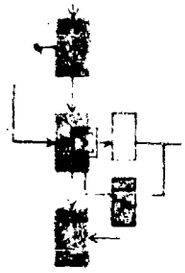
November 1976

Report ESL-R-718

Contract AFOSR-72-2273



Approved for public release;
distribution unlimited.

INFORMATION, CONSISTENT ESTIMATION AND DYNAMIC SYSTEM IDENTIFICATION

Yoram Baram

Electronic Systems Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS 02139

Department of Electrical Engineering and Computer Science

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DDC
This technical report has been reviewed and is
approved for public release IAW AFR 190-12 (7b).
Distribution is unlimited.
A. D. BLOSE
Technical Information Officer

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR - TR - 77 - 0283	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): INFORMATION, CONSISTENT ESTIMATION AND DYNAMIC SYSTEM IDENTIFICATION	5. TYPE OF REPORT & PERIOD COVERED: Interim
6. PERFORMING ORG. REPORT NUMBER: ESL-R-718	
7. AUTHOR(s): Yoram Baram	8. CONTRACT OR GRANT NUMBER(s): AFOSR [redacted]-2273-[illegible]
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Massachusetts Institute of Technology, Electronic Systems Laboratory, Cambridge, Massachusetts 02139	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F, 2304/A1
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, DC 20332	12. REPORT DATE: Nov[redacted]976
	13. NUMBER OF PAGES: 129
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The asymptotic behaviour of parameter estimates and the identification and modeling of dynamical systems are investigated. Measures of the relevant information in a given sequence of observations are defined and shown to possess useful properties, such as the metric property on the parameter set. The convergence of maximum likelihood and related Bayesian estimates for general observation sequences is investigated. The situation where the true parameter is not a member of a given parameter set is considered as well as the situation where the parameter set includes the true model. The finite

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE 127 200 UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

parameter set case is emphasized for simplicity in the convergence analysis, but the results are extended in general terms to the infinite parameter case. It is shown that under uniqueness conditions or the output statistics of linear dynamical systems identification procedures converge to the true model if it is a member of a given model set. If the true model is not a member of the set, then the estimates converge to a model in the set, closest to the actual system in the information metric sense. Stationary and non-stationary systems are considered. Rates of convergence in the mean are obtained, and the separate contributions of the stochastic and the deterministic parts of the input to the convergence rates are shown. The analysis also suggests methods for approximating a high order system by a low order model and for selecting a representative model from a given model set, applicable to infinite and even non-compact model sets.

0263

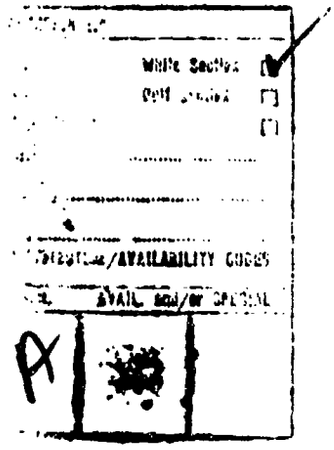
November 1976 Report ESL-R-718

INFORMATION, CONSISTENT ESTIMATION AND DYNAMIC SYSTEM IDENTIFICATION

by

Yoram Baram

This report is based on the unaltered thesis of Yoram Baram, submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology in November, 1976. The research was conducted at the Decision and Control Sciences group of the M.I.T. Electronic Systems Laboratory, with partial support provided by the Air Force Office of Scientific Research under Contract No. AFOSR-72-2273.



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

INFORMATION, CONSISTENT ESTIMATION AND DYNAMIC SYSTEM IDENTIFICATION

by

Yoram Baram

B.S., Technion, Israel Institute of Technology
1972

S.M., Massachusetts Institute of Technology
1974

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

November, 1976

Signature of Author.....
Department of Electrical Engineering and Computer Science
November 30, 1976

Certified by.....
Thesis Supervisor

Accepted by.....
Chairman, Departmental Committee on Graduate Students

INFORMATION, CONSISTENT ESTIMATION AND
DYNAMIC SYSTEM IDENTIFICATION

by

Yoram Baram

Submitted to the Department of Electrical Engineering and Computer Science on November 30, 1976 in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

ABSTRACT

The asymptotic behaviour of parameter estimates and the identification and modeling of dynamical systems are investigated. Measures of the relevant information in a given sequence of observations are defined and shown to possess useful properties, such as the metric property on the parameter set. The convergence of maximum likelihood and related Bayesian estimates for general observation sequences is investigated. The siutation where the true parameter is not a member of a given parameter set is considered as well as the situation where the parameter set includes the true model. The finite parameter set case is emphasized for simplicity in the convergence analysis, but the results are extended in general terms to the infinite parameter case. It is shown that under uniqueness conditions on the output statistics of linear dynamical systems identification procedures converge to the true model if it is a member of a given model set. If the true model is not a member of the set, then the estimates converge to a model in the set, closest to the actual system in the information metric sense. Stationary and non-stationary systems are considered. Rates of convergence in the mean are obtained, and the separate contributions of the stochastic and the deterministic parts of the input to the convergence rates are shown. The analysis also suggests methods for approximating a high order system by a low order model and for selecting a representative model from a given model set, applicable to infinite and even non-compact model sets.

THESIS SUPERVISOR: Nils R. Sandell, Jr.

TITLE: Assistant Professor of Electrical Engineering

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to Professor Nils Sandell for his invaluable advice, patience and constructive criticism throughout this effort. I thank Professors Peter Caines, Nils Sandell, Fred Schweppe and Alan Willsky for reading the thesis and making very useful comments and suggestions. I also thank Professor Willsky for serving as my academic advisor. Discussions with Professors Herman Chernoff, Richard Dudley, Adrian Segall and Michael Woodroofe at different stages of the research were very helpful. The M.I.T. Electronic Systems Laboratory provided an intellectually stimulating environment and financial support for my studies and this research.

I thank my wife Susan for her understanding and encouragement during the course of my studies.

This work was supported by the Air Force Office of Scientific Research under Contract AFOSR 72-2273.

TABLE OF CONTENTS

CONDITIONS, THEOREMS, LEMMAS AND COROLLARIES

GLOSSARY OF SYMBOLS

Symbol	Page	Symbol	Page
Γ_j^i	79	$\|x\|$	15
$\psi_{j,n}^i$	78	x^T	20
		$\|A\|$	80
ψ_j^i	79	$\|\|A\|\|$	103
Λ_n	104		
$\Phi_{j,n}^*$, $\nabla_{j,n}^*$, $\Theta_{j,n}^*$	111		
$\overline{\Phi}_{j,n}^*$	112		
$\Pi_{j,n}^*$	113		
μ, ν	13		
λ_n	13		
ν_o	17		
ν_n, λ	17		
$\delta_n(s;t)$	41		
μ_i	43		
γ_s	44		
$\lambda_{n,i}$	104		
*	12		
l_{s_j}	17		

CHAPTER I

INTRODUCTION

This thesis is concerned with some fundamental questions associated with the common problem of assigning a mathematical model to a physical phenomenon, using a set of observations. The situation is complicated by the fact that the relationship between the observations and the sought mathematical model is uncertain and can only be specified in a probabilistic framework. For mathematical tractability the problem is formulated as one of selecting via some criterion the "best" model from a specified set of models. The formulation of the mathematical problem requires, then, the choice of a model set on the one hand and the choice of a model selection criterion, on the other. The first choice presents an obvious tradeoff. The more strictly the model set is specified, the more tractable is the mathematical solution, but the less probable is the case that a correct model is included in the specified model set. As an illustration, consider the two extreme situations. If the model set consists of a single model, then the selection is trivial, but the model may not be an adequate representative of the observed phenomenon. On the other hand, if the model set is the abstract "set" of "all models", then it obviously contains the correct model, but a mathematical solution (or formulation) of the model selection problem is then not feasible.

The model set can be naturally specified in terms of a parameter set, such that to each parameter there corresponds a model and vice versa. The terms model set and parameter set will be used interchangeably and precise relationships between them are defined in the thesis. The model selection problem can then be naturally defined as a parameter estimation problem. Given a parameter set the problem formulation requires the selection of a parameter estimation criterion. The true parameter cannot, in general, be assumed to belong to the prespecified parameter set, as asserted above. It turns out that the maximum likelihood estimate, defined in Chapter 2 is most adequate for this situation. On the other hand, the Bayesian methods of maximum *a posteriori* probability and least squares, also defined in Chapter 2, intrinsically assume that the true parameter is a member of the model set.

One objective of this thesis is to provide in a very general setting answers to the following questions: Under what conditions do the maximum likelihood and the Bayesian estimates converge to some parameter in the parameter set? What distinguishes the selected model from the other models in the model set and what is its relationship to the true model? For the selection of an estimation procedure is it reasonable to assume that the true parameter is a member of the set when it is not? Is the true model selected when it is a member of the model set? A question that arises naturally in this setting is: what is the best approximation of a complex model by a simple one?

A particular problem of considerable practical significance is that

of dynamic system identification. The situation described above, and the questions raised, naturally apply to the system identification problem. In fact, this research has been motivated by the problem of identifying the dynamic equations of an aircraft during its operation throughout the flight envelope for the purpose of adaptive control. We analyze the asymptotic behaviour of system identification procedures in the presence and in the absence of the true model in a given model set. The analysis also suggests a systematic approach to certain system modeling problems of practical significance.

A major part of the analysis in this thesis will be restricted to the case where the model set is finite. This restriction serves several purposes. We chose to emphasize the statistical properties of the observation sequences involved (such as their content of information) and to avoid considerations of topological conditions on the parameter set, which are unavoidable if results for e.g. infinite compact parameter sets are desired. This makes the analysis considerably simpler, and enables us to consider very general classes of observation sequences. It is nevertheless demonstrated in Chapter 7 that the results obtained in this thesis for finite parameter sets may be extended to compact sets by additional requirements on the topology of the set, such as uniform continuity of the density functions involved. Further research in this direction is recommended.

In addition to the above consideration, the case of finite parameter sets has a considerable practical significance as a method of approximation. Identification techniques for finite sets of models are considerably faster than those for infinite sets, as the search procedure for the parameter satisfying the estimation criterion is practically trivial. In fact, this thesis makes a strong case for the finite model set, taking the viewpoint that the true model is in most cases not included in any prespecified set of models. Identification is thus a procedure of finding an approximate model whether a finite or an infinite model set is considered. The approximation is nevertheless "coarser" when fewer models are included in the model set.

It should, however, be emphasized that a substantial portion of the thesis applies to parameter sets that may be infinite and even noncompact. This is the case in the derivation of distance measures on the parameter set and the consideration of system modeling problems.

For comparison with earlier results we note that the convergence of the parameter estimates is considered in this thesis in the probabilistic senses of convergence almost everywhere (a.e.) and convergence in the mean square (m.s.), which will be defined in Chapter 2. Consistency is traditionally defined as convergence a.e. of the estimates to the true parameter when it is included in the parameter set.

1.1 Historical Review

Parameter estimation techniques have been studied ever since the

introduction of the maximum *a posteriori* probability (MAP) and the least squares (LS) criteria by Gauss [1809], and Laplace [1820] and their later studies by Edgeworth [1908]. Fisher [1922] proposed the maximum likelihood (ML) estimate, which has since gained considerable popularity due to its intuitive appeal and its asymptotic properties (e.g. LeCam [1953]).

The consistency of ML estimates for sequences of independent and identically distributed (i.i.d.) observations was proved by Cramer [1946] who assumed differentiability to 4'th order of the probability density functions involved. Differentiability assumptions were dispensed with in proofs by Doob [1934] and Wald [1949]. The main tool in proving consistency for i.i.d. observations, is, naturally, the strong law of large numbers. Roussas [1965] proved the consistency of ML estimates for the case of ergodic Markov observation sequences, employing the ergodic theorem. The m.s. convergence of LS estimates given i.i.d. observations was considered by Liporace [1971], who showed, via the multiplication rule for independent random variables, that the mean square error of these estimates is exponentially diminishing. In the case where the true parameter is not included in the parameter set, the estimates were shown to converge to a parameter in the set, which is most similar to the true parameter. The measure of similarity suggested by Liporace is related to the information measures introduced in this thesis. Caines [1975a] proved and applied the submartingale property of sequences of maximized likelihood ratios on finite parameter sets to prove the con-

sistency of ML estimates on such sets for a general class of observation sequences, satisfying a certain probabilistic condition. Baram and Sandell [1976] extended Caines' results to Bayesian estimates, which were shown to be consistent a.e. and in the mean square, and showed that Caines' condition applies to stationary Gaussian linear systems.

The identification of linear dynamical systems employing parameter estimation techniques has been studied intensively for over a decade. However, several consistency proofs that have appeared in the early literature have overlooked the fact that for consistent estimation on compact parameter sets, uniform convergence of the associated probability densities on the parameter set is necessary, while pointwise convergence only provides consistency for finite parameter sets. Correct consistency proofs have appeared in the laterature in recent years. Caines and Rissanen [1974] (see also Rissanen and Caines [1974]) proved the consistency of ML estimates for autoregressive and moving average (ARMA) observation sequences. Ljung proved the consistency of a general class of stochastic approximation techniques [1974a] and the consistency of a class of prediction error techniques [1974b]. (see also Ljung [1975]) Caines [1975b] proved consistency for stationary processes of a more general class of prediction error techniques, which includes the maximum likelihood technique for the case of stationary Gaussain observation sequences. The topological requirements specified by Caines [1975b] reduce in the finite parameter set case to a requirement that there exist a 1 to 1 correspondence between the parameter set and the set of

system's impulse responses, corresponding to the system's innovations representation. Similar conditions were suggested by Tse and Weinert [1975] (see also Tse [1976]) and by Hawkes and Moore [1976] (see also Moore and Hawkes [1974]), who considered the convergence of Bayesian estimates on finite sets of stationary Gaussian linear systems. The condition suggested by Baram and Sandell [1976] is a uniqueness condition on the output statistics associated with the different models in the model set. Other statistical conditions are motivated and derived in this thesis. We shall comment on the correspondence between parametric and statistical conditions in Chapter 7 as we suggest further study of this subject.

Information methods have been suggested by many authors for the solution of the related problems of hypothesis testing, signal selection and model indentification. In recent years Kullback's information measure (Kullback [1959]) has proved to be useful in the analysis of parameter estimation and model identification techniques. Akaike ([1972], [1974]) has related Kullback's information with certain versions of the ML criterion. Kullback's information measure was employed by Liporace [1971], and, following Liporace, by Hawkes and Moore [1976] in their studies of parameter estimates given i.i.d. and stationary Gaussian observations. In this thesis we define and employ information measures, which prove to posses valuable properties lacked by Kullback's information measure, such as the metric property on the parameter space. Other information measures defined and employed in the litera-

ture, will be mentioned in Chapter 3 as they are compared with the information measures defined in this thesis.

1.2 Organization and Results

In the first part of the thesis (Chapters 2, 3 and 4) we consider general classes of observation sequences and parameter sets. The results are specialized to linear dynamical systems in Chapters 5 and 6. Familiarity with advanced concepts of probability theory is only required in Chapter 2 and parts of Chapter 4. The sequence of Chapter 3, sections 4.1 and 4.4, Chapter 5 and section 6.3 provides a consistent discussion of the information approach to system identification and modeling, which is the mainstream of the thesis. The rest of Chapter 4 is believed to be of theoretical interest and also of practical value, which is demonstrated in sections 6.1 and 6.2.

In Chapter 2 we present the underlying probabilistic set up for the thesis and recall definitions and results from probability and estimation theory used in the thesis. Since parameter estimates may be based on the possibly incorrect assumption that the true parameter is a member of a given parameter set, we define the different probability spaces in which the estimates are defined and in which the analysis is performed.

In Chapter 3 we define two measures of the relevant information in each observation favoring one parameter in the parameter set against another. Both measures will prove useful in later analysis. The information measures are shown to be metrics, or distance measures on the para-

meter set and to provide a measure of closeness of each parameter in the set to the true parameter which is not necessarily a member of the set. The information measures defined in this chapter are compared with other measures of information common in statistics and information theory.

In Chapter 4 we investigate the convergence of maximum likelihood and Bayesian parameter estimates for general classes of observation sequences. Consistency conditions are derived in terms of the information in the observations and extended to the case where the true parameter is not a member of the parameter set. Rates of convergence in the mean for the ML and MAP procedures are also derived.

In Chapter 5 we analyze the identification and modeling of stationary Gaussian linear systems. We show that the identification procedures under consideration converge under a certain uniqueness condition to the true model if it is included in the model set. If the true model is not a member of the model set the identification procedures converge to the model in the set whose output statistics are best matched to those of the true model. The selected model is also shown to be closest to the true model in the information metric used. It is then shown that under the uniqueness condition likelihood ratios and *a posteriori* probability ratios converge in the mean at rates faster than exponential. The analysis also suggests solutions to other modeling problems, such as the approximation of a complex system by a simple model and an optimal representation of a model set by a single model.

In Chapter 6 we consider general classes of time varying linear systems. In particular, we interpret for such systems the information

conditions derived in Chapter 4, and obtain consistency conditions in terms of the output statistics associated with the different models in the model set. The L_1 convergence of the likelihood and the *a posteriori* probability ratios is investigated and the separate contributions of the stochastic and the deterministic parts of the input to the information and, consequently, to the L_1 convergence rates are shown.

In Chapter 7 we suggest further research of possible extension and application of the theory. In particular, we show how the convergence results obtained in this thesis for finite sets of parameters may be extended to compact parameter sets. We also suggest further investigation of the problem of existence and uniqueness of a solution to the estimation, or identification problem. Then we suggest further study of the identifiability of dynamic systems via application of deterministic input sequences. Finally, we suggest applications of the theory to classes of problems, not directly addressed in this thesis, such as the identifiability of non-linear systems and periodically varying linear systems.

CHAPTER II

PRELIMINARIES: PROBABILITY SPACES, PARAMETER ESTIMATES AND STOCHASTIC CONVERGENCE

The purpose of this chapter is to present the underlying mathematical set up for this thesis and to recall definitions and results from probability and estimation theory that will be used in the following chapters.

Since a major objective of this thesis is to analyze, using correct assumptions, parameter estimates that may be based on incorrect assumptions, it is essential to define at the outset the different probabilistic frameworks in which the estimates are defined and in which the analysis is performed. We first introduce the correct framework in which the analysis is performed. It consists of an underlying probability space and a separate parameter space, of which the true parameter may or may not be a member. Likelihood ratios and maximum likelihood estimates are naturally defined in this framework. On the other hand, Bayesian parameter estimates are defined in a different framework where the parameter space is a part of the underlying sample space. Consequently, the existence of a probability measure defined on the parameter space (i.e. assigning to each set in the parameter space the probability that it includes the true parameter) is postulated. The Bayesian framework then inherently includes the assumption that the true parameter is a member of the given parameter space, and is inadequate for the ana-

of the general case considered in this study. Thus, while the Bayesian set up is assumed in the definition of Bayesian estimates, the analysis of these estimates, as well as the maximum likelihood estimate, is performed using the underlying, non-Bayesian framework.

Readers unfamiliar with the notion of measure and probability spaces may identify here, and in the following chapters, the functions $f(z^n)$, $f(z_n|z^{n-1})$ and $f(s|z^n)$ with the familiar probability density, conditional probability density and *a posteriori* probability density functions on Euclidean observation and parameter spaces. Several symbols and terms, mostly standard in probability and estimation theory, are introduced in this chapter. For other terms and symbols, defined throughout the thesis where they are used, the reader is referred to the symbol list.

2.1 Observations, Parameters and Likelihood Ratios

Consider a measurable space (Ω, U) where Ω is some sample space and U is a σ-algebra of subsets of Ω. The observation sequence (z_n) is a stochastic process on a probability space (Ω, U, P_*) with values in a measurable space $(D, \mathcal{D})$, called the observation space. We shall be interested in the case $(D, \mathcal{D}) = (R^\ell, B^\ell)$ where R^ℓ is the ℓ-dimensional Euclidean space and B^ℓ is the σ-algebra of Borel sets in R^ℓ. We call P_* the true measure and * the true parameter.

The parameter space S is a set such that for each $s \in S$ there exists a probability measure P_s defined on (Ω, U). Let $T \equiv (* \cup S)$.

Obviously, $* \in T$, but $*$ need not belong to the set S.

For each $s \in T$ we denote by E_s expectation taken with respect to P_s. We use the notation a.e. (almost everywhere) to denote events of P_* measure one. Events of P_s measure one will be denoted a.e. P_s.

Recall that the conditional expectation of a random variable x on (Ω, U, P) given $A \in U$ is a U-measurable random variable denoted $E^A(x)$ such that

$$E\, E^A(x) = E(x) \tag{2.0}$$

For each $s \in S$ we shall denote by E_s^A the conditional expectation given A, taken with respect to P_s.

If μ and ν are measures defined on (Ω, U) then μ is said to be absolutely continuous with respect to ν if for any set $A \in U$ $\nu(A) = 0$ implies $\mu(A) = 0$. μ is said to be singular[*] with respect to ν if it is not absolutely continuous with respect to ν.

