SELF TUNING CONTROL

AND

SYSTEM IDENTIFICATION

by

D. L. PRAGER

A thesis submitted to

THE VICTORIA UNIVERSITY OF MANCHESTER

for the degree of

DOCTOR OF PHILOSOPHY

in the Faculty of Technology

Control Systems Centre
University of Manchester
Institute of Science and Technology

October 1980

DECLARATION

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

D. L. Prager

ABSTRACT

New techniques in the related areas of system identification and self-tuning control are discussed. In the first part of the thesis an approach to the maximum likelihood estimation of the parameters of an ARMA model for single input single output systems based upon hypothesis testing is described. This is then extended to cater for multivariable systems where the minimization of the trace of the residual covariance matrix is also considered. The advantage of the technique over conventional hill-climbing or recursive approximate methods is the large amount of graphic information, including sections through the cost function, that is available. This allows the analyst to see the likelihood function evolve and make subjective decisions based upon the available information. The algorithm has excellent convergence properties making it ideal for short data runs. Simulation examples demonstrate its properties and compare the technique with alternative algorithms yielding consistent parameter estimates.

The second part of the thesis introduces two new multivariable self-tuning regulators, namely the detuned minimum variance and pole-shifting regulators and shows that they possess the 'self-tuning property'. The regulators are intended to overcome some of the limitations of existing designs. Both regulators offer a real alternative to the conventional self-tuning optimal control objective by allowing the specification of closed loop system poles. The pole-shifting regulator is exceptionally robust and can be applied to non-minimum phase systems. It is often suitable for systems with unknown or time-varying time delays and systems in which the input-output relations have differing pure time delays. Simulation examples and an example of the control of a hydraulic system illustrate the features of the algorithm.

ACKNOWLEDGEMENTS

I am greatly indebted to my supervisor, Dr. P.E. Wellstead for the invaluable guidance and assistance he has given me. His ready help and advice at every stage of this work is sincerely appreciated.

The work was carried out while I was employed as a Special Research Assistant in Control Systems Centre and I should like to record my gratitude for the opportunity this appointment afforded me. Thanks are also due to the Staff and Students at Control Systems Centre for the many stimulating and valuable discussions.

Finally I wish to thank my wife Siobhan for her patience and encouragement during the drafting of the manuscript and her assistance in preparing the final copy.

BIOGRAPHICAL DETAILS

The author was born in 1952 in Johannesburg, South Africa and obtained the degree of B.Sc. (Eng.) (Cum Laude) in Electrical Engineering from the University of the Witwatersrand, Johannesburg in 1974. Following the award of the Chamber of Mines of South Africa Research Scholarship the author continued his studies at Control Systems Centre, UMIST, obtaining the degree of M.Sc. in 1977. The work leading to this thesis was carried out from October 1976 to December 1978 while the author was employed as a Special Research Assistant in Control Systems Centre.

TO SIOBHAN

CONTENTS

CHAPTER 1

INTRODUCTION

System Identification is becoming an increasingly valuable technique in the field of control engineering and indeed in all sciences where the inference of process model parameters from observed data offers an attractive alternative to purely theoretical or physical models. The identification of stochastic linear dynamical systems in particular has attracted a great deal of interest. In the field of control, linear models are useful because linear design techniques are well established. If the process is actually a non-linear one, then it is common practice to generate a set of models corresponding to different operating points. Many estimation algorithms are designed to operate on-line and clearly it is a fairly small step to propose simultaneous on-line identification and controller design. Combined Identification and Control algorithms open the way to the inviting prospect of adaptive controls, modifying themselves to changes in system characteristics so as to achieve some pre-programmed control objective.

If one views the broader field of Identification and Control, there are therefore perhaps two primary topics of importance. The first of these is Identification per se, including the techniques for structural estimation and parameter estimation (both on-line and off-line). The aim here is to determine a process model in a form suitable for further off-line system analysis or controller design using standard techniques. The second topic is concerned with combined, on-line, Identification and Control.

Aspects of both these topics are discussed in this thesis, and the work therefore divides naturally into two related, but distinct parts. In the first part (Chapters 3 - 5) a new approach to parameter estimation based on a recursive hypothesis testing procedure is discussed. The method (named 'RHYP') is developed first for Single-Input Single-Output (SISO) systems under the assumption of an Auto-Regressive Moving-Average (ARMAX) model including both deterministic and noise inputs, and is designed to generate maximum likelihood estimates. It is extended for multivariable systems (when the algorithm is called 'MVHYP') with the objective of minimizing either a cost function based on the trace of the model residual covariance matrix or the maximum likelihood cost function. The work was motivated by the observation that present techniques for obtaining consistent parameter estimates (which are generally based on non-linear maximum likelihood estimation) are not always satisfactory. Approximate linear recursive methods can be unreliable, especially when only short data records are available, and the more exact hill-climbing approach involves a great deal of computation but gives little information about the nature of the likelihood function or insight into the estimation problem. The objective therefore was to develop an estimation tool which in addition to the conventional numerical diagnostics (for example parameter covariances) increased the analysts interaction with the estimation algorithm by giving graphic information about the nature of the cost function and produced reliable estimates even with short data records.

The second part of the thesis (Chapters 6 - 8) is devoted to combined Identification and Control, in particular, the family of

methods known collectively as 'Self Tuning Controllers'. The contribution here is the development of two new multivariable self tuning regulators, the first being an extension of the minimum variance regulator, namely the detuned minimum variance regulator, and the second, a pole-shifting regulator. Both regulators allow (with certain restrictions) the specification of the closed loop system poles. Although the detuned minimum variance regulator is related to optimal control methods and is intended to add flexibility to the standard minimum variance regulator which may be used to avoid difficulties due to excessive control excursions in this type of regulator, the pole shifting regulator deviates from the existing multivariable self tuning approach and its roots lie more in classical control concepts. The general robustness of the latter algorithm makes it particularly attractive and gives it certain unique features amongst self tuning regulators which are discussed in the thesis. In particular, it is capable of regulating non-minimum phase systems and frequently may be used to regulate systems in which there are differing pure time delays in the various input-output relationships.

The thesis is organised as follows. Chapter 2 surveys the literature relevant to the topics of Estimation and on-line Identification and Control discussed in the thesis. However, not all the review material is contained in this chapter and particularly pertinent material is covered in greater depth elsewhere.

Chapter 3 marks the start of the first part of the thesis which is concerned with parameter estimation. The properties of maximum likelihood estimation are discussed and existing techniques for obtaining maximum likelihood estimates reviewed.

Chapter 4 describes the new estimation algorithm, RHYP, for obtaining maximum likelihood estimates of the parameters of a SISO system ARMAX model by recursive hypothesis testing. Two non-recursive variants of the method, intended to reduce computation, are also described, and examples are given illustrating the advantages of the method over techniques such as 'Recusive Maximum Likelihood' (Soderstrom, 1973) and 'Instrumental Variables' (Wong and Polak, 1967, etc.)

The multivariable extension of the algorithm (MVHYP) is given in Chapter 5. The minimization of both the maximum likelihood cost function and the trace of the model residual covariance matrix are discussed.

Two papers (Prager and Wellstead 1979 ; Prager and Wellstead, 1981 (to appear)) have been written based on the material in Chapters 4 and 5.

The second part of the thesis, concerned with Self Tuning Control, begins with Chapter 6. This chapter explores control strategies (assuming off-line controller design using known or estimated system models) which are suitable for implementation in a self tuning configuration. Certain review material is included, in particular a description of the multivariable minimum variance regulator (Borisson, 1975) and the generalized minimum variance controller (Clarke and Gawthrop, 1975; Gawthrop, 1977). Two new designs, namely the multivariable detuned minimum variance regulator and multivariable pole-shifting regulator are introduced. The features of these strategies are discussed and compared with existing methods.

Chapter 7 discusses the self-tuning versions of the laws presented in Chapter 6. Review material corresponding to that given in the previous chapter is included. However, the main purpose of the chapter is to present the self-tuning multivariable detuned minimum variance and pole shifting regulators. The proofs of the self tuning properties of these laws are derived from a common 'Self Tuning Lemma'. Simulation examples highlighting the features of the new algorithms, in particular, those of the self tuning pole shifting regulator are given.

The practical aspects of self tuning are extremely important, and a brief introduction to the implementation of self tuning laws is given in Chapter 8. The comments are drawn from the authors observations in numerous simulation runs, both digital and analogue. The performance of the multivariable pole shifting regulator in a 'real' application is documented. The process under control is a coupled tank system which exhibits typical non-linear characteristics such as saturation, and dynamics varying significantly with operating point. Encouraging results were obtained and point to the usefulness of the self tuning technique.

A number of publications have been based on the material in Chapters 6, 7 and 8. (Wellstead, Edmunds, Prager and Zanker, 1979, 1980 ; Wellstead, Prager and Zanker, 1979 ; Prager and Wellstead, 1979a).

Finally, Chapter 9 concludes by summarizing the results and discusses possible future areas of research.

CHAPTER 2

A SURVEY OF THE LITERATURE

The subjects of estimation and self-tuning controller theory both support a growing volume of literature. Activity in the field of self-tuning controllers in particular has received great impetus from advances in micro-electronics technology which has encouraged the practical development of the theory. This chapter gives an overview of the most significant developments so as to place the material in the ensuing pages in its proper context. Throughout the thesis particularly relevant work is reviewed in greater detail so that, for example, Chapter 3 contains a more thorough presentation of maximum likelihood estimation, and Chapters 6 and 7 discuss the broader class of digital controllers which lend themselves to self-tuning.

Clearly, as an estimator is at the heart of the self-tuning algorithm, the subjects of estimation and self-tuning are closely related, the latter exploiting recursive methods of estimation. In this chapter however, in keeping with the rest of the thesis, the subjects are discussed separately.

2.1 Estimation of the Parameters of a Model

In the field of control engineering, the techniques of parameter estimation are relatively new being popularized to a large extent by the well known work of Kalman (1960). So it is particularly sobering to observe that the fundamental principles of least squares estimation

were known and used as long ago as the early 19th century by Legendre (1806) and Gauss (1809). The more recent interest in engineering estimation is to a large extent due to the need for better control of processes and improved techniques in forecasting which in turn require an improvement in system modelling. Parallel work in econometrics, where complex economic models are attracting considerable attention (e.g. Bray, 1975), biological modelling (e.g. Beck, 1977) and statistics has clearly led to an interchange of ideas and techniques.

The subject of parameter estimation as treated in this thesis is centred about the Auto-Regressive Moving-Average (ARMAX) model whereby the output (y_t) of a system is related to its deterministic input (u_t) and a non-measurable stochastic noise input (e_t) by a discrete-time difference equation of the form

$$(1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_{n_a} z^{-n_a}) y_t \qquad 2.1$$
$$= z^{-k}(b_1 z^{-1} + \ldots + b_{n_b} z^{-n_b}) u_t + (1 + c_1 z^{-1} + \ldots + c_{n_c} z^{-n_c}) e_t$$

where z^{-1} denotes the backward shift operator. The disturbance input $\{e_t\}$ is a white noise sequence and the equation is written here for Single-Input Single-Output (SISO) systems. The above equation may be written more concisely as

$$\{1+A(z^{-1})\}\, y_t = z^{-k} B(z^{-1})\, u_t + \{1+C(z^{-1})\}\, e_t \qquad 2.2$$

where the backward shift polynomial notation is self explanatory. The use of this model has been actively promoted by Astrom and others and can model any linear system if the polynomial orders are appropriately selected. Its relationship with the form of the s-domain

transfer-function representation is clear if eqn. 2.2 is re-written as

$$y_t = \frac{z^{-k}B(z^{-1})}{1+A(z^{-1})} u_t + \frac{1+C(z^{-1})}{1+A(z^{-1})} e_t \qquad 2.3$$

and thus provides a natural transition for the classical control engineer to the world of discrete time modelling and control.

Of course the ARMAX model is not the only model one might want to use to describe the behaviour of a system. A far more intuitive approach is to postulate a model

$$y_t = \frac{B(z^{-1})}{1+A(z^{-1})} u_t + \varepsilon_t \qquad 2.4$$

for a given system and estimate the relevant parameters of $A(z^{-1})$ and $B(z^{-1})$ such that the sum of the squares of the errors $\left(\sum_{i=1}^{N} \varepsilon_i^2 \right)$ between the model and the actual system output is minimized over the given set of N observations of input-output data. Indeed, such a 'direct' model is particularly useful in obtaining the 'best' (in the least squares sense) match to a given set of data when the model order chosen is less than the true system order. Like the ARMAX model, the parameter estimation involved is a non-linear operation. However, it has the distinct disadvantage that, even when the deterministic model has the same order as the true system, the parameter estimates will generally be biased, depending on the colouration of the residual sequence ε_t. The results obtained are therefore extremely data dependant and lack the statistical advantages of the ARMAX model.

A crucial difference is that the ARMAX model is a prediction model. It is quite easy to show that the optimum one-step-ahead least

squares prediction can be computed directly from the ARMAX model as follows: (see for example Astrom and Bohlin, 1965)

$$\hat{y}_{t/t-1} = -A(z^{-1})y_t + z^{-k}B(z^{-1})u_t + C(z^{-1})e_t \qquad 2.5$$

and $$e_t = y_t - \hat{y}_{t/t-1}$$

The general ARMAX model in eqn. 2.2 is often termed the 'maximum likelihood' model because asymptotically unbiased estimates of its parameters may be derived using the maximum likelihood technique, a method that is discussed in detail later. The maximum likelihood estimator is in general non-linear. There is however, an extremely useful exception, namely if polynomial $C(z^{-1})$ is zero. Not only is the estimator linear in the parameters of $A(z^{-1})$ and $B(z^{-1})$ but the estimates can be computed recursively. The recursive least squares parameter estimator that is used is directly equivalent to the well known Kalman Filter (Kalman, 1960), a 'state estimator'. The relationship is extremely straight forward to illustrate, for the 'least squares' ARMAX model

$$y_t = -A(z^{-1})y_t + z^{-k}B(z^{-1})u_t + e_t \qquad 2.6$$

can clearly be re-written as follows:

$$\theta_{t+1} = \theta_t \qquad 2.7$$

$$y_t = x_t^T\theta_t + e_t \qquad 2.8$$

where θ_t is a vector of parameters drawn from $A(z^{-1})$ and $B(z^{-1})$ and x_t is a vector of past inputs and outputs, u_t and y_t. Thus the parameters of the ARMAX model become the states of the constant state system of eqn. 2.7 and can be estimated by direct application

of the Kalman Filter.

Broadly speaking, work in the field of recursive parameter estimation may be divided into two main parts: †

a) Development of faster numerical techniques for realizing existing estimators, in particular linear least squares.

b) The development of linear algorithms which approximate non-linear estimators.

These topics, as well as some aspects of model order testing, are discussed in the following sections, where, given a set of N data records, the vector equation

$$Y_N = X_N\hat{\theta} + E_N \qquad 2.9$$

is often used. Y_N is the vector formed by collecting all N output observations,

$$Y_N = (y_N, y_{N-1}, \ldots, y_1)^T \qquad 2.10$$

$$X_N = (x_N, x_{N-1}, \ldots, x_1)^T \qquad 2.11$$

and the residual vector is E_N. Then the least squares estimate $\hat{\theta}$ is found from

$$\hat{\theta} = (X_N^T X_N)^{-1} X_N^T Y_N \qquad 2.12$$

(see for example Eykhoff, 1974).

---

† In addition to these there remains the vitally important field of model order testing and structural identification which is not explicitly discussed in this thesis.

2.1.1 Numerical Aspects of Parameter Estimation

One of the earliest examples of work in this category is Levinson's (1947) recursion for solving the normal least squares equations (eqn. 2.12). The method makes use of the symmetric Toeplitz structure of the correlation matrix X^TX for an all pole auto-regressive (AR) model where the entire signal y_t, t=0,1, ... ,∞ is used. In practice only finite data is available, and the elements of the matrix are computed from the autocorrelation coefficients of the available data. Durbin (1959) has modified Levinson's original algorithm and the method can also be extended to deal with the more general auto-regressive moving-average (ARMA) model. The importance of Levinson's algorithm is that it reduces the amount of computation necessary to solve the normal equations for an n-parameter problem from a quantity proportional to n^3 to an amount proportional to n^2.

However, perhaps the most well known contribution to numerical techniques in estimation is the Kalman Filter implementation of recursive least squares already mentioned above. This still remains a good technique for real time identification. The computational requirement is proportional to the square of the number of parameters estimated. The parameter set at time t+1, θ_{t+1}, is derived from that at time t, θ_t, by adding a value proportional to the prediction error

i.e. $$\hat{\theta}_{t+1} = \hat{\theta}_t + K_{t+1}(y_{t+1} - x_{t+1}^T \hat{\theta}_t) \qquad 2.13$$

where K_{t+1} is known as the Kalman Gain, and is computed from

$$K_{t+1} = P_{t+1} x_{t+1} \qquad 2.14$$

and $$P_{t+1} = (X_{t+1}^T X_{t+1})^{-1} \qquad 2.15$$

P_{t+1} (known as the covariance matrix) is itself computed recursively from

$$P_{t+1} = P_t - (P_t x_{t+1} x_{t+1}^T P_t)/(1+x_{t+1}^T P_t x_{t+1}) \qquad 2.16$$

Numerical errors in this latter recursion can lead to the covariance matrix becoming non-positive definite, resulting in disastrous instability in the parameter update equation as a consequence. In order to avoid such difficulties, "Square Root" algorithms have been derived which update the square root of P_t. (see for example Peterka, 1975). Unfortunately the extra degree of numerical stability afforded by square root algorithms must be weighed against their greater computational complexity.

Referring to the recursive least squares update equations 2.13-2.16 above it is evident that the evaluation of the gain K_{t+1} is the main computational task at each iteration. Kailath and his co-workers (Morf, 1974; Morf and Ljung, 1976; Morf, Ljung and Kailath, 1976; Ljung, Morf and Falconer, 1978) have exploited the 'shift invariance' property of the matrix P_t^{-1} to develop fast methods for generating the Kalman Gain in a manner which parallels the Levinson method. By 'shift invariance' is meant the property that element (i,j) of each block of the block Toeplitz matrix P_t^{-1} is merely shifted to position (i+1,j+1) at time t+1. Robins and Wellstead (1979) discuss extensions of the basic algorithm for multivariable ARMA models, instrumental variable type estimators, and highlight the structural decomposition that is inherent in the estimation algorithm. The numerical properties of the algorithm are as yet undocumented, and the fact that the programming of the algorithm is complex and error prone

might be disadvantageous. It is vital to appreciate that the efficiency of the algorithm is entirely dependent on the degree of 'shift invariance' so that it may be expected to perform well for a high order AR model, but would perform very poorly for a multivariable AR model with low order autoregression which has little or no shift property.

2.1.2 Estimation Techniques

The numerical developments described above are all associated in the main with linear least squares estimation. Unfortunately least squares is not always reliable. In particular, when the system $C(z^{-1})$ polynomial is non-zero, the estimates of the polynomials $A(z^{-1})$ and $B(z^{-1})$ will always be biased. (Eykhoff, 1974) A theoretically satisfying alternative is to use maximum likelihood estimation (ML) which, for SISO systems, reduces to a slightly more complicated version of least squares, namely the minimization of $\sum_{t=1}^{N} \varepsilon_t^2$, where

$$\{1+\hat{C}(z^{-1})\}\varepsilon_t = \{1+\hat{A}(z^{-1})\}y_t - z^{-k}\hat{B}(z^{-1})u_t \qquad 2.17$$

The non-linearity introduced by the incorporation of the $\hat{C}(z^{-1})$ polynomial alters, numerically, the whole nature of the estimation problem and has motivated an important effort in the field, namely research into good, preferably linear, numerically attractive estimation algorithms. Initial work, however, was based on hill climbing algorithms of which perhaps the best known is Astrom and Bohlin's (1965) discussed in the following chapter.

Clearly, as there is a definite link between the least squares and ML cost functions, the temptation is to modify the existing

Recursive Least Squares (RLS) method to accommodate the noise colouration polynomial. One popular approach is that of generalized least squares, (Clarke 1967; Hastings-James and Sage, 1969), RGLS. Basically, the method assumes a system equation

$$y_t = \frac{z^{-k}B(z^{-1})}{1+A(z^{-1})} u_t + \frac{1}{1+F(z^{-1})} e_t \qquad 2.18$$

where the symbols have their usual meaning and e_t is a zero mean white noise. The estimation is a two stage procedure the first part being the estimation of the parameters $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$ in the model

$$\{1+\hat{A}(z^{-1})\}y_t = z^{-k}\hat{B}(z^{-1})u_t + w_t \qquad 2.19$$

by least squares. (w_t is the residual sequence). Then w_t is itself employed in estimating the noise colouration polynomial $F(z^{-1})$ by using least squares to estimate $\hat{F}(z^{-1})$ in

$$w_t = -\hat{F}(z^{-1})w_t + \varepsilon_t \qquad 2.20$$

(ε_t is now the residual sequence, hopefully tending to the white noise e_t.) In the next iteration, the data $\{u_t\}$ and $\{y_t\}$ is filtered by $\hat{F}(z^{-1})$ such that

$$\tilde{y}_t = \{1+\hat{F}(z^{-1})\}y_t \qquad 2.21$$

$$\tilde{u}_t = \{1+\hat{F}(z^{-1})\}u_t \qquad 2.22$$

and the new estimates of $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$ are sought to minimize the sum of the squares of the residual sequence $\{\tilde{w}_t\}$ in the model

$$\{1+\hat{A}(z^{-1})\}\tilde{y}_t = z^{-k}\hat{B}(z^{-1})\tilde{u}_t + \tilde{w}_t \qquad 2.23$$

The method is computationally attractive. However, convergence

cannot be proved in general (Soderstrom, 1974a). Later variants of the algorithm have been proposed by Gertler and Banfasz (1974) and Sen and Sinha (1975).

An obvious extension of least squares is to include the past values of the residuals in the regression equation and this approach is commonly known as 'Extended Least Squares' (ELS). Variants of it are described by Panuska (1968 , 1969), Young (1968) , Kashyap (1974) and Soderstrom, Gustavsson and Ljung (1974). Basically, the model is 'extended' to

$$\{1+\hat{A}(z^{-1})\}y_t = z^{-k}\hat{B}(z^{-1})u_t + \hat{C}(z^{-1})\varepsilon_t + \varepsilon_t \qquad 2.24$$

where ε_t is the residual and past values of the residual are entered as an extra input to the system. The parameters are estimated in a RLS-type algorithm (or a stochastic approximation algorithm). The method can give good results and the estimates are consistent when the algorithm converges. However, Soderstrom, Gustavsson and Ljung (1974) have shown that the algorithm may not be stable about the true parameter vector and have demonstrated this phenomenon in an example.

