2.14)

denote the corresponding predicted observation error covariance matrix. It is well known (e.g. [1],[55]) that those quantities can be obtained by the Kalman filter corresponding to \( M(\alpha) \):

\[ \hat{x}(t+1;\alpha) = A_\alpha(t)\hat{x}(t;\alpha) + B_\alpha(t)u(t) + A_\alpha(t)H_\alpha(t)r(t;\alpha) \]  

(2.15)

\[ r(t;\alpha) = z(t) - \hat{z}(t;\alpha) = z(t) - C_\alpha(t)\hat{x}(t;\alpha) \]  

(2.16)

where

\[ H_\alpha(t) = C_\alpha(t)C_\alpha(t)'S_{\alpha}^{-1}(t) \]  

(2.17)
\[
S_{\alpha}(t) = C_{\alpha}(t)\Sigma_{\alpha}(t)C_{\alpha}(t) + \Theta_{\alpha}(t) \tag{2.18}
\]
\[
\Sigma_{\alpha}(t+1) = A_{\alpha}(t)\Sigma_{\alpha}(t)A_{\alpha}(t) + L_{\alpha}(t)\Sigma_{\alpha}(t)L_{\alpha}(t) - A_{\alpha}(t)H_{\alpha}(t)S_{\alpha}(t)H_{\alpha}^t(t)A_{\alpha}(t) \tag{2.19}
\]

and where, under the Gaussian assumption, the residuals \( r(t;\alpha) \) form an independent, zero mean, Gaussian sequence with respect to the probability measure corresponding to \( \alpha \) (In particular, \( E_{\alpha}\{r(t;\alpha)\} = 0 \)).

Therefore, under the same assumptions, the conditional probability density function of the measurement \( z(t) \) given \( Z_t \) and corresponding to the model \( M(\alpha) \), \( \alpha \in \{\ast\} \cup P \) is

\[
f_{\alpha}(z(t)|Z^{t-1}) = \frac{1}{(2\pi)^{\frac{K}{2}}(\text{det}[S_{\alpha}(t)])^{\frac{1}{2}}} e^{\frac{1}{2} r'(t;\alpha)S_{\alpha}^{-1}(t)r(t;\alpha)} \tag{2.20}
\]

and the maximum likelihood estimate \( \hat{\alpha}_t \) of \( \alpha \) at time \( t \) is that value which minimizes the negative log likelihood function (see e.g. [17]):

\[
\zeta(z^t;\alpha) = \sum_{\tau=0}^{t} \zeta(z(\tau)|Z^{\tau-1};\alpha) \tag{2.21}
\]

where

\[
\zeta(z(\tau)|Z^{\tau-1};\alpha) \triangleq \frac{1}{2} \log\text{det}[S_{\alpha}(\tau)] + \frac{1}{2} r'(\tau;\alpha)S_{\alpha}^{-1}(\tau)r(\tau;\alpha) \tag{2.22}
\]
In contrast to $S_\alpha(t)$ defined in (2.14) let

$$S^*_\alpha(t) \triangleq E_*\{x(t;\alpha)x'(t;\alpha)\} \tag{2.23}$$

denote the second moment of the observation error of the predictor corresponding to $\alpha$ when in fact the model corresponding to $*$ is the correct one. This quantity can be generated by solving the covariance equation for the $(n_\alpha+n_\alpha)$ linear model obtained by composing the "true" model with the Kalman filter corresponding to $\mathcal{M}(\alpha)$ as follows. Define

$$A^*_\alpha(t) \triangleq \begin{bmatrix} A_*(t) & 0 \\ A_\alpha(t)H_\alpha(t)C_\alpha(t) & A_\alpha(t)[I-H_\alpha(t)C_\alpha(t)] \end{bmatrix} \tag{2.24}$$

$$B^*_\alpha(t) \triangleq \begin{bmatrix} B_\alpha(t) \\ B_\alpha(t) \end{bmatrix} \tag{2.25}$$

$$L^*_\alpha(t) \triangleq \begin{bmatrix} L_*(t) & 0 \\ 0 & A_\alpha(t)H_\alpha(t) \end{bmatrix} \tag{2.26}$$

$$C^*_\alpha(t) \triangleq [C_*(t) ; -C_\alpha(t)] \tag{2.27}$$

$$\Xi^*_\alpha(t) \triangleq \begin{bmatrix} \Xi_*(t) & 0 \\ 0 & \Theta_\alpha(t) \end{bmatrix} \tag{2.28}$$

$$\hat{x}^*_\alpha(t) \triangleq \begin{bmatrix} \hat{x}(t;\alpha) \\ \hat{x}(t;\alpha) \end{bmatrix} \tag{2.29}$$
Then from (2.5), (2.6), (2.15) and (2.16)

\[ x^*_\alpha(t+1) = A^*_\alpha(t)x^*_\alpha(t) + B^*_\alpha(t)u(t) + L^*_\alpha(t) \begin{bmatrix} \xi(t) \\ \theta(t) \end{bmatrix} \] (2.30)

\[ r(t; \alpha) = C^*_\alpha(t)x^*_\alpha(t) + \theta(t) \] (2.31)

If we now let \( x^*_\alpha(t) \) and \( \Sigma^*_\alpha(t) \) respectively be the \((n_1+n_2)\) mean vector and the \((n_1+n_2) x (n_1+n_2)\) covariance matrix of \( x^*_\alpha(t) \) with respect to the true probability measure,

\[ x^*_\alpha(t+1) = A^*_\alpha(t)x^*_\alpha(t) + B^*_\alpha(t)u(t) \]

\[ \Sigma^*_\alpha(t+1) = A^*_\alpha(t)\Sigma^*_\alpha(t)A^*_\alpha(t) + L^*_\alpha(t) \Sigma^*_\alpha(t)L^*_\alpha(t) \] (2.32)

and

\[ S^*_\alpha(t) = C^*_\alpha(t)\Sigma^*_\alpha(t)C^*_\alpha(t) + \theta^* + C^*_\alpha(t)x^*_\alpha(t)x^*_\alpha(t)C^*_\alpha(t) \] (2.33)

Therefore, the expected value of the conditional likelihood function (2.22) with respect to the true probability measure is

\[ I^*_\alpha(t) \triangleq E_\alpha[\xi(z(t)|z^{t-1}, \alpha)] \] (2.34)

\[ = \frac{1}{2} \{ \log\det[S^*_\alpha(t)] + \text{tr}[S^{-1}_\alpha(t)S^*_\alpha(t)] \} \] (2.35)

and the information measure defined in section 2.1 between two models \( M(\alpha_1) \) and \( M(\alpha_2) \) is
Furthermore, in view of the remark following Theorem 2.1, we will consider $M(\alpha_1)$ closer to the true system than $M(\alpha_2)$ in the metric defined above if and only if $I_{\alpha_1}^*(t) < I_{\alpha_2}^*(t)$.

Now all 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 $H_\alpha(t)$ and residual covariance matrix $S_\alpha(t)$. However, under the assumptions of detectability and stabilizability of the model $M(\alpha)$ (see, e.g. [56]), the Riccati equation (2.19) has a limiting solution as $t \to \infty$ which in turn implies that $H_\alpha(t)$ and $S_\alpha(t)$ become constant. A very common practice in the case of time invariant systems is to use this steady state Kalman filter. This greatly simplifies the calculation of the likelihood function (2.21)-(2.22). Indeed, equation (2.19) can be then replaced by the algebraic Riccati equation

$$
\Sigma_\alpha = A_\alpha \Sigma_\alpha A_\alpha + L_\alpha \Sigma_\alpha L_\alpha - A_\alpha \Sigma_\alpha C_\alpha (C_\alpha \Sigma_\alpha C_\alpha + \Theta_\alpha)^{-1} C_\alpha \Sigma_\alpha A_\alpha
$$

which can be solved by methods more efficient than the iteration of (2.19) to steady state. Furthermore, the observations are now processed by a time-invariant filter. This obviously 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.

The same remarks can be also made about $\Sigma^*(t)$. It can be deduced from equation (2.32) that $E^*(t)$ has a steady state limit $\Sigma^*$ if $A^*$ (now a constant matrix under the above time invariant assumptions) is stable. This follows if both the true model and the Kalman filter corresponding to $\alpha$ are stable. Note however that

$$ S^*_\alpha(t) = C^*_\alpha E^*_\alpha C^*_\alpha + \Theta^* + C^*_\alpha \overline{x}^\alpha(t)\overline{x}^\alpha(t)C^*_\alpha $$

(2.38)

remains a time varying matrix due to the non zero mean $\overline{x}^\alpha(t)$.

In the more particular case where $u(t) = 0$, $x^\alpha(t) \sim \mathcal{N}(0, S^*_\alpha(t))$ and

$$ S^*_\alpha(t) = S^*_\alpha = C^*_\alpha E^*_\alpha C^*_\alpha + \Theta^* $$

(2.39)

This also makes

$$ I^*_{\alpha} = \frac{1}{2} \{ \log(\det[S^*_\alpha]) + \text{tr}[S^{-1}_\alpha S^*_\alpha] \} $$

(2.40)

and

$$ I(\alpha_1; \alpha_2) = I^*_{\alpha_2} - I^*_{\alpha_1} $$

(2.41)

time invariant functions of $\alpha$. 
2.3 Convergence Analysis for Stationary Systems

In this as well as the following section we restrict our attention to time invariant systems where \( u(t) = 0 \). We will also assume that the noise processes are stationary and ergodic, that each model \( M(\alpha) \) is stable, controllable and observable and that, as discussed above, the steady state filter is used to compute the likelihood function.

We now study the convergence of the ML estimates and its conditions via the following two lemmas:

**Lemma 2.1**

Consider the conditions:

\[ C_1: \zeta(z(t) | z_{t-1}; \alpha) \text{ is differentiable with respect to } \alpha (\forall \alpha \in P) \]

\[ \zeta(z(t) | z_{t-1}; \alpha) \text{ and } \frac{\partial}{\partial \alpha} \zeta(z(t) | z_{t-1}; \alpha) \text{ are stationary ergodic processes (} \forall \alpha \in P). \]

\[ C_2: \forall \alpha \in P, \forall t \left| \frac{\partial}{\partial \alpha} \zeta(z(t) | z_{t-1}; \alpha) \right| \leq k(z_t) \]

where \( k(\cdot) \) is a time invariant function measurable with respect to the \( \sigma \)-field generated by \( z^t \) such that \( E_x[k(z_t)] < K < \infty \).

Then, under \( C_1 \) and \( C_2 \),

\[
\lim_{t \to \infty} \frac{\alpha_2}{\alpha_1} (\mathbf{z}) = 0 \quad \text{a.e., uniformly in } \alpha_1, \alpha_2 \in P \tag{2.42}
\]

if \( I_{\alpha_1}^* < I_{\alpha_2}^* \) (i.e. \( I(\alpha_1; \alpha_2) > 0 \)) and only if \( I_{\alpha_1}^* \leq I_{\alpha_2}^* \).

(note that we mean by (2.42) the following:)
∀a₁ ∈ P, ∀ε > 0, ∀ρ > 0, ∃T(a₁, ε, ρ)s.t.

\[ h_{a₁}^{α₂}(z^t) < ε \quad ∀t > T(a₁, ε, ρ), \quad ∀α₂ ∈ P - B(a₁, ρ) \]

**Proof:**

\[
\log \frac{f_{α₂}(z^t)}{f_{α₁}(z^t)} = \log \sum_{s=0}^{t} \frac{f_{α₂}(z(s) | z^{s-1})}{f_{α₁}(z(s) | z^{s-1})}
\]

\[
= \sum_{s=0}^{t} [ζ(z(s) | z^{s-1}; α₁) - ζ(z(s) | z^{s-1}; α₂)]
\]

Define, for simplicity of notation,

\[
J_t(α, z^{t-1}) = \frac{1}{t+1} \sum_{s=0}^{t} ζ(z(s) | z^{s-1}; α) . \quad (2.43)
\]

Since ζ(z(s) | z^{s-1}; α) is assumed to be a stationary ergodic process (C1) it follows form the ergodic theorem (see e.g. [57]) that,

∀α ∈ P

\[
\lim_{t→∞} J_t(α, z^{t-1}) = E_α^* [ζ(z(s) | z^{s-1}; α)] = I^*_α \quad a.e.
\]

We now follow a similar analysis by Caines [25] in proving that this convergence is uniform in α ∈ P. Note, however, that here expectations are taken with respect to the true probability measure.
For the convergence to be uniform in $\alpha \in \mathcal{P}$ it is sufficient to show that $\forall \alpha \in \mathcal{P}, \forall \varepsilon > 0, \exists$ an open neighborhood $N_{1}(\alpha, \varepsilon)$ s.t.

$$\forall \alpha' \in N_{1}(\alpha, \varepsilon), \exists T_{1}(\alpha, \varepsilon, z), \text{ s.t. } \forall t_{1}, t_{2} > T_{1}(\alpha, \varepsilon, z)$$

$$| J_{t_{1}^{-1}}(\alpha', z^{1}) - J_{t_{2}^{-1}}(\alpha', z^{2})| < \varepsilon$$

(2.44)

where $z$ denotes the entire realized observation sample. Using the differentiability assumption (Cl) on $\zeta(z(t)|z^{-1}; \alpha)$ and the mean value theorem we obtain, for some neighborhood $N_{2}(\alpha)$ of $\alpha$

$$\forall \alpha' \in N_{2}(\alpha), \exists \alpha'' \in N_{2}(\alpha) \text{ s.t.}$$

$$| J_{t_{1}^{-1}}(\alpha', z^{1}) - J_{t_{2}^{-1}}(\alpha', z^{2})| \leq | J_{t_{1}^{-1}}(\alpha, z^{1}) - J_{t_{2}^{-1}}(\alpha, z^{2})|$$

$$+ \left| \frac{\partial J_{t_{1}}}{\partial \alpha} (\alpha'', z^{1}) - \frac{\partial J_{t_{2}}}{\partial \alpha} (\alpha'', z^{2}) \right| \alpha' - \alpha'' \right|$$

(2.45)

Now, using again (Cl) and the ergodic theorem

$$\lim_{\min(t_{1}, t_{2}) \to \infty} | J_{t_{1}^{-1}}(\alpha, z^{1}) - J_{t_{2}^{-1}}(\alpha, z^{2})| = 0 \text{ a.e.}$$

and this convergence is independent of $\alpha' \in N_{2}(\alpha)$, i.e.,

$$\forall \varepsilon > 0, \forall \alpha' \in N_{2}(\alpha), \exists T_{2}(\alpha, \varepsilon, z) \text{ s.t. } \forall(t_{1}, t_{2}) > T_{2}(\alpha, \varepsilon, z)$$

$$| J_{t_{1}^{-1}}(\alpha, z^{1}) - J_{t_{2}^{-1}}(\alpha, z^{2})| < \varepsilon/2$$

(2.46)
As for the second term on the r.h.s. of the inequality in (2.45)

\[ \left| \frac{\partial J}{\partial \alpha} (\alpha'', z^{t-1}) \right| \leq \frac{1}{t+1} \sum_{s=0}^{t} \left| \frac{\partial J}{\partial \alpha} (z(s)|z^{s-1}, \alpha'') \right| \]

and, in view of C2, and using the ergodic theorem again,

\[ \forall \alpha'' \in N_2(\alpha), \exists T_3(z) \text{ s.t. } \forall t > T_3(z) \]

\[ \left| \frac{\partial J}{\partial \alpha} (\alpha'', z^{t-1}) \right| < 2 E_4\{k(z^0)\} \]

So, taking \( K = 4E_4\{k(z^0)\} \)

\[ \forall \alpha'' \in N_2(\alpha), \exists T_3(z) \text{ s.t. } \forall t_1, t_2 > T_3(z) \]

\[ \left| \frac{\partial J}{\partial \alpha} (\alpha'', z^{t_1-1}) - \frac{\partial J}{\partial \alpha} (\alpha'', z^{t_2-1}) \right| < K \] (2.47)

Now take \( N_3(\alpha, \varepsilon) \) s.t. \( ||\alpha - \alpha'|| < \frac{\varepsilon}{2K} \) \( \forall \alpha' \in N_3(\alpha, \varepsilon) \) and define \( N_1(\alpha, \varepsilon) \) as \( N_2(\alpha) \cap N_3(\alpha, \varepsilon) \) and \( T_1(\alpha, \varepsilon, z) \) as \( \max(T_2(\alpha, \varepsilon, z), T_3(z)) \).

Then (2.44) follows from (2.46) and (2.47).

Now (2.44) implies that

\[ \lim_{t \to \infty} \frac{1}{t+1} \sum_{s=0}^{t} \log \frac{f_{\alpha_2}(z(s)|z^{s-1})}{f_{\alpha_1}(z(s)|z^{s-1})} = (I_{\alpha_1}^* - I_{\alpha_2}^*) \text{ a.e.} \] (2.48)

uniformly in \( \alpha_1, \alpha_2 \in \mathcal{P} \). But if \( I_{\alpha_1}^* < I_{\alpha_2}^* \), then
\[
\lim_{t \to \infty} \sum_{s=0}^{t} \log \frac{f_{\alpha_2}(z(s) | z^{s-1})}{f_{\alpha_1}(z(s) | z^{s-1})} = \lim_{t \to \infty} \log \frac{f_{\alpha_2}(z^t)}{f_{\alpha_1}(z^t)} = -\infty
\]

uniformly in $\alpha_1, \alpha_2 \in P$, which in turn implies

\[
\lim_{t \to \infty} \frac{f_{\alpha_2}(z^t)}{f_{\alpha_1}(z^t)} = 0 \quad \text{a.e.}
\]

uniformly in $\alpha_1, \alpha_2 \in P$.

To prove the necessary condition, suppose that (2.42) holds but

\[I_1^* > I_2^*,\]

then, in view of (2.48)

\[
\lim_{t \to \infty} \sum_{s=0}^{t} \log \frac{f_{\alpha_2}(z(s) | z^{s-1})}{f_{\alpha_1}(z(s) | z^{s-1})} = \lim_{t \to \infty} \log \frac{f_{\alpha_2}(z^t)}{f_{\alpha_1}(z^t)} = \infty
\]

implying

\[
\lim_{t \to \infty} \frac{f_{\alpha_2}(z^t)}{f_{\alpha_1}(z^t)} = \infty
\]

which contradicts (2.42).

\[\square\]

Note that conditions C1 and C2 are similar to those given by Caines [25] and are used to make the convergence in (2.42) uniform in the parameter space. They are in effect the extra conditions required to generalize Baram and Sandell's [36] corresponding results.
for finite parameter sets to continuous compact subsets of $\mathbb{R}$. Now those conditions follow from the assumptions made in the beginning of this chapter. Indeed, let $\alpha_i$ denote the $i$th component of $\alpha$ and differentiate $\zeta(z(t)|z^{t-1};\alpha)$ defined in (2.22):

$$\begin{align*}
\frac{\partial \zeta(z(t)|z^{t-1};\alpha)}{\partial \alpha_i} &= r'(t;\alpha) s^{-1}_{\alpha} \frac{\partial r(t;\alpha)}{\partial \alpha_i} \\
&\quad - \frac{1}{2} r'(t;\alpha) s^{-1}_{\alpha} \frac{\partial s_{\alpha}}{\partial \alpha_i} s^{-1}_{\alpha} r(t;\alpha) \\
&\quad + \frac{1}{2} \text{tr} \left[ s^{-1}_{\alpha} \frac{\partial s_{\alpha}}{\partial \alpha_i} \right] \\
&\quad + \frac{1}{2} \text{tr} \left[ s^{-1}_{\alpha} \frac{\partial s_{\alpha}}{\partial \alpha_i} \right] \tag{2.49}
\end{align*}$$

From the controllability and observability of each model $M(\alpha)$, $S^{-1}_{\alpha}$ exists ([55]). Furthermore, $\frac{\partial r(t;\alpha)}{\partial \alpha_i}$ and $\frac{\partial s_{\alpha}}{\partial \alpha_i}$ are generated by differentiating the Kalman filter equations (2.15)-(2.19) (see [17] for details):

$$\begin{align*}
\frac{\partial r(t;\alpha)}{\partial \alpha_i} &= - \frac{\partial C_{\alpha}}{\partial \alpha_i} \hat{x}(t;\alpha) - C_{\alpha} \frac{\partial \hat{x}(t;\alpha)}{\partial \alpha_i} \tag{2.50} \\
\frac{\partial \hat{x}(t+1;\alpha)}{\partial \alpha_i} &= \frac{\partial A_{\alpha}}{\partial \alpha_i} \hat{x}(t;\alpha) + \frac{\partial C_{\alpha}}{\partial \alpha_i} \hat{y}(t;\alpha) + \frac{\partial A_{\alpha} H_{\alpha}}{\partial \alpha_i} r(t;\alpha) \tag{2.51} \\
\frac{\partial H_{\alpha}}{\partial \alpha_i} &= \left[ \frac{\partial E_{\alpha}}{\partial \alpha_i} C_{\alpha} + \sum_{\alpha} \frac{\partial C_{\alpha}}{\partial \alpha_i} - H_{\alpha} \frac{\partial s_{\alpha}}{\partial \alpha_i} \right] S^{-1}_{\alpha} \tag{2.53}
\end{align*}$$
\[
\frac{\partial S}{\partial \alpha_i} = \frac{\partial \zeta}{\partial \alpha_i} \frac{\partial \alpha}{\partial \alpha_i} c'_\alpha + \frac{\partial \varphi}{\partial \alpha_i} \Sigma \alpha c'_\alpha + c \Sigma \alpha c'_\alpha + \frac{\partial c}{\partial \alpha_i} \Sigma \alpha c'_\alpha + \frac{\partial c'}{\partial \alpha_i} \Sigma \alpha c'_\alpha
\] (2.54)

\[
\frac{\partial \Sigma}{\partial \alpha_i} = \frac{\partial \Sigma}{\partial \alpha_i} \Sigma = \frac{\partial A}{\partial \alpha_i} - \frac{\partial A}{\partial \alpha_i} + \Omega_i(\alpha) + \Omega_i'(\alpha)
\] (2.55)

\[
\Omega_i(\alpha) = \left[ \frac{\partial \alpha}{\partial \alpha_i} - \frac{\partial c}{\partial \alpha_i} \right] \Sigma \alpha \alpha + \frac{1}{2} \frac{\partial (L' \Sigma L')}{\partial \alpha_i} + \frac{1}{2} \frac{\partial \Sigma}{\partial \alpha_i} + \frac{\partial \varphi}{\partial \alpha_i} \Sigma \alpha \alpha
\] (2.56)

So that, provided the model matrices are differentiable with respect to \( \alpha \), \( \zeta(z(t)|z^{t-1}, \alpha) \) is differentiable with respect to \( \alpha \). Since we are also assuming linear stable models driven by stationary ergodic noise processes the rest of condition C1 holds. As for the uniform bound in condition C2, it also follows from the above stability assumptions provided the model matrices and their partial derivatives are bounded uniformly in the parameter space.