Let $(U_n) \equiv (U_n(z^n))$ be the increasing family of σ-subalgebras of U, generated by

$$z^n \equiv (z_1, \ldots, z_n) \tag{2.1}$$

For each $s \in T$ and for each $n \geq 0$ let $P_{s,n}$ denote the restriction of P_s to U_n. Suppose that for each $n \geq 0$ the measures $P_{s,n}$ are absolutely continuous with respect to some measure λ_n defined on (Ω, U_n). Then

[*] This is not a standard definition. For a definition of mutually singular measures see Rudin [1966], p. 121.

$$f_{s,n} \equiv \frac{d P_{s,n}}{d \lambda_n} \quad ; \quad s \in T \tag{2.2}$$

are the Radon-Nikodym derivatives (or densities) between the respective measures. The likelihood ratio between two parameters s, t $\in T$ is defined as

$$h^s_{t,n} \equiv \frac{d P_{s,n}}{d P_{t,n}} = \frac{f_{s,n}}{f_{t,n}} \tag{2.3}$$

provided that $P_{s,n}$ is absolutely continuous with respect to $P_{t,n}$. When the time parameter n is included in the argument we shall use the somewhat shorter notation

$$f_s(a(n)) \equiv f_{s,n}(a(n)) \; ; \; s \in T \; ; \; a(n) \in U_n$$

in particular

$$f_s(z^n) \equiv f_{s,n}(z^n) \; ; \; s \in T \tag{2.4}$$

$$h^t_s(z^n) \equiv h^t_{s,n}(z^n) \; ; \; s, \; t \in T \tag{2.5}$$

For any c $\in U_n$ and b $\in U_n$ such that $f_{s,n}(b) \neq 0$ for all s $\in S$, the conditional densities of c given b are

$$f_{s,n}(c|b) \equiv \frac{f_{s,n}(c, b)}{f_{s,n}(b)}$$

in particular

$$f_s(z_n|z^{n-1}) = \frac{f_s(z^n)}{f_s(z^{n-1})} \quad ; \quad s \in T \tag{2.6}$$

The conditional likelihood ratios are then defined as

$$h_t^s(z_n|z^{n-1}) \equiv \frac{f_s(z_n|z^{n-1})}{f_t(z_n|z^{n-1})} \; ; \; s, t \in T \tag{2.7}$$

for any z^n such that $f_t(z_n|z^{n-1}) \neq 0$ for all $t \in T$.

The following condition will be assumed throughout the thesis

(c2.1) For all $s \in S$ the probability measures $P_{s,n}$ are mutually absolutely continuous.

2.2 Bayesian Probability Densities

Consider a measurable space $(\Omega, \mathcal{U})$, where Ω is some sample space and $\mathcal{U}$ is a σ-algebra of subsets of Ω, and a measurable space $(S, \mathcal{U}^S)$, where S is the parameter space and $\mathcal{U}^S$ is a σ-algebra of subsets of S. Let $(\Omega^b, \mathcal{U}^b)$ be a measurable space, where

$$\Omega^b \equiv \Omega \times S$$

and

$$\mathcal{U}^b \equiv \mathcal{U} \times \mathcal{U}^S$$

are the cartesian products of the respective sample spaces and σ-algebras. Let P^b be a measure on $(\Omega^b, \mathcal{U}^b)$. We denote by E^b expectation and by E^{b^A} ; $A \in \mathcal{U}$ conditional expectation given A, taken with respect to P^b. We call the restriction P_o^b of P^b to $(S, \mathcal{U}^S)$ the *a priori* probability measure on $(S, \mathcal{U}^S)$. Suppose that P_o^b is absolutely continuous with re-

spect to some measure ν_o on $(S, \mathcal{U}^S)$, then the density

$$f_o^b \equiv \frac{d\, P_o^b}{d\, \nu_o} \tag{2.8}$$

is well defined. In particular, we call

$$f_o^b(s) \; ; \; s \in S \tag{2.9}$$

the *a priori* probability density on S with respect to the measure ν_o.

Let (z_n) be a stochastic process on the probability space $(\Omega^b, \mathcal{U}^b, P^b)$ with values in a measurable space $(\mathcal{D}, \mathcal{D})$, and let $(\mathcal{U}_n^b) \equiv (\mathcal{U}_n^b(z^n))$ be the increasing sequence of σ-subalgebras of $\mathcal{U}^b$, generated by $z^n \equiv (z_1, \ldots, z_n)$. Let $P_n^b \; ; \; n \geq 1$ be the restriction of P^b to $\mathcal{U}_n^b$ and for each $n \geq 1$ let P_n^b be absolutely continuous with respect to some measure ν_n defined on $(\Omega^b, \mathcal{U}^b)$. Then the density

$$f_n^b \equiv \frac{d\, P_n^b}{d\, \nu_n} \tag{2.10}$$

is well defined. We shall be particularly interested in the *a posteriori* probability density of s, given z^n

$$f^b(s|z^n) \equiv f_n^b(s|z^n) \equiv \frac{f_n^b(s, z^n)}{f_n^b(z^n)} \; ; \; n \geq 1 \tag{2.11}$$

assuming $f_n^b(z^n) \neq 0$.

Let the parameter set S be finite, i.e.

$$S = \{s_j \; ; \; j \in K \equiv (0,\ldots,p)\} \tag{2.12}$$

For each $j \in K$ let

$$1_{s_j}(s) \equiv \begin{cases} 1 & s = s_j \\ 0 & s \neq s_j \end{cases}$$

Then

$$\nu_o \equiv \sum_{j=0}^{p} 1_{s_j}(s)$$

is a measure ("the counting measure") on $(S, U^{\mathcal{B}})$. Let λ be a measure on (Ω, U) then $\nu_n = \nu_o \cdot \lambda \;\; ; \; n \geq 1$ is the product measure on (Ω^b, U^b). Suppose that P_n^b is absolutely continuous with respect to ν_n (i.e. the entire measure P_n^b is concentrated on the set $\Omega \times \{s_i \; ; \; i \in K\}$) for all $n \geq 0$, then we have

$$f_o^b(s) = \sum_{i=0}^{p} f_o^b(s_i)\, 1_{s_i}(s) \tag{2.13}$$

and

$$f_n^b(s, z^n) = \sum_{i=0}^{p} f_n^b(s_i, z^n)\, 1_{s_i}(s) \tag{2.14}$$

Hence

$$f_n^b(z^n) = \int_S f_n^b(s, z^n)\, d\nu_o$$

$$= \sum_{i=0}^{p} f_n^b(s_i, z^n)$$

$$= \sum_{i=0}^{p} f_o^b(s_i)\, f^b(z^n|s_i) \tag{2.15}$$

where we have applied Bayes rule

$$f_n^b(s_i, z^n) = f_o^b(s_i)\, f_n^b(z^n|s_i) \tag{2.16}$$

Substituting (2.15) and (2.16) into (2.11) yields for each $j \in k$

$$f^b(s_j|z^n) = \frac{f_o^b(s_j)\, f_n^b(z^n|s_j)}{\sum_{i=0}^{p} f_o^b(s_i)\, f_n^b(z^n|s_i)} \tag{2.17}$$

Note that

$$f_n^b(z^n|s_j) = f_{s_j}(z^n) \tag{2.18}$$

where the right hand side is defined by (2.4). Thus, finally, for the finite parameter set S

$$f^{b}(s_j|z^n) = \frac{f_o^b(s_j)\, f_{s_j}(z^n)}{\sum_{i=0}^{p} f_o^b(s_i)\, f_{s_i}(z^n)} \tag{2.19}$$

2.3 Parameter Estimates and Stochastic Convergence

An estimate $\hat{s}_n$ on S is a U_n-measurable mapping from Ω onto S.

A maximum likelihood (ML) estimate on S is an estimate $\hat{s}_n \in S$ such that

$$\left\{f_s(z^n) \; ; \; s \in S\right\} \leq f_{\hat{s}_n}(z^n)$$

A maximum *a posteriori* probability (MAP) estimate on S is an estimate $\hat{s}_n \in S$ such that

$$\left\{f^b(s|z^n) \; ; \; s \in S\right\} \leq f^b(\hat{s}_n|z^n)$$

Let S be linear. Then a least-squares (LS) estimate on S is an

estimate $\hat{s}_n \in S$ such that

$$E^b \left\{ (s_n' - *)^T (s_n' - *) \right\} \geq E^b \left\{ (\hat{s}_n - *)^T (s - *) \right\}$$

for any estimate s_n' on S. x^T denotes x transposed and * denotes the true parameter, assumed to have the same dimension as s.

Let the true parameter be assumed to belong to a finite set $\{s_j \in R^m;\ j \in k\}$. Then the LS estimate on R^m at instant n is the conditional expectation

$$E^b \overset{U_n}{s} = \int_S s\, f^b(s|z^n)\, d\nu_o$$

$$= \sum_{j=0}^{p} s_j\, f^b(s_j|z^n) \qquad (2.20)$$

A stochastic sequence (x_n) on (Ω, U, P) is said to converge almost everywhere (a.e.) to a random variable x on (Ω, U, P) if

$$\lim_{n\to\infty} x_n = x \qquad \text{a.e.}$$

A stochastic sequence (x_n) on (Ω, U, P) is said to converge in the mean (or in L_1) to a random variable x on (Ω, U, P) if

$$\lim_{n\to\infty} E|x_n - x| = 0$$

A vector-valued stochastic sequence (x_n) on (Ω, U, P) is said to converge in the mean square (in m.s. or in L_2) to a random vector x on (Ω, U, P) if

$$\lim_{n\to\infty} E|x_n - x|^2 = 0$$

where $|x| \equiv (x^T x)^{\frac{1}{2}}$.

A sequence of parameter estimates $(\hat{s}_n)$ is said to be consistent a.e. or in the mean square if it converges a.e. or in the mean square to the true parameter.

We now present without proofs three well known results from the probability theory, which are used in this thesis.

Theorem 2.1 (Jensen's inequality, e.g. Bauer [1972], p. 322).

Let x be a real integrable random variable on a probability space (Ω, U, P) with values in R^1, and let g(x) be a convex integrable function on R^1, then

$$g(Ex) \leq E\, g(x)$$

Theorem 2.2 (Fatou's Lemma, e.g. Bauer [1972], p. 71)

Let (x_n) be an integrable stochastic sequence on (Ω, U, P) such that $x_n \geq 0$ a.e. for all n, then

$$E \liminf_{n\to\infty} x_n \leq \liminf_{n\to\infty} E\, x_n$$

Theorem 2.3 (Lebesgue's dominated convergence theorem, e.g. Chung [1974], p. 42)

Let (x_n) be an integrable stochastic sequence on (Ω, U, P). Then if

$$\lim_{n\to\infty} x_n = x \quad \text{a.e.}$$

where x is an integrable random variable on (Ω, U, P) and if there exists some integrable random variable y on (Ω, U, P) such that

$$E\, y < \infty$$

and

$$|x_n| \leq y \quad \text{a.e.} \quad \text{for all } n$$

then

$$\lim_{n\to\infty} E\, x_n = E\, x.$$

2.4 Martingales and Martingale Difference Sequences

Let (Ω, U, P) be a probability space and let (U_n) be an increasing family of σ-subalgebras in U. A U_n-measurable stochastic sequence (x_n) on (Ω, U, P) is called a U_n-martingale if for each n

(a) $E|x_n| < \infty$

(b) $E^{U_{n-1}} x_n = x_{n-1}$ a.e.

If the equality in (b) is replaced by $\leq$ then (x_n) is called a U_n-supermartingale.

It can be shown (Doob [1953], p. 93) that the likelihood ratio sequences $\left(\frac{d P_{s,n}}{d P_{*,n}}\right)$; $s \in S$, defined in section 2.2 are U_n-martingales according to the measure P_*. Hence

$$E_*^{U_{n-1}} \frac{d P_{s,n}/d P_{s,n-1}}{d P_{*,n}/d P_{*,n-1}} = \frac{d P_{*,n-1}}{d P_{s,n-1}} E_*^{U_{n-1}} \frac{d P_{s,n}}{d P_{*,n}}$$

$$= \frac{d P_{*,n-1}}{d P_{s,n-1}} \frac{d P_{s,n-1}}{d P_{*,n-1}}$$

$$= 1$$

Consequently, we have by (2.7), (2.6) and (2.2)

$$E_*^{U_{n-1}} h_*^s (z_n | z^{n-1}) = E_*^{U_{n-1}} \frac{f_s(z_n | z^{n-1})}{f_*(z_n | z^{n-1})} = 1 \quad \text{for each } s \in S \tag{2.21}$$

Theorem 2.4 (The martingale convergence theorem, e.g. Chung [1974], p. 334, Bauer [1972], pp. 341-343)

Let (x_n) be a U_n-martingale on (Ω, U, P) and let

$$\sup_{n \geq 0} E\, x_n^+ < \infty$$

where

$$x_n^+ = \sup (x_n, 0)$$

Then (x_n) converges a.e. to a finite limit.

Let (Ω, U, P) be a probability space and let (U_n) be an increasing family of σ-subalgebras in U. A U_n-measurable stochastic sequence (x_n) on (Ω, U, P) is called a U_n-martingale difference sequence if it is integrable and if

$$E^{U_{n-1}} x_n = 0 \quad \text{a.e.}$$

Let y_n be a stochastic sequence on (Ω, U, P) and let (U_n) be a sequence of σ-subalgebras of U, generated by $(y_1,\ldots,y_n)$. Then, clearly

$$(y_n - E^{U_{n-1}} y_n)$$

is a U_n-martingale difference sequence. Also note that if (x_n) is a U_n-martingale difference sequence then

$$X_n \equiv \sum_{m=1}^{n} x_m$$

is a martingale. Indeed

$$E^{U_{n-1}} X_n = \sum_{m=1}^{n} E^{U_{n-1}} x_m$$

$$= \sum_{m=1}^{n-1} x_m + E^{U_{n-1}} x_n$$

$$= \sum_{m=1}^{n-1} x_m = X_{n-1}$$

2.5 Stationarity and Ergodicity

The purpose of this section is to provide definitions and convergence results for ergodic sequences, which will be used in the thesis. It is not intended to provide an elaborate presentation of the concept of ergodicity. For a thorough development of ergodic theory the reader is referred to, e.g., Doob [1953], Halmos [1956] and Chacon and Ornstein [1959].

Consider a probability space (Ω, U, P). A transformation T from Ω to U is said to be measure preserving if

$$P(T^{-1}A) = P(A)$$

for all $A \in U$.

Given a measure preserving transformation T, a U-measurable event A is said to be invariant if

$$T^{-1}A = A$$

Let (x_n) be a stochastic sequence on (Ω, U, P) with values in (R^ℓ, B^ℓ), where R^ℓ is the ℓ-dimensional Euclidean space and B^ℓ is the σ algebra of Borel sets of R^ℓ. Let B_∞^ℓ be the σ-algebra of Borel sets of R_∞^ℓ where $R_\infty^\ell \equiv R^\ell \times R^\ell \times \ldots$ Then (x_n) is said to be stationary if for each $k \geq 1$

$$P\left[(x_1,\ldots,x_n) \in C\right] = P\left[(x_{k+1}, x_{k+2},\ldots) \in C\right]$$

for every $C \in B_\infty^\ell$.

A stationary sequence (x_n) on (Ω, U, P) is said to be ergodic if every invariant event in U has probability zero or one. It can be shown (e.g. Stout [1974], p. 168) that (x_n) is generated by a measure preserving transformation T (the shift operator), i.e.

$$x_n(\omega) = x_{n-1}(T\,\omega) \tag{2.22}$$

Let (x_n) be a vector valued stochastic sequence from (Ω, U, P) into (R^ℓ, B^ℓ) such that the probability density with respect to the Lebesgue measure on (R^ℓ, B^ℓ) of (x_n) is Gaussian on R^ℓ, with

$$E\,x_n = m_x, \text{ constant for all } n$$

and

$$E\left\{(x_n - m_x)(x_{n+k} - m_x)^T\right\} \text{ depends only on } k.$$

Then (x_n) is a stationary Gaussian sequence.

Proposition 2.1 (Grenander [1959], pp. 257-260 and Doob [1953, p. 494)

A zero mean stationary Gaussian process is ergodic is and only is

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} |R(k)|^2 = 0$$

where

$$R(k) \equiv E\left\{x_n\, x_{n+k}^T\right\}$$

and where $|R(k)|$ denotes the determinant of $R(k)$.

Theorem 2.5 (The ergodic theorem, e.g. Doob [1953], p. 464, Halmos [1956], p. 22, Weiner [1949], p. 16)

Let (x_n) be an ergodic sequence on (Ω, U, P) and let $f(x_n)$ be a U-measurable function such that $E|f(x_o)|$ is finite, then

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{j=0}^{n} f(x_j) = E\ f(x_o).$$

The following version of the central limit theorem of probability theory will prove useful in later chapters.

Theorem 2.6 (Billingsley [1961])

Let (x_n) be an ergodic stochastic process on (Ω, U, P) such that $E\ x_1^2$ is finite and

$$E^{U_{n-1}} x_n = 0 \quad \text{a.e.}$$

(i.e. (x_n) is an ergodic martingale difference sequence). Then the distribution of $n^{-\frac{1}{2}} \sum_{k=1}^{n} x_k$ approaches the Gaussian distribution with mean zero and variance $E\ x_1^2$.

2.6 Metric Spaces and Stochastic Metrics

Consider a set S and a real-valued function e on $S \times S$ which satisfied

(i) $e(s;s) = 0$ for any $s \in S$

(ii) $e(s;t) = e(t;s)$ for any $s, t \in S$

(iii) $e(s;t) \leq e(s;r) + e(r;t)$ for any $s, t, r \in S$.

Then e is called a pseudo metric on S. If in addition to (i), (ii) and (iii) e satisfies

(iv) $e(s;t) = 0$; $s, t \in S$ implies $s=t$

then e is called a metric on S. The pair (S, e) is called a metric space.

Now consider a probability space (Ω, U, P) and an increasing family (U_n) of σ-subalgebras of U. Let (e_n) be a (U_n)-measurable sequence of functions on $S \times S$ such that each e_n satisfies (i) - (iii) above. Then we shall call (e_n) a stochastic pseudo metric sequence(*) on S. If each e_n satisfies (i) - (iv) above, we shall call (e_n) a stochastic metric sequence(*) on S.

(*) These definitions do not seem to have appeared in the literature before.

CHAPTER III

INFORMATION

In this chapter we develop the notion of the information in a sequence of observations favoring one parameter in a given parameter set against another. We do not make the assumption, common in the derivation of other information measures in information theory, that the true parameter is included in a known set, or, equivalently, that the true measure belongs to a known set of measures. The mean and the conditional mean values of the discriminating information in a single observation are shown to possess properties that will prove useful in the following chapters. In particular, their absolute values are metrics, or distance measures, on the parameter space. This provides a meaningful measure of the relative closeness of parameters to the true parameter. The new information measures are then compared with other measures common in information theory.

3.1 The Information in a Single Observation

Let S be a parameter space and let $T = (* \cup S)$, where $*$ is the true parameter. If for some pair of parameters $s, t \in T$

$$f_s(Z^n) > f_t(Z^n)$$

or, equivalently,

$$\log f_s(Z^n) > \log f_t(Z^n)$$

we say that the parameter s is favored over the parameter t by the observations z^n. Then $\log f_s(z^n)$ may be regarded as a measure of the information in z^n for selecting a parameter from the set T. The difference

$$\log f_s(z^n) - \log f_t(z^n) = \log h_t^s(z^n) \tag{3.1}$$

is then a measure of the information in z^n for selecting between s and t. If (3.1) is positive then s is favored and if it is negative then t is favored. The difference

$$\log h_t^s(z^n) - \log h_t^s(z^{n-1}) = \log h_t^s(z_n|z^{n-1}) \tag{3.2}$$

is then a measure of the difference between the information favoring s against t at instant n and the information favoring s against t at instant n-1. It can then be regarded as a measure of the information favoring s against t in the observation z_n. We define

$$I_n(s;t) \equiv E_*^{U_{n-1}} \log h_t^s(z_n|z^{n-1}) \tag{3.3}$$

as the conditional mean information in z_n favoring s against t and

$$\bar{I}_n(s;t) \equiv E_* \log h_t^s(z_n|z^{n-1}) \tag{3.4}$$

as the mean information in z_n favoring s against t. (A more general form of (3.3) would be $\tilde{I}_n(s;t) \equiv E_*^{A_{n-1}} \log h_t^s(z_n|z^{n-1})$ for some sequence (A_n) such that $A_n \in U_n$. However, for the purposes of this thesis we

use the information defined by (3.4).)

3.2 Properties of Information

We now show some properties of the information measures defined above that will prove to be useful in the following chapters.

Theorem 3.1

Let S be a parameter space. Then for any $s \in S$ and for each $n \geq 0$ we have

$$I_n(*;s) \geq 0 \qquad \text{a.e.}$$

and

$$\bar{I}_n(*;s) \geq 0$$

with equality if and only if $f_s(z_n|z^{n-1}) = f_*(z_n|z^{n-1})$ a.e.