Soderstrom (1973) has developed an algorithm very similar in form to ELS but with superior convergence properties. His RML (Recursive Maximum Likelihood) algorithm is often used as a benchmark against which to measure new algorithms and for this reason is considered more fully in the following chapter. At this stage it is sufficient to note that the algorithm replaces the regression variables of the ELS algorithm by values filtered by the latest

estimate of the $1+C(z^{-1})$ polynomial, that is for example

$$y_{f_t} = \frac{y_t}{1+\hat{C}(z^{-1})} \qquad 2.25$$

The estimation is carried out using the well known RLS algorithm so that the update equations become

$$\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1}x_{t+1}^T(y_{t+1} - x_{t+1}^T\hat{\theta}_t) \qquad 2.26$$

$$P_{t+1} = P_t - \gamma_t P_t x_{t+1} x_{t+1}^T P_{t+1} \qquad 2.27$$

$$\gamma_t = (1+x_{t+1}^T P_t x_{t+1})^{-1} \qquad 2.28$$

where $x_t = (-y_{f_{t-1}},..,-y_{f_{t-n_a}},u_{f_{t-k-1}},..,u_{f_{t-k-n_b}},\varepsilon_{f_{t-1}},..,\varepsilon_{f_{t-n_c}})^T$ 2.29

and ε_t satisfies

$$\varepsilon_t = \{1+\hat{A}(z^{-1})\}y_t - z^{-k}\hat{B}(z^{-1})u_t - \hat{C}(z^{-1})\varepsilon_t \qquad 2.30$$

as in the ELS method.

It is not always necessary to estimate the noise colouration polynomial $C(z^{-1})$ coefficients. The 'Instrumental Variable' technique (Wong and Polak, 1967; Young, 1970, 1976; Young, Shellswell and Neethling, 1971) provides a means for obtaining asymptotically unbiased estimates of the $A(z^{-1})$ and $B(z^{-1})$ polynomial coefficients without considering the noise colouration. The principle of the method is extremely simple and relies on pre-multiplying eqn. 2.9 by W^T so that

$$W^TY = W^TX\hat{\theta} + W^TE \qquad 2.31$$

where W is called the 'Instrumental Matrix' and satisfies

$$E(W^T E) = 0$$

$$E(W^T Y) \quad \text{non-singular}$$

These properties require the elements of W to be uncorrelated with the residuals but not with the input-output data. (See for example Kendal and Stuart, 1961). Clearly, the asymptotically unbiased estimates $\hat{\theta}$ are obtained from

$$\hat{\theta} = \lim_{N\to\infty} \left(\frac{1}{N} W_N^T X_N \right)^{-1} \left(\frac{1}{N} W_N^T Y_N \right) \qquad 2.32$$

The recursive evaluation of $\hat{\theta}$ is very similar to recursive least squares (Eykhoff, 1974). In systems where the input u_t is not a function of the output y_t (i.e. no feedback) u_t may be used as an 'instrumental variable' and estimates of y_t are often obtained by using a cruder model ('instrumental model') derived from least squares estimation. Then W_N is a matrix of the same form as X_N except that it is built from u_t and estimates of y_t rather than the noisy measured output itself. The method does not converge generally for all kinds of systems, inputs and choices of instrumental model (Soderstrom, 1974b). Young (1974) has also described a method by which the Recursive Instrumental Variable (RIV) method may be combined with the RELS method to establish the noise dynamics. The $A(z^{-1})$ and $B(z^{-1})$ polynomial coefficients are first estimated by RIV and the residual then fed to a RELS algorithm for estimating the noise dynamics. A refinement of this technique is described in Young (1976).

A unified analysis of the RLS, RGLS, RELS, RIV and RML algorithms described above is given in Soderstrom, Ljung and Gustavsson (1978) where it is concluded that Soderstrom's RML technique gives the best

results, certainly over a long run of 2000 data samples.

An earlier analysis by Isermann, Baur, Bamberger, Kneppo and Siebert (1974) considers the RGLS, RLS, and RIV algorithms plus a two stage stochastic approximation algorithm (STA) due to Saridis and Stein (1968). The algorithm computes the impulse response from which the $A(z^{-1})$ and $B(z^{-1})$ parameters are estimated using RLS. A Fourier Analysis technique which approximates the process by a model of the form

$$\frac{k}{(1+Ts)^n} \; e^{-T_D s}$$

where n is assumed known and k, T and T_D must be estimated (see Isermann 1973) is also considered. The best overall results however were obtained with a correlation analysis cum least squares algorithm (COR) in which the ordinary least squares model is pre-multiplied by $u(k-\tau)$ and thus on taking mathematical expectations on both sides of the premultiplied model equation, an equation in terms of the auto- and cross-correlation functions $\phi_{uu}(\tau)$, $\phi_{uy}(\tau)$, and the model parameters emerges. The parameters are found by using least squares on the set of equations that can be generated in terms of ϕ_{uu} and ϕ_{uy}, quite analogous to the basic equations in terms of u_t and y_t. The authors claim that the algorithm always converges, and that the cross correlation function as an intermediate result is useful for detecting time delays and allows the structure and order of the parametric model to be preselected easily. However, one should not lose sight of the fact that the least squares step in the algorithm requires the inversion of a matrix and that RLS cannot be used. RIV is found to give good results when it converges, although convergence difficulties

have been encountered. This has also been the author's experience, difficulties occurring especially when the model order has not matched the true system order.

Most of the algorithms mentioned above, in particular, RLS, RIV, RGLS, RELS and RML will generally run in real time. However, a great deal of identification analysis is carried out off-line from logged system input-output data and it is therefore not necessarily a disadvantage to use an algorithm which is not suitable for real time processing. Mayne and Firoozan (1977) have developed an interesting off-line technique which is computationally more efficient than the standard hill-climbing approach.

Consider the ARMA model

$$\{1+\bar{A}(z^{-1})\}y_t = \{1+C(z^{-1})\}e_t \qquad 2.33$$

where, for simplicity, the deterministic input u_t is omitted on this occasion. The expression is equivalent to

$$\{1+\bar{A}(z^{-1})\}y_t = e_t \qquad 2.34$$

where $1+\bar{A}(z^{-1}) = \{1+C(z^{-1})\}^{-1}\{1+A(z^{-1})\}$, a polynomial of infinite order. Thus a reasonable estimate ε_t of e_t could be obtained by modelling equation 2.33 by an autoregressive model of high order (say order q)

$$\{1+\hat{\bar{A}}(z^{-1})\}y_t = \varepsilon_t \qquad 2.35$$

Here, the estimate $\hat{\bar{A}}(z^{-1})$ is chosen so as to minimize the sum of squares of the residual sequence $\{\varepsilon_t\}$. Estimates of $A(z^{-1})$ and $C(z^{-1})$

are now obtained by a least squares procedure in which the non-measurable noise input e_t is approximated by ε_t. This approach is described by Durbin (1961) who points out that, although computationally simple, the estimates are not efficient. Mayne and Firoozan have extended the above method and derived an algorithm for SISO systems which yields consistent and asymptotically efficient estimates. The above 2-step least squares procedure is used to obtain initial estimates $\hat{A}_1(z^{-1})$, $\hat{C}_1(z^{-1})$ and the estimate $\{\varepsilon_t\}$ of the white noise sequence $\{e_t\}$. The data $\{y_t,\varepsilon_t\}$ is then filtered to produce sequences $\{\bar{y}_t\}$ and $\{\bar{\varepsilon}_t\}$ where

$$\{1+\hat{C}_1(z^{-1})\}\bar{y}_t = y_t \qquad 2.36$$

and

$$\{1+\hat{C}_1(z^{-1})\}\bar{\varepsilon}_t = \varepsilon_t \qquad 2.37$$

Final estimates $\hat{A}_2(z^{-1})$ and $\hat{C}_2(z^{-1})$ are then obtained which minimize

$$J = \sum_{t=1}^{N} \{(1+\hat{A}_2(z^{-1}))\bar{y}_t - (1+\hat{C}_2(z^{-1}))\bar{\varepsilon}_t\}^2 \qquad 2.38$$

This is a simple linear least squares stage. The data $\{y_t,\varepsilon_t\}$ may be filtered using the latest estimate of $1+C(z^{-1})$, and eqn. 2.38 repeated as often as desired.

The method has the disadvantage that it cannot be used on-line but it requires only a number of simple passes through the data using linear estimation and must therefore be regarded as an attractive approach. The convergence properties however depend on the appropriate choice of q.

An entirely different approach to the problem of parameter estimation is that of Sequential Hypothesis Testing. Hypothesis testing

is well established in the field of statistics (see for example the classic work of Wald, 1945) and has also been applied to the detection problem in communication engineering (Van Trees, 1968). Two recent applications in parameter estimation are due to Nebeker (1976) and Tomanek (1976). Both base their technique on computing the maximum a posteriori estimate by setting up a set of hypotheses and recursively determining the a posteriori probabilities for each member of the set using Bayes Rule. This involves running a bank of Kalman Filters in parallel. Interestingly, the maximum a posteriori estimate is equivalent to the maximum likelihood estimate in the case where the a priori probabilities of the hypotheses are equal. This is often assumed if no a priori information is available. Hypotheses on the entire parameter vector are set up and consequently the algorithms demand enormous computational resources. In spite of the criteria for eliminating unpromising hypotheses and the introduction of 'stopping criteria' the algorithm is untenable in a real time environment except in the slowest of processes. Nevertheless, the fact that sequential hypothesis testing simultaneously considers many possible parameter vectors gives it a substantial headstart over traditional techniques which iterate about a potential parameter vector. The resulting improvement in convergence over Soderstrom's RML method is demonstrated by Nebeker for a first order model.

The concept of sequential hypothesis testing is related to the RHYP method developed in this thesis.

2.1.3 Model Order Testing

This review has up till now been concerned with the subject of parameter estimation only, the field that is the topic of the first part of this thesis. It would however, give an unduly limited view of the field if the vitally important question of the estimation of system structure were to be neglected as this forms such a fundamental problem in system identification.

In terms of SISO systems the question of structure reduces to that of system order. For multivariable systems, the structural identification problem is one of considerably greater complexity which relates crucially to the whole question of system representation and canonical forms. (Denham 1974; Dickenson, Kailath and Morf, 1974; Glover and Willems, 1974; Risannen, 1974; Hannan, 1975, 1976; Dunsmuir and Hannan, 1976).

Numerous techniques have been proposed which seek to determine the 'correct' model order. Perhaps the most intuitively obvious approach is the minimization of the least squares or maximum likelihood cost function for various model orders. (Unbehauen and Gohring, 1974; Van den Boom and Van den Enden, 1974). The latter have shown that the behaviour of the cost function changes significantly in the neighbourhood of correct noise and process system orders for large data samples. However, the cost function is always a monotonically decreasing function of order, and thus, when the observations are highly contaminated by noise, the breakpoint is difficult to ascertain.

Another fairly obvious approach is to monitor the singularity of

the matrix X^TX (see eqns. 2.11 and 2.12). In the absence of disturbances this matrix becomes singular if the true system order is exceeded. This is due to the linear constraint imposed on the observations. The singularity property can be exploited in determining correct model order (Chow 1972). A similar approach is to test the covariance matrix of the parameter estimates which also becomes singular for the case of an over-parametrized model with noiseless observations. The covariance matrix is related to $(X_N^TX_N)^{-1}$.

The above methods are classified as 'subjective' as the model cut off point is not clearly defined and is essentially left to the user to determine. Akaike's (1969) Final Prediction Error test removes the subjective element by defining the 'correct' model order as being that which minimizes the cost

$$J_{FPE} = \frac{N+n}{N-n} J \qquad 2.39$$

where N = total number of observations

n = number of parameters in the model

J = estimate value of the disturbance variance (obtained from the least squares cost)

and J_{FPE} = Final Prediction Error.

The method was unfortunately developed for independent observations only (i.e. $C(z^{-1})=0$) and becomes insensitive to changes in n as $N\to\infty$. A modified final prediction error criterion is given by Chan, Harris and Wellstead (1974) which allows for coloured noise and is explicitly a function of n_a, n_b and n_c and not just their sum n.

Akaike later proposed an Information Theoretic criterion, AIC,

(Akaike, 1972, 1974a, 1974b) which selects the correct model order as the minimizer of

$$AIC = -2 \ln L + 2n \qquad 2.40$$

where L is the likelihood function. A more sophisticated criterion which, unlike Akaike's, can select between system theoretically equivalent structures with the same number of parameters has been suggested by Risannen and Ljung (1976). The interesting link between maximum likelihood estimation and these information theoretic criteria is explored more fully in part of the following chapter.

Soderstrom (1977) discusses the relationship between AIC, FPE and the popular F-test and likelihood ratio methods. The F-test involves forming the quantity

$$t = \frac{J_1 - J_2}{J_2} \cdot \frac{N - n_2}{n_2 - n_1} \qquad 2.41$$

where $J_k = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i^2$ for model k

n_k = number of parameters in model k

and $n_2 > n_1$

Then t is asymptotically $F(N-n_2, n_2-n_1)$ distributed (Wilks, 1962; Astrom, 1967) and the hypotheses

H_o: order n_1

H_1: order n_2

may be selected according to the rule:

Accept H_o if $t \leq t_\alpha$

Accept H_1 if $t > t_\alpha$

where t_α is the 100α percent point of the distribution $F(N-n_2, n_2-n_1)$ given by $\text{Prob}(t > t_\alpha) = \alpha$ when H_o is true.

The likelihood ratio test (Woodside, 1971) is based on the likelihood ratio (λ) which may be expressed as

$$\lambda = \left[\frac{J_2}{J_1}\right]^{N/2} \qquad 2.42$$

H_o is accepted if λ is close to unity.

λ may also be expressed as

$$\lambda = \left[1 + \frac{(n_2 - n_1)t}{N - n_2}\right]^{-N/2} \qquad 2.43$$

where t is the F-distributed variable defined in eqn. 2.41. Thus a test on λ may be interpreted as an equivalent F-test on t.

If $J_1 - J_2 << J_2$, then

$$-2 \ln \lambda \simeq N \frac{J_1 - J_2}{J_2} = t' \qquad 2.44$$

where t' is asymptotically $\chi^2(n_2 - n_1)$ distributed. Thus the hypotheses H_o and H_1 may also be tested using the Chi Square distribution.

Soderstrom demonstrates the equivalence of the following inequalities:

a) $\text{AIC}(n_1) < \text{AIC}(n_2)$

b) $\text{FPE}(n_1) < \text{FPE}(n_2)$

c) $t < 2$

d) $t' < 2(n_2 - n_1)$

e) $-2 \ln \lambda < 2(n_2 - n_1)$

Interestingly, the Akaike criteria set a rational choice of $t < 2$

for the acceptance of the zero hypothesis for the F-test.

Many other techniques for model-order testing exist, for example, techniques based on pole-zero cancellation and whiteness of residuals and the reader is referred to Soderstrom's paper (1977) for a comprehensive list of references.

2.1.4 Other Aspects of Estimation Theory

There are clearly many other aspects of parameter estimation and identification that are important. They include subjects such as detection of feedback (see for example Caines and Chan, 1975), identification of systems under closed loop control (Wellstead and Edmunds, 1975; Soderstrom, Gustavsson and Ljung, 1975), the selection of optimum inputs (Mehra, 1974; Goodwin, Zarrop and Payne, 1974) and the vast field of state estimation which is so closely related to parameter estimation (see for example Kailath's survey paper, 1974). This review has therefore necessarily covered only those topics most relevant to the ensuing work.

2.2 Self Tuning Control

One of the prime purposes of system identification in the field of control is to construct system models which can be used in the design of controllers. The advance of real-time identification techniques has led researchers to consider the possibility of on-line estimation and control. Indeed, such a scheme could potentially realize the control engineer's dream of a 'black box' controller which adapts itself so as to successfully control any system, or which could

accommodate varying gain and dynamic characteristics of a plant. These variations might be due to time dependent changes in a plant (e.g. wear) or non-linearities resulting in operating point dependent characteristics.

Historically, this concept has its roots in a proposal by Kalman (1958) for the design of a 'self optimizing' control system. Kalman based his work on a deterministic pulse transfer function model, the parameters of which he estimated using least squares. Unlike present day configurations, Kalman limited the model order to second order, and solved the normal estimation equations not, ironically, by the recursive least squares technique he later pioneered, but by computing the required covariances and substituting these into 4 equations, one for each parameter. The parameters were updated every three time samples, and a dead-beat controller was synthesized. The computer Kalman used was externally digital, but internally analogue, using potentiometers to perform multiplication and storing numbers on potentiometers positioned by a servo arrangement.

Kalman's work was, typically, ahead of its time and it was only in the 1970's that research effort began to be directed to what have become known as 'self tuning controllers'. By that stage, faster and smaller digital machines were available which could be used to implement the new fast recursive estimators (the most popular formulation being derived directly from the Kalman Filter). The concept of stochastic control and the ARMA model formulation were also becoming more familiar to engineers and it was in this new setting that the present day self-tuning controller began to emerge.

The architecture of the self tuning controller is based on a discrete time ARMA model of the plant. Parameters of this model are estimated and the result used in deriving a control strategy. However, other stochastic adaptive controllers which rely on the familiar state space description of the plant, viz. of the type

$$x_{k+1} = Ax_k + Bu_k + Dw_k \qquad 2.45$$

$$y_k = Cx_k + v_k \qquad 2.46$$

where x_k is an n-vector of the system states

u_k is an m-vector of deterministic inputs

w_k is an m-vector of noise inputs

v_k is an 1-vector of noise inputs

y_k is an 1-vector of system outputs

and A, B, C and D are (time-varying) coefficient matrices, have also been developed and employ concepts which are similar enough to the self tuning controllers to warrant a brief review.

More specific fields of adaptive control, for example Model Reference Adaptive Control are not discussed. The reader is directed to the surveys of Landau (1974) and Hang and Parks (1973).

2.2.1 Approaches to Adaptive Control

Feldbaum (1960, 1961) has identified two basic classes of adaptive control, dual and non-dual. The dual controller is one in which the control signal performs the dual role of 'probing' and 'control'. There is thus an interaction between identification and control which takes into account that future uncertainties in the

parameters are functions of the control signal applied to the plant. A non-dual control strategy only takes into account the previous measurements and does not assume that future information will be available. The task of the dual control is therefore to find a happy compromise between the requirements of identification and control, providing a sufficiently stimulating control signal to promote good parameter estimation whilst not unduly impairing the quality of control. The minimization of a one-step-ahead cost function leads to non-dual control, whilst the minimization of a cost function several steps ahead leads to dual control.

In his survey paper, Wittenmark (1975) defines two categories of non-dual controllers, namely 'certainty equivalence' and 'cautious' controllers. In the first of these classes the certainty equivalence principle is invoked; in the second the controller is designed by appealling to the separation principle. 'Certainty equivalence' implies that it is possible to estimate the parameters of a model and then to treat these parameters as if they were exactly known in the design of a controller. It is not often strictly applicable, although it does hold for linear quadratic Gaussian control problems. The 'Separation Principle' is weaker in that the parameters of the controller may also be functions of, for instance, the uncertainties of the estimated model parameters. Bar Shalom and Tse (1974) discuss the concepts of Dual Control, Certainty Equivalence and Separation fully in their paper.

The basic idea of a certainty equivalence controller is to obtain estimates of the process model parameters by a real-time identification method, and then determine a control law from this model on the basis

that the parameter estimates are exact. This is similar to Kalman's (1958) controller although he did not consider noise. One of the recursive estimation algorithms discussed previously would typically be used to estimate the parameters of a least squares or maximum likelihood model. A popular approach to designing the control law is to minimize some loss function, for example the plant output variance, i.e. a minimum variance controller (Astrom, 1965, 1970; Peterka 1972; Schwartz and Steiglitz, 1971). When the system is modelled in state space form and the states are not directly measurable, the estimation problem becomes more complicated (non-linear) and is usually solved using an Extended Kalman Filter (e.g. Luxat and Lees, 1973). Typically, the cost function is of the form

$$J = E\left(x_N^T Q_o x_N + \sum_{t=0}^{N-1} x_t^T Q_1 x_t + u_t^T Q_2 u_t \right) \qquad 2.47$$

where the symbol x_t is the state vector, u_t is the input vector, Q_i are weighting matrices, and E(.) denotes the mathematical expectation operator. It is generally necessary to solve a new optimization problem each time the parameters change.

Another solution is to discretize the parameter space into a finite set, and then construct a Kalman Filter and obtain a minimizing control for each hypothesized parameter set. The final control signal is computed as the weighted sum of the individual controls, the problem then being reduced to a determination of weights (Saridis and Dao, 1972; Deshpande et. al., 1973).

The most relevant concept to this thesis in the class of Certainty Equivalence controllers is however the combination of

identification and control initially proposed by Peterka (1970). By using a prediction model and least squares estimator, he directly obtained the parameters of a controller which asymptotically converges to the same minimum variance control law that could have been computed from the parameters of the maximum likelihood model. The 'self tuning property' and a review of self-tuning controllers is discussed in a separate section of this chapter.

Cautious controllers, based on the Separation Principle differ from Certainty Equivalence controllers in that not only the model parameter estimates, but also the parameter covariances, are used to determine the control law. The effect of including the parameter covariance is to make a more 'cautious' controller (low gain) when the estimates are poor (high variances). This can have the unfortunate effect of inducing extremely small excitation signals resulting in even poorer estimates eventually leading to the inputs turning off altogether until such time as system noise disturbances excite the system so as to improve parameter estimates. Wieslander and Wittenmark (1971) working with a least squares type model give an example of this phenomenon. The turn-off phenomenon results from minimizing a one-step-ahead cost function. The controller is not rewarded if it produces system inputs which give better estimates which in turn may be used to improve control in future steps as is the case in the dual controller. Modifications to the scheme may be introduced which act to overcome this problem, and thus approximate the behaviour of a dual controller. These are discussed later.

In the case of state-space models, the loss function in eqn. 2.47 is generally to be minimized given that the parameters are unknown and

possibly stochastic. Thus there are two significant difficulties, firstly the non-linear estimation problem, and secondly the multi-step minimization. If the parameters are known, or if the stochastic process describing the parameters is independent between sampling times ('white' parameters) it turns out that the certainty equivalence principle holds. However, only approximate solutions to the more general case may usually be obtained.

Farison, Graham and Shelton (1967) simplify the estimation problem by assuming that the state vector may be measured exactly. The estimation of the parameters is then a linear problem. The control is determined so as to minimize a cost function on the basis of the present model of the system, say at time k. After applying control u_k the problem is repeated for k+1, and so on.

Tse and Athans (1972) considering a system in which only the zeros are unknown, and Ku and Athans (1973) who allow both unknown poles and zeros use the Open Loop Feedback Optimal (OLFO) approach. This controller does not take into account that future measurements will become available. The 'cost to go' depends on the time evolution of the expected value of future states ($\hat{x}_{j/k}$) and state error moment matrix ($\Sigma_x(j/k)$) conditioned on past data, and on the future values of the controls which are treated as deterministic. This deterministic cost-to-go, together with a set of deterministic equations describing the dynamics of $\hat{x}_{j/k}$ and $\Sigma_x(j/k)$ for j=k,k+1, ... ,N-1 defines a deterministic optimal control problem whose solution yields the optimum future open-loop controls. Unfortunately OLFO controllers can be over cautious because of the assumption that no future measurements will be available to correct for erroneous control actions.

The more complex 'Dual' controllers may be classified as 'optimal' and 'sub-optimal'. It is conceptually possible to solve the optimal dual control problem by solving the stochastic dynamic programming equation associated with the problem (Bellman 1961). But in practice, a numerical solution is prohibited by the 'curse of dimensionality'. Florentine (1962), Jacobs and Langdon (1970) and Astrom and Wittenmark (1971) have produced solutions for extremely simple systems.

