AFOSR-TR-90-0289

(2)

FINAL REPORT

ADAPTIVE TIME SERIES ANALYSIS

USING PREDICTIVE INFERENCE AND ENTROPY

Approved for public release;
distribution unlimited

AD-A222 337

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DTIC
This technical report has been reviewed and is
unlimited release IAW AFR 190-12.
MATTHEW J. KERPER
Chief, Technical Information Division

FEBRUARY 1990

DR. RAMAN K. MEHRA
DR. SHAH MAHMOOD

SCIENTIFIC SYSTEMS, INC.
500 W. CUMMINGS PARK
SUITE 3950
WOBURN, MA 01801

PREPARED FOR:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
BOLLING AIR FORCE BASE
WASHINGTON, D.C. 20332-6448

UNDER CONTRACT NO. F49620-87-C-0026

DTIC
ELECTE
MAY 30 1990
S B D

90 05 25 120

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) -	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR-90-0289
6a. NAME OF PERFORMING ORGANIZATION (6b. OFFICE SYMBOL (If applicable)) Scientific Systems, Inc.	7a. NAME OF MONITORING ORGANIZATION AFOSR/NM
6c. ADDRESS (City, State, and ZIP Code) 500 W. Cummings Park Woburn, MA 01801	7b. ADDRESS (City, State, and ZIP Code) AFOSR/NM Bldg 410 Bolling AFB DC 20332-6448
8a. NAME OF FUNDING / SPONSORING ORGANIZATION (8b. OFFICE SYMBOL (If applicable)) USAF Office of Scient. Res. NM	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F49620-87-C-0026
8c. ADDRESS (City, State, and ZIP Code) Building 410 Bollling AFB, DC 20332-6448	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61105 / PROJECT NO. 3005 / TASK NO. A1 / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

Adaptive Time Series Analysis Using Predictive Inference and Entropy

12. PERSONAL AUTHOR(S)
Dr. Raman K. Mehra and Dr. Shah Mahmood

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM 12/86 TO 10/89	90/2/13	

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

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

Research is reported on adaptive time series methods for detecting and tracking both abrupt and slow changes in both structure and parameters of dynamic systems. The methods are based on a unified statistical framework which is motivated by statistical inferences and entropy arguments. The method yields estimates of multivariate input/output dynamics and noise statistics. It also gives estimate of system order that is optimal in the sense of an information theoretic criterion. The integrated approach is known as CVA-AIC. Many theoretical issues have been explored under the scope of this project. The relationship between this technique and another powerful framework for estimation known as E-M algorithmic approach has been established. If the CVA-AIC technique is embedded properly in an E-M framework, it leads to maximum likelihood estimates and recursive algorithms for system identification.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL DR. JON A. SJOGREN	22b. TELEPHONE (Include Area Code) (202) 767-4940
	22c. OFFICE SYMBOL NM

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

ACKNOWLEDGEMENT

This is the final report on the Project "Adaptive Time Series Analysis Using Predictive Inference and Entropy." This is a Phase II SBIR project sponsored by the Air force Office of Scientific Research under Contract F49620-87-C-0026. The work was performed by researchers from Scientific Systems, Inc., Woburn, MA 01801. During the progress of the project, Dr. Donald E. Gustafson and Dr. Wallace E. Larimore made technical contributions to this project. Their contributions to the research effort are gratefully acknowledged.

DTIC COPY INSPECTED

Accession For
NTIS GRA&I ☑
DTIC TAB ☐
Unannounced ☐
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

ABSTRACT

The problem of Adaptive Time Series Analysis and Systen Identification is a very difficult one,particularly in an environment where the system characteristics are changing with time or the system is in a closed-loop configuration, such as the problem of Adaptive Flutter Suppression and Control of Large Space Structures. The main objective of this project was to investigate the theoretical issues related to a relatively new system identification technique, known as Canonical Variate Analysis (CVA), that has been successfully used in such environments. In this technique, a Markov Model is extracted from an observed data set on the basis of stochastic realization theory and statistical correlation analysis. The optimal model order is automatically selected using an information theoretic criterion known as Akaike Information Criteria (AIC). The overall technique is known as CVA-AIC System Identification Technique.

The CVA-AIC technique has been in use for some time primarily as an ad-hoc scheme. Although the underlying theories such as the Stochastic Realization Theory and Canonical Correlation Analysis are rigorously established, various aspect of CVA-AIC technique itself have not been based on rigorous theoretical justifications. For this reason, the overall effort of this project was devoted

towards theoretical understanding and relationship to other identification techniques. In particular, the relationship of the CVA-AIC Technique to Maximum Likelihood Identification using the E-M algorithmic approach has been investigated. It has been shown how Maximum Likelihood Estimates can be obtained starting from the CVA-AIC solution. The E-M Algorithmic approach also suggests real-time recursive algorithms.

The report starts with a clear and theoretically elegant interpretation of AIC in a discrete valued random variable framework. The relationship between the state-space model obtained by the CVA-AIC technique and the standard Kalman Filter form has been explored. In many situations, models of all orders are needed and, therefore, a simple algorithm that is recursive in the model order has been developed without matrix inversion at each step. The control engineers often rely on the confidence bands around the power spectral density function of noise and transfer function of the system. It has been shown how to compute this band at various confidence levels around the true spectral density and transfer function. A major emphasis has been placed upon time varying dynamic systems with slowly varying and abrupt changes. It has been shown that by selecting moving windows, the slowly time varying parameters in the system can be tracked and, by appropriately partitioning the data, the computed AIC from each partition

can be used to detect abrupt changes. Simulated examples have been provided in support of this result. It is further shown how the E-M Algorithmic approach can used to identify time-varying parameters.

1. INTRODUCTION

1.1 Overview of the Adaptive Time Series Analysis Problem

Adaptation in time series is an important problem in a number of DOD systems and has many applications in various commercial industries. This is an especially difficult problem in problems requiring realtime adaptation to process changes since such a procedure would have to be completely automatic and reliable. Adaptation is necessary in systems where the dynamical characteristics change with time in unpredictable ways, or where the noise disturbance process characteristics vary with time. Examples of systems that require adaptive time series analysis are the adaptive suppression of aircraft wing flutter, identification of the dynamics of large flexible space structures, detection of failures in aircraft from subsystem failures of battle damage, identification of missile aerodynamics, target tracking, and various signal processing problems.

The solution to the adaptive time series analysis requires several advances in current time series methods. At the core of the problem is the need for a fundamental statistical approach to the adaptation problem that poses the problem in a meaningful way and that leads to computable solutions. To solve the online adaptation problem, a reliable and automatic time series modeling procedure is required that is lacking in previous methods. The current research provides

- A sound statistical basis for posing and solving the adaptation problem

- A numerically and statistically reliable online computational procedure

This approach has been used in conjunction with a new high resolution system identification method utilizing canonical variate analysis (CVA) for the determination of the dynamics of high order multisensor systems with a small data length (Larimore, 1983b). This algorithm can be implemented on highly parallel processors such as a systolic array. This makes practical the consideration of many different system characteristics to determine the best for modeling the observed sensor data and correlational relationships between the many sensors. The system characteristics that have been successfully determined adaptively are the dynamical state order of the system, the presence of correlated disturbances, the optimal data length to use in tracking a time varying system, and the optimal data interval for detection of an abrupt change or other event in the data.

The CVA time series analysis method has been applied to the design of an adaptive flutter suppression problem for suppressing wing flutter or aero-structural vibration in aircraft. While considerable progress has been made in the problem of adaptation in terms of identification of time series models, adaptive time series methods which can efficiently track and detect time varying processes would further improve the system. In such a system the wing dynamical characteristics can change instantaneously when a wing store is dropped, and the new wing dynamics are unknown and may be unstable resulting in a growing oscillation. If the unstable mode is not detected, accurately identified, and stabilized by control feedback in less

than a second, then the aircraft can lose a wing. The CVA algorithm using entropy methods for deciding model state order are being implemented on a vector array processor which will identify high order systems with dozens of dynamical states and multiple inputs and outputs in fractions of a second. This system has been tested in real-time simulations, and was successfully demonstrated in wind tunnel tests at the NASA Langley Transonic Dynamics Wind Tunnel. It is expected that highly parallel processors such as systolic array processors could result in a speedup of many thousands of times which would be required for some very large scale real time adaptive problems.

1.2 Signal and Fault Detection

A Comprehensive survey of fault detection methods is given by Willsky (1976). See also Mehra and Peschon (1971), Willsky and Jones (1974), Willsky (1980), and Isermann (1984). The type of abrupt changes in a system that are considered are of the form

$$x(t+1) = \Phi x(t) + Gu(t) + w(t) + m(t) \tag{1.1}$$

$$y(t) = Hx(t) + Au(t) + Bw(t) + v(t) + N(t) \tag{1.2}$$

where u is the input vector process, y is the output vector, x is the state vector, and w and v are white noise processes that are independent with covariance matrices Q and R respectively. These white noise processes model the covariance structure of the error in predicting y from u. The abrupt changes are in the form of the time the functions m(t) and n(t) introduced into the state and observation equations. Fault detection is thus the detection of the presence of such nonzero functions.

For various hypothesized forms of the functions, i.e., for jumps in various components or specific combinations of the components, a particular detection computation is devised which requires implementation of a Kalman filter. This leads to statistically most powerful likelihood ratio tests of the various failure hypotheses. An optimal solution to the failure detection problem formulated in (1.1) and (1.2) is thus obtained.

There are however several more general failure detection problems not of the form of (1.1) and (1.2). The approach permits only the consideration of simple hypotheses, i.e., where the failure functions m(t) and n(t) are of the form of an unknown scalar amplitude parameter multiplying a function of known form. More general functional forms such as two components with different unknown amplitude parameters multiplying the known functions requires maximum likelihood parameter identification at considerable computational expense and loss of numerical reliability. Furthermore, the problem of unknown failure time leads to a considerable increase in the required computation, and no theoretically sound decision procedure has been proposed for choosing the failure time.

The general case of changes in the system dynamics or correlational characteristics of the disturbance or measurement noise processes cannot be handled. Such cases require general time series analysis parameter identification methods which are not reliable for online application to high state order multivariable systems as discussed in Section Multisensor System Identification. Isermann (1984) gives a survey of current fault detection methods and concludes that: "A unique calculation of the process

coefficients and a parameter estimation with high precision is only possible for low order elements between measured variables. Therefore the measured variables should be selected such that the process is divided in first order elements or, in other words, all state variables should be measurable. Easy to implement parameter estimation methods for continuous-time modles to be used on-line, real-time and in closed loop need to be developed." The requirement of measuring all of the states is not realistic in most situations especially in general multivariate time series and system identification problems. Fortunately, the CVA system identification method does not require this, but indeed is an online, real-time method that gives the same accuracy in either open or closed loop.

The issues of adaptation are not addressed in the fault detection literaure except in simplistic ways. The present state of the art in adaptation for failure detection appears to be the work of Hagglund (1983) discussed in the next section, and is just beginning of adaptive approaches which consider fundamental issues in adaptation.

1.3 Adaptation to Changing Processes

Concepts of adaptive systems have been around since the 1950's involving various senses of adaptation. The present literature on the subject includes a number of methods such as recursive computational schemes, exponential forgetting, lattice computational methods, etc., which have certain "knobs" that allow tuning of the algorithm to accommodate changes in the characteristics of the actual processes. Reviews of these and related methods are contained in several recent special issues of technical

journals and books (Special Issue on Adaptive Control, Automatica, Vol. 20, No. 5, 1985; Special Issue on Linear Adaptive Filtering, IEEE Trans. on Information Theory, Vol. 30, No 2, 1984; Honing and Messerschmitt, 1984). While these methods do permit some degree of adaptation to process changes, the methods of adaptation are ad hoc, and no sound underlying statistical principle for adaptation is proposed or demonstrated. As might be expected, these methods can work poorly on certain cases because of the lack of a sound statistical basis.

In particular, the recursive prediction error and lattice methods are convenient due to their recursive form and provide an estimate at every observation (Friedlander, 1982a, 1982b, 1983; Ljung and Soderstrom, 1983). Also, the recursive algorithms can be sued for adaptation by exponential weighting of the past data (Wellstead and Sanoff, 1981; Irving, 1979; Evans and Betz, 1982). But the rational for exponential weighting has not been given a sound fundamental justification, but is used largely due to its ease of use. the choice of the exponential weight has been ad hoc and susceptible to misinterpretation of changing noise variance levels as time varying changes in the dynamics (Hagglund, 1983).

The fundamental problem in adaptive time series analysis is adaptation to time varying processes. The essential problem is the determination of the characteristics describing the rate at which the process is changing. This problem has received very little in-depth treatment in the literature. Most of the difficulty can be attributed to the discrepancy between the true and assumed uncertainty in the measurements. Adaptive control schemes

are notoriously optimistic about the quality of the parameter estimates because the time varying nature of the process is ignored.

A notable exception is the recent work of Hagglund (1983) which takes an information handling point-of-view. This approach leads to a more realistic appraisal of the accuracy of the parameter estimates and consequently the value of new measurements which become available in time. Two classes of time varying systems are considered:

- Processes with abrupt changes
- Processes with slowly varying changes.

Within each of these classes, changes are considered in the process dynamics and/or noise variance.

For abrupt changes, the fault detection approach is taken. The central idea is to monitor differential changes in the parameter estimates to detect abrupt changes. A new procedure is derived by Hagglund which requires no apriori information and is very sensitive to jumps in the parameters. This procedure is shown to have very good properties in both theory and practice. This works well for parameters of the dynamics as well as those of the noise variances in the simple cases of low order systems.

The problem of slowly varying parameters has plagued many adaptive control schemes. Although the concept of discounting the old data using a forgetting factor has been in use for a long time, the problem of how to

relate this factor to the data has been elusive. The principal proposed by Hagglund is to discount past data in such a way that a constant amount of information would be retained if the parameters were constant. The quantitative measure of the information used is the inverse of the parameter estimation error covariance matrix which is the Fisher information matrix. Theory and simulations show that this works quite well in low order and well conditioned systems. However for high order and multisensor systems with illconditioned parametric structure, the algorithms are not so well behaved.

1.4 Multisensor System Identification

System parameter identification from observed measurements is a crucial part of the adaptive multivariate timeseries analysis problem. It is necessary to adapt anot only to changes in the input to output characteristics of a system, but the correlational characteristics of the disturbance and noise processes must simultaneously be determined. The feasibility of adaptive methods requires first that a reliable online multivariate time series identification procedure be available.

There are several difficulties with currently available methods and software for the identification of system dynamics and noise characteristics. Current methods include the self tuning regulator (STR) (Ljung, 1983; Astrom, 1973; Astrom et al, 1973, 1977), maximum likelihood estimation (MLE) (Mehra and Tyler, 1973; Larimore, 1981a), Box-Jenkins (BJ) methods (Box and Jenkins, 1976), and a variety of heuristic approaches. The current state of the art in both MLE and BJ require that an analyst be

involved in the procedure, and the required number of computational iterations is not bounded. The STR has been applied successfully to simple processes, but is not completely reliable for general processes particularly when multi-input, multi-ouput systems are involved. In addition, the recursive prediction error algorithm used in the STR requires a good initial estimate and so is not suitable for short data where no apriori data is available. The heuristic approaches tend to be special purposes and are rather unreliable in general applications.

Of the current approaches to multivariate time series identification which are high resolution, i.e., make efficient use of the observational information, most use the ARMA (autoregressive moving average) representation for the process. For multi-input multi-output systems this is not a globally well defined parameterization which is a major cause of the difficulties in the present identification methods (Gevers and Wertz, 1982). A consequence is that there is no single parameterization which is numerically well conditioned, and known algorithms can be made to fail for a particular choice of system. The system identification problem is well defined in that the class of models does have best models in a maximum likelihood sense (Larimore, 1981a), but the ARMA parameterization is not unique so that for cases such as pole-zero cancellation there is a whole equivalence class of models with equivalent characteristics. In the sequel this difficulty in parameterization will be resolved by the use of state space models, and stable numerical methods will be described for statistically reliable online identification of multivariable time series.

1.5 Adaptive Time Series Analysis Using Predictive Inference and Entropy

Recently a very general predictive inference approach to statistical modeling has led to a fundamental statistical inference justication of negative entropy as the natural measure of model approximation error (Larimore, 1983a). This development has a number of very attractive features:

- It applies to completely general modeling problems including nonparametric methods.
- It applies exactly to small samples.
- Only the fundamental statistical principlaes of sufficiency and repeated sampling are used.
- It applies to time correlated problems such as time series model identification and tracking.
- Statistical inference can be fundamentally viewed as model approximation.

Early developments in predictive distributions are very old, although modern approaches apparently begin with Jeffreys (1961, p.143) who used a Bayesian approach, as has much of the work following (Atchison and Dunsmore, 1975, preface and p. 39). The approach taken here has been stimulated by Murray (1977, 1979), the work of Akaike (1973) and model structure determination problems (Larimore, 1977a).

1.6 Initial Results Indicating Feasibility

SSI has been in the forefront in developing the CVA and entropy metnods. Here the related projects are discussed along with preliminary results indicating the feasibility of the proposed methods.

The original stochastic realization method of Akaike's (1975) was further developed into a commercial software package for mainframe and mini computers by Mehra (1978) and Mehra and Cameron (1976, 1980). Further generalizations to input output systems along with refinements in computational speed and accuracy were developed by larimore (1983b) and Goodrich and Larimore (1983) leading to the current timeseries analysis and forecasting package, Forecast Master (Trademark of SSI), for the IBM/PC. This package is in widespread use in utilities, banks companies and universities.

This algorithm has been the basis for several studies in online systems identification. The project "Basic Research in Adaptive Model Algorithmic Control" used the online CVA system identification algorithm. In the current study "Reconfiguration Control Strategies", the CVA method along with adaptive tracking and detection methods are being studied. The present theory on adaptation using entropy methods (Larimore, 1985a) was developed under the basic research study "traget Dynamic Modeling" and under the study "Development of Statistical Methods Using Predictive Inference and Entropy" which was Phase I of this proposed Phase II study.

A review of the technology in system identification and adaptive control for adaptive methods applicable to the suppression of aeroelastic

wing vibration (flutter) was done in Larimore and Mehra (1984). This study describes the deficiencies of current methods and suggets the feasibility of CVA and entropy methods for fully adaptive online detection and tracking of wing flutter. In a current study with General Dynamics sponsored by the Air Force Wright Aeronautical Laboratories, CVA has been analyzed extensively in computer simulations, real time tests, and demonstrated wind tunnel teste for adaptive flutter suppression. The ability of CVA to identify very complex flutter dynamics of high state order involving very closely spaced spectral peaks in the presence of correlated wind gust disturbances using short data lenghts demonstrated the consderable statistical accuracy of the method. The online CVA identification algorithm was demonstrated ina wind tunnel test at the NASA Langley Transonic Dynamics Wind Tunnel on a 1/4 scale model of an F-16 aircraft.

1.7 Synopsis of Report

In Section 2, we present a detailed and transparent derivation of an unbiased entropy measure which will be used in the sequel for adaptive estimation. This measure is asymptotically equal to Akaike's AIC criterion. In Section 3, we present a detailed description and derivation of linear least-squares prediction using canonical variates analysis (CVA). Several new forms for these predictors are given. In Section 4, a method for direct determination of the parameters of the Kalman filter in canonical form is given, and is shown to be equivalent to a truncated optimal linear predictor derived using CVA. Section 5 considers the model order selection problem, using an entropy-based approach. The problem of abrupt

change detection using entropy methods is considered in Section 6 and a specific algorithm is derived and tested. In Section 7 we consider the problem of slow change detection, specifically the problem of finding the optimal data length for model fitting when the time series coefficients are slowly varying. An entropy-based algorithm is developed and tested.

2. PREDICTIVE INFERENCE AND ENTROPY

2.1 Introduction

In this section we develop the necessary background for development of adaptive estimation algorithms in the sequel.

The problem under consideration is that of predicting the future evolution of a time series, given some observations of the past. The predictive inference framework may be described as follows.

We assume that the density function of interest is parametrized by a parameter vector $\Theta \in R^m$ and is denoted by $p(x \mid \Theta)$. For the purposes of discrimination between two alternatives Θ_1 and Θ_0 it can be shown (Akaike, 1973) that all necessary information is contained in the likelihood ratio

$$L(x) = \frac{p(x \mid \Theta_1)}{p(x \mid \Theta_0)} \tag{2.1}$$

Thus, the mean amount of information for discrimination when $p(x \mid \Theta_0)$ is the true density is of the form

$$I(\Theta_1, \Theta_0) = \int p(x \mid \Theta_0)\, \phi\left[\frac{p(x \mid \Theta_1)}{p(x \mid \Theta_0)}\right] dx \tag{2.2}$$

where $\phi(.)$ is a properly chosen function. It can be argued using information theoretic arguments (Akaike, 1973) that the only appropriate form is

$$\phi(y) = \log y \tag{2.3}$$

which leads directly to the measure

$$B(\Theta_1, \Theta_0) = \int p(x \mid \Theta_0) \log \left[\frac{p(x \mid \Theta_1)}{p(x)\Theta_0} \right] dx \qquad (2.4)$$

Note that $- B(\Theta_1, \Theta_0)$ is the Kullback-Liebler information for discrimination in favor of Θ_0. It can be easily shown that $B(\Theta_1, \Theta_0) \leq 0$ and equality holds if and only if $p(x \mid \Theta_1) = p(x \mid \Theta_0)$ almost everywhere (Aitchison and Dunsmore, 1975).

Note that $B(\Theta_1, \Theta_0)$ can be written as

$$B(\Theta_1, \Theta_0) = \int p(x \mid \Theta_0) \log p(x \mid \Theta_1) \, dx - \int p(x \mid \Theta_0) \log p(x \mid \Theta_0) \, dx \qquad (2.5)$$

Since Θ_0 represents the true (unknown) parameter, our objective is to find the parameter estimate $\hat{\Theta}$ which maximize $B(\hat{\Theta}, \Theta_0)$. From (2.5), we need only maximize

$$\int p(x \mid \Theta_0) \log p(x \mid \hat{\Theta}) \, dx$$

with respect to $\hat{\Theta}$ to produce our estimate. This estimate maximizes the expected log-likelihood and is thus a maximum - likelihood estimate.

2.2 Preliminaries

In order to present a clear development, we will work in a partitioned sample space. The random variable x is presumed to be in n - dimensional Euclidean space, $x \in R^n$, and R^n is partitioned into s mutually disjoint regions $\Omega_1, \Omega_2, \ldots, \Omega_s$ which cover R^n:

$$\Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_s = R^n$$

$$\Omega_j \cap \Omega_j = 0 \; ; \; i \neq j$$

We then define

$$p_i(\Theta) = \int_{\Omega_i} p(x \mid \Theta)\, dx \qquad (2.6)$$

$$i = 1, 2, \ldots, s$$

We consider two different samples, an informative sample q and a predictive sample r. The informative sample is

$$x_q = \{x_{q1}, x_{q2} \ldots, x_{qn_q}\}$$

which consists of n_q observations of x. The predictive sample is

$$x_r = \{x_{r1}, x_{r2}, \ldots, x_{rn_r}\}$$

consists of n_r observations of x. We assume that n_{qi} of the informative samples fall into Ω_i and that n_{ri} of the predictive sample fall into Ω_i. Then

$$\sum_{i=1}^{s} n_{qi} = n_q$$

$$\sum_{i=1}^{s} n_{ri} = n_r \qquad (2.7)$$

The two samples x_q and x_r are from the true distribution.

Thus we have, approximately, for sufficiently large samples,

$$p_{qi}(\Theta_0) \approx \frac{n_{qi}}{n_q} \tag{2.8}$$

and

$$p_{ri}(\Theta_0) \approx \frac{n_{ri}}{n_r} \tag{2.9}$$

and we assume regularity conditions throughout such that

$$p_q(x \mid \Theta_0) = \lim_{\substack{n_q \to \infty \\ s \to \infty}} p_{qi}(\Theta_0)$$

where

$$\lim_{\substack{s \to \infty \\ x \in \Omega_i}} \Omega_i = x$$

and similarly for $p_r(x \mid \Theta_0)$. The computation of the probabilities associated with the parametrized densities is different. Here we use the definition (2.6) and note that $p_i(\Theta)$ is computable from $p(x \mid \Theta)$ and knowledge of Ω_i. In practice, this computation need not be done, as become clear in the sequel.