Proof

$$I_n(*;s) = -E_*^{U_{n-1}} \log h_*^s(z_n|z^{n-1})$$

Using the inequality

$$\log a \leq a-1 \; ; \; \log a = a-1 \text{ if and only if } a = 1 \tag{3.5}$$

We get

$$I_n(*;s) \geq 1 - E_*^{U_{n-1}} h_*^s(z_n|z^{n-1}) = 0 \quad \text{a.e.} \tag{3.6}$$

where the second equality follows from (2.21). To show that equality holds only if $f_s(z_n|z^{n-1}) = f_*(z_n|z^{n-1})$ a.e. (sufficiency is trivial)

suppose that

$$I_n(*;s) = 1 - E_*^{U_{n-1}} h_*^s (z_n|Z^{n-1}) \quad \text{a.e.}$$

i.e.

$$E_*^{U_{n-1}} \left\{ h_*^s (z_n|Z^{n-1}) - \log h_*^s (z_n|Z^{n-1}) - 1 \right\} = 0 \quad \text{a.e.}$$

By (2.0) we then have

$$\int [h_*^s (z_n|Z^{n-1}) - \log h_*^s (z_n|Z^{n-1}) - 1] dP_* = 0 \qquad (3.7)$$

(3.7) and (3.5) together give

$$h_*^s (z_n|Z^{n-1}) = 1 \quad \text{a.e.}$$

or

$$f_s(z_n|Z^{n-1}) = f_*(z_n|Z^{n-1}) \quad \text{a.e.} \qquad (3.8)$$

Hence, equality in (3.6) holds if and only if (3.8) holds. Similarly, since $I_n(*;s) \geq 0$, we have

$$\bar{I}_n(*;s) = E_* I_n(*;s) \geq 0$$

with equality if and only if $I_n(*;s) = 0$ a.e., which, as shown above, occurs if and only if $f_s(z_n|Z^{n-1}) = f_*(z_n|Z^{n-1})$ a.e. ■

Corollary 3.1

Suppose that $r \in S$ is the true parameter. Then for any $t \in S$ $I_n(s;t)$ and $\bar{I}_n(s;t)$ are maximized on S at $s = r$. This maximum is

unique unless for some $s \in S$ $f_s(z_n|z^{n-1}) = f_r(z_n|z^{n-1})$ a.e.

Proof

By theorem 3.1 we have

$$I_n(r;t) - I_n(s;t) = I_n(r;s) \geq 0 \quad \text{a.e.}$$

and

$$\overline{I}_n(r;t) - \overline{I}_n(s;t) = \overline{I}_n(r;s) \geq 0$$

with equality if and only if $f_s(z_n|z^{n-1}) = f_r(z_n|z^{n-1})$ a.e. The assertion follows. ■

Theorem 3.2

The sequence $(|\overline{I}_n(s;t)|)$; $s, t \in S$ is a sequence of pseudo metrics on S. It is a sequence of metrics on S if an only if $\overline{I}_n(s;t) = 0$ implies $s = t$. The sequence $(|I_n(s;t)|)$; $s, t \in S$ is a stochastic sequence of pseudo metrics on S. It is a stochastic sequence of metrics if and only if $I_n(s;t) = 0$ implies $s = t$.

Proof

To prove that $|\overline{I}_n(s;t)|$ is a pseudo metric on S for each n we have to show (see section 2.6) that for each n it satisfies the following conditions.

(i) $|\overline{I}_n(s;s)| = 0$ for any $s \in S$

(ii) $|\overline{I}_n(s;t)| = |\overline{I}_n(t;s)|$ for any $s, t \in S$

(iii) $|\overline{I}_n(s;t)| \leq |\overline{I}_n(s;r)| + |\overline{I}_n(r;t)|$ for any $s, t, r \in S$.

We have

$$h_s^s\ (z_n|z^{n-1}) = 1$$

Hence

$$\log h_s^s\ (z_n|z^{n-1}) = 0$$

Then also

$$E \log h_s^s\ (z_n|z^{n-1}) = 0$$

and (i) follows. Also

$$\bar{I}(s;t) = -\bar{I}(t;s)$$

and (ii) follows.

Condition (iii) is proved as follows

$$|\bar{I}_n(s;r)| + |\bar{I}_n(r;t)|$$

$$= |E_* \log h_r^s\ (z_n|z^{n-1})| + |E_* \log h_t^r\ (z_n|z^{n-1})|$$

$$= |E_* \log f_s(z_n|z^{n-1}) - E_* \log f_r(z_n|z^{n-1})|$$

$$+ |E_* \log f_r(z_n|z^{n-1}) - E_* \log f_t(z_n|z^{n-1})|$$

$$\geq |E_* \log f_s(z_n|z^{n-1}) - E_* \log f_t(z_n|z^{n-1})|$$

$$= |E_* \log h_t^s\ (z_n|z^{n-1})| = |\bar{I}_n(s;t)|$$

If in addition to (i), (ii) and (iii) $|\bar{I}_n(s;t)|$ satisfies

(iv) $\quad |\bar{I}_n(s;t)| = 0$; $s, t \in S$ implies $s = t$

then $|\bar{I}_n(s;t)|$ is a metric on S. The assertion follows for $|\bar{I}_n(s;t)|$. The result for $|I_n(s;t)|$ is obtained by showing that conditions (i)-(iv) above hold a.e., replacing E_* by $E_*^{U_{n-1}}$ and following the same steps. ■

Theorem 3.3

For any $t, r \in S$ and for each $n \geq 0$ the sequences $(|\bar{I}_n(*;t)|)$ and $(|I_n(*;t)|)$ satisfy the properties (i) - (iii) above. They satisfy (iv) if and only if $f_t(z_n|z^{n-1}) = f_*(z_n|z^{n-1})$ a.e. implies $t = *$.

Proof

The proof of properties (i)-(iii) is obtained precisely as in the proof of theorem 3.2. (iv) is satisfied if and only if $f_t(z_n|z^{n-1}) = f_*(z_n|z^{n-1})$ a.e. implies $t = *$ by theorem 3.1. ■

The variables $|\bar{I}_n(*;t)|$ and $|I_n(*;t)|$; $t \in S$ are then distance measures from the true parameter * to points in the parameter set S. They can be regarded as extensions of the metrics $|\bar{I}_n(s;t)|$ and $|I_n(s;t)|$ on S to the set $T = (* \cup S)$.

Corollary 3.2

Let $s, t \in S$ be any pair of parameters in the parameter space S. Then s is closer to the true parameter * than t in the metric $|I_n(s;t)|$ if and only if $I_n(s;t) > 0$ a.e. and in the metric $|\bar{I}_n(s;t)|$ if and

only if $\bar{I}_n(s;t) > 0$.

Proof

s is closer to the true parameter than t in the metric $|I_n(s;t)|$ if and only if

$$|I_n(*;s)| < |I_n(*;t)| \quad \text{a.e.}$$

But by theorem 3.1

$$|I_n(*;s)| = I_n(*;s) \quad \text{a.e. for any } s \in S$$

Hence, s is closer to the true parameter than t if and only if

$$I_n(*;s) < I_n(*;t) \quad \text{a.e.}$$

or

$$I_n(*;t) - I_n(*;s) = I_n(s;t) > 0 \quad \text{a.e.}$$

To show that s is closer to the parameter than t in the metric $|\bar{I}_n(s;t)|$ an identical procedure can be followed using $\bar{I}_n(s;t)$ instead of $I_n(s;t)$. ■

Example 3.1

Let x be a random variable, whose probability density is known to belong to the set

$$f_i(x) = \frac{1}{\sqrt{2\pi}\,\sigma_i^2}\, e^{-\frac{x^2}{2\sigma_i^2}} \; ; \; i = 0,1,2 \qquad (3.9)$$

Suppose that i=0 is the true parameter, i.e., that x is actually distributed according to $f_o(x)$. The mean information in a single observation x favoring one parameter against the other is found to be

$$\bar{I}(1;0) = I(1:0) = \frac{1}{2}\log\frac{\sigma_0^2}{\sigma_1^2} + \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma_1^2}\right)$$

$$\bar{I}(2;0) = I(2:0) = \frac{1}{2}\log\frac{\sigma_0^2}{\sigma_2^2} + \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma_2^2}\right)$$

$$\bar{I}(1;2) = I(1:2) = \frac{1}{2}\log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_0^2}{2}\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)$$

Note that $I(i;j) \to 0$ as $\sigma_i \to \sigma_j$

Theorem 3.1 is verified as follows

$$I(1;0) \leq \frac{1}{2}\left[\log\frac{\sigma_0^2}{\sigma_1^2} + \log\frac{\sigma_1^2}{\sigma_0^2}\right] = 0$$

where we have used the inequality $1 - a \leq -\log a$.

Similarly

$$I(2;0) \leq 0$$

To verify corollary 3.1 we check whether

$$I(2;1) \geq I(2;0)$$

but

$$I(2;1) - I(2;0) = I(0;1) = -I(1;0) \geq 0$$

Similarly

$$I(1;2) \geq I(1;0).$$

Next, we check the conditions under which the parameter 1 is closer to the parameter 0 than the parameter 2, in the metric senses defined by theorem 3.2. By corollary 3.2 it suffices to have

$$I(1;2) > 0$$

i.e.

$$\log \frac{\sigma_2^2}{\sigma_1^2} + \sigma_0\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) > 0. \tag{3.10}$$

(3.10) relates the relative closeness of the parameters 1 and 2 to the true parameter 0 (see corollary 3.2) with the covariances associated with the parameters. It is interesting to check, then, whether the closeness of the covariances implies closeness of the parameters in the information metric sense, i.e. whether

$$|\sigma_1^2 - \sigma_0^2| < |\sigma_2^2 - \sigma_0^2| \tag{3.11}$$

implies that the parameter 1 is closer than the parameter 2 to the true parameter 0, i.e. that

$$I(1;2) > 0.$$

In general, (3.11) does not imply (3.10), which depends on the numerical values of σ_0, σ_1 and σ_2. However, (3.11) does imply (3.10) in two cases, namely:

Case 1: $\sigma_0 < \sigma_1 < \sigma_2$

Clearly, (3.11) is satisfied. Using the inequality $\log a < a - 1$ for $a \neq 1$ we have

$$\log \frac{\sigma_1^2}{\sigma_2^2} < \frac{\sigma_1^2}{\sigma_2^2} - 1 \tag{3.12}$$

or

$$\log \frac{\sigma_2^2}{\sigma_1^2} > 1 - \frac{\sigma_1^2}{\sigma_2^2} > 0 \tag{3.13}$$

and since $\frac{\sigma_0^2}{\sigma_1^2} < 1$ we further have

$$\log \frac{\sigma_2^2}{\sigma_1^2} > \frac{\sigma_0^2}{\sigma_1^2}\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)$$

Hence

$$I(1;2) = \log \frac{\sigma_2^2}{\sigma_1^2} + \sigma_0^2\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) > 0$$

Case 2: $\sigma_2 < \sigma_1 < \sigma_0$

(3.11) is again satisfied. By (3.12) we have

$$\log \frac{\sigma_2^2}{\sigma_1^2} > 1 - \frac{\sigma_1^2}{\sigma_2^2} < 0$$

and since $\frac{\sigma_0^2}{\sigma_1^2} > 1$ we further have

$$\log \frac{\sigma_2^2}{\sigma_1^2} > \frac{\sigma_0^2}{\sigma_1^2}\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)$$

Hence

$$I(1;2) = \frac{1}{2}\left[\log \frac{\sigma_2^2}{\sigma_1^2} + \sigma_0\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)\right] > 0.$$

However, if we have

Case 3: $\sigma_1 < \sigma_0 < \sigma_2$

or

Case 4: $\sigma_2 < \sigma_0 < \sigma_1$

Then $|\sigma_i - \sigma_0| < |\sigma_j - \sigma_0|$; $i, j = 1, 2$; $i \neq j$ does not necessarily imply $I(i;j) > 0$ or $\bar{I}(i;j) > 0$. For instance, let

$$\sigma_0^2 = 2$$

and

$$\sigma_2^2 = 1$$

Then if

$$\sigma_1^2 = 2.73$$

we have

$$|\sigma_1^2 - \sigma_0^2| < |\sigma_2^2 - \sigma_0^2|$$

Then

$$\bar{I}(1;2) = I(1;2) = \frac{1}{2}\left[-1 + 2\left(1 - \frac{1}{2.73}\right)\right] = 0.134 > 0$$

But if

$$\sigma_1^2 = 4$$

then we have

$$|\sigma_1^2 - \sigma_0^2| > |\sigma_2^2 - \sigma_0^2|$$

But

$$\bar{I}(2;1) = I(2;1) = \frac{1}{2}\left[1.386 + 2\left(\frac{1}{4} - 1\right)\right] = -0.057 < 0.$$

Hence closeness of the covariances to the true covariance does not imply closeness of the parameters to the true parameter in the metrics $|I(\cdot;\cdot)|$ and $|\bar{I}(\cdot;\cdot)|$ in general, except for cases 1 and 2 above. ▲

We shall use the notation

$$\delta_n(s;t) \equiv |I_n(s;t)|$$

and

$$d_n(s;t) \equiv |\bar{I}_n(s;t)|$$

Then we have sequences of metric spaces

$$(S, \delta_n) \; ; \; (S, d_n)$$

where S is the parameter set. Note that while $I_n(s;t)$ and $\delta_n(s;t)$ are U_{n-1}-measurable random variables, $\bar{I}_n(s;t)$ and $d_n(s;t)$ are not random variables. We shall see that $I_n(s;t)$ and $\delta_n(s;t)$ are useful for purposes of

analysis. The metric $d_n(s;t)$ will prove particularly useful when it is constant in time, as will prove to be the case for ergodic observation sequences. The parameter metric space can then be denoted (S, d).

3.3 Comparison with Other Information Measures

Attempts by statisticians and engineers to assign quantitative measures to the intuitive notion of information have resulted over the years in many different definitions of information. Information measures can, in essence, be classified in two different categories. One is characterized by the Shannon entropy, which has proved useful in communication and source-coding theory, sometimes termed information theory. The other is characterized by Fisher's and Kullback's information measures, which have been more popular in statistical circles. Our information measures fall in the second category. It seems that different permutations of Fisher's or Kullback's information measures result from different interpretations of a given set of data, which in turn reflect the intended application. Our version of information seems to be the most general, since, unlike other definitions, it does not assume that the true parameter belongs to the parameter set under consideration. However, special care must be taken in evaluating the advantages of one definition of information over another.

The information measures defined in this chapter prove very useful in the analysis of the asymptotic behavior of parameter estimates. They provide insight into the convergence of the estimates in the presence and in the absence of the true parameter. However, they can only be computed

if the true parameter is known. Nevertheless their application is not limited to analysis, as will be evident in Chapter 5 where we consider several model selection problems. On the other hand, several other information measures which are useful in given applications, such as signal detection, do not possess properties which are useful for analytical purposes, such as the metric property. In the rest of this section we briefly discuss a few information measures, common in the information theoretic and the statistical literature and relate them to the information measures defined in section 3.1.

3.3.1 Kullback's Information, the Divergence, the Bhattacharyya Distance and the Ambiguity Function

Kullback [1959] defined the mean information for discriminating in favor of one hypothesis H_1 against another, H_2, given an observation x as

$$I^k(1;2) = \int \log \frac{f_1(x)}{f_2(x)} \, d\mu_1(x)$$

where μ_1 is a probability measure corresponding to H_1. f_1 is the density of μ_1 with respect to some measure λ and f_2 is the density with respect to λ of μ_2, a probability measure corresponding to H_2. The divergence between H_1 and H_2, first introduced by Jeffreys [1946] and employed by Kullback [1959] is defined as:

$$\begin{aligned} J(1;2) &= I^k(1;2) + I^k(2;1) \\ &= \int [f_1(x) - f_2(x)] \log \frac{f_1(x)}{f_2(x)} \, d\lambda(x) \\ &= \int \log \frac{f_1(x)}{f_2(x)} \, d\mu_1(x) - \int \log \frac{f_1(x)}{f_2(x)} \, d\mu_2(x) \end{aligned}$$

In contrast, I(1;2), defined by (3.3) would be written as

$$I(1;2) = \int \log \frac{f_1(x)}{f_2(x)} \, d\mu(x)$$

where $\mu(x)$, the correct probability measure may be different from both $\mu_1(x)$ and $\mu_2(x)$.

The Bhattacharyya distance (Bhattacharyya [1943]) between two densities $f_1(x)$ and $f_2(x)$ of an observation x

$$B = - \ln \int [f_1(x) \; f_2(x)]^{\frac{1}{2}} \, d\lambda$$

where λ is the Lebesgue measure on the space of x. Properties of the Bhattacharyya distance and the divergence were studied and compared by Kailath [1967], and they were found to be particularly suitable for signal detection in communication. However, Kullback's information, the divergence and the Bhattacharyya distance do not satisfy the triangle inequality and thus fail as metrics on the parameter (or hypothesis) space. In contrast, the metric property of the information measures introduced in section 3.1 follows from the consistent use of the true probability measure throughout, whereas Kullback's information, the divergence and the Bhattacharyya distance are defined using different measures.

The ambiguity in an observation x between a parameter s and the true parameter * is defined as

$$\gamma_s = E_* \log f_s(x)$$

The ambiguity function γ_s has been found useful in the analysis of error in radar applications (Woodward [1953]). In fact

$$I(s;t) = \gamma_t - \gamma_s$$

Hence, the information between two parameters as defined in this thesis is the difference between their ambiguities.

3.3.2 Fisher's Information

We shall now show that the information measures introduced in section 3.1 are related to Fisher's information measure (Fisher [1956], Savage [1954]). We follow a similar comparison between Kullback's and Fisher's information measures (Kullback [1959]). However, in order to relate measures of the same quantity, we define Fisher's information in a single observation z_n.

Let $S \in R^k$ be the parameter space. Suppose that for any $s \in S$ the following regularity conditions (Cramer [1946], Gurland [1954], hold for all $i, j = 1,\ldots,k$

1) $$\left|\frac{\partial \log f_s(z_n|z^{n-1})}{\partial s^i}\right| < F_1(z^n); \quad \left|\frac{\partial^2 \log f_s(z_n|z^{n-1})}{\partial s^i\, \partial s^j}\right| < F_2(z^n) \quad \text{a.e.}$$

where the partial derivatives are assumed to exist and $F_1(z^n)$ and $F_2(z^n)$ are integrable random variables.

2) $$\int \frac{\partial f_s(z_n|z^{n-1})}{\partial s^i}\, dP_* = 0; \quad \int \frac{\partial^2 f_s(z_n|z^{n-1})}{\partial s^i\, \partial s^j}\, dP_* = 0$$

We define Fisher's information in a single observation at a parameter point s as a matrix $\overline{I}_n^F(s)$, whose elements are

$$\overline{I}^F_{i,j,n}(s) \equiv E_s\left\{\left(\frac{1}{f_s(z_n|z^{n-1})}\frac{\partial f_s(z_n|z^{n-1})}{\partial s^i}\right)\left(\frac{1}{f_s(z_n|z^{n-1})}\frac{\partial f_s(z_n|z^{n-1})}{\partial s^j}\right)\right\} \tag{3.9}$$

Consider a point $s \in S$ and a close point $s + \Delta s \in S$. Using Taylor's expansion to second order we have

$$\log f_{s+\Delta s}(z_n|z^{n-1}) - \log f_s(z_n|z^{n-1}) = \sum_{i=1}^{k} \Delta s^i \frac{\partial \log f_s(z_n|z^{n-1})}{\partial s^i} + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} \Delta s^i \Delta s^j \frac{\partial^2 \log f_s(z_n|z^{n-1})}{\partial s^i \partial s^j}$$

But

$$\frac{\partial \log f_s(z_n|z^{n-1})}{\partial s^i} = \frac{1}{f_s(z_n|z^{n-1})}\frac{\partial f_s(z_n|z^{n-1})}{\partial s^i}\ ; \qquad \frac{\partial^2 \log f_s(z_n|z^{n-1})}{\partial s^i \partial s^j}$$

$$= -\frac{1}{f_s(z_n|z^{n-1})}\frac{\partial^2 f_s(z_n|z^{n-1})}{\partial s^i \partial s^j} - \frac{1}{f_s^2(z_n|z^{n-1})}\frac{\partial f_s(z_n|z^{n-1})}{\partial s^i}\frac{\partial f_s(z_n|z^{n-1})}{\partial s^j}$$

The information in z_n favoring s against a close point $s + \Delta s$ as defined by (3.4) is

$$\overline{I}_n(s;s+\Delta s) = \int \log \frac{f_s(z_n|z^{n-1})}{f_{s+\Delta s}(z_n|z^{n-1})}\, dP_*$$

$$= -\int \sum_{i=1}^{k} \Delta s^i \frac{\partial \log f_s(z_n|z^{n-1})}{\partial s^i}\, dP_*$$

$$- \frac{1}{2}\int \sum_{i=1}^{k}\sum_{j=1}^{k} \Delta s^i \, \Delta s^j \, \frac{\partial^2 \log f_s(z_n|z^{n-1})}{\partial s^i \, \partial s^j} \, dP_*$$

$$= - \sum_{i=1}^{k} \Delta s^i \int \frac{\partial f_s(z_n|z^{n-1})}{\partial s^i} \, dP_*$$

$$- \frac{1}{2} \sum_{i=1}^{k}\sum_{j=1}^{k} \Delta s^i \, \Delta s^j \int \left[\frac{\partial^2 f_s(z_n|z^{n-1})}{\partial s^i \, \partial s^j} \right.$$

$$\left. - \frac{1}{f_s(z_n|z^{n-1})} \frac{\partial f_s(z_n|z^{n-1})}{\partial s^i} \frac{\partial f_s(z_n|z^{n-1})}{\partial s^j} \right] dP_*$$

$$= \frac{1}{2} \sum_{i=1}^{k}\sum_{j=1}^{k} \Delta s^i \, \Delta s^j \, \bar{I}^F_{i,j,n}(s)$$

where the last equality is obtained by the regularity condition 2) above. Hence, the information in a single observation is related to Fisher's information in a single observation by

$$\bar{I}_n(s; s + \Delta s) = \frac{1}{2} \Delta s^T \, \bar{I}^F_n(s) \, \Delta s$$

Defining similarly the conditional Fisher Information in a single observation as a matrix $I^F_n(s)$ whose elements are

$$I^F_{i,j,n}(s) \equiv E^{U_{n-1}}_* \left\{ \left(\frac{1}{f_s(z_n|z^{n-1})} \frac{\partial f_s(z_n|z^{n-1})}{\partial s^i} \right) \left(\frac{1}{f_s(z_n|z^{n-1})} \frac{\partial f_s(z_n|z^{n-1})}{\partial s^i} \right) \right\} .$$

We get, using a similar procedure

$$I_n(s;s + \Delta s) = \frac{1}{2} \Delta s^T I_n^F(s) \Delta s$$

3.3.3 Self Information and Entropy

To complete this discussion, placing the information measures motivated and defined in this chapter in perspective with respect to other measures found in the literature, we mention two other measures which are quite common in information theory, namely, the self information and the entropy (e.g. Fano [1961] and Gallager [1968]). The definition of these measures is based on the Bayesian assumption (see Chapter 2).