Since solutions to the optimal dual control problem are so difficult to compute, researchers have attempted to find sub-optimal solutions that nevertheless have dual properties. There are two approaches to this. The first is to try to find approximations to the functional equation; the second is to elaborate on the class of cautious controllers. Murphy (1968), Tse and Bar Shalom (1973) and Tse, Bar Shalom and Meier (1973) have made contributions in the first area using a state space description of the system. The second approach, namely the development of cautious controllers, results from a realization that although cautious control fulfils the 'control' function of a dual controller, it lacks the 'probing' component. It is this deficiency that leads to the problem of turn-off. In order to incorporate an element of information sensing a perturbation (or probing) signal may be introduced if some function of the parameter covariances exceeds a set limit. This is suggested by Wieslander and Wittenmark (1971) who use an ARMA type model. Alster and Belanger (1974), also using an ARMA model, propose that turn off can be avoided by modifying the cost function so that minimization is constrained by, for example, the trace of the inverse parameter covariance matrix not

falling below a set limit. Hughes and Jacobs (1974) place a lower limit on the magnitude of the control input. These techniques help to improve performance. It is interesting that some of these ideas are now being re-considered for use in conjunction with self-tuning controllers.

Attention is now focussed on the development of self-tuning controllers themselves.

2.2.2 The Self Tuning Approach

In order to present and motivate the new material in this thesis it has been found useful to discuss existing self tuning control laws in some mathematical detail within the main body of the volume. The main aim here therefore is to provide a brief resume of self tuning control and its applications, avoiding for the present any detailed descriptions which appear later.

The self tuning approach belongs, as has already been stated, to the family of certainty equivalence controllers. Probably the first stochastic self tuning controller was that proposed by Peterka (1970). He used the ARMAX formulation

$$\{1+A(z^{-1})\}y_t = z^{-k}B(z^{-1})u_t + \{1+C(z^{-1})\}e_t$$

(see eqn. 2.1) to define the system and chose as his control objective the minimization of output variance, i.e. $\dot{E}(y_t^2)$. The off line solution to this regulation problem, given that the parameters of $A(z^{-1})$, $B(z^{-1})$ and $C(z^{-1})$ are available is well known (Astrom, 1970). If the parameters are not known a priori they must be estimated, and

the conventional approach is to either identify the system off-line using a non-linear estimator (e.g. maximum likelihood estimation, Astrom and Bohlin, 1965) or on-line using an approximate maximum likelihood estimator (e.g. RML). Peterka's approach differed in that he used a simple linear least squares prediction model of the form

$$\hat{y}_{t/t-k-1} = -z^{-k}\hat{A}(z^{-1})y_t + z^{-k}\hat{B}(z^{-1})u_t$$

$$y_t = \hat{y}_{t/t-k-1} + \varepsilon_t$$

where $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$ were estimated by recursive least squares so as to minimize the sum of the squares of the residual sequence ε_t. The parameters of the control law are obtained directly and are chosen so that the-k+1 - step-ahead prediction of output y_t ($\hat{y}_{t/t-k-1}$) is zero, i.e.

$$u_t = \frac{\hat{A}(z^{-1})}{\hat{B}(z^{-1})}\, y_t$$

The interesting property of the algorithm is that asymptotically the controller converges to the true minimum variance control law.† This neat result may be termed the 'self tuning property' of the control algorithm, and clearly provides a simpler method for obtaining minimum variance control than the conventional approach.

The first stochastic self tuner was therefore conceived as a regulator, minimizing the effect on the output of stochastic disturbances and regulating the output to zero. It is important to note also that the under-lying assumption is one of a time-invariant system. This is true of later self tuners too, and it is only by making approximations and introducing ad hoc variants of the basic algorithm that time varying systems can be controlled by self tuners.

† N.B. Convergence here and elsewhere throughout the thesis implies convergence with probability one

Servo self tuners may be configured by introducing feed-forward from the reference input.

Peterka's work was followed up by a detailed study of minimum variance self tuning by researchers at Lund University in Sweden, the first important paper being by Astrom and Wittenmark (1973). Since then the minimum variance regulator has been successfully used in the control of several processes. Applications include moisture content control on a paper machine (Cegrell and Hedquist, 1975; Borisson and Wittenmark, 1974), control of an ore crusher (Borisson and Syding, 1976), an enthalpy exchanger (Jensen and Hansel, 1974) and a supertanker (Kallstrom, Astrom, Thorell, Eriksson and Sten, 1978).

The minimum variance regulator has two important disadvantages:

a) In some applications the control action is excessive

b) It cannot control non-minimum phase systems.

This has motivated researchers to find alternative self tuning laws.

The simplest variant is the 'detuned' minimum variance regulator (Edmunds, 1976) which can to some extent overcome the first of these difficulties without significantly complicating the control algorithm. The effect is to alter the optimum closed loop system output from

$$y_t = \{1+M(z^{-1})\}e_t$$

(where $M(z^{-1})$ is a kth order polynomial) for the minimum variance configuration to

$$y_t = \frac{1+M(z^{-1})}{1+T(z^{-1})}\ e_t$$

where the introduction of 'tailoring' or 'detuning' polynomial $T(z^{-1})$

can significantly reduce the controller gain. The concept is analogous to the introduction of a staleness weighting factor in the design of minimal response controllers (Bertram, 1956). The method has been successfully employed in the control of a diesel engine (Wellstead and Zanker, 1978) and its extension to multivariable systems is discussed in this thesis.

The second limitation is a serious one. Many systems which are minimum phase in the s-domain are non-minimum phase in the z-domain, i.e. the zeros of $B(z^{-1})$ lie outside the z-plane unit circle. The minimum variance method effectively uses controller poles to cancel system zeros and thus when the system is non-minimum phase the closed loop system will in practice be unstable. Furthermore, $B(z^{-1})$ can become non-minimum phase by the introduction of the computational delay in calculating the next control input. Astrom and Wittenmark (1974) have proposed a sub-optimal self tuning minimum variance law. This requires on-line polynomial factorization and solution of a set of linear simultaneous equations.

Clarke, Gawthrop and their co-workers at Oxford University have generalized the minimum variance controller, first to the 'Lambda Controller' which minimized a cost function of both output and control variances (Clarke and Gawthrop, 1975; Clarke, Cope and Gawthrop, 1975) and then a more generalized cost function which can lead to a configuration in which both closed loop poles and zeros are specified (Gawthrop, 1977). The controller is more powerful than standard model reference adaptive controls in that the structure of the stochastic system disturbances appears explicitly in the formulation. If the cost function parameters are carefully chosen, non-minimum phase systems can

be controlled. The incorporation of set-point following (i.e. the servo rather than just the regulation problem) is implicit in the architecture.

Amongst the approaches that avoid the limitations imposed by non-minimum phase systems is that of a pole-placement self tuner first suggested by Edmunds (1976) and developed here for both the SISO and multivariable cases. The approach is distinct from the other self-tuners which are based on optimality criteria and has its roots in classical control theory where the direct relationship between pole position and system transient response is often used.

The controllers discussed thus far apply only to SISO systems. Borisson (1975) has developed a multivariable version of the minimum variance self tuner which retains the simplicity of its SISO counterpart. However, it restricts the number of outputs to be the same as the number of inputs. A variant of the multivariable minimum variance self tuner which not only minimizes the error variance of the output signal but keeps a weighted average for a finite interval as close to the reference value as possible has been used by Keviczky, Hetthessy, Hilger and Kolostori (1978) for a cement material blending process.

There have also been suggestions for incorporating compensators such as those designed using Rosenbrock's (1969) Inverse Nyquist Array Method to obtain minimum interaction minimum variance controllers. Sinha (1977) suggests such an approach in which the diagonal dominance of the system is used to justify employing a set of single input single output minimum variance controllers, one for each loop. The idea requires some further development, particularly with regard to the

real time incorporation of the decoupling compensator design.

No multivariable version of the generalized minimum variance controller appears in the literature as yet. A multivariable pole-placement self tuning scheme is discussed in this thesis.

Studies on the stability and convergence of minimum variance regulators have been carried out. It can be shown that if the time delay k of the process is known and if the order of the system (assumed to be minimum phase) is not under-estimated, the minimum variance self tuning regulator will stabilize any linear time-invariant system (Ljung and Wittenmark, 1976). Furthermore, if the regulator converges, it must converge to the true minimum variance regulator (Astrom, Borisson, Ljung and Wittenmark 1977; Ljung and Wittenmark, 1974). The convergence analysis associates the algorithm with a differential equation that contains all the necessary information about the asymptotic behaviour, and is based on a general method of Ljung's (1977) for a variety of stochastic recursive algorithms (also estimators).

More recently, Gawthrop (1978) adapted a technique used to analyse a model reference identifier to give convergence results for the generalized minimum variance self tuning controller. First the self tuner is written as a feedback system driven by a moving average process. It is then shown to be stable for systems with unit sample time delay or white noise models. Finally, Gawthrop shows that if the conditions for input-output stability hold, then the mean square of his chosen system scalar output (the difference between the desired and self tuning prediction errors) converges to zero.

In conclusion it must be pointed out that the self tuning concept may be used to construct a Self-Tuning Predictor (Wittenmark, 1974). The relationship between prediction and control is clearly a very close one, so that this result is not unexpected.

CHAPTER 3

MAXIMUM LIKELIHOOD ESTIMATION

3.1 Introduction

The first part of this thesis is concerned with the estimation of the parameters of the ARMAX model. Possibly the most attractive theoretical solution to this estimation problem is to be found in the Maximum Likelihood Method (ML). In this chapter, the concept of Likelihood, the form of the likelihood function for the case of the ARMAX System with Gaussian disturbances, and the standard solution to the ML estimation problem are discussed. The interesting relationship between Maximum Likelihood and Information Theory is noted, and it is shown how the likelihood function forms part of more general criteria used for model structure as well as parameter estimation. Although the material is not new, its fundamental importance to later analysis justifies its inclusion.

3.2 A definition of likelihood

The general identification problem may be thought of as a combination of two exercises, firstly that of estimating the system structure, and secondly, the estimation of the parameters of the selected structure. It is assumed here, that the correct structure has already been determined, and that it remains to estimate the constant, but unknown vector of model parameters, θ. A set of system input-output records, denoted R is available.

On a purely intuitive basis, it is sensible to choose that parameter set which, according to the model, makes the data R the most "likely" outcome. This is the principle of maximum likelihood.

First however, a measure of likelihood must be found. Let $P(R/\theta)$ be the probability of obtaining data R given the hypothesized parameter set θ. Then the likelihood $L(\theta/R)$ of the hypothesis θ, given R and a specific model, may sensibly be defined to be proportional to $P(R/\theta)$, i.e.:

$$L(\theta/R) \propto P(R/\theta) \qquad 3.1$$

(Edwards 1972) The probability $P(R/\theta)$ is defined for every member of the set of possible results R, given any one hypothesis θ. Although it is a function of both R and θ, it will generally be considered to be a function of R alone, given some specified parameter set θ. The statistical properties of R are then well known.

The likelihood, $L(\theta/R)$, however, is a function of θ, considered to be the variable, and R which is considered to be constant. The Maximum Likelihood estimate is that value of θ for which the likelihood function $L(\theta/R)$ is a maximum. (Van Trees 1968; Edwards 1972). Expressed another way, it is the value of θ that lends most credence to the proposition that the observed data could be generated by the selected model.

If two independent sets of results, R_1 and R_2 are available, then by definition

$$\begin{aligned} L(\theta/R_1,R_2) &\propto P(R_1,R_2/\theta) \\ &= P(R_1/\theta)P(R_2/\theta) \end{aligned}$$

so that

$$L(\theta/R_1,R_2) = L(\theta/R_1)L(\theta/R_2) \qquad 3.2$$

and clearly, for n independent sets of results, the likelihood of θ given the combined data is simply defined by the product of the

individual likelihoods $L(\Theta/R_i)$, $i = 1,2,\ldots n$. The definition of likelihood holds equally well for continuous or discrete (quantised) observed data.

In comparing the likelihoods of two hypotheses, one must of course ensure that the comparison is made on the basis of the same data. It is as well to re-inforce the idea that likelihood is a function of the hypothesized parameters Θ and note that the likelihood may be plotted out against each value of the parameter set (which may be continuous). This graphed function is known as the Likelihood Function. In the following section, the function is derived for the case when the system under study is described by the ARMAX equation

$$(I + A(z^{-1}))y_t = z^{-k} B(z^{-1})u_t + (I + C(z^{-1}))e_t \qquad 3.3$$

where u_t is a r-vector of inputs, y_t is a p-vector of outputs, and e_t is a p-dimensional zero mean Gaussian white noise process with

$$E(e_{t+i}e_{t+j}^T) = Q\delta_{ij} \qquad Q \geqslant 0 \qquad 3.4$$

and $A(z^{-1})$, $B(z^{-1})$, $C(z^{-1})$ are matrix polynomials in the backward shift operator, z^{-1}, of the form

$$X(z^{-1}) = X_1 z^{-1} + X_2 z^{-2} + \ldots + X_{n_x} z^{-n_x} \qquad 3.5$$

where X_i, $i = 1,\ldots,n_x$

are matrix coefficients.

3.3 The Likelihood Function in the presence of Gaussian Disturbances

It is assumed that the system described above in eqns 3.3 - 3.5 is modelled as

$$(I + \hat{A}(z^{-1}))y_t = z^{-k} \hat{B}(z^{-1})u_t + (I + \hat{C}(z^{-1}))\varepsilon_t \qquad 3.6$$

where the matrix polynomials have the same order as their counterparts in eqn. 3.3. Let Y_N denote the vector string of outputs to date, i.e.

$$Y_N = (y_N^T, y_N^T, \ldots\ldots, y_1^T)^T \qquad 3.7$$

and correspondingly, for the inputs

$$U_N = (u_{N-k-1}^T, u_{N-k-2}^T, \ldots\ldots, u_1^T\ u_1^T) \qquad 3.8$$

Collect the parameters of $\hat{A}(z^{-1})$, $\hat{B}(z^{-1})$, and $\hat{C}(z^{-1})$ into vector $\hat{\theta}$. Further, let $p(Y_N/U_N,\hat{\theta})$ be the probability density function (pdf) of the outputs Y_N, given the past history of inputs and parameter estimates. Now, it is well-known that the pdf of a Gaussian random p-vector X with mean M and covariance Σ is given by

$$p(X) = ((2\pi)^{p/2}|\Sigma|^{\frac{1}{2}})^{-1}\exp(-\tfrac{1}{2}(x-M)^T\Sigma^{-1}(x-M)) \qquad 3.9$$

(Kendal + Stuart 1961) Thus, under the hypothesis that $\hat{\theta}$ is indeed the correct parameter set, $\{\varepsilon_t\}$ is clearly a sequence of independent Gaussian random numbers, and

$$\begin{aligned} p(Y_N/U_N,\hat{\theta}) &= \prod_{i=1}^{N} ((2\Pi)^{p/2}\ |Q|^{\frac{1}{2}})^{-1}\ \exp\ (-\tfrac{1}{2}(\varepsilon_i^T\ Q^{-1}\ \varepsilon_i) \\ &= ((2\Pi)^{p/2}\ |Q|^{\frac{1}{2}})^{-N}\ \exp\ \{-\tfrac{1}{2}\ (\sum_{i=1}^{N}\ \varepsilon_i^T\ Q^{-1}\ \varepsilon_i)\} \\ &= ((2\Pi)^{p/2}\ |Q|^{\frac{1}{2}})^{-N}\ \exp\ \{-\tfrac{1}{2}\ (\sum_{i=1}^{N}\ \mathrm{tr}\ \varepsilon_i\varepsilon_i^T\ Q^{-1})\} \end{aligned} \qquad 3.10$$

which by the definition of likelihood must be proportional to $L(\hat{\theta}/Y_N,U_N)$.

In order to change the multiplicative properties to additive ones, the logarithm of the likelihood function is frequently used. Setting the constant of proportionality between $L(\hat{\theta}/Y_N,U_N)$ and $p(Y_N/U_N,\hat{\theta})$ to unity, it is evident from eqn. 3.10 that

$$\ln L(\hat{\theta}/Y_N,U_N) = \frac{-N}{2}(p \ln 2\pi + \ln |Q| + \text{tr } JQ^{-1}) \qquad 3.11$$

where J is the sample convariance matrix

$$J = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i\varepsilon_i^T \qquad 3.12$$

and the residuals $\{\varepsilon_i\}$, are a function of the parameter set $\hat{\theta}$.

It is often useful to express the log-likelihood as a function of the parameters $\hat{\theta}$, maximized with respect to covariance Q. Given that:

$$\frac{\partial}{\partial\theta_i} \log |Q| = \text{tr}\left(\frac{\partial Q}{\partial\theta_i} Q^{-1}\right)$$

$$\frac{\partial}{\partial\theta_i} Q^{-1} = -Q^{-1}\frac{\partial Q}{\partial\theta_i} Q^{-1}$$

(Akaike 1973)

where Q is a non-singular matrix and is a function of a set of parameters θ_i, it is easy to show that this maximum occurs for

$$Q = J \qquad 3.13$$

and thus

$$\max_Q \ln L(\hat{\theta}/Y_N,U_n) = \frac{-N}{2}(p \ln 2\pi + \ln |J| + p) \qquad 3.14$$

It is also evident that in order to maximize the likelihood (or log likelihood) function, it is necessary to minimize

$$V = |J| = \left|\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i\varepsilon_i^T\right| \qquad 3.15$$

J, and thus V has a particularly simple form in the case of SISO systems, and the likelihood is maximized by minimizing the sum of squares of the residual sequence ε_i, $i=1,2, \ldots, N$.

3.4 Information Theory and Maximum Likelihood

It is re-assuring to note that the maximum likelihood principle does not stand alone as an isolated theory, but in fact relates very strongly to criteria originating from Information Theory. To illustrate this, consider again the problem of obtaining the estimate $\hat{\theta}$ of a vector of parameters θ of a probability density function $p(R/\theta)$, where R is an observed random variable. The maximum likelihood estimate is that value of $\hat{\theta}$ for which the log likelihood function $\ln p(R/\hat{\theta})$ is a maximum. Akaike (1972; 1974a; 1974b) has suggested the maximization of the expected log likelihood function which is by definition

$$E_{\hat{\theta}} \int_{-\infty}^{\infty} p(R/\theta) \ln p(R/\hat{\theta}) \, dR \qquad 3.16$$

This would appear to be a formal extension of the classical Maximum Likelihood Principle, but the maximization of the above expression is clearly also equivalent to minimizing the information theoretic quantity

$$E \ln \left(\frac{p(R/\theta)}{p(R/\hat{\theta})} \right) = E_{\hat{\theta}} \int_{-\infty}^{\infty} p(R/\theta) \ln \left(\frac{p(R/\theta)}{p(R/\hat{\theta})} \right) dR \qquad 3.17$$

where the integral in the right hand side of eqn. 3.17 is the Kullback-Leibler mean information for discrimination between $p(R/\theta)$ and $p(R/\hat{\theta})$ defined as

$$I(\theta;\hat{\theta}) = \int_{-\infty}^{\infty} p(R/\theta) \ln \left(\frac{p(R/\theta)}{p(R/\hat{\theta})} \right) dR \qquad 3.18$$

(Kullback 1959). This gives a measure of separation or distance between the two distributions and has a positive value unless $p(R/\theta) = p(R/\hat{\theta})$ holds almost everywhere, in which case $I(\theta;\hat{\theta})$ is equal to zero. There is therefore a clear link between the maximum

likelihood principle and information theoretic concepts. Furthermore, notice that if N independent realizations R_i, i=1,2,...,N of R are available, (-1) times the sample mean of the log likelihood ratio

$$\frac{1}{N} \sum_{i=1}^{N} \ln \left(\frac{p(R_i/\hat{\theta})}{p(R_i/\theta)} \right) \qquad 3.19$$

will be a consistent estimate of $I(\theta;\hat{\theta})$. Now eqn. 3.19 is maximized by the maximum likelihood estimate. Thus the maximum likelihood estimator is designed to minimize the mean information for discrimination between the true and estimated distributions, $I(\theta;\hat{\theta})$.

It is useful at this point to also introduce the 'Fisher Information Matrix'. It can be shown (Kullback 1959) that when $\hat{\theta}$ and θ are sufficiently close, $I(\theta;\hat{\theta})$ may be approximated by

$$I(\theta;\theta+\Delta\theta) = \tfrac{1}{2} \| \Delta\theta \|^2_{F(R/\theta)} \qquad 3.20$$

where $\Delta\theta = \hat{\theta} - \theta$

$$\| \Delta\theta \|^2_F = \Delta\theta^T F \Delta\theta$$

and F is the Fisher Information Matrix (Van Trees 1967) given by

$$\begin{aligned} F(R/\theta) &\triangleq \{F_{ij}(\theta)\} \\ &= \int_{-\infty}^{\infty} \{ \frac{\partial}{\partial\theta} \log p(R/\theta) \}\{ \frac{\partial}{\partial\theta} \log p(R/\theta) \}^T p(R/\theta)\, dR \\ &= E \left(\{ \frac{\partial}{\partial\theta} \log p(R/\theta) \}\{ \frac{\partial}{\partial\theta} \log p(R/\theta) \}^T \right) \\ &= -E \left(\frac{\partial^2 \log p(R/\theta)}{\partial\theta^2} \right) \end{aligned} \qquad 3.21$$

The importance of the Fisher Information Matrix is that, via the Cramer Rao Bound (Van Trees 1967) its inverse provides a lower bound to the covariance matrix of any unbiased estimate $\hat{\theta}$ of θ, viz

$$E\{(\theta-\hat{\theta})(\theta-\hat{\theta})^T\} \geq F^{-1}(R/\theta) \qquad 3.22$$

On its own, the maximum likelihood principle cannot be used for model order identification, or the selection of the correct system structure. The likelihood function will always achieve its maximum for the model with the largest number of parameters. However, information theoretic criteria (reviewed in Prager 1976) do exist which can be used to select system structure. It is particularly interesting to see that the likelihood function still plays an important role in these criteria. Akaike's Information Criterion for example, which is derived from the cost function in eqn. 3.16, is given by:

$$AIC = -2 \sum_{i=1}^{N} \log p(R_i/{}_k\hat{\theta}) + 2k \qquad 3.23$$

where ${}_k\hat{\theta}$ is a kth order parameter vector. The 'best' approximating model is that which minimizes AIC. For a given k (i.e. k parameters), the criterion is minimized by the Maximum Likelihood estimate.

Rissanen and Ljung (1976), have developed an information theoretic criterion that measures the fit between a model and observed data as well as determining the system structure. The criterion is an advance on AIC which clearly cannot distinguish between system theoretically equivalent structures with the same number of parameters. Rissanen and Ljung's criterion, termed the "Fundamental Estimation Criterion" requires the minimization with respect to both structure (s) and parameter vector θ of

$$V_N(s,\hat{\theta}) = \ln |R_s(\hat{\theta})| + \frac{1}{N} \ln |P_s(\hat{\theta})| + \frac{k}{N}(1 + \ln 2\pi) \qquad 3.24$$

where $$R_s(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \varepsilon_i^T$$

ε_t is defined by eqn. 3.3

and $P_s(\hat{\theta})$ is an estimate of the covariance matrix of $\hat{\theta}$ defined by

$$P_s(\hat{\theta}) \simeq \frac{4}{N^2} \left(\left\{\frac{\partial}{\partial \hat{\theta}} \ln |R_s(\hat{\theta})|\right\}\left\{\frac{\partial}{\partial \hat{\theta}} \ln |R_s(\hat{\theta})|\right\}^T\right)^{-1} \qquad 3.25$$

Note that the right hand side of eqn. 3.25 is the inverse of the Information Matrix of eqn. 3.21. For a given structure, s, this criterion too is minimized by the maximum likelihood estimate.