2.3 Entropy and Maximum Likelihood Estimation

The first step in our development is to form the maximum-likelihood estimate. This is done by maximuzing (2.5) on the informative sample:

$$\hat{\Theta} = \arg \max_{\Theta} B_q(\Theta, \Theta_0)$$

where

$$B_q(\Theta, \Theta_0) = \sum_{i=1}^{s} p_{qi}(\Theta_0) \log \left[\frac{p_{qi}(\Theta)}{p_{qi}(\Theta_0)} \right] \tag{2.10}$$

Thus

$$\sum_{i=1}^{s} p_{qi}(\Theta_0) \left. \frac{\partial \log p_{qi}(\Theta)}{\partial \Theta} \right|_{\hat{\Theta}} = 0 \tag{2.11}$$

We note here that an approximation for $B_q(\Theta_1, \Theta_0)$ is

$$B_q(\Theta, \Theta_0) \approx \sum_{i=1}^{n_q} \log \frac{p(x_i \mid \Theta)}{p(x_i \mid \Theta_0)} \tag{2.12}$$

and the two expressions are asymptotically equal as $n_q \to \infty$. This form was used by Akaike (1973) to derive the AIC criterion.

Solving (2.12) would, in principal, give the maximum-likelihood estimate if the dimension of Θ were known. However, in practice, the actual dimension, m, of θ is not known. Furthermore, there is an obvious tradeoff between the dimension of our estimate $\hat{\Theta}$ and prediction error. Assume $\hat{\Theta} \in R^k$. Then as we increase k, the fit error on the informative sample will decrease momotonically. However, at some point we are in danger of overfitting the model so that $\hat{\Theta}$ is a function of the sampling error on the informative sample. When this happens, the fit errors on the predictive sample will begin to increase.

If we assume that the true parameter vector dimension is m and that the estimated parameter dimension is $k < m$, then our objective is to evaluate the information measure on the predictive sample and select the model which

maximizes this measure. The discrimination measure is now separated into two parts in order to simplify the analysis:

$$B_r(\hat{\theta}^k, \theta_0) = \sum_{i=1}^{s} p_{ri}(\theta_0)\log \frac{p_{ri}(\hat{\theta}^k)}{p_{ri}(\theta_0)}$$

$$= \sum_{i=1}^{s} p_{ri}(\theta_0)\log \frac{p_{ri}(\hat{\theta}^k)}{p_{ri}(\theta^k)} - \sum_{i=1}^{s} p_{ri}(\theta_0)\log \frac{p_{ri}(\theta_0)}{p_{ri}(\theta^k)}$$

$$= B_r(\hat{\theta}^k, \theta^k) - B_r(\theta_0, \theta^k) \qquad (2.13)$$

where $\theta^k \in R^k$. Both entropy measures are measured with respect to the density $p_{ri}(\theta^k)$ and θ^k is arbitrary. We will in the sequel pick θ^k in a particular manner which clarifies and simplifies the development. The decomposition of (2.13) is done to clarify the exposition and to make clear the crucial role played by the number of parameters k. The summations in (2.13) are taken with respect to the true density on the predictive sample while $\hat{\theta}^k$ is the estimate computed on the informative sample. Thus, $B_r(\hat{\theta}^k, \theta_0)$ is a measure of the information between the estimated density and the true density on the predictive sample. Since the informative sample is known but the predictive sample is not we will use statistical mean values in the sequel.

In order to evaluate $B_r(\hat{\theta}^k, \theta^k)$ and $B_r(\theta_0, \theta^k)$ we will expand around the actual probabilities on the informative sample.

Evaluation of $B_r(\hat{\theta}^k, \theta^k)$

From (2.13):

$$B_r(\hat{\theta}^k, \theta^k) = \sum_{i=1}^{s} p_{ri}(\theta_0) \left[\log p_{ri}(\hat{\theta}^k) - \log p_{ri}(\theta^k)\right] \tag{2.14}$$

Define the sampling error between the informative and predictive probabilities as

$$e_i(\theta) = p_{ri}(\theta) - p_{qi}(\theta) \tag{2.15}$$

Expanding the log term to second order yields

$$\log p_{ri}(\hat{\theta}^k) = \log p_{qi}(\theta^k) + \frac{\partial \log p_{qi}(\theta^k)}{\partial p_{qi}} e_i(\theta^k)$$

$$+ \frac{1}{2} \frac{\partial^2 \log p_{qi}(\theta^k)}{\partial p_{qi}^2} e_i^2(\theta^k) + \frac{\partial \log p_{qi}(\theta^k)}{\partial \theta^k} (\hat{\theta}^k - \theta^k)$$

$$+ \frac{1}{2} (\hat{\theta}^k - \theta^k)^T \frac{\partial^2 \log p_{qi}(\theta^k)}{\partial \theta^{k2}} (\hat{\theta}^k - \theta^k) + (\hat{\theta}^k - \theta^k)^T \frac{\partial^2 \log p_{qi}(\theta^k)}{\partial \theta^k \, \partial p_{qi}} e_i(\theta^k) \tag{2.16}$$

Thus

$$B_r(\hat{\theta}^k, \theta^k) = \sum_{i=1}^{s} p_{ri}(\theta_0) \frac{\partial \log p_{qi}(\theta^k)}{\partial \theta^k} (\hat{\theta}^k - \theta^k)$$

$$+ \frac{1}{2} \sum_{i=1}^{s} p_{ri}(\theta_0) (\hat{\theta}^k - \theta^k)^T \frac{\partial^2 \log p_{qi}(\theta^k)}{\partial \theta^{k2}} (\hat{\theta}^k - \theta^k)$$

$$+ \sum_{i=1}^{s} p_{ri}(\theta_0) (\hat{\theta}^k - \theta^k)^T \frac{\partial^2 \log p_{qi}(\theta^k)}{\partial \theta^k \, \partial pqi} e_i(\theta^k) \tag{2.17}$$

This expression can be further simplified by utilizing the fact that, since $\hat{\theta}^k$ is a maximum-likelihood estimate on the informative sample:

$$\sum_{i=1}^{s} p_{qi}(\Theta_0) \frac{\partial \log p_{qi}(\hat{\theta}^k)}{\partial \theta^k} = 0 \qquad (2.18)$$

Expanding this around θ^k yields

$$\sum_{i=1}^{s} p_{qi}(\Theta_0) \left[\frac{\partial \log p_{qi}(\theta^k)}{\partial \theta^k} + \frac{\partial^2 \log p_{qi}(\theta^k)}{\partial \theta k2} (\hat{\theta}^k - \theta^k) \right] = 0 \qquad (2.19)$$

Using (2.15) and (2.19) and in (2.17) yields

$$B_r(\hat{\theta}^k, \theta^k) = \sum_{i=1}^{s} e_i(\Theta_0) \frac{\partial \log p_{qi}(\theta^k)}{\partial \theta^k} (\hat{\theta}^k - \theta^k)$$

$$+ \frac{1}{2} \sum_{i=1}^{s} e_i(\Theta_0) (\hat{\theta}^k - \theta^k)^T \frac{\partial^2 \log p_{qi}(\theta^k)}{\partial \theta k2} (\hat{\theta}^k - \theta^k)$$

$$- \frac{1}{2} \sum_{i=1}^{s} p_{qi}(\Theta_0) (\hat{\theta}^k - \theta^k)T \frac{\partial^2 \log p_{qi}(\theta^k)}{\partial \theta k2} (\hat{\theta}^k - \theta^k)$$

$$+ \sum_{i=1}^{s} [p_{qi}(\Theta_0) + e_i(\Theta_0)] (\hat{\theta}^k - \theta^k)T \frac{\partial^2 \log p_{qi}(\theta^k)}{\partial \theta^k \, \partial p_{qi}} e_i(\Theta_0) \qquad (2.20)$$

where we have assumed $e_i(\theta^k) \approx e_i(\Theta_0)$.

The error $e_i(\Theta_0)$ is the difference of two probabilities, which are binomially distributed, by construction:

$$e_i(\Theta_0) = p_{ri}(\Theta_0) - p_{qi}(\Theta_0)$$

Furthermore $\sum_{i=1}^{s} e_i(\Theta_0) = 1$, by definition.

Since we are assuming here that $p_{ri}(\Theta_0)$ and $p_{qi}(\Theta_0)$ are independent samples from the same underlying distribution, $e_i(\Theta_0)$ is unbiased:

$$E\{e_i(\Theta_0)\} = 0 \tag{2.21}$$

where $E\{\ \}$ denotes expectation with respect to all underlying random variables. Recalling that the informative sample is of size n_q and the predictive sample is of size n_r, $p_{qi}(\Theta_0)$ has approximate variance

$$\text{var}\ (p_{qi}(\Theta_0)) = \frac{1}{n_q}\, p_i(\Theta_0)\ [1 - p_i(\Theta_0)]$$

and $p_{ri}(\Theta_0)$ has variance

$$\text{var}\ (p_{ri}(\Theta_0)) = \frac{1}{n_r}\, p_i(\Theta_0)\ [1 - p_i(\Theta_0)]$$

Thus

$$\text{var}\ (e_i(\Theta_0)) = \frac{1}{\bar{n}}\, p_i(\Theta_0)\ [1 - p_i(\Theta_0)] \tag{2.22}$$

where $\bar{n} = n_q n_r / (n_q + n_r)$. The expected value of $B_r(\hat{\theta}^k, \theta^k)$ can now be written in simplified form by using

$$\frac{\partial^2 \log\, p_{qi}(\theta^k)}{\partial \theta^k\, \partial p_{qi}} \approx - \frac{1}{p_i(\Theta_0)} \quad \frac{\partial \log\, p_i(\theta^k)}{\partial \theta^k}$$

Furthermore $\sum_{i=1}^{s} e_i(\Theta_0) = 0$, by definition.

Since we are assuming here that $p_{ri}(\Theta_0)$ and $p_{qi}(\Theta_0)$ are independent samples from the same underlying distribution, $e_i(\Theta_0)$ is unbiased:

$$E\{e_i(\Theta_0)\} = 0 \tag{2.21}$$

where $E\{\ \}$ denotes expectation with respect to all underlying random variables. Recalling that the informative sample is of size n_q and the predictive sample is of size n_r, $p_{qi}(\Theta_0)$ has approximate variance

$$\text{var}\ (p_{qi}(\Theta_0)) = \frac{1}{n_q}\, p_i(\Theta_0)\ [1 - p_i(\Theta_0)]$$

and $p_{ri}(\Theta_0)$ has variance

$$\text{var}\ (p_{ri}(\Theta_0)) = \frac{1}{n_r}\, p_i(\Theta_0)\ [1 - p_i(\Theta_0)]$$

Thus

$$\text{var}\ (e_i(\Theta_0)) = \frac{1}{\bar{n}}\, p_i(\Theta_0)\ [1 - p_i(\Theta_0)] \tag{2.22}$$

where $\bar{n} = n_q n_r / (n_q + n_r)$. The expected value of $B_r(\hat{\theta}^k, \theta^k)$ can now be written in simplified form by using

$$\frac{\partial^2 \log\, p_{qi}(\theta^k)}{\partial \theta^k\, \partial p_{qi}} \approx - \frac{1}{p_i(\Theta_0)} \quad \frac{\partial \log\, p_i(\theta^k)}{\partial \theta^k}$$

The result is that the expected value of $B_r(\hat{\theta}^k, \theta^k)$ is

$$E\left\{B_r(\hat{\theta}^k, \theta^k)\right\}$$

$$= -\frac{1}{2} E\left\{\sum_{i=1}^{s} p_i(\Theta_0)\ (\hat{\theta}^k - \theta^k)^T \frac{\partial^2 \log\ p_i(\theta^k)}{\partial \theta^{k2}}\ (\hat{\theta}^k - \theta^k)\right\}$$

$$- \frac{1}{\bar{n}} \sum_{i=1}^{s} E\left\{(\hat{\theta}^k - \theta^k)^T \frac{\partial \log\ p_i(\theta^k)}{\partial \theta^k}\right\} \qquad (2.23)$$

In the sequel we will choose $\theta^k = \theta^{*k}$ so that θ^{*k} is a minimum-variance estimate of Θ_0. This results in the second term being much smaller than the first term for reasonably large values of $\bar{n}/s$. We will explicitly neglect this term in the sequel.

Evaluation of $B_r(\Theta_0, \theta^{*k})$

From (2.23)

$$B_r(\Theta_0, \theta^k) =$$

$$-\frac{1}{2} \sum_{i=1}^{s} p_i(\Theta_0)\ (\Theta_0 - \theta^k)^T \frac{\partial^2 \log\ p_i(\Theta_0)}{\partial p_i^2}\ (\Theta_0 - \theta^k) \qquad (2.24)$$

$$-\frac{1}{2} (\Theta_0 - \theta^k)^T\ I(\Theta_0)\ (\Theta_0 - \theta^k) \qquad (2.25)$$

where $I(\Theta_0)$ is the information matrix

$$I(\Theta_0) = \sum_{i=1}^{s} p_i(\Theta_0) \frac{\partial^2 \log p_i(\Theta_0)}{\partial p_i^2} \tag{2.26}$$

In (2.23) both Θ^k and $\hat{\Theta}^k$ are k-dimensional parameter vectors. Here, however, Θ_0 is an m-dimensional vector ($m > k$). To handle this situation, we write $\Theta_0 - \Theta^k \in R^m$ as

$$\Theta_0 - \Theta^k = \begin{bmatrix} \Theta_0^k - \Theta^k \\ \tilde{\Theta}_0 \end{bmatrix}$$

where $\Theta_0^k \in R^k$, $\tilde{\Theta}_0 \in R^{m-k}$

setting

$$J(\Theta^k) = \frac{1}{2} (\Theta_0 - \Theta^k)^T I(\Theta_0) \Theta_0 - \Theta^k)$$

and minimizing with respect to Θ^k yields

$$\Theta^{*k} = \Theta_0^k - I_{11}^{-1} (\Theta_0) I_{12} (\Theta_0) \tilde{\Theta}_0 \tag{2.27}$$

where we have partitioned $I(\Theta_0)$ as

$$I(\Theta_0) = \begin{bmatrix} I_{11}(\Theta_0) & I_{12}(\Theta_0) \\ I_{12}^T(\Theta_0) & I_{22}(\Theta_0) \end{bmatrix}$$

The minimum value of J is

$$J(\theta^{*k}) = \frac{1}{2}\ \tilde{\theta}_0^T\ [I_{22}(\theta_0) - I_{12}(\theta_0)^T\ I_{11}(\theta_0)^{-1}\ I_{12}(\theta_0)]\ \tilde{\theta}_0$$

If we partition the covariance matrix

$$P(\theta_0) = I(\theta_0)^{-1}$$

$$= \begin{bmatrix} P_{11}(\theta_0) & P_{12}(\theta_0) \\ P_{12}^T(\theta_0) & P_{22}(\theta_0) \end{bmatrix}$$

then

$$J(\theta^{*k}) = \frac{1}{2}\ \tilde{\theta}_0^T\ \ P_{22}(\theta_0)^{-1}\ \tilde{\theta}_0$$

where $P_{22}(\theta_0)^{-1} = I_{22} - I_{12}^T\ I_{11}^{-1}\ I_{12}$

Since

$$P(\theta_0) = \sum_{i=1}^{s} P_i(\theta)\ (\theta_0 - \theta^{*k})\ (\theta_0 - \theta^{*k})^T$$

$$\approx E\ [(\theta_0 - \theta^{*k})\ (\theta_0 - \theta^{*k})^T]$$

we get, finally,

$$E\ [J(\theta^{*k})] \approx \frac{1}{2}\ (m - k)$$

or

$$E\ [B_r(\theta_0,\ \theta^{*k})] = \frac{1}{2}\ (k-m) \qquad (2.28)$$

2.4 Unbiased Estimate of Entropy

From (2.23)

$$E\left\{B_r(\hat{\theta}^k, \theta^k)\right\} = -\frac{1}{2}(\hat{\theta}^k - \theta^k)^T I(\theta^k)(\hat{\theta}^k - \theta^k)$$

where $I(\theta^k)$ is the kxk information matrix

$$I(\theta^k) = E\left\{\sum_{i=1}^{s} p_i(\theta_0)\frac{\partial^2 \log p_i(\theta^k)}{\partial \theta^{k2}}\right\}$$

and $p_i(\theta_0)$ is given in (2.6). Using (1.5) and (1.6) we see that

$$E\left\{B_r(\hat{\theta}^k, \theta^k)\right\} = -\frac{1}{2}\,tr\, I_k$$

$$= -\frac{k}{2} \qquad (2.29)$$

where I_k is the kxk identity matrix.

Combining (2.29) and (2.28) yields

$$E\left\{B_r(\hat{\theta}^k, \theta_0)\right\} = \frac{m}{2} - k \qquad (2.30)$$

where $\hat{\theta}^k$ is the maximum likelihood estimate ($\hat{\theta}^k \in R^k$). This represents a bias in the maximized log-likelihood function, with the result that our goal is to pick k such that

$$\sum_{i=1}^{s} p_{qi}(\theta_0) \log p_{qi}(\hat{\theta}^k) + \frac{m}{2} - k$$

is maximized. By reference to (2.10) and (2.12), this is equivalant asymptotically to picking k such that

$$\sum_{i=1}^{n_q} \log p(x_i \mid \hat{\theta}^k) + \frac{m}{2} - k$$

is maximized. Since m is a constant here, the equivalent goal is to minimize

$$AIC(k) = -2 \sum_{i=1}^{n_q} \log p(x_i \mid \hat{\theta}^k) + 2k \qquad (2.31)$$

with respect to k, which is Akaike's AIC criterion.

CHAPTER 3

3. CANONICAL VARIATES ANALYSIS

We now consider the linear prediction problem using the canonical variates analysis approach.

Let the past be represented as a column vector P(t) defined by

$$P(t) = \begin{bmatrix} y(t) \\ y(t-1) \\ \vdots \\ \cdot \end{bmatrix}_{nx1}$$

and define the future as a column vector

$$F(t) = \begin{bmatrix} y(t+1) \\ y(t+2) \\ \vdots \\ \cdot \end{bmatrix}_{mx1} \qquad m \leq n$$

where y(t) is the r-dimensional observed output at time t. Our goal is to predict the future F(t) given P(t).

We now consider the canonical variate analysis in a form that allows us to explicitly show the optimality properties of the method.

Consider nonsingular transformations of the past and future

$$\underset{nx1}{c(t)} = \underset{nxn}{J}\ \underset{nx1}{P(t)} \tag{3.1}$$

$$\underset{mx1}{d(t)} = \underset{mxm}{L}\ \underset{mx1}{F(t)} \tag{3.2}$$

and form a $k^{\underline{th}}$ order estimate of F(t)

$$\hat{F}_k(t) = \sum_{i=1}^{k} a_i \, c_i(t) \tag{3.3}$$

where $\{a_i\}$ are mxl vectors and c_i is the $i^{\underline{th}}$ component of c (a scalar). Since a_i is fixed, only $c_i(t)$ depends on the data. Since J is only constrained to be nonsingular, we can use a very general form for it. Without loss of generality we can specify that

$$E\,[c(t)\; c(t)^T] = I_{nxn} \tag{3.4}$$

Let B be an orthonormal matrix:

$$B_{nxn}^{T} \; B_{nxn} = I_{nxn} \tag{3.5}$$

Then

$$J \; S_{pp} \; J^T = B^T \; B \tag{3.6}$$

where $S_{pp} = E\,[P(t)\; P(t)^T]$

This has a solution

$$J = B^T \; S_{pp}^{-1/2} \tag{3.7}$$

Now

$$c_i = J_i^T \, P(t) \tag{3.8}$$

where J_i^T is the $i^{\underline{th}}$ row of J;

$$J_i^T = b_i^T S_{pp}^{-1/2} \tag{3.9}$$

and

$$B = [b_1\ b_2\ \cdots\ b_n]_{nxn} \tag{3.10}$$

Thus

$$c_i = b_i^T S_{pp}^{-1/2} P(t) \tag{3.11}$$

and the estimate $\hat{F}_k(t)$ is

$$\hat{F}_k(t) = \left[\sum_{i=1}^{k} a_i\ b_i^T \right] S_{pp}^{-1/2} P(t) \tag{3.12}$$

$$\underline{\Delta}\ Q_k\ S_{pp}^{-1/2} P(t)$$

where

$$Q_k = \sum_{i=1}^{k} a_i\ b_i^T \tag{3.13}$$

Note that Q_k has maximum rank k.

The prediction error is

$$e_k(t) = Q_k S_{pp}^{-1/2} P(t) - F(t) \tag{3.14}$$

We now form a quadratic cost function

$$L_k = E\,[e_k^T(t)\, W^{-1}\, e_k(t)]$$

$$= \mathrm{tr}\, W^{-1}\, E\,\{[Q_k\, S_{pp}^{-1/2}\, P(t) - F(t)]\,[P(t)^T\, S_{pp}^{-1/2}\, Q_k^T - F(t)^T]\}$$

$$= \mathrm{tr}(W^{-1}\, Q_k\, Q_k^T)$$
$$- 2\mathrm{tr}(W^{-1}\, Q_k\, S_{pp}^{-1/2}\, S_{pf}) + \mathrm{tr}(W^{-1}\, S_{ff}) \qquad (3.15)$$

where $S_{pf} = E\,[P(t)\, F(t)^T]$, $S_{ff} = E\,[F(t)\, F(t)^T]$

In order to handle the orthonormality constraints we add the constraint equations via Lagrange multipliers to form the augmented cost

$$\tilde{L}_k = L_k + \sum_{i=1}^{k} \lambda_i\, (b_i^T\, b_i - 1) \qquad (3.16)$$

where $\{\lambda_i\}$ are Lagrange multipliers. Thus

$$\tilde{L}_k = \mathrm{tr}\,\{W^{-1} \sum_{i=1}^{k} a_i\, b_i^T \sum_{j=1}^{n} b_j\, a_j^T\}$$

$$- 2\,\mathrm{tr}\,\{W^{-1} \sum_{i=1}^{k} a_i\, b_i^T\, S_{pp}^{-1/2}\, S_{pf}\}$$

$$+ \mathrm{tr}\,\{W^{-1}\, S_{ff}\}$$

$$+ \sum_{i=1}^{k} \lambda_i\, (b_i^T\, b_i - 1) \qquad (3.17)$$

Using $b_i^T\, b_j = \delta_{ij}$, with δ the Kroneker delta function, (3.18)

and rearranging gives

$$\bar{L}_k = \sum_{i=1}^{k} a_i^T W^{-1} a_i$$

$$- 2 \sum_{i=1}^{k} b_i^T S_{pp}^{-1/2} S_{pf} W^{-1} a_i$$

$$+ \mathrm{tr}(W^{-1} S_{ff}) + \sum_{i=1}^{k} \lambda_i (b_i^T b_i - 1) \qquad (3.19)$$

Taking partial derivatives:

$$\frac{\partial \bar{L}_k}{\partial a_i} = 2 a_i^T W^{-1} - 2 b_i^T S_{pp}^{-1/2} S_{pf} W^{-1} \qquad (3.20)$$

$$\frac{\partial \bar{L}_k}{\partial b_i} = -2 a_i^T W^{-1} S_{pf}^T S_{pp}^{-1/2} + 2 \lambda_i b_i^T \qquad (3.21)$$

Thus, the first order necessary conditions for minimizing $\bar{L}_k$ are

$$a_i^* = S_{pf}^T S_{pp}^{-1/2} b_i^* \qquad (3.22)$$

$$\lambda_i b_i^* = S_{pp}^{-1/2} S_{pf} W^{-1} a_i^* \qquad (3.23)$$

for $i = 1, 2, \ldots, k$.

Eliminating a_i^*:

$$\lambda_i b_i^* = S_{pp}^{-1/2} S_{pf} W^{-1} S_{pf}^T S_{pp}^{-1/2} b_i^* \qquad (3.24)$$

which is an eigenequation.