Consider a parameter set S. The self information in the measurements z^n about a parameter $s \in S$ is defined as

$$I_n^S(s) \equiv - \log f^b(s|z^n) \tag{3.14}$$

A comparative measure of information can then be obtained by taking the difference of the self information corresponding to two parameters $s, t \in S$

$$\Delta I_n^S(s;t) \equiv I_n^S(s) - I^S(t) = - \log \frac{f^b(s|z^n)}{f^b(t|z^n)}$$

The self information difference between s and t in a single observation z_n can be obtained, using (2.6) and (2.19) as

$$\Delta I_n^S(s;t) - \Delta I_{n-1}^S(s;t) = -\log \frac{f_s(z_n|z^{n-1})}{f_t(z_n|z^{n-1})} = -\log h_t^s(z_n|z^{n-1}) \quad (3.15)$$

Taking expectation and conditional expectation of (3.15) with respect to the true measure one gets

$$E_* \left\{ \Delta I_n^S(s;t) - \Delta I_{n-1}^S(s;t) \right\} = -\bar{I}_n(s;t) \quad (3.16)$$

and

$$E_*^{U_{n-1}} \left\{ \Delta I_n^S(s;t) - \Delta I_{n-1}^S(s;t) \right\} = -I_n(s;t) \quad (3.17)$$

Hence, the mean and the conditional mean values of the self information difference in a single observation are the negative values of the information measures defined in section 3.1. (The sign is, of course, of no significance since the self information defined by (3.14) is in fact lack of information, and would become positive information, in the sense meant in this chapter, by inverting the sign.)

Note that in (3.17) the expectation is taken with respect to the correct probability measure P_*, independently of whether the correct parameter even belongs to the set S. If, on the other hand one makes the assumption that the true parameter belongs to a finite set, say $\{s_j; j \in K \equiv (0,\ldots,p)\}$, and takes a conditional expectation given Z^n of (3.14), then one gets

$$E^{b\,U_n} I_n^S(s) = -\sum_{j=0}^{p} f^b(s_j|Z^n) \log f^b(s_j|Z^n) \equiv H(Z^n) \quad (3.18)$$

(3.18) is the entropy in Z^n. Note that the entropy differs from the other information measures considered in this section in the sense that it is not a comparative measure between parameters. It does provide some measure of the average information gained from the observations, with respect to the *a priori* assumptions (Berger [1971]).

CHAPTER IV

CONVERGENCE OF MAXIMUM LIKELIHOOD AND BAYESIAN ESTIMATES ON FINITE SETS OF PARAMETERS

In this chapter we study the convergence of maximum likelihood and Bayesian parameter estimates for general classes of observation sequences. The convergence of the estimates follows from the convergence of the likelihood ratios over the parameter set. Consistency conditions are derived in terms of the information in the observations. The case where the true parameter is not necessarily a member of the parameter set is also considered. Rates of convergence in the mean for the ML and the MAP procedures are derived.

4.1 Convergence of Parameter Estimates

Let (z_n) be a stochastic process on a probability space (Ω, U, P_k) and let $S = K \equiv \{0,\ldots,p\}$ be a parameter set such that $\{P_j;\ j \in K\}$ is a family of probability measures on (Ω, U). Let (U_n) be an increasing sequence of σ-subalgebras of U generated by (Z^n) and let $P_{j,n}$ be the restriction of P_j to U_n for each $j \in K$. Consider the following condition:

(c4.1) For some $k \in K$ and for each $j \in K;\ j \neq k$

$$\lim_{n\to\infty} h_k^j (Z^n) = 0 \quad \text{a.e.} \tag{4.1}$$

In the sequel we show that the convergence a.e. of the ML and the MAP parameter estimation procedures and the convergence a.e. and in m.s. of the LS procedure follow from condition (c4.1). Of course, the major difficulty in proving convergence of the parameter estimates is to verify condition (c4.1). In the following sections we give conditions for general classes of observation sequences under which condition (c4.1) is satisfied when k is the true parameter and extend the results to the case where the true parameter is not necessarily a member of the parameter set. The latter case is treated specifically in the following chapter where the following theorems will prove very useful.

Theorem 4.1

Suppose that (c4.1) is satisfied, then ML estimates on K converge a.e. to k as $n \to \infty$.

Proof

Since the set $j \in K ; j \neq k$ is finite, (c4.1) implies

$$\lim_{n\to\infty} \sup_{j} \left\{ h_k^j (Z^n); \; j \in K ; \; j \neq k \right\} = 0 \quad \text{a.e.}$$

Hence

$$\lim_{n\to\infty} \sup_{j} \left\{ h_k^j (Z^n); \; j \in K \right\} = h_k^k (Z^n) = 1 \quad \text{a.e.}$$

or

$$\lim_{n\to\infty} \hat{k}(Z^n) = k \quad \text{a.e.} \blacksquare$$

Theorem 4.2

Under condition (c4.1) MAP estimates on K converge a.e. to k.

Proof

By (2.19) we have for each $j \in K$

$$f^b(j|z^n) \leq \frac{f_o^b(j)\ f_j(z^n)}{f_o^b(k)\ f_k(z^n)} = \frac{f_o^b(j)}{f_o^b(k)}\ h_k^j\ (z^n)$$

By (c4.1) for each $j \in K$; $j \neq k$ we have

$$\lim_{n\to\infty} f^b(j|z^n) \leq \frac{f_o^b(j)}{f_o^b(k)} \lim_{n\to\infty} h_k^j\ (z^n) = 0 \quad \text{a.e.}$$

implying

$$\lim_{n\to\infty} f^b(j|z^n) = 0 \quad \text{a.e.} \quad \text{for each } j \in K\ ;\ j \neq k \tag{4.2a}$$

But since

$$\sum_{j=o}^{p} f^b(j|z^n) = 1$$

we have

$$\lim_{n\to\infty} f^b(k|z^n) = 1 \quad \text{a.e.} \tag{4.2b}$$

yielding the assertion. ■

Theorem 4.3

Suppose that a parameter vector s is assumed to belong to a finite

set $s_j \in R^m$; $j \in K$ in the calculation of the estimates (but is not necessarily a member of the set). Suppose further that for some $k \in K$ condition (c4.1) is satisfied. Then LS estimates of s on R^m converge a.e. to s_k.

Proof

By (2.20) and (4.2) we have

$$\lim_{n\to\infty} \hat{s}_n = \sum_{j=o}^{p} s_j \lim_{n\to\infty} f^b(j|z^n) = s_k \quad \text{a.e.} \blacksquare$$

Theorem 4.4

For the situation given in theorem 4.3 LS estimates converge to s_k in the mean-square.

Proof

We follow in part Liporace [1971] who treated the case of independent and identically distributed observations. Consider the norm

$$N_n = E_*\left\{(\hat{s}_n - s_k)^T(\hat{s}_n - s_k)\right\}$$

$$= E_*\left\{\sum_{j=o}^{p}(s_j - s_k)^T f^b(s_j|z^n) \sum_{i=o}^{p}(s_i - s_k) \quad f^b(s_i|z^n)\right\}$$

$$= \sum_{j=o}^{p}\sum_{i=o}^{p} (s_j - s_k)^T (s_i - s_k)\ E_*\left\{f^b(s_j|z^n)\ f^b(s_i|z^n)\right\}$$

$$\leq p^2R^2\ E_*\ f^b(s_j|z^n) \qquad \text{for some } j \in K\ ;\ j \neq k$$

where

$$R^2 = \max \left\{ (s_j - s_k)^T (s_j - s_k) \; ; \; j \in K \; ; \; j \neq k \right\}$$

since obviously

$$E_* \left\{ f^b(s_j | Z^n) \right\} \geq E_* \left\{ f(s_j | Z^n) \; f(s_i | Z^n) \right\}$$

(because $f(s_i | Z^n) \leq 1$).

By (2.19) we have for each $j \in K$

$$f^b(s_j | Z^n) \leq \frac{f_o^b(s_j) \; f_j(Z^n)}{f_o^b(s_k) \; f_k(Z^n)} = \frac{f_o^b(s_j)}{f_o^b(s_k)} h_k^j(Z^n)$$

By (c4.1) we have for each $j \in K \; ; \; j \neq k$

$$\lim_{n \to \infty} f^b(s_j | Z^n) \leq \frac{f_o^b(s_j)}{f_o^b(s_k)} \lim_{n \to \infty} h_k^j(Z^n) = 0 \quad \text{a.e.}$$

Hence

$$\lim_{n \to \infty} f^b(s_j | Z^n) = 0 \quad \text{a.e.}$$

Now since

$$f^b(s_j | Z^n) \leq 1$$

we have by the dominated convergence theorem (theorem 2.3) that for each $j \in K \; ; \; j \neq k$

$$\lim_{n \to \infty} E_* f^b(s_j | Z^n) = E_* \lim_{n \to \infty} f^b(s_j | Z^n) = 0$$

and thus

$$\lim_{n\to\infty} N_n \leq p^2R^2 \lim_{n\to\infty} E_* f^b(s_j|Z^{n-1}) = 0$$

yielding finally

$$\lim_{n\to\infty} N_n = 0. \blacksquare$$

4.2 Consistency of the Estimates

In Chapter 3 we defined for each pair k, j ∈ K

$$I_n(k;j) \equiv E_*^{U_{n-1}} \log h_j^k (z_n|Z^{n-1})$$

Let us also define

$$J_n(k;j) \equiv \log h_j^k (z_n|Z^{n-1}) - I_n(k;j)$$

$J_n(k;j)$ is the error in the incremental information $I_n(k;j)$, or the information residual. Denote

$$Y_n(k;j) \equiv \sum_{m=1}^{n} I_m(k;j)$$

and

$$V_n(k;j) \equiv \sum_{m=1}^{n} J_m(k;j)$$

Note that for each j, k ∈ K $(J_n(k;j))$ is a U_n-martingale difference sequence according to the true measure P_*, and consequently $(V_n(k;j))$

is a U_n-martingale sequence according to P_*.

Suppose that * ∈ K, i.e. that the true parameter is a member of the parameter set and consider the following conditions:

(c4.2) For some k ∈ K and for each j ∈ K ; j ≠ k

$$\limsup V_n(k;j) > -\infty \quad \text{a.e.}$$

(c4.3) For some k ∈ K and for each j ∈ K ; j ≠ k

$$\lim_{n\to\infty} Y_n(k;j) = \infty \quad \text{a.e.}$$

Lemma 4.1

Suppose that conditions (c4.2) and (c4.3) hold for k = *. Then for each j ∈ K ; j ≠ * one has

$$\lim_{n\to\infty} h_*^j (Z^n) = 0 \quad \text{a.e.}$$

Proof

We have noted (see section 2.4) that for each j ∈ K the sequence $(h_*^j (Z^n))$ is a U_n-martingale according to the measure P_*. Furthermore

$$E_* h_*^j (Z^n) = 1$$

It follows from the martingale convergence theorem (theorem 2.4) that the sequence $(h_*^j (Z^n))$ converges to a finite limit. Thus, the sequence $(\log h_*^j (Z^n))$ converges to some $a < \infty$. We have

$$\log h_j^*(Z^n) = Y_n(*;j) + V_n(*;j) \tag{4.3}$$

Suppose that

$$\lim_{n\to\infty} \log h_*^j(Z^n) = a > -\infty \quad \text{a.e.}$$

or

$$\lim_{n\to\infty} \log h_j^*(Z^n) < \infty \quad \text{a.e.}$$

Then by condition (c4.3) and by (4.3) we have

$$\lim_{n\to\infty} V_n(*;j) = -\infty \quad \text{a.e.}$$

contradicting condition (c4.2). Hence, we have

$$\lim_{n\to\infty} \log h_j^*(Z^n) = \infty \quad \text{a.e.}$$

or

$$\lim_{n\to\infty} \log h_*^j(Z^n) = -\infty \quad \text{a.e.}$$

yielding

$$\lim_{n\to\infty} h_*^j(Z^n) = 0 \quad \text{a.e.} \blacksquare$$

Theorem 4.5

Suppose that some $k \in K$ is the true parameter. Then under conditions (c4.2) and (c4.3) ML and MAP estimates are consistent a.e. and LS estimates are consistent a.e. and in the mean square.

Proof

The assertion follows directly from lemma 4.1 and from theorems 4.1 through 4.4.■

Consider the following condition

(c4.4) For some $k \in K$ and for each $j \in K$; $j \neq k$ there exists some $\varepsilon_j > 0$ and a subsequence (n_{i_j}) of n such that

$$I_{n_{i_j}}(k;j) \geq \varepsilon_j \text{ a.e. for all } n_{i_j}$$

Theorem 4.6

Suppose that some $k \in K$ is the true parameter. Then under condition (c4.2) and (c4.4) ML and MAP estimates are consistent a.e. and LS estimates are consistent a.e. and in the mean square.

Proof

By theorem 3.1 we have

$$I_n(k;j) \geq 0 \quad \text{a.e.} \quad \text{for all } n \geq 0$$

Thus, condition (c4.4) implies condition (c4.3). The assertion follows from theorem 4.5.■

In the following chapters we shall see certain important cases to which the information condition (c4.3) applies. We now examine condition (c4.2). We have noted that for each pair $j, k \in K$ the sequence

$(J_n(k;j))$ is a martingale difference sequence according to the true measure P_*. The following special case is of particular interest.

Lemma 4.2

For any pair $j, k \in K$ let $(J_n(k;j))$ be an ergodic sequence. Then

$$\limsup V_n(k;j) = \infty \quad \text{a.e.}$$

Proof

We have by (2.21) for each $\omega \in \Omega$

$$V_n(k;j,\omega) = J_1(k;j,\omega) + V_{n-1}(k;j, T\,\omega)$$

where T is a measure preserving transformation. It follows that the event

$$\left\{ \lim_{n\to\infty} \sup V_n(k;j) < \infty \right\}$$

is invariant. Thus, either

$$P\left\{\limsup V_n(k;j) < \infty\right\} = 0$$

or

$$P\left\{\limsup V_n(k;j) < \infty\right\} = 1$$

Obviously, we have that if

$$\limsup V_n(k;j) < \infty$$

then

$$\limsup \frac{V_n(k;j)}{\sqrt{n}} < \infty$$

But by theorem 2.6

$$P\left\{\limsup \frac{V_n(k;j)}{\sqrt{n}} < \infty\right\} < 1$$

Hence

$$P\left\{\limsup V_n(k;j) < \infty\right\} < 1$$

yielding

$$P\left\{\limsup V_n(k;j) < \infty\right\} = 0$$

Thus

$$\limsup V_n(k;j) = \infty \quad \text{a.e.} \blacksquare$$

Example 4.1

Let (x_n) be a sequence of independent identically distributed observations. Suppose that each x_n is distributed according to the density

$$f(x_n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_n^2}{2\sigma^2}}$$

Let the covariance σ^2 be given on a set $\{\sigma_i, i=1,2\}$, and suppose that $\sigma^2 = \sigma_1^2$, i.e. that 1 is the true parameter. As in example 3.1 we have for all $n \geq 0$

$$I_n(1;2) = \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{2}\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)$$

and

$$J_n(1;2) = \frac{1}{2} x_n^2 \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2} \right) - \frac{\sigma_1^2}{2} \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2} \right)$$

Since (x_n) is an ergodic sequence, so is $(J_n(1;2)$. Thus, by lemma 4.2 condition (c4.2) is satisfied. It follows from (3.5) that if $\sigma_1 \neq \sigma_2$ then $I_n(1;2) \neq 0$ and then, by theorem 3.1 we have

$$I_n(1;2) = I(1;2) > 0 \quad \text{for all } n \geq 0$$

Thus, condition (c4.3) is satisfied for k = 1. Hence, by theorems 4.1 through 4.4, the ML and the MAP estimates of σ will converge a.e. and the LS estimates will converge a.e. and in the mean-square to σ_1. ▲

The following general result provides a sufficient condition satisfying condition (c4.2). Although it will not be used directly in the following chapters, it seems to have useful implications (see example 4.2).

Lemma 4.3

Suppose that for any j, k ∈ K we have for any positive scalar a

$$E_* \left\{ V_{n_a}(k;j) - a \right\} < \infty \tag{4.4}$$

where

$$n_a = \inf \left\{ n : V_n(k;j) > a \right\} \tag{4.5}$$

then for each $\omega \in \Omega$ either

$$\lim_{n\to\infty} V_n(k;j,\omega) \text{ exists and is finite}$$

or

$$\limsup V_n(k;j,\omega) = \infty$$

Proof

Let

$$R_a(k;j) \equiv V_{n_a}(k;j) - a$$

and

$$V_n^a(k;j) \equiv V(k;j)_{\min(n,n_a)}$$

Note that

$$V_n^a(k;j) \leq a + R_a(k;j)$$

Since $Y_n(k;j)$ is a U_n-martingale, so is $(Y_n^a(k;j))$. Obviously, we have

$$E_*\left\{V_n^{a^+}(k;j)\right\} \leq a + E_* R_a(k;j)$$

Hence, under (4.4)

$$E_*\left\{V_n^{a^+}(k;j)\right\} \leq \infty$$

It follows from theorem 2.4 that the sequence $(V_n^a(k;j))$ converges to a

finite limit. Let

$$A_a \equiv \left\{ \omega \in \Omega : \sup V_n(k;j,\omega) < a \right\}$$

and

$$A \equiv \bigcup_{a=1}^{\infty} A_a$$

If $\omega \in A$ then $\omega \in A_a$ for some a, say, a_o. Then,

$$V_n(k;j,\omega) = V_n^{a_o}(k;j,\omega) \text{ for all n}$$

and then

$$\lim_{n\to\infty} V_n(k;j,\omega) = \lim_{n\to\infty} V_n^{a_o}(k;j,\omega) \text{ is finite.}$$

If $\omega \notin A$, then

$$\lim \sup V_n(k;j,\omega) = \infty. \blacksquare$$

Example 4.2

Let (x_n) be a sequence of real valued random variables taking values in the interval [0, 3]. Suppose that the sequence (x_n) is not necassarily independent or identically distributed. Consider two hypotheses (or two parameters) 1 and 2, according to which (x_n) is i.i.d. with probability densities

$$f_1(x_n) = \begin{cases} \frac{1}{2} & 0 \leq x_n \leq 1 \\ \frac{1}{4} & 1 \leq x_n \leq 3 \\ 0 & \text{elsewhere} \end{cases}$$

and

$$f_2(x_n) = \begin{cases} \frac{1}{4} & 0 \leq x_n \leq 2 \\ \frac{1}{2} & 2 \leq x_n \leq 3 \\ 0 & \text{elsewhere} \end{cases}$$

It is easy to see that

$$J_n(1;2) \leq 2 \log 2 \quad \text{for all } n$$

independently of the actual values the sequence (x_n) might take in the interval [0, 3]. Now since for any $a > 0$

$$V_{n_a}(1;2) = V_{n_a-1} + J_{n_a}(1;2) \leq a + J_{n_a}(1;2)$$

we have

$$E_*\left\{V_{n_a}(1;2) - a\right\} \leq E_* J_{n_a}(1;2) \leq 2 \log 2$$

for all n. Hence (4.4) holds. It follows from lemma 4.2 that condition (c4.2) is satisfied for this case independently of the actual probability measure generating the sequence (x_n).

4.3 Convergence in the Absence of the True Parameter

Consider the probability and parameter spaces given in section 4.1. While the absolute continuity of the restrictions $P_{1,n}$ and $P_{2,n}$ of two measures P_1 and P_2 to the σ-subalgebra U_n of U is possible to verify in practical situations (it follows e.g. from the absolute continuity of

the corresponding conditional densities $f_{1,n}(z_n|z^{n-1})$ and $f_{2,n}(z_n|z^{n-1})$ for each n), the absolute continuity of P_1 and P_2 does not follow and is, in general, more difficult to verify. The following results are nevertheless interesting from a theoretical viewpoint.