3.5 Properties of Maximum Likelihood Estimates

The three most fundamental properties of any estimate which can be used to assess its quality are bias, consistency and efficiency.

An estimate is said to be unbiased if the relationship

$$E(\hat{\theta}) = \theta \qquad 3.26$$

holds. Thus, the average value of an unbiased estimate of a parameter is the parameter itself. If $\hat{\theta}$ converges in probability to θ the estimate is said to be consistent. Clearly, consistent estimates are asymptotically also unbiased. The third property concerns the covariance of the estimation error. An estimate is efficient if the covariance matrix equals the Cramer Rao lower bound, a lower bound on the covariance of any unbiased estimate.

Astrom + Bohlin (1965) have shown that for the SISO ARMAX model (eqn. 2.2) the Maximum Likelihood estimator is asymptotically both consistent and efficient. The condition† on the system input sequence is that the following limits exist for all finite t:

† See the original Astrom and Bohlin paper for the full conditions.

a) $$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} u_t \qquad 3.27$$

b) $$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} u_t u_{t+\tau}^T \qquad 3.28$$

Astrom + Bohlin establish (for SISO systems) that if the zeros of $1+A(z^{-1})$ and $1+C(z^{-1})$ lie within the z-plane unit circle, $\frac{1}{N}\ln L(\theta/Y_N,U_N)$ converges to its ensemble average, which is a differentiable function in the parameters θ. They also show that the estimates $\hat{\theta}$ converge into the set S_o of parameters, θ, which are equivalent to θ_o (the true parameters) in the sense that any model with $\theta \epsilon S_o$ will generate outputs y_t having the same likelihood function. If S_o contains only one point, the estimates are strongly consistent and converge to θ_o.

The question of uniqueness has been studied by Astrom + Soderstrom (1974). Their work shows that if the model is adequately parametrized, and if the orders of either the autoregressive (AR) or moving average (MA) parts of the model equation are correct, the likelihood has a unique local and global maximum. If the orders of the AR and MA parts are both too high, there will be many maxima, with the property that the AR and MA polynomials will have common factors. There may be several local maxima if the model is under parametrized.

Maximum likelihood estimates have particularly good statistical properties. Following from Astrom and Bohlin (1965) it is known that:

1) The variable $L_{\theta\theta}(\theta) \triangleq \lim_{N\to\infty} \frac{1}{N} E \frac{\partial^2 \log L(\theta)}{\partial\theta^2}$ exists

and

$$\lim_{N\to\infty} \frac{1}{N} \frac{\partial^2 \log L(\theta)}{\partial\theta^2} = \lim_{N\to\infty} \frac{1}{N} E \frac{\partial^2 \log L(\theta)}{\partial\theta^2}$$

2) $L_{\theta\theta}(\dot{\theta}) \sqrt{N} (\hat{\theta}_N - \dot{\theta})$ (where $\dot{\theta}$ is the true parameter vector) is asymptotically distributed $N(0, -L_{\theta\theta}(\dot{\theta}))$.

3) If $L_{\theta\theta}(\dot{\theta})$ is non-singular, then $\hat{\theta}_N$ is asymptotically distributed $N(\dot{\theta}, -\frac{1}{N} L^{-1}_{\theta\theta}(\dot{\theta}))$.

Thus the estimate $\hat{\theta}_N$ is asymptotically normally distributed with mean $\dot{\theta}$ and covariance $-\frac{1}{N}L^{-1}_{\theta\theta}(\dot{\theta})$. This latter quantity will be recognized as the Cramer Rao lower bound (see eqn. 3.21) on the variance of any estimator, and by definition then, the Maximum Likelihood estimate is efficient.

The consistency property is more complex in the case of multivariable systems. Caines and Rissanen (1974) discuss the problem and consider the representation

$$y_t = \phi_o e_t + \phi_1 e_{t-1} + \ldots \qquad 3.29$$

where ϕ_o is invertible, and y_t is now a full rank, p-component stationary process. In addition, y_t is finitely generated in the sense that the relation

$$\{I+A(z^{-1})\}y_t = \{C_o+C(z^{-1})\}e_t \qquad 3.30$$

holds, where the matrix polynomials $A(z^{-1})$ and $C(z^{-1})$ are of order n and follow their previous definition. C_o is a matrix of constants. Now, there are in general several pairs $\{I+A(z^{-1}), C_o+C(z^{-1})\}$ of matrix polynomials that generate y_t from the orthonormal process e_t. They are all those such that

$$\{I+A(z^{-1})\}^{-1}\{C_o+C(z^{-1})\} = \Phi(z^{-1}) = \phi_o+\phi_1 z^{-1}+\ldots \qquad 3.31$$

It is clear that the ML estimates of the parameters in $\{A(z^{-1}), C_o+C(z^{-1})\}$ cannot be expected to be consistent without a prior knowledge of the Kronecker indices, itself a complex problem.

However, Caines and Rissanen show that if $\{\hat{A}(z^{-1}),\ \hat{C}_o+\hat{C}(z^{-1})\}$ maximize the likelihood function, then $(I+\hat{A}(z^{-1}))^{-1}\ (\hat{C}_o+\hat{C}(z^{-1}))\rightarrow\Phi(z^{-1})$.

Caines and Ljung (1976) have shown that Astrom and Bohlin's (1965) SISO results of asymptotic normality and efficiency of ML estimates (for Gaussian processes) are valid in the multivariable case. Their results are part of a broader study considering asymptotic normality and accuracy of prediction error estimates, and will be considered in more detail later.

3.6 Computation of Maximum Likelihood Estimates

In Chapter 2, it was pointed out that maximum likelihood estimation is a non-linear procedure, and for this reason is numerically far more complex than standard least squares. Two solutions to the problem are discussed here. The first is a conventional hill climbing technique; the second is an approximate, recursive approach. Both techniques are applicable to SISO systems.

3.6.1 ML estimation by hill climbing

The conventional approach to maximising the likelihood function (for SISO systems) is due to Astrom and Bohlin (1965) and involves the use of the well-known Newton Raphson hill climbing algorithm

$$\theta_{k+1} = \theta_k - \left(\frac{\partial^2 V(\theta_k)}{\partial\theta^2}\right)^{-1} \frac{\partial V(\theta_k)}{\partial\theta} \qquad 3.32$$

where k denotes the iteration number,

and

$$V(\theta_k) = \tfrac{1}{2}\sum_{i=1}^{N}\varepsilon_i^2 \qquad 3.33$$

which for convenience, differs by a factor ($\frac{1}{2}N$) from the cost function given in eqn. 3.15. The computation of the partial derivatives of the cost function is fairly straightforward, and is detailed below. Differentiating $V(\theta)$ with respect to the i^{th} component of θ yields:

$$\frac{\partial V}{\partial \theta_i} = \sum_{t=1}^{N} \varepsilon_t \frac{\partial \varepsilon_t}{\partial \theta_i} \qquad 3.34$$

$$\frac{\partial^2 V}{\partial \theta_i \partial \theta_j} = \sum_{t=1}^{N} \frac{\partial \varepsilon_t}{\partial \theta_i} \frac{\partial \varepsilon_t}{\partial \theta_j} + \sum_{t=1}^{N} \varepsilon_t \frac{\partial^2 \varepsilon_t}{\partial \theta_i \partial \theta_j} \qquad 3.35$$

It is convenient at this stage to introduce the filtered variables y_{f_t}, u_{f_t}, ε_{f_t}, and the "doubly filtered" variables y_{ff_t}, u_{ff_t} and ε_{ff_t}, where these filtered variables are derived according to:

$$x_{f_t} = x_t / \{1+\hat{C}_k(z^{-1})\} \qquad 3.36$$

$$x_{ff_t} = x_{f_t} / \{1+\hat{C}_k(z^{-1})\} \qquad 3.37$$

and $\hat{C}_k(z^{-1})$ implies the estimate of $C(z^{-1})$ at iteration k. Then the SISO model equation (from eqn. 3.3) can be re-written as

$$\varepsilon_t = \{1+\hat{A}(z^{-1})\}y_{f_t} - z^{-k}\hat{B}(z^{-1})u_{f_t} \qquad 3.38$$

from which the partial derivatives may be deduced as follows:

$$\begin{aligned}
\frac{\partial \varepsilon_t}{\partial a_i} &= z^{-i} y_{f_t} \\
\frac{\partial \varepsilon_t}{\partial b_i} &= -z^{-i-k} u_{f_t} \\
\frac{\partial \varepsilon_t}{\partial c_i} &= -z^{-i} \varepsilon_{f_t} \\
\frac{\partial^2 \varepsilon_t}{\partial a_i \partial c_j} &= -z^{-j-i} y_{ff_t} \\
\frac{\partial^2 \varepsilon_t}{\partial b_i \partial c_j} &= z^{-j-i-k} u_{ff_t} \\
\frac{\partial^2 \varepsilon_t}{\partial c_i \partial c_j} &= 2z^{-j-i} \varepsilon_{ff_t}
\end{aligned} \qquad 3.39$$

Thus, if the parameter vector $\hat{\theta}$ is

$$\hat{\theta} = (\hat{a}_1,\ldots,\hat{a}_{n_a},\hat{b}_1,\ldots,\hat{b}_{n_b},\hat{c}_1\ldots,\hat{c}_{n_c})^T \qquad 3.40$$

$$Y_{f_{t-i}} = (y_{f_{t-i}}, y_{f_{t-i-1}},\ldots\ldots, y_{f_{1-i}})^T \qquad 3.41$$

and similar definitions hold for U_{f_t} (the vector of inputs) and E_{f_t} (the vector of residuals)

$$\frac{\partial V}{\partial \theta} = \begin{bmatrix} Y^T_{f_{t-1}} \\ \vdots \\ Y^T_{f_{t-n_a}} \\ -U^T_{f_{t-k-1}} \\ \vdots \\ -U^T_{f_{t-k-n_b}} \\ -E^T_{f_{t-1}} \\ \vdots \\ -E^T_{f_{t-n_c}} \end{bmatrix} E_t = X^T E_t \qquad 3.42$$

Gustavsson (1969) indicates that the second term in the expansion of $\frac{\partial^2 V}{\partial \theta_i \partial \theta_j}$ has little influence on the minimization of the cost function, and if it is ignored, the cost function may be approximated as follows:

$$\frac{\partial^2 V}{\partial \theta^2} \simeq X^T X \qquad 3.43$$

The evaluation of the partial derivatives is therefore neatly given by equations 3.42 and 3.43. The algorithm then consists of choosing some initial estimate of θ, evaluating eqn. 3.42 and 3.43 to obtain the partial derivatives, and iterating on eqn. 3.32. Clearly, any other hill-climbing method, for example, that of Rosenbrock (1960), may equally well be used.

The greatest disadvantage of this technique is its substantial computational requirement. Furthermore, as the method relies solely on "number crunching" there is negligible user interaction and the user gains only a minimal "feel" for the problem in hand.

Accuracy is the prime advantage of the method. The excellent properties of the cost function, especially the uniqueness of the global and local maximum, ensure that if the correct model structure is chosen, and if there are sufficient observations the hill-climber will eventually converge to the true ML parameter estimates.

In the multivariable case, where a determinantal cost function is to be minimized, the computation of the partial derivatives becomes more complicated.Kashyap (1970) has suggested a method by which Lagrange multipliers are used to incorporate the system equations into the loss function and Akaike (1973) discusses the numerical maximization of the likelihood function for multivariable systems.

3.6.2. Recursive Maximum Likelihood

The computational complexity of the hill-climbing approach has motivated some considerable interest in approximate, recursive, real-time algorithms for ML estimation. Many of these have already been discussed in Chapter 2. As was mentioned there, the most prominent technique is Soderstrom's (1973) RML algorithm, which is now discussed in some detail. The cost function given in eqn. 3.33 is again used, and the vector of parameter estimates after N observations (t=1,2,...,N) is denoted $\hat{\theta}_N$.

If $\hat{\theta}_N$ minimizes the cost V_N obtained after N observations, then

$\hat{\theta}_{N+1}$ can be computed from a Taylor expansion of $V_{N+1}(\theta)$ about $\hat{\theta}_N$. A second order expansion is deemed sufficient, viz:

$$V_{N+1}(\hat{\theta}) \simeq V_{N+1}(\hat{\theta}_N)+V'_{N+1}(\hat{\theta}_N)(\hat{\theta}-\hat{\theta}_N)+\tfrac{1}{2}(\hat{\theta}-\hat{\theta}_N)^T V''_{N+1}(\hat{\theta}_N)(\hat{\theta}-\hat{\theta}_N) \qquad 3.44$$

where V' and V'' denote the first and second partial derivatives of V with respect to θ, respectively. $V_{N+1}(\hat{\theta})$ is minimized when

$$\hat{\theta}_{N+1} = \hat{\theta}_N - V''^{-1}_{N+1}(\hat{\theta}_N)\, V'_{N+1}(\hat{\theta}_N) \qquad 3.45$$

If it is assumed that $\hat{\theta}_N$ is the maximum likelihood estimate at stage N then $V'_N(\hat{\theta}_N) = 0$

and

$$V'_{N+1}(\hat{\theta}_N) = \varepsilon(N+1,\hat{\theta}_N)\varepsilon'(N+1,\hat{\theta}_N) \qquad 3.46$$

(see eqn. 3.34). In the evaluation of $V''_{N+1}(\hat{\theta}_N)$ the following assumptions are made:

1) The term $\sum_{i=i}^{N+1} \varepsilon(t,\hat{\theta}_N)\,\varepsilon''(t,\hat{\theta}_N)$ in the expression for $V''_{N+1}(\hat{\theta}_N)$ (see eqn. 3.35) has neglible effect on the minimization. (This is found in the case of off-line minimization by Gustavsson (1969)).

2) $V''_N(\hat{\theta}_N) \simeq V''_N(\hat{\theta}_{N-1})$

Then:

$$V''_{N+1}(\hat{\theta}_N) \simeq V''_N(\hat{\theta}_{N-1}) + \varepsilon'(N+1,\hat{\theta}_N)\varepsilon'(N+1,\hat{\theta}_N)^T \qquad 3.47$$

Now define $P_N = V''_N(\hat{\theta}_{N-1})^{-1}$ 3.48

and $x_t^T = (-y_{f_{t-1}},\ldots,-y_{f_{t-n_a}},u_{f_{t-k-1}},\ldots,u_{f_{t-k-n_b}},\varepsilon_{f_{t-1}},\ldots,\varepsilon_{f_{t-n_c}})$

$= -\varepsilon'(N,\theta_{N-1})^T$ (from eqn. 3.39) 3.49

where the filter variable notation (x_{f_t}) is defined in eqn. 3.36. Then, substituting eqns. 3.48 and 3.49 into eqn. 3.46 and 3.47, and then substituting eqns. 3.46 and 3.47 into eqn. 3.45, and expressing ε_t in terms of y_N, x_N and $\hat{\theta}_N$, eqn. 3.45 becomes

$$\hat{\theta}_{N+1} = \hat{\theta}_N + P_{N+1}x_{N+1}(y_{N+1} - x_{N+1}^T\hat{\theta}_N) \qquad 3.50$$

From the matrix inversion lemma (and the familiar RLS formulation) it is easy to show that P_N can be computed recursively as follows:

$$P_{N+1} = P_N - \gamma_N P_N x_{N+1} x_{N+1}^T P_N \qquad 3.51$$

where $\gamma_N = (1+x_{N+1}^T P_N x_{N+1})^{-1}$

Exact computation of the prediction error $\varepsilon(t,\hat{\theta}_N)$ requires the solution of eqn. 3.38 from t=0 for every new parameter vector $\hat{\theta}_N$, i.e. every new measurement. Thus, to speed up computation, it is generally approximated, for example by

$$\varepsilon_N = y_N - (-y_{N-1},\cdots,-y_{N-n_a},u_{N-k-1},\cdots,u_{N-k-n_b},\varepsilon_{N-1},\cdots,\varepsilon_{N-n_c})\hat{\theta}_N \qquad 3.52$$

Soderstrom, Ljung + Gustavsson (1978) show that using RML, the estimate $\hat{\theta}$ converges with probability one to a local minimum of V. It is evident that the recursive algorithm 3.50 - 3.51 may be interpreted as giving the recursive least squares solution to the set of equations

$$\begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix} \hat{\theta} = \begin{bmatrix} \varepsilon_1 + x_1^T\hat{\theta}_o \\ \vdots \\ \varepsilon_N + x_N^T\hat{\theta}_{N-1} \end{bmatrix} \qquad 3.53$$

Thus, the normal equation

$$\frac{1}{N}\sum_{t=1}^{N}(x_t x_t^T)\hat{\theta}_N = \frac{1}{N}\sum_{t=1}^{N} x_t(\varepsilon_t + x_t^T\hat{\theta}_{t-1}) \qquad 3.54$$

must hold. Now if the algorithm converges, and $\hat{\theta}_N$ and $\hat{\theta}_{t-1}$ are asymtotically replaced by θ^*, then eqn. 3.54 implies

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} x_t \varepsilon_t = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \varepsilon_t' \varepsilon_t = 0 \qquad 3.55$$

or, invoking ergodic theory, $E\ \varepsilon_t' \varepsilon_t = 0$ 3.56

which is precisely the condition necessary to minimize the cost

$$V_\infty = E(\varepsilon^2) \qquad 3.57$$

Thus, intuitively, the true parameter set is seen as a possible convergence point, if the usual conditions required for the uniqueness of the ML estimate are fulfilled.

This algorithm has superior convergence properties to RELS. Simulation studies, results of which are presented later, have however indicated that the method often requires a very large number of observations in order to obtain good reliable estimates. Furthermore, it has been found essential to check the stability of the inverse noise filter $(1+\hat{C}(z^{-1}))^{-1}$ at every iteration and to modify the estimates to ensure stability. Modification is usually necessary only at the start of a run.

Soderstrom suggests other devices to improve performance. In particular, it was observed that the algorithm rapidly lost gain, even before the $\hat{C}(z^{-1})$ polynomial estimates had approached their correct values. In order to assist convergence, one might monitor "convergence rate" and "restart" the algorithm mid-run (with a higher value of P_t) thus boosting the gain. Clearly, many criteria for assessing convergence can be found, and endless research time, of dubious value, can be expended in the search for further

"modifications".

One particularly useful device, however, is that of the forgetting factor as used also in RLS. It has been found that a forgetting factor which forgets itself can enhance results, and is a point made by Soderstrom, Ljung + Gustavsson (1978). The forgetting factor λ is updated according to

$$\lambda_{t+1} = \beta\lambda_t + (1-\beta) = 1 - \beta(1-\lambda_t)$$

where β is typically chosen to be 0.99.

Ultimately, it must be conceded that the algorithm does not perform nearly as well as the off-line hill-climbing ML method which can achieve good estimates with relatively few observations. Admittedly, it does so at some considerable computational cost.

3.7 Conclusion

This chapter has introduced the concept of likelihood and considered fundamental properties of maximum likelihood estimation with particular reference to the ARMAX model. Attention has been focussed on the interesting link between Information Theoretic criteria and Maximum likelihood. Finally, the review of the two most dominant approaches to the evaluation of maximum likelihood estimates has prepared the ground for the presentation of a new approach to ML estimation in the following chapter.

CHAPTER 4

MAXIMUM LIKELIHOOD ESTIMATION BY HYPOTHESIS TESTING

4.1 Introduction

This chapter is concerned with the Maximum Likelihood estimation of the parameters of the ARMAX model for SISO systems described in eqns. 2.1 - 2.2. Caines and Ljung (1976) have shown that the maximization of the likelihood function leads to asymptotically consistent and efficient estimates, and that the parameter estimates are asymptotically normally distributed. Details were given in Chapter 3. The attractiveness of the maximum likelihood method is lessened only by the property that the parameters enter into the maximization in a non-linear manner. As has been discussed previously, this greatly complicates their evaluation.

In Chapter 3 examples were given of the two established approaches to the computation of maximum likelihood estimates, viz.

1) Optimization techniques which employ numerical hill climbing methods (e.g. Astrom + Bohlin 1965)
2) Quasi-linear approximation methods which use variations of linear least squares (Panuska 1968, 1969; Soderstrom 1973; Gertler and Banfasz 1974; Young 1968, 1970; Young, Shellswell and Neethling 1971).

Direct hill-climbing methods have two major disadvantages. Firstly, they require considerable computational effort, and secondly, because the final result is simply a set of estimates and

covariances, little information about the nature of the likelihood function or deeper insight into the estimation problem is gained by the analyst.

On the other hand, the computationally more attractive, recursive, approximate maximum likelihood methods often experience convergence difficulties which are difficult to diagnose. (Soderstrom, Ljung + Gustavsson, 1975, 1978). Although the methods can work extremely well, especially when large data samples are available, they are not generally reliable for the off-line analysis of short data records. In spite of the apparent computational saving associated with a linear estimator, repeated iterations in the case of short data records are often involved with a concommitant increase in computational effort.

These shortcomings led to a fundamental re-appraisal of how best to approach the evaluation of maximum likelihood estimates. In particular, due consideration was taken of the advantages to be gained from increasing the analyst's interaction with the estimation algorithm via graphic information as opposed to purely numerical diagnostics supplied by existing methods. The natural realization of this concept is the visual presentation and interpretation of the likelihood function or related cost function, and marks a return to the original proposals of R.A. Fisher (Edwards 1972). Computationally, this involves the discretization of the parameter space, and the setting up and testing of multiple hypotheses in a manner familiar in communications theory (Van Trees 1968) and related to proposals by Nebeker (1976) and Tomanek (1976) discussed in Chapter 2. The implementation of the method, which is the subject of this chapter,

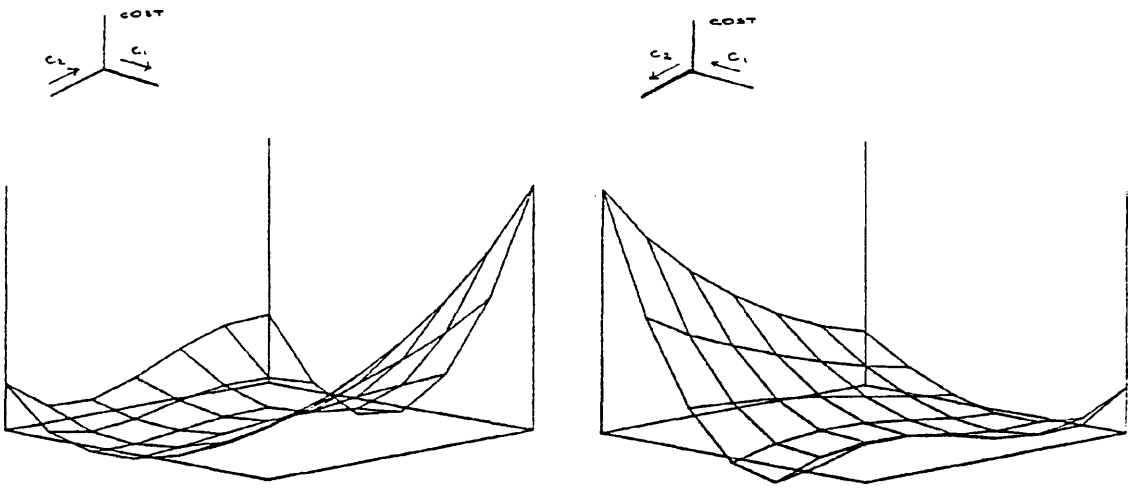
is achieved recursively, and this has led to the acronym RHYP (estimation by Recursive Hypothesis testing).