The first term of (3.19) becomes

$$\sum_{i=1}^{k} a_i^{*T} W^{-1} a_i^{*}$$

$$= \sum_{i=1}^{k} b_i^{*T} S_{pp}^{-1/2} S_{pf} W^{-1} S_{pf}^{T} S_{pp}^{-1/2} b_i^{*}$$

$$= \sum_{i=1}^{k} \lambda_i \qquad (3.25)$$

The second term of (3.19) becomes

$$-2 \sum_{i=1}^{k} b_i^{*T} S_{pp}^{-1/2} S_{pf} W^{-1} S_{pf}^{T} S_{pp}^{-1/2} b_i^{*}$$

$$= -2 \sum_{i=1}^{k} \lambda_i \qquad (3.26)$$

Thus, the optimized cost is

$$L_k^{*} = \mathrm{tr}(W^{-1} S_{ff}) - \sum_{i=1}^{k} \lambda_i \qquad (3.27)$$

Now let

$$R = S_{pp}^{-1/2} S_{pf} W^{-1/2} \qquad (nxm) \quad n>m \qquad (3.28)$$

From (3.24),

$$b_i^{*T} R R^{T} b_i^{*} = \lambda_i \qquad (3.29)$$

By using a singular value decomposition on R:

$$R = U\ D\ V^T \tag{3.30}$$

$$V^T\ V = I,\ U^T\ U = I \tag{3.31}$$

$$D = \begin{bmatrix} \gamma_1 & & 0 \\ & \ddots & \\ 0 & & \gamma_m \\ \hline & 0 & \end{bmatrix} \tag{3.32}$$

where $\gamma_1 > \gamma_2 > \cdots \gamma_m$

Then $R\ R^T = U\ D\ V^T\ V\ D^T\ U^T$

$$= U\ D\ D^T\ U^T \tag{3.33}$$

Then, from (3.29)

$$b_i^{*T}\ U\ D\ D^T\ U^T\ b_i^* = \lambda_i \tag{3.34}$$

Thus b_i^* is the eigenvector of $U\ D\ D^T\ U^T$ whose eigenvalue is λ_i.

Now let

$$U = [U_1\ U_2\ \cdots\ U_n] \tag{3.35}$$

where the U_i are mutually orthogonal unit vectors by construction. But the matrix $U\ D\ D^T\ U^T$ has eigenvectors U_i and associated eigenvectors γ_i^2 since

$$U_i^T\ U\ D\ D^T\ U^T\ U_j = \gamma_i^2\ \delta_{ij} \tag{3.36}$$

Thus

$$\gamma_i^2 = \lambda_i, \quad U_i = b_i^* \tag{3.37}$$

and

$$a_i^* = S_{pf}^T \; S_{pp}^{-1/2} \; U_i \tag{3.38}$$

By using (3.37) in (3.27) we get

$$L_k^* = \mathrm{tr}(W^{-1} S_{ff}) - \sum_{i=1}^{k} \gamma_i^2 \tag{3.39}$$

and we see that the cost is minimized by using the k largest canonical variances, $\gamma_1^2 > \gamma_2^2 > \gamma_3^2 > \ldots > \gamma_k^2$.

We can now write the optimal forecast as

$$\begin{aligned} F_k^*(t) &= \sum_{i=1}^{k} a_i^* \; c_i^*(t) \\ &= \sum_{i=1}^{k} S_{pf}^T \; S_{pp}^{-1/2} \; U_i \; U_i^T \; S_{pp}^{-1/2} \; P(t) \\ &= S_{pf}^T \; S_{pp}^{-1/2} \; \left(\sum_{i=1}^{k} U_i \; U_i^T \right) S_{pp}^{-1/2} \; P(t) \end{aligned} \tag{3.40}$$

Thus, if we denote the optimal weighting matrix by A_k^*:

$$F_k^*(t) = A_k^* \; P(t) \tag{3.41}$$

$$A_k^* = S_{pf}^T \; S_{pp}^{-1/2} \left[\sum_{i=1}^{k} U_i \; U_i^T \right] S_{pp}^{-1/2} \tag{3.42}$$

Note that

$$A_n^* = S_{pf}^T S_{pp}^{-1} \tag{3.43}$$

To determine L (cf (3.2)), we can use the condition

$$E(cd^T) = D \tag{3.44}$$

or

$$J^* S_{pf} L^T = D \tag{3.45}$$

From (3.7),

$$U^T S_{pp}^{-1/2} S_{pf} L^T = D \tag{3.46}$$

But

$$\begin{aligned} D &= U^T R V \\ &= U^T S_{pp}^{-1/2} S_{pf} W^{-1/2} V \end{aligned} \tag{3.47}$$

Comparing (3.46) and (3.47) gives

$$L^T = W^{-1/2} V \text{ , or}$$

$$L = V^T W^{-1/2} \tag{3.48}$$

Note that A_k^*, the optimal gain matrix is of dimension mxn but has a maximum rank of k.

Note that $k \leq m$ since the symmetric matrix in the eigenequation (3.24) has rank $\leq m$. This is very important, as it implies that we need to make the dimension of the future vector (m) at least as large as the maximum expected order of the estimator.

An efficient computation of $\overset{*}{A}_k$ is

$$\overset{*}{d}_i = S_{pp}^{-1/2} U_i \; ; \; S_{pp}^{-1/2} \text{ symmetric}$$

$$\overset{*}{a}_i = S_{pf}^{T} \overset{*}{d}_i$$

$$\overset{*}{A}_k = \sum_{i=1}^{k} \overset{*}{a}_i \overset{*}{d}_i^{T}$$

Cholesky Form

The cholesky factorization of a positive-definite matrix is an attractive way of computing a square root matrix. Let

$$S_{pp}^{-1} = Q \, Q^{T}$$

Then we get the following relations

$$\overset{*}{a}_i = S_{pf}^{T} Q \, U_i$$

$$\overset{*}{\lambda}_i \, U_i = Q^{T} S_{pf} \, W^{-1} \, S_{pf}^{T} Q \, U_i$$

$$R = Q^{T} S_{pf} \, W^{-1/2}$$

$$\overset{*}{F}_k(t) = S_{pf}^{T} \, Q \left(\sum_{i=1}^{k} U_i \, U_i^{T} \right) Q^{T} \, P(t)$$

$$\overset{*}{A}_k = S_{pf}^{T} \, Q \left(\sum_{i=1}^{k} U_i \, U_i^{T} \right) Q^{T}$$

Trucated Predictor

In the sequel, we will be restricting the total number of parameters allowed in the predictor. The question arises as how to best truncate the prediction equations. Our approach is to use only the most recent past values. For example, suppose we have used m = 5 in our analysis, but wish only to use a one-step-ahead predictor with k parameters. Then our predictor uses only the first k elements of the first row of A_k^*.

Inclusion of Known Inputs

If we have an unknown system with measured outputs y(t) and measured inputs u(t), the analysis of this section holds with only slight modifications. If we augment the past vector as

$$P(t) = \begin{bmatrix} y(t) \\ u(t) \\ y(t-1) \\ u(t-1) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \tag{3.49}$$

then all of the analyes of this section holds and the predicted values of y(t) depend on both past values of y(t) and on past values u(t).

CHAPTER 4

STATE SPACE MODELS USING CVA-AIC TECHNIQUE

We now consider the problem of determining the state-space matrices directly from the linear prediction solution. Recall from Section 3:

$$P(t) = \begin{bmatrix} y(t) \\ y(t-1) \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}_{nx1} \tag{4.1}$$

$$\hat{F}(t) = A\ P(t) \tag{4.2}$$

If we restrict our problem to a one-step-ahead prediction of y(t), then

$$\hat{y}(t+1 \mid t) = A\ P(t) \tag{4.3}$$

where A is mxn.

We can write a recursion for P(t) as follows:

$$P(t+1) = M\ P(t) + T\ y(t+1) \tag{4.4}$$

where

$$M = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ I & 0 & \cdots & 0 \\ 0 & I & 0 \cdots & 0 \\ \cdot & & \cdot & \cdot \\ \cdot & & & \cdot \\ 0 & \cdots & 0\ I & 0 \end{bmatrix} \tag{4.5}$$

where all submatrices are mxm.

$$T = \begin{bmatrix} I_{m \times m} \\ 0_{(n-m) \times m} \end{bmatrix} \tag{4.6}$$

so that P(t) is in recursive form with a driving term y(t+1).

The state-space formulation employs a Kalman filter for updating. The equations for the time-invariant filter at steady state are

$$\hat{y}(t+1 \mid t) = H \hat{x}(t+1 \mid t) \tag{4.7}$$

$$\hat{x}(t+1 \mid t) = \Phi \hat{x}(t \mid t) \tag{4.8}$$

$$\hat{x}(t+1 \mid t+1) = \hat{x}(t+1 \mid t) + K [y(t+1) - \hat{y}(t+1 \mid t)] \tag{4.9}$$

Combining (4.7) - (4.9) yields

$$\begin{cases} \hat{y}(t+1 \mid t) = H \Phi \hat{x}(t \mid t) & (4.10) \\ \hat{x}(t+1 \mid t+1) = (I-KH) \Phi \hat{x}(t \mid t) + K y(t+1) & (4.11) \end{cases}$$

as the state-space equation set. The linear prediction set is

$$\begin{cases} \hat{y}(t+1 \mid t) = A P(t) & (4.12) \\ P(t+1) = M P(t) + T y(t+1) & (4.13) \end{cases}$$

What we seek to do is <u>match</u> these two pairs of equations by finding the state-space matrices H, Φ, K which give the best "fit" to the linear prediction equations.

Equations (4.7) - (4.9) can also be put in the form

$$\hat{y}(t+1 \mid t) = H\,\hat{x}(t+1 \mid t) \tag{4.14}$$

$$\hat{x}(t+1 \mid t) = \Phi(I - KH)\hat{x}(t \mid t-1) + \Phi K\, y(t) \tag{4.15}$$

We can solve the problem operationally by using solutions involving delay operators.

We will first solve the linear prediction equations. From (4.13),

$$P(t+1) = zM\,P(t+1) + T\,y(t+1) \tag{4.16}$$

where z is the delay operator: $z\,P(t+1) = P(t)$

Solving (4.16):

$$P(t+1) = (I-zM)^{-1}\,T\,y(t+1) \tag{4.17}$$

Thus, from (4.12):

$$\hat{y}(t+1 \mid t) = A\,(I-zM)^{-1}\,T\,y(t) \tag{4.18}$$

We can also solve the state-space equations in the same way.

From (4.10) and (4.11) we get

$$\hat{x}(t \mid t) = [I - (I-KH)\,\Phi z]^{-1}\,K\,y(t) \tag{4.19}$$

so that

$$\hat{y}(t+1 \mid t) = H\,\Phi\,[I - (I-KH)\,\Phi z]^{-1}\,K\,y(t) \tag{4.20}$$

while from (4.14) and (4.15) we get

$$\hat{x}(t+2 \mid t+1) = [I - \Phi(I-KH)z]^{-1} \Phi K\, y(t+1) \tag{4.21}$$

$$\hat{y}(t+1 \mid t) = H\,[I - \Phi(I-KH)z]^{-1} \Phi K\, y(t) \tag{4.22}$$

If we could get a perfect match between the linear prediction and the state-space predictors, then the following equation would be satisfied

$$A\,(I-Mz)^{-1}\,T = H\,\Phi[I-(I-KH)\Phi z]^{-1}\,K \tag{4.23}$$

or, equivalently

$$A\,(I-Mz)^{-1}\,T = H[I-\Phi(I-KH)z]^{-1}\,\Phi\,K \tag{4.24}$$

Using (4.23), we see that exact matching occurs, for dim (x) = n, if

$$\begin{cases} A = H\,\Phi & (4.25) \\ M = (I-KH)\,\Phi & (4.26) \\ T = K & (4.27) \end{cases}$$

Equation (4.26) can be written as

$$M = \Phi - TA, \tag{4.28}$$

so that

$$\Phi = M + TA \tag{4.29}$$

In addition, (4.25) gives

$$H = A\,\Phi^{-1} \tag{4.30}$$

Since H must satisfy A = H (M + TA), it is easily shown that H is in canonical form

$$H = [I_{mxm}\ 0_{mx(n-m)}]$$

If Φ is a valid transition matrix, it is guaranteed to be invertible. Thus we only need to guarantee the invertibility of M + TA. Using the definition of M and partitioning T and A appropriately, we see

$$\Phi_{nxm} = \begin{bmatrix} 0_{mx(n-m)} & 0_{mxm} \\ I_{(n-m)x(n-m)} & 0_{(n-m)xm} \end{bmatrix} + \begin{bmatrix} I_{mxm} \\ 0_{(n-m)xm} \end{bmatrix} \begin{bmatrix} A_{1mx(n-m)} & A_{2mxm} \end{bmatrix}$$

$$= \begin{bmatrix} A_1 & A_2 \\ I & 0 \end{bmatrix} \tag{4.31}$$

The inverse is

$$\Phi^{-1} = \begin{bmatrix} 0 & I \\ A_2^{-1} & -A_2^{-1} A_1 \end{bmatrix} \tag{4.32}$$

Thus Φ^{-1} exists if A_2^{-1} exists. To check this, write equation (4.15) in partitioned form as

$$[A_1 \quad A_2] = \begin{bmatrix} S_{pf1}^T & S_{pf2_{pxp}}^T \end{bmatrix} \begin{bmatrix} W_{11} & W_{12} \\ W_{12}^T & W_{22} \end{bmatrix} \tag{4.33}$$

where

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{12}^T & W_{22} \end{bmatrix} = S_{pp}^{-1} \tag{4.34}$$

Then

$$A_2 = S_{pf1}^T W_{12} + S_{pf2}^T W_{22} \tag{4.35}$$

Now partition S_{pp} as

$$S_{pp} = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix} \tag{4.36}$$

Then

$$W_{12} = - S_{11}^{-1} S_{12} (S_{22} - S_{12}^T S_{11}^{-1} S_{12})^{-1} \tag{4.37}$$

$$W_{22} = (S_{22} - S_{12}^T S_{11}^{-1} S_{12})^{-1} \tag{4.38}$$

so that

$$A_2 = (S_{pf2}^T - S_{pf1}^T S_{11}^{-1} S_{12}) (S_{22} - S_{12}^T S_{11}^{-1} S_{12})^{-1} \tag{4.39}$$

Therefore

$$A_2^{-1} = (S_{22} - S_{12}^T S_{11}^{-1} S_{12}) (S_{pf2}^T - S_{pf1}^T S_{11}^{-1} S_{12})^{-1} \tag{4.40}$$

Solving for A_1 yields

$$A_1 = S_{pf1}^T W_{11} + S_{pf2}^T W_{12}^T \tag{4.41}$$

Using

$$W_{11} = S_{11}^{-1} + S_{11}^{-1} S_{12} (S_{22} - S_{12}^{T} S_{11}^{-1} S_{12})^{-1} S_{12}^{T} S_{11}^{-1} \qquad (4.42)$$

we get

$$A_1 = S_{pf1}^{T} S_{11}^{-1} +$$

$$[S_{pf1}^{T} S_{11}^{-1} S_{12} - S_{pf2}^{T}] (S_{22} - S_{12}^{T} S_{11}^{-1} S_{12})^{-1} S_{12}^{T} S_{11}^{-1} \qquad (4.43)$$

In summary, we can use the direct solution if the matrix

$$S_{ce} = S_{pf2}^{T} - S_{pf1}^{T} S_{11}^{-1} S_{12} \qquad (4.44)$$

is invertible.

This exact solution is restricted to the case dim (x) = n. This implies that

$$\dim(x(t)) = \dim(P(t))$$

Truncated Filter

Given the matrices for the full-order Kalman filter Φ_{nxn}, H_{mxn}, K_{nxm}, the question arises as to whether there is a suitable truncation to a lower-oder form required to meet restrictions on the total number of parameters. Using the forms of (4.27), (4.29) and (4.30) and assuming an order $k < n$, and prediction of the first p future values ($p < m$), we truncate as follows:

(1) Φ_{kxk} is the upper left kxk submatrix of Φ_{nxn}

(2) H_{pxk} is the upper left pxk submatrix of H_{mxn}

(3) K_{kxp} is the upper left kxp submatrix of K_{nxm}

Then the Kalman filter using Φ_{kxk}, H_{pxk}, K_{kxp} yields <u>exactly</u> the same predictions as the truncated linear predictor of Section 3.

CONFIDENCE BAND AND ACHIEVABLE IN SPECTRAL ESTIMATION

5.1 Spectral Estimation Problem

The problem of determining the statistical accuracy in identifying a model for a stationary multiple time series is considered in this chapter. The cases of the presence or absence of an exogenous input or additive measurement noise are included. Consider the general case where the vector x(t) is the exogenous input and the vector y(t) is the observed endogenous output of a system which may include other unknown excitations and measurement noise. Thus consider the jointly stationary gaussian vector time series x(t) and y(t), t = ..., with power cross–spectral matrices $S_{xx}(\omega,\theta)$, $S_{xy}(\omega,\theta)$, $S_{yy}(\omega,\theta)$ parameterized by θ, and denote the power cross–spectral matrices of the joint vector $(x^T(t),y^T(t))^T$ as $S(\omega,\theta)$.

Statistical inference is considered on a class of linear Gaussian processes parameterized by θ. Specifying a parametric model for the conditional process y(t), t < s, given x (t), t < s, implies a causal linear model of the form

$$y(t) = q(t) + \sum_{\tau=0}^{1} h(t-\tau;\theta)x(\tau) = q(t) + r(t)$$

where $h(t;\theta)$ is a causal linear system giving the response r(t) due to the past exogenous input x(t) and where q(t) is the error in predicting y(t) by r(t). From linear prediction theory, the transfer function of $h(t;\theta) = S_{yx}(\omega,\theta)S_{xx}^{-1}(\omega,\theta)$, and the error q(t) in predicting y(t) is uncorrelated with r(t) with power spectrum $S_{qq}(\omega;\theta) = S_{yy}(\omega,\theta) - H(\omega,\theta)S_{xx}(\omega,\theta)H^*(\omega,\theta)$. Note that any class of parameterized models $S(\omega,\theta)$ can be equivalently specified by the parameterized models $(S_{qq}(\omega,\theta),H(\omega,\theta))$ which will prove

more convenient.

It is convenient to work entirely in the frequency domain and specify the probability distribution and likelihood functions in terms of the power cross-spectral density matrices and Fourier coefficients. For simplicity the time series case with t a scalar is developed below, however the results generalize easily to the random field case of a vector t. Then asymptotically the log likelihood function is given following Whittle (1953) and Larimore (1977) with $Q(\omega) = Y(\omega) - R(\omega)$ and using the relationship $E\{Q(\omega) = Y(\omega) - R(\omega)$ and using the relationship $E\{Q(\omega)X^{*}(\omega)\} = 0$ by

$$\log p(x,\theta) = -\frac{N}{2}\log 2\pi - \frac{N\pi}{2}\int_{-\pi}^{} [\log|\, S_{qq}(\omega)| \; + Q^{*}(\omega)S_{qq}^{-1}(\omega)]\frac{d\omega}{2\pi}$$

and the elements of the gradient vector $\partial \log p/\partial\theta$ and Fisher information matrix $F(\hat{\theta})$ are

$$\frac{\partial \log p}{\partial \theta_i} = -\frac{N}{2}\int_{-\pi}^{\pi} \mathrm{tr}[\{I - S_{qq}^{-1}(\omega)Q(\omega)Q^{*}(\omega)\}S_{qq}^{-1}(\omega)\frac{\partial S_{qq}(\omega)}{\partial \theta_i}$$

$$- X(\omega)Q^{*}(\omega)S_{qq}^{-1}(\omega)\frac{\partial\, H(\omega)}{\partial \theta_i}]\frac{d\omega}{2\pi}$$

$$F_{ij}(\theta) = -E\{\frac{\partial^2 \log p}{\partial \theta_i \partial \theta_j}\}$$

$$= \frac{N}{2}\int_{-\pi}^{\pi} \mathrm{tr}[S_{qq}^{-1}(\omega)\frac{\partial\, S_{qq}(\omega)}{\partial\, \theta_i}\{S_{qq}^{-1}(\omega)\frac{\partial\, S_{qq}(\omega)}{\partial\, \theta_j} + S_{xx}^{-1}(\omega)\frac{\partial\, H(\omega)}{\partial\, \theta_j}\}]\frac{d\omega}{2\pi}$$

5.2 Simultaneous Confidence Bands

Let $\gamma \epsilon \Gamma$ be a variable such as frequency or time, and consider a p–dimensional

complex vector $f(\gamma,\theta)$ with components that are functions of γ and θ having continuous second derivatives with respect to the parameters θ. For example, the elements of the vector function $f(\gamma,\theta)$ could be the elements of the spectral matrix S, the squared magnitude coherencies, the impulse response functions of a spectral factor, or the covariance functions of the process. Asymptotically

$$f(\gamma,\hat{\theta}) - f(\gamma,\theta) = f_{\theta}(\gamma,\hat{\theta})(\hat{\theta} - \theta)$$

where $f_{\theta}(\gamma,\hat{\theta})$ denotes the matrix of partials $\partial f(\gamma,\theta)/\partial\theta^{T}$ evaluated at $\theta = \hat{\theta}$. This expansion and the Scheffe method (Scheffe, 1953, 1959, p. 68–70) of simultaneous confidence intervals as applied in Newton & Pagano (1984) lead to simultaneous confidence bands in the univariate case. For multivariate processes, it is of considerable interest to extend these results to simultaneous confidence bands on vector and matrix functions of the parameters, e.g. the spectral matrix. The extension that we will consider is the quadratic form

$$\{f(\gamma,\hat{\theta}) - f(\gamma,\theta)\}^{*} P(\gamma)\{f(\gamma,\hat{\theta}) - f(\gamma,\theta)\} \qquad (2.1)$$

which will be bounded as a function of γ. In the multivariate case, there is a choice to be made for P. For reasons of invariance and to obtain an equally tight confidence bound on any linear combination of $f(\gamma,\hat{\theta}) - f(\gamma,\theta)$, P is naturally chosen as the inverse of the covariance of (2.1).

In the sequel, a general P is used and then specialized to this natural choice. The basic mathematical result needed for such an extension is given in the Appendix and is used to prove the following theorem on simultaneous confidence intervals.

Theorem 1. Consider a parametric family of stationary Gaussian vector processes with power cross–spectral density matrices $S(\gamma,\theta)$ for $\theta\epsilon\Theta$ satisfying regularity conditions (Whittle, (1953), and for which the parameters are locally identifiable so that the Fisher information matrix $F(\theta)$ as given by (1.1) is full rank. Let y(1), y(2),...,y(N) be a sample realization and $\hat{\theta}$ be an asymptotically normal and efficient estimator of θ. Let $P(\gamma,\hat{\theta})$ be a Hermitian Matrix. Then as $N \to \infty$, the probability is at least $1-\alpha$ that simultaneously for all $\gamma\epsilon\Gamma$ the true p–vector function $f(\gamma,\theta)$ is bounded by

$$\{f(\gamma,\hat{\theta}) - f(\gamma,\theta)\}^{*} P(\gamma,\hat{\theta})\{\gamma,\hat{\theta}) - f(\gamma,\theta)\}$$

$$\leq X^2_{a,q}\ \mathrm{tr}\ f_{\theta}(\gamma,\hat{\theta})\ F^{-1}|\ (\hat{\theta})f^{*}_{\theta}\ (\gamma,\hat{\theta})P|\ (\gamma,\hat{\theta})$$

where q is the dimension of the vector θ and where $X^2_{aa,q}$ is the upper a critical point of the chisquared distribution on q degrees of freedom.

Proof: As shown by Rothenberg (1971), the parameters are locally identifiable if and only if the Fisher information is full rank. Let $f(\gamma)$ and $f(\gamma)$ denote $f(\gamma,\theta)$ evaluated at θ and $\hat{\theta}$ respectively. The vector random variable $N^{1/2}\{f(\gamma) - f(\gamma)\}$ is asymptotically distributed as the normal random vector $N^{1/2} f_{\theta}(\gamma,\hat{\theta})(\hat{\theta} - \theta)$. Asymptotically $(\theta - \theta)^{T} F(\hat{\theta})(\hat{\theta} - \theta)$ is a $X^2_{a,q}$ random variable, where $F(\theta)$ is proportional to sample size N as in (1.1). So the probability is $1-a$ that the true θ satisfies $(\hat{\theta} - \theta)^{T} M(\hat{\theta} - \theta) \leq 1$ where $M = F(\hat{\theta})/X^2_{a,q}$. From the Appendix, this inequality is satisfied if and only if $\|H(\hat{\theta} - \theta)\|^2 \leq \mathrm{tr} HM^{-1}H^{*}$ for all pxq –dimensional matrices H. Since the set $\{H = P^{1/2}(\gamma,\hat{\theta}) f_{\theta}(\gamma,\theta)$ for $\gamma 2\Gamma\}$ is possibly a proper subset of all pxq –dimensional matrices H, it follows that asymptotically with probability at least $1-a$ the inequality.

$$N\{\hat{f}(\gamma) - f(\gamma)\}^{*} P(\gamma\hat{\theta})\{\hat{f}(\gamma) - f(\gamma)\}$$

$$= N\{f_\theta(\gamma,\hat{\theta})(\hat{\theta}-\theta)\}^* P(\gamma,\hat{\theta})\{f_\theta(\gamma,\hat{\theta})(\hat{\theta}-\theta)\}$$

$$\leq\leq N X^2_{a,q} \operatorname{tr} f_\theta(\gamma,\hat{\theta})\, F^{-1}(\hat{\theta})\, f^*_\theta(\gamma,\hat{\theta}) P(\gamma,\hat{\theta}) \qquad (2.3)$$

is satisfied simultaneously for all $\gamma\epsilon\Gamma$.