Theorem 4.8

Let conditions (c4.2) and (c4.3) hold for some parameter $k \in K$. Furthermore, suppose that the true measure P_* is absolutely continuous with respect to the measure P_k. Then for each $j \in K$; $j \neq k$ one has

$$\lim_{n\to\infty} h_k^j (z^n) = 0$$

and, consequently, the parameter estimates will converge to the parameter k in the senses specified in theorems 4.1 through 4.4.

Proof

Since the sequence $(h_k^j (z^n))$ is a (U_n, P_k)-martingale and since

$$E_k h_k^j (z^n) = 1$$

it follows from theorem 2.4 that $(h_k^j (z^n))$ convergence a.e. P_k to a finite random variable. Since P_* is absolutely continuous with respect to P_k, then $(h_k^j (z^n))$ converges to a finite random variable a.e. P_*. The remainder of the proof is identical to the proof of lemma 4.1, and the convergence of the estimates follows from the convergence of the likelihood ratios by theorems 4.1 through 4.4. ■

In the following chapter we shall treat the case where the true parameter is not necessarily a member of the parameter set for a case of practical interest, namely, linear dynamical system. We shall not, however, investigate the absolute continuity of the probability measures P_* and P_k, but rather use simpler arguments, enabled by the particular problem under consideration.

Condition (c2.1) requires that for any parameter $k \in K$, the restrictions $P_{j,n}$ of the measures P_j, $j \in K$; $j \neq k$ be absolutely continuous with respect to the restriction $P_{k,n}$ of the measure P_k. An interesting observation is given in the following theorem.

Theorem 4.9

Suppose that condition (c4.1) holds for the parameter $k \in K$. Then the measures P_j, $j \in K$; $j \neq k$ are singular with respect to the measure P_*.

Proof

For each $j \in K$ the likelihood ratio sequence $\left(\frac{f_k(Z^n)}{f_j(Z^n)} \right)$ is a martingale according to the measure P_j (Doob [1953], p. 93). In addition, we have

$$E_j \left\{ \frac{f_k(Z^n)}{f_j(Z^n)} \right\} = 1 \qquad (4.6)$$

By the martingale convergence theorem we then have

$$\lim_{n\to\infty} \frac{f_k(Z^n)}{f_j(Z^n)} = \text{finite r.v.} \qquad \text{a.e. } P_j$$

(where r.v. denotes random variable).

But under condition (c4.1)

$$\lim_{n\to\infty} \frac{f_k(Z^n)}{f_j(Z^n)} = \infty \qquad \text{a.e. } P_*$$

So we have

$$P_j\left\{\lim_{n\to\infty} \frac{f_k(Z^n)}{f_j(Z^n)} = \text{finite r.v.}\right\} = 1$$

and also

$$P_*\left\{\lim_{n\to\infty} \frac{f_k(Z^n)}{f_j(Z^n)} = \text{finite r.v.}\right\} = 0$$

Hence, under condition (c4.1) the measures P_j; $j \in K$; $j \neq k$ are singular with respect to the measure P_*. ■

4.4 L_1 Convergence

The L_1 convergence of the likelihood and *a posteriori* probability ratios follows directly if condition (c4.1) holds. We show that under a certain condition on the information in the observations the convergence rates are bounded by exponentials of the number of samples. The true parameter is not assumed to belong to the parameter set. These results provide performance measures for the ML and MAP estimation

methods. In the following chapters we show that bounds of the L_1 convergence rates can be computed in common situations for linear systems.

Theorem 4.10

Suppose that condition (c4.1) holds for some $k \in K$.

Then for each $j \in K$; $j \neq k$ we have

$$\lim_{n\to\infty} E_* h_j^k (Z^n) = \infty$$

and

$$\lim_{n\to\infty} E_* \frac{f^b(k|Z^n)}{f^b(j|Z^n)} = \infty$$

Proof

We have by (c4.1) for each $j \in K$: $j \neq k$

$$\lim_{n\to\infty} h_j^k (Z^n) = \infty \quad \text{a.e.}$$

and by (4.2)

$$\lim_{n\to\infty} \frac{f^b(k|Z^n)}{f^b(j|Z^n)} = \infty \quad \text{a.e.}$$

Since both sequences are non-negative, we have by Fatou's lemma (theorem 2.2)

$$\lim_{n\to\infty} E_* h_j^k (Z^n) \geq \liminf_{n\to\infty} E_* h_j^k (Z^n) \geq E_* \liminf_{n\to\infty} h_j^k (Z^n) = \infty$$

and

$$\lim_{n\to\infty} E_* \frac{f^b(k|Z^n)}{f^b(j|Z^n)} \geq \liminf_{n\to\infty} E_* \frac{f^b(k|Z^n)}{f^b(j|Z^n)} \geq E_* \liminf_{n\to\infty} \frac{f^b(k|Z^n)}{f^b(j|Z^n)} = \infty$$

Now consider the following condition

(c4.5) There exists a parameter $k \in K$ such that for each $j \in K$; $j \neq k$ there exists a positive scalar α_j and a positive integer N_j such that

$$\bar{I}_n(k;j) \geq \alpha_j \quad \text{for all } n \geq N_j \tag{4.7}$$

Theorem 4.11

Under condition (c4.5) there exists some positive integer N such that for each $j \in K$; $j \neq k$ the sequence $(h_j^k(z^n))$ and $\left(\frac{f^b(k|z^n)}{f^b(j|z^n)}\right)$ diverge in L_1 at rates no slower than exponential for all $n \geq N$.

Proof

$$\bar{I}_n(k;j) = E_* \log \frac{f_k(z_n|z^{n-1})}{f_j(z_n|z^{n-1})}$$

$$= E_* \log \frac{f_k(z^n)}{f_j(z^n)} - E_* \log \frac{f_k(z^{n-1})}{f_j(z^{n-1})}$$

By (c4.5) we then have

$$E_* \log \frac{f_k(z^n)}{f_j(z^n)} - E_* \log \frac{f_k(z^{n-1})}{f_j(z^{n-1})} \geq \alpha_j$$

yielding

$$E_* \log \frac{f_k(z^n)}{f_j(z^n)} \geq a_j + (n - N_j)\alpha_j \quad \text{for all } n \geq N_j \tag{4.8}$$

where

$$a_j \equiv E_* \log \frac{f_k(z_{N_j} | z^{N_j-1})}{f_j(z_{N_j} | z^{N_j-1})} = \bar{I}_{N_j}(k;j) \geq \alpha_j \tag{4.9}$$

Since log (•) is a concave function, we have by Jensen's inequality (theorem 2.1)

$$\log E_* \frac{f_k(z^n)}{f_j(z^n)} \geq E_* \log \frac{f_k(z^n)}{f_j(z^n)} \tag{4.10}$$

(4.8), (4.9) and (4.10) give

$$E_* h_j^k(z^n) = E_* \frac{f_k(z^n)}{f_j(z^n)} \geq e^{a_j} e^{(n - N_j)\alpha_j}$$

$$\geq e^{(n-N_j+1)\alpha_j} \tag{4.11}$$

for all $n \geq N_j$, for each $j \in K$; $j \neq k$

Hence for each $j \in K$; $j \neq k$ the likelihood ratio $h_j^k(z^n)$ tends in the mean to infinity faster then an exponential with a rate of α_j.

By (2.19) we have for each $j \in K$

$$\frac{f^b(k|z^n)}{f^b(j|z^n)} = \frac{f^b(k)}{f^b(j)} \frac{f_k(z^n)}{f_j(z^n)}$$

Thus, by (4.11) for each $j \in K$; $j \neq k$

$$E_* \frac{f^b(k|z^n)}{f^b(j|z^n)} = \frac{f^b(k)}{f^b(j)} E_* \frac{f_k(z^n)}{f_j(z^n)}$$

$$\geq \frac{f^b(k)}{f^b(j)} e^{(n-N_j+1)\alpha_j} \quad \text{for all } n \geq N_j \tag{4.12}$$

Hence, for each $j \in K$; $j \neq k$ the *a posteriori* probability ratio $\frac{f^b(k|z^n)}{f^b(j|z^n)}$ tends in the mean to infinity faster than an exponential with a rate of α_j.

Finally, taking $N = \max \{N_j ; j \in K ; j \neq k\}$ and $\alpha = \min \{\alpha_j ; j \in K ; j \neq k\}$ we have that the sequences $(h_j^k (z^n))$ and $\left(\frac{f^b(k|z^n)}{f^b(j|z^n)}\right)$ converge in L_1 to infinity faster than an exponential with a rate of α for all $n \geq N$. ■

At instant n the ML estimation method will select the parameter k if

$$\frac{f_k(z^n)}{f_j(z^n)} \geq 1 \quad \text{for all } j \in K ; j \neq k \tag{4.13}$$

The MAP method will select k if

$$\frac{f^b(k|z^n)}{f^b(j|z^n)} \geq 1 \quad \text{for all } j \in K ; j \neq k \tag{4.14}$$

Hence, the L_1 convergence bounds established in theorem 4.11 provide a qualitative measure of performance for the ML and the MAP estimates in terms of rates at which (4.13) and (4.14) are attained in the mean. Of course, the bounds cannot be computed unless the true measure i

known. Yet, if the true parameter can be assumed to belong to a finite set, then bounds can be computed over the set. This will be demonstrated in the following chapters, where we consider linear systems.

CHAPTER V

STATIONARY LINEAR SYSTEMS

In this chapter we restrict our attention to linear systems driven by white Gaussian inputs having time-invariant statistics. We make the assumption that the system has attained steady state, i.e. that all signals of interest are stationary. We first study the convergence of identification procedures. The convergence conditions are obtained in terms of the second order statistics associated with the models in the model set. If the true model is included in the set, it will be identified under a verifiable uniqueness condition. If the true model is not included in the model set, then the identification procedures converge to a model in the set which is closest to the true model in the information metric sense, introduced in Chapter 3, and in the sense of the second-order statistics associated with the models. Then we treat the L_1 convergence of the likelihood ratios and the ratios of *a posteriori* probabilities. We show that under a simple uniqueness condition the sequences of likelihood and *a posteriori* probability ratios are bounded in L_1 by simple exponentials. If the true system belongs to the given model set, then the bounds can be easily computed using the *a priori* data. The L_1 convergence results provide performance measures for the ML and the MAP identification methods. Finally, the analysis is extended to other modeling problems. Methods are suggested for selecting a reduced order

model to represent a high order system and for selecting a representative model from a set from which the true system, or an appropriate model of it, are known to take their values.

The convergence of the identification procedures is proved by direct application of the ergodic theorem. This chapter then depends only on the results of Chapter 3 and section 4.1 and the more advanced probabilistic arguments used in Chapter 4 are omitted. (Note that since we consider here a very specific class of observation sequences, we are, in fact, able to treat a more interesting class of problems than that considered in section 4.2, as the true parameter is not assumed to belong to the parameter set.)

5.1 Models and Densities

Consider the system

$$x_{n+1} = F_* x_n + G_* w_n$$

$$z_n = H_* x_n + v_n \tag{5.1a}$$

initialized at $n = n_o$ with

$$E\, x_{n_o} = 0 \qquad E\, x_{n_o} x_{n_o}^T = \Psi_{*,n_o}$$

where (w_n) and (v_n) are uncorrelated and mutually uncorrelated Gaussian

sequences with

$$E\, w_n = E\, v_n = 0$$

$$E\{w_n w_n^T\} = Q_* \quad ; \quad E\{v_n v_n^T\} = R_* \tag{5.1b}$$

The model set is a finite set of models for (5.1) denoted by

$$M_1 \equiv \left\{(F_j,\ G_j,\ H_j,\ Q_j,\ R_j)\ ;\ j \in K \equiv (o,\ldots,p)\right\} \tag{5.2}$$

Let

$$K' \equiv (* \cup K)$$

(As in Chapter 4, the restriction to a finite set is done for the analysis of convergence and consistency. In section 5.4 we consider other modeling problems and there the model set is allowed to be infinite. Also note that the results of this chapter can easily be extended to the case where the system (5.1a) is driven by an additional deterministic inputs sequence.)

Let

$$\hat{z}_{j,n} \equiv E_j^{U_{n-1}} z_n \quad ; \quad j \in K'$$

denote the one-step least squares prediction of z_n, given the past observations Z^{n-1}, assuming that the j'th model is the true one. For each i, j $\in$ K' let

$$\Sigma_{j,n} = \Sigma_j(n,\ n_o) \equiv E_j\left\{(z_n - \hat{z}_{j,n})^T(z_n - \hat{z}_{j,n})\right\} \tag{5.3}$$

and

$$\Gamma^i_{j,n} \equiv \Gamma^i_j (n, n_o) \equiv E_i\left\{(z_n - \hat{z}_{j,n})^T(z_n - \hat{z}_{j,n})\right\} \tag{5.4}$$

denote the prediction error covariance matrices according to the respective measures. (For each $j \in K'$, $\hat{z}_{j,n}$ and $\Sigma_{j,n}$ are computed, in essence, by a Kalman-Bucy filter corresponding to the j'th model.) Denote

$$\Sigma_j \equiv \lim_{n_o \to -\infty} \Sigma_j(n, n_o) \tag{5.5}$$

provided that the limit exists.

We shall use the following condition:

(c5.1) For each $j \in K'$ Σ_j exists and has a finite positive definite value.

A sufficient condition for (c5.1) is that each model corresponding to $j \in K'$ is detectable and controllable. For each $j \in K'$ Σ_j is obtained by running a Riccati equation, or equivalently, a Kalman-Bucy filter corresponding to $(F_j, G_j, H_j, Q_j, R_j)$.

Also denote

$$\Gamma^i_j \equiv \lim_{n_o \to -\infty} \Gamma^i_j (n, n_o) \tag{5.6}$$

provided that the limit exists. Γ^i_j is obtained by the following procedure. First, assuming $n_o = -\infty$, take $\Sigma_{j,n} = \Sigma_j$ for each $j \in K'$ and

for all $n \geq n_1$, where n_1 is any fixed integer. Then the dynamic equation generating simultaneously, according to the measure P_i, the state x_n and its' one step prediction by the j'th Kalman-Bucy filter $\hat{x}_{j,n}$ is

$$\begin{bmatrix} x_{i,n+1} \\ \hat{x}_{j,n+1} \end{bmatrix} = \begin{bmatrix} F_i & 0 \\ F_j K_j H_i & F_j(I-K_j H_j) \end{bmatrix} \begin{bmatrix} x_{i,n} \\ \hat{x}_{j,n} \end{bmatrix} + \begin{bmatrix} G_i & 0 \\ 0 & F_j K_j \end{bmatrix} \begin{bmatrix} w_n \\ v_n \end{bmatrix}$$

where

$$K_j = \Sigma_j H_j^T (H_j \Sigma_j H_j^T + R_j)^{-1}$$

Let

$$F_j^i \equiv \begin{bmatrix} F_i & 0 \\ F_j K_j H_i & F_j(I-K_j H_j) \end{bmatrix} \; ; \; G_j^i \equiv \begin{bmatrix} G_i & 0 \\ 0 & F_j K_j \end{bmatrix}$$

$$Q^i \equiv \begin{bmatrix} Q_i & 0 \\ 0 & R_i \end{bmatrix} \qquad ; \; H_j^i \equiv \begin{bmatrix} H_i & -H_j \end{bmatrix}$$

Then the matrix

$$\Psi_{j,n}^i \equiv E\left\{ \begin{bmatrix} x_{i,n+1} \\ \hat{x}_{j,n+1} \end{bmatrix} \begin{bmatrix} x_{i,n+1} \, , \, \hat{x}_{j,n+1} \end{bmatrix} \right\}$$

is generated by the Lyapunov equation

$$\Psi_{j,n+1}^i = F_j^i \, \Psi_{j,n}^i \, F_j^{i^T} + G_j^i \, Q^i \, G_j^i \tag{5.7}$$

Initialized at n_1 by any initial value. We can write

$$\Psi^i_{j,n} = \Psi^i_j (n, n_1)$$

Then let

$$\Psi^i_j = \lim_{n_1 \to -\infty} \Psi^i_j (n, n_1) \tag{5.8}$$

Finally

$$I^i_j = H^i_j \Psi^i_j H^{i^T}_j + R_i \tag{5.9}$$

It is well known that the limit (5.8) of (5.7) exists and is finite if the matrix F^i_j has all its' eigenvalues inside the unit circle. This is the case if for each $j \in K'$ F_j has all its' eigenvalues inside the unit circle and (F_j, H_j) is observable. Note, however, that these conditions are only sufficient, not necessary, for Γ^i_j to be finite, since (5.9) may be finite even if Ψ^i_j, obtained as the limit value of (5.7) is not finite.

Theorem 5.1

For each $j \in K'$ let the corresponding model be stable and observable and let $n_o = -\infty$. Then the residuals sequences $(z_n - \hat{z}_{j,n})$; $j \in K'$; $n \geq 0$ are ergodic according to the true probability measure.

Proof

We have by (5.5) $\Sigma_{j,n} = \Sigma_j$ for all $n \geq 0$. Since both (z_n) and

$(\hat{z}_{j,n})$ are linear operators on a zero mean Gaussian sequence (x_n), they are zero mean Gaussian, and so is the sequence $(z_n - \hat{z}_{n,j})$ for each $j \in K'$. Hence, $(z_n - \hat{z}_{j,n})$ is a zero mean stationary Gaussian sequence for each $j \in K'$. By proposition 2.1 we have that $(z_n - \hat{z}_{j,n})$ is ergodic if (and only if)

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} |R(k)|^2 = 0 \tag{5.10}$$

where $|R(k)|$ denotes the determinant of the matrix

$$R(k) = E\left\{(z_n - \hat{z}_{j,n})(z_{n+k} - \hat{z}_{j,n+k})^T\right\} = \begin{cases} H_j^i \Psi_j^i H_j^{iT} + R_i & k=0 \\ H_j^i \Psi_j^i H_j^{iT} (F_j^i)^k & k>0 \end{cases} \tag{5.11}$$

We have for any $k > 0$

$$|R(k)| = |H_j^i H_j^{iT}| \, |\Psi_j^i| \, |F_j^i|^k$$

Since all eigenvalues of F_j^i are inside the unit circle, then

$$|F_j^i| < 1$$

Hence

$$\lim_{n\to\infty} \sum_{k=1}^{n} |R(k)|^2 = |H_j^i H_j^{iT}|^2 \, |\Psi_j^i|^2 \lim_{n\to\infty} \sum_{k=1}^{n} |F_j^i|^{2k}$$

$$= \frac{|H_j^i H_j^{iT}|^2 \, |\Psi_j^i|^2}{1 - |F_j^i|^2} < \infty \tag{5.12}$$

Yielding

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{j=0}^{n} |R(k)|^2 = 0$$

The assertion follows. ■

Note that the stability and observability of the models, assumed in theorem 5.1 are only sufficient, not necessary. In fact, we have proved ergodicity of the state residuals $(x_n - \hat{x}_{j,n})$, which is not necessary, to show the ergodicity of $(z_n - \hat{z}_{j,n})$. In the sequel we shall directly use the following condition.

(c5.2) For each $j \in K'$ the residuals sequence $(z_n - \hat{z}_{j,n})$ is ergodic.

5.2 Information, Convergence and Consistency

Consider the system (5.1) and the model set (5.2). Let condition (c5.1) hold. Then the conditional probability density of z_n given the past observations z^{n-1}, corresponding to each model is

$$f_j(z_n|z^{n-1}) = \left[(2\pi)^{\ell}|\Sigma_j|\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(z_n - \hat{z}_{j,n})^T \Sigma_j^{-1} (z_n - \hat{z}_{j,n})\right\}; \quad j \in K' \qquad (5.13)$$

where ℓ is the dimension of z_n.