**Lemma 2.2:**

Suppose that \( \exists \alpha_0 \in \mathbb{P} \) s.t. \( \forall \rho > 0 \)

\[
\lim_{t \to \infty} h^{\alpha_0}_t(z^t) = 0 \quad \text{a.e. uniformly in } \alpha \in \mathbb{P} - B(\alpha_0, \rho)
\] (2.57)

Then, the ML estimate \( \hat{\alpha}_t \) on \( \mathbb{P} \) converges to \( \alpha_0 \), a.s. as \( t \to \infty \).

**Proof:**

Choose \( 0 < \varepsilon_1 < 1 \) and \( \beta_1 > 0 \). Then, by (2.57)
3T₁ s.t. \( \sup_{\alpha \in \mathcal{P} - B(\alpha₀, \rho₁)} h_\alpha^\alpha(z^t) \leq \varepsilon₁ < 1 \) a.e. \( \forall t > T₁ \)

Furthermore,

3T₂ s.t. \( \alpha \in B(\alpha₀, \rho₂) \)

where the sup is attained only at \( \alpha₀ \). (To prove that, assume by way of contradiction that

\( \forall T₂ \exists \alpha'₀ \in B(\alpha₀, \rho₂), \alpha'₀ \neq \alpha₀ \), s.t. \( h_\alpha^\alpha₀(z^t) > h_\alpha^\alpha₀(z^t) \) some \( t > T₂ \)

and choose \( \varepsilon₂ < ||\alpha₀ - \alpha'₀|| \), \( \varepsilon₂ < h_\alpha^\alpha₀ = 1 \). Then

\( \forall T \exists \alpha₀' \in \mathcal{P} - B(\alpha₀, \rho₂) \) s.t. \( h_\alpha^\alpha₀(z^t) > \varepsilon₂ \) some \( t > T \)

which contradicts (2.57).

Now let \( T = \max(T₁, T₂) \). We then have

\[ \sup_{\alpha \in \mathcal{P}} h_\alpha^\alpha(z^t) = h_\alpha^\alpha₀(z^t) = 1 \] a.e. \( \forall t > T \)

or

\[ \lim_{t \to \infty} \sup_{\alpha \in \mathcal{P}} h_\alpha^\alpha(z^t) = h_\alpha^\alpha₀(z^t) = 1 \] a.e.

which implies \( \lim_{t \to \infty} \hat{\alpha}_t = \alpha₀ \). a.s.

It now follows from the two previous lemmas that the maximum likelihood estimate converges with probability one into the set \( \mathcal{P}_₀ \) of parameters minimizing
as a function of $\alpha$. By virtue of our comment following (2.36),
such a parameter $\alpha_0 \in P_0$ can be described as closest in the
information distance $|I(,;.)|$ to the true parameter; and the above
convergence result means that the maximum likelihood method selects
the models closest to the true system in that information distance.

In the case where the true system is a member of the model set,
the existence of a unique true parameter is often described
through identifiability conditions. Consistency of the parameter
estimates then follows from their convergence to this identifiable
true parameter. In the present context we will borrow the term
identifiability to describe the uniqueness of a parameter $\alpha_0$
iminimizing $I^*_\alpha$.

2.4. Identifiability Analysis for Stationary Systems

2.4.1 Global Identifiability

Recall that $\hat{\alpha}_t$ minimizes $\zeta(,;\alpha)$ which depends on $\alpha$ through the
Kalman filter residuals $r(t;\alpha)$ as well as their covariance $S_\alpha$.

Rewrite (2.49) as:

$$
\frac{\partial \zeta(z(t)\mid z^{t-1}\alpha)}{\partial \alpha_i} = r'(t;\alpha)S^{-1}_\alpha \frac{\partial r(t;\alpha)}{\partial \alpha_i}
+ \frac{1}{2} tr [(I - S^{-1}_\alpha r(t;\alpha) r'(t;\alpha))S^{-1}_\alpha \frac{\partial S}{\partial \alpha_i}] 
$$

(2.58)
where the first term is the $i^{th}$ component of the gradient of
\[ \zeta(z(t)|z^{-1};\alpha) \] holding $S_\alpha$ fixed and the second term is the $i^{th}$ component of the gradient holding $r(t;\alpha)$ 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 $S_\alpha$, precomputed through the algebraic matrix Riccati equation corresponding to $\alpha$. One would therefore expect that if, for some $\alpha \neq \alpha_0$, $r(t;\alpha)$ and $S_\alpha$ are respectively identical to $r(t;\alpha_0)$ and $S_{\alpha_0}$, $\alpha_0$ will not be identifiable.

Now by virtue of the stationarity assumptions made above, consider the frequency domain description of the steady state Kalman filter and let

\[ H(y;\alpha) = C_\alpha (y - A_\alpha)^{-1} A_\alpha H + I \quad (2.59) \]

Then

\[ r(y;\alpha) = H(y;\alpha)^{-1} z(y) \quad (2.60) \]

Therefore, in view of the above discussion, it is necessary for identifiability of $\alpha_0$ to have

\[ \begin{cases} S_\alpha \neq S_{\alpha_0} \\ or \\ H(y;\alpha) \neq H(y;\alpha_0) \end{cases} \quad (2.61) \]

\[ \begin{cases} S_\alpha \neq S_{\alpha_0} \\ or \\ H(y;\alpha) \neq H(y;\alpha_0) \end{cases} \quad (2.62) \]
for all $\alpha \neq \alpha_0$, $\alpha \in P$ since otherwise some other parameter would have the same likelihood function as $\alpha_0$, and would be, in particular, as close in the information distance (2.3) to the true parameter.

In the case where the true system is part of the model set, we can show that (2.61) or (2.62) for all $\alpha \neq \alpha_0$ is sufficient as well as necessary for $\alpha$ to be a unique minimizer of $I_\alpha$ in the following manner:

Let $\phi_z(y;*)$ be the discrete (assumed full rank) rational spectral density matrix of the observation sequence $\{z(t)\}$. It can be shown (see e.g. [3]) that the true Kalman filter transfer function $H(y;*)$ and the corresponding positive definite prediction error covariance matrix $S_*$ are uniquely determined by the unique (Youla [58]) factorization:

1) $\phi_z(y;*) = H(y;*)S_*H^*(y^{-1};*)$

2) $H(y;*)$ analytic for $|y| > 1$

3) $H^{-1}(y;*)$ analytic for $|y| > 1$

4) $\lim_{y \to \infty} H(y) = I$

Now, for all $\alpha \in P$

$$ I_\alpha = \frac{1}{2} \log \det S_\alpha + \frac{1}{2} \text{tr}[S_\alpha^{-1}E_\alpha \{r(t;\alpha)r^\prime(t;\alpha)\}] $$

$$ \geq \frac{1}{2} \log \det \{E_\alpha \{r(t;\alpha)r^\prime(t;\alpha)\}\} + \frac{\epsilon}{2} $$

$$ \geq \frac{1}{2} \log \det \{E_* \{r(t;\alpha)r^\prime(t;\alpha)\}\} + \frac{\epsilon}{2} $$

$$ = \frac{1}{2} \log \det S_* + \frac{1}{2} \text{tr}[S_*^{-1}E_* \{r(t;\alpha)r^\prime(t;\alpha)\}] \triangleq I_* $$
where the first inequality follows from lemma A.1 of Appendix A and the second inequality from the optimality property of the Kalman filter. Furthermore, the first inequality is tight if and only if (Lemma A.1)

\[ S_\alpha = E_\star [r(t;\alpha)r'(t;\alpha)] = S_\alpha^* \]

and the second inequality is tight if and only if (Lemma A.2)

\[ S_\alpha^* = E_\star [r(t;\alpha)r'(t;\alpha)] = E_\star [r(t;*)r'(t;*)] = S_* \]

Therefore, and in view of the above properties of the true Kalman filter

\[ I_\alpha^* = I_*^* \quad \text{for some } \alpha \neq * \quad \iff \quad \begin{cases} S_\alpha = S_* \\ S_\alpha^* = S_*^* \end{cases} \]

\[ \iff \phi_z(y;\alpha) \equiv \phi_z(y;*) \]

or, equivalently stated, \( I_\alpha^* > I_*^* \) for all \( \alpha \neq * \) if and only if

\[
\begin{cases}
S_\alpha \neq S_* \\
\text{or} \\
H(y;\alpha) \neq H(y;*)
\end{cases}
\]

for all \( \alpha \neq * \).

\[(2.63)\]

\[(2.64)\]
The following simple example illustrates the role played by condition (2.63).

**Example 2.1**

Consider the system

\[
x(t+1) = \xi(t) ; \quad \xi(t) \sim N(0, \Xi_*)
\]

\[
z(t) = x(t) + \theta(t) ; \quad \theta(t) \sim N(0, \Theta_*)
\]

And the model set

\[
\mathcal{M} = \{ (\xi_\alpha, \Theta_\alpha) ; \alpha \in \mathcal{P} \}
\]

where \( \mathcal{P} \) contains \( * \).

Clearly,

\[
\hat{z}(t; \alpha) = \hat{x}(t; \alpha) = 0 \quad \forall \alpha \in \mathcal{P}
\]

so that,

\[
H(y; \alpha) = I \quad \forall \alpha \in \mathcal{P}
\]

However

\[
S_\alpha = \Xi_\alpha + \Theta_\alpha
\]

and \( \hat{\alpha}_t \) will converge to \( * \) if and only if

\[
S_\alpha \neq S_* \quad \forall \alpha \neq *, \alpha \in \mathcal{P}
\]

For example, if

\[
\Xi_\alpha = \alpha_1 ; \quad \Theta_\alpha = \alpha_2
\]
\[ \hat{a}_t \] will not be consistent whereas if
\[ \bar{a} = a_1, \quad \Theta = 2a_1 \]
\[ \hat{a}_t \] will be consistent.

Note also, that, in general, obtaining a parametrization of state space models which satisfies the above conditions is no trivial task. As mentioned in Chapter one, canonical forms and a priori structural knowledge could be used (e.g. Caines [43], Glover and Willems [44]). However, it is sometimes the case that the parametrization is naturally given by physical laws (e.g. time constants) and that the problem of a unique parametrization is avoided from the outset.

Recall now the discussion in the beginning of this section. In the case where the true system is in the model set and has a unique parametrization in the sense of (2.63)-(2.64), \(a_0 = \ast\) will obviously be the only parameter to match the residuals \(r(t; a)\) to the true innovations sequence \(r(t; \ast)\) as well as their presumed covariance \(S_a\) to the minimum prediction error covariance \(S_\ast\). This provides insight to the role of a unique \(a_0 \in P_0\) when the true system is not part of the model set. In this case, the above match cannot be made in general, and the minimum information distance parameter \(a_0\) provides a tradeoff between the best fit of the sequence of residuals \(r(t; a)\)
and the best approximation of the corresponding $S_\alpha$ to the "true" covariance $S_*$. The following example illustrates this point.

Example 2.2:

Consider the scalar model set

$$M = \{ A_\alpha = \alpha, L_\alpha = 1, C_\alpha = 1, E_\alpha = (1-\alpha^2), \Theta_\alpha = 1 \mid \alpha \in [0,1]\}$$

for which we can find $S_\alpha = 1 + \sqrt{1-\alpha^2}$.

Note that such a dependence of the plant noise model on the plant dynamics model could occur when modelling a process with known state covariance but unknown time constant $\alpha$.

If the true model has the same form with

$$A_* = 0.5; \quad E_* = 0.75$$

it is clearly contained in the model set ($\alpha = 0.5$) and has

$$S_* = S_{0.5} = 1.866$$

After tedious but straightforward manipulations we can also find

$$S_*^2 = 2 + \frac{2(\alpha-1)(\alpha^2-1+\sqrt{1-\alpha^2})}{(2\alpha-1+\sqrt{1-\alpha^2})}$$

$$I_*^2 = \frac{1}{2} \left[ \log(1+\sqrt{1-\alpha^2}) + \frac{S_*^2}{1+\sqrt{1-\alpha^2}} \right]$$

Both $S_*^2$ and $I_*^2$ are exactly minimized at $\alpha = 0.5$ which achieves
\[ S^*_{0.5} = S^* = S_{0.5} = 1.866 \]

If, on the other hand, the true model is

\[ A^* = 0.5; \quad \xi^* = 1.5 \]

then it is not contained in the model set. In this case

\[ S^* = 2.656 \]

and

\[ S^*_\alpha = 3 + \frac{(\alpha^2-1 + \sqrt{1-\alpha^2})(6\alpha - 7 - \sqrt{1-\alpha^2})}{2(2\alpha - 1 + \sqrt{1-\alpha^2})} \]

Here, \( S^*_\alpha \) is minimized and therefore takes on its closest value to

to \( S^*_\alpha \) at \( \alpha = 0.59 \), for which \( S^*_{0.5} = 2.664 \); whereas \( S^*_\alpha \) takes on its closest

value to \( S^*_\alpha \) at \( \alpha = 0 \), for which \( S^*_\alpha = 2 \).

The ML compromise solution is at the minimizer of \( I^*_{\alpha} \),\n
\[ \alpha_0 = 0.49 \]

for which

\[ S^*_{0.49} = 1.872 \]

\[ S^*_0.49 = 2.676 \]

It is also interesting to compare the minimum information distance

parameter \( \alpha_0 \) with the asymptotic parameter estimate selected by other
prediction error identification methods. This term has been used by Ljung (e.g. [29]) and Caines ([25]) to designate the general class of methods which involve minimizing criteria of the form

\[ h \left[ \frac{1}{t+1} \sum_{\tau=0}^{t} \lambda(\tau; \alpha, r(\tau; \alpha)) \right] \] (2.65)

where \( \lambda \) is a general function from \( \mathbb{R} \times \mathbb{R} \times \mathbb{R} \) to the space of symmetric matrices subject to some regularity conditions, and where \( h \) is some "order preserving" scalar function of those symmetric matrices.

The convergence of general prediction error parameter estimates was studied in broad terms by Ljung [29] under general stability assumptions on the system, differentiability assumptions on the model matrices and a search over models leading to stable predictors (consistent with our present assumptions). It was shown in particular ([29]) that when \( h \) is increasing, \( \lambda \) is scalar and satisfies

\[ E_\star \{ |r(t; \alpha_1)|^2 \} > E_\star \{ |r(t; \alpha_2)|^2 \} \]

\[ \Rightarrow E_\star \{ \ell(t; \alpha_1, r(t; \alpha_1)) \} > E_\star \{ \ell(t; \alpha_2, r(t; \alpha_2)) \} \] (2.66)

the corresponding parameter estimates converge into the set of models that give the "best" output predictions in the sense:

\[ \min_{\alpha \in \mathcal{SP}} \lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} E_\star \{ |\hat{z}(t; \alpha) - \hat{z}(t; \alpha)|^2 \} \] (2.67)
or equivalently

\[
\min_{\alpha \in \mathcal{P}} \lim_{t \to \infty} \frac{1}{t+1} \sum_{t=0}^{t} E_{*}\{ |r(t; \alpha) - r(t; \alpha) |^2 \} \quad (2.70)
\]

Now the maximum likelihood criterion, although a special case of (2.65), does not satisfy (2.66) in general. This emphasizes the difference between the asymptotic models obtained by the ML method and those obtained by a least squares type criterion which satisfies (2.66). More specifically, the latter give "good" approximate predictors in the sense of (2.70) whereas the former give "close" models in the sense of the information distance defined above, drawing on the tradeoff between matching the residuals and matching their second order moment as discussed above.

Example 2.3

In Example 2.2, the asymptotic parameter selected by criterion (2.70) is

\[ \alpha = \arg\min_{\alpha} S_{\alpha}^* = 0.59 \]

placing

\[ A_{0.59} = 0.59 \quad \text{and} \quad S_{0.59} = 1.807 \]

"further away" from

\[ A_* = 0.5 \quad \text{and} \quad S_* = 2.656 \]
than the tradeoff model

\[ A_{0.49} = 0.49 \quad \text{and} \quad S_{0.49} = 1.872 \]

2.4.2. Local Identifiability

An element \( \alpha_0 \) of \( P_0 \) is locally identifiable if there exists an open neighborhood \( N_{\alpha_0} \) of \( \alpha_0 \) in \( P \) where \( I^*_{\alpha} > I^*_{\alpha_0} \quad \forall \alpha \in N_{\alpha_0} \). Under the current stationarity assumptions, we formally have

\[
I^*_{\alpha} - I^*_{\alpha_0} = (\alpha - \alpha_0) \left( \frac{\partial I^*_{\alpha}}{\partial \alpha} \right)_{\alpha = \alpha_0} + \frac{1}{2} (\alpha - \alpha_0)^T \left( \frac{\partial^2 I^*_{\alpha}}{\partial \alpha^2} \right)_{\alpha = \alpha_0} (\alpha - \alpha_0) + \text{h.o.t.}
\]

and a sufficient condition for local identifiability of \( \alpha_0 \) would be

\[
\left. \frac{\partial^2 I^*_{\alpha}}{\partial \alpha^2} \right|_{\alpha = \alpha_0} > 0 \quad (2.71)
\]

This is a generalization of the corresponding local identifiability condition in the case where the true system is part of the model set and where (2.71) is replaced by the positive definiteness of the Fisher information matrix in a single observation \( z(t) \).

Indeed, for \( \alpha_0 = \star \) and under some regularity conditions,

\[
\left. \frac{\partial^2 I^*_{\alpha}}{\partial \alpha^2} \right|_{\alpha = \alpha_0 = \star} = E_{\star} \left\{ \frac{\partial^2 z(t) | z(t-1) | \mu}{\partial \alpha^2} \right\} |_{\alpha = \alpha_0 = \star}
\]
which is, by definition the Fisher information matrix in observation 
z(t).

Note also, at this point, that the relation in [59], section 3.3.2

\[ I(\alpha; \alpha + \Delta \alpha) = \frac{1}{2} \Delta \alpha \int \left\{ \frac{\partial^2 \xi(z(t)|z^{-1}; \alpha)}{\partial \alpha^2} \right\} \Delta \alpha \]

between the information measure \( I(\alpha_1; \alpha_2) \) and the Fisher information in a single observation, is valid only at \( \alpha = \star \) (i.e. at the true parameter if it is included in \( P \)). The relation does not hold in general and the error seems to result from erroneous regularity conditions.

### 2.5 Convergence Analysis for Time Invariant Systems with Deterministic Inputs

In this section, we keep our assumptions of time-invariant, stable, controllable and observable models and their corresponding steady state filters to process the observations but we will assume the presence of deterministic inputs \( u(t) \). In this case \( S^*_\alpha(t) \) as obtained from (2.38) depends on \( u(t) \), so does

\[ I^*_\alpha(t) = \frac{1}{2} \log(\det[S^*_\alpha]) + \frac{1}{2} \text{tr}[S^{-1}_\alpha S^*_\alpha(t)] \]  \hspace{1cm} (2.72)

and the previous stationarity is lost in general. However, the convergence analysis of section 2.3 can be generalized in the following manner.
In view of (2.30)-(2.33), the sequence of residuals \( r(t;\alpha) \) can be expressed as

\[
 r(t;\alpha) = \tilde{r}(t;\alpha) + \bar{r}(t;\alpha) \tag{2.73}
\]

where \( \tilde{r}(t;\alpha) \) is a zero mean stationary ergodic sequence given by

\[
 \dot{x}_\alpha^*(t+1) = A_{\alpha} \dot{x}_\alpha^*(t) + L_{\alpha} \xi(t) \tag{2.74}
\]

\[
 \tilde{r}(t;\alpha) = C_{\alpha} \dot{x}_\alpha^*(t) + \theta(t) \tag{2.75}
\]

with

\[
 E_\alpha [ \tilde{r}(t;\alpha) \tilde{r}'(t;\alpha) ] = C_{\alpha}^* \Sigma_{\alpha} C_{\alpha}^* + \Theta_{\alpha}^* \Lambda_\alpha \tag{2.76}
\]

and where \( \bar{r}(t;\alpha) \) is a deterministic sequence given by

\[
 \dot{x}_\alpha^*(t+1) = A_{\alpha} \dot{x}_\alpha^*(t) + B_{\alpha} u(t) \tag{2.77}
\]

\[
 \bar{r}(t;\alpha) = C_{\alpha} \dot{x}_\alpha^*(t) \tag{2.78}
\]

Note, in particular, that \( ^* S_{\alpha}(t) \) as obtained from (2.38) can now be reexpressed as

\[
 ^* S_{\alpha}(t) = S_{\alpha} + \bar{r}(t;\alpha) \bar{r}(t;\alpha) \tag{2.79}
\]

We will also use the following lemma from Dunsmuir [33] (the proof is due to Hannan).
Lemma 2.3:

Let \( w(t) \) be a scalar zero mean, ergodic, purely non deterministic process and let \( v(t) \) be a scalar deterministic sequence satisfying

\[
\lim_{t \to \infty} \frac{1}{t} \log \frac{1}{\frac{1}{t^\beta} \sum_{\tau=1}^{t} v^2(\tau)} = 0 \quad 0 < \beta < \frac{1}{2}
\]

Then

\[
\lim_{t \to \infty} \left( \frac{1}{t^\beta} \sum_{\tau=1}^{t} v^2(\tau) \right)^{1/2} = 0 \quad \text{a.s.}
\]

Note that if \( v(t) \) also satisfies the stronger condition

\[
\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} v^2(\tau) < \infty \tag{2.80}
\]

then

\[
\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} v(t)w(t) = 0 \quad \text{a.s.} \tag{2.81}
\]

Recall now \( J_t(\alpha, z^{t-1}) \) defined in the proof of lemma 2.1, equation (2.43). We now have

\[
J_t(\alpha, z^{t-1}) = \frac{1}{t+1} \sum_{\tau=0}^{t} \xi(z(\tau)) |z^{\tau-1}; \alpha|
\]

\[
= \frac{1}{2} \log \det S_\alpha + \frac{1}{2} \text{tr} \left[ S^{-1}_\alpha \left( \frac{1}{t+1} \sum_{\tau=0}^{t} r(\tau; \alpha)r'(\tau; \alpha) \right) \right]
\]
\[
\begin{aligned}
&\frac{1}{2} \log \det S_\alpha \\
= &\left\{ \begin{array}{l}
\frac{1}{2} \operatorname{tr} \left[ S_\alpha^{-1} \left( \frac{1}{t+1} \sum_{\tau=0}^{t} \tilde{r}(\tau; \alpha) \tilde{r}'(\tau; \alpha) \right) \right] \\
+ \frac{1}{2} \operatorname{tr} \left[ S_\alpha^{-1} \left( \frac{1}{t+1} \sum_{\tau=0}^{t} \tilde{r}(\tau; \alpha) \tilde{r}'(\tau; \alpha) \right) \right] \\
+ \operatorname{tr} \left[ S_\alpha^{-1} \left( \frac{1}{t+1} \sum_{\tau=0}^{t} \tilde{r}(\tau; \alpha) \tilde{r}'(\tau; \alpha) \right) \right]
\end{array} \right. \\
&\text{(2.82)}
\end{aligned}
\]

where we substituted for \( r(t; \alpha) \) from (2.73).