For the natural choice of $P = \{f_\theta(\gamma,\hat{\theta})\, F^{-1}(\hat{\theta}) f^*_\theta(\gamma,\hat{\theta})\}^\dagger$, using $\dagger$ to denote the pseudo inverse of the covariance of (2.2), the inequality (2.3) becomes

$$\{\hat{f}(\gamma) - f(\gamma)\}^* \{f_\theta(\gamma,\hat{\theta})\, F^{-1}(\hat{\theta}) f^*_\theta(\gamma,\hat{\theta})\}^\dagger \{\hat{f}(\gamma) - f(\gamma)\} \leq r X^2_{a,q}$$

where r = Rank (P)

The relative squared spectral error $\operatorname{tr}[\hat{S}^{-1}(\omega)\{\hat{S}(\omega) - S(\omega)\}]^2$ is a fundamental quantity in measuring the accuracy of a spectral estimation procedure. The integral of this quantity is asymptotically the Kullback–Leibler information of negative entropy (Larimore, 1983) which is a fundamental statistical measure of model approximation error. The expected value of the integral is proportional to the number of estimated parameters divided by the sample size (Larimore, 1982). From Theorem 1, simultaneous confidence bands on the sample relative squared spectral error are given by the following theorem.

Theorem 2. Under the conditions of Theorem 1, as $N \to \infty$, the probability is at least $1 - a$ that simultaneously for all $\omega\epsilon\Omega$ the sample squared relative spectral error is bounded as

$$\operatorname{tr}[\hat{S}^{-1}(\omega)\{\hat{S}(\omega) - S(\omega)\}]^2$$

$$\leq \chi^2_{a,q} \operatorname{tr}_{k,\ell}^{q} S^{-1}(\omega,\theta)\{\hat{S}(\omega) - S(\omega)\}]^2$$

where $\{g_{k,\ell}(\theta)\} = G = F^{-1}(\theta)$.

Proof: Asymptotically $\hat{S}(\omega)$ and $S(\omega)$ are equal so that we may consider its inverse in (2.3) a constant denoted $\bar{S}(\omega)$. To apply Theorem 1, we consider the Hermitian matrix $A(\omega) = S^{-1/2}\{S(\omega) - \hat{S}(\omega)\}S^{-1/2}(\omega)$ and express the squared relative error symmetrically as

$$\operatorname{tr}[\hat{S}^{-1}(\omega)\{\hat{S}(\omega) - S(\omega)\}]^2 = \operatorname{tr}[S^{-1/2}(\omega)\{S^{-1/2}(\omega)]^2$$

$$= \operatorname{tr} AA = \operatorname{tr} AA^{*} = \sum_{i,j} a_{ij} a^{*}_{ij} = f^{*}(\omega)f(\omega)$$

where $f(\omega) = \operatorname{vec}A(\omega)$ is a vector containing the elements of the matrix $A(\omega)$. Application of Theorem 1 to the vector function $f(\omega)$ and rearrangement as in (2.4) proves the inequality. Expanding $S(\omega,\theta)$ as in (2.4), the equality follows from

$$E \operatorname{tr}[\hat{S}^{-1}(\omega)\{\hat{S}(\omega) - S(\omega)\}^2$$

$$= \operatorname{tr} \sum_{k,j} S^{-1}(\omega,\theta) \frac{\partial S(\omega,\theta)}{\partial \theta_k} E(\hat{\theta} - \theta)(\hat{\theta} - \theta)^{T} S^{-1}(\omega,\theta) \frac{\partial S(\omega,\theta)}{\partial \theta_\ell}$$

and using $E(\hat{\theta} - \theta)(\hat{\theta} - \theta)^{T} = F^{-1}$ from the asymptotic efficiency of $\hat{\theta}$

In principle any quadratic form in the components of the spectral matrix could be used as in Theorem 1 by introducing a weighting matrix $P(\omega,\theta)$. For confidence intervals

on the spectral matrix, the weighting of the inverse covariance of the error in estimating the spectral matrix gives tightest confidence bands which can be expressed as

$$\mathrm{vec}^{*}\{\hat{S}(\omega)-S(\omega)\}\ \{\sum_{k,\ell}\frac{\partial \mathrm{vec}S(\omega)}{\partial\theta_k}g_{k\ell}(\hat{\theta})\frac{\partial \mathrm{vec}^{*}S(\omega)}{\partial\theta_\ell}\}^{\dagger}\mathrm{vec}\{\hat{S}(\omega)\}\le X^2_{a,q}$$

For a given confidence level a, this gives a simultaneous confidence band for all frequencies $\omega 2\Omega$ as a quadratic form in the elements of $\hat{S}(\omega)-S(\omega)$,

5.3 Entropy and Spectral Accuracy

Consider the following predictive inference setting involving an observed informative sample $u^{T}=(x^{T}(1),y^{T}(1),\ldots,x^{T}(N),y^{T}(N))$ of size N used to estimate the process model, and similarly consider a conceptual predictive sample v of size M used to evaluate the accuracy of the estimated model. The predictive sample is assumed to be identically distributed by independent of the informative sample. Consider the problem of inference on the parametric class $\{p(v,\theta),\theta\epsilon\Theta\}$ of models with probability densities $p(v,\theta)$ based upon the informative sample u. Consider the conceptual repeated sampling experiment where on each trial the samples u and v are each drawn independently from the process $S(\omega,\theta_*)$ with θ_* assumed to be the true parameter value. An estimative model $\hat{p}=p(v,\hat{\theta}(u))$ is chosen for the density of v by some parameter estimation scheme $\hat{\theta}(u)$. The expected negative entropy, also known as the expected Kullback–Leibler discrimination information or expected I–divergence, is a measure of the error in approximating the true density p_* of v by the estimate $\hat{p}$ and is given by

$$R(p_*,\hat{p})=E_u K(p_*,\hat{p})=E_u\int p(v,\theta_*)\log\frac{p(v,\theta_*)}{p(v,\hat{\theta}(u))}dv \qquad (3.1)$$

where E_u denotes expectation relative to u and K denotes the Kullback–Leibler information. The negative entropy measure follows as the natural measure in the predictive inference setting from the fundamental principles of sufficiency and repeated sampling (Larimore (1983)). This approach applies to very general modeling methods such as nonparametric, semi parametric or parametric procedures as well as methods including decisions on model structure or order such as those used for AR and ARMA modeling.

Let lower case variables denote a sample of size M of the predictive sample, e.g. $y=(y^T(1),y^T(2),\ldots,y^T(M))^T$ and qyy denote the covariance matrix of y. By expressing the density $p(y,x;\theta)=p(y-r;\theta)p(x;\theta)$ in terms of the conditional random process $q(t)=y(t)-r(t)$, the log likelihood separates with the density of x(t) in many problems not a function of the unknown parameters or at least a function of a separate set of parameters. The I–divergence (3.1) thus becomes

$$K(p_*\hat{p}) = \int p(q,\theta_*) \log\frac{p(q,\theta_*)}{p(q,\hat{\theta})}\,dq + \int p(x,\theta_*) \log\frac{p(x,\theta_*)}{p(x,\hat{\theta})}\,dx$$

$$= K(p_*(q),\hat{p}(q)) + K(p_*(x),\hat{p}(x)) \qquad (3.2)$$

This conditional viewpoint is tken in the following where only the first term of the I–divergence is considered. Inclusion of the second term is tantamount to modeling the joint vector time series involving the two series x(t) and y(t) jointly rather than as exogeneous and endogeneous respectively. The joint case is included as a special case of y(t) a vector process with no input x(t) which will be discussed as a particular instance of the model throughout the paper. One further expression for the I–divergence will be very useful

$$K(p_*(q),\hat{p}(q)) = E \log p(y-r,\Sigma_{qq}) + E \log p(y-\hat{r}, \hat{\Sigma}_{qq})$$

$$= E \log p(y-r,\Sigma_{qq}) + E \log p((y-r)+(r-\hat{r}),\hat{\Sigma}_{qq})$$

$$= E \log p(y-r,\Sigma_{qq}) + E \log g((y-r), + E(r-\hat{r})^T \hat{\Sigma}_{qq}^{1}(r-\hat{r})$$

where E denotes expectation with respect to the density p_*.

Let $\hat{S}$ denote an estimate of S. We will need to assume tht $S(\omega)$ is continuous and that $S_{qq}(\omega,\theta)$ and $\hat{S}_{qq}(\omega, \theta)$ are positive definite for $\omega\epsilon[-\pi,\pi]$. In the discussion, the redictive sample v will be considered to be conditional on x(t) and to have an infinite sample size M. This will require the normaalization of the negative entropy and I–divergence by the sample size M. The I–divergence per sample time conditional on x(t), which we will denote $I(S,\hat{S})$ and call I–divergence for brevity, can be expressed using (3.2) as (Kazakos & Papantoni–Kazakos, 1980)

$$I(S,\hat{S}) = \lim_{M\to\infty} \frac{1}{M} K(p(v_M,\theta_*),p(v_M,\hat{\theta}(u_N))$$

$$= -\frac{1}{2}\int_{-\pi}^{\pi} \{\log|S_{qq}(\omega)\hat{S}_{qq}^{-1}(\omega)| + tr[I - S_{qq}(\omega)\hat{S}_{qq}^{-1}(\omega)|\}\frac{d\omega}{2\pi}$$

$$-\frac{1}{2}\int_{-\pi}^{\pi} tr\{\hat{S}_{qq}^{-1}[H(\omega) - \hat{H}(\omega)]SS_{xx}(\omega) - \hat{H}(\omega)]^*\}\frac{d\omega}{2\pi} \qquad (3.4)$$

where the subscript emphasizes that the sample of size M of v becomes infinite. The negative entropy per sample, or negentropy for brevity, is defined as $N(S,\hat{S})$ =

$\frac{1}{M!}\lim R(p_*,\hat{p})$. Note that the I–divergence is composed of two terms, the last due to the error estimaating the transfer function $H(\omega)$ and the first due to the error in estimating the spectrum $S_{qq}(\omega)$ of the noise q(t). A useful approximation for the first term in (3.4) is

$$-\frac{1}{2}\int_{-\pi}^{\pi}\{\log|S_{qq}(\omega)\hat{S}_{qq}^{-1}(\omega)| + \mathrm{tr}[I - S_{qq}(\omega)\hat{S}_{qq}^{-1}(\omega)|\}\frac{d\omega}{2\pi}$$

$$\downarrow \frac{1}{4}\int_{-\pi}^{\pi}\mathrm{tr}\{S_{qq}^{-1}(\omega)[\hat{S}_{qq}(\omega) - S_{qq}(\omega)|\}^2\frac{d\omega}{2\pi}$$

which holds to second order in the elements of $\hat{S}_{qq}$ as is easily shown by comparing first and second derivatiives of the integraands. This is a generalization to the multivariate case of the integral of the squared relaative error. Thus the I–divergence is approximately a quadratic form in the estimation errors of $S_{qq}(\omega)$ and H (ω), and these quadratic forms do *not intereact*, i.e. there are no cross terms.

5.4 Normalized Spectral Error in Principal Components

In the mulple time series case, the spectral measure has an intuitive interpretation in terms of principal components of the power spectrum in the frequency domain. Principle component representations of the spectral matricies $S_{xx}(\omega)$ and $S_{qq}(\omega)$ have the form.

$$J(\omega)S_{qq}(\omega)J^*(\omega) = D(\omega)\ ,\ L(\omega)S_{xx}(\omega)L^*(\omega) = E(\omega) \qquad (4.1)$$

where $J(\omega)$ and $L(\omega)$ given as a functionof frequency ω are unitary matrix transformatons so $J(\omega)J^*(\omega)=I=L(\omega)L^*(\omega)$ which diagonalize $S_{xx}(\omega)$ and $S_{qq}(\omega)$ repectively and where

where $G(\omega)=J(\omega)H(\omega)L^{*}(\omega)$ is the transfer function $H(\omega)$ expressed inthe coordinate frame of the principal component series $\bar{x}(t)$ and $\bar{y}(t)$. The squared magnitude error $|\hat{G}_{ij}(\omega) - G_{ij}(\omega)|^2$ in the i, j element of the transfer function is weighted by the input signal to output noise ratio $D_{ii}E_{jj}$ for the pair i,j.

6. DETECTION OF ABRUPT MODEL CHANGES

In this section we apply the tools developed so far to the problem of abrupt change detection. We then present some experimental results.

6.1 Algorithm Development

Consider the situation depicted in Figure 6.1, in which the true time-series model changes from M_0 to M_1 at time $t-d_1$, where t is the present time. Suppose

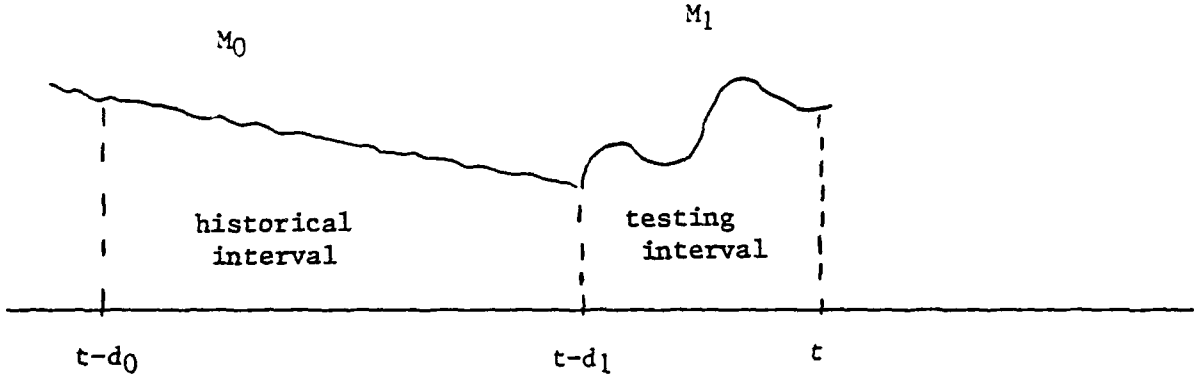

Figure 6.1 Changing Time Series Model

that we have data back to time $t-d_0$ and that the true model is M_0 in the interval $(t-d_0, t-d_1)$.

We wish to detect this change in the model. In this example, fitting a single model to data over the interval $(t-d_0, t)$ should result in greater fit errors than fitting one model over the interval $(t-d_0, t-d_1)$ and another model over the interval $(t-d_1, t)$. The crucial issue is to determine an appropriate selection measure so as to be sensitive to

changing models, while at the same time not being too sensitive to noise. Over-sensitivity to noise will result in deciding that model changes have occurred when, in fact, they have not. Low sensitivity to model changes will result in missing changes which have occurred in the model. Another obvious problem is how best to select the testing intervals, $(t-d_0, t)$ and $(t-d_1, t)$, to minimize the time required to achieve accurate detection. We consider first the selection measure.

Over the interval $(t-d_0, t)$ we can find the model which minimizes the AIC:

$$\text{AIC}(k) = -2 \sum_{i=1}^{d_0} \log p(e(t-d_0+i) \mid \hat{\theta}^k) + 2\,M(k) \qquad (6.1)$$

where M(k) is the number of independently adjustable parameters and where we have assumed a sampling time increment of one, for convenience.

If we now divide the interval $(t-d_0, t)$ into two subintervals, $(t-d_0, t-d_1)$ and $(t-d_1, t)$, we determine minimum AIC models for each subinterval

$$\text{AIC}_0(k) = -2 \sum_{i=1}^{d_0-d_1} \log p(e(t-d_0+i) \mid \hat{\theta}^k) + 2\,M(k) \qquad (6.2)$$

$$\text{AIC}_1(k) = -2 \sum_{i=d_0-d_1+1}^{d_0} \log p(e(t-d_0+i) \mid \hat{\theta}^k) + 2\,M(k) \qquad (6.3)$$

Now assume that

$k^* = \arg\min AIC(k)$

$k_0^* = \arg\min AIC_0(k)$

$k_1^* = \arg\min AIC_1(k)$

and let the corresponding models be parametrized by $\hat{\theta}^{k^*}$, $\hat{\theta}_0^{k_0^*}$, $\hat{\theta}_1^{k_1^*}$ respectively.

Then the model selection criterion is based on comparing $AIC(k^*)$ with $AIC_0(k_0^*) + AIC(k_1^*)$ and selecting the model(s) which give the least value. We can simplify the calculation in the case that $d_0 >> d_1$ and the model does not change too much, in which case we expect that $k^* \approx k_0^*$, $\theta^{k^*} \approx \theta_0^{k_0^*}$. In this case we can define the AIC difference as

$$\Delta AIC^* = AIC(k^*) - AIC_0(k_0^*) - AIC_1(k_1^*)$$

$$= -2 \sum_{i=d_0-d_1+1}^{d_0} \log p(e(t-d_0+i) \mid \hat{\theta}^{k^*})$$

$$+ 2 \sum_{i=d_0-d_1+1}^{d_0} \log p(e(t-d_0+i) \mid \theta_1^{k_1^*}) - 2\,M(k_1^*) \qquad (6.4)$$

and the decision rule is

$$\Delta AIC^* \begin{cases} < 0 \text{ ; declare "no change"} \\ > 0 \text{ ; declare "change"} \end{cases} \qquad (6.5)$$

Note that ΔAIC^* may be written as

$$\Delta AIC^* = 2 \sum_{i=1d_0-d_1+1}^{d_0} \log \frac{p(e(t-d_0+i) \mid \hat{\theta}_1^{k_1^*})}{p(e(t-d_0+i) \mid \hat{\theta}^{k^*})} - 2\,M(k_1^*) \qquad (6.6)$$

which is the likelihood ratio in favor of the best model in the interval $(t-d_1, t)$ to the best historical model evaluated over the same interval, but biased off by the number of parameters of the best model in the interval $(t-d_1, t)$.

If we specialize this result to the linear prediction problem under study here, we see that

$$\begin{aligned}\Delta AIC^* \approx\ & d_1 \left\{\log \left| S(k^*) \right| + \operatorname{tr} \tilde{S}(k^*)\, S(k^*)^{-1}\right\} \\ & - d_1 \left\{\log \left| S_1(k_1^*) \right| + m\right\} \\ & - 2\, mk_1^*\end{aligned} \tag{6.7}$$

where $S(k^*)$ is the theoretical covariance matrix of prediction errors for the historical model fitted on the interval $(t-d_0, t-d_1)$, $S_1(k_1^*)$ is the theoretical covariance matrix of prediction errors for the model fitted to the data on the interval $(t-d_1, t)$, and $\tilde{S}(k^*)$ is the actual covariance matrix of prediction errors for the historical model, evaluated on the interval $(t-d_1, t)$. Now let $\Delta\tilde{S}(k^*) = \tilde{S}(k^*) - S(k^*)$.

Then

$$\Delta AIC^* = d_1 \log \left\{ \frac{\left| S(k^*) \right|}{\left| S_1(k_1^*) \right|} \exp \left[\frac{-2\, mk_1^*}{d_1} + \operatorname{tr} \Delta\tilde{S}(k^*) S(k^*)^{-1} \right] \right\} \tag{6.8}$$

Thus our decision parameter is

$$\gamma = \log \frac{|S(k^*)|}{|S_1(k_1^*)|} - \frac{2\ mk_1^*}{d_1} + \text{tr}\ \tilde{\Delta S}(k^*)S(k^*)^{-1} \qquad (6.9)$$

and the decision rule is

$$\gamma \begin{cases} < 0 \text{ ; declare "no change"} \\ > 0 \text{ ; declare "change"} \end{cases} \qquad (6.10)$$

6.2 Experimental Results

The abrupt change detector was tried on a changing autoregressive model. On the interval $t \in [1, d_0]$, the actual model was

$$y(t) = 1.65\ y(t-1) - 0.665\ y(t-2) + u(t) \qquad \text{(Model 1)} \qquad (6.11)$$

where u(t) was zero-mean white Gaussian noise with variance of 1. This model has two real stable poles at 0.95 and 0.7. The actual model was then changed to

$$y(t) = 2.5\ y(t-1) - 2.11\ y(t-2) + 0.595\ y(t-3) + u(t) \qquad \text{(Model 2)} \qquad (6.12)$$

on the interval $t \in [d_0 + 1, d_0 + d,]$. This model has three poles at 0.7, 0.9 + 0.2i, 0.9 - 0.2i.

The first trial used $d_0 = 80$, $d_1 = 20$. The resulting covariance matrices on the interval $[1, d_0]$ were

$$S_{pp1} = \begin{Bmatrix} 11.0808 & 10.9642 & 10.7662 & 10.5378 & 10.2860 \\ 10.9642 & 11.0018 & 10.8992 & 10.7144 & 10.4733 \\ 10.7662 & 10.8992 & 10.9484 & 10.8566 & 10.6614 \\ 10.5378 & 10.7144 & 10.8566 & 10.9144 & 10.8144 \\ 10.2860 & 10.4733 & 10.6614 & 10.8144 & 10.8619 \end{Bmatrix}$$

$$S_{pf1} = \begin{Bmatrix} 11.0439 & 10.9089 & 10.7124 & 10.4822 & 10.2222 \\ 10.8318 & 10.6534 & 10.4500 & 10.2258 & 9.9723 \\ 10.5900 & 10.4015 & 10.1993 & 9.9753 & 9.7098 \\ 10.3510 & 10.1607 & 9.9542 & 9.7122 & 9.4267 \\ 10.0978 & 9.9061 & 9.6859 & 9.4296 & 9.1241 \end{Bmatrix}$$

$$S_{ff1} = \begin{Bmatrix} 11.1612 & 11.1214 & 10.9673 & 10.7430 & 10.4764 \\ 11.1214 & 11.2356 & 11.1760 & 10.9937 & 10.7336 \\ 10.9673 & 11.1760 & 11.2697 & 11.1831 & 10.9711 \\ 10.7430 & 10.9937 & 11.1831 & 11.2537 & 11.1474 \\ 10.4764 & 10.7336 & 10.9711 & 11.1474 & 11.2135 \end{Bmatrix}$$

By performing an SVD, we obtain

$$U_1 = \begin{Bmatrix} 0.6655 & 0.4685 & -0.4465 & 0.3351 & -0.1607 \\ 0.4239 & -0.2456 & 0.7166 & 0.4911 & 0.0726 \\ 0.3888 & 0.2135 & 0.3803 & -0.7320 & -0.3504 \\ 0.3596 & -0.1033 & -0.1198 & -0.3149 & 0.8640 \\ 0.3112 & -0.8148 & -0.3579 & -0.1072 & -0.3157 \end{Bmatrix}$$

$$D_1 = \begin{Bmatrix} 7.2182 & 0 & 0 & 0 & 0 \\ 0 & 0.1863 & 0 & 0 & 0 \\ 0 & 0 & 0.1032 & 0 & 0 \\ 0 & 0 & 0 & 0.0207 & 0 \\ 0 & 0 & 0 & 0 & 0.0165 \end{Bmatrix}$$

$$V_1 = \begin{Bmatrix} 0.4605 & -0.6945 & 0.5094 & -0.1666 & 0.1355 \\ 0.4569 & -0.2407 & -0.4337 & 0.4939 & -0.5489 \\ 0.4493 & 0.0706 & -0.5097 & -0.0174 & 0.7301 \\ 0.4398 & 0.3337 & -0.0833 & -0.7382 & -0.3787 \\ 0.4289 & 0.5860 & 0.5344 & 0.4279 & 0.0627 \end{Bmatrix}$$

The resulting values of AIC(k) for different orders k are, neglecting constants,

AIC(1) = - 2.150
AIC(2) = - 2.264 ←
AIC(3) = - 2.243
AIC(4) = - 2.195

so that $k^* = 2$, which is the correct order, is selected. The estimated model is $y(t) = 1.843\ y(t-1) - 1.0081\ y(t-2)$.