In Chapter 3 we have defined for each pair $j, k \in K'$

$$\bar{I}_n(k;j) \equiv E_* \log \frac{f_k(z_n|z^{n-1})}{f_j(z_n|z^{n-1})}$$

and

$$d_n(k;j) \equiv |\bar{I}_n(k;j)|$$

We have for each $j \in K'$

$$E_* \log f_j(z_n|z^{n-1}) = -\frac{\ell}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}\operatorname{tr}\Sigma_j^{-1}\Gamma_j^* \qquad (5.14)$$

and for each pair $j, k \in K'$

$$\bar{I}_n(k;j) = \bar{I}(k;j) = \frac{1}{2}\log|\Sigma_j| + \frac{1}{2}\operatorname{tr}\Sigma_j^{-1}\Gamma_j^* - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}\operatorname{tr}\Sigma_k^{-1}\Gamma_k^* \qquad (5.15)$$

Let

$$L_j^i \equiv \log|\Sigma_j| + \operatorname{tr}\Sigma_j^{-1}\Gamma_j^i \qquad i, j \in K' \qquad (5.16)$$

Then we have

$$\bar{I}_n(k;j) = \frac{1}{2}[L_j^* - L_k^*] \qquad j, k \in K \qquad (5.17)$$

Also, by theorem 3.1

$$\bar{I}_n(*;j) \geq 0 \quad \text{for each } j \in K$$

Hence

$$d(*;j) = d_n(*;j) = \bar{I}_n(*;j) \quad \text{for each } j \in K$$

Thus, for any $j, k \in K$

$$d(*;j) - d(*;k) = \bar{I}_n(*;j) - \bar{I}_n(*;k)$$

$$= \bar{I}_n(k;j)$$

$$= \frac{1}{2}\,[L_j^* - L_k^*] \tag{5.18}$$

Hence

$$d(*;j) \geq d(*;k)$$

if any only if

$$L_j^* \geq L_k^*$$

Lemma 5.1

Let (z_n) be generated by (5.1) and let condition (c5.1) hold. Then, under condition (c5.2), for any $j, k \in K$

$$\lim_{n\to\infty} h_k^j(Z^n) = 0 \quad \text{a.e.} \tag{5.19}$$

if

$$L_k^* < L_j^* \tag{5.20}$$

and only if

$$L_k^* \leq L_j^* \tag{5.21}$$

Proof

$$\log h_k^j (Z^n) = \sum_{m=1}^{n} \log h_k^j (z_m | Z^{m-1}) \tag{5.22}$$

We have

$$\log h_k^j (z_n | Z^{n-1}) = \frac{1}{2} \log \frac{|\Sigma_k|}{|\Sigma_j|}$$

$$- \frac{1}{2} (z_n - \hat{z}_{j,n})^T \Sigma_j^{-1} (z_n - \hat{z}_{j,n})$$

$$+ \frac{1}{2} (z_n - \hat{z}_{k,n})^T \Sigma_k^{-1} (z_n - \hat{z}_{k,n}) \tag{5.23}$$

Since for each $j \in K$ the residuals $(z_n - \hat{z}_{j,n})$ are ergodic, it follows from the ergodic theorem (theorem 2.5) that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \log h_k^j (z_m | Z^{m-1}) = E_* \log h_k^j (z_m | Z^{m-1}) \quad \text{a.e.}$$

$$= \bar{I}_m(j;k)$$

$$= \frac{1}{2} (L_k^* - L_j^*) \tag{5.24}$$

Now if

$$L_k^* < L_j^* \tag{5.25}$$

Then obviously

$$\lim_{n\to\infty} \sum_{m=1}^{n} \log h_k^j (z_m | Z^{m-1}) = \lim_{n\to\infty} \log h_k^j (Z^n) = -\infty \quad \text{a.e.}$$

yielding

$$\lim_{n\to\infty} h_k^j (z^n) = 0$$

To prove that (5.19) implies (5.21), suppose that (5.19) holds, but (5.21) does not, then

$$L_k > L_j \tag{5.26}$$

and by (5.24)

$$\lim_{n\to\infty} \sum_{m=1}^{n} \log h_k^j (z_m | z^{m-1}) = \lim_{n\to\infty} \log h_k^j (z^n) = \infty$$

implying

$$\lim h_k^j (z^n) = \infty \tag{5.27}$$

which contradicts (5.19). Thus, (5.19) implies (5.21). ■

Consider the following condition

(c5.3) There exists some parameter $k \in K$ such that

$$L_k^* < L_j^* \quad \text{for all } j \in K \; ; \; j \neq k \tag{5.28}$$

Theorem 5.2

Consider the system (5.1) and the model set M_1, and let (c5.1) hold. Under conditions (c5.2) and (c5.3) the ML, the MAP and the LS identification methods will converge a.e. and the LS method will also converge in m.s. to the model $(F_k, G_k, H_k, Q_k, R_k)$.

Proof

By lemma 5.1 condition (c4.1) is satisfied for the parameter k. The assertion then follows from theorems 4.1 through 4.4. ■

Note that by (5.18) the identified model is the closest to the true model in the metric d.

Corollary 5.1

The convergence specified in theorem 5.2 will be to a model in M_1, such that

$$|L_k^* - L_*^*| = \min \left\{ |L_j^* - L_*^*| \; ; \; j \in K \right\} \tag{5.29}$$

Proof

By theorem 3.1 we have

$$\bar{I}\,(*;j) \geq 0 \qquad \text{for each } j \in K$$

Hence

$$L_j^* - L_*^* > 0$$

if

$$L_j^* \neq L_*^*$$

So the assertion follows from lemma 5.1 and theorem 5.2. ■

The identification methods will then converge to a parameter in K, closest to the true parameter in the scalar L, which in turn implies closeness of the corresponding models in terms of their output statistics.

Corollary 5.2

Suppose that the true system belongs to the set M_1, i.e. let $(F, G, H, Q, R) = (F_r, G_r, H_r, Q_r, R_r)$ for some $r \in K$. Let conditions (c5.1) and (c5.2) hold and suppose that for each $j \in K$; $j \neq r$ we have

$$L_j^* = L_j^r \neq L_r^r = L_*^* \tag{5.30}$$

Then the identification procedures will converge to the true model in the senses specified in theorem 5.2.

Proof

The result follows immediately from corollary 5.1. ■

To compute the scalars L_j^*, $j \in K$ one must compute the matrices Σ_j and Γ_j^*. While the matrix Σ_j can be computed by running a Riccati equation corresponding to the j'th model to steady-state, the matrix Γ_j^* cannot be computed unless the true measure or, equivalently, the true system is known. If $r \in K$ is the true parameter, then

$$\Gamma_r^* = \Gamma_r^r = \Sigma_r$$

and consequently

$$L_r^* = L_r^r = \log |\Sigma_r| + \ell \tag{5.31}$$

In the identification problem the true parameter is unknown. If the true parameter can be assumed to belong to the parameter set, then (5.30) will have to be checked for all pairs of parameters in the set, namely

(c5.4) For all pairs i, j ∈ K ; i ≠ j

$$L_j^i \neq L_i^i = \log |\Sigma_i| + \ell \tag{5.32}$$

Theorem 5.3

Let the system (5.1) belong to the set M_1, and let conditions (c5.1) and (c5.2) hold. Then under conditions (c5.4) the true model is identifiable a.e. by the ML and the MAP estimates and identifiable a.e. and in m.s. by the LS estimate.

Proof

Under condition (c5.4) we have (5.30). The assertion then follows directly from corollary 5.2. ■

5.3 L_1 Convergence

We have shown in section 4.4 the L_1 convergence of the likelihood ratios and the *a posteriori* probability ratios under condition (c4.1). Furthermore, it was shown that under condition (c4.5) bounds on the L_1 convergence rates can be established, thus providing a measure of performance of the ML and the MAP estimation methods. We now show L_1 convergence and derive L_1 convergence bounds for the identification of stationary linear systems treated in this chapter.

Consider the system (5.1) and the model set M_1 and let condition (c5.1) hold. We have shown ((5.17)) that under condition (c5.2)

$$\bar{I}_n(k;j) = \bar{I}(k;j) = \frac{1}{2}[L_j^* - L_k^*] \text{ for all } n$$

for each pair k, j ∈ K where L_j^* ; j ∈ K are constants.

Theorem 5.4

Consider the system (5.1) and the model set M_1 given by (5.2). Under conditions (c5.1) and (c5.3) for each j ∈ K ; j ≠ k the sequences $(h_j^k(z^n))$ and $\left(\frac{f^b(k|z^n)}{f^b(j|z^n)}\right)$ converge in L_1 to infinity. Furthermore, the sequences converge at rates no slower than exponential.

Proof

By lemma 5.1 condition (c5.3) implies condition (c4.1). The L_1 convergence of both sequences follows from theorem 4.10.

Now let

$$\alpha_j \equiv \frac{1}{2}[L_j^* - L_k^*] \quad \text{for each } j \in K ; j \neq k$$

then following the proof of theorem 4.11, we get by (4.11) and (4.12) for each j ∈ K ; j ≠ k

$$E\, h_j^k(z^n) \geq e^{(n+1)\alpha_j} \tag{5.33}$$

and

$$E\,\frac{f^b(k|z^n)}{f^b(j|z^n)} \geq \frac{f^b(k)}{f^b(j)}\, e^{(n+1)\alpha_j} \tag{5.34}$$

■

The rates $\alpha_j = \bar{I}(k;j)$; j ∈ K ; j ≠ k can only be computed, as discussed in section 5.2, if the true model is known. If the true model

is only known to belong to the set M_1, then the rates can be bounded as follows. We have seen that if $k \in K$ is the true parameter, then ((5.31))

$$L_k^* = \log |\Sigma_k| + \ell$$

Now since

$$\alpha_j = \frac{1}{2} [L_j^* - L_k^*] \quad \text{for each } j \in K ; j \neq k$$

where k is now the true parameter, we have

$$\alpha_j = \frac{1}{2} [L_j^* - \log |\Sigma_k| - \ell]$$

Then

$$\alpha_j \geq \alpha \equiv \min\left\{\min\left\{\frac{1}{2} [L_j^k - \log |\Sigma_k| - \ell] \; ; \; L_j^k - \log |\Sigma_k| - \ell = 0 \right.\right.$$

$$\left.\left. \text{for all } j \in K ; j \neq k \right\} k \in K\right\} \qquad (5.35)$$

(5.35) reads as follows: For each $k \in K$ suppose that k is the true parameter. If

$$L_j^k - \log |\Sigma_k| - \ell \geq 0 \quad \text{for all } j \in K ; j \neq k \qquad (5.36)$$

then take the min over j of (5.36). Continue the procedure over all $k \in K$, discarding such k for which (5.36) does not hold (since then k cannot be the true parameter, for which (5.36) always holds). Then take the least of all the minimum values of (5.36) found above. Note that this procedure does not identify the true parameter, but rather finds a lower bound for α_j over $j \in K$.

The above discussion is summarized in the following theorem.

Theorem 5.5

Consider the system (5.1) and suppose that its' true model belongs to the set M_1 given by (5.2). Then under condition (c5.3) we have

$$E_j^r (Z^n) \geq e^{(n+1)\alpha} \tag{5.37}$$

$$E \frac{f^b(r|Z^n)}{f^b(j|Z^n)} \geq \frac{f^b(r)}{f^b(j)} e^{(n+1)\alpha} \tag{5.38}$$

for each $j \in K$; $j \neq r$ where r is the true parameter and where α is given by (5.35).

As discussed in section 4.4 the bounds (5.37) and (5.38) provide performance measures for the ML and the MAP estimation methods. We have shown that bounds on the L_1 convergence rates of the indicated ratios can actually be computed for stationary Gaussian linear systems.

5.4 Model Selection

In practice, when a mathematical model of a dynamical system is required for purposes of estimation and control, one often knows, to certain approximation, what the model should be. However, because of implementation constraint one has to select a different model. Such is the case when the actual system is of high order, but the available computation and storage capabilities are such that only a low order model can be handled. Another modeling problem arises when the actual

system's model is known to take its' values, which may be time-varying, from a given set, but only a single model can be considered. An example of practical significance is the dynamical model of an aircraft, whose parameters vary considerably over its' flight envelope. However, the airborne computation and storage capabilities are limited and a single model of the aircraft dynamics must be used throughout its operation.

These are not identification problems in the strict sense. Nevertheless the analysis in Chapter 3 and sections 5.1 and 5.2 suggests a natural extension of the results into the model selection problems introducad above. It should be emphasized that unlike the investigation of convergence and consistency of parameter estimate the results of this section apply to infine and even non-compact parameter sets.

5.4.1 The Selection of a Reduced Order Model

Suppose that the true system or an approximate model of it are known, but their dimensions are too high for implementation of estimation and control prodecures. A model of lower dimension is then desired. Let the true system, or an approximate model of it be given by (5.1) and let

$$M \equiv \left\{(F_s, G_s, H_s, Q_s, R_s) \; ; \; s \in S\right\} \tag{5.39}$$

be a model set of dimension lower than that of (5.1). The system coefficients in M depend on a parameter vector s belonging to a parameter set S. It is desired to find the model in the set M which is closest to the true system $(F_*, G_*, H_*, Q_*, R_*)$ is some meaningful distance

sense.

For each $s \in S$ let

$$L_s^* \equiv \log |\Sigma_s| + \text{tr } \Sigma_s^{-1} \Gamma_s^* \tag{5.40}$$

where

$$\Gamma_s^* \equiv E_* \left\{ (z_n - \hat{z}_{s,n}) \quad (z_n - \hat{z}_{s,n})^T \right\}$$

$\hat{z}_{s,n}$ is the one-step least-square prediction of z_n given the past observations z^{n-1} assuming that s is the true parameter value, and Σ_s is the correspinding prediction error covariance matrix. Σ_s is obtained by running a Riccati equation corresponding to the model (F_s, G_s, H_s, Q_s, R_s) to steady-state. The computation of Γ_s^* was discussed in the previous section. Let $s^\circ \in S$ be a parameter which satisfies the following criterion

$$L_{s_o}^* < \{L_s^* ; s \in S, s \neq s^\circ\}$$

Then, following the reasoning of section 5.2 the model (F_{s°, G_{s°, H_{s°, Q_{s°, R_{s°) satisfies the following equivalent criteria:

1) The model which is closer to the true model than any other model in M in the sense $d(*;s^\circ) \leq \{d(*;s);\ s \in S\}$.
2) The model which would be favored over any other model in M by the incoming information.
3) The model which would identified as the true model among any finite set of models from the set M by the maximum

likelihood and Bayesian estimation techniques.

The model selection problem reduces then to the minimization problem

$$\min_{s} \left\{ L_s^* \; ; \; s \in S \right\} \tag{5.41}$$

We do not address the algorithmic problem of solving (5.41) or the existence of a unique minimum of L_s^* on S. These problems are suggested for further research.

5.4.2 The Selection of a Representative Model

Suppose that the model of a linear system whose parameters may be time-varying is known to take its values from a set

$$M \equiv \left\{ (F_s, G_s, H_s, Q_s, R_s) \; ; \; s \in S \right\}$$

Two different cases may be considered.

1) The model takes a certain constant value in the set M and there is no prior knowledge even in a probabilistic sense on what value it might be.
2) During the system's operation its' mathematical model varies over the model set M. However, it is not possible to consider the model's time program.

In either case it is desired to select a single model from the set M to represent the system throughout its' operation. One criterion for the selection of such a model is that the maximum possible distance d between the representing model and the true model (whatever it might be)

will be minimal.

The procedure for selecting the representative model from M will then be as follows. First, for each parameter $s \in S$ find the parameter t whose distance from s is maximal, and the corresponding maximum distance. Then find the parameter s for which the maximal distance found is the first step is minimal.

The distance between a parameter s and the parameters t of the set S is maximized over t by maximizing with respect to t

$$L_s^t = \log |\Sigma_s| + \text{tr } \Sigma_s^{-1} \Gamma_s^t \tag{5.42}$$

where, as before

$$\Sigma_s \equiv E_s \left\{ (z_n - \hat{z}_{s,n}) \quad (z_n - \hat{z}_{s,n})^T \right\}$$

is obtained by running a Kalman-Bucy filter to steady-state, and

$$\Gamma_s^t \equiv E_t \left\{ (z_n - \hat{z}_{s,n}) \quad (z_n - \hat{z}_{s,n})^T \right\}$$

is obtained by running a Lyapunov equation to steady-state, as shown in the previous section.

The representative model is then found by solving the minimax problem

$$\min_s \max_t \left\{ L_s^t \; ; \; s, \; t \in S \right\} \tag{5.43}$$

The uniquenes: of the solution of (5.43) is suggested for further research.

Remarks

1) The procedures described in this chapter find, in general, a model in the model set, whose output (or observations) statistics are best matched with those of the true system. However, for the modeling problems considered above, the role of the output can be played by any linear function of the state variables. If, for instance, it is desired to emphasize certain variables that affect the system's performance more than the others, or that can can be measured better than the others, then these variables can be selected as outputs for the model selection procedures described above.

2) The problem of selecting a single model from a model set, considered in sections 5.4.1 and 5.4.2 can be generalized to a problem of selecting a number of models from the set, so that the model set is approximately represented by a finite set of models. An identification procedure can then be employed "on-line" to find the model in the finite set which is closest to the true system. The selection of a finite model set would require, as a first step, the division of the infinite parameter set into a finite number of subsets. The way in which the parameter space should be divided would depend on considerations of the physical problem involved, but it seems obvious that the division could employ the metric topology of the parameter space introduced in Chapter 3. (Just as interval lengths are used in R^n, say, to divide a rectangle into equal parts.) The selection of a representative model for each subset

is then performed as described in sections 5.4.1 and 5.4.2 above. Further research of this seemingly promising approach to system modeling and identification problems is recommended.

CHAPTER VI

NON-STATIONARY LINEAR SYSTEMS

The assumption of stationarity made in the previous chapter is now removed, as we consider the general case of non-stationary, time-varying linear systems. We first derive expressions for the information in the observations, discriminating one model in the model set against another. The information conditions for the consistency of the estimates are interpreted in terms of the second-order statistics associated with the different models and computed by solving the corresponding Riccati equations (or, equivalently, running Kalman-Bucy filters). The consistency result for time varying systems is not, however, as explicit as in the stationary case. The L_1 convergence of the likelihood and the *a posteriori* probability ratios is investigated. The separate contributions of the stochastic and the deterministic parts of the input to the information and, consequently, to the L_1 convergence rates are shown.

6.1 Models

Consider the system

$$x_{n+1} = F_{*,n}x_n + G_{*,n}w_n$$

$$z_n = H_{*,n}x_n + v_n \qquad (6.1a)$$

initialized at $n = n_o$ with

$$E_* x_{n_o} = 0 \; ; \; E_* x_{n_o} x_{n_o}^T = \Psi_*$$

where (w_n) and (v_n) are uncorrelated and mutually uncorrelated Gaussian sequences with

$$E\, w_n = E\, v_n = 0$$

$$E\left\{w_n w_n^T\right\} = Q_{*,n} \; ; \; E\left\{v_n v_n^T\right\} = R_{*,n} \tag{6.1b}$$

Consider a finite set of families of models

$$M_2 \equiv \Big\{(F_{j,n},\, G_{j,n},\, H_{j,n},\, \Psi_j,\, Q_{j,n},\, R_{j,n}) \; ; \; j \in K = (o,1,\ldots,p)\Big\} \tag{6.2}$$

Let (z_n) be an ℓ dimensional observation sequence. The conditional probability density of z_n given the past observations z^{n-1} and corresponding to each model is given by

$$f_j(z_n | z^{n-1}) = \left[(2\pi)^{\ell} |\Sigma_{j,n}|\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(z_n - \hat{z}_{j,n})^T \Sigma_{j,n}^{-1} (z_n - \hat{z}_{j,n})\right\}$$

$$; \; j \in K \tag{6.3}$$

where, as before, $\hat{z}_{j,n}$ is the one-step prediction of z_n given the past observations z^{n-1}, assuming that the j'th model is the true one (i.e. assuming that the observations are generated by the j'th model), and $\Sigma_{j,n}$ is the corresponding error covariance matrix. Both $\hat{z}_{j,n}$ and $\Sigma_{j,n}$ are generated by a Kalman-Bucy filter corresponding to the j'th model.

6.2 Information, Convergence and Consistency

The information in a single observation z_n, favoring the k'th model against the j'th model will now be derived.

$$I_n(k;j) = E_*^{U_{n-1}} \log \frac{f_k(z_n|z^{n-1})}{f_j(z_n|z^{n-1})}$$

$$= E_*^{U_{n-1}} \left\{ -\frac{1}{2}\log|\Sigma_{k,n}| - \frac{1}{2}(z_n - \hat{z}_{k,n})^T \Sigma_{k,n}^{-1} (z_n - \hat{z}_{k,n}) \right\}$$

$$+ \frac{1}{2}\log|\Sigma_{j,n}| + \frac{1}{2}(z_n - \hat{z}_{j,n})^T \Sigma_{j,n}^{-1} (z_n - \hat{z}_{j,n})$$

$$= \frac{1}{2}\log\frac{|\Sigma_{j,n}|}{|\Sigma_{k,n}|} - \frac{1}{2}\operatorname{tr}\Sigma_{k,n}^{-1} E_*^{U_{n-1}}\left\{z_n z_n^T\right\} + \frac{1}{2}\hat{z}_{k,n}^T \Sigma_{k,n}^{-1}\hat{z}_{*,n}$$

$$+ \frac{1}{2}\hat{z}_{*,n}^T \Sigma_{k,n}^{-1}\hat{z}_{k,n} - \frac{1}{2}\hat{z}_{*,n}^T \Sigma_{k,n}^{-1}\hat{z}_{*,n} + \frac{1}{2}\operatorname{tr}\Sigma_{j,n}^{-1} E_*^{U_{n-1}}\left\{z_n z_n^T\right\}$$

$$- \frac{1}{2}\hat{z}_{j,n}^T \Sigma_{j,n}^{-1}\hat{z}_{*,n} - \frac{1}{2}\hat{z}_{*,n}^T \Sigma_{j,n}^{-1}\hat{z}_{j,n} + \frac{1}{2}\hat{z}_{*,n}^T \Sigma_{j,n}^{-1}\hat{z}_{*,n} \quad \text{a.e.}$$

But

$$E_*^{U_{n-1}} \left\{ z_n z_n^T \right\} = \Sigma_{*,n} + \hat{z}_{*,n} \hat{z}_{*,n}^T \qquad \text{a.e.}$$

Hence

$$I_n(k;j) = \frac{1}{2} \log \frac{|\Sigma_{j,n}|}{|\Sigma_{k,n}|} + \frac{1}{2} \operatorname{tr} \Sigma_{*,n} (\Sigma_{j,n}^{-1} - \Sigma_{k,n}^{-1})$$

$$- \frac{1}{2} (\hat{z}_{*,n} - \hat{z}_{k,n})^T \Sigma_{k,n}^{-1} (\hat{z}_{*,n} - \hat{z}_{k,n})$$

$$+ \frac{1}{2} (\hat{z}_{*,n} - \hat{z}_{j,n})^T \Sigma_{j,n}^{-1} (\hat{z}_{*,n} - \hat{z}_{j,n}) \quad \text{a.e.} \qquad (6.4)$$

Let

$$I_n^{(1)}(k;j) \equiv \frac{1}{2} \log \frac{|\Sigma_{j,n}|}{|\Sigma_{k,n}|} + \frac{1}{2} \operatorname{tr} \Sigma_{*,n} (\Sigma_{j,n}^{-1} - \Sigma_{k,n}^{-1}) \qquad (6.5)$$

and

$$I_n^{(2)}(k;j) \equiv - \frac{1}{2} (\hat{z}_{*,n} - \hat{z}_{k,n})^T \Sigma_{k,n}^{-1} (\hat{z}_{*,n} - \hat{z}_{k,n})$$

$$+ \frac{1}{2} (\hat{z}_{*,n} - \hat{z}_{j,n})^T \Sigma_{j,n}^{-1} (\hat{z}_{*,n} - \hat{z}_{j,n}) \qquad (6.6)$$

Hence

$$I_n(k;j) = I_n^{(1)}(k;j) + I_n^{(2)}(k;j) \qquad (6.7)$$

Suppose that condition (c4.2) is satisfied. Then, by lemma 4.2 and by theorems 4.1 through 4.4 conditions (c4.3) or (c4.4) are sufficient

for convergence of the estimates to the k'th model in M_1. However, it is not difficult to see that the verification of conditions (c4.3) or (c4.4) is not possible for the general case considered here under any conditions imposed on the deterministic part $I_n^{(1)}(k;j)$, due to the random part $I_n^{(2)}(k;j)$. In section 6.3 we shall show that $I_n^{(2)}(k;j)$ can be further separated into deterministic and stochastic parts. We now show that under the assumption that the true model belongs to the given set M_2, the information expression for the time varying system under consideration is simplified and consequently some explicit conditions for identification can be obtained.