Under the same assumptions of stable, controllable, observable and differentiable models, with the partial derivatives of the model matrices bounded uniformly in the parameter space, \( \tilde{r}(t; \alpha) \) has the same properties as \( r(t; \alpha) \) in Section 2.3. Therefore, the same steps as those in the proof of Lemma 2.1 can be used to show that the second term of (2.82) converges uniformly in \( \alpha \in \mathcal{P} \) to

\[
\frac{1}{2} \operatorname{tr} \left[ S_\alpha^{-1} S_\alpha^{-1} \right]
\]

Furthermore, if we restrict the inputs \( u(t) \) to satisfy the "finite average power" condition

\[
\lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} u(\tau-i)u'(\tau-j) < \frac{1}{\delta} I \quad \delta > 0, \forall i, j \quad (2.83)
\]

\( \tilde{r}(t; \alpha) \) and \( \tilde{r}(t; \alpha) \), being the outputs of the LTI systems (2.74)-(2.75) and (2.77)-(2.78) respectively, satisfy under the same above assumptions (see e.g. [27] for similar arguments)
\[
\lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} \bar{r}(\tau; a) \bar{r}'(\tau; a) \leq K_1 < \infty
\] (2.84)

as well as
\[
\lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} \frac{3}{2a} (\bar{r}(\tau; a) \bar{r}'(\tau; a)) \leq K_2 < \infty
\] (2.85)

Therefore, in view of (2.84), each component of \( \bar{r}(t; a) \) and \( \bar{r}(t; a) \)
has the properties of respectively \( w(t) \) and \( v(t) \) of Lemma 2.3 and

\[ t \to \infty \]

(2.86)

With (2.85) playing the role of \( C_2 \) in Lemma 2.1 we can use, here
again the same technique to show that this convergence is uniform
in \( \alpha \) and hence, that the last term of (2.82) converges to zero
uniformly in \( \alpha \in \mathcal{P} \).

In consequence
\[
\lim_{t \to \infty} J_t(\alpha, \beta^{-1}) = \frac{1}{2} \log \det S + \frac{1}{2} \text{tr} \left[ S^{-1} \bar{S}^* + \lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} \bar{r}(\tau; \alpha) \bar{r}'(\tau; \alpha) \right]
\] (2.87)

or, recalling (2.79),
It now follows, from an analysis similar to that of Section 2.3, that the maximum likelihood parameter estimate converges with probability one into the set \( P_0(t) \) of parameters minimizing

\[
J^*_\alpha(t) \overset{\Delta}{=} \lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} E_* \{ \zeta(z(\tau) | z^{T-1}; \alpha) \}
\]

as a function of \( \alpha \).

This result is consistent with that obtained by Ljung ([29]) for the general prediction error parameter estimates mentioned in Section 2.4. Note also from (2.87) that \( J^*_\alpha(t) \) has a time invariant part

\[
\frac{1}{2} \log \det S_\alpha + \frac{1}{2} \text{tr} [S^{-1}_{\alpha} S_{\alpha}^*]
\]

(2.89)

(which plays the same role as that of \( I^*_\alpha \) in Sections 2.3-2.4) as well as a time average of the square of means part

\[
\frac{1}{2} \text{tr} \left[ S^{-1}_{\alpha} \left( \frac{1}{t+1} \sum_{\tau=0}^{t} \overline{r}(\tau; \alpha) \overline{r}^*(\tau; \alpha) \right) \right]
\]

(2.90)
(due to the deterministic input $u(t)$). Without further assumptions on
$u(t)$, (2.90) cannot be simplified. This emphasizes the role of a
particular input in biasing the asymptotic choice of a model, and, in
the next section, we extend our analysis of section 2.4 to take into
account this additional term.

2.6 Identifiability Analysis for Time Invariant Systems with
Deterministic Inputs

As in Section 2.4, consider the frequency domain description of
the steady state Kalman filter, let $H(y;\alpha)$ be defined as in (2.59)
and let

$$G(y;\alpha) = C_\alpha(yI-A)^{-1}B_\alpha$$

(2.91)

Now,

$$r(y;\alpha) = H(y;\alpha)^{-1}z(y) - H(y;\alpha)^{-1}G(y;\alpha)u(y)$$

and we add

$$G(y;\alpha) \neq G(y;\alpha_0)$$

(2.92)

for all $\alpha \neq \alpha_0$, $\alpha \in P$ as a possible necessary condition for $\alpha_0$ to
be identifiable in the sense of Section 2.4.

Here again, in the case where the true system is part of the
model set, (2.92) also provides a sufficient condition for identifying
those unknown parameters which only affect $G(y;\alpha)$ provided the input
$u(y)$ is general enough to excite all modes of the system. Indeed,
following steps similar to those of section 2.4 we have, for all \( \alpha \in \mathcal{P} \)

\[
J_{\alpha}^*(t) \triangleq \frac{1}{2} \log \det S_{\alpha} + \frac{1}{2} \operatorname{tr} \left[ S_{\alpha}^{-1} (S_{\alpha}^* + \lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} \overline{r}(\tau;\alpha)\overline{r}'(\tau;\alpha)) \right]
\]

\[
\geq \frac{1}{2} \log \det \left[ S_{\alpha}^* + \lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} \overline{r}(\tau;\alpha)\overline{r}'(\tau;\alpha) \right] + \frac{r}{2}
\]

\[
\geq \frac{1}{2} \log \det \left[ E_{\alpha} \{r(t;\alpha)r'(t;\alpha)\} \right] + \frac{r}{2}
\]

\[
= \frac{1}{2} \log \det S_{\alpha} + \frac{1}{2} \operatorname{tr} \left[ S_{\alpha}^{-1} E_{\alpha} \{r(t;\alpha)r'(t;\alpha)\} \right] \triangleq J_{\alpha}^*
\]

where, as before, the first inequality is tight if and only if (Lemma A.1)

\[
S_{\alpha} = S_{\alpha}^* + \lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} \overline{r}(\tau;\alpha)\overline{r}'(\tau;\alpha)
\]

and the second inequality is tight if and only if (Lemma A.2)

\[
S_{\alpha}^* + \lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} \overline{r}(\tau;\alpha)\overline{r}'(\tau;\alpha) = S_{\alpha}
\]

In view of the optimal properties of the true Kalman filter, (2.94) is equivalent to

\[
\begin{align*}
- S_{\alpha}^* & \triangleq E_{\alpha} \{\overline{r}(t;\alpha)\overline{r}'(t;\alpha)\} = E_{\alpha} \{r(t;\alpha)r'(t;\alpha)\} \triangleq S_{\alpha}^*
\end{align*}
\]

and

\[
\begin{align*}
\lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} \overline{r}(\tau;\alpha)\overline{r}'(\tau;\alpha) = 0
\end{align*}
\]
Furthermore, under our present assumptions on $u(t)$ and using our frequency domain description again

$$-1 \quad \mathcal{F}(y;\alpha) = H(y;\alpha)[G(y;\ast) - G(y;\alpha)]u(y)$$

Therefore, it is easy to agree with e.g. Ljung [27] that sufficient conditions for identifiability here would be

$$G(y;\alpha) \neq G(y;\ast) \quad \forall \alpha \neq \ast, \quad \alpha \in \mathcal{P}$$

and $u(t)$ independent of the process noise and persistently exciting\(^1\).

This point will be further developed in the next chapter. Let us only look here at the following example to illustrate the role of persistently exciting inputs.

**Example 2.4:**

Consider the scalar model set

$$x(t+1) = \alpha u(t) + \xi(t) ; \quad \xi(t) \sim N(0,1)$$

$$z(t) = x(t) + \theta(t) ; \quad \theta(t) \sim N(0,1)$$

and where the true system has the same form with $\alpha = 1$.

Clearly,

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

\(^1\) $u(t)$ is persistently exciting if, for all $M$, there exists $\delta(M)$ and $N_0(M)$ such that

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

for $N > N_0$ and where $u'_M(t) \triangleq [u'(t) \ldots u'(t-M)]$. 
and

\[ S_{\alpha} = S_* = 2 \quad \forall \alpha \]

\[ H(y; \alpha) = H(y; *) = I \quad \forall \alpha \]

\[ G(y; \alpha) = \frac{\alpha}{\beta} \neq G(y; *) = \frac{1}{\beta} \quad \forall \alpha \neq * \]

But

\[ S_{\alpha}^*(t) = 2 + (1-\alpha)^2 u^2(t-1) \]

and

\[ J_{\alpha}^*(t) = \frac{1}{2} \left( \log 2 + 1 \right) + \frac{(1-\alpha)^2}{2} \frac{1}{t+1} \sum_{\tau=0}^{t} u^2(\tau-1) \]

so that, provided

\[ 0 < \lim_{t \to \infty} \frac{1}{t+1} \sum_{\tau=0}^{t} u^2(\tau-1) < \infty \]

\[ J_{\alpha}^*(t) \text{ is uniquely minimized at } \alpha = 1 \text{ as } t \to \infty. \]

On the other hand, when the true system is not included in the model set, the deterministic inputs enter as a third factor in the tradeoff discussed in Section 2.4.1. In particular, the search for a good approximation of the stochastic part of the system could be dominated by that of its deterministic part when the input excites only certain modes of the system, say, dwarfing the excitation of the remaining modes by the noise input \( \xi(t) \).
2.7 Asymptotic Normality

In this section we discuss the other standard important property of maximum likelihood estimators but in the context of the systems and model sets of this chapter. Along with convergence, normality is an important asymptotic property because it can help determine the amount of data to be gathered in order to achieve a certain accuracy of estimation.

Briefly, the classical results on the asymptotic normality and efficiency of ML estimators as they are applied to independent identically distributed observations and where the parameter set contains the true parameter \( \theta_0 = \ast \) can be formulated as follows:

i) As \( t \to \infty \), \( \hat{\theta}_t - \theta_0 \) is asymptotically normally distributed with zero mean and covariance matrix

\[
I^{-1}_t(\theta_0)
\]

where

\[
I_t(\theta) \triangleq \mathbb{E}_z \left\{ \frac{\partial^2 \zeta(z_t; \theta)}{\partial \theta^2} \right\}
\]

is the Fisher information matrix in the set of observations \( z_t \).

ii) \( I^{-1}_t(\theta_0) \) is the smallest error covariance that any unbiased estimator can achieve and is known as the Cramer-Rao lower bound. Now because the observations are assumed to be independent and identically distributed we also have
\[ I^*_t(\alpha) = (t+1)I^*_z(\alpha) \]  
(2.97)

where

\[ I^*_z(\alpha) = E^*_z \left\{ \frac{\partial^2 \xi(z(t);\alpha)}{\partial \alpha^2} \right\} \]  
(2.98)

is the time independant information matrix in a single independent observation \( z(t) \). This means that as \( t \to \infty \)

\[ E^*_z\{ (\hat{\alpha}_t - \alpha_0) (\hat{\alpha}_t - \alpha_0)' \} = \frac{1}{t} I^{-1}_z(\alpha_0) \]  
(2.99)

and this characteristic \( \frac{1}{\sqrt{t}} \) behavior of the estimate error standard deviation implies that a tenfold increase in accuracy requires a hundred-fold increase in the number of observations.

This result can be generalized to ML parameter estimates based on observations generated by a linear time invariant system. In the stationary ergodic case of Section 2.3, and when the system is contained in the model set, one can show that, here again, \( (\hat{\alpha}_t - \alpha_0) \) is asymptotically normally distributed with zero mean and covariance \( I^{-1}_z(\alpha_0) \) defined in (2.95)-(2.96). The proof is similar to that used by Caines and Ljung [30] for a set of identification criteria similar to, but not containing, our present ML criterion. (see also Dunsmuir and Hannan [34,35]). It is based on the martingale property of the process.
at the true parameter $a=a_0^*$, with respect to the true probability measure, and uses Billingsley's [60] version of the central limit theorem for martingales. In this case, because of the stationarity assumption, we also have

$$I_t^{(a)} = (t+1)I_t^{(a)}$$

where

$$I_t^{(a)} = E_t^* \left\{ \frac{\partial^2 \zeta(z(t)|z^{t-1};a)}{\partial a^2} \right\}$$

is the time invariant Fisher information matrix in a single observation, and the same comments made after (2.99) are valid here.

On the other hand, when the system is not contained in the model set, it was shown in section 2.3 that $\hat{a}_t$ could still converge to $a_0$ minimizing $I_{\alpha}^*$. However, the martingale property of the process in (2.100) does not hold for this parameter value $a_0^*$ and another approach has to be used to show the asymptotic normality of the parameter estimates. The alternate approach, due to Ljung and Caines [31], is quite general and covers both the stationary and non stationary cases of Sections 2.3 and 2.5.
By using the same general criterion (2.65) with some additional smoothness assumptions on the models as well as the functions \( h \) and \( \ell \), and by drawing on Ljung's [29] general convergence result, they obtain [31] a corresponding general asymptotic normality result which, when specialized to our present maximum likelihood case reduces to the following:

Let \( \hat{\alpha}_t \) be the maximum likelihood parameter estimate at time \( t \). Let \( \alpha_0, t \) be the global minimum, assumed unique, of \( J^*_\alpha(t) \) defined in (2.88). (i.e. \( \alpha_0, t \) is the unique element of \( P_{0,t} \)). From Section 2.5 we already have \( \hat{\alpha}_t \to \alpha_0, t \) with probability one as \( t \to \infty \).

Assume further that

\[
\frac{\partial^2 J^*_\alpha(t)}{\partial \alpha^2} > \delta I ; \quad \delta > 0, \; \forall \alpha \in P
\]  

(2.103)

and that the compact set \( P \) is a neighborhood of \( \alpha_0, t \) (i.e. contains an open set containing \( \alpha_0, t \)).

Then, \( (\hat{\alpha}_t - \alpha_0, t) \) is asymptotically normal with zero mean and covariance matrix

\[
I_t(\hat{\alpha}_t) \to^t I_t(\hat{\alpha}_t) \to^t \text{as} t \to \infty
\]  

(2.104)

where

\[
I_t(\alpha) \triangleq E^* \left\{ \frac{\partial^2 \zeta(z, \alpha)}{\partial \alpha^2} \right\} = (t+1) \frac{\partial^2 J^*_\alpha(t)}{\partial \alpha^2}
\]  

(2.105)
was already defined in (2.96) and

\[
\tilde{I}_t^* (z) = E_* \left\{ \left[ \frac{\partial^2 \xi (z^t ; \alpha)}{\partial \alpha \partial \alpha} \right] \right\}
\] (2.106)

Note that this result is consistent with the one presented above for the stationary case and where the system is contained in the model set. Indeed, it is easy to show that for \( \alpha = * \)

\[
\tilde{I}_t^* (z) = I_t (z) \]

so that (2.104) reduces to

\[
I_t^{-1} (z)
\]

Note also that in the stationary case, but where the model set does not contain the system, \( I_t (\alpha) \) reduces to

\[
(t+1) \frac{\partial^2 I_t^* (z)}{\partial \alpha^2}
\] (2.107)

which, recalling (2.71) in Section 2.4.2, seems to play a role similar to that of the information matrix in classical maximum likelihood theory. It is unfortunate, however, that \( \tilde{I}_t^* (z) \) remains more cumbersome to evaluate.
2.8 Summary

In this chapter we discussed different asymptotic properties of maximum likelihood parameter estimates in a linear time invariant state space model environment.

In Section 2.1, we adopted Baram and Sandell's [36] information measure as a tool for comparing different models in a model set which does not necessarily contain what is assumed to be a better description of the real system: a "true" state space model. This has provided us with a metric which describes the asymptotic models provided by the ML method as closest to the true system. We have also attempted to get more insight into the nature of such approximate models in a state space framework by showing them as tradeoff choices (Section 2.4). One of our goals in doing so was to interpret, in the state space framework, general theoretical convergence results (Caines [25], Ljung [29]) from which we drew heavily in Sections 2.3 through 2.6. The same goal was also pursued in Section 2.7 where we looked at the asymptotic normality of maximum likelihood estimates.
CHAPTER 3

IDENTIFICATION OF SYSTEMS UNDER FEEDBACK

As mentioned in Chapter 1, systems operating in closed loop present special identifiability problems, and many theoretical results on the behavior of identification methods for systems operating in open loop are no longer valid. Furthermore, it is often the case that one cannot remove the feedback from such systems either for safety reasons (as in industrial processes) or simply because it is intrinsic to the system (as in biological or economic systems). In this chapter, we will focus our attention on the special case where the input-output data is collected for parameter identification from a time-invariant system operating in closed loop. In particular, we review some of the results of Chapter 2 to determine what problems arise, if any, within our present framework of maximum likelihood identification for linear state space models. Throughout the chapter, we will restrict attention to cases where the parameters to be estimated are only in the forward path, assuming the feedback loop is known or can be chosen separately. Note however that several authors (e.g. Caines and Chan [61],[62]) have also studied the case where determining the feedback parameters (or even the presence of feedback) are part of the identification goal. Note also that knowledge of, or control over, the feedback mechanism lead us to distinguish between direct and indirect identification.
In the latter, the closed loop system is identified as a whole and knowledge of the feedback parameters is then used to deduce the appropriate forward path parameters. In the former those parameters are estimated directly from the input-output data which makes this approach more appealing.

We will first recall some of the examples of the type of difficulties encountered with direct identification algorithms in the presence of feedback. Then several treatments of those difficulties will be briefly reviewed arriving finally at a discussion of those problems from our maximum likelihood and state space models point of view.

3.1 Review of Some Examples

One of the earlier problems pointed out by e.g. Akaike [63], and Åström and Eykhoff [4], is connected with identification methods, such as spectral analysis, where no causal structure is enforced. We will use the example of [4] to illustrate this.

Example 3.1

Consider the simple closed loop configuration of Figure 3.1. In the frequency domain we have

\[ z(j\omega) = \frac{1}{1 + G_p(j\omega)G_R(j\omega)} \xi(j\omega) \]  

(3.1)
Figure 3.1: A Simple Closed Loop Configuration.

Figure 3.2: A Simple Closed Loop Configuration.
\[ u(\mathcal{f}) = \frac{-G_P(\mathcal{f})}{1 + G_P(\mathcal{f})G_R(\mathcal{f})} \xi(\mathcal{f}) \] (3.2)

\[ \phi_{zu}(\omega) = \frac{G_R^*(\omega)}{|1 + G_P(\omega)G_R(\omega)|^2} \phi_{\xi\xi}(\omega) \] (3.3)

\[ \phi_{uu}(\omega) = \frac{|G_R(\omega)|^2}{|1 + G_P(\omega)G_R(\omega)|^2} \phi_{\xi\xi}(\omega) \] (3.4)

so that if we attempt a direct identification of \( G_P \) using correlation and spectral information from input \( u \) and output \( z \) data, and assuming the model

\[ z(\mathcal{f}) = G_P(\mathcal{f}) u(\mathcal{f}) + \xi(\mathcal{f}) \] (3.5)

where \( u \) and \( \xi \) are uncorrelated, we get

\[ \hat{G}_P(\omega) = \frac{\phi_{zu}(\omega)}{\phi_{uu}(\omega)} = \frac{1}{G_R(\omega)} \]

The problem comes of course from the false assumption of independence between \( u \) and \( \xi \) when we ignore the feedback in (3.5).

Note for further reference that for the configuration of Figure 3.2 where the noise \( \xi \) is in the feedback loop we get

\[ z(\mathcal{f}) = \frac{G_P(\mathcal{f})}{1 + G_R(\mathcal{f})G_P(\mathcal{f})} \xi(\mathcal{f}) \] (3.6)

\[ u(\mathcal{f}) = \frac{1}{1 + G_R(\mathcal{f})G_P(\mathcal{f})} \xi(\mathcal{f}) \] (3.7)
\[ \phi_{zu}(\omega) = \frac{G_p(\omega)}{|1 + G_r(\omega)G_p(\omega)|^2} \phi_{\xi\xi}(\omega) \] (3.8)

\[ \phi_{uu}(\omega) = \frac{1}{|1 + G_r(\omega)G_p(\omega)|^2} \phi_{\xi\xi}(\omega) \] (3.9)

and

\[ \hat{G}_p(\omega) = \frac{\phi_{zu}(\omega)}{\phi_{uu}(\omega)} = G_p(\omega) \] (3.10)

Another widely used example is the following, given by e.g. Åström and Wittenmark [64] and illustrates the identifiability problem associated with feedback even when a so-called parametric identification method is used.

**Example 3.2**

Consider the system

\[ z(t) + az(t-1) = bu(t-1) + \xi(t) \] (3.11)

with a linear feedback regulator

\[ u(t) = dz(t) \] (3.12)

In this case, \( \hat{a} = a + \gamma d \) and \( \hat{b} = b + \gamma \) (where \( \gamma \) is arbitrary) are respectively equivalent to \( a \) and \( b \) from a noise-output point of view. This makes \( a \) and \( b \) unidentifiable even with the knowledge of \( d \). 

\[ \Box \]
Example 3.2 points out, the causality problem associated with feedback where the unique roles of input and output are lost.

Here again, for further reference, note that if the feedback regulator of example 3.2 contained a time delay, e.g.

\[ u(t) = dz(t-1) \] (3.13)

then a and b would become identifiable.