On the interval $[d_0 + 1,\ d_0 + d_1]$, the covariance matrices were

$$S_{pp2} = \begin{Bmatrix} 1.7647 & 1.7624 & 1.6008 & 1.3214 & 1.1104 \\ 1.7624 & 1.9451 & 1.9171 & 1.7039 & 1.4781 \\ 1.6008 & 1.9171 & 2.0772 & 2.0067 & 1.8375 \\ 1.3214 & 1.7039 & 2.0067 & 2.1332 & 2.0981 \\ 1.3214 & 1.7039 & 2.0067 & 2.1332 & 2.0981 \end{Bmatrix}$$

$$S_{pf2} = \begin{Bmatrix} 1.6394 & 1.4902 & 1.4677 & 1.6660 & 2.1622 \\ 1.4905 & 1.2481 & 1.1833 & 1.3743 & 1.8518 \\ 1.2626 & 1.0238 & 1.0057 & 1.2397 & 1.7310 \\ 0.9897 & 0.8111 & 0.8554 & 1.1288 & 1.6058 \\ 0.8331 & 0.7084 & 0.7783 & 1.0178 & 1.4219 \end{Bmatrix}$$

$$S_{ff2} = \begin{Bmatrix} 1.7259 & 1.7733 & 1.9101 & 2.2212 & 2.7981 \\ 1.7733 & 2.1250 & 2.5776 & 3.1739 & 3.9927 \\ 1.9101 & 2.5776 & 3.4770 & 4.5339 & 5.7802 \\ 2.2212 & 3.1739 & 4.5339 & 6.1817 & 8.0268 \\ 2.7981 & 3.9927 & 5.7802 & 8.0268 & 10.5912 \end{Bmatrix}$$

Performing the SVD yields.

$$U_2 = \begin{Bmatrix} 0.9394 & -0.1382 & -0.2689 & 0.0560 & -0.1513 \\ 0.0073 & -0.8145 & 0.4181 & 0.3600 & 0.1791 \\ 0.1124 & -0.1507 & 0.5108 & -0.7606 & -0.3538 \\ 0.2936 & 0.5410 & 0.6957 & 0.3074 & 0.2065 \\ 0.1361 & -0.0455 & -0.0891 & -0.4408 & 0.8816 \end{Bmatrix}$$

$$D_2 = \begin{Bmatrix} 3.5041 & 0 & 0 & 0 & 0 \\ 0 & 0.6770 & 0 & 0 & 0 \\ 0 & 0 & 0.2473 & 0 & 0 \\ 0 & 0 & 0 & 0.0326 & 0 \\ 0 & 0 & 0 & 0 & 0.0094 \end{Bmatrix}$$

$$V = \begin{Bmatrix} 0.3352 & -0.7554 & 0.4483 & -0.2558 & 0.2250 \\ 0.3611 & -0.3926 & -0.3766 & 0.6815 & -0.3305 \\ 0.4033 & -0.0185 & -0.6095 & -0.6546 & -0.1926 \\ 0.4766 & 0.2819 & -0.1621 & 0.2039 & 0.7909 \\ 0.6062 & 0.4422 & 0.5093 & 0.0107 & -0.4213 \end{Bmatrix}$$

The resulting values of AIC(k) are, again neglecting constants,

AIC(1) = - 0.960
AIC(2) = - 2.266
AIC(3) = - 2.323 ←
AIC(4) = - 2.223
AIC(5) = - 2.123

so that $k^* = 3$, as desired. The estimated model is

$$y(t) = 1.9456\ y(t-1) - 1.1485\ y(t-2) + 0.0483\ y(t-3)$$

Note that, with the sparse amount of data available, the coefficient errors are relatively large and the two estimated models are relatively close to each other.

The ΔAIC criterion was used to test for a change in the time series coefficients. Since we have only one output, the criterion is

$$\Delta AIC = \log \frac{S(k^*)}{S_1(k_1^*)} + \frac{\tilde{S}(k^*) - S(k^*)}{S(k^*)} - \frac{2\ k_1^*}{d1} \qquad (6.13)$$

Using $k^* = 2$, $k_1^* = 3$, $S(k^*) = .0940$, $S_1(k_1^*) = .0726$, $S(k^*) = .0903$, $d_1 = 20$ yields,

$$\Delta AIC = 0.2583 - 0.0394 - 0.3 = - 0.0811$$

so that a "no change" decision is made, but just barely. Note that the actual covariance on the second interval using Model #1 is actually less than for the first interval, as a result of using only a small testing interval.

We next tried the test over larger intervals, keeping a 4:1 ratio between the historical interval and the testing interval. The intervals used were 160 for the historical interval and 40 for the testing interval.

The covariance matrices for the historical interval were

$$S_{ppl} = \begin{Bmatrix} 6.4177 & 6.3504 & 6.2021 & 6.0182 & 5.8335 \\ 6.3504 & 6.4177 & 6.3504 & 6.2022 & 6.0183 \\ 6.2021 & 6.3504 & 6.4170 & 6.3494 & 6.2006 \\ 6.0182 & 6.2022 & 6.3494 & 6.4153 & 6.3470 \\ 5.8335 & 6.0183 & 6.2006 & 6.3470 & 6.4119 \end{Bmatrix}$$

$$S_{pfl} = \begin{Bmatrix} 6.3504 & 6.2014 & 6.0149 & 5.8244 & 5.6495 \\ 6.2013 & 6.0147 & 5.8242 & 5.6493 & 5.4869 \\ 6.0169 & 5.8282 & 5.6544 & 5.4919 & 5.3310 \\ 5.8316 & 5.6606 & 5.4999 & 5.3387 & 5.1657 \\ 5.6655 & 5.5089 & 5.3502 & 5.1770 & 4.9871 \end{Bmatrix}$$

$$S_{ffl} = \begin{Bmatrix} 6.4187 & 6.3529 & 6.2057 & 6.0202 & 5.8294 \\ 6.3529 & 6.4251 & 6.3642 & 6.2194 & 6.0332 \\ 6.2057 & 6.3642 & 6.4436 & 6.3856 & 6.2387 \\ 6.0202 & 6.2194 & 6.3856 & 6.4671 & 6.4058 \\ 5.8294 & 6.0332 & 6.2387 & 6.4058 & 6.4835 \end{Bmatrix}$$

The results of the SVD were

$$U_1 = \begin{Bmatrix} 0.6995 & 0.4349 & -0.2706 & 0.4797 & 0.1353 \\ 0.3997 & -0.7769 & 0.1851 & 0.3162 & -0.3202 \\ 0.3573 & -0.0762 & 0.5794 & -0.3109 & 0.6589 \\ 0.3481 & 0.3521 & 0.3443 & -0.4460 & -0.6613 \\ 0.3197 & -0.2785 & -0.6620 & -0.6118 & 0.0879 \end{Bmatrix}$$

$$D = \begin{Bmatrix} 5.3574 & 0 & 0 & 0 & 0 \\ 0 & 0.1611 & 0 & 0 & 0 \\ 0 & 0 & 0.1026 & 0 & 0 \\ 0 & 0 & 0 & 0.0233 & 0 \\ 0 & 0 & 0 & 0 & 0.0070 \end{Bmatrix}$$

$$V = \begin{Bmatrix} 0.4693 & -0.8139 & 0.1915 & -0.2626 & 0.1082 \\ 0.4616 & -0.1011 & -0.5037 & 0.6123 & -0.3847 \\ 0.4489 & 0.3182 & -0.4470 & -0.2359 & 0.6647 \\ 0.4342 & 0.3743 & 0.1072 & -0.5518 & -0.5962 \\ 0.4204 & 0.2933 & 0.7059 & 0.4426 & 0.2075 \end{Bmatrix}$$

The values of AIC (k) were

AIC(1) = - 2.3072
AIC(2) = - 2.4871 ←
AIC(3) = - 2.4795
AIC(4) = - 2.467
AIC(5) = - 2.4545

so that $k^* = 2$ is selected. The estimated model is

$$y(t) = 1.6777\ y(t-1) - 0.7178\ y(t-2)$$

which is much closer to the actual model (6.11), due to the increased data length.

The model over the testing interval was next found. The covariance matrices were

$$S_{pp2} = \begin{Bmatrix} 20.6940 & 20.2655 & 19.2378 & 17.6822 & 15.7286 \\ 20.2655 & 20.4909 & 20.0699 & 19.0219 & 17.4532 \\ 19.2378 & 20.0699 & 20.3055 & 19.8666 & 18.8081 \\ 17.6822 & 19.0219 & 19.8666 & 20.0831 & 19.6334 \\ 15.7286 & 17.4532 & 18.8081 & 19.6334 & 19.8397 \end{Bmatrix}$$

$$S_{pf2} = \begin{Bmatrix} 20.4906 & 19.6758 & 18.3328 & 16.5805 & 14.5504 \\ 19.4580 & 18.1453 & 16.4322 & 14.4431 & 12.2933 \\ 17.9266 & 16.2349 & 14.2789 & 12.1690 & 9.9935 \\ 15.9897 & 14.0540 & 11.9786 & 9.8473 & 7.7296 \\ 13.7888 & 11.7306 & 9.6332 & 7.5625 & 5.5657 \end{Bmatrix}$$

$$S_{ff2} = \begin{Bmatrix} 20.9398 & 20.7220 & 19.8651 & 18.4736 & 16.6762 \\ 20.7220 & 21.1467 & 20.8735 & 19.9608 & 18.5264 \\ 19.8651 & 20.8735 & 21.2276 & 20.8927 & 19.9440 \\ 18.4736 & 19.9608 & 20.8927 & 21.1853 & 20.8230 \\ 16.6762 & 18.5264 & 19.9440 & 20.8230 & 21.0965 \end{Bmatrix}$$

The SVD yielded

$$U_2 = \begin{Bmatrix} 0.8905 & 0.4123 & 0.0935 & -0.1344 & -0.1008 \\ 0.3705 & -0.4959 & -0.6851 & 0.2227 & 0.3128 \\ 0.2259 & -0.4936 & 0.5293 & 0.5778 & -0.3023 \\ 0.1280 & -0.4076 & 0.4613 & -0.5170 & 0.5808 \\ 0.0473 & -0.4175 & -0.1701 & -0.5755 & -0.6807 \end{Bmatrix}$$

$$D_2 = \begin{Bmatrix} 9.7032 & 0 & 0 & 0 & 0 \\ 0 & 1.7256 & 0 & 0 & 0 \\ 0 & 0 & 0.0533 & 0 & 0 \\ 0 & 0 & 0 & 0.0158 & 0 \\ 0 & 0 & 0 & 0 & 0.0117 \end{Bmatrix}$$

$$V_2 = \begin{Bmatrix} 0.4557 & -0.6682 & 0.5738 & -0.1019 & 0.0784 \\ 0.4670 & -0.2754 & -0.5643 & 0.6055 & -0.1447 \\ 0.4617 & 0.0760 & -0.4483 & -0.6411 & 0.4112 \\ 0.4410 & 0.3630 & 0.1371 & -0.2325 & -0.7752 \\ 0.4081 & 0.5832 & 0.3640 & 0.3974 & 0.4504 \end{Bmatrix}$$

The resulting values of AIC(k) were

AIC(1) = 0.3771
AIC(2) = - 2.7253
AIC(3) = - 2.7595 ←
AIC(4) = - 2.6753
AIC(5) = - 2.6253

so that $k_1^* = 3$ was selected. The resulting estimated model was

$$y(t) = 2.3931\ y(t-1) - 1.977\ y(t-2) + 0.7421\ y(t-3)$$

which is much closer to the actual model (6.12) than the estimate based on half as many data points.

The ΔAIC criterion was then applied to test for a model change. Using (6.13) with $k^* = 2$, $k_1^* = 3$, $S(k^*) = 0.0811$, $S_1(k_1^*) = 0.0564$, $\bar{S}(k^*) = 0.1119$, we get

$$\Delta AIC = 0.3632 + 0.3801 - 0.15 = 0.5933 \qquad (6.14)$$

which yields a "change" decision. By comparing (6.14) to (6.13) we note several things. The first term, which is log $S(k^*)$ - log $S_1(k_1^*)$ now more strongly indicates a change, due to better model fit. The second term, which is the effect of modeling error on the measured error covariances, also more strongly indicates a change due to increased data length, which produces a more accurate estimate of the true error covariance during the testing interval, using the "no change" hypothesis. Finally, the last term more strongly indicates a change, since the bias for a "no change" decision is reduced due to increased data length. Thus we see that all three terms in the ΔAIC criterion contribute to the final decision, and each one is of importance in achieving an accurate decision.

OPTMAL ADAPTIVE IDENTIFICATION OF CHANGING SYSTEMS

7.1. Introduction

A fundamental problem in tracking a target performing evasive manuvers is adaptation to changes in the dynamical characteristics of the target motion. In previous approaches to target modeling, simplistic models have largely been used which do not take into account the changing characteristics of the target motion as different manuvers are performed by the target. To improve upon these methods requires the development of new methods that are able to adapt to the changes in target motion characteristics which may be either slowly varying or abrupt.

Previous approaches to adaptive identification and detection of abrupt changes in systems have had a number of deficiencies. Adaptive tracking of slow changes has been largely adhoc and not based upon a sound statistical theory. Much of this work has been done in the context of recursive identification using exponential weighting. This is of course very attractive from a computational point of view. For detecting abrupt changes the literature of failure detection is applicable primarily to the comparison of a limited number of specific simple hypotheses involving jumps in the states or simple actuator or sensor failures. This does not include the case of a dynamical system where abrupt changes can occur in the characteristics of the dynamics. The difficulty is the vast number of possible changes in dynamical structure and order that can occur. A further problem is that the change can occur at an arbitrary time which requires the comparison of a multitude of nonnested hypotheses which is not delt with by classical hypothesis testing theory.

The objective of this paper is to discuss these current unsolved problems in adaptive systems and to propose a new approach which is believed to solve the problem of adaptation for changing systems in a fundamental way. These current problems have been discussed to some extent in the recent dissertation of Hagglund (1983), with some approaches proposed in the context of recursive prediction error identification. In that work, a number of particular adaptive identification and detection problems were defined, and the desired properties of the solution were discussed. The proposed solutions were shown to work on simple low order systems. However, a number of problems were not addressed that occur in higher order and multivariable systems. These difficulties in include the lack of invariance of the procedures in general, the nonexistence of a global, and the lack of a procedure of the determination of an appropraiate model state order. These are indeed very difficult problems some of which have not been completely solved even for the offline case.

In this paper we share much of the intent stated in Hagglund, but take a much more general approach to solving these problems. The Kullback–Leibler information (1959) or entropy measure of model approximation error has recently been shown to follow naturally from the fundamental statistical principles of sufficiency and repeated sampling in a predictive inference context (Larimore, 1983). The Akaike information criterion (Akaike, 1973) is an unbiased estimate of the entropy measure which is optimal for large samples (Shibata, 1981). In this paper, the entropy and predictive inference approach is applied to the problems of adaptive tracking of slowly changing processes and abrupt changes. This requires the extension of the AIC procedure to the case of comparing different models over different data intervals whereas the AIC procedure was originally developed for comparisons of different models on the same data interval. The concepts and notation developed in Larimore (1983, 1985) is used in the development.

The structure of the paper is as follows. The approach to the general problem of adaptation is discussed in the next section. To generalize the AIC procedure, some properties of maximum likelihood estimators are derived for nonnested classes of models in section 7.3 Section 7.4 gives a generalization of the AIC estimate of negentropy. Section 7.5 discusses the use of this generalized AIC in adaptative tracking while section 7.6 describes its use in abrupt change detection.

7.2. Approach to Adaptation

The use of predictive inference and entropy concepts and methods is the basic approach taken in adaptation to system changes. To formulate the problem, consider a division of a time span into a set of disjoint intervals whose union is the whole time span. With each time interval we associate a model for the dynamical system which is determined from the observed data using a particular model selection method. In this approach, different divisions of the time span into time intervals are considered as well as different model selection procedures. This allows the consideration of very general models that include slow as well as abrupt changes as described below. The details of this are given in Sections 7.5 and 7.6, with a brief overview of the concepts given here.

Consider first the case of tracking slow changes. Suppose a given time span is divided into a set of time intervals. Several different intervals sets will be considered involving divisions of the time span using different interval lengths. On each time interval of each interval set, a best model will be determined by choosing a model from a class of models using say the AIC procedure. The class of models considered can include different state orders and other structural characteristics. For each set of intervals, a composite model consists of the models associated with the various intervals of the interval set. Composite models for several interval sets can be compared to determine which division of

the time span into intervals leads to the bet composite model. If the intervals are chosen to be too long, then changes in the process during the intervals will lead to increased prediction error. On the other hand if the time intervals are too short, then variability in observations and the use of an increased number of total parameters in the composite model will also increase the prediction error. As a result, there will be an optimal choice of division of the time span into time intervals. In practice it is not necessary to obtain the optimum division but only a good approximation which will determine an approximately optimal model update rate for re–identifying the model dynamics and noise process characteristics.

Now consider the case of abrupt change detection. suppose that the above tracking of slowly changing process characteristics is done and the optimal slowly changing model is identified. Then suppose that an abrupt change occurs at an unspecified time following the end of the time span used in tracking the slow changes but within the update rate used in slow tracking. We consider the new time span that includes the failure and consider divisions of it into intervals. On each interval, a model is fitted using say AIC to determine a model from a class of models. Now the principle question is whether the fitted models on the various intervals are significantly different from the last model chosen by the slowly changing adaptive procedure.

To actually make the proposed comparisons for adaptive tracking and abrupt change detection requires the development of new results in predictive inference since different models are being compared across different time intervals, and non–nested multiple comparisons are also involved. Neither of these cases is considered in the previous theory of AIC and related predictive measures. This theory is developed in the following sections.

7.3. Constrained Maximum Likelihood Estimation

In this section, properties of the maximum likelihood parameter estimates are developed for the case that the true probability model is not contained in the class of parameterized densities that are considered for inference. The classical development of the asymptotic consistency and minimum variance of maximum likelihood estimators is for the case where the true density id contained in the parametric class.

The predictive inference framework as in Larimore (1985) is adopted here with $p(q,\theta)$ the parameterized probability density where θ is a vector of parameters, q is the informative sample and r is the predictive sample. Suppose that the parameter vector $\theta^T=(\theta_1,\theta_2,...)$ is a finite or infinite set of parameters, and for each subset of distinct positive integers $k=(k_1,...,k_m)$ consider the subspaceΘ_k of θ such that only the corresponding $\theta_{k_1}, ...,\theta_{k_m}$ are nonzero where θ^k denotes a member ofΘ_k, and let $\mathbb{C}_k$ be the class of models $\mathbb{C}_k=\{p(q,\theta^k),\epsilon\Theta_k\}$, These classes of models are in general nonnested so that we do not in general have $\mathbb{C}_k c \mathbb{C}_j$ or $\mathbb{C}_j c \mathbb{C}_k$. The maximum likelihood estimator for the class $\mathbb{C}_k$ will be denoted as $\hat{\theta}q)$.

The development of the maximum likelihood theory is straight forward for the case where Taylor series expansions are possible. This holds under the following regularity conditions (Cox & Hinkley, p. 281):

(i) The parameter space is closed and compact.

(ii) The probability distributions defined by any two different values of 0 are distinct.

(iii) The first three derivatives of the log likelihood $\ell q,\theta)$ with respect to 0 exists in the neighborhood of the true parameter value almost surely. Further, in such a neighborhood, n^{-1} times the absolute value of the third derivative is bounded above by a function of q, whose expectation exists.

In particular, these conditions permit the interchange of expectation and differentiation up to second order.

In the discussion various order models are considered, and the relationships between the various orders is developed. The log likelihood function of the informative sample q will be denoted by $\ell(q,\theta)$, and the gradient row vector and Hessian matrix denoted $\ell'(q,\theta)$ and $\ell''(q,\theta)$ respectively. Expectation, denoted E, will be with respect to the true density $p(q,\tilde{\theta})$ unless stated otherwise where $\tilde{\theta}$ onto Θ_k as the parameters $\theta^k \in \Theta_k$ minimizing the negentropy R_q relative to the informative sample q

$$R_q(\tilde{\theta},\theta^k) = E\ell(q,\theta^k) \qquad (7.1)$$

At the minimum $\tilde{\theta}^k$, the gradient of (3–2) is zero so from the regularity conditions

$$E\ell'(q,\tilde{\theta}^k) = \theta,$$

and the minimum is unique if and only if the expected Hessian, denoted $D_q^k = E\ell''(q,\tilde{\theta}^k)$. Expanding (7–1) in a Taylor series gives a second order expression for the information distance which holds asymptotically for large sample size of the informative sample

$$R_q(\tilde{\theta},\theta^k) = E[\ell(\tilde{\theta}^k) - \ell(\tilde{\theta}^k)] + E[\ell(\tilde{\theta}) - \ell(\tilde{\theta}^k)]$$

$$= -E[\ell'(\bar{\theta}^k)(\theta^k - \bar{\theta}^k)] - E[\frac{1}{2}(\theta^k - \bar{\theta}^k)^T \ell''(\bar{\theta}^k)(\theta^k)(\theta^k - \bar{\theta}^k)] + E\ell(\bar{\theta}) - \ell(\bar{\theta}^k)]$$

$$= -\frac{1}{2}\| \theta^k - \bar{\theta}^k \|^2_{D^k_q} + Rq(\bar{\theta}, \bar{\theta}^k)$$

To determine the moments of the maximum likelihood estimates $\hat{\theta}^k$, consider the first order equality

$$0 = \ell'(q, \bar{\theta}^k) = \ell'(q, \bar{\theta}^k - \bar{\theta}^k)^T \ell''(q, \bar{\theta}^k)$$

Taking expectation with respect to the true density and using (7–2) gives the equation

$$D^k_q(E\hat{\theta}^k - \bar{\theta}^k) = 0$$

that holds asymptotically for large informative sample N. For θ^k identifiable, i.e. $\bar{\theta}^k$ unique, D^k_q is nonsingular which implies that to first order

$$E\hat{\theta}^k = \bar{\theta}^k$$

Now using (7–4), the covariance of the estimation error is

$$E(\hat{\theta}^k - \bar{\theta}^k)(\hat{\theta}^k - \bar{\theta}^k)^T = (D^k_q)^{-1} E\{\ell'^T(q, \bar{\theta}^k)\ell'(q, \bar{\theta}^k)\}(D^k_q)^{-1}$$

Note that in the unconstrained case, the middle term which is the Fisher information matrix is equal to minus the expected Hessian D^k_q, but this is not in general true for the constrained case.

7.4 Estimation of Entropy

For decision on model parametric order and structure, it it necessary to estimate the negative entropy based on the sample. One such procedure is due to Akaike (1973. We consider the case where the informative sample q and the predictive sample r are independent. For each selection of a parameter subset $k = (k_1,....,k_m)$, the Akaike information criterion for comparing the maximum likelihood estimators is

$$AIC(k) = -2\log p(q,\hat{\theta}^k(q)) + 2K(k)$$

where K (k) is the number of parameters, i.e. the dimension on θ^k. The minimum AIC estimator (MAICE), denoted $\hat{\theta}_A(q)$, is $\hat{\theta}_A(q)=\hat{\theta}^{\hat{k}(q)}(q)$ where $\hat{k}(q)$ is the parameter set minimizing AIC (k) is an unbiased estimator of the negative entropy (7–1) based upon the informative sample and the assumed model structure. The predictive sample is essentially replaced by the information sample, and the term 2K(k) is an adjustment for the bias due to the correlation between the informative sample q and the estimate $\hat{\theta}^k(q)$.