Suppose that some $k \in K$ is the true parameter, then by (6.4)

$$I_n(k;j) = I_n'(k;j) + I_n''(k;j) \tag{6.8}$$

where

$$I_n'(k;j) \equiv \frac{1}{2} \log \frac{|\Sigma_{j,n}|}{|\Sigma_{k,n}|} + \frac{1}{2} \operatorname{tr} (\Sigma_{k,n} \Sigma_{j,n}^{-1} - I) \tag{6.9}$$

and

$$I_n''(k;j) \equiv \frac{1}{2} (\hat{z}_{k,n} - \hat{z}_{j,n})^T \Sigma_{j,n}^{-1} (\hat{z}_{k,n} - \hat{z}_{j,n}) \tag{6.10}$$

Consider the following condition

(c6.1) For some $k \in K$ and for each $j \in K$; $j \neq k$ there exists some scalar $\alpha_j > 0$ and a subsequence (n^j) of (n) such that

$$||\Sigma_{k,n^j} - \Sigma_{j,n^j}|| \geq \alpha_j \quad \text{for all } n^j \tag{6.11}$$

where

$$||A|| \equiv |\det A|$$

Lemma 6.1

Let some $k \in K$ be the true parameter, i.e. let

$$(F_{*,n}, G_{*,n}, H_{*,n}, \Psi_*, Q_{*,n}, R_{*,n}) = (F_{k,n}, G_{k,n}, H_{k,n}, \Psi_k, Q_{k,n}, R_{k,n})$$

Then condition (c6.1) implies condition (c4.4) for k.

Proof

Clearly

$$I_n''(k;j) \geq 0 \quad \text{for all } n \text{ for each } j \in K$$

Thus

$$I_n(k;j) \geq I_n'(k;j) \quad \text{for all } n \text{ for each } j \in K.$$

It will suffice then to show that condition (c6.1) implies the existence of a subsequence (n_r^j) of (n^j) and some $\varepsilon_j > 0$ such that

$$I'_{n_r^j}(k;j) \geq \varepsilon_j \quad \text{for all } n_r^j \tag{6.12}$$

Consider the following equation

$$|\Sigma_{k,n} - \lambda_n \Sigma_{j,n}| = 0 \tag{6.13}$$

For positive definite $\Sigma_{k,n}$ and $\Sigma_{j,n}$ there exists a nonsingular matrix A_n such that (Anderson, [1958], p.341):

$$A_n^T \Sigma_{k,n} A_n = \Lambda_n \tag{6.14}$$

and

$$A_n^T \Sigma_{j,n} A_n = I \tag{6.15}$$

where Λ_n is a diagonal matrix whose elements are $\lambda_{n,i}$; $i=1,\ldots,\ell$, the roots of (6.13). In addition, we have $\lambda_{n,i} \geq 1$ for all $i=1,\ldots,\ell$ and $n \geq 0$.

It is easy to verify that $I_n'(k;j)$ remains invariant under the transformations (6.14) and (6.15). Hence

$$\begin{aligned} I_n'(k;j) &= -\frac{1}{2} \log |\Lambda_n| + \frac{1}{2} \operatorname{tr} (\Lambda_n - I) \\ &= \frac{1}{2} \sum_{i=1}^{\ell} \left[\lambda_{n,i} - \log \lambda_{n,i} - 1 \right] \end{aligned} \tag{6.16}$$

Suppose that for some subsequence (n^j) of (n)

$$||\Sigma_{k,n_j} - \Sigma_{j,n_j}|| \geq \alpha_j > 0 \quad \text{for all } n^j \tag{6.17}$$

Then there exists some $\zeta_j > 0$ and a subsequence (n_r^j) of (n^j) such that

$$|\lambda_{n_r^j,i} - 1| \geq \zeta_j \quad \text{for all } n_r^j \text{ for each } i=1,\ldots,\ell \tag{6.18}$$

since if such ζ_j and such (n_r^j) do not exist, then

$$\lambda_{n^j,\, i_{n^j}} \to 1 \quad \text{as } n \to \infty \tag{6.19}$$

where

$$\lambda_{n^j,\, i_{n^j}} \equiv \min\left\{ \lambda_{n^j,\, i} ;\ i=1,\ldots,\ell \right\}$$

and then

$$\left| \Sigma_{k,n_r^j} - \Sigma_{j,n_r^j} \right| \to \left| \Sigma_{k,n_r^j} - \lambda_{n_r^j,\, i_{n_r^j}} \Sigma_{j,n_r^j} \right| = 0$$

as $n_r^j \to \infty$, contradicting (6.17). Hence, (6.17) implies (6.18).

Now consider (6.16). Since

$$a - \log a - 1 \geq 0 \tag{6.20}$$

with equality if and only if $a = 1$, and since the function on the left hand side of (6.20) is convex in a, it follows that given $\zeta > 0$ there exists some $\alpha > 0$ such that

$$a - \log a - 1 > \alpha$$

whenever $|a - 1| > \zeta$

Thus, finally, (6.18) implies that there exists some $\epsilon_j > 0$ such that

$$I^{*}_{n_r^j}(k;j) \geq \epsilon_j \quad \text{for all } n_r^j \tag{6.21}$$

The assertion follows. ■

We have shown in Chapter 4 that consistency of the parameter estimates (or, equivalently, identifiability of the dynamical system) follows from conditions (c4.2) and (c4.4). Condition (c4.4) (or, more generally, (c4.3)) seems to be, for obvious reasons, the "crucial" condition for the strong consistency of the estimates. We show below that condition (c4.2) holds for the case of time invariant stationary linear systems. It seems, however, that condition (c4.2) would hold for very general classes of observation sequences. For the general case of time varying systems we condition the consistency result on condition (c4.2) which has to be checked for each case under consideration. It seems, in particular, that condition (c4.2) would not be difficult to verify for the class of periodically varying linear systems and for systems driven by bounded deterministic inputs. This, however, is left for future research.

Theorem 6.1

Suppose that the system (6.1) belongs to the set M_2 specified by (6.2). Furthermore, suppose that condition (c4.2) holds. Then the system is identifiable a.e. by the ML and the MAP estimates and identifiable a.e. and in m.s. by the LS estimate on the set if condition (c6.1) is satisfied.

Proof

The assertion follows from lemma 6.1 and theorem 4.6. ∎

Now consider the case, treated in Chapter 5, where the true system, given by (5.1), is assumed to belong to the set M_1, given by (5.2). Under conditions (c5.1) and (c5.2) condition (c6.1) simplifies to the the following condition:

(c6.2) For each $j \in K$; $j \neq k$

$$|\Sigma_j - \Sigma_k| \neq 0$$

Suppose that $k \in K$ is the true parameter. We have for each $j \in K$; $j \neq k$

$$I_n(k;j) = -\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}\operatorname{tr}\Sigma_j^{-1} E_*^{U_{n-1}}\left\{(z_n - \hat{z}_{j,n})(z_n - \hat{z}_{j,n})^T\right\}$$

$$+\frac{1}{2}\log|\Sigma_k| + \frac{1}{2}\operatorname{tr}\Sigma_k^{-1} E_*^{U_{n-1}}\left\{(z_n - \hat{z}_{k,n})(z_n - \hat{z}_{k,n})^T\right\} \tag{6.22}$$

where, for each $j \in K$

$$E_*^{U_{n-1}}\left\{(z_n - \hat{z}_{j,n})(z_n - \hat{z}_{j,n})^T\right\}$$

$$= E_*^{U_{n-1}}\left\{z_n z_n^T\right\} - \hat{z}_{*,n}\hat{z}_{j,n}^T - \hat{z}_{j,n}\hat{z}_{*,n}^T + \hat{z}_{j,n}\hat{z}_{j,n}^T$$

$$= \Sigma_* + (\hat{z}_{*,n} - \hat{z}_{j,n})(\hat{z}_{*,n} - \hat{z}_{j,n})^T \tag{6.23}$$

and, since $k = *$,

$$J_n(k;j) \equiv \log h_j^k(z_n|z^{n-1}) - I_n(k;j)$$

$$= \frac{1}{2}\operatorname{tr}\Sigma_k(\Sigma_j^{-1} - I)$$

$$+ \frac{1}{2} \operatorname{tr} \Sigma_k^{-1} \left[(z_n - \hat{z}_{k,n})(z_n - \hat{z}_{k,n})^T \right]$$

$$- \frac{1}{2} \operatorname{tr} \Sigma_j^{-1} \left[(z_n - \hat{z}_{j,n})(z_n - \hat{z}_{j,n}) - (\hat{z}_{k,n} - \hat{z}_{j,n})(\hat{z}_{k,n} - \hat{z}_{j,n})^T \right] \tag{6.24}$$

Since the sequences $(z_n - \hat{z}_{j,n})$ and $(\hat{z}_{k,n} - \hat{z}_{j,n})$ are ergodic for all $j, k \in K$, so are the sequences $(I_n(k;j))$ and $(J_n(k;j))$. It follows from lemma 4.2 that condition (c4.2) is satisifed. Condition (c4.3) is satisfied if condition (c5.4) is satisfied, by theorem 3.1 and the ergodicty of $(I_n(k;j))$. The identifiability of the system under condition (c5.4) thus follows from theorem 4.5.

6.3 L_1 Convergence

We have shown in section 4.4 that by bounding the information in the observations away from zero, bounds on the L_1 convergence rates of the likelihood and the *a posteriori* probability ratios can be established, which in turn provides performance measures for the ML and the MAP estimation procedures. In this section we consider the identification of a general class of time-varying systems driven by stochastic and deterministic inputs. The fact that only convergence in L_1 and not in the stronger senses of a.e. and m.s. is sought enables us to obtain rather explicit results. The stochastic and the deterministic parts of the input are shown to contribute separately to the convergence rates of the identification procedures.

Consider the system

$$x_{n+1} = F_{*,n}\, x_n + G_{*,n}\, u_n + J_{*,n}\, w_n$$

$$z_n = H_{*,n}\, x_n + v_n \tag{6.25}$$

where (u_n) is a deterministic (known) input sequence and the other elements are as specified in section 6.1. Also consider a model set

$$M_3 \equiv \left\{ (F_{j,n}, G_{j,n}, J_{j,n}, H_{j,n}, \Psi_j, Q_{j,n}, R_{j,n}) \;;\; j \in K \right\} \tag{6.26}$$

where $Q_{j,n}$ and $R_{j,n}$ are the covariance matrices of (u_n) and (v_n) respectively, corresponding to each model.

The incremental information for favoring a parameter k over a parameter j in the set K, is given by (6.8). For each $j \in K$ we have

$$I_n^m (*;j) \equiv \frac{1}{2} (\hat{z}_{*,n} - \hat{z}_{j,n})^T \Sigma_{j,n}^{-1} (\hat{z}_{*,n} - \hat{z}_{j,n}) \tag{6.27}$$

Let

$$\hat{z}^*_{j,n} \equiv (\hat{z}_{*,n} - \hat{z}_{j,n}) = H^*_{j,n}\, \hat{x}^*_{j,n} \tag{6.28}$$

where

$$H^*_{j,n} \equiv (H_{*,n}\,,\; H_{j,n}) \tag{6.29}$$

and

$$\hat{x}^*_{j,n} \equiv (\hat{x}^T_{*,n},\; \hat{x}^T_{j,n})^T \tag{6.30}$$

For each $j \in K'$ we have

$$\hat{x}_{j,n+1} = F_{j,n}(I - K_{j,n}H_{j,n})\,\hat{x}_{j,n} + G_{j,n}\,u_n + F_{j,n}\,K_{j,n}\,H_{*,n}\,x_n + F_{j,n}\,K_{j,n}\,v_n \qquad (6.31)$$

where

$$K_{j,n} \equiv \Sigma_{j,n}\,H_{j,n}^T\,(H_{j,n}\,\Sigma_{j,n}\,H_{j,n}^T + R_{j,n})^{-1}$$

Let

$$\bar{\hat{x}}_{j,n+1} \equiv E_*\hat{x}_{j,n+1} = F_{j,n}(I - K_{j,n}\,H_{j,n})\bar{\hat{x}}_{j,n} + G_{j,n}\,u_n + F_{j,n}\,K_{j,n}\,H_{*,n}\,\bar{x}_{*,n}$$

where

$$\bar{x}_{*,n} \equiv E_*x_n = E_*E_*^{U_{n-1}}x_n = E_*\hat{x}_{*,n} = \bar{\hat{x}}_{*,n}$$

Also let

$$\tilde{\hat{x}}_{j,n+1} \equiv \hat{x}_{j,n+1} - \bar{\hat{x}}_{j,n+1} = F_{j,n}(I - K_{j,n}\,H_{j,n})\tilde{\hat{x}}_{j,n} + F_{j,n}\,K_{j,n}\,H_{*,n}(x_n - \bar{x}_{*,n}) + F_{j,n}\,K_{j,n}\,v_n$$

and

$$\tilde{x}_{*,n+1} \equiv x_{n+1} - \bar{x}_{*,n+1} = F_{*,n}\,\tilde{x}_{*,n} + J_{*,n}\,w_n$$

Now let

$$\tilde{\hat{z}}^{*}_{j,n} \equiv H^{*}_{j,n}\, \tilde{\hat{x}}^{*}_{j,n} \tag{6.32}$$

where

$$\tilde{\hat{x}}^{*}_{j,n} \equiv (\tilde{\hat{x}}^{T}_{*,n}\,,\ \tilde{\hat{x}}^{T}_{j,n})^{T} \tag{6.33}$$

and let

$$\bar{\hat{z}}^{*}_{j,n} \equiv H^{*}_{j,n}\, \bar{\hat{x}}^{*}_{j,n} \tag{6.34}$$

where

$$\bar{\hat{x}}^{*}_{j,n} \equiv (\bar{\hat{x}}^{T}_{*,n},\ \bar{\hat{x}}^{T}_{j,n})^{T} \tag{6.35}$$

Then we can write

$$I''_{n}\,(*;j) = \frac{1}{2}\,(\tilde{\hat{z}}^{*}_{j,n} + \bar{\hat{z}}^{*}_{j,n})^{T}\, \Sigma^{-1}_{j,n} (\tilde{\hat{z}}^{*}_{j,n} + \bar{\hat{z}}^{*}_{j,n})$$

$$= \frac{1}{2}\, \tilde{\hat{z}}^{*T}_{j,n}\, \Sigma^{-1}_{j,n}\, \tilde{\hat{z}}^{*}_{j,n} + \frac{1}{2}\, \bar{\hat{z}}^{*T}_{j,n}\, \Sigma^{-1}_{j,n} \bar{\hat{z}}^{*}_{j,n} + \tilde{\hat{z}}^{*T}_{j,n}\, \Sigma^{-1}_{j,n} \bar{\hat{z}}^{*}_{j,n}$$

Let

$$\Phi^{*}_{j,n} \equiv \tilde{\hat{z}}^{*}_{j,n}\, \tilde{\hat{z}}^{*T}_{j,n} \tag{6.36}$$

$$\nabla^{*}_{j,n} \equiv \bar{\hat{z}}^{*}_{j,n}\, \bar{\hat{z}}^{*T}_{j,n} \tag{6.37}$$

and

$$\Theta^{*}_{j,n} \equiv \tilde{\hat{z}}^{*}_{j,n}\, \bar{\hat{z}}^{*T}_{j,n} \tag{6.38}$$

Then we have

$$I_n^{"}(*;j) = \frac{1}{2} \operatorname{tr} \Sigma_{j,n}^{-1} (\Phi_{j,n}^{*} + \nabla_{j,n}^{*} + \Theta_{j,n}^{*}) \tag{6.39}$$

We shall use

$$\overline{\Phi}_{j,n}^{*} \equiv E_{*} \Phi_{j,n}^{*}$$

which is obtained via the following procedure.

Define

$$\tilde{x}_{j,n}^{*} \equiv (\tilde{x}_{*,n}, \tilde{\hat{x}}_{*,n}, \tilde{\hat{x}}_{j,n}) \tag{6.40}$$

Then $\tilde{x}_{j,n}^{*}$ is generated by the following equation

$$\tilde{x}_{j,n+1}^{*} = F_{j,n}^{*} \tilde{x}_{j,n}^{*} + G_{j}^{*} w_n \tag{6.41}$$

where

$$F_{j,n}^{*} \equiv \begin{bmatrix} F_{*,n} & 0 & 0 \\ F_{*,n}K_{*,n}H_{*,n} & F_{*,n}(I - K_{*,n}H_{*,n}) & 0 \\ F_{j,n}K_{j,n}H_{*,n} & 0 & F_{j,n}(I - K_{j,n}H_{h,n}) \end{bmatrix} \tag{6.42}$$

$$G_{j,n}^{*} \equiv \begin{bmatrix} J_{*,n} & 0 \\ 0 & F_{*,n}K_{*,n} \\ 0 & F_{j,n}K_{j,n} \end{bmatrix} \qquad W_n \equiv \begin{bmatrix} w_n \\ v_n \end{bmatrix} \tag{6.43}$$

Also let

$$Q_n^* \equiv \begin{bmatrix} Q_{*,n} & 0 \\ 0 & R_{*,n} \end{bmatrix} \tag{6.44}$$

Then

$$\overline{\Phi}_{j,n}^* = (\underline{0},\ H_{j,n}^*)\Pi_{j,n}^* (\underline{0},\ H_{j,n}^*)^T \tag{6.45}$$

where

$$\Pi_{j,n}^* \equiv E_* \left\{ \tilde{x}_{j,n}^* \tilde{x}_{j,n}^{*T} \right\} \tag{6.46}$$

is generated by the equation

$$\Pi_{j,n+1}^* = F_{j,n}^* \Pi_{j,n}^* F_{j,n}^{*T} + G_{j,n}^* Q_n^* G_{j,n}^{*T} \tag{6.47}$$

initialized at

$$\Pi_j^* = \begin{bmatrix} \Psi_* & 0 & 0 \\ 0 & \Psi_* & 0 \\ 0 & 0 & \Psi_j \end{bmatrix}$$

Next consider $\nabla_{j,n}^*$. We have

$$\overrightarrow{x}_{j,n+1}^* = \overrightarrow{F}_{j,n}^* \overrightarrow{x}_{j,n}^* + \overrightarrow{G}_{j,n}^* u_n \ ; \ \overrightarrow{x}_{j,o}^* = 0 \tag{6.48}$$

where

$$\vec{F}_{j,n} \equiv \begin{bmatrix} F_{*,n} & 0 \\ F_{j,n}K_{j,n}H_{*,n} & F_{j,n}(I - K_{j,n}H_{j,n}) \end{bmatrix}; \quad \vec{G}_{j,n} \equiv \begin{bmatrix} G_{*,n} \\ G_{j,n} \end{bmatrix} \tag{6.49}$$

(6.48) can be written as

$$\vec{x}_{j,n} = \sum_{m=1}^{n-1} A_j^*(n,m)\, \vec{G}_{j,m}\, u_m \tag{6.50}$$

where

$$A_j^*(n,m) \equiv \prod_{i=1}^{n-m-1} \vec{F}_{j,i} \tag{6.51}$$

We have

$$E_* \Theta_{j,m} = 0$$

Thus, for $k = *$ and each $j \in K$; $j \neq k$

$$\bar{I}_n(k;j) = \bar{I}_n'(k;j) + \bar{I}_n''(k;j)$$

$$= \frac{1}{2} \log \frac{|\Sigma_{j,n}|}{|\Sigma_{k,n}|} + \frac{1}{2} \operatorname{tr} (\Sigma_{k,n} \Sigma_{j.n}^{-1} - I)$$

$$+ \frac{1}{2} \operatorname{tr} \Sigma_{j,n}^{-1} (\bar{\Phi}_{j,n}^k + \nabla_{j,n}^k)$$

Note that while $\Sigma_{j,n}$ and $\Sigma_{k,n}$ (or $I_n'(k;j)$) depend only on the stochastic part of the input the term $\nabla^k_{j,n}$ represents its' deterministic part. In the sequel we examine the separate contributions of these elements to the L_1 convergence rates of the likelihood ratios and the *a posteriori* probability ratios on the set K.