3.2 Overview of Standard Treatments

As mentioned previously there have been several approaches to the problem of identifying open loop dynamics from closed loop data. The main guiding principle is to overcome the loss of input output causal structure discussed in Section 3.1 by insuring a variety of different conditions. Those have ranged from the inclusion of noise in the feedback path (e.g. Vorchik et al. [50], Ng et al. [65]), to the addition of a small perturbation to the input (e.g. Saridis and Lobbia [49]); to having the regulator satisfying certain complexity requirements (e.g. Söderström et al. [66], Wellestead and Edmunds [67]) being nonlinear in the states (Fisher [47]) or even shifting between an appropriate number of linear regulators (Ljung et al. [46]). The above approaches, as well as many others including indirect identification approaches (such as that of Glover [41]) have been surveyed in an excellent paper by Gustavsson et al. [51]. We will take a closer look for our present purposes at two of the above studies.
In the vein of example 3.1 and the note following it, Ng, Goodwin and Anderson [65] have studied necessary and sufficient conditions to identify uniquely the forward and feedback path transfer functions of multi-input multi-output linear dynamic systems using spectral factorization techniques. They have considered the general configuration of Figure 3.3 where the joint process \((z(t), u(t))\) is a full rank stationary stochastic process with joint spectral density \(\phi_{zu}(\omega)\) and have obtained the following main result: From Figure 3.3

\[
\phi_{zu}(\omega) = K(\omega)D K' (\omega^{-1}) ; \quad \omega = e^{j\omega}
\]  

(3.14)

where

\[
K = \begin{bmatrix}
(I-G_1G_3)^{-1}G_2 & (I-G_1G_3)^{-1}G_1G_4 \\
(I-G_3G_1)^{-1}G_3G_2 & (I-G_3G_1)^{-1}G_4
\end{bmatrix}
\]

and where one can uniquely recover \(G_1, G_2, G_3\) and \(G_4\) from any particular spectral factor \(K\). Let \(\overline{K}(\omega)\) be the unique minimum phase spectral factor of \(\phi_{zu}(\omega)\) as specified by the requirements that \(\overline{K}(\omega)\) be stable, nonsingular for \(|\omega|>1\), \(\overline{K}(\omega) = I\) and \(\phi_{zu}(\omega) = \overline{K}(\omega)D \overline{K}'(\omega^{-1})\) for some \(D>0\). Let \(\overline{G}_1, \overline{G}_2, \overline{G}_3\) and \(\overline{G}_4\) be the corresponding transfer functions. The original open loop transfer functions are recoverable if

\[
G_1 = \overline{G}_1 ; \quad G_3 = \overline{G}_3 ; \quad G_2 = \overline{G}_2 \gamma ; \quad G_4 = \overline{G}_4 \gamma
\]
Figure 3.3: A General Closed Loop Configuration.
where $V(\mathcal{J})$ and $W(\mathcal{J})$ are rational matrices with

$$V(\mathcal{J})D_{11}V'(\mathcal{J}^{-1}) = \overline{D}_{11} \quad W(\mathcal{J})D_{22}W'(\mathcal{J}^{-1}) = \overline{D}_{22} \quad .$$

**Theorem 3.1:**

i) a sufficient condition for recoverability of $G_1$ to $G_4$ is that $D$ in (3.14) be block diagonal.

ii) a necessary and sufficient condition for recoverability of $G_1$ to $G_4$ is that the transformation from $K$ to $\overline{K}$ be block diagonal.

This result means that if the noise sequence in the forward and reverse path are uncorrelated, then the correct values of $G_1$ and $G_3$ can be obtained from the stable minimum phase spectral factor of $\phi_{zu}(\mathcal{J})$. As mentioned in [65], (see also Vorchik et al. [50, Case 1]) one can also argue that in this case, the likelihood function factors into two parts, one for each of the forward and backward paths.

Therefore, provided $G_1$ and $G_2$ have no common parameters with $G_3$ and $G_4$, the maximum likelihood method can be applied to the data $z(t)$ and $u(t)$ as if it were open loop data and $G_1$ and $G_3$ can be uniquely determined independantly.

It is essential, however, for the above argument to hold that a delay be assumed in both $G_1(\mathcal{J})$ and $G_3(\mathcal{J})$ (see assumptions with Figure 3.3). If the delay is absent from either $G_1$ or $G_3$, or in the equivalent situation where $g_1(t)$ and $g_2(t)$ are correlated, one could
still recover $G_1$ to $G_4$ in the non generic cases covered by (ii) of the above theorem.

Another result of [65] is for the feedback noise free model where $G_4 \equiv 0$ and $G_3$ is fully known and does not have to be strictly proper (i.e. could contain instantaneous transmission). In this case it was found that $G_1$ is uniquely determined if the feedback regulator is sufficiently complex. To avoid getting too far in side issues we will only give, in the vein of Example 3.2, the following single input single output example from Söderström et al [66]. (Similar results have been obtained by Vorchik et al. [50, case 3] and Wellstead and Edmunds [67]).

Consider the system given by

$$A(\mathcal{J})z(\mathcal{J}) = \mathcal{J}^{-k}B(\mathcal{J})u(\mathcal{J}) + C(\mathcal{J})\xi(\mathcal{J}) \quad ; \quad k \geq 0 \quad (3.15)$$

and the regulator

$$F(\mathcal{J})u(\mathcal{J}) = D(\mathcal{J})z(\mathcal{J}) \quad (3.16)$$

where

$$A(\mathcal{J}) = 1 + a_1 \mathcal{J}^{-1} + \ldots + a_n \mathcal{J}^{-n} \quad ; \quad a_n \neq 0 \quad (3.17)$$

$$C(\mathcal{J}) = 1 + c_1 \mathcal{J}^{-1} + \ldots + c_n \mathcal{J}^{-n} \quad ; \quad c_n \neq 0 \quad (3.18)$$

$$F(\mathcal{J}) = 1 + f_1 \mathcal{J}^{-1} + \ldots + f_n \mathcal{J}^{-n} \quad ; \quad f_n \neq 0 \quad (3.19)$$
and

\[ B(\mathcal{Y}) = b_1 \mathcal{Y}^{-1} + \ldots + b_n \mathcal{Y}^{-n_b} \quad ; \quad b_1 \neq 0, b_n \neq 0 \]  

(3.20)

\[ D(\mathcal{Y}) = d_0 + d_1 \mathcal{Y}^{-1} + \ldots + d_{n_d} \mathcal{Y}^{-n_d} \quad ; \quad d_{n_d} \neq 0 \]  

(3.21)

and assume that \( F(\mathcal{Y}) \) and \( D(\mathcal{Y}) \) are relatively prime and that there is no common factor to \( A(\mathcal{Y}) \), \( B(\mathcal{Y}) \) and \( C(\mathcal{Y}) \). Further, assume that \( C(\mathcal{Y}) \) and \( \left[ A(\mathcal{Y})F(\mathcal{Y}) - \mathcal{Y}^{-k}B(\mathcal{Y})D(\mathcal{Y}) \right] \) have exactly \( n_p \) common factors. Rewrite (3.15) as

\[ z(\mathcal{Y}) = G_1(\mathcal{Y})u(\mathcal{Y}) + G_2(\mathcal{Y})\xi(\mathcal{Y}) \]  

(3.22)

where

\[ G_1(\mathcal{Y}) = \mathcal{Y}^{-k}A^{-1}(\mathcal{Y})B(\mathcal{Y}) \]  

(3.23)

and

\[ G_2(\mathcal{Y}) = A^{-1}(\mathcal{Y})C(\mathcal{Y}) \]  

(3.24)

are the input-output and noise-output transfer functions respectively.

Now by using a prediction error type identification method those open loop transfer functions \( G_1(\mathcal{Y}) \) and \( G_2(\mathcal{Y}) \) are recoverable if and only if

(see [66])

\[ \max(n_{f - n_b}, n_{d + k - n_a}) - n_p > 0 \]  

(3.25)

This result determines the order of complexity required from the regulator.
(3.15) to make the open loop transfer functions identifiable.

Note in passing that in Example 3.2 we had

\[ n_a=1, \quad n_b=1, \quad k=0, \quad n_f=0, \quad n_d=0, \quad n_p=0 \]

violating (3.25) whereas with the regulator (3.13) \( n_d=1 \) and (3.25) is satisfied.

In the next section we return to linear time invariant state space models to discuss identifiability of open loop dynamics from data collected under time invariant feedback.

### 3.3 LTI State Space Models Under Feedback

Recall, as in Section 2.6 above, the frequency domain description of the steady state Kalman filter

\[ r(y;\alpha) = H(y;\alpha)z(y) - H(y;\alpha) G(y;\alpha)u(y) \quad (3.36) \]

where

\[ G(y;\alpha) = C_\alpha(y(I-A_\alpha)\alpha^{-1}B_\alpha \quad (3.37) \]

and

\[ H(y;\alpha) = C_\alpha(y(I-A_\alpha)^{-1}A_\alpha H_\alpha + I \quad (3.38) \]

are the input-output and residual-output transfer functions respectively.

It will be useful to examine first the following result obtained by Söderström, Ljung and Gustavsson [48] for:
a) the feedback structure

\[ u(y) = F_1(y)z(y) + L_1(y)v(y) \quad 1 \leq i \leq \gamma \]  

(3.39)

where \( v(t) \in \mathbb{R}^q \) is a signal vector with properties given below and where the feedback law is assumed to shift between \( \gamma \) different regimes with \( \gamma \geq 1 \).

b) model sets containing the true system

c) a general prediction error identification method which satisfies (2.66) and therefore, as mentioned in Section 2.4, does not contain our present ML identification method.

Theorem 3.2 ([48]):

Assume that

i) \( G(\infty, *)F_1(\infty) = 0 \)  

(\( 1 \leq i \leq \gamma \))

(i.e. there is at least one delay around the loop)

ii) the closed loop system is asymptotically stable

iii) \( v(y) = K_1(y) r(y;*) + \tilde{v}(y) \)

where \( K_1(y) \) is a causal, asymptotically stable filter, \( \tilde{v}(t) \) is independent of the residuals \( r(t;*) \) and persistently exciting (i.e. at least some part of the feedback is independent persistently

---

\(^1\) Those shifts actually make the feedback time-variant but we will assume that each regime is used long enough for initial effects to be neglected and that stationary operation is maintained.
exciting noise)

\[ G(\omega; a)L_i(\omega)K_i(\omega) = 0 \quad 1 \leq i \leq \gamma, \quad \alpha \in \mathcal{P} \]

Then, the open-loop transfer functions \( G(y; \ast) \) and \( H(y; \ast) \) are recoverable from any class of models satisfying (b) above if and only if

\[
\text{rank} \begin{bmatrix}
I & \cdots & I & 0 & \cdots & 0 \\
F_1(y) & \cdots & F_\gamma(y) & L_1(y) & \cdots & L_\gamma(y)
\end{bmatrix} = m+r \quad \text{a.e.} y. \quad (3.40)
\]

where \( I \) is \( r \times r \), \( 0 \) is \( r \times q \), \( F_i \) is \( m \times r \) and \( L_i \) is \( m \times q \) so that the above composite matrix is \( (m+r) \times (r+q) \).

This result summarizes the conditions required to make the open loop transfer functions identifiable from the closed loop data \( u(t) \) and \( z(t) \).

For example if there is no shift of feedback laws but the number of injected independent inputs is equal to the number of inputs \( u(t) \) to the system (i.e. \( \gamma = 1 \), \( q = m \), \( L_1(y) \) is nonsingular), then

\[
\text{rank} \begin{bmatrix}
I & 0 \\
F_1(y) & L_1(y)
\end{bmatrix} = m+r
\]

and \( G(y; \ast) \) and \( H(y; \ast) \) are recoverable from any class of models containing the true system no matter what the feedback \( F_1(y) \) might be. (see results of [65] above for configuration of figure 3.3). If on the other hand there are no injected independent inputs \( (q=0) \) then a shift between
\[ y > 1 + \frac{m}{r} \]

appropriately chosen feedback laws \( F_i(y) \) is required (see Ljung et al. [46]). Finally if there is no shift of feedback laws and no independent injected input (i.e. \( y=0 \) and \( q=0 \)) then the above rank condition cannot be met and \( G(y;*) \) and \( H(y;*) \) are not always recoverable. This was the last case described in Section 3.2 where extra conditions were imposed on the model set (typically feedback of "higher order" than open loop model) to obtain identifiability of the open loop transfer functions.

But note also, as we did in Section 2.4, that the prediction error estimator used above only selects the right predictor model asymptotically via \( G(y; \alpha) \) and \( H(y; \alpha) \). It will not necessarily give however, as does the maximum likelihood estimator, a good "second order model" \( S_{\alpha} \) for the residuals \( r(t;*) \).

Recall now the identifiability condition

\[ H(y; \alpha) \neq H(y; \alpha_0) \quad (3.41) \]

or

\[ G(y; \alpha) \neq G(y; \alpha_0) \quad (3.42) \]

or

\[ S_{\alpha} \neq S_{\alpha_0} \quad (3.43) \]

for all \( \alpha \neq \alpha_0 \), \( \alpha \in \mathcal{P} \). If (3.41) and (3.42) were violated for some \( \alpha \neq \alpha_0 \), (3.43) does not depend on the closed loop nature at all and
the arguments used in Chapter 2 still apply. We therefore conclude that, in the case where the model set contains the true system, the true parameter $\alpha_0 = \ast$ will be identifiable with our present maximum likelihood method and with feedback laws satisfying the assumptions of Theorem 3.2, if and only if the rank condition (3.40) and one of the conditions (3.41)-(3.43) hold.

Similarly, in the case where the model set does not contain the true system and $\alpha_0$ is the tradeoff choice discussed in Chapter 2, this choice will be unique at any particular large value of $t$ if the feedback law regulating the true system satisfies the conditions described above and only if at least one the conditions (3.41)-(3.43) holds.

3.4. Summary

In this chapter we concentrated our attention on the problems of identifying open loop dynamics from data collected under closed loop operation. In Section 3.1 we looked at two archetypal examples of such problems. In Section 3.2 standard treatments were reviewed in more detail than in Chapter 1. Section 3.3 finally draws on a general result by Söderström et al. [48] to look at conditions of identifiability of linear state space models from maximum likelihood estimates.
Recall from Chapter 2 that the maximum likelihood estimate $\hat{\theta}_t$ can be evaluated by minimizing the likelihood function defined in (2.21) and (2.22). To carry out the minimization, one of a host of numerical optimization techniques can be used. These generally require the gradient and sometimes the Hessian (or an approximation thereof) of the likelihood function. In this chapter we will describe standard as well as new ways to obtain those quantities. The likelihood function gradient is first discussed and three approaches to its evaluation are given. A similar analysis is then applied to the Hessian. Then the evaluation of the information matrix as an approximation to the Hessian is discussed. Note that some of this work has already been reported in Sandell and Yared [17] but is included here for completeness.

4.1 Gradient Evaluation

4.1.1 The forward filter

As mentioned in Chapter 2, we obtain by straightforward differentiation of the likelihood function

$$
\zeta(z^t; \alpha) = \sum_{\tau=0}^{t} \zeta(z(\tau) | z^{T-1}; \alpha)
$$

(4.1)

where
\[ \zeta(z(T)|z_{\tau}^{-1};\alpha) = \frac{1}{2} \log \det S_{\alpha}(\tau) + \frac{1}{2} r'(\tau;\alpha)S_{\alpha}^{-1}(\tau)r(\tau;\alpha) \] (4.2)

its gradient

\[ \frac{\partial \zeta(z_{\tau};\alpha)}{\partial \alpha_i} = \sum_{\tau=0}^{T} \frac{\partial \zeta(z(\tau)|z_{\tau}^{-1};\alpha)}{\partial \alpha_i} \] (4.3)

where \( \alpha_i \) denotes the \( i \)th component of \( \alpha \) and where

\[ \frac{\partial \zeta(z(\tau)|z_{\tau}^{-1};\alpha)}{\partial \alpha_i} = r'(\tau;\alpha)S_{\alpha}^{-1}(\tau) \frac{\partial r(\tau;\alpha)}{\partial \alpha_i} \]

\[ - \frac{1}{2} r'(\tau;\alpha)S_{\alpha}^{-1}(\tau) \left[ \frac{\partial S_{\alpha}(\tau)}{\partial \alpha_i} S_{\alpha}^{-1}(\tau)r(\tau;\alpha) \right] \]

\[ + \frac{1}{2} \text{tr} \left[ S_{\alpha}^{-1}(\tau) \frac{\partial S_{\alpha}(\tau)}{\partial \alpha_i} \right] \] (4.4)

Now \( \frac{\partial r(\tau;\alpha)}{\partial \alpha_i} \) and \( \frac{\partial S_{\alpha}(\tau)}{\partial \alpha_i} \) are in turn evaluated by differentiating the Kalman filter equations (2.15)-(2.19) to obtain:

From (2.16)

\[ \frac{\partial r(t;\alpha)}{\partial \alpha_i} = - \frac{\partial c_{\alpha}(t)}{\partial \alpha_i} \hat{x}(t;\alpha) - c_{\alpha}(t) \frac{\partial \hat{x}(t;\alpha)}{\partial \alpha_i} \] (4.5)

From (2.15), the filter sensitivity equation

\[ \frac{\partial \hat{x}(t+1;\alpha)}{\partial \alpha_i} = \frac{\partial x(t;\alpha)}{\partial \alpha_i} + \omega(t;\alpha) \] (4.6)
where

\[ \tilde{A}_\alpha(t) \triangleq A_\alpha(t) [I - H_\alpha(t) C_\alpha(t)] \]  

\[ \omega_i(t;\alpha) \triangleq \frac{\partial A_\alpha(t)}{\partial \alpha_i} \hat{x}(t;\alpha) + \frac{\partial B_\alpha(t)}{\partial \alpha_i} u(t) + \frac{\partial [A_\alpha(t) H_\alpha(t)]}{\partial \alpha_i} z(t) \]

\[ = \begin{bmatrix} \frac{\partial A_\alpha(t)}{\partial \alpha_i} - A_\alpha(t) H_\alpha(t) \frac{\partial C_\alpha(t)}{\partial \alpha_i} \end{bmatrix} \hat{x}(t;\alpha) + \frac{\partial B_\alpha(t)}{\partial \alpha_i} u(t) + \frac{\partial [A_\alpha(t) H_\alpha(t)]}{\partial \alpha_i} r(t;\alpha) \]

(4.8)

and where, from (2.17)

\[ \frac{\partial h_\alpha(t)}{\partial \alpha_i} = \begin{bmatrix} \frac{\partial S(t)}{\partial \alpha_i} C_\alpha(t) + S_\alpha(t) \frac{\partial \alpha_i}{\partial \alpha_i} - H_\alpha(t) \frac{\partial C_\alpha(t)}{\partial \alpha_i} \end{bmatrix} s^{-1}(t) \]

(4.9)

From (2.18)

\[ \frac{\partial s(t)}{\partial \alpha_i} = C_\alpha(t) \frac{\partial s(t)}{\partial \alpha_i} + C_\alpha(t) + \frac{\partial \theta(t)}{\partial \alpha_i} + \frac{\partial C_\alpha(t)}{\partial \alpha_i} \Sigma(t) C_\alpha(t) + C_\alpha(t) \Sigma(t) \frac{\partial C_\alpha(t)}{\partial \alpha_i} \]

(4.10)

And from (2.19), the Riccati sensitivity equation

\[ \frac{\partial s(t+1)}{\partial \alpha_i} = \tilde{A}_\alpha(t) \frac{\partial s(t)}{\partial \alpha_i} + \tilde{A}_\alpha(t) + \Omega_i(t;\alpha) + \Omega_i^t(t;\alpha) \]

(4.11)
where

\[ \Omega_i(t; \alpha) = \left[ \frac{\partial \alpha_i(t)}{\partial \alpha_1} - \frac{\partial \alpha_i(t)}{\partial C_i(t)} \right] S_i(t) A_i(t) + \frac{1}{2} \frac{\partial}{\partial \alpha_1} \left[ L_i(t) Z_i(t) L_i'(t) \right] + \frac{1}{2} \frac{\partial}{\partial \alpha_1} \left[ \frac{\partial H_i(t)}{\partial \alpha_1} A_i(t) \right] + \frac{\partial}{\partial \alpha_1} \left[ \frac{\partial H_i'(t) A_i'(t)}{\partial \alpha_1} \right] \]

(4.12)

The derivation of (4.12) is given in Appendix B, and figure 4.1 summarizes the steps described above in the evaluation of \( \frac{\partial \xi(z(t))}{\partial \alpha_1} \).

Since the recursive equations (4.6) and (4.11) run forward in time, this approach can be referred to as the forward filter evaluation of \( \frac{\partial \xi}{\partial \alpha} \).

Recalling that \( \alpha \) is \( \ell \)-dimensional, this evaluation then requires:

1 Riccati equation (2.19) or \( \frac{n(n+1)}{2} \) equations

1 filter equation (2.15) or \( n \) equations

\( \ell \) Lyapunov equations (Riccati sensitivity equations)

(4.11) or \( \frac{n(n+1)}{2} \) \( \ell \) equations

\( \ell \) filter sensitivity equations (4.6) 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 straightforward fashion.