Following Akaike, we use the maximized log likelihood $\ell_q(\hat{\theta}^k) = \ell(q,\hat{\theta}^k(q))$ as an estimate of the relative entropy and compute the bias in the procedure. Consider the expected log likelihood difference using (7–3)

$$E[\ell(\tilde{\theta}) - \ell(\hat{\theta}^k)] = E[\ell(\tilde{\theta}^k) - \ell(\hat{\theta}^k)] + E[\ell(\tilde{\theta}) - \ell(\tilde{\theta}^k)]$$

$$= - E[\ell'(\tilde{\theta}^k)(\hat{\theta}^k - \tilde{\theta}^k)] - E[\tfrac{1}{2}(\hat{\theta}^k - \hat{\theta}^k)^T \ell''(\tilde{\theta}^k)(\hat{\theta}^k - \tilde{\theta}^k)] + E[\ell(\tilde{\theta}) - \ell(\tilde{\theta}^k)]$$

$$E[(\hat{\theta}^k - \tilde{\theta}^k)^T \ell''(\tilde{\theta}^k)((\hat{\theta}^k - \hat{\theta}^k - \tilde{\theta}^k)] + R(\tilde{\theta},\tilde{\theta}^k)$$

$$= -\mathrm{tr}I_{\dim(\theta^k)} + R(\bar{\theta},\bar{\theta}^k) = -\dim(\theta^k) + R(\bar{\theta},\bar{\theta}^k)$$

where the third equality follows using Equation (7–4) which is satisfied by the maximum likelihood estimate, using the asymptotic equivalence of the negative Fisher information and the Hessian in (7–7), and the expression (7–3) for the negentropy. Consider the case of fitting two models $\hat{\theta}^k$ and $\hat{\theta}^j$, and consider the expected difference of the maximized log likelihoods

$$E[\ell(\bar{\theta}) - \ell(\hat{\theta}^k)] = E[\ell(\bar{\theta}^k) - \ell(\hat{\theta}^j)] - E[\ell(\bar{\theta}) - \ell(\hat{\theta}^k)]$$

$$= + \dim(\theta^k) - \dim(\theta^j) + R(\bar{\theta},\hat{\theta}^j) - R(\bar{\theta},\hat{\theta}^k)$$

Thus for relative comparisons among hypotheses based on a given sample, an unbiased estimate of twice the negentropy $E[\ell(\bar{\theta}) - \ell(\hat{\theta}^k)]$ is given by the Akaike information criterion. Note that the proof of this is much more general than that originally given by Akaike (1973) since it applies to the general case of comparisons of nonnested structures. Also, the true parameter $\bar{\theta}$ need not be contained in the structures being compared so long as the Fisher information matrix is a constant in a neighborhood including the true parameter and its projection onto the subspaces of these structures.

7.5 Adaptation to Slow Changes

The key issue in adaptation to slow changes is the choice of the rate of adaptation. Previous approaches to slow adaptation have been largely heuristic and not based upon sound statistical principles. In this section the principles and concepts of predictive inference are used to derive a procedure for choosing an adaptation rate which minimizes the prediction error for an independent sample from the process.

Consider the problem of choosing the optimal rate at which to identify the system. we consider the problem as stated in Section 7.2 where it is assumed that the system is slowly changing, and that based upon the observed data we wish to determine a best rate to identify the system. Consider dividing the data over an interval of $L = 2^{\ell}$ samples, where ℓ is an integer, into $H = 2^{h}$ subintervals each with length $2^{\ell-h}$. Suppose for a given h that on each of the intervals I_j for 1,2,3,...2^{h} the best state space model $M_{h,j}$ is determined using the minimum AIC estimate (MAICE) criterion,

To provide some motivation, we consider the negentropy as expressed in (7–3). For the case of a constant parameter model, consider the effect of the number of the parameters and the bias in choosing too low an order model. The negative entropy depends upon the particular parameter estimation procedure, but for large samples it is bounded from below by

$$ER(\tilde{\theta},\hat{\theta}^k) = \frac{-1}{2}\|\hat{\theta}^k - \tilde{\theta}^k\|_D^2 k + ER(\tilde{\theta},\tilde{\theta}^k) \geq \frac{1}{2}K(k) + \sum_{j=1}^{2^h} ER_{,}(\tilde{\theta},\tilde{\theta}^k)$$

where j is a interval index. This lower bound is accurate for $\tilde{\theta}^k$ near $\tilde{\theta}$. The first term K_k is the sampling variability of the optimal estimator. The second term is the bias due to constraining the parameter estimates $\hat{\theta}^k$ to lie in the subspace Θ_k which increases with increasing sample size since the parameters of the time varying process are not constant.

On the interval L we consider the model selection procedures M_h for h = 1,...H as above, i.e., procedure M_h fits the composite model consisting of the time invariant models $M_{h,j}$ for j =1,2,3,...2^{h} fitted on each of the 2^{h} subintervals. For a given h the model selection procedure M_h is identical to the maximum likelihood problem of estimating the parameters $\theta_{h,j}$ of the joint models $M_{h,j}$ for j =1,...,2^{h}. Then for a given h the MAICE for

the model selection procedure M_h is

$$\text{MAICE}(M_h) = -2 \log p\,(I_L, \hat{\theta}_h) + 2K_h(k_h)$$

$$= \sum_{j=1}^{2^h} [-2 \log p\,(I_j, \hat{\theta}_{h,j} \mid I_{j-1}) + 2K_{h,j}] = \sum_{j=1}^{2^h} \text{MAICE}\,(M_{h,j})$$

where the interval $I_{h,j}$ denotes the appropriate data.

We wish to choose the data length $\hat{D} = L2^{-\hat{h}}$ corresponding to the $\hat{h}$ minimizing $\text{MAICE}(M_h)$. This procedure gives a very sensitive comparison of different rates 2^{-h} for identifying the model, or equivalently the data length $D = 2^{\ell-h}$ for identifying a model. The tradeoff between sampling variability form using too small a data length and bias from using too long a data length is seen by the effect of data length on the two terms in the MAICE criterion. Too little data introduces variability in the log likelihood function and a penalty for more parameters, while too much data reduces these but introduces bias in the model due to modeling a changing process as one with constant parameters. The optimum is achieved when the respective rates of decrease and increase are balanced.

7.6 Detecting Model Changes Across Different Data Sets

In Section 7.4 the AIC was shown to give an unbiased estimate for choosing a model structure for a give set of data. In this section, we consider the problem of determining if there is a change in the process between tow different data sets. The detection problem that we consider is where the process is modeled as a slowly changing process using some efficient procedure such as given in Section 7.5. Suppose that from such a procedure a model is given for an interval of data Q_1 and that over a later interval of data set Q_2

second model is fitted. The data lengths of the two sets are generally different with the second set usually much shorter. In the context of the scheme in Section 5 for fitting optimal fixed parameter models over many intervals of various lengths, the two intervals we wish to determine if there has been a significant departure in the process characteristics between the two data sets.

Since the model, denoted M_1, fitted on data set Q_1involves a near optimal selection of the data length, the model M_1 provides a best prior model when Q_1 is chosen as the most recent optimal length interval preceding Q_2. To detect any abrupt changes in the system, we wish to compare on the joint data set (Q_1,Q_2) the fit of the composite model (M_1,M_2) with the composite model (M_1,M_1) taking into account the fact that the models M_1 and M_2 are fitted over the respective intervals Q_1 and Q_2. To make this comparison, we seek an unbiased estimate of the difference between the negentropies of these two composite models.

The Markov stucture can be used to make the two samples essentially independent by conditioning the observations on the past. The joint distribution of the two data intervals is

$$p(Q_1,Q_2) = p(Q_1)p(Q_2|\ Q_1)$$

Thus we suppose that the models on each of the two sets are fitted to the conditional data given the past. Since the model on the first data set Q_1 is the same for both composite hypotheses, the first term $p(Q_1)$ is the same for both composite models. Thus we need only compare the negentropy difference between the conditional model $M_1(Q_2|\ Q_1)$ and the conditional model $M_2(Q_2|\ Q_1)$ both fitted on the interval Q_2.

Denoting the parameter estimates of the models Q_1 and Q_2 by θ^1 and θ^2 respectively, we compare the log likelihoods on the random variables Q_2 given Q_1. The expected difference of the log likelihoods of the two models is

$$E[\ell(\hat{\theta}^1) - \ell(\hat{\theta}^2)] = E[\ell(\bar{\theta}) - \ell(\hat{\hat{\theta}}^2)] - E[\ell(\bar{\theta}) - \ell(\hat{\theta}^1)]$$

$$= -\dim(\theta^2) + R(\bar{\theta},\hat{\theta}^2) - R(\bar{\theta},\hat{\theta}^1)$$

where the term $\dim(\theta^1)$ is not present since the estimate $\hat{\theta}^1$ is a function only of the sample Q_1 which is independent of the sample $Q_2 | Q_1$. Thus an unbiased estimate of the difference of negentropies $R(\bar{\theta},\hat{\theta}^2)$ of the two models is

$$\ell(\hat{\theta}^1) - \ell(\hat{\theta}^2) + \dim(\theta^2)$$

This gives a test for the occurrence of an abrupt change between the two data intervals. Depending upon the nature of the change and the process characteristics, the bet detection interval will vary. Some changes give most of the information about the change over a short interval while others have a cumulative effect and require a long time interval to detect.

7.7 Detection of Slow Model Changes

We now consider the detection of slow, essentially continuous, model changes. What we wish to achieve in this case is an appropriate data length over which to fit models. If the data length is too short, then there will be a tendency to over–fit the model to the noisy data, leading to larger prediction errors. If the data length is too long, then the effects of parameter variations will begin to dominate the prediction errors.

In order to generate an appropriate measure by which to trade off these two characteristics, we again use the AIC criterion, but in a different way. Assume we have data over a time interval $I = \{1,2,...,n\}$ and suppose we divide this interval into subintervals of length W:

$$I_1 = \{1,2,..., W\}$$

$$I_2 = \{W+1,\ W+2, ..., 2W\},$$

etc.

Then an appropriate measure for data length determination is the average per sample entropy. In terms of the AIC criterion we define

$$\overline{AIC}_W = \frac{1}{N_W} \sum_i \frac{1}{W} AIC_p(k_i^*)$$

where $AIC_p(k_i^*)$ is the minimum prediction AIC for the i^{th} interval I_i, k_i^* is the optimal model order for the i^{th} interval, and N_W is the number of intervals of W over the whole data interval I. The prediction AIC uses the forward prediction error variance (cf eq. 5.9)) rather than the fit error variance, since we are interested in the error over the next interval, not the one over which the model was fitted. This has the effect of increasing the penalty on the number of parameters in the AIC criterion.

The form of AIC_p is

$$AIC_p(k_i^*) = AIC(k_i^*) + M(k_i^*)$$

Experimental Results

In order to test this criterion as the basis for data length selection, we considered a second–order AR model

$$y(t) = a_1(t)\, y(t-1) + a_2(t)\, y(t-2) + n(t)$$

where n(t) was zero–mean white gaussian noise with unit variance. The time varying coefficients were selected so that the two system poles were on the unit circle in the z–plane. This yields: $a_1(t) = 2 \cos \theta(t)$, $a_2 = -1$ where the roots are: $\cos \theta(t) - i \sin \theta(t)$ and $\cos \theta(t) - i \sin \theta(t)$. The time–variation of $\theta(t)$ was selected as $\theta(t) = \theta(0) + 2 \pi f t$, where f is a selected frequency. Two values of f (.0001, .001) were used in the experiments. Total data length was 1000 time points. The results for Case 1 (f = 0.001) are shown in Table 7.1, using $\theta(0) = 0.2$. The result is that the optimal indicated data length is 10–12 samples and corresponds to the case in which the average coefficient change over the fit window is in the range of 0.80 – 0.96. Over the entire data length of 1000 samples, the value of a_1 starts at 1.64, deceases to 1.90 at t = 400 and then increases to 1.90 at t = 1000. Thus, the average coefficient change over the optimum data length is generally more than 40% of the coefficient value.

Table 2 shows the results for Case 2 (f = 0.0001) in which the optimal data length is found to be 30 samples. This is, of course, increased over that of Case 1 since the coefficients vary much less rapidly – on the order of 0.019, on the average. Note that the rms prediction error σ_e evaluated over the fit set generally decreases monotonically with data length and cannot be used as a selection criterion.

CHARTER 8

E–M Algorithm for Adaptive Time Series Analysis

In this chapter, we explain the basic properties of the E–M Algorithm based on Dempster, Laird and Rubin (1977) and describe its application to Adaptive Time Series Analysis The E–M Algorithm involving iteration of E (Estimation) and M (Maximization) steps is a general procedure for maximum likelihood estimation (MLE) of models from incomplete data. We show that CVA–Regression algorithm of previous chapters implements E and M steps in a special way. Based on the recognition of this fact, we show how the CVA–Regression approach can be generalized to obtain MLE of the state space model. In addition, we show how the algorithm can be implemented recursively to allow for real–time identification and extension to time–varying systems. We also consider extensions to missing data, nongaussianstatistics and ARMA models.

8.1 E–M algorithm – Basic Properties:

The basic motivation for the E–M algorithm comes from the fact that in numerous estimation problems, only partial observations of all the states or underlying causal factors are available. If the observations of the complete state or factors were available, the estimation problem would become simple. The E–M algorithm estimates the complete state given the observed data and then estimates parameters after construction of the complete data set. This process is repeated in such a way that the likelihood function increases with each iteration of the E–M algorithm, till convergence to a local or global maximum of the likelihood function is achieved.

The simplest example of the application of the E–M algorithm is for the case of missing data. The E–step involves estimating the missing data points and the M step is based on the procedure used when there is no missing data. Regression models with

missing data points can be estimated in this fashion.

The Markovian modeling approach to Time Series Analysis lends itself to the E–M approach. In state space models for Markov processes, all the states are generally not observed. It is well known that if all the states are observed, the system identification problem is solved easily by Regression or Ordinary Least Squares (OLS). For partial state observations, estimation of the full state using a Kalman Filter requires knowledge of the parameters. A natural approach based on the E–M algorithm is to assume the parameters, estimate states and update the parameters using the estimated states. Even though the concept is simple and has been proposed earlier in the literature of system identification, both the E and M steps must be carried out properly so that the likelihood function increases at each iteration and the parameters converge to the ML estimates. Therefore the conditions under which the E–M algorithm converges must be examined carefully.

We state here the basic results from Dempster, Laird & Rubin (1977) on the application and convergence of the E–M algorithm. We, then, apply these results to the problem of state space model identification using CVA and E–M algorithms.

Assume two sample spaces S and Y and a many–to–one mapping from S to Y. The observed data y is a realization from Y. The corresponding state s in S is not observed directly, but only indirectly through y. The sampling density defined for all s in S is $f(s|\phi)$, where ϕ denotes a parameter vector. The corresponding sampling density for y is Y is obtained as

$$g(y|\phi) = \int_{S(y)} f(s)|\phi)ds \qquad (8.1)$$

The EM algorithm attempts to find $\hat{\phi}$ which maximizes $g(y|\phi)$ given y by making an essential use of the density $f(s|\phi)$.

One of the key relationship used in the derivation of the E–M algorithm is

$$g(y|\phi) = \frac{f(s|\phi)}{p(s|y,\phi)} \qquad (8.2)$$

where $p(s|y,\phi)$ denotes conditional distribution of s given y and ϕ. Eq. (8.2) follows from the fact that

$$g(y|\phi)\, p(s|y,\phi) = p(s,y|\phi)$$

and $$p(s,y|\phi) = f(s|\phi)$$

since y is a subset of s. Taking logs on both sides of Eq. (8.2), we obtain the log–likelihood function

$$L(\phi) = \log g(y|\phi)$$

$$= \log f(s|\phi) - \log p(s|y,\phi) \qquad (8.3)$$

For the case of exponential family of probability distributions, which includes the gaussion case, Eq. (8.3) simplifies greatly and leads to very useful results. We first note that an exponential distribution can be written in the form,

$$f(s|\phi) = b(s) \exp(\phi\, t(s)^T)/a(\phi) \qquad (8.4)$$

where t(s) is a 1 x r row vector of sufficient statistics for the complete data. The 1 x r row vector parameterization ϕ is unique up to an arbitrary non–singular r x r linear

transformation, as is the corresponding choice of $t(s)$. Based on a property of the exponential distributions, $p(s|y,\phi)$ is also exponential and has the form

$$p(s|y,\phi) = b(s) \exp(\phi t(s)^T)/a(\phi(y)) \qquad (8.5)$$

$$\text{where } a(\phi(y)) = \int_{S(y)} b(s) \exp(\phi t(s)^T) ds \qquad (8.6)$$

The key property here is that both $f(s|\phi)$ and $p(s|y,\phi)$ are from the same exponential family with the same natural parameters ϕ and the same sufficient statistics $t(s)$, but are defined over different sample spaces S and $S(y)$.

Eq. (8.3) can now be written as

$$L(\phi) = -\log a(\phi) + \log a(\phi(y)) \qquad (8.7)$$

The first partial derivative of $L(\phi)$ or the score function is obtained from Eq. (8.6) & (8.7) as

$$\frac{\partial L(\phi)}{\partial \phi} = -E(t(s)|\phi) + E(t(s)|y,\phi) \qquad (8.8)$$

where E(.) denotes expectation over the appropriate sample spaces viz. S and S(y) respectively in Eq. (8.8).

Since $\frac{\partial L}{\partial \phi} = 0$ is a necessary condition at the maximum of the likelihood function, the MLE $\hat{\phi}$ must satisfy the condition.

$$E\,(t(s)|\hat{\phi}) = E\,(t(s)|y,\hat{\phi}) \qquad (8.9)$$

The E–M algorithm achieves (8.9) iteratively as follows:

E–Step: Estimate t(s) during kth iteration as,

$$\hat{t}^{(k)}(s) = E\,(t(s)|y,\phi^{(k)}) \qquad (8.10)$$

M–Step: Obtain updated parameter vector, $\phi^{(k+1)}$ by solving the eqn.

$$E\,(t(s)|\phi) = \hat{t}^{(k)}(s) \qquad (8.11)$$

The E–M algorithm of Eq. (8.10) and (8.11) defines a parameter mapping $\phi^{(k)} \rightarrow \phi^{(k+1)}$ as

$$\phi^{(k+1)} = M\,(\phi^{(k)}) \qquad (8.12)$$

This mapping has several interesting properties which are given in Dempster, Laird & Rubin (1977) and Wu (1983). In particular, it leads to a monotonic increase in the likelihood function $L(\phi)$. The conditions under which the parameter sequences $\phi^{(k)}$ converges to the MLE $\hat{\phi}$ and the rates of convergence are given in the above papers. These conditions are verifiable for exponential family. The Jacobian of the mapping (8.12) is given by the expression.

$$\frac{\partial M(\hat{\phi})}{\partial \phi} = V\,(t|y,\hat{\phi})\;V\,(t|\hat{\phi})^{-1} \qquad (8.13)$$

where V(.) denotes covariance operator. The rate of convergence is determined by the eigenvalues of the above Jacobian which in turn depend on the information loss due to incompleteness of the data.

8.2 MLE of State Space Model Parameters Using E–M Algorithm:

Consider a state space model in the usual notation:

$$x(k+1) = F\,x(k) + G\,w\,(k) \qquad (8.14)$$

$$y(k) = H\,x\,(k) + u(k) \qquad (8.15)$$

$$k = 1,2,.....N$$

Let ϕ denote all unknown parameters in the above model. It is required to estimate ϕ given output data, $Y_1^N = \{y(1),...y(N)\}$.

It is obvious from Eq. (8.14) and (8.15) that if the full state $\{x(k)\}_{k=1,N}$ was known, matrices F,G and H could be estimated using Regression or OLS. We, therefore, define the complete data set as:

$$S_1^N = \{ Y_1^N, X_1^N \} \qquad (8.16)$$

The complete – data density function is

$$f(S_1^N|\phi) = f(Y_1^N, X_1^N|\phi)$$

$$= f(Y_1^N | X_1^N, \phi)\, f(X_1^N | \phi)$$

$$= \prod_{j=1}^{N} f(Y(j) | X(j), \phi)(X(j) | X(j-1), \phi) \qquad (8.17)$$

Using Eq. (8.14) and (8.15), we can write

$$f(y(j) | x(j), \phi) \sim N(Hx(j), R) \qquad (8.18)$$

$$f(x(j) | x(j-1), \phi) \sim N(Fx(j-1), GQG^T) \qquad (8.19)$$

We now try to express $f(S_1^N | \phi)$ in the form (8.4) to identify $t(S_1^N)$ and ϕ.

$$f(S_1^N | \phi) = \frac{1}{(2\pi)^{\frac{Nn}{2}} |R|^{\frac{N}{2}} |GQG|^{\frac{N}{2}}} \exp -1/2 \left\{ \sum_{j=1}^{N} ||y(j) - Hx(j)||^2_{R^{-1}} \right.$$

$$\left. + ||x(j) - Fx(j-1)||^2_{(GQG^T)^{-1}} \right\} \qquad (8.20)$$

The term outside the exponent represents $b/a(\phi)$. The term in the exponent within brackets represents $\phi\, t(S_1^N)^T$ as follows:

$$\phi\, t(S_1^N)^T = \mathrm{Tr}\{R^{-1} \sum_{j=1}^{N} (y(j) - Hx(j))\,(y(j) - Hx(j))^T\}$$

$$+ \mathrm{Tr}\,(GQG^T)^{-1} \sum_{j=1}^{N} \{(x(j) - Fx(j-1))(x(j) - Fx(j-1))^T\} \qquad (8.21)$$

The next step is to redefine the unknown parameters in such a way that they appear

linearly in the above equation. The terms multiplying a given parameter, then, define the associated sufficient statistic for the estimation of that parameter. As an example, consider the case in which only the noise covariances R and GQG^T are unknown.

This is a very common case in Adaptive Kalman Filtering and correlation type methods for the estimation of these matrices were considered in Mehra (1970). E–M algorithm leads to a new method for the MLE of noise covariances, which should have great practical significance. Based on Eq. (8.21), it is better to estimate R^{-1} and $(GQG^T)^{-1}$. The sufficient statistics are:

$$t_1 = \sum_{j=1}^{N} \{(y(j)-Hx(j))(y(j)-H\,x(j))^T\} \quad \text{for } R^{-1}$$

This is intuitively obvious since the above quantities are sample covariances of v and $\underset{\sim}{G}\underset{\sim}{w}$.

Before proceeding to the more general case of additional unknowns {F,H}, we derive the E–M algorithm for the above case.

E–Step:

$$\hat{t}_1^{(k)}(S_1^N) = E[t_1(S_1^N) \mid y^N, \phi^{(k)}]$$

$$= \sum_{j=1}^{N} \{(y(j) - H\hat{x}(j|N))\,(y(j) - H\hat{x}(j|N))^T$$

$$+ H\,P\,(j|N)\,H^T\} \qquad (8.22)$$

where $\hat{x}(j|N)$ denotes the smoothed estimate of $x(j)$ based on Y_1^N and $\phi^{(k)}$. The associated covariance is denoted by $P(j|N)$.

The equations for the computation of $\hat{x}(j|N)$ and $P(j|N)$ are well known in the control literature. See (Bryson and Ho (1975)). The computations can be carried out recursively using a Kalman Filter and a backward sweep. Alternatively, $\hat{x}(j|N)$ can be written as a weighted average of a forward Kalman filter state estimate $\hat{x}(j|Y_1^j)$ and a backward Kalman filter estimate $\hat{x}(j|Y_{j+1}^N)$, (see Mehra (1968)).

Similarly t_2 is estimated as

$$\hat{t}_2^{(k)}(S_1^N) = \sum_{j=1}^{N} (\hat{x}(j|N) - F\hat{x}(j-1|N-1))\,(\hat{x}(j|N) - F\hat{x}(j-1|N-1))^T + P(j|N) + FP(j-1|N)\,F^T - C_N(j,j-1)\,F^T - F\,C_N^T(j,j-1) \qquad (8.23)$$

where $C_N(j,j-1)$ denotes the correlation between the smoothing errors at j and j–1. An expression for $C_N(j,j-1)$ can be found from the smoothing equations.

The E–step, therefore, consists of running two Kalman filters or a filter/smoother and solving Eq. (8.22) and (8.23). The calculations can be made recursive in data length N.

M–Step: For this step, we need to evaluate $E(t_1|\phi)$ and $E(t_2|\phi)$, expressing them as functions of ϕ, equating them to the values for $\hat{t}_1$ and $\hat{t}_2$ obtained in the E–step and solving for ϕ. This is quite straight–forward based on the state

$$E(t_1|\phi) = \sum_{j=1}^{N} E[v(j)\,v^T(j)] = NR \qquad (8.23)$$

$$E(t_2|\phi) = \sum_{j=1}^{N} E\{Gw(j)w^T(j)G^T\} = N\,GQG^T \qquad (8.24)$$

Therefore,

$$\hat{R}^{(k+1)} = \frac{1}{N}\sum_{j=1}^{N}\{(y(j) - H\hat{x}(j|N))\,(y(j) - H\hat{x}(j|N))^T$$

$$+ H\,P(j|N)\,H^T\} \qquad (8.25)$$

$$(G\hat{Q}G^T)^{(k+1)} = \frac{1}{N}\,\Sigma\{(\hat{x}(j|N) - F\hat{x}(j-1|N))\cdot$$

$$(\hat{x}(j|N) - F\hat{x}(j-1|N))^T + P(j|N) + F\,P(j-1|N)\,F^T$$

$$- C_N(j,j-1)\,F^T - F\,C_N^T(j,j-1)\} \qquad (8.26)$$

We now consider the case of unknown $\{F,R\}$. In this case, the estimation of R remains unchanged, but the sufficient statistics for the estimation of F are

$$t_3 = \sum_{j=1}^{N} x(j)\,x^T(j-1) \text{ and } t_4 = \sum_{j=1}^{N} x(j)x^T(j).$$

$$E(t_3|\phi) = NF\,\Sigma_{xx}$$

$$E(t_4|\phi) = N\,\Sigma_{xx}$$

where Σ_{xx} is the covariance of $x(j)$.