An initial objection to the method would be the problems of dimensionality in computing the likelihood function. However, referring to eqn. 2.2, it should be noted that only the parameters of $C(z^{-1})$ occur in the estimation in a non-linear manner, and hence only those cross-sections of the likelihood function which correspond to the co-efficients of $C(z^{-1})$ need to be examined, the $A(z^{-1})$and $B(z^{-1})$ polynomial parameters being constrained in the RHYP procedure.

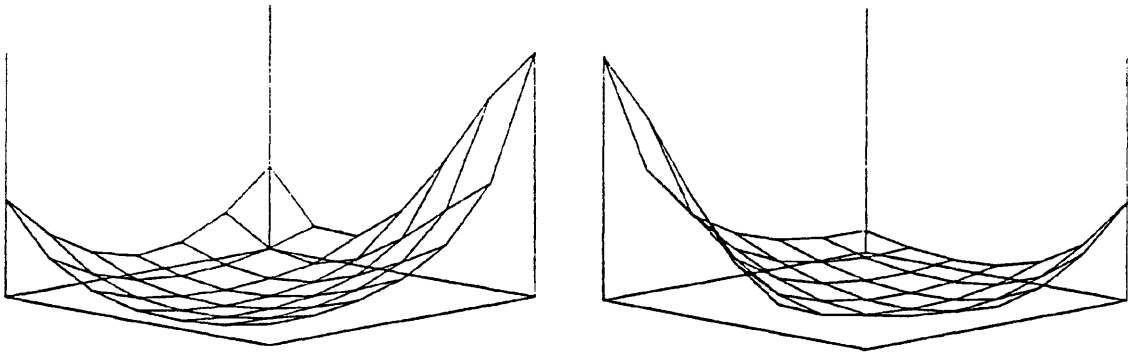
The difficulties of dimensionality are more than compensated for by the insight afforded the analyst by a physical inspection of the likelihood surface and its evolution with time. It is important to appreciate that, although the likelihood function is endowed with excellent asymptotic properties (Astrom + Soderstrom 1974; Caines + Ljung 1976) the sample properties for short data records can be very different, as it is likelihood conditioned on available data which is obtained. To illustrate this point, data was generated according to the moving average rule

$$y_t = (1 - 0.8z^{-1} + 0.6z^{-2})e_t \qquad 4.1$$

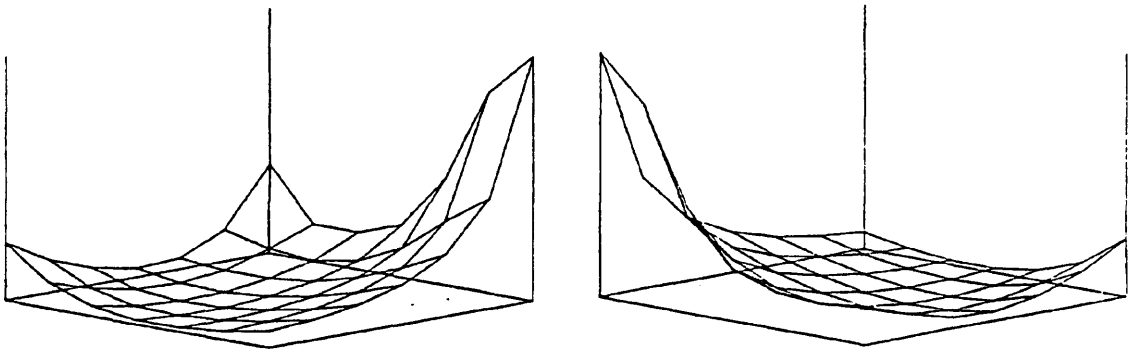
where e_t is a Gaussian white noise. Fig. 4.1 shows the time evolution of a section of the cost function V (given in eqn.3.15 and which is related to the likelihood function) as the amount of data is increased. The diagrams (produced using RHYP) clearly illustrate the conditioning of the likelihood surface in a manner which would not be possible by other methods.



After 10 observations



After 20 observations



After 30 observations

Fig. 4.1 Time evolution of section of cost function

This chapter describes the new algorithm, and suggests methods for reducing computation. Examples illustrating the use of the procedure, and highlighting its merits are given.

4.2 The Basic RHYP Algorithm

Consider the linear time invariant SISO system defined by the ARMAX equation

$$\left[1 + A(z^{-1})\right] y_t = z^{-k} B(z^{-1}) u_t + \left[1 + C(z^{-1})\right] e_t \qquad 4.2$$

where u_t and y_t are the system input and output respectively and e_t is a Gaussian white noise sequence with statistics

$$E(e_t) = 0 \qquad 4.3$$

$$E(e_t^2) = R \ , \ R > 0. \qquad 4.4$$

$A(z^{-1})$, $B(z^{-1})$ and $C(z^{-1})$ are polynomials in the backward shift operator z^{-1} of order n_a, n_b and n_c respectively, and follow previous definitions. Assume that N observations of the input/output data (y_1,u_1), (y_2,u_2), ..., (y_N,u_N) are available. Then the maximum likelihood estimates of $A(z^{-1})$, $B(z^{-1})$ and $C(z^{-1})$ are those values of $\hat{A}(z^{-1})$, $\hat{B}(z^{-1})$ and $\hat{C}(z^{-1})$ in the model

$$\left[1 + \hat{A}(z^{-1})\right] y_t = z^{-k} \hat{B}(z^{-1}) u_t + \left[1 + \hat{C}(z^{-1})\right] \varepsilon_t \qquad 4.5$$

which minimize the cost function

$$V_N = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i^2 \qquad 4.6$$

$\hat{A}(z^{-1})$, $\hat{B}(z^{-1})$ and $\hat{C}(z^{-1})$ are assumed to be of order n_a, n_b and n_c respectively.

Assume now, that the maximum likelihood estimate $\hat{C}_{ml}(z^{-1})$ of $C(z^{-1})$ were known a priori. The filtered variables y_{f_t} and u_{f_t} defined by

$$y_{f_t} = y_t / (1 + \hat{C}_{ml}(z^{-1})) \qquad 4.7$$

$$u_{f_t} = u_t / (1 + \hat{C}_{ml}(z^{-1})) \qquad 4.8$$

could be computed, and model eqn 4.5 reformulated as:

$$\left[1 + \hat{A}(z^{-1})\right] y_{f_t} = z^{-k}\hat{B}(z^{-1})\, u_{f_t} + \varepsilon_t \qquad 4.9$$

The evaluation of the maximum likelihood estimates $\hat{A}_{ml}(z^{-1})$ and $\hat{B}_{ml}(z^{-1})$ is then a very simple task indeed, and may be accomplished by applying linear least squares estimation to the minimization of V_N.

Unfortunately, $\hat{C}_{ml}(z^{-1})$ is not known a priori, but the above property may be exploited as follows.

A set of hypothesized $1+\hat{C}(z^{-1})$ polynomials is employed to filter the data, and linear least squares estimation used on each filtered data record to obtain estimates of $A(z^{-1})$ and $B(z^{-1})$. The cost function is also evaluated for each hypothesis. The approximate maximum likelihood solution is that which minimizes the cost. A more detailed exposition is now given.

Construct a set of "allowable polynomials"

$$C_o = \{(1 + \hat{C}_i (z^{-1})),\ i = 1,2,....,M\} \qquad 4.10$$

where the elements of C_o serve as hypothesized estimates of the $1+C(z^{-1})$ polynomial, and it is assumed that $1+\hat{C}_{m1}(z^{-1})$ is a member of C_o. An "allowable" polynomial is defined so that its zeros lie inside the z-plane unit disc (i.e. the polynomial is 'inverse stable'). The elements of C_o,

$$1+\hat{C}_i(z^{-1}) = 1+c_1^i z^{-1} + ... + c_{n_c}^i z^{-n_c}\ ,\ i=1,2,...,M$$

are generated by discretizing the parameter space of the $1+\hat{C}(z^{-1})$ polynomial according to Procedure A.

Procedure A

1) Select quantisation intervals q_j, $j = 1,2,...,n_c$ for each c_j, $j = 1,2,..,n_c$
2) Determine the values $\bar{c}_j$, $j = 1,2,...,n_c$ where $\bar{c}_j$ is the absolute value of the corresponding coefficient of $(1+z^{-1})^{n_c}$. If

$$c_j^+ = \bar{c}_j$$

and $c_j^- = -\bar{c}_j$

then c_j^+ and c_j^- determine the maximum boundary limits of c_j which could possibly allow the zeros of $1+\hat{C}_i(z^{-1})$ to lie within the z-plane unit disc. There are then a maximum of

$$K_j = \text{INT}\left[\frac{c_j^+ - c_j^-}{q_j}\right] - 1 \qquad 4.12$$

possible values of c_j, where

$$\text{INT}(x) \triangleq \text{integer part of } x \qquad 4.13$$

3) Generate the set of quantized coefficients S_{c_j}, $j = 1,2,\ldots,n_c$ where

$$S_{c_j} = \{c_j / \bar{c_j} < c_j < c_j^{+} \text{ and } c_j = \bar{c_j} + \ell q_j,\ \ell \in (1,2,\ldots,k_j)\} \qquad 4.14$$

4) Having thus quantised the coefficients c_j, construct all allowable polynomials $1+C_i(z^{-1})$, i=1,2,..,M, in C_o such that

$$c_j \in S_{c_j},\ j = 1,2,\ldots,n_c$$

This procedure quantises the parameters of the $1+\hat{C}(z^{-1})$ polynomial and sets up the set of hypothesized polynomials, C_o. As $q_j \to o, j=1,2,..n_c$, so set C_o expands and the assumption that $(1+\hat{C}_{ml}(z^{-1})) \in C_o$ gains validity.

As a simple example of the quantization procedure, consider the first order polynomial $1+C(z^{-1}) = 1+c_1 z^{-1}$. Select $q_1=0.2$.

Now $\bar{c}_1=1$, and hence $c_1^{+}=1$, $c_1^{-}=-1$, and $k_1=9$. The quantised coefficient set $S_{c_1} = \{-0.8,-0.6,\ldots,0.6,0.8\}$ and the M=9 elements of the allowable polynomial set are

$$C_o = \{(1-0.8z^{-1}),\ (1-0.6z^{-1}),\ \ldots,\ (1+0.6z^{-1}),\ (1+0.8z^{-1})\}$$

The algorithm for approximate maximum likelihood estimation then follows, and is given in Procedure B.

Procedure B

1) Generate M sets of filtered data $\{(y^i_{f_1}, u^i_{f_1}) .. (y^i_{f_N}, u^i_{f_N})\}i=1,2..,M$ where a filtered variable $x^i_{f_t}$ of x^i_t is defined by

$$x^i_{f_t} = x_t/(1+\hat{C}_i(z^{-1})) \quad , \quad i=1,2,..,M \qquad 4.15$$

and $x_t = (y_{t-1},\cdots,y_{t-n_a},u_{t-k-1},\cdots,u_{t-k-n_b})$

2) For each i, i=1,2,....,M hypothesize that $(1+\hat{C}_i(z^{-1}))$ is in fact the maximum likelihood estimate $(1+C_{ml}(z^{-1}))$, and find (by linear least squares) the maximum likelihood estimates of the parameters of the model

$$[1+\hat{A}_i(z^{-1})]\, y^i_{f_t} = z^{-k}\hat{B}_i(z^{-1})u^i_{f_t} + \varepsilon^i_t \qquad 4.16$$

i.e. those parameters minimizing

$$V^i_N = \sum_{j=1}^{N} (\varepsilon^i_j)^2 \qquad 4.17$$

3) Evaluate V^i_N, i=1,2,...,M

4) Determine $V^{i^*}_N = \min_i V^i_N$, and record the number i* of the minimizing hypothesis. Then, under the assumption that $1+\hat{C}_{ml}(z^{-1}) \in C_o$, the Maximum Likelihood Estimates are given by

$$\hat{A}_{i^*}(z^{-1}),\ \hat{B}_{i^*}(z^{-1}),\ \hat{C}_{i^*}(z^{-1})$$

The effect of this procedure is to resolve the non-linear estimation problem into a set of parallel linear least-squares exercises. Both parameter estimates and cost function may be computed recursively, using the standard Recursive Least Squares Algorithm (Eykhoff 1974). For convenience, the equations are listed here.

Let $$\hat{\Theta}^i_t = (\hat{a}^i_1,\ldots,\hat{a}^i_{n_a},\hat{b}^i_1,\ldots,\hat{b}^i_{n_b})^T_t \qquad 4.18$$

and following the previously defined filtered-variable notation

$$x^i_{f_t} = (-y^i_{f_{t-1}},..,-y^i_{f_{t-n_a}}, u^i_{f_{t-k-1}},..,u^i_{f_{t-k-n_b}})^T \qquad 4.19$$

Then for the i^{th} hypothesis, the model equation (4.16) may be rewritten as

$$y^i_{f_t} = x^{iT}_{f_t}\hat{\Theta}^i_t + \varepsilon^i_t \qquad 4.20$$

The recursion for $\hat{\Theta}^i_t$ is then

$$\hat{\Theta}^i_{t+1} = \hat{\Theta}^i_t + P^i_{t+1}x_{f_{t+1}}\left[y^i_{f_{t+1}} - x^{iT}_{f_{t+1}}\hat{\Theta}^i_t\right] \qquad 4.21$$

where $$P^i_{t+1} = P^i_t - \gamma^i_t P^i_t x^i_{f_{t+1}} x^{iT}_{f_{t+1}} P^i_t \qquad 4.22$$

and

$$\gamma^i_t = (1 + x^{iT}_{f_{t+1}} P^i_t x^i_{f_{t+1}})^{-1} \qquad 4.23$$

The recursion for the cost V^i_t is

$$V^i_{t+1} = V^i_t + (\gamma^i_t)^{-1}\varepsilon^2_{t+1} \qquad 4.24$$

where $$\varepsilon^i_t = y^i_{f_t} - x^{iT}_{f_t}\hat{\Theta}^i_t \qquad 4.25$$

At each time iteration, eqns 4.21 - 4.25 are updated for i=1,2,..,M. Also, $V^{i^*}_t$ is evaluated, and the parameter values corresponding to hypothesis i* accepted as the optimal estimates at that iteration time.

The accuracy of the approximation to true maximum likelihood is determined by the quantization intervals q_j, $j=1,2,\ldots,n_c$. Whilst small intervals are desirable to improve accuracy, larger values of q_j result in fewer hypotheses and minimize computation.

The basic algorithm described thus far is the nucleus of a multi-pass interactive computer procedure which allows the analyst to interpret the results and analyse the cost function. This scheme is the subject of the next section.

4.3 An Interactive Computer Scheme

The data derived from the above algorithm contains a considerable amount of information, including an important section through the likelihood function, or cost function V. Although it is not possible to plot out the entire likelihood function as a function of all the parameters, it may be mapped as a function of the $C(z^{-1})$ polynomial coefficients (or equivalently the hypotheses of set C_o) for constrained values of $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$. This section of the likelihood function possesses the useful property that, assuming $\hat{C}_{ml}(z^{-1}) \in C_o$, it includes the unique maximum of the overall likelihood. This constitutes the first visual aid and is important in that it illustrates the sensitivity of the likelihood to changes in hypothesis. Even a coarsely quantised (discretized) display can be used to define more precisely the region in which the minimum of the cost is likely to occur. Thus, new smaller quantization intervals, $q_j, j=1,2,\ldots,n_c$ can be selected and a new set of hypotheses (set C_o) chosen to span the more narrowly defined region.

Clearly, this interactive process may be repeated until a sufficiently small quantization has been achieved.

The approach is not without its graphical difficulties. Whilst representation in the case of first and second order $1+C(z^{-1})$ polynomials is simple, it is more difficult (though by no means impossible) to interpret results for higher order polynomials. The display techniques are shown later. Fortunately, however, second order descriptions of $1+C(z^{-1})$ are adequate in many practical identification problems.

The second graphical aid is concerned with the sensitivity of the $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$ polynomial parameters to the choice of the hypothesized $1+\hat{C}(z^{-1})$ polynomial. These parameters may be plotted as a function of the hypothesis. Where only estimates of $A(z^{-1})$ and $B(z^{-1})$ are required, examination of this plot will reveal whether or not the quantization level is small enough to determine the parameters within the required accuracy. Often, very coarse quantization of the $\hat{C}(z^{-1})$ polynomial coefficients, and $\hat{C}(z^{-1})$ polynomials of reduced order will still result in dramatic improvements over least squares estimation.

Finally, the use of a recursive least squares algorithm makes it possible for the analyst to display the time evolution of the parameter estimates. This plot is in itself a valuable guide in assessing the quality of parameter estimates, and in particular, to decide whether the estimates have converged.

In summary, three graphical aids are available to the analyst.

1) A representation of a section through the likelihood function (or cost function) enabling new, smaller regions enclosing the extremum to be defined at each pass
2) A display illustrating the sensitivity of the $A(z^{-1})$ and $B(z^{-1})$ polynomial parameter estimates to changes in the hypothesized $1+\hat{C}(z^{-1})$ polynomial and effectively indicating the scale of error induced through the quantization approximation
3) A display of the time evolution of the parameter estimates themselves.

The interactive estimation procedure is illustrated by means of an example.

4.4 Illustrative Example

Consider the system described by the discrete-time equation

$$(1-1.5z^{-1}+0.7z^{-2})y_t = (z^{-1}+0.3z^{-2})u_t+(1-0.8z^{-1}+0.6z^{-2})e_t \qquad 4.26$$

where e_t and u_t are both white noise inputs with zero mean and unit variance. 200 observations of the input-output data (u_t and y_t) generated by this system were available. The system was modelled by

$$(1+a_1z^{-1}+a_2z^{-2})y_t = (b_1z^{-1}+b_2z^{-2})u_t+(1+c_1z^{-1}+c_2z^{-2})\varepsilon_t \qquad 4.27$$

Initially the quantisation intervals q_1, q_2 were both chosen to be 0.2. The candidate coefficients were then

$$Sc_1 = \{-1.8,-1.6,.....,0,.....1.6,1.8\} \qquad 4.28$$

$$Sc_2 = \{-0.8,-0.6,\ldots 0,\ldots.0.6,0.8\} \qquad 4.29$$

resulting in a set C_o of 87 allowable hypothesized polynomials, $1+\hat{C}_i(z^{-1})$, i=1,2,..,87. For example, for c_2=-0.8, the allowable polynomials are

$$\{(1+0.0z^{-1}-0.8z^{-2}),(1+0.2z^{-1}-0.8z^{-2})\} \qquad 4.30$$

The cost function ($\frac{1}{N}\sum_{\ell=1}^{N}\varepsilon_\ell^2$) is depicted in Fig. 4.2 where sections through it are plotted side by side against the corresponding c_1 and c_2 parameters and hypothesis number. The minimizing hypothesis is number 58, viz,

$$1+\hat{C}_{58}(z^{-1}) = 1-0.8z^{-1} + 0.6z^{-2} \qquad 4.31$$

The process of narrowing down the region of interest now begins. Examination of the cost function suggests that it is reasonable to confine the search to:

$$-1 < c_1 < -.4 \qquad 4.32$$

$$0.4 < c_2 < 0.8 \qquad 4.33$$

This enables the quantisation interval to be reduced to 0.1 for both q_1 and q_2, whilst simultaneously reducing the number of hypotheses to M=35 for the second pass. A similar analysis of the cost function obtained after this pass leads to the even smaller region

$$-0.9 < c_1 < -0.7$$

$$0.5 < c_2 < 0.7$$

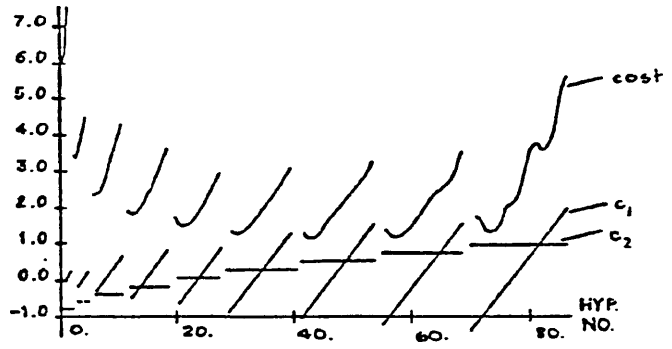


Fig. 4.2

Cost function and noise polynomial parameters vs. hypothesis number (Pass 1)

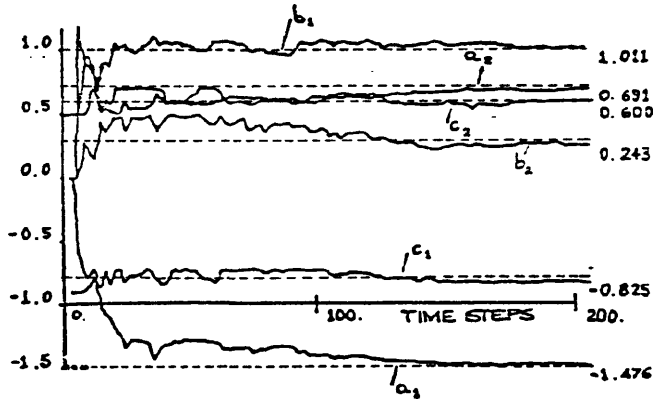


Fig. 4.3

Time evolution of the parameter estimates (pass 3)

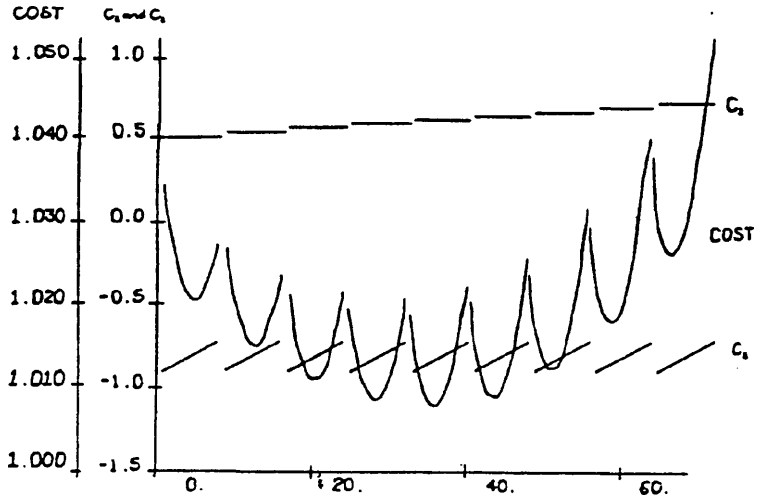


Fig. 4.4

Cost function and noise polynomial parameters vs. hypothesis number (pass 3)

being specified for the third pass. Quantization intervals q_1 and q_2 may be reduced to 0.025, giving 72 hypotheses.

Fig. 4.3 shows the time evolution of the parameter estimates after this run, whilst Fig. 4.4 depicts the cost function in a manner similar to Fig. 4.2. From this plot, it is evident that the minimum does indeed lie in the specified region. Since only a second order $1+C(z^{-1})$ polynomial is involved, it is possible to present this section of the cost function in a 3-dimensional perspective drawing as shown in Fig. 4.5. It is clear that the function is in fact extremely smooth and possesses a well-defined minimum.