Now there are alternate approaches, with the possibility of reduced computation, to the problem of evaluating the gradient of the likelihood function. Before presenting them in the next sections,
Figure 4.1: Forward Filter Evaluation of $\frac{\partial \mathbf{r}}{\partial \alpha}$. 

\[ \Omega_i(t;\alpha), i=1,\ldots,l \]

\[ \Sigma_{\alpha}(t) \]

\[ \hat{x}(t;\alpha) \]

\[ u(t) \]

\[ x(t;\alpha) \]

\[ \Sigma_{\alpha}(t) \]

\[ \frac{\partial \Sigma_{\alpha}(t)}{\partial \alpha_i}, i=1,\ldots,l \]

\[ \frac{\partial \hat{x}(z(T);\alpha)}{\partial \alpha_i} \]

\[ \frac{\partial \Sigma_{\alpha}(t)}{\partial \alpha_i} \]

\[ \gamma(z^T;\alpha) \]

\[ \Gamma(z^T;\alpha) \]

\[ \frac{\partial \mathbf{z}(z^T;\alpha)}{\partial \alpha} \]
reexpress the forward filter equations in the following convenient form:

\[
\frac{\partial \zeta(z^t; \alpha)}{\partial \alpha_i} = \sum_{\tau=0}^{t} \beta_i(z^\tau; \alpha) + \sum_{\tau=0}^{t} \gamma'(z^\tau; \alpha) \frac{\partial \mathbf{x}(\tau; \alpha)}{\partial \alpha_i} + \sum_{\tau=0}^{t} tr \left[ \Gamma(z^\tau; \alpha) \frac{\partial \Sigma(\tau)}{\partial \alpha_i} \right] \tag{4.14}
\]

where

\[
\frac{\partial \hat{x}(t+1; \alpha)}{\partial \alpha_i} = \bar{\alpha}(t) \frac{\partial \mathbf{x}(t; \alpha)}{\partial \alpha_i} - \bar{\alpha}(t) \frac{\partial \Sigma(t)}{\partial \alpha_i} \gamma(z^t; \alpha) + \omega_i(t; \alpha) \tag{4.15}
\]

\[
\frac{\partial \Sigma(t+1)}{\partial \alpha_i} = \bar{\alpha}(t) \frac{\partial \Sigma(t)}{\partial \alpha_i} - \bar{\alpha}(t) + \Omega_i(t; \alpha) + \Omega'(t; \alpha) \tag{4.16}
\]

and where, dropping the \( t \) and \( \alpha \) arguments from matrix notations for clarity,

\[
\beta_i(z^t; \alpha) \triangleq \frac{1}{2} \text{tr} \left\{ S^{-1}(I - r(t; \alpha)r'(t; \alpha)S^{-1}) [2CS \frac{\partial c'}{\partial \alpha_i} + \frac{\partial \theta}{\partial \alpha_i}] - r'(t; \alpha)S^{-1} \frac{\partial c}{\partial \alpha_i} \right\} \tag{4.17}
\]
\[ \gamma(z^t; \alpha) \triangleq - C' S^{-1} r(t; \alpha) \] (4.18)

\[ \Gamma(z^t; \alpha) \triangleq \frac{1}{2} C' S^{-1} [I - r(t; \alpha) r'(t; \alpha) S^{-1}] C \] (4.19)

\[ \tilde{\omega}_i(t; \alpha) = \left[ \frac{\partial A}{\partial \alpha_i} - \lambda \frac{\partial C}{\partial \alpha_i} \right] x(t; \alpha) + \frac{\partial B}{\partial \alpha_i} u(t) \]

\[ + \left\{ \frac{\partial A}{\partial \alpha_i} H + A \left[ \sum \frac{\partial C'}{\partial \alpha_i} - H \left( \frac{\partial \theta}{\partial \alpha_i} - \frac{\partial C}{\partial \alpha_i} \Sigma C' - C \Sigma \frac{\partial C'}{\partial \alpha_i} \right) S^{-1} \right] r(t; \alpha) \] (4.20)

The derivation of the above equations is given in Appendix C. Based on the form (4.14)-(4.16), the theory of adjoint equations, briefly summarized in Appendix D suggests the possibility of replacing the 2 forward filter equations (4.15) and (4.16) with 2 adjoint equations running backward in time and using the adjoint variables to evaluate the second and third terms of (4.14). However, for further reference we first develop the following:

4.1.2 The mixed filter

This approach consists of running forward in time the Lyapunov equation (4.16) to evaluate \( \frac{\partial \Sigma}{\partial \alpha_i} (t) \) and form the driving terms

\[ \omega_i(t; \alpha) = - \frac{\partial A}{\partial \alpha_i} \frac{\partial \Sigma}{\partial \alpha_i} (t) \gamma(z^t; \alpha) + \tilde{\omega}_i(t; \alpha) \] (4.21)

of the filter sensitivity equations (4.15); then running backward in time the adjoint of (4.15); then using the solutions of those to
evaluate the second term of (4.14). In other terms we have by direct application of Corollary 1 of Appendix D.

\[
\frac{\partial \zeta(z^T;\alpha)}{\partial \alpha_i} = \sum_{\tau=0}^{t} \beta_1(z^T;\alpha) + \lambda'(0,\alpha) \frac{\partial \hat{X}(0;\alpha)}{\partial \alpha_i} + \sum_{\tau=1}^{t} \lambda'(\tau;\alpha) \omega_1(\tau-1;\alpha) + \sum_{\tau=0}^{t} \text{tr} \left[ \Gamma(z^T;\alpha) \frac{\partial \Sigma(\tau)}{\partial \alpha_i} \right]
\]

(4.22)

where

\[
\lambda(\tau;\alpha) = \frac{\lambda'}{\alpha_1}(\tau) \lambda(\tau+1;\alpha) + \gamma(z^T;\alpha)
\]

(4.23)

with

\[
\lambda(t;\alpha) = \gamma(z^T;\alpha)
\]

replaces (4.15) in the forward form (4.14-4.16).

Because (4.23) runs backward in time, we refer to (4.22, 4.23, 4.16) as the mixed filter evaluation of \( \frac{\partial \zeta}{\partial \alpha_i} \), summarized in Figure 4.2. Furthermore, in comparison with the forward filter approach, this approach requires

1 Riccati equation (2.19) or \( \frac{n(n+1)}{2} \) equations
1 filter equation (2.15) or \( n \) equations
1 Lyapunov equation (4.16) or \( \frac{n(n+1)}{2} \) equations
1 adjoint filter sensitivity equation (4.23) or \( n \) equations.

This apparently represents less computation but the storage burden is clearly increased.
Figure 4.2: Mixed Filter Evaluation of $\frac{\partial \mathcal{L}}{\partial \alpha}$.
4.1.3 The backward filter

This is the approach, mentioned at the end of Section 4.1.1, where we have, by application of Corollary 3 of Appendix D on the form

(4.14-4.16)

\[
\frac{\partial \xi(z; t; \alpha)}{\partial \alpha_i} = \sum_{\tau=0}^{t} \beta_{\tau} (z^\tau; \alpha)
+ \lambda'(\tau; \alpha) \frac{\partial \hat{x}(0; \alpha)}{\partial \alpha_i} + \sum_{\tau=1}^{t} \lambda'(\tau; \alpha) \omega_{\tau}(\tau-1; \alpha)
+ \text{tr} \left[ \Lambda(0; \alpha) \frac{\partial \Sigma(0)}{\partial \alpha_i} \right] + \sum_{\tau=1}^{t} \text{tr} \left[ \Lambda(\tau; \alpha) \left( \Omega_{\tau}(\tau-1; \alpha) + \Omega'_{\tau}(\tau-1; \alpha) \right) \right]
\]

(4.24)

where

\[
\Lambda(\tau; \alpha) = \Lambda'_{\tau}(\tau) \lambda(\tau+1; \alpha) + \gamma(z^\tau; \alpha)
\]

(4.25)

with

\[
\lambda(t; \alpha) = \gamma(z^t; \alpha)
\]

and

\[
\Lambda(\tau; \alpha) = \Lambda'_{\tau}(\tau) \lambda(\tau+1; \alpha) \Lambda_{\tau}(\tau) - \Lambda'_{\tau}(\tau) \lambda(\tau+1; \alpha) \gamma'(z^\tau; \alpha) + \Gamma(z^\tau; \alpha)
\]

(4.26)

with

\[
\Lambda(t; \alpha) = \Gamma(z^t; \alpha)
\]
Equations (4.25) and (4.26) replace (4.15) and (4.16) respectively. Because they both run backward in time we refer to (4.24-4.26) as the backward filter evaluation of $\frac{\partial \xi}{\partial q}$, summarized in Figure 4.3. In comparison with the previous two filters, this approach requires.

- 1 Riccati equation (2.19) or $\frac{n(n+1)}{2}$ equations
- 1 filter equation (2.15) or $n$ equations
- 1 adjoint Lyapunov equation (4.26) or $n^2$ equations
- 1 adjoint filter sensitivity equation (4.25) or $n$ equations

or roughly the equivalent of 2 (1 forward and 1 backward) Kalman filters.

Note however that we have lost the symmetry in equation (4.26) due to the coupling between the sensitivity equations. Here again we find the same computation versus storage tradeoff and the relative advantages of the three approaches depend heavily on the particular application.

Note also that $\lambda(t)$ and $\Lambda(t)$ in the adjoint equations have interpretations as Lagrange multipliers or costates of an optimal control problem (see e.g. [11]).

4.1.4 Gradient evaluation for time invariant systems

As was the case for the evaluation of the likelihood function (see Section 2.2), the equations above simplify significantly for a time invariant system where the observations are processed by a steady state Kalman filter. In particular, the time varying matrix equation (4.16)
RICCATI EQUATION \((a)\)

\[ u(T) \times (1; a) \]

\[ z(T) \times (T) \]

FILTER \((a)\)

\[ \Sigma_a(T) \]

STORE

\[ \hat{x}(T; \alpha) \]
\[ u(T) \]
\[ z(T) \]
\[ \Sigma_a(T) \]
\[ \tau = 1, \ldots, t \]

ADJOINT RICCATI SENSITIVITY EQUATION \((a)\)

\[ \Gamma(z^T; \alpha) \]

\[ \gamma(z^T; \alpha) \]

\[ \Lambda_a(T) \]

ADJOINT FILTER SENSITIVITY EQUATION \((a)\)

\[ \lambda(T) \]

GRADIENT COMPUTATION

\[ \frac{\partial \zeta(z^T; \alpha)}{\partial \alpha} \]

Figure 4.3: Backward Filter Evaluation of \( \frac{\partial \zeta}{\partial \alpha} \).
reduces to the steady state version, the corresponding algebraic Lyapunov equation

\[
\frac{\partial E}{\partial \alpha_i} = \frac{\partial A}{\partial \alpha_i} \bar{A} + \frac{\partial A}{\partial \alpha_i} A^T + \Omega_i (\alpha) + \Omega_i' (\alpha)
\]  

Note that there exists an efficient method (the Bartels-Stewart method [68]) for obtaining directly the solution of this equation. The major part of the computations in this algorithm involve only operations on \( \bar{A} \) which do not have to be repeated when the equation is resolved for the various \( \Omega_i \), \( i=1,...,\lambda \). Note also that the forward filter approach now involves solving the same \((1+\lambda)\) filter equations (2.15) and (4.15) but only one algebraic Riccati equation (2.37) and \( \lambda \) algebraic Lyapunov equations (4.27).

With the backward filter approach however, note that the adjoint Lyapunov equation (4.26) is driven by \( \lambda(t+1;\alpha) \), the solution of the adjoint filter sensitivity equation (4.25), as well as by the time varying \( \Gamma(z;\alpha) \). Therefore, no steady state version exists in this case and we still have to solve 2 filter equations (2.15) and (4.25), one algebraic Riccati equation (2.37) but one time varying Lyapunov equation (4.26).

As a consequence, it might be significantly more economical in the time invariant case to use the mixed filter approach where we need to solve 2 filter equations (2.15) and (4.23), one algebraic Riccati
equation (2.37) and \textit{algebraic} Lyapunov equations (4.27).

4.2 Hessian Evaluation

Here again, it is conceptually straightforward to obtain the Hessian \( \frac{\partial^2 \xi(t;\alpha)}{\partial \alpha^2} \) of the likelihood function by differentiating one more time the above sensitivity equations. One could also collect terms to obtain the convenient form

\[
\begin{align*}
\frac{\partial^2 \xi(t;\alpha)}{\partial \alpha_i \partial \alpha_j} &= \sum_{\tau=0}^{t} \beta_{ij}(z^\tau;\alpha) \\
&+ \sum_{\tau=0}^{t} \gamma'(z^\tau;\alpha) \frac{\partial^2 x(\tau;\alpha)}{\partial \alpha_i \partial \alpha_j} \\
&+ \sum_{\tau=0}^{t} \text{tr} \left[ \Gamma(z^\tau;\alpha) \frac{\partial^2 \Sigma(\tau)}{\partial \alpha_i \partial \alpha_j} \right] \\
&= \sum_{\tau=0}^{t} \beta_{ij}(z^\tau;\alpha) \\
&+ \sum_{\tau=0}^{t} \gamma'(z^\tau;\alpha) \frac{\partial^2 x(\tau;\alpha)}{\partial \alpha_i \partial \alpha_j} \\
&+ \sum_{\tau=0}^{t} \text{tr} \left[ \Gamma(z^\tau;\alpha) \frac{\partial^2 \Sigma(\tau)}{\partial \alpha_i \partial \alpha_j} \right] \\
\end{align*}
\]  

(4.28)

where

\[
\begin{align*}
\frac{\partial^2 x(t+1;\alpha)}{\partial \alpha_i \partial \alpha_j} &= \bar{\alpha}_i(t) \frac{\partial^2 x(t;\alpha)}{\partial \alpha_i \partial \alpha_j} - \bar{\alpha}_j(t) \frac{\partial^2 \Sigma(t)}{\partial \alpha_i \partial \alpha_j} \gamma(z^t;\alpha) + \bar{\omega}_{ij}(t;\alpha) \\
\frac{\partial^2 \Sigma(t+1)}{\partial \alpha_i \partial \alpha_j} &= \bar{\alpha}_i(t) \frac{\partial^2 \Sigma(t)}{\partial \alpha_i \partial \alpha_j} - \bar{\alpha}_j(t) \frac{\partial^2 \Sigma(t)}{\partial \alpha_i \partial \alpha_j} \bar{\alpha}'_i(t) + \Omega_{ij}(t;\alpha) + \Omega'_{ij}(t;\alpha) \\
\end{align*}
\]  

(4.29)

and where \( \gamma(z^t;\alpha) \) and \( \Gamma(z^t;\alpha) \) are as defined in (4.18) and (4.19) respectively, and
$$\beta_{ij}(z^t; \alpha) = \frac{1}{2} \text{tr} \left\{ s^{-1} [I - r' s^{-1}] \left[ 2c \frac{\partial^2 C'}{\partial \alpha_i \partial \alpha_j} + 2c \frac{\partial C}{\partial \alpha_i} \frac{\partial C'}{\partial \alpha_j} + 2c \frac{\partial C}{\partial \alpha_j} \frac{\partial C'}{\partial \alpha_i} + 2 \frac{\partial C}{\partial \alpha_i} \frac{\partial C'}{\partial \alpha_j} + \frac{\partial^2 \theta}{\partial \alpha_i \partial \alpha_j} \right] \right\}$$

$$- r's^{-1} \left[ \frac{\partial^2 C}{\partial \alpha_i \partial \alpha_j} x + \frac{\partial C}{\partial \alpha_i} \frac{\partial x}{\partial \alpha_j} + \frac{\partial C}{\partial \alpha_j} \frac{\partial x}{\partial \alpha_i} \right]$$

$$+ \text{tr} \left\{ s^{-1} \left[ \frac{\partial x}{\partial \alpha_i} \frac{\partial x}{\partial \alpha_j} - \frac{\partial s}{\partial \alpha_j} s^{-1} \frac{\partial x}{\partial \alpha_i} r' - \frac{\partial s}{\partial \alpha_i} s^{-1} \frac{\partial x}{\partial \alpha_j} r' \right. \right.$$

$$+ \left. \frac{\partial s}{\partial \alpha_i} s^{-1} \frac{\partial s}{\partial \alpha_j} s^{-1} rr' - \frac{1}{2} \frac{\partial s}{\partial \alpha_j} s^{-1} \frac{\partial s}{\partial \alpha_i} \right\} \right\}$$

$$= \frac{\partial A}{\partial \alpha_j} \frac{\partial \hat{x}}{\partial \alpha_i} + \frac{\partial A}{\partial \alpha_i} \frac{\partial \hat{x}}{\partial \alpha_j} + \left[ \frac{\partial^2 A}{\partial \alpha_i \partial \alpha_j} - \frac{\partial A}{\partial \alpha_i} \frac{\partial H}{\partial \alpha_j} - \frac{\partial A}{\partial \alpha_j} \frac{\partial H}{\partial \alpha_i} - \frac{\partial (AH)}{\partial \alpha_i \partial \alpha_j} \right] x$$

$$+ \left[ \frac{\partial^2 A}{\partial \alpha_i \partial \alpha_j} H + \frac{\partial A}{\partial \alpha_i} \frac{\partial H}{\partial \alpha_j} + \frac{\partial A}{\partial \alpha_j} \frac{\partial H}{\partial \alpha_i} - A \frac{\partial H}{\partial \alpha_i} \frac{\partial s}{\partial \alpha_j} s^{-1} \right.$$

$$- A \frac{\partial H}{\partial \alpha_i} \frac{\partial s}{\partial \alpha_j} s^{-1}$$

$$+ \frac{\partial (\partial C}{\partial \alpha_i} \frac{\partial C'}{\partial \alpha_j} + \frac{\partial C}{\partial \alpha_j} \frac{\partial C'}{\partial \alpha_i} + \frac{\partial C}{\partial \alpha_i} \frac{\partial C'}{\partial \alpha_j} + \frac{\partial C}{\partial \alpha_j} \frac{\partial C'}{\partial \alpha_i} \right) s^{-1}$$

$$- \frac{\partial C}{\partial \alpha_i} \frac{\partial C'}{\partial \alpha_j} s^{-1} - \frac{\partial C}{\partial \alpha_i} \frac{\partial C'}{\partial \alpha_j} H$$

$$- A \frac{\partial C}{\partial \alpha_j} \frac{\partial C'}{\partial \alpha_i} c' + \frac{\partial C}{\partial \alpha_i} \frac{\partial C'}{\partial \alpha_j} c' + \frac{\partial C}{\partial \alpha_i} \frac{\partial C'}{\partial \alpha_j} + \frac{\partial C}{\partial \alpha_j} \frac{\partial C'}{\partial \alpha_i} \right] r$$

$$+ \frac{\partial^2 B}{\partial \alpha_i \partial \alpha_j} u \right\}$$

(4.31)
\[ \Omega_{ij}(t; \alpha) = \left[ \frac{\partial^2 A}{\partial \alpha_i \partial \alpha_j} - AH \frac{\partial^2 C}{\partial \alpha_i \partial \alpha_j} - \frac{\partial (AH)}{\partial \alpha_i} \frac{\partial C}{\partial \alpha_j} \right] \Sigma A' \]

\[ + \left[ \frac{\partial A}{\partial \alpha_i} - AH \frac{\partial C}{\partial \alpha_i} \right] \left[ \frac{\partial L}{\partial \alpha_j} A' + \Sigma \frac{\partial A'}{\partial \alpha_j} \right] \]

\[ + \frac{\partial A}{\partial \alpha_j} \frac{\partial L}{\partial \alpha_i} A', \]

\[ + \frac{1}{2} \frac{\partial^2 (L^2 L')}{\partial \alpha_i \partial \alpha_j} + \frac{\partial A}{\partial \alpha_j} H \frac{\partial \theta}{\partial \alpha_i} H' A' + A \frac{\partial H}{\partial \alpha_j} \frac{\partial \theta}{\partial \alpha_i} H' A' \]

\[ + \frac{1}{2} AH \frac{\partial^2 \theta}{\partial \alpha_i \partial \alpha_j} H' A' \]  

(4.33)

where the first order sensitivity quantities \( \frac{\partial \hat{x}(t; \alpha)}{\partial \alpha_i}, \frac{\partial \hat{r}(t; \alpha)}{\partial \alpha_i} \), etc. are obtained as described in Section 4.1.

The above expressions for \( \beta_{ij}(z^t; \alpha), \omega_{ij}(t; \alpha) \) and \( \Omega_{ij}(t; \alpha) \) are rather involved but might simplify significantly in specific instances where the unknown parameter does not enter all the model matrices.

In practice however, the main reason for not using the above direct method to evaluate the Hessian is because it involves solving

1 Riccati equation (2.19)

1 filter equation (2.15)

1 Riccati first order sensitivity equations (4.6)

1 filter first order sensitivity equations (4.11)
\[ Z^2 \text{ Riccati second order sensitivity equations (4.30)} \]

\[ Z^2 \text{ filter second order sensitivity equations (4.29)} \]

or roughly the equivalent of \((Z^2 + Z + 1)\) Kalman filters, which is a heavy computational cost.\(^1\)

A common alternate approach is to use the information matrix

\[ I_t(\alpha) = E_\alpha \left\{ \frac{\partial^2 \zeta (z_t; \alpha)}{\partial \alpha^2} \right\} \]

or, as is usually done, an approximation thereof, and this will be further discussed in Section 4.3. On the other hand, and noting the similarity in structure of (4.28)-(4.30) to the forward filter form (4.14)-(4.16) for evaluating the likelihood function gradient, one could also resort to a backward filter evaluation of the Hessian as well. This would consist of running forward in time the Kalman filter and the first order sensitivity filter (required for evaluating \( \beta_{ij} (z_t; \alpha) \) and the driving terms \( \bar{\omega}_{ij} (t; \alpha) \) and \( \bar{\Omega}_{ij} (t; \alpha) \)); then backward in time the adjoint to the second order sensitivity filter. Thus, by direct application of Corollary 3 of Appendix D on the form (4.28)-(4.30) we obtain:

\(^1\)This cost is not as formidable as it might appear at first: the symmetrical nature of the second order sensitivity equations can be indeed used here to great advantage.
\[
\frac{\partial^2 z(t;\alpha)}{\partial \alpha_i \partial \alpha_j} = \sum_{\tau=0}^{t} \beta_{ij}(z^T;\alpha) \\
+ \lambda'(0;\alpha) \frac{\partial^2 x(0;\alpha)}{\partial \alpha_i \partial \alpha_j} + \sum_{\tau=1}^{t} \lambda'(t;\alpha) \bar{\omega}_{ij}(t+1;\alpha) \\
+ \text{tr} \left[ \Lambda(0;\alpha) \frac{\partial^2 \Sigma(0)}{\partial \alpha_i \partial \alpha_j} \right] + \sum_{\tau=1}^{t} \text{tr} \left[ \Lambda(t;\alpha) \left( \Omega_{ij}(t;\alpha) + \bar{\Omega}_{ij}'(t;\alpha) \right) \right]
\]

where

\[
\lambda(t;\alpha) = \lambda'(0;\alpha) + \gamma(z^T;\alpha)
\]

with

\[
\lambda(t;\alpha) = \gamma(z^T;\alpha)
\]

and

\[
\Lambda(t;\alpha) = \Lambda'(0;\alpha) \Lambda(t+1;\alpha) \Lambda(0;\alpha) - \Lambda'(0;\alpha) \Lambda(t+1;\alpha) \gamma'(z^T;\alpha) + \Gamma(z^T;\alpha)
\]

with

\[
\Lambda(t;\alpha) = \Gamma(z^T;\alpha)
\]

Note that (4.35) and (4.36) are exactly the same equations as (4.25) and (4.26) respectively. Note also that this approach would now require roughly the equivalent of (\ell+2) Kalman filters in computation which is a significant improvement over the (\ell^2+\ell+1) cost mentioned above. In consequence, if one were willing to compute the gradient via the forward filter approach at a cost of (\ell+1) Kalman filters computations, one could also obtain the Hessian for an
extra cost of 1 (backward) Kalman filter computation and of course an increased storage burden. This also assumes that the cost of computing $\beta_{ij}(z^T;\alpha)$, $\bar{w}_{ij}(\tau;\alpha)$ and $\Omega_{ij}(\tau;\alpha)$ does not dominate. Here again, the relative advantages depend on the particular model set being used.

In the time invariant case, where the observations are processed by a steady state Kalman filter the same analysis as that of Section 4.1.4 applies shifting the potential advantage in computing the Hessian to the corresponding mixed filter approach.

4.3 Information Matrix Evaluation

For every parameter $\alpha$, let $I_t(\alpha)$ be the information matrix defined by the equivalent expressions

$$I_t(\alpha) = E_{\alpha}\left\{ \frac{\partial^2 \zeta(z^T;\alpha)}{\partial \alpha^2} \right\} = E_{\alpha}\left\{ \frac{\partial \zeta(z^T;\alpha)}{\partial \alpha} \right\} \left[ \frac{\partial \zeta(z^T;\alpha)}{\partial \alpha} \right]'$$

(4.37)

where the expectation is taken with respect to the probability measure corresponding to $\alpha$.