Using the E–step.

$$\hat{F}^{(k+1)} = (\hat{t}_4^{(k)})^{-1}\, \hat{t}_3^{(k)} \qquad (8.27)$$

where $\hat{t}_3^{(k)} = \sum_{j=1}^{N} \{\hat{x}(j|N)\hat{x}^T(j-1|N) + C_N(j,j-1)\}$

$$\hat{t}_4^{(k)} = \sum_{j=1}^{N} \{\hat{x}(j|N)\hat{x}^T(j|N) + P(j|N)\}$$

The above results are easily generalized to the case of unknown $\{F,H,R, GQG^T\}$. However, identifiability issues must be considered. In particular, F and H must be put in a canonical form. If output canonical form is used, H consist of zeros and ones and F has no more than np parameters where n is the state dimension and p is the output y dimension.

8.2.1 Relationship of E–M algorithm to direct MLE:

Direct MLE of parameters ϕ is given in Appendix B involves maximization of

$$L(\phi) = \log p\,(Y_1^N|\phi)$$

The computational aspects are discussed in Gupta and Mehra (1974). It is interesting to note the similarities and differences between the two approaches. In particular, if the smoothing form of $L(\phi)$ given in Schweppe (1973) is used, the similarities are quite striking. The main difference is that the Kalman Filter/Smoother sensitivity equations involving $\frac{\partial \hat{x}}{\partial \phi}$ and $\partial P/\partial \phi$ do not have to solved, resulting in significant computational savings. On the other hand, the rate of convergence of the Gauss–Newton iteration is probably faster,

though the convergence may be more problematic, especially where some parameters are unidentiable resulting in a singular Fisher Information Matrix. A comparison of the two methods on practical problems would be of great interest.

8.3 E–M Algorithm for ARMA models:

ARMA (p,q) model has the form

$$y(t) = \sum_{i=1}^{p} a_i\, y(t-i) + u(t) + \sum_{i=1}^{q} b_i\, u(t-i) \qquad (8.28)$$

$$t = 1,2,.....N$$

where $\{u(t)\}$ represents a random shock series which is zero mean, Gaussian and white with variance σ_u^2. Parameter vector ϕ is a (p+q+1) vector of unknowns (a_1, ..a_p, b_1,...b_q, σ_u^2).

We define the complete data set as consisting of the sequence $U_1^N = \{u(t), t=1, N\}$ since given U_1^N, the observed data set Y_1^N can be constructed exactly from Eq. (8.28). The sequence U_1^N is the innovation sequence and is related to Y_1^N through a causal and causally invertible transfer function. It is assumed that N is large so that the effect of initial conditons (i.e. values of y(t) and u(t) for t<o) is negligible.

$$P(U_1^N|\phi) = \prod_{t=1}^{N} p(u(t)|\phi) \qquad (8.29)$$

where $p(u(t)|\phi) \sim N(o,\sigma_u^2)$.

u(t) can be obtained from Eq. (8.28) as

$$u(t) = y(t) - \sum_{i=1}^{p} \alpha_i y(t-i) - \sum_{i=1}^{q} b_i\, u(t-i) \qquad (8.40)$$

Using this equation,

$$P(U_1^N|\phi) = \frac{1}{(2\pi\sigma_u)^N} \exp - \frac{1}{2\sigma_u^2} \sum_{t=1}^{N} [y(t) - \sum_{i=1}^{p} \alpha_i y(t-i)$$

$$- \sum_{i=1}^{q} b_i u((t-i)]^2$$

The sufficient statistics $t(U_1^N)$ for estimation of parameters $\{\alpha_1 \ldots \alpha_p, b_1 \ldots b_q, \sigma_u^2\}$ are $\sum_{t=1}^{N} y(t-i)\, u(t-j)$, $\sum_{t=1}^{N} y^2(t)$, $\sum_{t=1}^{N} y(t-i)\, y(t-j)$, $\sum_{t=1}^{N} y(t-i)\, u(t-j)$ and $\sum_{t=1}^{N} u^2(t)$. In order to obtain conditional mean of $t(U_1^N)$ given Y_1^N, u(t) is expressed in terms of y(t) sequence by inverting Eq. (8.28). The inversion is done most easily by using lag operator, z.

$$u(t) = \frac{A(z)}{B(z)}\, y(t) = C(z)\, y(t) \qquad (8.31)$$

where $A(z) = 1 - \sum_{i=1}^{p} \alpha_i z^i$

$$B(z) = 1 - \sum_{i=1}^{q} b_i z^i$$

$$C(z) = 1 - \sum_{i=1}^{\ell} C_i z^i$$

C(z) represents the equivalent AR model of a high order, ℓ.

The above relationships suggest the following E–M algorithm:

E–Step: Given Y_1^N and $\phi^{(k)}$, estimate sequence $\{u(t)\}$ from Eq. (8.31), which corresponds to a long autoregression. The order of the AR model can be determined using AIC.

M–Step: Update ϕ parameters $\{a_1 \ldots a_p, b_1 \ldots b_q, \sigma_u^2\}$ using regression of $Y(t)$ on lagged values of $y(t)$ and $u(t)$. The parameter estimation involves use of the same sufficient statistics as identified above.

An alternative procedure to estimate parameters which is similar to the CVA method is to first obtain predictors $\hat{y}(t+1|t), \hat{y}(t+2|t), \ldots \hat{y}(t+n|t)$ using orthogonal projections or conditional expectations, where $n = \max(p,q)$. Then from Eq. (8.38),

$$\hat{y}(t+n|t) = \sum_{i=1}^{p} \hat{\alpha}_i \hat{y}(t+n-i|t) \qquad (8.32)$$

Equation (8.32) is used to estimate $\{\hat{\alpha}_i\}_{1,p}$.

Then $\{b_i\}$ are obtained from

$$B(z) = A(z)/C(z) \qquad (8.33)$$

σ_u^2 is estimated from the variance of the residuals from the long autoregression that estimates coefficients of $C(z)$.

8.4 Relationship between CVA – Regression and E–M Algorithm:

The procedure used in section 8.3 for ARMA models can be generalized to state

space models which are equivalent to multiple ARMA models. The role of sequence u(t) is played by the innovation sequence $\nu(t)$ in the state space modeling framework. We use the Kalman Filter representation of the state model,

$$\hat{x}(k+1|k) = F[\hat{x}(k|k-1) + K\,\nu(k)] \tag{8.34}$$

$$y(k) = H\hat{x}(j|k-1) + \nu(k) \tag{8.35}$$

It is well–known that the innovation sequence $\{\nu(k)\}$ and the output sequence $\{y(k)\}$ are related through a causal and causally invertible transfer function. Therefore, given $\{\nu(k)\}$, $\{y(k)\}$ can be obtained from Eq. (8.34) and (8.35). To obtain $\nu(k)$ from y(k), we rewrite Eqs. (8.34) and (8.35) as

$$\begin{aligned}\hat{x}(k+1|k) &= F\hat{x}(k|k-1) + FK(y(k)-H\hat{x}(k|k-1)) \\ &= F(I-KH)\,\hat{x}(k|k-1) + FKy(k)\end{aligned} \tag{8.36}$$

$$\nu(k) = y(k) - H\hat{x}(k|k-1) \tag{8.37}$$

Defining the complete data set as $S_1^N = \{\nu(k)\}_{1,N}$, the E–M algorithm can be implemented in the same way as for the ARMA case. We are assuming here that N is large and the system is stable so that the effect of initial state x(o) is negligible.

E–Step: Given Y_1^N and parameter values in {F,H,K}, use Eq. (8.36) and (8.37) to estimate the sequence ν_1^N. At the same time, the sequence $\{\hat{x}(k|k-1)\}$ is estimated and the sufficient statistics are computed.

M–Step: Update parameters in {F.H.K} by equating sufficient statistics from the E–step to their expected values, which are only a function of ϕ. Alternatively, perform regressions using Eq. (8.34) and (8.35) and treating $\nu(k)$ as white noise.

Notice that the regression step of the CVA algorithm of the previous chapters is similar to the above M–step. The major difference lies in the E–step, where the CVA algorithm estimates $\hat{x}(k|k-1)$ nonparametically assuming no a priori model structure. In practice, this is a necessary first step since it provides model structure and initial parameter estimates.

The alternative procedure described in section (8.3) for implementing M–step has been generalized by Akaike (1976) and Mehra (1982) to the current situation. The covariance of $\nu(k)$ is obtained by fitting a long auto–regressive model. The F–parameters are obtained from the CVA calculation after model order has been determined. K matrix is obtained by equating the transfer functions from the AR model and the state space eqn. model (8.34) and 8.35).

The above relationship between CVA and E–M algorithm shows that CVA only implements one step of the E–M algorithm. By repeating these steps, as outlined above, one can increase the likelihood function and achieve convergence to MLE. Furthermore, the recognition of the above relationships shows us how to make CVA recursive.

It should be remarked that the results of the sections 8.2, 8.3 and 8.4 are extended easily to the case of exogenous inputs or forcing functions which are known.

8.5 Recursive ML Identification:

8.5.1 ARMA Case:

In this section, we show how the E and M steps can be implemented recursively for adaptive time series analysis. For simplicity, consider first the ARMA model of Eq. (8.28). The recursive estimation of AR models is well–known (Ljung and Soderstrom (1983)). The computations of $\{u(t)\}$ from $(y(t)\}$ can be made recursive both in model order and data length for the univariate as well as the multivariate cases. In order to perform recursively M–step involving regression, we create a new state vector $\phi(t)$ consisting of $\{a_1 \ldots a_p$, $b_1, \ldots b_q\}$. For constant coefficient ARMA models,

$$\phi(t+1) = \phi(t) \qquad (8.38)$$

$$y(t) = H(t)\,\phi(t) + u(t) \qquad (8.39)$$

where $H(t) = [y(t-1),\ \ y(t- \ ),\ \ u(t-1),\ \ u(t-q)]$

On–line estimation of $\phi(t)$ can be performed recursively using a Kalman Filter. The variance parameter σ_u^2 can also be updated recursively. We can generalize to the time varying parameter case in which

$$\phi(t+1) = A\,\phi\,(t) + w(t) \qquad (8.40)$$

If A and Cov(w) are known, a Kalman Filter still provides the best estimates of $\phi(t)$, along with the covariance of the estimates. For A and Cov(w) unknown, E–M algorithm needs generalization. This can be done by defining A and Cov(w) as the hyper–parameters and redefining the complete data vector to consist of $\{u(\ ), \phi(t)\}$ sequence.

The recursive MLE computations involve the following steps, when a new data point $y(N+1)$ is received.

1. Update AR model parameters (coefficients of C(z)) by using $y(N+1)$.

2. Recompute $\{u(t)\}$, t=1, N+1 using the new AR model.

3. Solve the KF equations for the parameter state vector $\phi(t)$, using the new values of $\{u(t)\}$ and $\{y(t)\}$, t=1, N+1.

4. Solve for AR model parameters from the estimated ϕ parameters and repeat the above steps.

In the above procedure, steps 2 and 3 are repeated over the whole data set. In cases where the parameter changes are small, it may be sufficient to simply compute u(N+1) and run the KF only for one step, without any further iterations. This procedure is recommended in any case to check if the changes in the parameters are large enough to warrant recomputation of the whole $\{u(t)\}$ sequence and iteration.

For the time varying case, it is also possible to implement a fixed data window KF (see Mahmood (1989)). The optimal window length can be determined using a generalized AIC criterion.

8.5.2 State Space Models:

The procedure for recursive estimation in state space models is similar to the one used above for ARMA models. For the case in which the model structure is unknown, CVA is

used first to determine model order and a system state vector. Then matrices F,H,K and $\Sigma_{\nu\nu}$ are determined using methods proposed by Akaike (1974) and Mehra (1982). In this approach F and H are chosen in a canonical form with a parsimonious representation. This initial step is nonrecursive and requires a batch of data $\{y(1),..y(\ )\}$. After this step, the algorithm can be made recursive. It is desirable to supplement CVA procedure with an E–M iteration on the initial batch of data to get MLE. Then the purpose of the recursive algorithm will be to update parameters with a new data point (N+1) without repeating all the previous computations.

The recursion consists of the following steps:

1. Solve Eq. (8.36) and (8.37) using parameters from CVA and store $\{\hat{x}(k|k-1),\nu(k)\}_{k=1,N}$

2. Rewrite Eq. (8.34) and (8.35) as regression equations with ϕ as dependent variable $\{\hat{x}(k+1|k), y(k)\}$ and independent variables $\hat{x}(k|k-1)$. Define a state equation for ϕ as (8.48) and use a Kalman Filter to estimate ϕ using the regression equations as the measurement equations for ϕ. We omit the details since the procedure is similar to the one used for the ARMA model in writing Eq. (8.39) and the Kalman Filter equations are well–known.

3. Repeat the above steps till convergence is achieved.

Depending on the change in the parameter estimates for applying the steps 1 and 2 to (N+1), a decision can be made to implement steps 1 and 2 on the entire data set or a suitable window. Similarly step 3 should be used if the changes in parameters is large.

The major benefit of using the above procedure ~~our~~ repeated use of CVA is that the time consuming CVA and complete regression calculations do not have to be repeated at each step.

8.6 Other Extensions:

8.6.1 Missing Data:

The E–M algorithm is ideally suited for handling missing data points. The complete data set is simply expanded by the missing y(.) data points. If an initial model is available, and the E–step is carried out using a KF, the missing data points presents no problem as such since the KF can handle data points taken at arbitrary times. In fact, the KF produces state estimates at every step of the state transition equation. The missing outputs are then estimated from the corresponding state estimates and used in computing the sufficient statistics. The M–step is based on the use of the estimated complete–data sufficient statistics in a regression or other type of parameter estimation process. As the E–M steps are repeated, the missing data points are replaced by their best estimates based on the observed data and the model parameters. The same procedure can also be used to handle outliers or bad data points.

8.6.2 Nongaussin Statistics:

E–M algorithm, like MLE, is applicable to the nongaussin case, as long as the form of the distribution function or the likelihood function are known except for the unknown parameters. However, for the nonexponential family of distributions, the M–step involves maximization of a function. The details of the M–step and its properties are well covered in Dempster etal. (1977).

CHAPTER 9

SUMMARY, CONCLUSION AND FUTURE RECOMMENDATIONS

9.1 INTRODUCTION:

The main goal of this project was to investigate theoretical issues about Adaptive Time Series Analysis including a system identification technique known as CVA-AIC. A predictive inference and entropy framework was selected for the analysis. The CVA-AIC technique was also be compared with other techniques of system identification.

This technique, developed over the last decade, has been found successful in many practical applications. Such success can be attributed to the fact that CVA-AIC has better statistical properties than many other existing techniques and allows automatic optimal model order selection using concepts of entropy and predictive inference. The latter capability relieves a user from the burden of applying subjective decisions regarding model order selection. The CVA-AIC technique is based upon the stochastic realization theory, statistical theory of canonical correlation analysis and the order selection procedure based upon Akaike's Information Criterion (AIC). Although these underlying theories are based upon rigorous

mathematical justification, various specific steps of the CVA-AIC technique have not been founded upon strict mathematical rigor. Instead, the technique has been used on an adhoc basis in practical applications. For this reason, the effort of this project was devoted to the strengthening of theoretical aspects of the CVA-AIC technique and exploring its relationship to other techniques, instead of producing empirical results using extensive simulation runs. Before presenting the conclusions of this effort, we provide a summary of the report.

9.2 SUMMARY:

The report start with an overview of the adaptive time series problems. The necessary mathematical preliminaries have been presented in Chapter 2. It has been shown at the beginning of this chapter (Section 2.3), that if the dimension of the parameter vector is known, then the AIC criterion is asymptotically equivalent to the maximum likelihood estimation problem. In the second half of this chapter, the estimation technique for model entropy has been presented. These results are not new, but have been presented here for the sake of completeness of the report. Moreover, the analysis has been carried out in a discrete valued random variable framework which is new and more transparent. The CVA theory, in its most general form, has been presented in Chapter 3. Here the case of modelling

stationary processes with possibly singular covariance matrices has been considered. It has been shown that this type of problem can be solved using a generalized singular value decomposition process leading to a generalized canonical variate theory. The analysis in this chapter includes the standard CVA theory when the respective covariance matrices are of full rank. Several schemes for generating state space models have been presented in Chapter 4. In Section 4.1, the technique for computing the Kalman Filter form of the state space model has been developed where it has also been shown how to transform this form to the standard state space model. The technique for computing the latter model has been presented in Section 4.2 and finally, a technique that is recursive in model order has been developed for the standard state model in Section 4.3. The motivation behind this effort is that, on many occasions, models of increasing order are computed until a desired characteristic (such as a desired mode) is detected in the model. The problem of confidence interval estimation around the power spectrum and systems transfer function is dealt with in Chapter 5. The motivation for undertaking this analysis is that both the power spectra as well as confidence bands are needed in many design problems. For example, in control system design, a control law is designed for a nominal plant to obtain a prespecified performance level and stability margin that will also domain valid for the entire set of plants in the neighborhood of the nominal

plant in the frequency remain. Therefore, the uncertainty region around the identified model must lie within the region allowed by the control law. It is interesting to note that the size of the confidence interval around the identified model can be adjusted from the data length.

The abrupt change in a model is encountered frequently in a real world situation. For example, sensor failure or a component failure in a system may induce an abrupt change in the system model. This change must be detected quickly and corrective action must be taken to prevent any catastrophic failure. It has been demonstrated in Chapter 6 that CVA-AIC technique can be used for such fault detection by comparing the value of AIC on each successive data interval. The technique developed in this chapter has been demonstrated on a simulation example. As shown in Chapter 7, an entirely different approach is taken for a slowly varying system. In this case, the data is divided into various subintervals and a separate model that is optimal in the sense of AIC criterion is identified for each segment. In the last section of this chapter, the technique has been illustrated through a simulation example.

Chapter 8 is devoted to the application of a technique known as the E-M algorithm for extension of the previous results. Although the technique is relatively new to the engineering community, the researchers in the field of

statistics have used this technique from the mid seventies. This is a very powerful and general framework in which many classes of estimation problems can be formulated and solved. In addition theoretical issues such as convergence rates can be analyzed. It has been shown in this chapter how the CVA-AIC technique fits naturally into the E-M framework and how the maximum likelihood estimates (MLE) of the parameters can be obtained starting from the CVA-AIC estimatse. In addition, extension to time varying parameters and recursive parameter estimation schems are described. The E-M algorithm presents a new way of analyzing and extending the CVA technique. At the same time, it unifies different techniques of adaptive time series analysis and shows clearly the relationship between them.

9.3 CONCLUSIONS:

The investigations under the scope of this project have enhanced our understanding of how to analyze a time varying system in a predictive inference and entropy framework. Our attention was focused mainly on the CVA-AIC technique and its relationship to other techniques for system identification. The major conclusions of this project are:

(i) The problem posed by Akaike (for computing AIC) is asymptotically equivalent to maximum likelihood estimation problem when the parameter dimension is known.

(ii) The CVA-AIC technique can be extended using the E-M Algorithm in such a way that the model will converge monotonically towards the ML estimates.

(iii) The problem of entropy maximization can be posed either for the continuous time or discrete time stochastic processes.

(iv) The CVA theory has been extended to include more general type of time series. The extended theory is known as "Generalized CVA Theory" and can handle time series with singular covariance matrices.

(v) In the CVA-AIC framework, the model can be identified either in the standard state space form or in the Kalman Filter form, depending upon the type of application at hand.

(vi) The standard state-space model of various orders can be computed recursively starting from order one.

(vii) Once a state-space model is identified, the system transfer function, noise power spectrum and

associated bands of various confidence levels can be computed.

(viii) An abrupt change in a model can be detected in a CVA-AIC framework by partitioning the data into various segments and comparing the value of the optimal AIC from various segments. This technique can be used as a generalized fault-detection scheme.

(ix) E-M algorithmic approach provides a very general framework where most of the estimation problem can be formulated. If the CVA-AIC technique is embedded properly in an E-M framework, it would leads to maximum likelihood estimates.

9.4 FUTURE DIRECTION:

Considerable theoretical analysis has been done in this project, yet more is needed to understand the CVA-AIC and the E-M techniques fully. Also future efforts should be directed towards practical implementation issues. It is recommended that the future work in this area should include, but not necessarily be limited to, the following items:

(a) It has been reported that the CVA-AIC technique produces estimates that are close to MLE estimates. In

fact, it can be shown for independent and identically distributed random variables, that the estimate generated by the AIC criterion is asymptotically equal to the MLE estimate. More theoretical investigation is needed to assess the performance of the generalized CVA-AIC technique relative to the MLE technique in the general Adaptive Time Series setting.

(b) Currently, CVA-AIC technique has been implemented in a batch mode and is computationally demanding. The technique is suitable for a slowly time varying system where the model may need periodic updating. For systems with fast parameter changes, a recursive form of CVA-AIC technique needs to be developed so that the it can be used in real time. We have developed approaches in this direction using the E-M Algorithmic approach.

(c) It has been theoretically demonstrated here how to compute the bounds at various confidence levels on the systems transfer function and the noise power spectral density. These algorithms need to be implemented and verified via Monte Carlo simulations.

(d) The state space matrices obtained from the existing CVA-AIC algorithm are not in any particular canonical form and may be over-parametrized. The effect of this over-parametrization on the efficiency of the estimates

finding the model - it should be investigated whether this can be replaced by other algorithms with lesser computational burden.

(i) The combination of CVA and E-M Algorithm based approaches derived in this report for time varying and constant systems should be tested on practical examples of increasing complexity. These techniques are extremely promising for solving the general Multivariate Adaptive Times Series Identification problem

REFERENCES

Aitchison J. and I.R. Dunsmore (1975). Statistical Prediction Analysis, Cambridge University Press.

Akaike, H. (1976). "Canonical Correlation Analysis of Times Series and the Use of an Information Criterion." Systems Identification: Advances and Case Stidies, R.K. Mehra and D.G. Lainiotis, eds., New York: Academic Press, pp. 27-96

Akaike, H. (1975). "Markovian Representation of Stochastic Precesses by Canonical Variables.: SIAM J. Contr., Vol 13, pp. 1620-173

Akaike, H. (1974a). "Stochastic Theory of Minimal Realization.: IEEE Trans. Automat. Contr., Vol. 19, pp. 667-674.

Akaike, H. (1974b). "A New Look at Statistical Model Identification." IEEE Automatic Control, Vol 19, pp. 667-674.

Akaike, H., (1973). "Information Theory and and Extension of the Maximum Likelihood Principle.: In 2nd International Symposium on Information Theory., Eds. B.N. Petrov and F. Csaki, pp. 267-281. Budapest: Adakemiai Kiado.

Anderson, T.W. (1971). The Statistical Analysis of Time Series. New York: Wiley.

Ansley, C.F., & Kohn, R. (1983). Exact Likelihood of Vector Autorefressiv - Moving Average Process with Moving or Aggregated Data. Biometrika 70, 275-8.

Astrom, K.J. (1973). "On Self-tuning Regulators." Automatica, Vol 9, pp. 185.

Astrom, K.J. (1983). :Theory and Applications of Adaptive Control-A Survey." Automaitca, Vol 19, pp. 471-86

Astrom, K.J., Borisson, L. Ljung and B. Wittenmark, (1977). "Theory and Applications of Self-tuning Regulators." Automatica, Vol 9, pp. 185.

Bhansali, R.J. and Downham, D.Y. (1977), "Some Properties of the Order of an Autoregressive Model Selected by a Beneralization of Akaike's FPE Criterion.: Biometrika, Vol 64, ppl 547-71

Box, G.E.P. and G.M. Jenkins (1970), Times Series Analysis Forecasting and Control, San Francisco: Holden-Da.

Gevers, M. and V. Wertz (1982). "On the Problems of Structure Selection for the Identification of Stationary Stochastic Process,: Papers of the IFAC Symp. on Identification and System Parameter Estimation, G. Bekey and G. Sardis (eds), Was. D.C.: McGregor-Wener, pp. 387 -92.