Fig. 4.6 illustrates the sensitivity of the b_1 and b_2 parameter estimates to changes in hypothesis. Similar graphs can obviously also be drawn for $A(z^{-1})$ polynomial parameters. When using the actual interactive computer suite it is often useful to assess these variations in conjunction with the cost function (Fig. 4.4). Actual values of parameters or cost function are read by selecting points with a cursor facility (or light pen) on the graphics display terminal. In particular, to assess the errors induced by quantization it may be useful to examine changes in parameter values for hypotheses neighbouring the optimal choice. Fig. 4.7 shows the grid about the chosen hypothesis together with the value of parameter a_1. The maximum deviation from the chosen value is 0.0113.

The entire method is geared towards affording the analyst maximum interaction with the estimation procedure. A great deal of information is available, and its presentation could be further enhanced by greater programming effort.

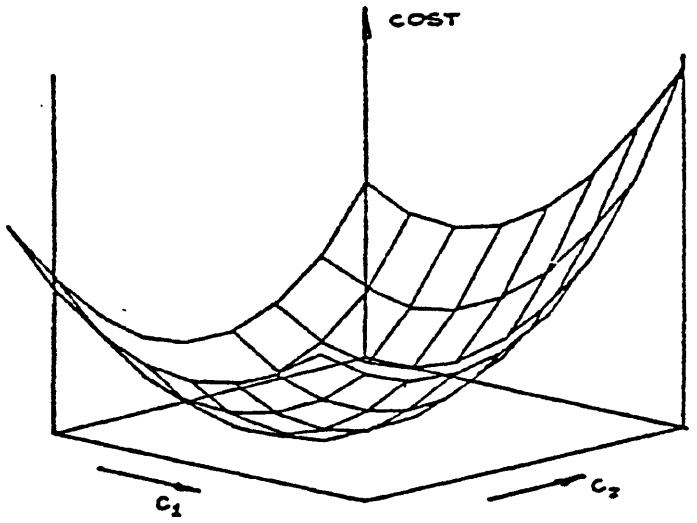


Fig. 4.5

3-D view of cost function (pass 3)

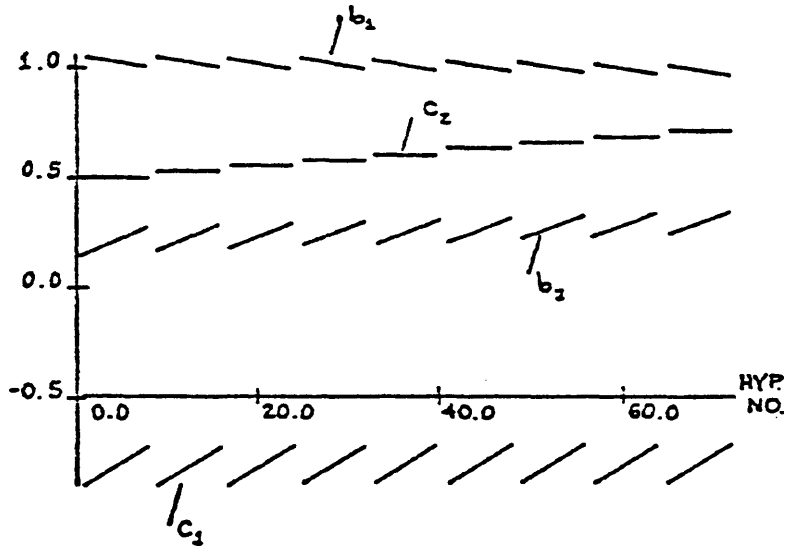


Fig. 4.6

Parameters b_1 and b_2 vs. hypothesis number (parameter sensitivity, pass 3)

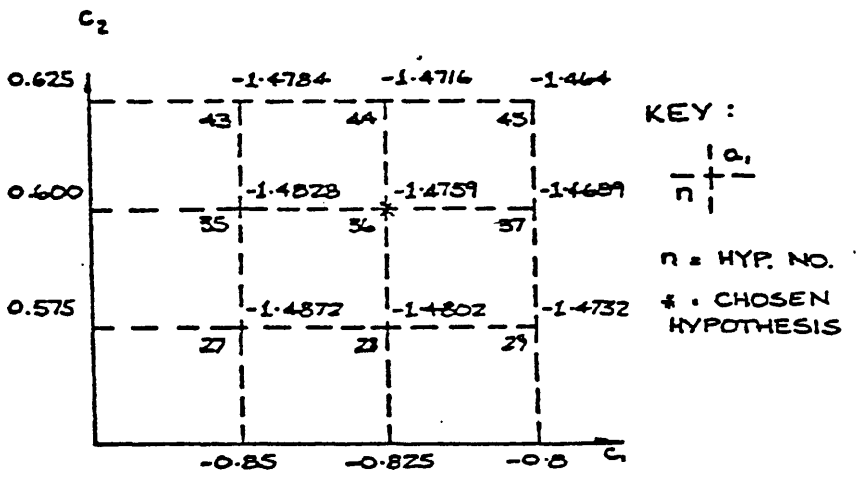


Fig. 4.7

Parameter a_1 sensitivity

For the above example, the computation time used on a PDP-10 system was:

Pass 1:	(87 hypotheses)	94s
Pass 2:	(35 hypotheses)	42s
Pass 3:	(72 hypotheses)	82s

This averages to approximately 5.3 ms/hypothesis/iteration.

4.5 Implementation of the Recursive Algorithm

The implementation of the algorithm is extremely simple and is briefly outlined here. The first task is to define the maximum permissible range of the coefficients $c_j, j=1,2,\ldots n_c$, in order for the hypothesized polynomials to be inverse stable. Maximum magnitudes of the coefficients can be found from Pascal's Triangle, and depend on the polynomial order. For example, for

i) $C(z^{-1})$ first order: $|c_1| < 1$

ii) $C(z^{-1})$ second order: $|c_1| < 2,\ |c_2| < 1$

iii) $C(z^{-1})$ third order: $|c_1| < 3,\ |c_2| < 3,\ |c_3| < 1$

and so on, following the rows of Pascals Triangle

1

1 2 1

1 3 3 1

etc.

Once the range of $c_j, j=1,2,\ldots,n_c$ has been established, the polynomial set C_o is constructed by forming $C(z^{-1})$ polynomials from all permutations of the given coefficient sets and accepting only

those which are inverse stable. The Jury Criterion (Jury, 1964) is suggested as a suitable test for establishing whether the zeros of $C(z^{-1})$ lie outside the unit disc.

Finally, the estimation problem is resolved into M parallel linear recursive least squares routines, for example, the 'Kalman Filter' implementation given in eqns. 4.21 - 4.25. In order to maintain accuracy, especially when only limited data is available, it is suggested that the initial conditions for the recursive algorithm are found by ordinary least squares using sufficient initial data to ensure that the information matrix has full rank.

4.6 Modified Implementations

The recursive method outlined in the previous sections is valuable when dealing with relatively short data records. Even with limited data it is usually possible to extract good parameter estimates. There are however instances where larger data sets are available and where the computational cost of running M parallel recursive least squares estimators may be unacceptable. Two alternative methods are described which rely on an approximation to the filter function $\frac{1}{1+C_i(z^{-1})}$. They are non recursive and thus the information in the time evolution of the parameter estimates is lost. However, they significantly reduce computational requirements.

From eqn 4.20, define the model equation for N observations of the input and output data as follows:

$$Y^i_{f_N} = X^i_{f_N} \hat{\theta}^i_N + E^i_N \qquad 4.34$$

where

$$Y^i_{f_N} = \frac{1}{1+\hat{C}_i(z^{-1})}[y_N,..,y_1]^T = \frac{1}{1+\hat{C}_i(z^{-1})}Y_N \quad 4.35$$

$$X^i_{f_N} = \frac{1}{1+\hat{C}_i(z^{-1})}[x_N,..,x_1]^T = \frac{1}{1+\hat{C}_i(z^{-1})}X_N \quad 4.36$$

$$E^i_N = [\varepsilon^i_N,...,\varepsilon^i_1]^T \quad 4.37$$

and $\hat{\theta}^i_N$ is the vector of least squares parameter estimates under the ith hypothesis, viz

$$\hat{\theta}^i_N = [X^{iT}_{f_N} X^i_{f_N}]^{-1} X^{iT}_f Y_{f_N} \quad 4.38$$

4.6.1 Correlation Method

As the number of observations $N \to \infty$, so the information matrix $\frac{1}{N}X_N^T X_N$ approaches more closely the block Toeplitz correlation matrix structured as follows:

$$\frac{1}{N}X_N^T X_N \to \begin{bmatrix} R_{yy} & \vdots & R_{yu} \\ \cdots & \cdots & \cdots \\ R^T_{yu} & \vdots & R_{uu} \end{bmatrix} = \bar{R} \quad \text{w.p.1 as } N \to \infty \quad 4.39$$

where R_{yy}, R_{yu} and R_{uu} are block matrices of orders $n_a x n_a$, $n_a x n_b$, and $n_b x n_b$ respectively, given by

$$R_{yy} = \begin{bmatrix} R_{yy}(0) & R_{yy}(1) & - & - & - & R_{yy}(n_a-1) \\ & \ddots & \ddots & \ddots & \ddots & \vdots \\ R_{yy}(n_a-1) & - & - & - & - & R_{yy}(0) \end{bmatrix} \quad 4.40$$

$$R_{yu} = \begin{bmatrix} R_{yu}(-k) & - & - & - & - & R_{yu}(-k-n_b+1) \\ \vdots & & & & & \vdots \\ R_{yu}(-k-1+n_a) & - & - & - & - & R_{yu}(-k-n_b+n_a) \end{bmatrix} \quad 4.41$$

and

$$R_{uu} = \begin{bmatrix} R_{uu}(0) & \cdots & R_{uu}(n_b-1) \\ \vdots & \ddots & \vdots \\ R_{uu}(n_b-1) & \cdots & R_{uu}(0) \end{bmatrix} \qquad 4.42$$

The correlation between two variables w and z is defined by

$$R_{wz}(k) = E[w_t z_{t+k}] \qquad 4.43$$

where E denotes the expectation operator.

Similarly as $N \to \infty$, so the vector

$$\frac{1}{N} X_N^T Y_N \to [R_{yy}(1),..,R_{yy}(n_a),R_{uy}(k+1),..,R_{uy}(k+n_b)]^T = \bar{W} \qquad 4.44$$

so that in the limit as $N \to \infty$, the vector of parameter estimates $\hat{\theta}_N$ for the unfiltered data is given by

$$\hat{\theta}_N = \bar{R}^{-1} \bar{W} \qquad 4.45$$

Now in order to avoid having to filter the <u>data</u> M times corresponding to the M hypotheses on the noise filter $1+C(z^{-1})$, a method is devised which allows the information matrix of the filtered data, $\frac{1}{N} X_{f_N}^{i\,T} X_{f_N}^{i}$, (denoted $\bar{R}_f^i$) and the vector $\frac{1}{N} X_{f_N}^{i\,T} Y_{f_N}^{i}$ (denoted $\bar{W}_f^i$) to be derived directly from correlation information on the unfiltered data.

Approximate the inverse filter $\frac{1}{1+\hat{c}_i(z^{-1})}$ by the pth order polynomial $H_i(z^{-1})$ such that

$$\frac{1}{1+\hat{C}_i(z^{-1})} \simeq H_i(z^{-1}) = \sum_{k=0}^{p} h_k^i z^{-k} \qquad 4.46$$

where $h_o = 1$, $h_\ell = -\sum_{m=1}^{\ell} c_m h_{\ell-m}$ 4.47

and p is the chosen truncation point of the series approximation. Thus $H_i(z^{-1})$ is simply the truncated version of the infinite order inverse of $1+\hat{C}_i(z^{-1})$.

It is then easy to relate the cross-correlations of two filtered variables $w^i_{f_t}$ and $z^i_{f_t}$ to the cross correlations of w_t and z_t, as follows:

$$R^i_{f_{wz}}(k) = R_{w^i_f z^i_f}(k) = \sum_{\ell=0}^{p}\sum_{m=0}^{p} h^i_\ell h^i_m R_{wz}(k+m-\ell) \qquad 4.48$$

where

$$w^i_{f_t} = H_i(z^{-1})w_t \simeq \frac{1}{1+\hat{C}_i(z^{-1})}\, w_t \qquad 4.49$$

$$z^i_{f_t} = H_i(z^{-1})z_t \simeq \frac{1}{1+\hat{C}_i(z^{-1})}\, z_t \qquad 4.50$$

Clearly then, the elements of the information matrix of <u>filtered data</u> given by

$$\lim_{N\to\infty} \frac{1}{N} X^{i^T}_{f_N} X^i_{f_N} = \begin{bmatrix} R^i_{f_{yy}} & R^i_{f_{yu}} \\ R^{i^T}_{f_{yu}} & R^i_{f_{uu}} \end{bmatrix} = \bar{R}^i_f \qquad 4.51$$

may be obtained approximately by operating on the <u>correlations</u> of the <u>unfiltered data</u> in the manner of eqn. 4.48, and similarly for the vector

$$\begin{aligned} \bar{W}^i_f &= \lim_{N\to\infty} \frac{1}{N} X^{i^T}_{f_N} Y^i_{f_N} \\ &= \{R^i_{f_{yy}}(1),..,R^i_{f_{yy}}(n_a),R^i_{f_{uy}}(k+1),..,R^i_{f_{uy}}(k+n_b)\}^T \end{aligned} \qquad 4.52$$

The parameter estimates for the ith hypothesis may then be computed from

$$\hat{\theta}_N^i = \bar{R}_f^{i\,-1} \bar{W}_f^i \tag{4.53}$$

and the cost function from

$$\frac{1}{N} V_N^i = R_{f_{yy}}^i(0) - \bar{W}_f^{i\,T} \hat{\theta}_N^i \tag{4.54}$$

The correlations of the unfiltered data required are:

$$R_{uu}(\tau), \quad \tau = -p, -p+1, \ldots\ldots, n_b+p-1 \tag{4.55}$$

$$R_{yy}(\tau), \quad \tau = -p, -p+1, \ldots\ldots, n_a+p-1 \tag{4.56}$$

$$R_{yu}(\tau), \quad \tau = k+1-n_b-p, \ldots\ldots, -k+1+n_a+p \tag{4.57}$$

and may be estimated from an expression of the form

$$R_{wz}(\tau) = \frac{1}{N-\tau} \sum_{m=1}^{N-\tau} w_m z_{m+\tau} \tag{4.58}$$

Of course advantage should be taken of the symmetry of correlation functions, viz

$$R_{wz}(\tau) = R_{zw}(-\tau) \tag{4.59}$$

The advantage of this method is that the parameter estimates under any filter hypothesis may be obtained without reprocessing the entire natural data, but by operating only on a set of stored correlations. The transformations involved are relatively fast and simple. All the information available in the recursive procedure

described for RHYP is accessible here, except for the time evolution of the parameter estimates. Estimates for each new hypothesis are obtained very quickly. However, the limitations of the approximations made should be borne in mind. In particular:

i) The 'block Toeplitz' approximation is only valid for large N. Good sample estimates of the correlation functions also require N to be large.

ii) The quality of the inverse filter approximation $H_i(z^{-1})$ is dependent both on the magnitude of p and the location of the zeros of $1+\hat{C}_i(z^{-1})$. The approximation, for a given p, deteriorates as the zeros of $1+\hat{C}_i(z^{-1})$ approach the boundary of the unit disc. As a rule of thumb, it is suggested that p be chosen as $10xn_c$. More elaborate schemes could clearly be devised which allow p to change as a function of the zeros of $1+\hat{C}_i(z^{-1})$, but this is not generally necessary.

iii) When there is little noise, the cost $\frac{1}{N}V_N^i$ becomes very small. Since both terms $R_{f_{yy}}^i(0)$ and $\bar{W}_f^{iT}\hat{\theta}_N^i$ involve approximations (due to approximate filtering and the correlation approximation) large errors may be generated. Computational error is minimized when p and N are large. An alternative method of computing the cost function is to substitute the parameter estimates into the model equation, and evaluate the residuals and their sum of squares by running the model. This method of course significantly affects computation time which then becomes directly proportional to N for each of the M hypotheses.

4.6.2 Covariance Method

This method also employs the inverse filter approximation $H_i(z^{-1})$ but avoids the approximation of a block Toeplitz structure for $\frac{1}{N} X_N^T X_N$.

Note that $Y^i_{f_N}$ may be approximated by

$$Y^i_{f_N} \simeq \{Y_N, Y_{N-1}, \ldots, Y_{N-p}\} t^i_p \qquad 4.60$$

where $$t^i_p = \{1, h^i_1, h^i_2, \ldots, h^i_p\}^T \qquad 4.61$$

Defining $$X^y_e = \{Y_{N-1}, \ldots, Y_{N-n_a-p}\} \qquad 4.62$$

$$X^u_e = \{U_{N-k-1}, \ldots, U_{N-k-n_b-p}\} \qquad 4.63$$

and $$X_{e_N} = \{X^y_e \;\; X^u_e\} \qquad 4.64$$

and following eqn. 4.60 it is easy to see that

$$X^{i^y}_{f_N} = \{Y^i_{f_{N-1}}, \ldots, Y^i_{f_{N-n_a}}\} \simeq X^y_e \begin{bmatrix} t^i_p & & \\ & \ddots & \\ & & t^i_p \end{bmatrix}_{n_a} \qquad 4.65$$

$$X^{i^u}_{f_N} = \{U^i_{f_{N-k-1}}, \ldots, U^i_{f_{N-k-n_b}}\} \simeq X^u_e \begin{bmatrix} t^i_p & & \\ & \ddots & \\ & & t^i_p \end{bmatrix}_{n_b} \qquad 4.66$$

Then $$X^i_{f_N} = \left[X^{i^y}_{f_N} \;\vdots\; X^{i^u}_{f_N}\right] \qquad 4.67$$

If the transformation matrix is defined by

$$T^i_{p,n_x} = \begin{bmatrix} t^i_p & & \\ & \ddots & \\ & & t^i_p \end{bmatrix}_{n_x} \qquad 4.68$$

eqns. 4.65 and 4.66 may be written more simply as

$$X_{f_N}^{i\,y} \simeq X_e^y T_{p,n_a}^i \tag{4.69}$$

$$X_{f_N}^{i\,u} \simeq X_e^u T_{p,n_b}^i \tag{4.70}$$

The information matrix of the filtered data is approximated by

$$\begin{bmatrix} X_{f_N}^{i\,T} & X_{f_N}^{i} \end{bmatrix} = \left[\begin{array}{c|c} (X_{f_N}^{i\,y})^T X_{f_N}^{i\,y} & (X_{f_N}^{i\,y})^T X_{f_N}^{i\,u} \\ \hline (X_{f_N}^{i\,u})^T X_{f_N}^{i\,y} & (X_{f_N}^{i\,u})^T X_{f_N}^{i\,u} \end{array}\right] \tag{4.71}$$

Now each block entry in 4.71 may be computed by operating on unfiltered data. For example

$$(X_{f_N}^{i\,y})^T X_{f_N}^{i\,y} \simeq T_{p,n_a}^{i\,T} X_e^{y\,T} X_e^y T_{p,n_a}^i \tag{4.72}$$

Similar results follow for the other block entries in eqn. 4.71. Thus if the "extended information matrix" of unfiltered data $X_{e_N}^T X_{e_N}$ is known, the information matrix of the <u>filtered data</u> (eqn. 4.71) may be obtained by transformations of the type shown in eqn. 4.72. Furthermore, the 'shift invariance' property and symmetry of the information matrix may be used to reduce computation. The 'shift invariance' property for the $(X_e^y)^T X_e^y$ matrix, for example, is that element $Y_{N-i}^T Y_{N-j}$ computed when N=t, is the same as element $Y_{N-i+1}^T Y_{N-j+1}$ computed when N=t-1. This means that elements appearing in rows other than the first row can be computed in the process of computing the first row covariances.

The same principle can be carried through to the computation of the vector $X_{f_N}^{i\,T}\, Y_{f_N}^{i}$ which is computed by transforming $X_{e_N}^{T}\{Y_N,..,Y_{N-p}\}$ as follows:

$$X_{f_N}^{i\,T}\, Y_{f_N}^{i} = \begin{bmatrix} (X_{f_N}^{i\,y})^T \\ \hline\hline (X_{f_N}^{i\,u})^T \end{bmatrix} Y_{f_N}^{i} \qquad 4.73$$

$$\simeq \left[\begin{array}{c:c} T_{p,n_a}^{i\,T} & \bigcirc \\ \hdashline \bigcirc & T_{p,n_b}^{i\,T} \end{array}\right] \left[\begin{pmatrix} X_e^{y\,T} \\ \\ X_e^{u\,T} \end{pmatrix} (Y_N,..,Y_{N-p})\right] t_p^i \qquad 4.74$$

The parameter estimates under each hypothesis are computed from

$$\hat{\theta}_N^i = \{X_{f_N}^{i\,T}\, X_{f_N}^{i}\}^{-1}\, X_{f_N}^{i\,T}\, Y_{f_N}^{i} \qquad 4.75$$

and the cost function from

$$V_N^i = Y_{f_N}^{i\,T}\, Y_{f_N}^{i} - \hat{\theta}_N^{i\,T}\{X_{f_N}^{i\,T}\, Y_{f_N}^{i}\} \qquad 4.76$$

Analogous to the correlation method, the parameter estimates for each new hypothesis are available from stored covariance information (independent of colouration polynomial) without recourse to the input-output data. The advantage of the covariance method is that, but for the inverse filter approximation, it involves no other approximation. As the 'block Toeplitz' approximation has been avoided, this method does not depend as crucially on a large data record, and generally is the more reliable algorithm. In particular there is a marked improvement in the estimation of the cost function. It does however still rely on choosing a large value of p, preferably of order $10xn_c$ (or larger if $1+C(z^{-1})$ appears to have zeros near the boundary of the unit disc) to obtain good results.

The covariance method involves more computation than the correlation method.

4.6.3 Modified Hill Climber

It is possible that more exact estimates of the parameters of the $1+C(z^{-1})$ polynomial are required than can reasonably be obtained by quantization. In such cases the covariance method may be used in conjunction with a hill climber. The standard hypothesis testing procedure could be used to obtain excellent initial conditions for the hill climber. This still enables the analyst to examine the likelihood function and gather parameter sensitivity information. The hill-climber then selects subsequent 'hypotheses' and the covariance method could be used to compute the $A(z^{-1})$ and $B(z^{-1})$ polynomial parameters and cost. By using the 'filtering' approach and obtaining the estimates $\hat{\theta}_N^i$ using simple linear least squares, the dimensionality of the optimization procedure is reduced from the general approach in which the hill climber operates on all parameters (Astrom and Bohlin, 1965) thus reducing computation. Rosenbrock's method (Rosenbrock, 1960) is a particularly robust optimization technique and has been successfully used in this application. Convergence tends to be fairly fast, typically 5 "rotations" when a stopping criterion is a change of less than 0.1% in the parameter estimates after a change in search direction.

4.6.4 Computational Implications

In order to assess the computational implications of the various approaches, the Illustrative Example was used as a basis for comparing the methods. The computation time used on a PDP10 system was recorded, giving the following results:

a) Recursive Method - 5.3 ms/hypothesis/iteration

b) Correlation Method - For p = 20

200 observations

Overhead to complete correlations: 1.9 seconds

To compute parameters and cost: 94ms/hypothesis

c) Covariance Method - For p = 20

200 observations

Overhead to compute covariances: 5.6 seconds

To compute parameters and cost: 133ms/hypothesis

Thus, if 194 hypotheses are tested (as in the Illustrative Example) and 200 observations are available, run times are:

Method	Run Time (s)
Recursive	205.6
Correlation	20.1
Covariance	31.4

Considerable run time savings can therefore be made by using an approximate technique, although the magnitude of p will determine the extent of the saving.