Therefore, from (4.1), we can write the information matrix as

$$I_t(\alpha) = E_{\alpha}\left\{ \sum_{\tau=0}^{t} \frac{\partial^2 \zeta(z(\tau)|z^{T-1};\alpha)}{\partial \alpha^2} \right\}$$

(4.38)

It is also useful to define the conditional information matrix $I_{z(t)|z^{T-1}}(\alpha)$ by the equivalent expressions
\[
\hat{I}_{z(t) \mid z^{t-1} (\alpha)} = E_\alpha \left\{ \frac{\partial^2 \zeta(z(t) \mid z^{t-1} ; \alpha)}{\partial \alpha^2} \right\} \mid z^{t-1} \\
= E_\alpha \left\{ \left[ \frac{\partial \zeta(z(t) \mid z^{t-1} ; \alpha)}{\partial \alpha} \right] \left[ \frac{\partial \zeta(z(t) \mid z^{t-1} ; \alpha)}{\partial \alpha} \right]' \right\} \mid z^{t-1} 
\]
so that

\[
I_t (\alpha) = E_\alpha \left\{ \sum_{\tau=0}^{t} \hat{I}_{z(\tau) \mid z^{\tau-1} (\alpha)} \right\} \\
= \sum_{\tau=0}^{t} E_\alpha \left\{ \hat{I}_{z(\tau) \mid z^{\tau-1} (\alpha)} \right\} 
\]

(4.39)

And in order to derive an expression for \( I_t (\alpha) \) one can first obtain an expression for \( \hat{I}_{z(t) \mid z^{t-1} (\alpha)} \) and then compute its expected value.

The calculation of \( \hat{I}_{z(\tau) \mid z^{\tau-1} (\alpha)} \) is carried out in Appendix E where the equation

\[
\begin{bmatrix}
\hat{I}_{z(\tau) \mid z^{\tau-1} (\alpha)} \\
\end{bmatrix}
= \text{tr} \left\{ \begin{bmatrix}
\frac{\partial r(\tau; \alpha)}{\partial \alpha_i} & \frac{\partial r'(\tau; \alpha)}{\partial \alpha_j} \\
\end{bmatrix} s^{-1}_\alpha (\tau) \right\} \\
+ \frac{1}{2} \text{tr} \left\{ \begin{bmatrix}
\frac{\partial^2 s(\tau)}{\partial \alpha_i} & \frac{\partial^2 s(\tau)}{\partial \alpha_j} \\
\end{bmatrix} \begin{bmatrix}
s^{-1}_\alpha (\tau) & s^{-1}_\alpha (\tau) \\
\end{bmatrix} \right\} 
\]

(4.41)

for its \( ij^{th} \) element is derived. Now let \( s_{\alpha, i, j} (\tau) \) denote the covariance matrix between \( \frac{\partial r(\tau; \alpha)}{\partial \alpha_i} \) and \( \frac{\partial r(\tau; \alpha)}{\partial \alpha_j} \) with respect to the probability measure corresponding to \( \alpha \), then
\[
E_{\alpha} \left\{ \frac{\partial \mathbf{r}(\tau; \alpha)}{\partial \alpha_i} \frac{\partial \mathbf{r}'(\tau; \alpha)}{\partial \alpha_j} \right\} = S_{\alpha, i, j}(\tau) + E_{\alpha} \left\{ \frac{\partial \mathbf{r}(\tau; \alpha)}{\partial \alpha_i} \right\} E_{\alpha} \left\{ \frac{\partial \mathbf{r}'(\tau; \alpha)}{\partial \alpha_j} \right\}
\] (4.42)

and
\[
E_{\alpha} \left\{ I_{z(\tau)} \right\} = \text{tr} \left[ E_{\alpha} \left\{ \frac{\partial \mathbf{r}(\tau; \alpha)}{\partial \alpha_i} \right\} E_{\alpha} \left\{ \frac{\partial \mathbf{r}'(\tau; \alpha)}{\partial \alpha_j} \right\} S_{\alpha}^{-1}(\tau) \right]
+ \text{tr} \left[ S_{\alpha, i, j}(\tau) S_{\alpha}^{-1}(\tau) + \frac{1}{2} \frac{\partial S_{\alpha}(\tau)}{\partial \alpha_i} S_{\alpha}^{-1}(\tau) \frac{\partial S_{\alpha}(\tau)}{\partial \alpha_j} S_{\alpha}^{-1}(\tau) \right]
\] (4.43)

and
\[
[I_{z(\tau)}]_{ij} = \sum_{\tau=0}^{t} \text{tr} \left[ E_{\alpha} \left\{ \frac{\partial \mathbf{r}(\tau; \alpha)}{\partial \alpha_i} \right\} E_{\alpha} \left\{ \frac{\partial \mathbf{r}'(\tau; \alpha)}{\partial \alpha_j} \right\} S_{\alpha}^{-1}(\tau) \right]
+ \sum_{\tau=0}^{t} \text{tr} \left[ S_{\alpha, i, j}(\tau) S_{\alpha}^{-1}(\tau) + \frac{1}{2} \frac{\partial S_{\alpha}(\tau)}{\partial \alpha_i} S_{\alpha}^{-1}(\tau) \frac{\partial S_{\alpha}(\tau)}{\partial \alpha_j} S_{\alpha}^{-1}(\tau) \right]
\] (4.44)

Evaluating \( S_{\alpha, i, j}(\tau) \) and \( E_{\alpha} \left\{ \frac{\partial \mathbf{r}(\tau; \alpha)}{\partial \alpha_i} \right\} \) is conceptually straightforward although computationally expensive. They can be obtained by propagating the mean and covariance equations for the \( 4n \) dimensional linear composite system shown in Figure 4.4. Then the sum (4.44) is accumulated to determine one element of the information matrix and the whole process must be repeated \( \frac{l(l+1)}{2} \) times since there are that many distinct elements of the symmetric information matrix.

Now one of the commonly used approximations to evaluate \( I_{z(\tau)} \) is to sum up the stochastic quantity \( \tilde{I}_{z(t)} \) rather than

\[
I_{z(t)}|z^{-1}(\alpha)
\]
Figure 4.4: Linear System for Information Matrix Element Computation.
its expected value (see 4.40), therefore obtaining the approximation

\[
\begin{bmatrix}
I_{z^t}(\alpha) \\
\end{bmatrix}
= \sum_{\tau=0}^{t} \text{tr} \left[ \frac{\partial \tilde{r}(\tau; \alpha)}{\partial \alpha_i} \frac{\partial \tilde{r}(\tau; \alpha)}{\partial \alpha_j} S^{-1}(\tau) \right]
+ \frac{1}{2} \frac{\partial S_{\alpha}(\tau)}{\partial \alpha_i} S^{-1}(\tau) \frac{\partial S_{\alpha}(\tau)}{\partial \alpha_j} S^{-1}(\tau)
\]

(4.45)

Since this is an instance of using a particular realization of a stochastic quantity as an approximation to its expected value, (4.45) would make a good approximation only if the standard deviation of \( \tilde{z}(t) | z^{t-1}(\alpha) \) is much smaller than its expected value.

Note that the advantage in using the approximation (4.45) is that it only involves quantities which are already evaluated when using the forward filter approach for computing the gradient.

When one is using the backward filter approach to computing the gradient, \( \frac{\partial \tilde{r}(\tau; \alpha)}{\partial \alpha_i} \) and \( \frac{\partial S_{\alpha}(\tau)}{\partial \alpha_i} \) are not available and the above approximation to computing \( I_{z^t}(\alpha) \) cannot be carried out. However, note that if the model set includes the true parameter, the Kalman filter residuals \( r(t; \alpha) \) come close to being independent near the correct parameter. Therefore, for parameters \( \alpha \) close to \( * \), the random variables \( \frac{\partial \tilde{z}(z(t)) | z^{t-1}; \alpha}{\partial \alpha_i} \) also become independent and the random variable
I\left[\frac{\partial \zeta(z(T)|z^{T-1};\alpha)}{\partial \alpha} \right] \left[\frac{\partial \zeta(z(T)|z^{T-1};\alpha)}{\partial \alpha} \right]

has a standard deviation that grows only as $\sqrt{t}$ while its expected value $I_t^{(a)}$ grows linearly with $t$. So for large values of $t$, and $z$ close to the true parameter, one could use the approximation

$$I_t^{(a)} = \sum_{\tau=0}^{t} \left[\frac{\partial \zeta(z(T)|z^{T-1};\alpha)}{\partial \alpha} \right] \left[\frac{\partial \zeta(z(T)|z^{T-1};\alpha)}{\partial \alpha} \right]$$

(4.46)

Furthermore, since

$$\frac{\partial \zeta(z(t)|z^{t-1};\alpha)}{\partial \alpha} = \frac{\partial \zeta(z^t;\alpha)}{\partial \alpha} - \frac{\partial \zeta(z^{t-1};\alpha)}{\partial \alpha}$$

this approximation is useful when the backward filter approach is being used to evaluate the gradient $\frac{\partial \zeta(z^t;\alpha)}{\partial \alpha_1}$.
CHAPTER 5

NUMERICAL EXAMPLES AND RESULTS

In this chapter, we present a few examples with simple linear time-invariant models and computer simulated data to illustrate some of the points made in chapters 2 and 4. The computer programs we used are first briefly described. Those include Fortran subroutines which

- evaluate ML parameter estimates from sets of simulated observations using methods described in Chapter 4.

- evaluate asymptotic minimum information measure parameters by calculating and minimizing the information function $J^*_\alpha(t)$ of Chapter 2.

Estimation runs with an increasing number of unknown parameters and with models of increasing order are then presented, using both forward and mixed filter evaluations of the likelihood function gradient. The relative computing times are compared for each case. Then, the information function $J^*_\alpha(t)$ is evaluated and minimized for simple truth models and model sets of same and lower order. Finally, convergence of the time scaled likelihood function to the corresponding information function is also illustrated.

5.1 Description of Programs

5.1.1 ML parameter estimates:

As mentioned above, the ML estimate $\hat{\alpha}_t$ is evaluated by minimizing
the likelihood function defined in (2.21)-(2.22).

(i) The minimization program (MINOPZ)

The minimization program used is directly adapted from Dennis and Mei's [69] subroutine MINOP, which implements a quasi-Newton algorithm with two main features: the switch from gradient to quasi-Newton steps uses a modification of Powell's [70] dogleg strategy; and the building up of the approximate Hessian uses Davidon's [71] updating scheme, thus maintaining only symmetric and positive definite updates.

MINOPZ calls two subroutines:

- CALCF which returns the value of the function $\xi(z^t;\alpha)$ corresponding to a fixed set of input output values $\{u(\tau), z(\tau), \tau=0,1,...,t\}$ (provided by the main calling program) and a current iteration value of the parameter ALPHA (passed by MINOPZ).

- Externally defined CALCG which returns the corresponding value of the gradient $\frac{\partial \xi(z^t;\alpha)}{\partial \alpha}$.

The algorithm stops either when the norm $G$ of the gradient becomes smaller than a user provided accuracy value, or when the number of calls to CALCF reaches a user provided maximum MAXFUN.

(ii) The function evaluation program (CALCF)

This program, as well as all the subsequently described programs, is coded for linear time invariant state space models with the corresponding steady state Kalman filters processing the set of input
output values. It is described by the flow chart in Figure 5.1 where:

- MODEF is an example dependent subroutine returning model matrices $A_{\alpha}, B_{\alpha}, C_{\alpha}, L_{\alpha}, \Sigma_{\alpha}$ and $\theta_{\alpha}$ corresponding to a value of the parameter vector $\alpha$.

- DKF returns the corresponding Kalman filter matrices $\bar{A}_{\alpha}, H_{\alpha}$ and $S_{\alpha}$.

- FRUN runs forward in time the Kalman filter and accumulates the corresponding sum:

$$X_{\alpha}^t = \sum_{\tau=0}^{t} r(\tau;\alpha)r'(\tau;\alpha)$$  \hspace{1cm} (5.1)

- CALF simply computes the corresponding value of the likelihood function (4.1)-(4.2)

$$\xi(z^t;\alpha) = \frac{t+1}{2} \log \det S_{\alpha} + \frac{1}{2} \text{tr}[S_{\alpha}^{-1} X_{\alpha}^t]$$  \hspace{1cm} (5.2)

The LIDS library ([72]) of Fortran subroutines is used to solve the algebraic matrix Riccati equation for DKF, as well as execute many of the other matrix manipulations.

(iii) The forward filter gradient evaluation program (CALCGF)

This program is based on the forward filter approach of section 4.4.1 and is described by the flow chart in Figure 5.2 where:
Figure 5.1: Flow chart describing subroutine CALCF.
Figure 5.2: Flow chart describing subroutine CALCGF.

Figure 5.3: Flow chart describing subroutine CALCGM.
- SENDEF is an example dependent subroutine returning model sensitivity matrices $\frac{\partial A}{\partial \alpha_i}, \frac{\partial B}{\partial \alpha_i}, \frac{\partial C}{\partial \alpha_i}, \frac{\partial D}{\partial \alpha_i}$ and $\frac{\partial E}{\partial \alpha_i} \; ; i=1, \ldots, l$

corresponding to a value of the parameter vector ALPHA.

- KFSEN1 returns the corresponding Kalman filter sensitivity matrices $\frac{\partial A}{\partial \alpha_i}, \frac{\partial (A_H)}{\partial \alpha_i}$ and $\frac{\partial S}{\partial \alpha_i} \; ; i=1, \ldots, l$.

- FRUN1 runs, forward in time, the Kalman filter as well as the filter sensitivity equations and accumulates the corresponding sums:

$$x_1 = \frac{\Delta}{\alpha_i} \sum_{\tau=0}^{t} r(\tau; \alpha) r'(\tau; \alpha)$$  \hspace{1cm} (5.3)

and

$$x_2 = \frac{\Delta}{\alpha_i} \sum_{\tau=0}^{t} \frac{\partial r(\tau; \alpha)}{\partial \alpha_i} r'(\tau; \alpha) \; ; i=1, \ldots, l$$  \hspace{1cm} (5.4)

- CALG simply computes the corresponding value of the gradient (4.3)-(4.4)

$$\frac{\partial \zeta(z^t; \alpha)}{\partial \alpha_i} = \text{tr} \left\{ S^{-1}_{\alpha} \left( \frac{1}{2} \left[ (t+1) I - x_1 S^{-1}_{\alpha} \right] \frac{\partial S}{\partial \alpha_i} + x_2 \right) \right\}$$  \hspace{1cm} (5.5)

where, similarly to (5.2), the time invariance of $S_{\alpha}$ and $\frac{\partial S}{\partial \alpha_i}$ was used to factor those quantities out of their corresponding summation terms in (4.3)-(4.4).
Here again, subroutines from [72] are used and the algebraic Lyapunov equations (4.27) are solved for KFSEN1 by the efficient Bartels-Stewart method mentioned in Section 4.4.1.

(iv) The mixed filter gradient evaluation program (CALCGM)

This program is based on the mixed filter approach of Section 4.4.2 and is described by the flow chart in Figure 5.3 where:

- FRUNZ runs, forward in time, the Kalman filter equations and stores the corresponding values of \( \hat{x}(\tau;\alpha) \) and \( r(\tau;\alpha) \); \( \tau=0,1,...,t \).
- BRUN runs, backward in time, the adjoint to the filter sensitivity equations and accumulates the corresponding sums:

\[
\begin{align*}
X_1 & \overset{\Delta}{=} \sum_{\tau=0}^{t} r(\tau;\alpha) r'(\tau;\alpha) & (5.6) \\
X_4 & \overset{\Delta}{=} \sum_{\tau=0}^{t} \hat{x}(\tau;\alpha) r'(\tau;\alpha) & (5.7) \\
X_5 & \overset{\Delta}{=} \sum_{\tau=1}^{t} \hat{x}(\tau-1;\alpha) \lambda'(\tau;\alpha) & (5.8) \\
X_6 & \overset{\Delta}{=} \sum_{\tau=1}^{t} u(\tau-1;\alpha) \lambda'(\tau;\alpha) & (5.9) \\
X_7 & \overset{\Delta}{=} \sum_{\tau=1}^{t} z(\tau-1;\alpha) \lambda'(\tau;\alpha) & (5.10)
\end{align*}
\]
CALGM computes the corresponding value of the gradient from (4.22)

$$\frac{\partial \xi}{\partial \alpha_i} = \text{tr} \left\{ S_{\alpha}^{-1} \left( \frac{1}{2} \left[ (t+1)I - X_{1\alpha}^{-1} \right] \frac{\partial S_{\alpha}}{\partial \alpha_i} - \frac{\partial C_{\alpha}}{\partial \alpha_i} X_{4\alpha} \right) \right\}$$

$$+ \text{tr} \left\{ \frac{\partial A_{\alpha}}{\partial \alpha_i} X_5 \right\} + \text{tr} \left\{ \frac{\partial B_{\alpha}}{\partial \alpha_i} X_6 \right\} + \text{tr} \left\{ \frac{\partial (A_{\alpha}H_{\alpha})}{\partial \alpha_i} X_7 \right\}$$

(5.11)

where, once again, time invariant quantities were factored out of their corresponding summation terms in (4.22). Note also that, in view of (4.17)-(4.19) and (4.8), the first trace term in (5.11) corresponds to the first and last term in (4.22) whereas the last three trace terms in (5.11) corresponds to the second term in (4.22) (with \( \frac{\partial \xi(0;\alpha)}{\partial \alpha_i} \) set to zero).

Although some attempt was made to code the above programs efficiently (e.g. factoring out time invariant matrices in (5.5) and (5.11); use of the Bartels-Stewart algorithm in KFSEN1) we make no claim to great efficiency. Our purpose was to illustrate the savings obtained when using the mixed filter approach as compared to the forward filter approach in evaluating the likelihood function gradient. Therefore, great care was taken only in keeping the same coding techniques for both CALCGF and CALCGM. For example, \( X_{2\alpha_i} \) defined in (5.4), used in
(5.5), is computed by PRUN1 as:

\[
- \frac{\partial}{\partial \alpha_i} \left( \sum_{\tau=0}^{t} \hat{x}(\tau;\alpha) r'(\tau;\alpha) \right) - C_{\alpha} \left( \sum_{\tau=0}^{t} \frac{\partial^2 x(\tau;\alpha)}{\partial \alpha_i} r'(\tau;\alpha) \right)
\]

where we recognize in the first term \(X_{\alpha}^4\), used to compute the corresponding term in (5.11). Computing \(X_{\alpha}^2\) as suggested by its definition (5.4) would have caused an additional burden to the forward filter approach, biasing computation time comparisons to its disadvantage. In other words, the computation times, reported in the examples of Section 5.2 are not to be taken as absolute measures, it is in their relative values that we are rather interested.

### 5.1.2 Minimum information measure parameters

As mentioned in Chapter 2, the minimum information measure parameter \(\alpha_{0,t}\) is evaluated by minimizing the function \(J^*_{\alpha}(t)\) defined in (2.88)

(i) **The minimization program (ZXMIN)**

This program is provided by the IMSL library ([73]) of Fortran subroutines. It implements a quasi-Newton algorithm but does not require the user to supply a gradient evaluation program for the objective function.

ZXMIN calls an externally defined subroutine which, in our case, returns the value of the function \(J^*_{\alpha}(t)\) corresponding to a fixed
set of input values \{u(\tau); \tau=0,1,\ldots,t\}, truth model matrices \(A_x, B_x, L_x, C_x, E_x, \Theta_x\) (provided by the main calling program) and a current iteration value of the parameter \(\text{ALPHA}\) (passed by ZXMIN).

The algorithm stops either when, on two successive iterations, the parameter estimates agree component by component to a user specified number of digits; or when the number of calls to the objective function evaluation program reaches a user provided maximum \(\text{MAXFN}\).

(ii) The function \(J^*_x(t)\) evaluation program (INFOD)

This program follows the steps suggested in Section 2.2 to evaluate \(J^*_x(t)\). In particular, MODEF and DKF are called to obtain the model matrices and Kalman filter matrices corresponding to a value of the parameter \(\text{ALPHA}\) passed by ZXMIN. Then the truth model (passed by the main calling program) is composed with the Kalman filter, and a corresponding augmented algebraic Lyapunov equation is solved to obtain \(\overline{S^*_x}\). The term

\[
\frac{1}{t+1} \sum_{\tau=0}^{t} \overline{r(\tau;\xi)} \overline{r'(\tau;\xi)}
\]

is of course obtained from the particular set of inputs \(\{u(\tau); \tau=0,1,\ldots,t\}\) also passed by the main program.

5.2 Comparison of Forward and Mixed Filter Approaches

In order to illustrate the relative savings in computing time obtained with the mixed filter approach, the programs described in
Section 5.1.1 were used with models of increasing order and with an increasing number of unknown parameters, on a CDC Cyber 70 machine.

For each one of the following cases data was generated by a simulation of the corresponding model with specified parameter values. The components of the deterministic input $u(t)$ were simply chosen to be sequences alternating between plus and minus an arbitrary value. The models used were also arbitrary, chosen, if for any reason, so that the estimation runs were done under good conditions (e.g. relatively large plant noise and relatively small measurement noise). Finally, the number of data points generated and used in all the examples of this section was $t=1000$.

The central processor time elapsed from the start of each job was clocked before and after one gradient evaluation using each approach; and again, before and after a typical estimation run using each approach for gradient evaluations. The time values reported below are simply the corresponding differences. Not only are those values machine dependent and of relative nature as discussed above, but they are also slightly influenced by the total user load at the particular time of day the examples were run. Their relationship to one another was consistent, however, judged on the basis of identical jobs submitted several times as well as similar runs on an IBM/370.

For reasons independent of our present purposes, a double precision version of the above programs was first used with the following second order LTI model:
Results obtained for one unknown parameter (Z=1), two unknown parameters (Z=2) and three unknown parameters (Z=3) are respectively shown in Figures 5.4, 5.5, and 5.6. Each one of those figures shows:

- The corresponding model matrix entries which are considered unknown (first line)
- The likelihood function gradient vector corresponding to an arbitrary parameter value and evaluated twice, once with the forward filter and once with the mixed filter, along with the respective times.
- Estimation runs following identical paths but the first using forward filter gradient evaluations and the second using mixed filter gradient evaluations. On each of the lines corresponding to those runs we show the iteration number (IT), the corresponding invocation count of CALCF (NF), the corresponding decreased value of the likelihood function (F), the corresponding gradient norm (G) and the corresponding values of the unknown parameter components. Each run is followed by its corresponding time cost.