Godolphin, E.J. & Unwin, J.M (1983). Evaluation of the Convariance Matrix for the Maximum Likelihood Estimator of a Gaussian Autoregressive-Moving Average Process. Biometrika 70, 279-84.

Goldstein, J.D., and W.E. Larimore, (1980). "Application of Kalman Filtering and Maximum Likelihood Parameter Identification to Hydrologic Forecasting." The Analytic Sciences Corporation, Report No. TR-1480-1, March 1980. Available as Report N. AD-A113347 through Defense Technical Information Center, Alexadria VA 23314.

Golub, G.H. (1969). Matrix Decompostitions and Statistical Calculations. Statistical Computation, R.C. Milton and J.A. Nelder, eds., New York: Academic Press, pp. 365 -379.

Goodrich, R., W.E. Larimore and R.K. Mehra (1983). "New Results in State Space Forcesting.: Intenational Syposium on Forecasting, Philadelphia, PA, June 5-8.

Granger, C.W. J. and G. McCollister (1979). "Comparison of Forecasts of Selected Series by Adaptive.: Box-Jenkins and State Space Mehtods, ORSA/TIMS, Los Angeles, California.

Habermann, S.J. (1984). Adjustment by Minimum Dicreiminant Information Estimation. Biometrics, 24, 707-713.

Hagglund, T. (1983). "New Estimatin Techniques for Adaptive Control." Report CODEN:LUTFD2/(TFRT-1025)/1-120/(1983), Department of Automatic Control, Lund Instutite of Technology. Doctoral Dissertation.

Hart, P.E. (1971). "Entrophy and Other Measures of Concentration," J. Roy Statist. Soc., A, Bol. 134, pp. 73-85.

Honig, M.L. and D.G. Messerschmitt (1984). Adaptive Filters: Structures, Algorithms, and Applications. Boston: Kluwer Academic Publishers.

Hotelling, H. (1936). "Relations between Two Sets of Variates." Biomrytika Vol. 28, pp. 321-377.

Ireland, C.T., and Kullback, S. (1968). Minimum Discriminate Information Estimation. Biometrics, 24, 707-713.

Irving, E. (1979). "New Development in Improving Power Network Stability with Adaptive Control." Proc. Workshop on Applications of Adaptive Control. Yale University, New Haven.

Isermann, R. (1984). "Process Fault Detection Based on Modeling and Estimation Methods - A Survey," Automatica, Vol. 20, pp. 387-404.

Jefferys, H. (1961). Theory of Probabilty, Clarendon Press.

Jones, R.H. (1980). Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations. Technometrics 22, 389-95.

Kendall, M.G. (1973). "Entropy, Probability and Information," International Statistical Review, Vol. 41, pp. 59-68.

Kiefer, J. and J. Wolfowritz (1956). Consistency of the Maximum Likelihood Estimator in the Presenence of Infinitely Many Incidental Parameters. Ann. Math. Statis. 42, 594-606.

Kullback, S (1959). Information Theory and Statistics, Dover.

Kullback, S. and R.A. Leibler (1951). "On Ingormation and Sufficency,: An. Math. Statistics, 22, pp. 79-86.

Kung, S.Y. and S.W. Llin (1981). "Optimal Hankel-Norm Model Reductins: Multivariable Systems", Trans. Auto. Control, Vol. 26, pp. 832-852.

Larimore, W.E. and R.K. Mehra (1984). :Technical Assesment of Adaptive Flutter Suppresion Research." Air Force Wright Aeronautical Lab Report No.-AFWAS-TR-84-3052, SSI.

Larimore, W.E., S. Mahmood, and R.K. Mehra (1983). "Adaptive Model Algorithms Control.: Proc. IFAC Workshop on Adaptive Systems in Control and Signal Processing, San Francisco, CA June 1983.

Larimore, W.E. (1983a). "Predictive Inference, Sufficiency, Entropy, and an Asympotic Likelihood Principal." Biometrika, Vol. 70, pp. 175-81.

Larimore, W.E. (1983b) "Systems Identification, Reduced -Order Filtering and Modeling Via Canonical Variate Analysis.: Proc. 1983 American Control Conference, H.S. Rao and T. Dorato, eds., New York: IEEE. pp. 445-51

Larimore, W.E. (1981a). "Small Sample Methods for Maximum Likelihood Identification of Dynamical Processes." Applied Time Series Analysis, Proceeding of the Fifth International Time Series Meeting. Houston Texas, August 1981. Amsterdam, North Holland, pp. 167-174.

Larimore, W.E. (1981b). "Recursive Maximum Likelihood and Related Algorithms for Parameter Identification of Dynamical Processes." Proceedings fo the 20th IEEE Congerence on Decision and Control, Vol. 1, pp. 50-55, San Diego, California, December 1981.

Larimore, W.E. (1977). "Nontested Tests on model Structures.: Proceedings Joint Automatic Control Conf. (San Francisco, CA). New York: IEEE, pp. 686-690.

Larimore, W.E. (1977b). "Statistical Inference on Stationary Random Fields." Proc. IEEE. Vol. 65, pp. 961 -70.

Loy, X.C., A.S. Willsky, and G.C. Verghese (1983). "Failure with Uncerain Models." Proc. Amer. Control Conf. San Francisco, California.

Ljung, J. and T. Soderstrom (1983). "Theory and Practice of Recursive Identification." Cambridge: MIT Press.

Ljung, L. (1979). "Asymptotic Behavior of the Extended Kalman Filter as a parameter Estimator for Linear Systems." IEEE Trans. Auto. Control, Vol. 24 pp. 36 -50.

Ljung, L., I. Gustavsson and T. Soderstrom (1974). "Identification of Linerar Multivariable Systems Operating Under Linear Feedback Control." IEEE Trans. Auto control, AC-19, pp. 836-840.

Mahmood, S. (1989). "Kalman Filtering for a Moving Window Data System." Interim Progress Report to NATC, Patuxent River, MD, Contract No. N00421-89-C-0142.

Mehra, R.K. (1982). "Identification In Control and Econometrics." Current Development in the Interface: Economics, Econometrics, Mathematics. M. Hasewinkel and A.H.G. Rinnooy Kan, Eds., Dordrecht, Holland: Reidel, pp. 261-288.

Mehra, R.K. (1981). Choice of Input Signals. Trends and Pregress in Sustem Identification, P. Eykhoff, ed., IFAC, Oxford: Pergamon Press.

Mehra, R.K. (1978). A Survey of Time-Series Modeling and Forecasting Methodology. Modeling, Identification and Control in Environmental Systems, E. Vansteenkist, ed., North Holland Publishing Co.

Mehra, R.K. and A. Cameron (1980). "Handbook on Business and Exonomic Forecasting for Single and Multiple Time Series." Scientific Systems, Inc., Notes for the Institute of Professional Education Seminar.

Mehra, R.K. and A. Cameron (1976). "A Multidimensional Identification and Forecasting Technique Using State Space Models." ORSA/TIMS Conf. Miami, FL, November 1976.

Mehra, R.K. and J.S. Tyler (1973). "Case Studies in Aircraft Parameter Identification.: Proc. 3rd IFAC Conf. on Identification and System Parameter Estimation, P. Eykoff, ed., Oxford: Perfamon Press, 117-144.

Mehra, R.K. and J. Peschon (1971). "An Innovations Approcach to Fault Detection and Diagnosis in Dynamic Systems." Automatica, Vol. 7, pp. 657.

Mehra, R.K. (1970). "On the Identification of Variances and Adaptive Kalman Filtering." IEEE Trans. on Auto. Cont., Vol AC-15 pp. 175-184.

Mehra R.K. (1968). "On Optimal and Suboptimal Linear Smoothing." Proc. National Electronic conf., Vol. XXIV, pp. 119-124

Murray, G.D. (1977). "A Note on the Estimation of Probability Density Functions." Biometrika, Vol. 64, pp. 150-2.

Murray, G.D. (1979). " The Estimation of Multivariate Normal Density Functions Using Incomplete Data." Biometrika, Vol. 66, pp. 375-80.

Peloubet, R.P., Jr., R.L. Haller and R.M Bolding, (1980). "F-16 Flutter Suppression System Investigation," presented at the AIAA/ASME/ASCE/AHS 21st Structures, Structural Dynamics amd Materials Conference, Seattle, Washington, May 1980.

APPENDIX

MAXIMUM LIKELIHOOD ESTIMATION

We present here some basic background on maximum likelihood estimation, which is used throughout this report.

The likelihood function for a sample $x_1, x_2, \ldots x_n$ parametrized by a parameter Θ is

$$L = \prod_{i=1}^{n} p(x_i \mid \Theta) \tag{A.1}$$

Assume the x_i are drawn independently from the true distribution $p(x \mid \Theta_0)$. Then L is the joint distribution function of $x_1, x_2, \ldots, x_n$ and

$$\int \ldots \int L \, dx, \ldots dx_n = 1 \tag{A.2}$$

Differentiating wrt Θ:

$$\int \ldots \int \left(\frac{\partial L}{\partial \Theta}\right) dx, \ldots dx_n = 0 \; ; \; \frac{\partial L}{\partial \Theta} = \text{row vector}$$

or

$$\int \ldots \int \left(\frac{\partial \log L}{\partial \Theta}\right) L \, dx, \ldots dx_n = 0$$

or

$$E\left(\frac{\partial \log L}{\partial \Theta}\right) = 0 \tag{A.3}$$

Differentiating again

$$\int \cdots \int \left(\frac{\partial^2 L}{\partial \Theta^2} \right) dx_1 \cdots dx_n = 0$$

Now

$$\frac{\partial^2 \log L}{\partial \Theta^2} = - \frac{1}{L^2} \left(\frac{\partial L}{\partial \Theta} \right)^T \left(\frac{\partial L}{\partial \Theta} \right) + \frac{1}{L} \frac{\partial^2 L}{\partial \Theta^2}$$

so that

$$\frac{\partial^2 \log L}{\partial \Theta^2} = - \left(\frac{\partial \log L}{\partial \Theta} \right)^T \left(\frac{\partial \log L}{\partial \Theta} \right) + \frac{1}{L} \frac{\partial^2 L}{\partial \Theta^2}$$

Thus

$$E \left[\frac{\partial^2 \log L}{\partial \Theta^2} \right] = - E \left[\left(\frac{\partial \log L}{\partial \Theta} \right)^T \left(\frac{\partial \log L}{\partial \Theta} \right) \right]$$

Now

$$\left. \frac{\partial \log L}{\partial \Theta} \right|_{\Theta_0} = \left. \frac{\partial \log L}{\partial \Theta} \right|_{\hat{\Theta}} + (\Theta_0 - \hat{\Theta})^T \left. \frac{\partial^2 \log L}{\partial \Theta^2} \right|_{\hat{\Theta}}$$

where $\hat{\Theta}$ is the maximum likelihood estimate which satisfies

$$\left. \frac{\partial \log L}{\partial \Theta} \right|_{\hat{\Theta}} = 0$$

and Θ_0 is the true parameter value.

Thus

$$\left.\frac{\partial \log L}{\partial \theta}\right|_{\theta_0} = (\theta_0 - \hat{\theta})^T \left.\frac{\partial^2 \log L}{\partial \theta^2}\right|_{\hat{\theta}} \tag{A.4}$$

Now define the covariance matrix

$$C = E\left[\left(\frac{\partial \log L}{\partial \Theta}\right)\Big|_{\hat{\Theta}}^{T} \left(\frac{\partial \log L}{\partial \Theta}\right)\Big|_{\hat{\Theta}}\right] = -E\left[\left.\frac{\partial^2 \log L}{\partial \Theta^2}\right|_{\hat{\Theta}}\right] \tag{A.5}$$

and factor C as $C = W W^T$.

Write (A.4) as

$$\left.\frac{\partial \log L}{\partial \Theta}\right|_{\Theta_0} W^{-T} = (\Theta_0 - \hat{\Theta})^T \left.\frac{\partial^2 \log L}{\partial \Theta^2}\right|_{\hat{\Theta}} W^{-T}$$

The right hand side is approximated as

$$(\Theta_0 - \hat{\Theta})^T C W^{-T} = (\Theta_0 - \hat{\Theta})^T W$$

The left hand side is a normalized gaussian variate since

$$E\left\{\left[\left(\frac{\partial \log L}{\partial \Theta}\right)\Big|_{\hat{\Theta}} W^{-T}\right]^T \left[\left.\frac{\partial \log L}{\partial \Theta}\right|_{\hat{\Theta}} W^{-T}\right]\right\} = I$$

Thus, the right hand side is also a normalized gaussian variate and

$$E\left\{[(\Theta_0 - \hat{\Theta})^T W]^T [(\Theta_0 - \hat{\Theta})^T W]\right\} = I$$

which yields

$$E\left[(\theta_j - \hat{\theta})(\theta_j - \hat{\theta})^T\right] = C^{-1} \qquad (A.5)$$

C is the Fisher information matrix, which is the inverse of the covariance matrix of the parameter estimation errors.

APPENDIX B: An Innovations Approach to Maximum Likelihood Identification of Linear and NonLinear Dynamic Systems

This appendix presents an approach to maximum likelihood identification of multi–input multi–output linear and nonlinear dynamic systems with arbitrary inputs. The approach is based on state vector formulation and uses the innovation properties of optimal filters for these systems. Application to the identification of the transfer function of a chemical reactor is considered.

1. Introduction

The maximum likelihood estimation of autoregressive and moving average parameters in time series analysis has been considered by several investigators [1,2].* The related problem of linear system identification can often be cast in this framework, though the parameter transformations invloved may be nonlinear and nonunique. Special difficulties are encountered in handling multi–input multi–output linear models and nonlinear models using the time–series approach. The author [3,4] has tried to circumvent these difficulties by working directly with the physical models and using the innovations approach of Kailath [5,6]. A schematic diagram of this method is shown in Figure B.1.

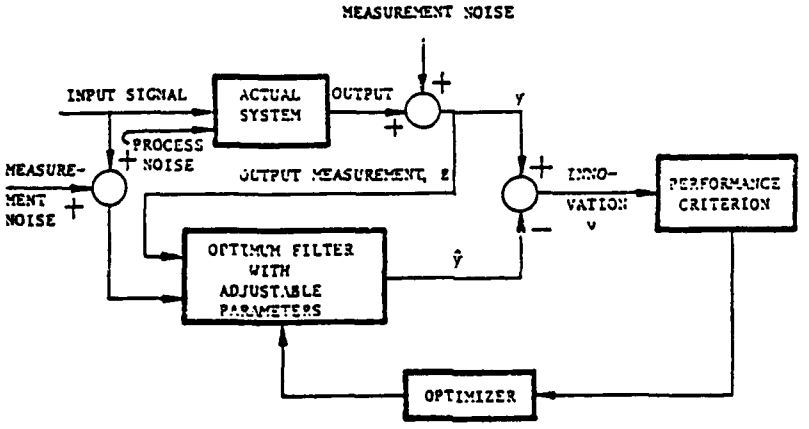

Figure B.1 Implementation of maximum likelihood estimator

*References for Appendix B are given separately at the end.

2. Linear Systems

Consider a discrete–time linear system*

$$x(t+1) = Fx(t) + Gu(t) + \Gamma w(t) \tag{B.1}$$

$$y(t) = Hx(t) + v(t) \tag{B.2}$$

where

$x(t)$ = n x 1 state vector; $u(t)$ = p x 1 input vector;

$w(t)$ = q x 1 vector of random forcing functions;

$y(t)$ = r x 1 output vectors; and $v(t)$ = r x 1 vector of output errors

and

$$E\{w(t)\} = o, \quad E\{w(t)w^T(\tau)\} = Q\delta_{t,\tau}$$

where δ, is the Kronecker delta function.

$$E\{w(t)v^T(t)\} = 0$$

$$E\{v(t)\} = 0, \quad E\{v(t)v^T(\tau)\} = R\delta_{t,\tau}$$

*References for Appendix B are given separately at the end.

It is assumed that the structure of the model is known. The vector of unknown parameters from F,G,Γ,H,Q and R isdenoted by θ. It is assumed tht θ is identifiable.

The ML estimate of θ is given by

(B.3) $$\hat{\theta} = \text{Arg } \{\max_{\theta} \log p(Y_N/\theta)\}$$

where

$$Y_N = \{y(1),.....,y(N)\}$$

and

$$p(Y_N/\theta) = \text{conditional probaability density of } Y_N \text{ given } \theta.$$

An expression for $p(Y_N/\theta)$ is derived as

$$p(Y_N/\theta) = p(y(1),.....,y(N)/\theta)$$

$$= p(y(N)|Y_{N-1},\theta)\ p(Y_{N-1}|\theta)$$

$$= p(y(N)|Y_{N-1},\theta)p(y(N-1)|Y_{N-2},\theta)p(Y_{N-2}|\theta)$$

$$= \sum_{j=1}^{N} p(y(j)|Y_{j-1},\theta).$$

Therefore

$$\log p(Y_N|\theta) = \sum_{j=1}^{N} \log p(y(j)|Y_{j-1},\theta) \tag{B.4}$$

Consider the case in which x(0), w(t) and v(t) are normally distributed. Then $p(y(j)|Y_{j-1},\theta)$ by a well–known property of normal distributions is also normal.

Let

$$E\{y(j)|Y_{j-1},\theta\} = \hat{y}(j|j-1) \tag{B.5}$$

and

$$\text{Cov}\{y(j)|Y_{j-1},\theta\} = B(j|j-1). \tag{B.6}$$

It is known that $\hat{y}(j|j-1)$ and $B(j|j-1)$ can be obtained from a Kalman filter [7] of the following form:

$$\hat{x}(t+1/t) = F\hat{x}(t/t) + Gu(t) \tag{B.7}$$

$$\hat{x}(t/t) = \hat{x}(t/t-1) + K(t)\nu(t) \tag{B.8}$$

$$\nu(t) = y(t) - H\hat{x}(t/t-1) \tag{B.9}$$

$$K(t) = P(t/t-1)H^T B^{-1}(t/t-1) \tag{B.10}$$

(B.11) $$B(t) = HP(t/t-1)H^T + R$$

(B.12) $$P(t/t = (I - K(t)H)P(t/t-1)$$

(B.13) $$P(t+1/t) = FP(t/t)F^T + \Gamma Q \Gamma^T.$$

The likelihood function (B.4) can now abe written as

(B.14) $$\log p(Y_N|\theta) = -\frac{1}{2} \sum_{j=1}^{N} [\nu^T(j)B^{-1}(j/j-1)\nu(j) + \log|B(j/j-1)|].$$

Here $\nu(t)$ denotes the innovation sequence which is zero mean, Gaussian and white [5]. ML estimate θ is obtained by maaximiing (B.14) with respect to θ subject to the constraints (B.7)–(B.13). This is a very difficult optimization problem. An approximation suggested in Ref. (3) simplifies the problem tremendously. It is assumed that the filter gain K(t) and covariance B(t/t − 1) have reached constant values K and B and the vector θ consists of unknown parameters from F,G,K and B only. Then

(B.15) $$\log p(Y_N|\theta) = -\frac{1}{2} \sum_{j=1}^{N} [\nu^T(j)B^{-1}\nu(j) + \log|B|].$$

Maximizing (B.15) over B, produces

(B.16) $$\hat{B} = \frac{1}{N} \sum_{j=1}^{N} \nu(j|\hat{\alpha})v^T(j|\hat{\alpha})$$

where a is the ML estimate of unknowns in F, G and K. It is given by the root of the equation

(B.17) $$\sum_{j=1}^{N} \nu^T(j) B^{-1} \frac{\partial \nu(j)}{\partial \alpha} = 0$$

where $(\partial v(j))/\partial\alpha$ is calculated from equations (B.7)–(B.9). The root of equation(B.17) is found by a Newton–Raphson or Gauss–Newton iteration. Once $\hat{\alpha}$ is obtained, Γ,Q and R are obtained from equations (10)–(13). In this way, the non–linear constraints of equations (10)–(13) are avoided during optimization. The above method is no more complicated than the well–known output error method. In fact, it reduces to the output error method when there is no process noise, i.e., $w(t) = 0$. In that case, $Q = 0$, $K = 0$ and $\nu(t) = y(t) - Hx(t)$ is the output error. A flow chart of the method is shown in Figure B.2.

3. Nonlinear Systems

Consider a nonlinear dynamic system

(B.18) $$x(t + 1) = f(x(t), \theta, u(t)) + \Gamma w(t)$$

(B.19) $$y(t) = h(x(t)) + v(t)$$

where f(.) and h(.) are n x 1 and r x 1 vectors of nonlinear functions. Also, w(t) and v(t) are Gaussian white noise sequences with zero mean and covariances Q and R.

The evaluation of the trule ML estimate would require the calculation of $p(y(j) | Y_{j-1}, \theta)$ using an optimal nonlinear filter. Since this is computaationally infeasible, we approximate $p(y(j) | Y_{j-1}, \theta)$ by a Gaussian density with mean and covariance obtained from an Extended Kalman Filter [8] of the following form:

(B.20) $$\hat{x}(t + 1/t) = f(\hat{x}(t/t), \theta, u(t))$$

(B.21) $$\hat{x}(t/t = \hat{x}(t/t - 1) + K(t)\nu(t)$$

(B.22) $$\nu(t) = y(t) - h(\hat{x}(t/t - 1))$$

K(t) is calculated from equations (B.10)–(B.13) by using time–varying matrices F(t) and H(t).

(B.23) $$H(t) = \left.\frac{\partial h}{\partial x}\right|_{x=\hat{x}(t/t-1)}$$

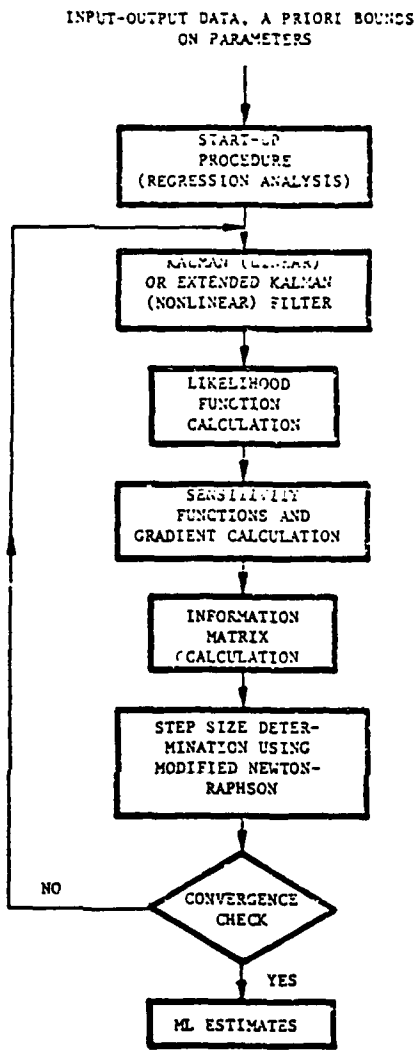

Figure B.2 Flow chart of the maximum likelihood algorithm

(B.24) $$F(t) = \left.\frac{\partial f}{\partial x}\right|_{x=\hat{x}(t/t)}$$

Kailath [6] has shown that the density of the innovation $v(t)$ tends to a Gaussian density as the sampling rate is increased. Thus the above approximation is quite good for high sampling rates.

REFERENCES (APPENDIX B)

1. E. E. P. Box and G. M. Jenkins, Times Series Analysis, Forecasting and Control, Holden Day, 1970.

2. K. J. Astrom and S. Wenmark. "The Numerical Identification of Stationary Times Series," 6th International Instruments and Measurements Congress, Stockholm, Sweden, Sept. 1964.

3. R. K. Mehra, "Identification of Stochastic Linear Dynamic Systems Using Kalm Filter Representation," AIAA Journal, 9, No. 1, 28-31 January 1971.

4. R. K. Mehra, "On-Line Identification of Linear Dynamics Systems with Applications to Kalman Filtering," IEEE Trans. Automatic control, AC-16, No 1 February 1971.

5. T. Kailath, "A General Likelihood - Ratio Formula for Random Signals in Gaussian Noise," IEEE Trans, IT, May 1969.

6. R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," Trans ASME J. Basic Eng. 82, 34-45.

7. R. E. Larson, R. M Dressler, and R. S. Ratner, "Application of the Extended Kalman Filter to Ballistic Trajectory Estimation," Final Report SRI, Proj. 5188-103 January 1967.

8. R.K. Mehra and J. S. Tyler, "Case Studies in Aircraft Parameter Identification," 1973 IFAC Sym. on Identification, The Hague, Netherlands.