4.7 Simulation and Comparative Results

4.7.1 Comparison of RHYP with other algorithms for Illustrative Example

It is valuable to compare the performance of other algorithms with that of RHYP, and the parameter estimates obtained using RML (Recursive Maximum likelihood) and RIV (Recursive Instrumental Variables) for the data generated by the Illustrative Example are presented here.

RML yields fairly good results for this example. Fig. 4.8 shows the evolution of the parameter estimates with the following initial conditions:

a) The parameters were initialized to the least squares estimates.
b) The 'covariance' matrix was set to 100I.
c) The 'forgetting factor' was 0.98.

Note in particular that, although the $A(z^{-1})$ and $B(z^{-1})$ polynomial parameter estimates are good, those for the noise colouration filter (polynomial $C(z^{-1})$) are not. This phenomenon has been noticed often especially when using short data records. By comparison (Fig. 4.3) RHYP gives good estimates of the $C(z^{-1})$ polynomial parameters after only a few steps, as is typical of this method. It is particularly interesting to note the effect of changing the initial conditions on the parameter estimates for the RML run. Fig. 4.9 shows the results when the initial estimates were set to zero. The final estimates bear no relation to the true system parameters.

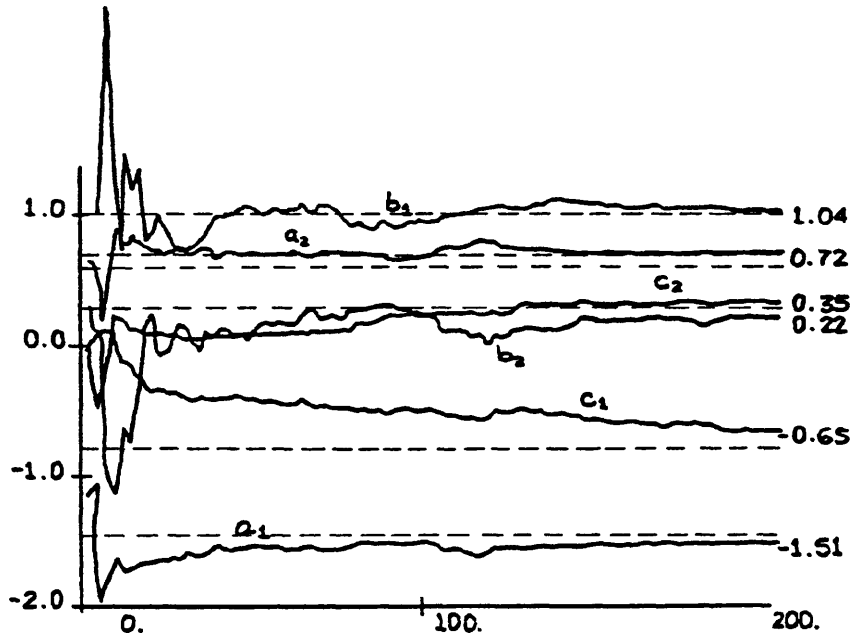


Fig. 4.8 Time evolution of RML parameter estimates (Initialized to least squares estimates)

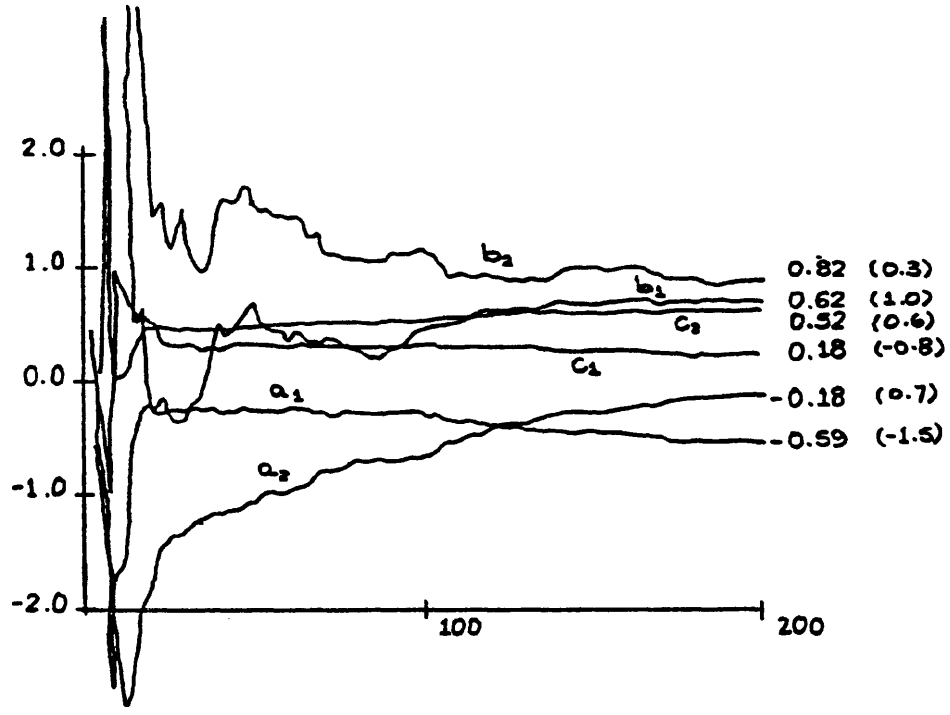


Fig. 4.9 Time evolution of RML parameter estimates (Initialised to zero)

In implementing the RML algorithm, it has been found necessary to test the inverse stability of the $1+\hat{C}(z^{-1})$ polynomial at every iteration. The Jury Criterion may be used to do this. Polynomials which are found to have zeros outside the z-plane unit circle are modified so as to move the zeros inside the unit circle. It should be recognised that results such as those shown in Fig. 4.9 would not be apparent from the comparative study of recursive identification methods carried out by Soderstrom, Ljung and Gustavsson (1974). This is because their results are averaged over 10 runs.

Fig. 4.10 illustrates the time evolution of the $A(z^{-1})$ and $B(z^{-1})$ polynomial parameter estimates using RIV. The instrumental model was formed using the least squares estimates for the data. The initial conditions were:

a) 'Covariance' matrix set to 100I.

b) 'Forgetting' factor 0.98.

In this run the results are good, although it has been demonstrated that convergence difficulties can occur using RIV (Soderstrom, 1974_b). A disadvantage of the basic RIV algorithm is that the $C(z^{-1})$ polynomial parameters are not estimated.

4.7.2 RHYP Using an under-parametrized noise polynomial model

Consider the system

$$(1-1.8z^{-1}+1.5z^{-2}-0.3z^{-3})y_t = (z^{-1}+0.7z^{-2}+0.4z^{-3})u_t + (1-1.2z^{-1}+0.4z^{-2}-0.1z^{-3})e_t \qquad 4.77$$

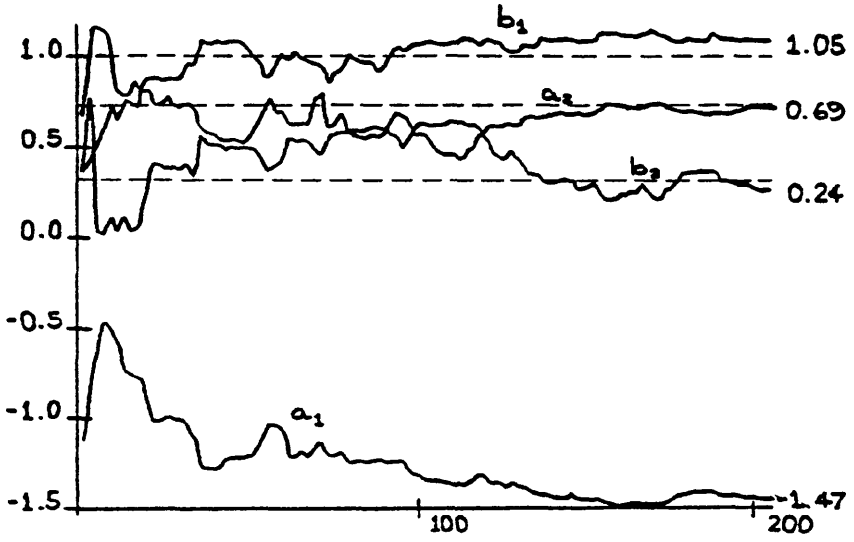


Fig. 4.10

Time evolution of RIV parameter estimates

where e_t is a non-measurable zero mean white noise input of variance 0.1 and u_t is a measurable input of unit variance. In order to illustrate that the hypothesis testing technique can lead to good estimates of the $A(z^{-1})$ and $B(z^{-1})$ polynomial parameters (using the usual notation) even when the noise colouration polynomial $1+C(z^{-1})$ is inadequately modelled, the following model was chosen.

$$(1+a_1z^{-1}+a_2z^{-2}+a_3z^{-3})y_t = (b_1z^{-1}+b_2z^{-2}+b_3z^{-3})u_t + (1+c_1z^{-1})\varepsilon_t \qquad 4.78$$

The data record consisted of 100 observations.

Fig. 4.11 shows the time evolution of the parameter estimates which are generally close to the true system values. A portion of the cost function is plotted against hypothesis number in Fig. 4.12 together with the hypothesized value of c_1. The minimizing cost (for hypothesis 5) was 0.08696 which is somewhat less than the minimum of 0.1 (corresponding to the noise variance) expected. The discrepancy is due to the small data sample used. Parameter sensitivity to changes in hypothesis is illustrated in Fig. 4.13. The parameter values do not vary substantially about the chosen hypothesis (the maximum change in any parameter by changing the chosen hypothesis by 1 being approximately 0.02) and this leads to a fair degree of confidence in their accuracy.

The results of the RHYP procedure may be compared with Recursive Instrumental Variables (Young et al 1968, 1970, 1971) in Fig. 4.14 and Recursive Maximum likelihood (Soderstrom 1973) in Fig. 4.15. The latter allowed for three $C(z^{-1})$ polynomial parameters. For the RML run the initial conditions were

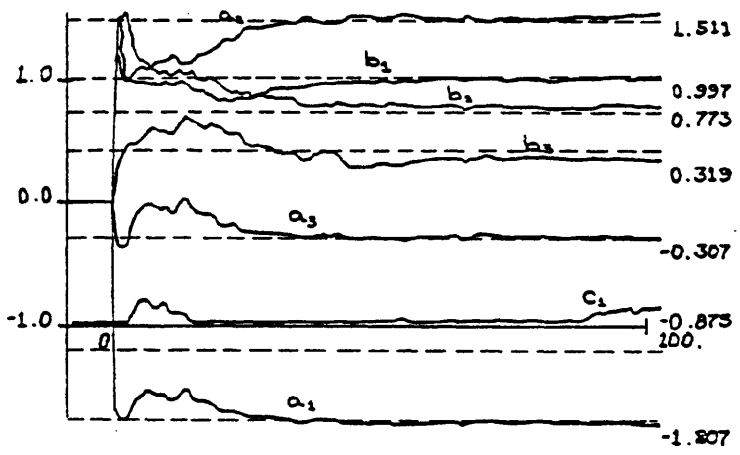


Fig. 4.11 Time evolution of RHYP parameter estimates

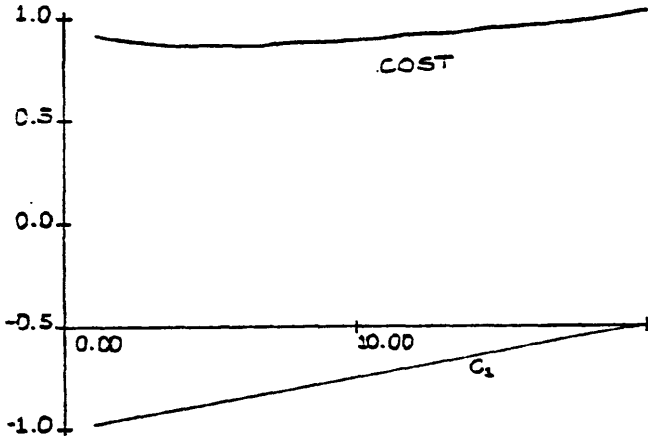


Fig. 4.12 Cost vs. hypothesis

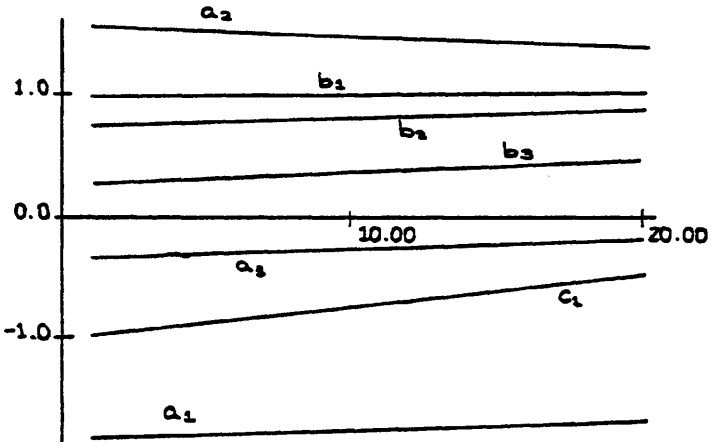


Fig. 4.13 Parameters a_1, a_2, b_1, b_2 vs hypothesis

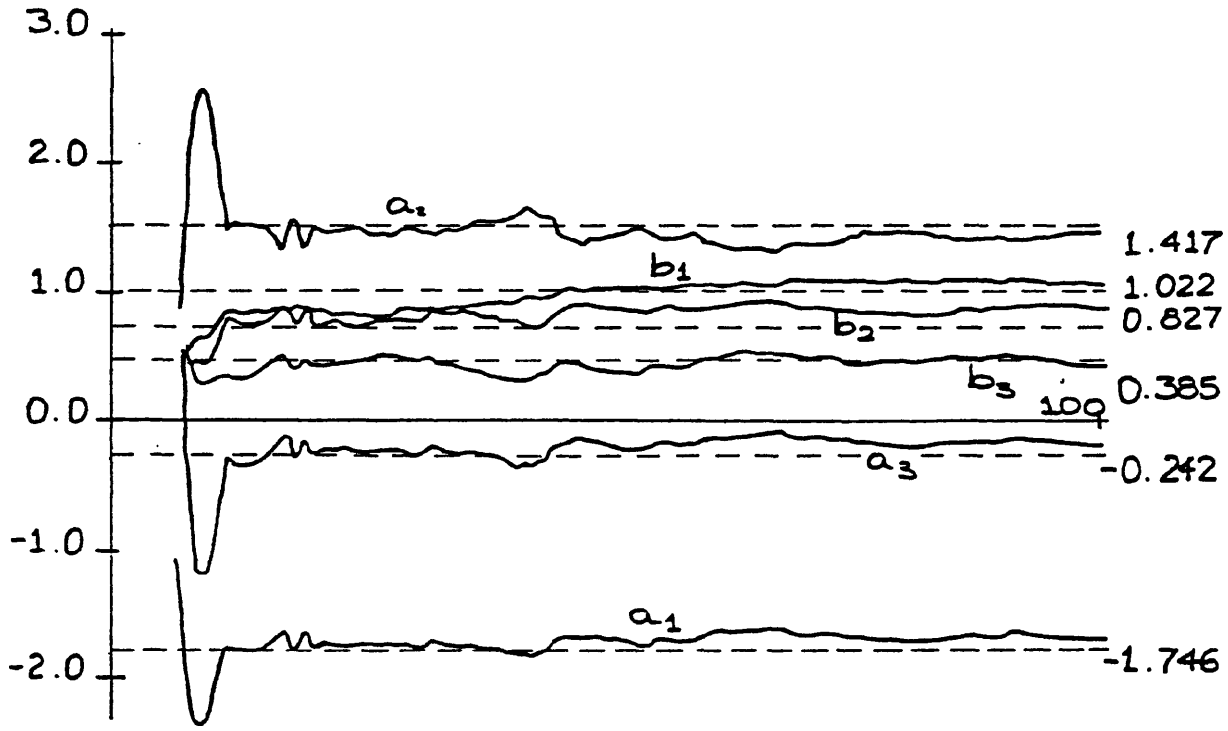


Fig. 4.14

Time evolution of RIV parameter estimates

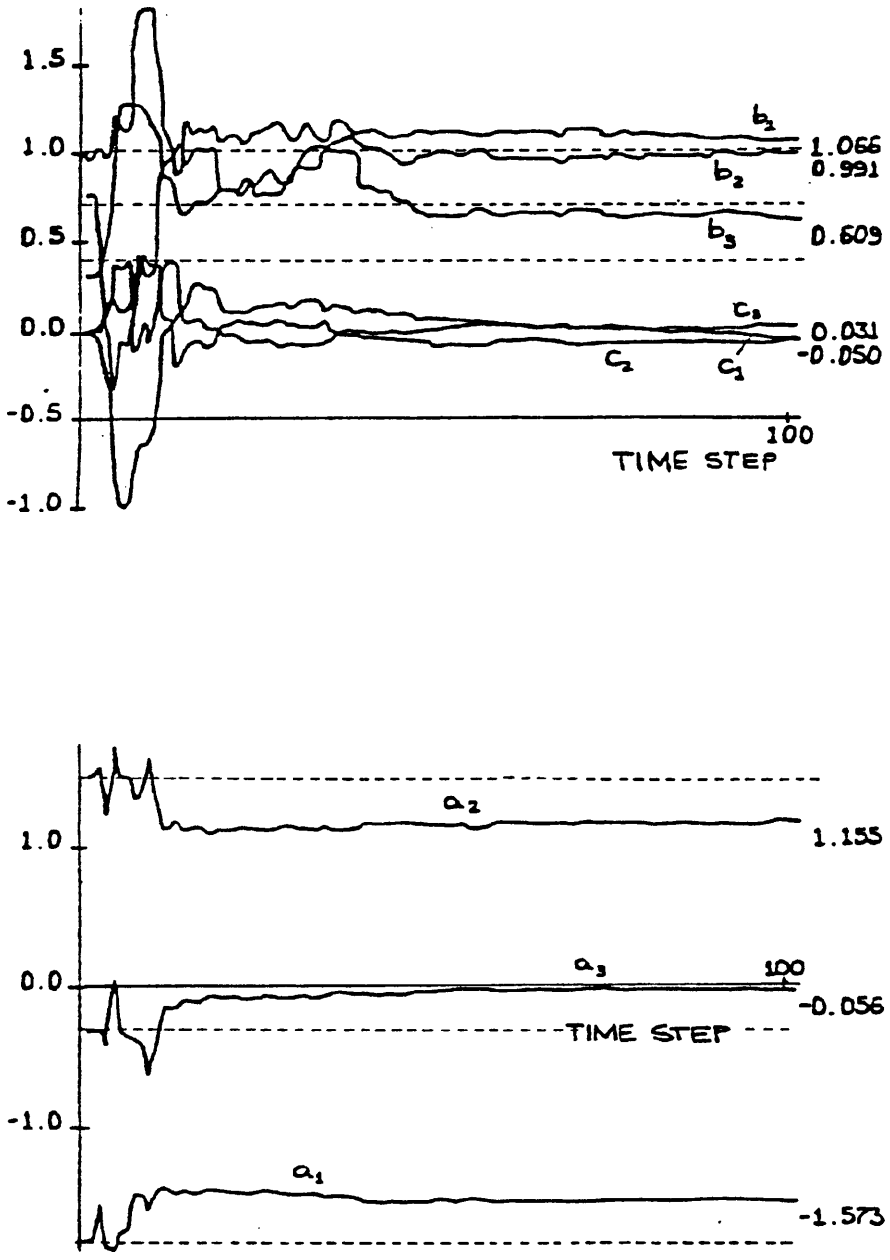


Fig. 4.15

Time evolution of RML parameter estimates

a) Covariance matrix set to 100I

b) Forgetting factor 0.98

c) $A(z^{-1})$ and $B(z^{-1})$ polynomial parameters set to RHYP estimates, $C(z^{-1}) = 0$

The Instrumental Model used for the RIV run was obtained from the least squares parameter estimates.

Table 4.1 compares the results obtained using several techniques:

Parameter	System	RLS	RIV	RML	RHYP
a_1	-1.8	-1.551	-1.746	-1.573	-1.807
a_2	1.5	1.122	1.417	1.155	1.51
a_3	-0.3	-0.0327	-0.242	-0.056	-0.307
b_1	1.0	1.022	1.022	1.066	0.997
b_2	0.7	1.011	0.827	0.991	0.773
b_3	0.4	0.593	0.385	0.609	0.319
c_1	-1.2	-	-	-0.05	-0.875
c_2	0.4	-	-	-0.05	-
c_3	-0.1	-	-	0.031	-

Table 4.1

In spite of adequately modelling the noise colouration polynomial, the RML estimates are poor. By contrast with RHYP, the $C(z^{-1})$ polynomial parameters tend to converge rather slowly.

The success of RHYP in this example is probably aided by the low noise variance and the fact that there is one dominant $1+C(z^{-1})$ polynomial zero, thus minimizing the effect of the under parametrized $C(z^{-1})$ polynomial. However, simulation runs have often demonstrated the superiority of the hypothesis testing technique and its success in achieving good estimates with short data records.

4.7.3 Identification of a Power Station Superheater

The performance of RHYP using real data is illustrated in this example. The system under analysis is the superheater of a power-station. A much simplified diagram of the system is given in Fig. 4.16. Steam passes from the boiler to the superheater. The spray valve allows a jet of cold water to be mixed with the steam and the spray valve position is a control input enabling the superheater output temperature to be controlled.

The estimation goal was to find a third order model to serve as a simulation model for testing various controllers. A PRBS sequence was used to excite the input and this data together with the output temperature was logged for 279 sampling periods.

Whilst least squares gave a satisfactory prediction model (in which the next prediction was computed from measured past inputs and outputs) it failed dismally when the output of the model was computed

from measured inputs and records of past outputs predicted by the model itself, (see Fig. 4.17). This is to be expected, as it is the one-step-ahead prediction error that is minimized, not the true output error. Bias in the parameter estimates leads to considerable errors in predicting further ahead.

The results of runs using RLS, RIV, RML and RHYP are best summarized by comparing the model impulse responses with the measured system impulse response. (Fig. 4.17). The instrumental model used the least squares parameter estimates which were also used as an initial condition for the RML estimator. The estimation model using RHYP included only a 2nd order $1+C(z^{-1})$ polynomial, yet resulting $A(z^{-1})$ and $B(z^{-1})$ polynomial estimates produce a remarkably good estimate of the impulse response.

It is interesting that when the model obtained by RHYP is used as an instrumental model, the RIV estimate of the impulse response improves dramatically. (Fig. 4.17)

Overall, however, the example testifies to the reliability and usefulness of the hypothesis testing procedure.