Note that the time required for one forward filter gradient evaluation increases roughly as \( \lambda \) (1.89 sec., 2.77 sec and 3.6 sec
Sample size: \( T = 1000 \)

Unknown parameters: \( A(2,2) \)

One forward filter grunt: Evaluation: \( 1.84 \times 10^{-4} \) seconds

\[-1.139164925177955 \times 10^{-7} \]

One mixed filter grunt: Evaluation: \( 2.47 \times 10^{-7} \) seconds

\[-1.1391649135177955 \times 10^{-17} \]

**Estimation run using forward filter**

<table>
<thead>
<tr>
<th>IT</th>
<th>NF</th>
<th>( F )</th>
<th>( G )</th>
<th>Alpha(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 1.40 \times 10^{-4} )</td>
<td>( 1.1990 \times 10^{-7} )</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 2.50 \times 10^{-4} )</td>
<td>( 7.81 \times 10^{-7} )</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 5.80 \times 10^{-4} )</td>
<td>( 2.50 \times 10^{-7} )</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 9.40 \times 10^{-4} )</td>
<td>( 3.79 \times 10^{-7} )</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 7.70 \times 10^{-4} )</td>
<td>( 3.79 \times 10^{-7} )</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 1.00 \times 10^{-4} )</td>
<td>( 3.34 \times 10^{-7} )</td>
</tr>
<tr>
<td>0</td>
<td>15</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 9.00 \times 10^{-5} )</td>
<td>( 3.34 \times 10^{-7} )</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 1.30 \times 10^{-4} )</td>
<td>( 3.34 \times 10^{-7} )</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 1.70 \times 10^{-4} )</td>
<td>( 3.34 \times 10^{-7} )</td>
</tr>
</tbody>
</table>

Time: \( 30.126 \) seconds

**Estimation run using mixed filter**

<table>
<thead>
<tr>
<th>IT</th>
<th>NF</th>
<th>( F )</th>
<th>( G )</th>
<th>Alpha(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 1.40 \times 10^{-4} )</td>
<td>( 1.1990 \times 10^{-7} )</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 2.50 \times 10^{-4} )</td>
<td>( 7.81 \times 10^{-7} )</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 5.80 \times 10^{-4} )</td>
<td>( 2.50 \times 10^{-7} )</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 9.40 \times 10^{-4} )</td>
<td>( 3.79 \times 10^{-7} )</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 7.70 \times 10^{-4} )</td>
<td>( 3.79 \times 10^{-7} )</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 1.00 \times 10^{-4} )</td>
<td>( 3.34 \times 10^{-7} )</td>
</tr>
<tr>
<td>0</td>
<td>15</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 9.00 \times 10^{-5} )</td>
<td>( 3.34 \times 10^{-7} )</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 1.30 \times 10^{-4} )</td>
<td>( 3.34 \times 10^{-7} )</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>( 8.4.7259376 \times 10^{-7} )</td>
<td>( 1.70 \times 10^{-4} )</td>
<td>( 3.34 \times 10^{-7} )</td>
</tr>
</tbody>
</table>

Time: \( 33.527 \) seconds

Figure 5.4: One unknown parameter in second order model.
SAMPLE SIZE T=100G

ONE FORWARD FILTER GRADIENT EVALUATION 1.7660 SECONDS

- .357414 x 10^3
- .319790 x 10^2

ONE MIXED FILTER GRADIENT EVALUATION 1.9650 SECONDS

- .357414 x 10^3
- .319790 x 10^2

ESTIMATION RUN USING FORWARD FILTER:

<table>
<thead>
<tr>
<th>IT</th>
<th>NF</th>
<th>F</th>
<th>G</th>
<th>ALPHA(1)</th>
<th>ALPHA(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>+1.176412 x 10^3</td>
<td>0.3560</td>
<td>+1.4060 x 10^4</td>
<td>+1.4060 x 10^4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>+1.176412 x 10^3</td>
<td>0.2421</td>
<td>+2.1987 x 10^4</td>
<td>+2.1987 x 10^4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>+1.176412 x 10^3</td>
<td>0.1990</td>
<td>+2.1987 x 10^4</td>
<td>+2.1987 x 10^4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>+1.176412 x 10^3</td>
<td>0.1500</td>
<td>+2.1987 x 10^4</td>
<td>+2.1987 x 10^4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>+1.176412 x 10^3</td>
<td>0.1100</td>
<td>+2.1987 x 10^4</td>
<td>+2.1987 x 10^4</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>+1.176412 x 10^3</td>
<td>0.0770</td>
<td>+2.1987 x 10^4</td>
<td>+2.1987 x 10^4</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>+1.176412 x 10^3</td>
<td>0.0490</td>
<td>+2.1987 x 10^4</td>
<td>+2.1987 x 10^4</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>+1.176412 x 10^3</td>
<td>0.0260</td>
<td>+2.1987 x 10^4</td>
<td>+2.1987 x 10^4</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>+1.176412 x 10^3</td>
<td>0.0130</td>
<td>+2.1987 x 10^4</td>
<td>+2.1987 x 10^4</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>+1.176412 x 10^3</td>
<td>0.0070</td>
<td>+2.1987 x 10^4</td>
<td>+2.1987 x 10^4</td>
</tr>
</tbody>
</table>

TIME: 50.3340 SECONDS

ESTIMATION RUN USING MIXED FILTER:

<table>
<thead>
<tr>
<th>IT</th>
<th>NF</th>
<th>F</th>
<th>G</th>
<th>ALPHA(1)</th>
<th>ALPHA(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>+1.356412 x 10^3</td>
<td>0.7960</td>
<td>+1.7000 x 10^4</td>
<td>+1.7000 x 10^4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>+1.356412 x 10^3</td>
<td>0.6420</td>
<td>+2.3190 x 10^4</td>
<td>+2.3190 x 10^4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>+1.356412 x 10^3</td>
<td>0.5100</td>
<td>+2.3190 x 10^4</td>
<td>+2.3190 x 10^4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>+1.356412 x 10^3</td>
<td>0.3970</td>
<td>+2.3190 x 10^4</td>
<td>+2.3190 x 10^4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>+1.356412 x 10^3</td>
<td>0.3000</td>
<td>+2.3190 x 10^4</td>
<td>+2.3190 x 10^4</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>+1.356412 x 10^3</td>
<td>0.2170</td>
<td>+2.3190 x 10^4</td>
<td>+2.3190 x 10^4</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>+1.356412 x 10^3</td>
<td>0.1500</td>
<td>+2.3190 x 10^4</td>
<td>+2.3190 x 10^4</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>+1.356412 x 10^3</td>
<td>0.1050</td>
<td>+2.3190 x 10^4</td>
<td>+2.3190 x 10^4</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>+1.356412 x 10^3</td>
<td>0.0720</td>
<td>+2.3190 x 10^4</td>
<td>+2.3190 x 10^4</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>+1.356412 x 10^3</td>
<td>0.0490</td>
<td>+2.3190 x 10^4</td>
<td>+2.3190 x 10^4</td>
</tr>
</tbody>
</table>

TIME: 41.6730 SECONDS

Figure 5.5: Two unknown parameters in second order model.
SAMPLE SIZE T=1200  

ONE FORWARD FILTER GRADIENT EVALUATION: 3.60409 SECONDS

-118-

ONE MIXED FILTER GRADIENT EVALUATION: 2.30240 SECONDS

ESTIMATION RUN USING FORWARD FILTER:

<table>
<thead>
<tr>
<th>TIME</th>
<th>SECONDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>105.9420</td>
</tr>
<tr>
<td>2</td>
<td>105.9420</td>
</tr>
<tr>
<td>3</td>
<td>105.9420</td>
</tr>
<tr>
<td>4</td>
<td>105.9420</td>
</tr>
<tr>
<td>5</td>
<td>105.9420</td>
</tr>
<tr>
<td>6</td>
<td>105.9420</td>
</tr>
<tr>
<td>7</td>
<td>105.9420</td>
</tr>
<tr>
<td>8</td>
<td>105.9420</td>
</tr>
<tr>
<td>9</td>
<td>105.9420</td>
</tr>
<tr>
<td>10</td>
<td>105.9420</td>
</tr>
<tr>
<td>11</td>
<td>105.9420</td>
</tr>
<tr>
<td>12</td>
<td>105.9420</td>
</tr>
<tr>
<td>13</td>
<td>105.9420</td>
</tr>
<tr>
<td>14</td>
<td>105.9420</td>
</tr>
<tr>
<td>15</td>
<td>105.9420</td>
</tr>
<tr>
<td>16</td>
<td>105.9420</td>
</tr>
</tbody>
</table>

TIME 78.4499 SECONDS

ESTIMATION RUN USING MIXED FILTER:

<table>
<thead>
<tr>
<th>TIME</th>
<th>SECONDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>105.9420</td>
</tr>
<tr>
<td>2</td>
<td>105.9420</td>
</tr>
<tr>
<td>3</td>
<td>105.9420</td>
</tr>
<tr>
<td>4</td>
<td>105.9420</td>
</tr>
<tr>
<td>5</td>
<td>105.9420</td>
</tr>
<tr>
<td>6</td>
<td>105.9420</td>
</tr>
<tr>
<td>7</td>
<td>105.9420</td>
</tr>
<tr>
<td>8</td>
<td>105.9420</td>
</tr>
<tr>
<td>9</td>
<td>105.9420</td>
</tr>
<tr>
<td>10</td>
<td>105.9420</td>
</tr>
<tr>
<td>11</td>
<td>105.9420</td>
</tr>
<tr>
<td>12</td>
<td>105.9420</td>
</tr>
<tr>
<td>13</td>
<td>105.9420</td>
</tr>
<tr>
<td>14</td>
<td>105.9420</td>
</tr>
<tr>
<td>15</td>
<td>105.9420</td>
</tr>
<tr>
<td>16</td>
<td>105.9420</td>
</tr>
</tbody>
</table>

TIME 51.0820 SECONDS

Figure 5.6: Three unknown parameters in second order model.
respectively) whereas the time required for one mixed filter gradient evaluation increases at a much slower rate (1.96 sec, 1.965 sec and 2.01 sec respectively). Note also that the forward filter is actually faster than the mixed filter when \( l=1 \) which is consistent with our expectations since the adjoint equation (4.23) is replacing only one filter sensitivity equation and still requires some additional computational overhead.

Interesting savings start showing up however with the estimation runs corresponding to \( l=2 \) and \( l=3 \). Those were done with \textsc{maxfun} (see Section 5.1) set to 20 and provide acceptable estimates with savings of 8.86 sec. (12 \textsc{calcg} calls) and 26.79 sec. (17 \textsc{calcg} calls) respectively. The dramatic savings that one could achieve with the use of the mixed filter approach are however better illustrated by the results corresponding to higher order models.

Runs similar to those corresponding to figures 5.4-5.6 were made with a single precision version of the above programs, and LTI models of order 5, 10 and 15 (see Appendix F) were used. The corresponding results are tabulated in Figure 5.7. Note that, for the higher order models, gradient values computed at the same point but using each one of the two filters differed enough to make the minimization algorithm follow slightly different paths. This is reflected here by the difference between the number of calls to \textsc{calcg} and the number of calls to \textsc{calcgm} in the corresponding estimation runs. In each run however,
<table>
<thead>
<tr>
<th></th>
<th>n=5</th>
<th>n=10</th>
<th>n=10</th>
<th>n=15</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l = 5 )</td>
<td>7.0</td>
<td>15.4</td>
<td>28.1</td>
<td>30.0</td>
</tr>
<tr>
<td>( l = 5 )</td>
<td>3.1</td>
<td>7.4</td>
<td>8.0</td>
<td>15.1</td>
</tr>
</tbody>
</table>

One Estimation Run
Using Forward Filter
Time: (sec):
\((\# \text{ calls to } \text{CALCGF})\)
\((16) \quad (23) \quad (26) \quad (15)\)

One Estimation Run
Using Mixed Filter
Time: (sec):
\((\# \text{ calls to } \text{CALCGM})\)
\((16) \quad (20) \quad (25) \quad (20)\)

**Figure 5.7:** Higher order models.
the final estimates agreed to two decimal places with those obtained by the corresponding run. At any rate, the results in Figure 5.7 illustrate very well the computational savings obtained with the mixed filter approach.

5.3 Minimum Information Measure Parameters

This section provides a simple illustration to the remarks following example 2.4 in Chapter 2, concerning the effect of deterministic inputs on the minimum information measure parameter when the true system is not included in the model set.

The programs described in section 5.1.2 were first used with the following truth model

\[
A_* = \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & -0.2 \end{bmatrix} \quad B_* = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} \quad L_* = \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix}
\]

\[C_* = \begin{bmatrix} 1.0 & 1.0 \end{bmatrix} \quad E_* = 0.5 \quad \Theta_* = 0.5 \times 10^{-5}
\]

and model set

\[
A_{\alpha} = \alpha \quad B_{\alpha} = \begin{bmatrix} 1.0 & 1.0 \end{bmatrix} \quad L_{\alpha} = 1.0
\]

\[C_{\alpha} = 1.0 \quad E_{\alpha} = 0.5 \quad \Theta_{\alpha} = 0.5 \times 10^{-5}
\]

Note that both modes of the truth model are excited by the input noise
ξ(t); but that only the first mode \((0.5)\) is excited by the first component \(u_1(t)\) of the deterministic input, and only the second mode \((-0.2)\) is excited by the second component \(u_2(t)\) of the deterministic input.

The first three lines of the table in Figure 5.8 show the minimum information measure parameters obtained for:

**Case 1:** \(u_1(t) = u_2(t) = 0; \quad t = 0, 1, \ldots, 1000\)

**Case 2:** \(u_1(t) = 1; \quad u_2(t) = 0; \quad t = 0, 1, \ldots, 1000\)

**Case 3:** \(u_1(t) = 0; \quad u_2(t) = 1; \quad t = 0, 1, \ldots, 1000\)

respectively. It comes as no surprise that \(\alpha_0\) is closer to the mode excited by \(u(t)\) (0.388 when mode 0.5 is excited and 0.081 when mode -0.2 is excited) than the value 0.174 when no deterministic inputs are present.

Those trends were dramatized when the input noise covariance matrix \(\Xi_q\) and its model \(\bar{\Xi}_q\) were reduced by a factor of \(10^3\), thus increasing the significance of the deterministic input excitation. The corresponding results for cases 1 through 3 above are respectively shown in lines 4 through 6 of Figure 5.8. When no deterministic inputs are present (case 1) \(\alpha_0\) is 0.177, close to the above corresponding value. But with mode 0.5 excited \(\alpha_0\) is now 0.4998; and with mode -0.2 excited \(\alpha_0\) is -0.1989.
<table>
<thead>
<tr>
<th>NF</th>
<th>DIST</th>
<th>G</th>
<th>ALPHA(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>-1.8627*1120+61</td>
<td>.5150-15</td>
<td>.7434*157+3</td>
</tr>
<tr>
<td>22</td>
<td>-1.8719*6663+61</td>
<td>.259-14</td>
<td>.3827*333+6</td>
</tr>
<tr>
<td>25</td>
<td>-1.7926*1104+91</td>
<td>.865-12</td>
<td>.8137*361+61</td>
</tr>
<tr>
<td>19</td>
<td>-1.7725*7975+91</td>
<td>.713-16</td>
<td>.1773*310+6</td>
</tr>
<tr>
<td>25</td>
<td>-1.5312*5454+91</td>
<td>.969-12</td>
<td>.4952*828+33</td>
</tr>
<tr>
<td>35</td>
<td>-1.3186*6390+91</td>
<td>.726-07</td>
<td>.1948*3157+3</td>
</tr>
</tbody>
</table>

NF: Number of function evaluations performed
DIST: Value of $J^*(t)$ at the final parameter estimate.
G: Norm of the corresponding gradient
ALPHA: Final parameter estimate as determined by ZXMIN.

**Figure 5.8:** Minimum Information Measure Parameters.
5.4 Convergence of the Scaled Likelihood Function

The above truth model (5.13) was used to simulate successively 10, 100 and 1000 data points with no deterministic input \( u(t) \). Then the corresponding scaled likelihood functions (2.43)

\[
\frac{1}{t+1} \zeta(z^T; \alpha) = \frac{1}{t+1} \sum_{T=0}^{t} \zeta(z(T) \mid z^{T-1}; \alpha)
\]  

(5.15) were plotted (along with the information function \( I^*_\alpha \) to which (5.15) theoretically converges with \( t \to \infty \)) as functions of \( \alpha \) for the lower order model set (5.14) (figure 5.9) and the full order model set

\[
A_\alpha = \begin{bmatrix} \alpha & 0 \\ 0 & -0.2 \end{bmatrix}, \quad B_\alpha = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}, \quad L_\alpha = \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix}
\]

\[
C_\alpha = \begin{bmatrix} 1.0 & 1.0 \end{bmatrix}, \quad \Xi_\alpha = 0.5, \quad \Theta_\alpha = 0.5 \times 10^{-5}
\]

(figure 5.10).

Note that, as expected, \( I^*_\alpha \) is minimized at \( \alpha = 0.17 \) in figure 5.9 (consistent with corresponding result in Section 5.3), and at \( \alpha_0 = * = 0.5 \) in figure 5.10. The convergence of the scaled likelihood function to the corresponding function \( I^*_\alpha \) is also clear.

5.5 Summary

In this chapter we briefly described software developed and successfully used to illustrate some of the points of previous chapters.
Figure 5.9: Convergence of the Scaled Likelihood Function (α)}

1: \( \frac{1}{11} \zeta(z^{10}; \alpha) \) versus \( \alpha \)

2: \( \frac{1}{101} \zeta(z^{100}; \alpha) \) versus \( \alpha \)

3: \( \frac{1}{1001} \zeta(z^{1000}; \alpha) \) versus \( \alpha \)
Figure 5.10: Convergence of the Scaled Likelihood Function (*eP).
Then, some results comparing computation times required by forward filter evaluations of the likelihood function gradient to those required by mixed filter evaluations are presented. In spite of the simplistic and artificial nature of the examples used, these results make a strong case in favor of the use of adjoint computations. Note however that the time savings were obtained at the familiar cost of increased storage requirement (e.g. an additional 24K for the second order example of section 5.2). Similar simple examples were also used to complement our discussion of the role and asymptotic nature of minimum information measure parameters.
CHAPTER 6
SUMMARY AND SUGGESTIONS

In this thesis, we have studied the maximum likelihood method as applied to the identification of linear state space models.

We first analyzed the asymptotic properties of ML parameter estimates for state space models. Previous information theoretic concepts were adapted to compare different models in a model set which does not necessarily contain what is assumed to be a better description of the real system: a "true" state space model. We have then attempted to get more insight into the nature of the asymptotic models selected by the ML method from such model sets; thus interpreting in the state space framework some general theoretical results in system identification. We have also briefly reviewed the special problem of identification under feedback.

In the second part of the thesis we have turned our attention to some practical aspects of computing the above estimates. In particular we have investigated the use of adjoint computations to evaluate the gradient and Hessian of the likelihood functions. We have thus derived a mixed filter and a backward filter as alternatives to the standard forward filter approach to evaluate those quantities. The evaluation of the information matrix was also briefly discussed.

Finally, we have shown with simple examples and computer simulated data that significant computational savings can be achieved with the
above adjoint computations. There is an additional storage burden however. We also took advantage of the developed software to further illustrate some of the points made in the first part of the thesis.

In conclusion, we note that although the major theoretical properties of ML identification of state space models are now well understood (except for the parametrization problem which has developed into a field in its own right), many an improvement on the implementation level has yet to be achieved. For all its attractive properties, the ML method is still an expensive proposition. In the spirit of this thesis we would therefore suggest further research in improving the efficiency of gradient, Hessian or information matrix evaluations. In particular, one could take advantage of sparsity, and special structure in certain cases. At the level of the minimization program used, we would suggest further development of algorithms more suited to the specific problem of identifying state space models, rather than resort to algorithms designed for general function minimization.
We present here, for completeness, two standard results.

**Lemma A.1:**

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

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

(A.1)

with equality holding if and only if $A = B$.

**Proof:**

(A.1) is equivalent to

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

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

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

Let $A^{1/2}$ denote the real symmetric positive square root of $A$ and $X = A^{-1/2}BA^{-1/2}$; then $X$ is a real symmetric positive definite matrix with eigenvalues $\lambda_i(X) > 0$, $i = 1, \ldots, r$, $\text{tr}[X] = \sum_{i=1}^{r} \lambda_i(X)$ and

$$
\det[X] = \prod_{i=1}^{r} \lambda_i(X)
$$

so that (A.1) is equivalent to

$$
\text{tr}[X] - \ln(\det[X]) \geq r
$$
or

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

Since for any scalar \( x > 0 \)

\[ x - \ln x \geq 1 \]

with equality holding if and only if \( x = 1 \) (A.1) follows, with equality holding if and only if

\[ \lambda_i(X) = 1 \quad \forall i = 1, \ldots, r. \]

Since \( X \) is real symmetric equality holds if and only if

\[ X = A^{-1/2} B A^{-1/2} = I \]

or

\[ A = B. \]

**Lemma A.2**

Let \( A \) and \( B \) be two real symmetric \( r \times r \) positive definite matrices satisfying \( A \succeq B \) and \( \det[A] = \det[B] \), then \( A = B \).

**Proof:**

Let \( X \) be as in proof of Lemma A.1. Then, \( A \succeq B \) implies

\[ I \succeq X. \]
and \( \det(A) = \det(B) \) implies
\[
\det(X) = 1
\]

Therefore
\[
X^{1/2}A^{-1/2}B^{1/2} = I
\]
or
\[
A = B
\]
APPENDIX B

Derivation of the Riccati Sensitivity Equation

Consider the Riccati Equation (2.19)

\[ \Sigma(t+1) = \alpha(t) \Sigma(t) \alpha'(t) + l(t) \Sigma(t) l'(t) \]

- \[ \alpha(t) H \alpha(t) S(t) H' \alpha(t) \alpha'(t) \]

Dropping the arguments from the matrix notations for clarity, differentiate with respect to \( \alpha_i \) and use (4.9) for \( \frac{\partial H}{\partial \alpha_i} \)

\[ \frac{\partial \Sigma(t+1)}{\partial \alpha_i} = A \frac{\partial \Sigma(t)}{\partial \alpha_i} A' + \frac{\partial}{\partial \alpha_i} (LEL') \]

- \[ \alpha H \frac{\partial \Sigma(t)}{\partial \alpha_i} A' - A \frac{\partial \Sigma(t)}{\partial \alpha_i} C'H'A' \]

- \[ \alpha H \frac{\partial C}{\partial \alpha_i} \Sigma(t) A' - A \Sigma(t) \frac{\partial C'}{\partial \alpha_i} H'A' \]

+ \[ \alpha H \frac{\partial S}{\partial \alpha_i} H'A' \]

+ \[ \frac{\partial A}{\partial \alpha_i} \Sigma(t) [I-C'H']A' + A[I-HC] \Sigma(t) \frac{\partial A}{\partial \alpha_i} \]

(B.1)

Substituting for \( \frac{\partial S}{\partial \alpha_i} \) from (4.10) and using the definition of \( \bar{A} \) in (4.7):

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

\[ - \bar{A} \frac{\partial \Sigma(t)}{\partial \alpha_i} \bar{A}' \]
the eighth and ninth terms into:

\[ \frac{\partial A}{\partial \alpha_1} \Sigma(t)A' + \overline{A \Sigma}(t) \frac{\partial A'}{\partial \alpha_1} \]

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

\[ -AH \frac{\partial C}{\partial \alpha_1} \Sigma(t)A' - \overline{\alpha \Sigma}(t) \frac{\partial C'}{\partial \alpha_1} H'A' \]

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

\[ \frac{\partial}{\partial \alpha_1} (L \Sigma') + AH \frac{\partial \Theta}{\partial \alpha_1} H'A' \]

Equations (4.11) and (4.12) now follow.
APPENDIX C

Derivation of the Forward Filter Form (4.14) - (4.20)

As in Appendix B, the arguments are dropped from matrix notations for clarity.