Theorem 6.2

Suppose that the true system (6.25) belongs to the set M_3 given by (6.26). Let $k \in K$ be the true parameter. Suppose that for each $j \in K$; $j \neq k$ there exist a positive scalar α_j and a positive integer N_j such that

$$||\Sigma_{k,n} - \Sigma_{j,n}|| \geq \alpha_j \quad \text{for all } n \geq N_j \tag{6.52}$$

Then we have for each $j \in K$; $j \neq k$

$$E_* h_j^k(Z^n) \geq e^{(n-N_j+1)\alpha_j} \quad \text{for all } n \geq N_j$$

and

$$E_* \frac{f^b(k|Z^n)}{f^b(j|Z^n)} = \frac{f^b(k)}{f^b(j)} \; e^{(n-N_j+1)\alpha_j} \quad \text{for all } n \geq N_j$$

Proof

The proof follows from arguments similar to those made in the proof

of lemma 6.1. We first show that (6.52) implies that there exists some $\varepsilon > 0$ such that

$$|\lambda_{n,i} - 1| \geq \varepsilon \quad \text{for each } i=1,\ldots,\ell, \text{ for all } n \geq N_j \tag{6.53}$$

where $\lambda_{n,i}$; $i=1,\ldots,\ell$ are the solutions of (6.13). Suppose that (6.53) does not hold, then for any $\varepsilon > 0$ there exists some $n_\varepsilon \geq N_j$ such that

$$|\lambda_{n_\varepsilon,i_{n_\varepsilon}} - 1| < \varepsilon \tag{6.54}$$

where

$$\lambda_{n_\varepsilon,i_{n_\varepsilon}} \equiv \min\left\{\lambda_{n_\varepsilon,i} \; ; \; i=1,\ldots,\ell\right\}$$

and then, by continuity of the left hand side of (6.13) in λ_n, given $\alpha_j > 0$ one can take ε such that

$$||\Sigma_{k,n_\varepsilon} - \lambda_{n,i_{n_\varepsilon}} \Sigma_{j,n}| - |\Sigma_{k,n} - \Sigma_{j,n}|| < \alpha_j$$

yielding

$$||\Sigma_{k,n} - \Sigma_{j,n}|| < \alpha_j$$

contradicting (6.52). Hence (6.52) implies (6.53). Now by (6.16) and by the convexity of (6.20) we have that (6.53) implies that for each $j \in K$; $j \neq k$ there exists some $\alpha_j > 0$ such that

$$I_n'(k;j) \geq \alpha_j \quad \text{for all } n \geq N_j$$

Since $\bar{I}_n''(k;j) \geq 0$ we then have

$$\bar{I}_n(k;j) \geq \alpha_j \quad \text{for all } n \geq N_j$$

Condition (c4.5) is then satisfied and the assertion follows from equations (4.11) and (4.12) in the proof of theorem 4.11. ■

Corollary 6.1

Let the set M_2 be time invariant and let the true system belong to M_2. Suppose that for each $j \in K$ Σ_j given by (6.23) is finite and non-singular. Then the L_1-convergence bounds asserted in theorem (6.2) holds under condition (c6.2), where k is the true parameter.

Proof

Condition (c6.2) implies that for each $j \in K$; $j \neq k$ there exists some $\zeta_j > 0$ such that

$$||\Sigma_j - \Sigma_k|| \geq \zeta_j$$

clearly,

$$\lim_{n\to\infty} ||\Sigma_{j,n} - \Sigma_{k,n}| = ||\Sigma_j - \Sigma_k|| \geq \zeta_j$$

Hence, for any positive scalar α_j such that $0 < \alpha_j < \zeta_j$ there exists some positive integer N_j such that

$$||\Sigma_{j,n} - \Sigma_{k,n}|| \geq \alpha_j \quad \text{for all } n \geq N_j$$

The assertion then follows from theorem 6.2. ∎

Theorem 6.3

Suppose that the true system (6.25) belongs to the set M_3 given by (6.26). Let $k \in K$ be the true parameter. Suppose that for each $j \in K$; $j \neq k$ there exists a positive scalar α_j and a positive integer N_j such that

$$\text{tr } \Sigma_{j,n}^{-1} \nabla_{j,n}^{k} \geq 2\alpha_j \quad \text{for all } n \geq N_j \tag{6.55}$$

Then the L_1 convergence rates asserted in theorem 6.2 hold.

Proof

For each $j \in K$; $j \neq k$ we have

$$I_n'(k;j) \geq 0 \quad \text{for all } n \geq 0$$

(It follows from (3.5). Also see the proof of lemma 6.1) and

$$\text{tr } \Sigma_{j,n}^{-1} \Phi_{j,n}^{k} \geq 0 \quad \text{for all } n \geq 0$$

Hence

$$\overline{I}_n(k;j) \geq \frac{1}{2} \operatorname{tr} \Sigma_{j,n}^{-1} \nabla_{j,n}^{k} \geq \alpha_j \quad \text{for all } n \geq N_j$$

Condition (c4.5) is then satisfied and the assertion follows from equations (4.11) and (4.12) in the proof of theorem 4.11.■

Theorem 6.2 guarantees a certain L_1 convergence rate of the likelihood ratios and the *a posteriori* probability ratios under a certain condition involving the stochastic characteristics of the inputs to the systems. Theorem 6.3 means that the convergence rates can be improved by application of certain deterministic inputs, satisfying (6.55).

CHAPTER VII

SUGGESTIONS FOR FURTHER RESEARCH

7.1 Extension to Compact Parameter Sets

As mentioned in Chapter 1, the extension of parameter estimation convergence results from finite to infinite sets can, in general, be obtained via the addition of topological conditions on the parameter set. Let S be a compact metric space with metric δ. In the previous sections we have studied conditions under which one has for some $r \in S$.

$$(c7.1) \quad \lim_{n\to\infty} h_r^s(Z^n) = 0 \quad \text{a.e. for each } s \in S;\ s \neq r \tag{7.1}$$

We have seen in Chapter 4 that if the true parameter is a member of the parameter set, say, $* = r \in S$, then (c7.1) is implied by the following conditions

$$(c7.2) \quad \lim_{n\to\infty} \sum_{m=1}^{n} I_m(r;s) = \infty \quad \text{a.e. for each } s \in S;\ s \neq r$$

and

$$\limsup \sum_{m=1}^{n} J_m(r;s) > -\infty \quad \text{a.e. for each } s \in S;\ s \neq r$$

The pointwise convergence in (7.1) is not sufficient for convergence a.e. of, say, the ML estimates on S to r (although mistakenly considered to be by several authors).

To obtain convergence a.e. of the estimates to r it must be shown that for any open neighborhood V(r) of r one has

$$\lim_{n\to\infty} \sup_{s \in V^c(r)} h_r^s (Z^n) = 0 \quad \text{a.e.} \tag{7.2}$$

where $V^c(r)$ is the complement of V(r) in S. Consider the following condition.

(c7.3) At each $s \in S$ the ratios $h_r^s (Z^n)$ are continuous in s uniformly in n. This means that for any realization of the sequence (z_n) given $\varepsilon > 0$ there exists for each $s \in S$ a neighborhood

$$V(s) = \left\{ t : |t - s| < \delta_s \right\} \tag{7.3}$$

for some $\delta_s > 0$, such that

$$\sup_{t \in V(s)} |h_r^t (Z^n) - h_r^s (Z^n)| < \varepsilon \quad \text{for all } n \geq 0 \tag{7.4}$$

Theorem 7.1

Suppose that conditions (c7.1) and (c7.3) hold, then ML estimates on S converge a.e. to the parameter r.

Proof

Choose $\varepsilon < 1$. Then for each $s \in V^c(r)$ there exists an open neighborhood V(s) satisfying (7.3) and (7.4). Since V(r) is open, $V^c(r)$ is a closed subset of a compact set, hence, compact. Thus, there exists a

finite number of points s_i, $i \in I = (1,\ldots,q)$ such that

$$V^c(r) \subset \cup_i \left\{ V(s_i) \; ; \; i \in I \right\}$$

Now

$$\limsup_{n\to\infty} \sup_{s \in V^c(r)} h_r^s(Z^n) \leq \lim_{n\to\infty} \max_i \left\{ \sup_{t \in V(s_i)} h_r^t(Z^n) \; ; \; i \in I \right\}$$

$$\leq \lim_{n\to\infty} \max_i \left\{ h_{s_i}(Z^n) + \varepsilon \; ; \; i \in I \right\}$$

$$= \max_i \left\{ \lim_{n\to\infty} \left[h_{s_i}(Z^n) + \varepsilon \right] ; \; i \in I \right\}$$

$$= \varepsilon < 1 \qquad \text{a.e.}$$

But since

$$\limsup_{n\to\infty} \sup_{s \in S} h_r^s(Z^n) \geq \lim_{n\to\infty} h_r^r(Z^n) = 1$$

the ML estimates on S converge a.e. to r. ■

The proof of convergence a.e. of MAP estimates on S to r is similar, as by (2.11) we have

$$\frac{f^b(s|Z^n)}{f^b(r|Z^n)} = \frac{f^b(s)}{f^b(r)} \; h_r^s(Z^n)$$

Condition (c7.3) and its applicability to cases of interest are suggested for further research. Two guiding questions seem to be:

1) When can (c7.3) be replaced by continuity conditions on the conditional ratios $h_r^s(z_n|Z^{n-1})$?

2) How can (c7.3) be relaxed and still provide the transition from (7.1) to (7.4)?

7.2 Existence and Uniqueness

Aström and Söderström [1974], considering the identification of the parameters of stationary Gaussian ARMA processes, presented the problem of consistency of the ML estimate as a problem of existence of a unique maximum over $s \in S$ of the scalar function $\lim_{n\to\infty} f_s(Z^n)$. An equivalent problem for state space models was posed in section 5.3.1 as the existence of a unique minimum of the scalar function L_s^*, defined by (5.40), on S. A related problem is the existence of a unique minmax point of the scalar function L_s^t, defined by (5.42), for the solution of the modeling problem proposed in section 5.3.2.

The existence and uniqueness problem has also been treated in the literature in terms of the parameters of certain realizations of the system to be identified. Caines [1975b] has proposed the condition that there exist a homeomorphism between the parameter set and the set of impulse responses of the system's innovations representations for the identifiability of stationary linear systems. Similar conditions were suggested by Tse and Weinert [1975] and, for the finite parameter set case, by Moore and Hawkes [1974]. The advantages of statistical uniqueness conditions such as the one suggested by Baram and Sandell [1976] and in this thesis is that they apply to any given set of state

space models and not to certain canonical representations of the system, and they are verifiable by standard computations (such as the steady-state solutions of Riccati and Lyapunov equations). Their disadvantage is that the actual parametrization of the system gets lost in the statistical conditions. The homeomorphism condition presented above seems to correspond to conditions (c7.2), which requires uniqueness, and (c7.3) which requires continuity, put together. More elaborate investigation of the correspondence between these conditions is suggested for future research. The finite parameter set case should be addressed first.

7.3 Identifiability by Deterministic Inputs

The application of deterministic inputs to dynamic systems for the purpose of identification and their optimal selection have been addressed by several authors (Levadi [1966], Gagliardi [1967], Nahi and Wallis [1969], Aoki and Staley [1970], Mehra [1972], Goodwin, et al [1973], Lopez-Toledo and Athans [1975]). The analysis of section 6.3 suggests a new approach to the problem. It follows from theorem 6.3 that any input sequence that satisfies (6.3) will provide convergence in the mean of the identification procedures at a certain rate. The condition in (6.3) also involves the system's coefficients and thus, the selected deterministic input sequence will obviously depend on the nature of the system under consideration. The problem can then be presented as follows. Under what conditions on the true system generating the observations and on the model set will the identification procedures converge to a

model in the set using some input sequence, and what class of input sequences will then provide identifiability?

7.4 Other Application Areas

In Chapters 5 and 6 we have applied the general theory derived in Chapters 3 and 4 to certain aspects of linear system identification and modeling. Further investigation of modeling aspects has been suggested in remarks 1 and 2 in section 5.4. Other general areas of application which have not been specifically addressed in this thesis are:

1) Application to certain classes of time varying systems, such as periodically varying linear systems.
2) Application to non-linear system identification problems.
3) Application to signal detection problems in communication systems.

REFERENCES

Akaike, H. [1972]: "Information Theory and an Extension of the Maximum Likelihood Principle," Proc. of the 2'nd International Symposium on Information Theory, Budapest.

Akaike, H. [1974]: "A New Look at the Statistical Model Identification," IEEE Trans. on Automatic Control, Vol. AC-19, N° 6, Dec.

Anderson, T. W. [1958]: An Introduction to Multivariate Statistical Analysis, Wiley, New York.

Aoki, M. and Staley, R. [1970]: "On Input Signal Synthesis in Parameter Identification," Automatica, Vol. 6, pp. 431-440.

Åström, K. J. and Söderström, T. [1974]: "Uniqueness of the Maximum Likelihood Estimates of the Parameters of an ARMA Model," IEEE Trans. on Automatic Control, Vol. AC-19, N° 6, December.

Baram, Y. and Sandell, N. R., Jr., [1976]: "Consistent Estimation on Finite Parameter Sets with Application to Dynamic System Identification". Submitted to IEEE Trans. on Automatic Control, July.

Bauer, H. [1972]: Probability Theory and Elements of Measure Theory, Holt, Rinehart and Winston, Inc.

Berger, T. [1971]: Rate Distortion Theory, Prentice-Hall, Inc.

Bhattacharyya, A. [1943]: "On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions," Bull. Calcutta Math. Soc., Vol. 35, pp. 99-109.

Billingsley, P. [1961]: "The Lindberg-Lévy Theorem for Martingales," Am. Math. Soc. Proc., Vol. 12, pp. 788-792.

Caines, P. E. [1975a]: "A Note on the Consistency of Maximum Likelihood Estimates for Finite Families of Stochastic Processes," Ann. Statistics, Vol. 3, N° 2, pp. 539-546.

Caines, P. E. [1975b]: "Prediction Error Identification Methods for Stationary Stochastic Processes," IEEE Trans. on Automatic Control, Vol. AC-21, N° 4, August.

Caines, P. E. and Rissanen, J. [1974]: "Maximum Likelihood Estimation of Parameters in Multivariate Gaussian Stochastic Processes," IEEE Trans. on Information Theory, Vol. IT-20, N° 1, January.

Chacon, R. V. and Ornstein, W. [1960]: "A General Ergodic Theorem," J. of Math., Vol. 4, pp. 153-160.

Chung, K. L. [1974]: A Course in Probability Theory, Academic Press, New York.

Cramer, H. [1946]: Mathematical Methods of Statistics, Princeton University Press.

Doob, J. L. [1934]: "Probability and Statistics," Trans. Amer. Math. Society, Vol. 36.

Doob, J. L. [1953]: Stochastic Processes, John Wiley and Sons, Inc., New York.

Edgeworth, F. Y. [1908]: "On the Probable Error of Frequency Constants," J. Royal Stat. Soc., London, Vol. 71, 72.

Fano, R. M. [1966]: Transmission of Information, Wiley, New York.

Fisher, R. A. [1922]: "On the Mathematical Foundation of Theoretical Statistics," Phil. Trans. Royal Soc. London, Ser. A, Vol. 222, pp. 309-368.

Fisher, R. A. [1956]: Statistical Methods and Scientific Inference, Oliver and Boyd, London.

Gagliardi, R. M. [1967]: "Input Selection for Parameter Identification in Discrete Systems," IEEE Trans. on A tomatic Control, Vol. AC-12, October.

Gallager, R. G. [1968]: Information Theory and Reliable Communication, Wiley, New York.

Gauss, C. F. [1809]: "Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium," Hamburg.

Gernander, U. [1950]: "Stochastic Processes and Statistical Inference," Arkiv für Mathematic, Vol. 1, pp. 195-277.

Goodwin, G., Murdock, J. and Payne, R. [1973]: "Optimal Test Signal Design for Linear SISO System Identification," Int. J. Cont., Vol. 17, pp. 45-55.

Gurland, J. [1954]: "On Regularity Conditions for Maximum Likelihood Estimators," Skand. Aktuar. Tidskr., Vol. 37, pp. 71-76.

Halmos, P. R. [1956]: Lectures on Ergodic Theory, Chelsea Publ. Co., New York.

Hawkes, R. M. and Moore, J. B. [1976]: "Performance Bounds for Adaptive Estimation," Proc. IEEE, August.

Jeffreys, H. [1946]: "An Invariant Form for the Prior Probability in Estimation Problems," Proc. Royal Soc., London, Ser. A, Vol. 186, pp. 453-461.

Kailath, T. [1967]: The Divergence and the Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. on Comm. Tech., Vol. COM-15, N° 1, February.

Kullback, S. [1959]: Information Theory and Statistics, John Wiley and Sons, Inc., New York.

Laplace, P. S. [1820]: Theorie Analytique des Probabilities, 3rd Ed., Paris, pp. 309-354.

LeCam, L. [1953]: "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates," Univ. of Calif. Publ. Statistics, Vol. 1, pp. 277-330.

Levadi, V. S. [1966]: "Design of Input Signals for Parameter Estimation," IEEE Trans. on Automatic Control, Vol. AC-11, April.

Liporace, L. A. [1971]: "Variance of Bayes Estimates," IEEE Trans. on Information Theory, Vol. IT-17, N° 6, November.

Ljung, L. [1974a]: "Convergence of Recursive Stochastic Algorithms," Report 7403, Div. of Auto. Cont., Lund Inst. of Tech., February.

Ljung, L. [1974b]: "On Consistency for Prediction Error Identification Methods," Report 7405, Div. of Auto. Cont., Lund Inst. of Tech., March.

Ljung, L. [1975]: "On Consistency and Identifiability," Report 7521(c), Div. of Auto. Cont., Lund Inst. of Tech., June. Also Proc. of Symp. Stoch. Sys., Lexington, Kentucky, June 1975. (North-Holland Publ.).

Lopez-Toledo, A. A. and Athans, M. [1975]: "Optimal Policies for Identification of Stochastic Linear Systems," IEEE Trans. on Automatic Control, Vol. AC-20, N° 6, December.

Mehra, R. [1972]: "Optimal Inputs for Linear System Identification," JACC Conf., pp. 811-820.

Meyer, P. A. [1966]: Probability and Potentials, Blaisdell Publishing Co.

Moore, J. B. and Hawkes, R. M. [1974]: "Performance of Bayesian Parameter Estimators for Linear Signal Models," Report EE7410, Univ. of Newcastle, Australia, July.

Nahi, N. and Wallis, D. [1969]: "Optimal Inputs for Parameter Estimation in Dynamic Systems with White Observation Noise," Proc. JACC Conf., pp. 506-512.

Rissanen, J. and Caines, P. E. [1974]: "Consistency of Maximum Likelihood Estimators for ARMA Processes," Control Systems Rpt. N° 7424, Dept. of EE, Univ. of Toronto, December.

Roussas, G. G. [1965]: "Extension to Markov Processes of a Result by A. Wald about the Consistency of Maximum Likelihood Estimates," Z. Wahrscheinlichkeitstheorie, Vol. 4, pp. 69-73.

Rudin, W., [1966]: Real and Complex Analysis, McGraw-Hill, New York.

Savage, L. J. [1954]: The Foundations of Statistics, John Wiley and Sons, New York.

Stout, W. F. [1974]: Almost Sure Convergence, Academic Press, New York.

Tse, E. and Weinert, H. L. [1975]: "Structure Determination and Parameter Identification for Multivariable Stochastic Linear Systems," IEEE Trans. on Automatic Control, Vol. AC-20, N° 5, October.

Tse, E. [1976]: "Identification and Estimation of Dynamic Linear Models with Unknown Structure," Proc. JACC Conf., Purdue University.

Wald, A. [1949]: "Note on the Consistency of the Maximum Likelihood Estimate," Ann. Mat. Statis., Vol. 20, pp. 595-601.

Weiner, N. [1949]: Extrapolation, Interpolation and Smoothing of Stationary Time Series, The M.I.T. Press, Cambridge, Mass.

Woodward, P. K. [1953]: Probability and Information Theory with Application to Radar, McGraw-Hill, New York.