Consider equation (4.4)

\[
\frac{\partial \xi (z(t) | z^{t-1}; \alpha)}{\partial \alpha_i} = r' s^{-1} \frac{\partial r}{\partial \alpha_i} - \frac{1}{2} r' s^{-1} \frac{\partial s}{\partial \alpha_i} s^{-1} r + \text{tr}[s^{-1} \frac{\partial s}{\partial \alpha_i}]
\]

\[
= r' s^{-1} \frac{\partial r}{\partial \alpha_i} + \frac{1}{2} \text{tr}[s^{-1}(I - rr' s^{-1}) \frac{\partial s}{\partial \alpha_i}]
\]

and substitute for \( \frac{\partial r}{\partial \alpha_i} \) and \( \frac{\partial s}{\partial \alpha_i} \)

\[
\frac{\partial \xi (z(t) | z^{t-1}; \alpha)}{\partial \alpha_i} = r' s^{-1} \left[ - \frac{\partial c}{\partial \alpha_i} \hat{x} - c \frac{\partial \hat{x}}{\partial \alpha_i} \right]
\]

\[
+ \frac{1}{2} \text{tr}[s^{-1}(I - rr' s^{-1}) \left( c \frac{\partial c}{\partial \alpha_i} c' + \frac{\partial \theta}{\partial \alpha_i} \right.]
\]

\[
+ \frac{\partial c}{\partial \alpha_i} \Sigma c' + c \Sigma \frac{\partial c'}{\partial \alpha_i} \right)\]

\[
- \frac{1}{2} \text{tr}[s^{-1}(I - rr' s^{-1}) (2c \Sigma \frac{\partial c'}{\partial \alpha} + \frac{\partial \theta}{\partial \alpha})]
\]

\[
- r' s^{-1} \frac{\partial c}{\partial \alpha_i} \hat{x}
\]

\[
- r' s^{-1} c \frac{\partial \hat{x}}{\partial \alpha_i}
\]

\[
+ \frac{1}{2} \text{tr} \left[ c' s^{-1}(I - rr' s^{-1}) c \frac{\partial \Sigma}{\partial \alpha_i} \right]
\]
where for the first term the identity
\[ \text{tr}(AB) = \text{tr}(A'B') \]
was used, and for the last term the identity
\[ \text{tr}(AB) = \text{tr}(BA) \]
was used.

Equations (4.14), (4.17), (4.18) and (4.19) now follow.

Consider now equations (4.6) and (4.8). Using (4.9) and (4.10) we can reexpress the last term in (4.8) as

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

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

\[
- \frac{\partial \Sigma}{\partial \alpha_i} C'S^{-1} - AH \frac{\partial \Sigma}{\partial \alpha_i} S^{-1} - \frac{\partial C}{\partial \alpha_i} \Sigma C'S^{-1}
\]

\[
- \frac{\partial \Sigma}{\partial \alpha_i} S^{-1} \right] r
\]

\[
= \left[ \frac{\partial \Sigma}{\partial \alpha_i} C'S^{-1} - AH \frac{\partial \Sigma}{\partial \alpha_i} C'S^{-1} \right] r
\]

\[
+ \left\{ \frac{\partial A}{\partial \alpha_i} H + A \left[ \frac{\partial C}{\partial \alpha_i} - H \left( \frac{\partial \Theta}{\partial \alpha_i} - \frac{\partial C}{\partial \alpha_i} \Sigma C' - C \frac{\partial C}{\partial \alpha_i} \right) \right] S^{-1} \right\} r
\]

Equations (4.15) and (4.20) now follow.
APPENDIX D

ADJOINT COMPUTATIONS

Some properties of adjoint equations are discussed.

Lemma:

Suppose \( J = \langle c, x \rangle \) where \( Ax = b \), \( A \) invertible.

Then \( J = \langle \lambda, b \rangle \) where \( A^\ast \lambda = c \).

Proof:

\[
J = \langle c, A^{-1}b \rangle = \langle (A^{-1})^\ast c, b \rangle = \langle \lambda, b \rangle
\]

where

\[
(A^{-1})^\ast c = (A^\ast)^{-1}c = \lambda
\]

or \( A^\ast \lambda = c \).

\[\Box\]

Corollary 1

Suppose

\[
J = \sum_{t=0}^{T} c_t^\ast x_t
\]

where

\[
x_{t+1} = A_t x_t + u_t, \quad x_0 \text{ given.}
\]

Then

\[
J = \lambda_0^\ast x_0 + \sum_{t=1}^{T} \lambda_t^\ast u_{t-1}
\]

where

\[
\lambda_t = A_t^\ast \lambda_{t+1} + c_t, \quad \lambda_T = c_T
\]
Proof:

\[
\begin{bmatrix}
1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
-A_0 & 1 & 0 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\
-A_{T-2} & 1 & 0 & \cdots & 0 & 0 & 0 \\
0 & -A_{T-1} & 1 & \cdots & 0 & 0 & 0 \\
\end{bmatrix}
\begin{bmatrix}
x_0 \\
x_1 \\
\vdots \\
x_T \\
\end{bmatrix}
=
\begin{bmatrix}
x_0 \\
u_0 \\
\vdots \\
u_{T-1} \\
\end{bmatrix}
\]

From the lemma

\[
J = \begin{bmatrix}
\lambda'_0 & \lambda'_1 & \cdots & \lambda'_T \\
\end{bmatrix}
\begin{bmatrix}
x_0 \\
u_0 \\
\vdots \\
u_{T-1} \\
\end{bmatrix}
\]

where

\[
\begin{bmatrix}
1 & -A'_0 & 0 & \cdots & 0 & 0 \\
0 & 1 & -A'_1 & \cdots & 0 & 0 \\
0 & 0 & 1 & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \ddots & \ddots & \ddots \\
0 & 0 & 0 & \cdots & 1 & -A'_{T-1} \\
\end{bmatrix}
\begin{bmatrix}
\lambda'_0 \\
\lambda'_1 \\
\vdots \\
\lambda'_T \\
\end{bmatrix}
=
\begin{bmatrix}
c'_0 \\
c'_1 \\
\vdots \\
c'_T \\
\end{bmatrix}
\]

i.e.

\[
\lambda'_T = c'_T \\
\lambda'_t = A'_t \lambda'_{t+1} + c'_t
\]

\[\square\]
Corollary 2

Suppose \( J = \sum_{t=0}^{T} \text{tr}[C_t X_t], \) \( C_t \) symmetric

where \( X_{t+1} = A_t X_t A_t' + U_t, \) \( X_0 \) given and symmetric, \( U_t \) symmetric.

Then \( J = \text{tr}(A_0 X_0) + \sum_{t=1}^{T} \text{tr}[A_t U_{t-1}] \)

where \( A_{t-1} = A_t A_{t-1} + C_{t-1}, \) \( A_T = C_T \)

Proof:

Define formally the operator

\[
\mathcal{L}(X_0, X_1, \ldots, X_T) = (X_0, -A_0 X_0 A_0' + X_1, \ldots, -A_{T-1} X_{T-1} A_{T-1}' + X_T)
\]

and the inner product

\[
\langle (C_0, C_1, \ldots, C_T), (X_0, X_1, \ldots, X_T) \rangle = \sum_{t=0}^{T} \text{tr}[C_t X_t]
\]

Evaluate \( J = \langle (C_0, C_1, \ldots, C_T), (X_0, X_1, \ldots, X_T) \rangle \)

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

From the lemma,

\[
J = \langle (A_0, A_1, \ldots, A_T), (X_0, U_0, \ldots, U_{T-1}) \rangle
\]

where \( \mathcal{L}^*(A_0, A_1, \ldots, A_T) = (C_0, C_1, \ldots, C_T) \)
But \( \mathcal{L}^*(\Lambda_0', \Lambda_1, \ldots, \Lambda_T) = (\Lambda_0 - \Lambda_1' \Lambda_0', \ldots, \Lambda_{T-1} - \Lambda_T' \Lambda_{T-1} \Lambda_T' \Lambda_T) \)

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

\[
= \text{tr}[X_0 \Lambda_0'] + \sum_{t=1}^{T} \text{tr}[(-A_{t-1}'X_{t-1}A_{t-1} + X_t)\Lambda_t] \\
= \text{tr}[X_0 \Lambda_0'] + \sum_{t=1}^{T} \text{tr}[-X_{t-1}'A_{t-1}A_tA_{t-1} + X_t\Lambda_t] \\
= \text{tr}[X_0 \Lambda_0'] - X_0' \Lambda_1' \Lambda_0' + \text{tr}[X_1' \Lambda_1' - X_0'A_1' \Lambda_0'] + \ldots \\
\ldots + \text{tr}[X_T' \Lambda_T' - X_{T-1}'A_{T-1}' \Lambda_{T-1}' \Lambda_T + \text{tr}[X_T' \Lambda_T] \\
= \sum_{t=0}^{T-1} \text{tr}[X_t (\Lambda_t' - A_t' \Lambda_t' \Lambda_t)] + \text{tr}[X_T' \Lambda_T] \\
= <(X_0, X_1', \ldots, X_T), \mathcal{L}^*(\Lambda_0', \Lambda_1', \ldots, \Lambda_T)> \\
\square

Corollary 3

Suppose \( J = \sum_{t=0}^{T} \{c_t' x_t + \text{tr}[C_t x_t]\} \), \( C_t \) symmetric

where \( x_{t+1} = A_t x_t - A_t' c_t + u_t \); \( x_0 \) given

and \( x_{t+1} = A_t x_t A_t' + U_t \); \( x_0 \) given and symmetric, \( U_t \) symmetric.

Then \( J = \lambda_0' x_0 + \text{tr}[\Lambda_0' x_0'] + \sum_{t=1}^{T} \{\lambda_t' u_{t-1} + \text{tr}[\Lambda_t U_t]\} \)
where \( \lambda_t = A_t' \lambda_{t+1} + c_t \), \( \lambda_T = c_T \)

and \( A_t = A_t'A_{t+1} - A_t' c_{t+1} + c_t \), \( A_T = c_T \).

**Proof:**

Define formally the operator

\[
\mathcal{L}(x_0, x_1, \ldots, x_T, x_0', x_1', \ldots, x_T') \overset{\Delta}{=} (x_0', -A_0' x_0 + c_0, x_1', \ldots, -A_{T-1}' x_{T-1}) \\
+ A_{T-1}' x_{T-1} - x_0' x_0' - A_0' x_0' x_0 + x_1' \ldots, -A_{T-1}' x_{T-1} - x_0' x_0 + x_1' \ldots
\]

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

\[
\langle (c_0, c_1', \ldots, c_T', C_0, C_1', \ldots, C_T'), (x_0, x_1, \ldots, x_T, x_0', x_1', \ldots, x_T') \rangle
\]

\[
\Delta \overset{\Delta}{=} \sum_{t=0}^T \{ c_t' x_t + \text{tr}[C_t x_t] \}
\]

Evaluate \( J = \langle (c_0, c_1', \ldots, c_T', C_0, C_1', \ldots, C_T'), (x_0, x_1, \ldots, x_T, x_0', x_1', \ldots, x_T') \rangle \)

subject to

\[
\mathcal{L}(x_0', \ldots, x_T', x_0, \ldots, x_T) = (x_0', u_0, \ldots, u_{T-1}' x_0', u_0, \ldots, u_{T-1})
\]

From the lemma,

\[
J = \langle (\lambda_0, \lambda_1, \ldots, \lambda_T, A_0, A_1, \ldots, A_T), (x_0', u_0, \ldots, u_{T-1}', x_0', u_0, \ldots, u_{T-1}) \rangle
\]

where

\[
\mathcal{F}^*(\lambda_0, \ldots, \lambda_T, A_0, \ldots, A_T) = (c_0, \ldots, c_T, C_0, \ldots, C_T)
\]

But

\[
\mathcal{F}^*(\lambda_0, \ldots, \lambda_T, A_0, \ldots, A_T) = (\lambda_0 - A_0' \lambda_1, \ldots, \lambda_{T-1}' \lambda_T, \lambda_T, A_0 - A_0' A_0' + A_0' \lambda_0, \ldots, A_{T-1}' \lambda_T + A_{T-1}' \lambda_T c_{T-1} + A_T)
\]
Indeed \( \langle \mathcal{L}(x_0, \ldots, x_T, x_0, \ldots, x_T), (\lambda_0, \ldots, \lambda_T, \Lambda_0, \ldots, \Lambda_T) \rangle \)

\[
= x_0^t\lambda_0 + \sum_{t=1}^{T} \left( -A_{t-1}x_{t-1} + A_t x_{t-1} c_{t-1} + x_t \right)^t\lambda_t
+ \text{tr}[x_0\Lambda_0] + \sum_{t=1}^{T} \text{tr}[(-A_{t-1}x_{t-1}A_t' + x_t)\Lambda_t]
= x_0^t\lambda_0 + \sum_{t=1}^{T} \left( -x_{t-1}A_{t-1}' \lambda_t + \text{tr}[x_{t-1}A_{t-1}' \lambda_{t-1} c_{t-1}'] + x_t^t\lambda_t \right)
+ \text{tr}[x_0\Lambda_0] + \sum_{t=1}^{T} \text{tr}[-x_{t-1}A_{t-1}' \Lambda_{t-1} A_t + x_t^t\Lambda_t]
= x_0^t(\lambda_0 - A_0'^t\lambda_1) + x_1^t(\lambda_1 - A_1'^t\lambda_2) + \ldots + x_T^t\lambda_T
+ \text{tr}[x_0(A_0' - A_1' A_0 + A_0' \lambda_0 c_0')] + \ldots + \text{tr}[x_T A_T]
= \sum_{t=0}^{T-1} x_t^t(\lambda_t - A_t'^t\lambda_{t+1}) + x_T^t\lambda_T
+ \sum_{t=0}^{T-1} \text{tr}[x_t(A_t' - A_{t+1}' A_t + A_t' \lambda_{t+1} c_t')] + \text{tr}[x_T A_T]
= \langle (x_0, \ldots, x_T, x_0, \ldots, x_T), \mathcal{F}^*(\lambda_0, \ldots, \lambda_T, \Lambda_0, \ldots, \Lambda_T) \rangle
APPENDIX E

Derivation of the Information Matrix Equation (4.41)

Here again, the arguments are dropped from matrix notations for clarity. From (4.39)

\[
\left[ \frac{\mathbf{T}}{\mathbf{z}(\tau)} \right] \tau^{-1} = \mathbf{E}_{\alpha} \left\{ \begin{array}{c|c} \frac{\partial z(\tau)}{\partial \alpha_i} & \frac{\partial z(\tau)}{\partial \alpha_j} \\ \hline \end{array} \right\} \left( \tau^{-1} \right) \quad (D.1)
\]

and from (4.4)

\[
= \mathbf{E}_{\alpha} \left( \begin{array}{c} r's^{-1} \frac{\partial r}{\partial \alpha_i} - \frac{1}{2} r's^{-1} \frac{\partial s}{\partial \alpha_i} \quad s^{-1}r + \frac{1}{2} \text{tr} \left[ s^{-1} \frac{\partial s}{\partial \alpha_i} \right] \\
\end{array} \right) \left( \begin{array}{c} r's^{-1} \frac{\partial r}{\partial \alpha_j} - \frac{1}{2} r's^{-1} \frac{\partial s}{\partial \alpha_j} \quad s^{-1}r + \frac{1}{2} \text{tr} \left[ s^{-1} \frac{\partial s}{\partial \alpha_j} \right] \\
\end{array} \right) \left( \tau^{-1} \right) \quad (D.2)
\]

\[
= \mathbf{E}_{\alpha} \left( \begin{array}{c} r's^{-1} \frac{\partial r}{\partial \alpha_i} \left( r's^{-1} \frac{\partial r}{\partial \alpha_j} \right) \\
+ \frac{1}{4} \left( r's^{-1} \frac{\partial s}{\partial \alpha_i} \quad s^{-1}r \right) \left( r's^{-1} \frac{\partial s}{\partial \alpha_j} \quad s^{-1}r \right) \\
+ \frac{1}{4} \text{tr} \left[ s^{-1} \frac{\partial s}{\partial \alpha_i} \right] \text{tr} \left[ s^{-1} \frac{\partial s}{\partial \alpha_j} \right] \\
+ \left( r's^{-1} \frac{\partial r}{\partial \alpha_i} \right) \left( - \frac{1}{2} r's^{-1} \frac{\partial s}{\partial \alpha_j} \quad s^{-1}r \right) \\
+ \left( r's^{-1} \frac{\partial r}{\partial \alpha_i} \right) \left( \frac{1}{2} \text{tr} \left[ s^{-1} \frac{\partial s}{\partial \alpha_j} \right] \right) \\
\end{array} \right)
\]
Recall now from Chapter 2 that, conditioned on $z^{T-1}$, \( r(T;\alpha) \)
has zero mean and covariance $S_\alpha(T)$ and that \( \frac{\partial r(T;\alpha)}{\partial \alpha_i} \) is deterministic, with respect to the probability measure corresponding to $\alpha$. Therefore, using the identities

\[
\mathbb{E}\{(c'x)(x'Ax)\} = 0
\]

\[
\mathbb{E}\{(x'Ax)(x'Ax)\} = (\text{tr}[\Sigma A])^2 + 2 \text{tr}[\Sigma A\Sigma A]
\]

for \( x \sim N(0,\Sigma) \), \( A = A' \)

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

\[
\begin{pmatrix}
I \\
 z(T) \mid z^{T-1}(\alpha)
\end{pmatrix}_{ij} = \text{tr} \begin{bmatrix}
\frac{\partial r}{\partial \alpha_i} & \frac{\partial r}{\partial \alpha_j} & S^{-1}
\end{bmatrix}
\]

\[
+ \frac{1}{4} \left( \text{tr} \begin{bmatrix}
\frac{\partial S}{\partial \alpha_i} & S^{-1}
\end{bmatrix} \text{tr} \begin{bmatrix}
\frac{\partial S}{\partial \alpha_j} & S^{-1}
\end{bmatrix} + 2 \text{tr} \begin{bmatrix}
\frac{\partial S}{\partial \alpha_i} & S^{-1} & \frac{\partial S}{\partial \alpha_j} & S^{-1}
\end{bmatrix} \right)
\]
\[ + \frac{1}{4} \text{tr} \left[ s^{-1} \frac{\partial s}{\partial \alpha_i} \right] \text{tr} \left[ s^{-1} \frac{\partial s}{\partial \alpha_j} \right] \]

\[ - \frac{1}{4} \text{tr} \left[ \frac{\partial s}{\partial \alpha_i} s^{-1} \right] \text{tr} \left[ \frac{\partial s}{\partial \alpha_j} s^{-1} \right] \]

\[ - \frac{1}{4} \text{tr} \left[ \frac{\partial s}{\partial \alpha_i} s^{-1} \right] \text{tr} \left[ \frac{\partial s}{\partial \alpha_j} s^{-1} \right] \]  

(D.4)

\[
\begin{bmatrix}
\tilde{I}
\end{bmatrix}_{ij} = \text{tr} \left[ \frac{\partial x}{\partial \alpha_i} \frac{\partial x}{\partial \alpha_j} s^{-1} \right] + \frac{1}{2} \text{tr} \left[ \frac{\partial s}{\partial \alpha_i} s^{-1} \frac{\partial s}{\partial \alpha_j} s^{-1} \right] \quad (D.5)
\]

which is (4.41).
APPENDIX F

LTI Models used for results of Figure 5.7

Model of order 5

\[
A = \begin{bmatrix}
-0.5 & 0 & 0 & 0 & 0 \\
0 & -0.2 & 0 & 0 & 0 \\
0 & 0 & 0.1 & 0 & 0 \\
0 & 0 & 0 & 0.4 & 0 \\
0 & 0 & 0 & 0 & 0.7
\end{bmatrix} \quad L = \begin{bmatrix}
1.0 & 0 & 0 \\
1.0 & 0 & 0 \\
0 & 1.0 & 0 \\
0 & 1.0 & 0 \\
0 & 0 & 1.0
\end{bmatrix}
\]

\[
C = \begin{bmatrix}
1.0 & 1.0 & 0 & 0 & 0 \\
0 & 0 & 1.0 & 1.0 & 0 \\
0 & 0 & 0 & 0 & 1.0
\end{bmatrix}
\]

\[
B = I_5 \quad \Xi = 0.5 I_3 \quad \Theta = 0.5 \times 10^{-5} I_3
\]

Model of order 10

\[
A = \text{diag}(-0.5,-0.2,0.1,0.4,0.7,0.5,0.3,0.1,-0.3,-0.6)
\]

\[
L = \begin{bmatrix}
U & 0 & 0 \\
U & 0 & 0 \\
0 & U & 0
\end{bmatrix} \quad \text{where} \quad U \Delta \begin{bmatrix}
1.0 \\
1.0
\end{bmatrix}
\]
\[ C = \begin{bmatrix} \mathbf{U}' & \mathbf{0} \\ \mathbf{0} & \mathbf{U}' \\ \mathbf{0} & \mathbf{U}' \\ \mathbf{0} & \mathbf{U}' \end{bmatrix} \]

\[ \mathbf{B} = \mathbf{I}_{10} \quad \Xi = 0.5 \mathbf{I}_5 \quad \Theta = 0.5 \times 10^{-5} \mathbf{I}_5 \]

**Model of order 15**

\[ \mathbf{A} = \text{diag}(-0.5,-0.2,0.1,0.4,0.7,-0.6,0.4,0.6,0.2,0.5,-0.2,0.7, \\
0.3,0.5,0.4) \]

\[ \mathbf{L} = \begin{bmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \\ \mathbf{U} & \mathbf{U} \\ \mathbf{U} & \mathbf{U} \\ \mathbf{0} & \mathbf{U} \end{bmatrix} \]
\[
C = \begin{bmatrix}
  u' & 0 & & \\
  0 & u' & & \\
  & 0 & u' & \\
  & & & u'
\end{bmatrix}
\]

\[
B = I_{15} \quad \Xi = 0.5 I_8 \quad \Theta = 0.5 \times 10^{-5} I_8
\]
REFERENCES