4.8 Conclusion

This chapter has presented a new method of obtaining Maximum Likelihood estimates by employing a Hypothesis Testing technique. The essential advantages of the method are its reliability and the interactive procedures which form an integral part of the overall algorithm and through which the analyst can assess the nature of the

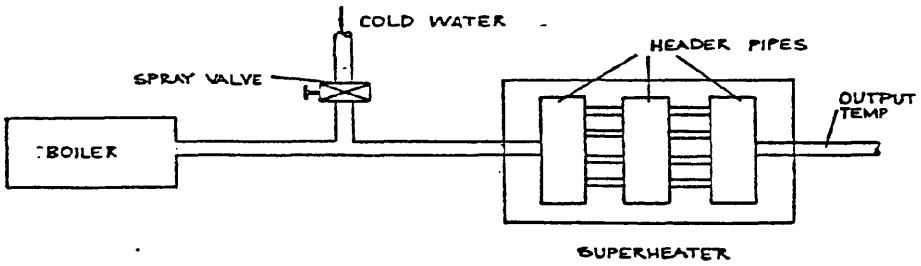


Fig. 4.16 Schematic diagram of superheater system

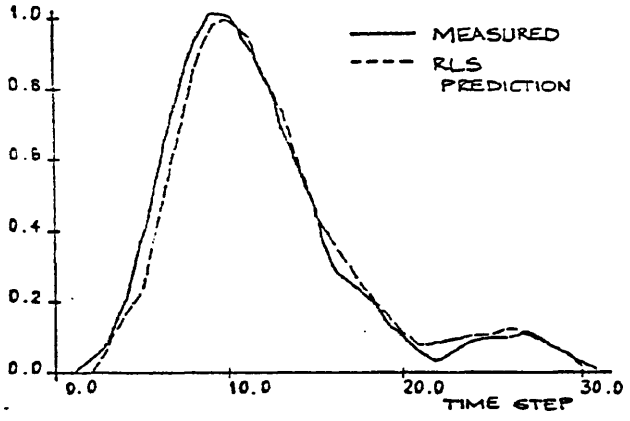


Fig. 4.17(a)

RLS one-step ahead predictor

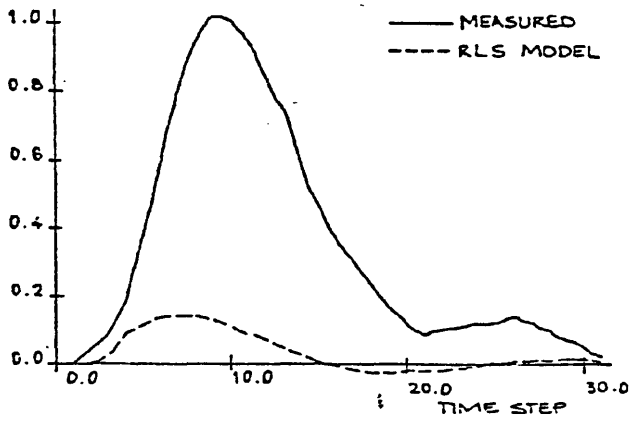


Fig. 4.17(b)

RLS model impulse response (output based on past predicted outputs)

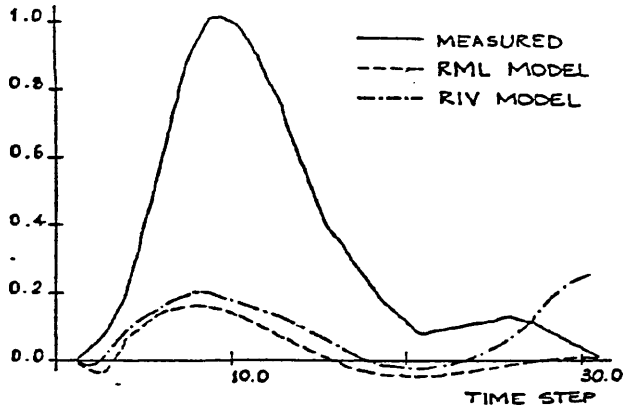


Fig. 4.17(c)

Impulse response of RML and RIV models

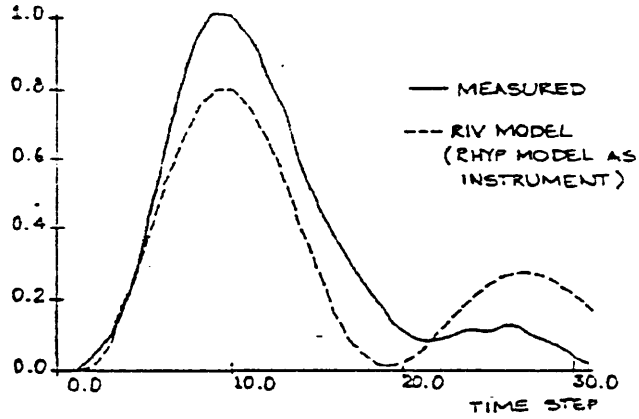


Fig. 4.17(d)

Impulse response of RIV model when using RHYP parameters in instrumental model

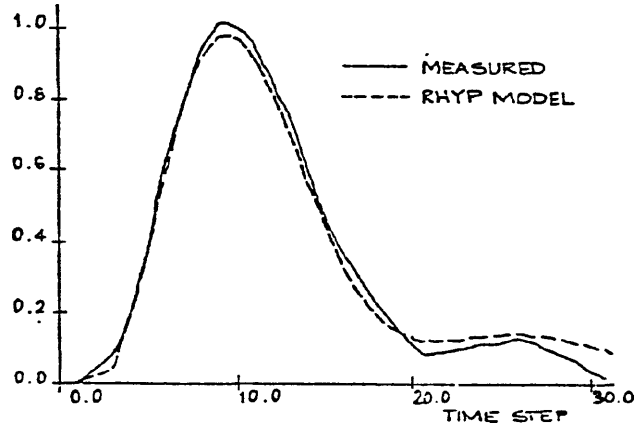


Fig. 4.17(e)

Impulse response of RHYP model

cost function and establish that minimizing parameters have been found. The estimation problem is viewed more broadly leading to a visual examination of the cost function rather than just seeking the minimum. Significantly, it may be found that in the region of the minimum the cost function is extremely 'flat' and that a wide range of models would give quite satisfactory solutions, or, that it is extremely sensitive to parameter variations. The method is particularly useful when short data records are available and the approximate recursive methods are likely to fail. Convergence of the noise colouration polynomial parameters is usually swift.

The key disadvantage is that the number of hypotheses grows exponentially with increasing order of the noise colouration polynomial. However, most single output systems are amenable to analysis. Furthermore, it has been observed that good estimates of the $A(z^{-1})$ and $B(z^{-1})$ polynomial parameters can often be obtained even when the $C(z^{-1})$ polynomial is under-parametrized in the estimation model.

Although the basic technique is recursive, alternative implementations have been described which are non-recursive and substantially reduce the computational requirement.

CHAPTER 5

PARAMETER ESTIMATION FOR MULTIVARIABLE SYSTEMS BY HYPOTHESIS TESTING

5.1 Introduction

This chapter is concerned with the extension of the principles established in Chapter 4 to the estimation of the parameters of a multivariable ARMAX model using a hypothesis testing approach. It was shown in Chapter 3 (eqn. 3.15) that the maximum likelihood estimates are to be found by minimizing a determinantal cost function. In the case of single output systems this cost function simplifies substantially and reduces to the sum of squares of the residual errors. The minimization of even this function presents no mean computational task. Thus Chapter 4 was dedicated to the development of a new algorithm that was aimed both at reducing computation and providing the analyst with a greater degree of insight into the nature of the likelihood function and sensitivity of the problem than is possible using direct hill-climbing techniques.

A little thought will show that the extension of RHYP to multivariable maximum likelihood estimation is not straightforward. Firstly, the fact that matrices do not generally commute makes it impossible to directly filter the system input and output records with the inverse noise colouration matrix polynomial as in the scalar case. To overcome this obstacle it is necessary to propose an alternative system parametrization.

Furthermore, the scalar method derives computational advantage by resolving the maximum likelihood estimation problem into a set of parallel linear least squares estimation exercises. A similar approach is possible in the multivariable case but is slightly more complicated because of the determinantal cost function. It is also possible to minimize another one-step-ahead prediction error cost function, namely the trace (rather than the determinant) of the residual covariance matrix. The implications of this choice on the statistical properties of the parameter estimates will be investigated and compared to the statistical properties of maximum likelihood estimates.

The prime objective of the multivariable hypothesis testing procedure is to identify a model which asymptotically produces the correct impulse response models for both the measurable (u_t) and non-measurable (e_t) inputs. Minimality of the representation is not required. It is with this in mind, together with the need for simplifying computation, that a suitable parametrization for the model is discussed.

The new algorithm is given the acronym MVHYP (Multi-Variable Hypothesis testing method). The layout of the chapter is such that the fundamental concepts of parametrization and cost function are discussed first, followed by the algorithmic details and simulation examples. It is also shown how a model with an orthogonal input noise (i.e. with uncorrelated elements of the noise vector) may be derived. A technique for obtaining a more concise model representation than that required for MVHYP is proposed.

5.2 The System Parametrization

A multivariable system with p outputs and r inputs is assumed to be governed by the ARMAX model

$$\{I+A(z^{-1})\}y_t = z^{-k}B(z^{-1})u_t + \{I+C(z^{-1})\}e_t \qquad 5.1$$

where $A(z^{-1})$, $B(z^{-1})$ and $C(z^{-1})$ are matrix polynomials of the form

$$X(z^{-1}) = X_1 z^{-1} + X_2 z^{-2} + ... + X_{n_x} z^{-n_x} \qquad 5.2$$

and e_t is a p-vector white driving noise with statistics

$$E(e_t) = 0 \qquad 5.3$$

$$E(e_t e_t^T) = Q \qquad Q > 0 \qquad 5.4$$

The model embodies the basic assumption of a generic structure (Denham, 1974; Dickenson et al, 1974). No attempt is made to estimate the structural indices, a problem currently enjoying a large degree of research interest. It is not assumed that the noise covariance matrix Q is diagonal. The elements of noise vector $e_t=(e_t^{(1)},....e_t^{(P)})^T$ may be correlated with one another. However, e_t may be derived from a noise vector v_t with uncorrelated components and covariance matrix $\text{diag}(\sigma_1^2,....\sigma_p^2)$ as follows:

$$e_t = \bar{K}\, v_t \qquad 5.5$$

where $\bar{K}$ is a lower triangular unit matrix selected such that

$$E(e_t e_t^T) = \bar{K} \begin{bmatrix} \sigma_1^2 & & \bigcirc \\ & \ddots & \\ \bigcirc & & \sigma_1^2 \end{bmatrix} \bar{K}^T = Q \qquad 5.6$$

If $\bar{C}(z^{-1})$ is now defined such that

$$\bar{C}(z^{-1}) = C(z^{-1})\bar{K} \qquad 5.7$$

it is clear that by substituting eqn. 5.5 into eqn. 5.1 and noting eqn. 5.7, an equivalent model

$$\{I+A(z^{-1})\}y_t = z^{-k}B(z^{-1})u_t + \{\bar{K}+\bar{C}(z^{-1})\}v_t \qquad 5.8$$

holds. It is sometimes useful to represent the system in this form, and in fact little extra work is required in the estimation routine to compute $\bar{K}$, as is shown later. However the optimal one-step-ahead prediction error remains

$$y_t - \hat{y}_{t/t-1} = e_t \qquad 5.9$$

Further manipulation, this time pre-multiplying eqn. 5.8 by the unit lower triangular matrix $K=\bar{K}^{-1}$, yields

$$\{\bar{A}_o+\bar{A}(z^{-1})\}y_t = z^{-k}\bar{B}(z^{-1})u_t + \{I+\bar{\bar{C}}(z^{-1})\}v_t \qquad 5.10$$

where $\bar{A}_o = K = \bar{\bar{K}}^{-1}$ (unit lower triangular)

$$\bar{A}(z^{-1}) = K\,A(z^{-1})$$
$$\bar{B}(z^{-1}) = K\,B(z^{-1})$$
$$\bar{\bar{C}}(z^{-1}) = K\,C(z^{-1})\,\bar{K}$$

and $E(v_t v_t^T)$ is a diagonal covariance matrix. 5.11

None of the above representations (5.1, 5.8 or 5.10) are in themselves suitable for extending the RHYP algorithm to the multivariable case. This is because the lack of a commutativity property for matrices does not permit the simple filtering of the input-output data y_t and u_t by the inverse noise colouration polynomial. However, by pre-multiplying eqn. 5.10 by adj $\{I+\bar{\bar{C}}(z^{-1})\}$ the following model results:

$$\{A_o+A(z^{-1})\}y_t = z^{-k}B(z^{-1})u_t + \{1+c(z^{-1})\}v_t \qquad 5.12$$

where

$$A_o = \bar{A}_o \quad \text{is unit lower triangular}$$

$$A_o + A(z^{-1}) = \text{adj}\{I+\bar{\bar{C}}(z^{-1})\}\{\bar{A}_o+\bar{A}(z^{-1})\}$$

$$B(z^{-1}) = \text{adj}\{I+\bar{\bar{C}}(z^{-1})\}\bar{B}(z^{-1})$$

and

$$1+c(z^{-1}) = \det\{I+\bar{\bar{C}}(z^{-1})\} \qquad 5.13$$

The structure of 5.12 is now more suited to an extended RHYP method, as the input-output data can be filtered directly by $1+c(z^{-1})$, a scalar polynomial, to give:

$$\{A_o+A(z^{-1})\}y_{f_t} = z^{-k}B(z^{-1})u_{f_t} + v_t \qquad 5.14$$

where, analogous to the SISO case,

$$y_{f_t} = y_t/\{1+c(z^{-1})\} \qquad 5.15$$

$$u_{f_t} = u_t/\{1+c(z^{-1})\} \qquad 5.16$$

The penalty of course is the increase in the order of the 'A' and 'B' polynomials, and thereby the number of parameters that must be estimated. The order of the $A_o+A(z^{-1})$ polynomial is now $n_a+n_c(p-1)$, and that of the $B(z^{-1})$ polynomial $n_b+n_c(p-1)$. The order of the scalar polynomial $1+c(z^{-1})$ becomes pn_c. The number of free parameters has increased from

$$(n_a+n_c)p^2 + \tfrac{1}{2}p(p-1) + n_b pr$$

in eqn. 5.10 to

$$(n_a+n_c(p-1))p^2 + \tfrac{1}{2}p(p-1) + (n_b+n_c(p-1))pr + pn_c$$

in eqn. 5.14. However, it turns out that by suitable choice of the cost function (discussed later) all but pn_c of these parameters may be estimated by linear least squares under a given hypothesis on the $1+c(z^{-1})$ polynomial. This is an important computational advantage.

At this point it is as well to note that the foregoing holds equally well for the representation

$$\{I+\bar{A}(z^{-1})\}y_{f_t} = z^{-k}\bar{B}(z^{-1})u_{f_t} + e_t \qquad 5.17$$

obtained by premultiplying eqn. 5.14 by $\bar{K}$, except that there are now $\frac{1}{2}p(p-1)$ fewer free parameters. The elements of the driving noise vector may now not be assumed to be uncorrelated.

The model structures suggested in eqns. 5.14 and 5.17 are in no way minimal, a factor which in the context of this work is not viewed as a particular disadvantage. The important feature is that the realization is capable of producing the correct system impulse

response, and can be used for optimal prediction. Any disadvantage of the representation is, in the author's view overshadowed by the simplicity (certainly where n_c and p are small) of the hypothesis-testing estimation procedure facilitated.

However, where it is essential to reduce the representation 5.14 or 5.17 to that of 5.1, the following procedure is suggested.

a) Use the estimates of $\bar{A}(z^{-1})$, $\bar{B}(z^{-1})$ in eqn. 5.17 or $A(z^{-1})$ and $B(z^{-1})$ in eqn. 5.14 to extract the estimate of the driving noise:

$$\varepsilon_t = \{I+\hat{\bar{A}}(z^{-1})\}y_{f_t} - z^{-k}\,\hat{\bar{B}}(z^{-1})u_{f_t} \qquad 5.18$$

where the caret indicates the estimate of the quantity.

b) Use ε_t as a known input in the model eqn

$$\{I+\hat{A}(z^{-1})\}y_t = \left[z^{-k}\hat{B}(z^{-1}) \quad I+\hat{C}(z^{-1})\right]\begin{bmatrix} u_t \\ \varepsilon_t \end{bmatrix} + w_t \qquad 5.19$$

and employ standard recursive least squares to minimize

$$\sum_{t=1}^{N} w_t^T w_t \qquad 5.20$$

The technique is demonstrated in the simulation example in section 5.7.2.

5.3 The Cost Function

The extension of RHYP to the multivariable case requires two fundamental properties:

a) It must be possible to filter input-output data with the inverse noise colouration polynomial

b) It is then essential that the cost can be minimized by using linear least squares.

The first criterion is satisfied by employing the model representations described above. This section is dedicated to the implications of the choice of cost function.

Consider first the system representation given in eqn. 5.17. Here the driving noise vector e_t is assumed to have correlated components (i.e. its covariance matrix is not diagonal). Following the arguments for the RHYP method, the maximum likelihood estimates $\hat{\bar{A}}_i(z^{-1})$ and $\hat{\bar{B}}_i(z^{-1})$ of $\bar{A}(z^{-1})$ and $\bar{B}(z^{-1})$ respectively under the hypothesis that $1+\hat{c}_i(z^{-1})$ is the maximum likelihood estimate of $1+c(z^{-1})$ are found by minimizing

$$V_N^i = \left| \sum_{t=1}^{N} \varepsilon_t^i \varepsilon_t^{iT} \right| \qquad 5.21$$

where $$\varepsilon_t^i = \{I+\hat{\bar{A}}_i(z^{-1})\}y_{f_t}^i - z^{-k}\hat{\bar{B}}_i(z^{-1})u_{f_t}^i \qquad 5.22$$

and $$y_{f_t}^i = y_{f_t}/\{1+\hat{c}_i(z^{-1})\} \qquad 5.23$$

$$u_{f_t}^i = u_{f_t}/\{1+\hat{c}_i(z^{-1})\} \qquad 5.24$$

The following analysis shows how the cost in eqn. 5.21 may be minimized by linear least squares. In order to simplify the notation the superscript 'i' denoting the ith hypothesis number and subscript 'f' will be omitted. Instead, the superscript will identify the element of the input, output or noise vector, for example

$$y_t = (y_t^{(1)}, y_t^{(2)}, \ldots\ldots, y_t^{(p)})^T \qquad 5.25$$

Similarly the subscript 'i' in $\hat{\hat{A}}_i(z^{-1})$ and $\hat{\hat{B}}_i(z^{-1})$ will be omitted and the polynomials will be expanded as follows:

$$\hat{\hat{A}}(z^{-1}) = \begin{bmatrix} a_{111} & a_{1p1} \\ a_{p11} & a_{pp1} \end{bmatrix} z^{-1} + \ldots + \begin{bmatrix} a_{11n_a} & a_{1pn_a} \\ a_{p1n_a} & a_{ppn_a} \end{bmatrix} z^{-n_a} \qquad 5.26$$

and similarly for $\hat{\hat{B}}(z^{-1})$ where the matrix elements will be denoted

$$b_{ij\ell}\ , \quad i=1,2,..,p\ ,\ j=1,2,..,r\ ,\ \ell=1,2,..,n_b$$

Now for the cost to be minimized

$$\frac{\partial}{\partial a_{ij\ell}} V_N = 0 \qquad i=1,2,..,p\ ,\ j=1,2,..,p\ ,\ \ell=1,2,..,n_a \qquad 5.27$$

and

$$\frac{\partial}{\partial b_{ij\ell}} V_N = 0 \qquad i=1,2,..,p\ ,\ j=1,2,..,p\ ,\ \ell=1,2,..,n_b \qquad 5.28$$

From the rules for differentiating determinants:

$$\frac{\partial}{a_{ij\ell}} V_N = \left| \begin{array}{ccc} \frac{\partial}{\partial a_{ij\ell}} \sum_t \varepsilon_t^{(1)}\varepsilon_t^{(1)} & \cdots\cdots\cdots & \frac{\partial}{\partial a_{ij\ell}} \sum_t \varepsilon_t^{(1)}\varepsilon_t^{(p)} \\ \sum_t \varepsilon_t^{(p)}\varepsilon_t^{(1)} & \cdots\cdots\cdots & \sum_t \varepsilon_t^{(p)}\varepsilon_t^{(p)} \end{array} \right| \longleftarrow$$

$$+ \ldots\ldots\ldots\ldots +$$

$$\left| \begin{array}{ccc} \sum_t \varepsilon_t^{(1)}\varepsilon_t^{(1)} & \cdots\cdots\cdots & \sum_t \varepsilon_t^{(1)}\varepsilon_t^{(p)} \\ \frac{\partial}{\partial a_{ij\ell}} \sum_t \varepsilon_t^{(p)}\varepsilon_t^{(1)} & \cdots\cdots\cdots & \frac{\partial}{\partial a_{ij\ell}} \sum_t \varepsilon_t^{(p)}\varepsilon_t^{(p)} \end{array} \right| \longleftarrow \quad \text{total of p determinants}$$

5.29

and similarly for $\frac{\partial}{\partial b_{ij\ell}} V_N$.

Now, for m=1,2,....,p and n=1,2,....,p

$$\frac{\partial}{\partial a_{ij\ell}} \sum_t \varepsilon_t^{(m)} \varepsilon_t^{(n)} = \sum_t (y_{t-\ell}^{(j)} \delta_{im} \varepsilon_t^{(n)} + \varepsilon_t^{(m)} y_{t-\ell}^{(j)} \delta_{in}) \qquad 5.30$$

$$\frac{\partial}{\partial b_{ij\ell}} \sum_t \varepsilon_t^{(m)} \varepsilon_t^{(n)} = \sum_t (-u_{t-k-\ell}^{(j)} \delta_{im} \varepsilon_t^{(n)} - \varepsilon_t^{(m)} u_{t-k-\ell}^{(j)} \delta_{in}) \qquad 5.31$$

$$\delta_{im} = \begin{cases} 1 & \text{for } i=m \\ 0 & \text{otherwise} \end{cases} \qquad 5.32$$

From eqns. 5.29, 5.30 and 5.31, it is evident that if the residuals $\varepsilon_t^{(i)}$, i=1,2,....,p are orthogonal to the data in the regression equation, then in fact one row in each of the p determinants of eqn. 5.29 will be zero, and thus

$$\frac{\partial}{\partial a_{ij\ell}} V_N = 0 \qquad 5.33$$

and similarly

$$\frac{\partial}{\partial b_{ij\ell}} V_N = 0 \qquad 5.34$$

for all i,j,ℓ.

But,these orthogonality conditions are precisely those for least squares estimation taking the regression equation for each output in turn and minimizing the cost

$$\sum_t (\varepsilon_t^{(i)})^2 \quad , \quad i=1,2,....,p.$$

Therefore linear least squares can indeed be used to minimize the cost in equation 5.21 <u>under a given hypothesis</u>.

However, in order to find the overall minimum of the cost function over all hypotheses, it is still necessary to evaluate the determinantal cost function for each hypothesis, and therefore the off-diagonal terms of the residual covariance matrix must be evaluated. This is a simple task, as is shown later.

In those instances where a model driven by a vector noise sequence with orthogonal elements is required, the parameter estimates $(I+\hat{A}_i(z^{-1}))$ and $\hat{B}_i(z^{-1})$ are transformed by premultiplying eqn. 5.22 by a lower triangular unit matrix designed to orthogonalize the residual covariance matrix. This matrix is in effect the estimate of A_o (=K) in the system representation of equation 5.12 - 5.14. Since $\hat{K}$ is unit lower triangular, the determinant of the residual covariance matrix is unchanged. However, the matrix is diagonal and therefore the determinant is evaluated very easily by multiplying together the diagonal elements. The algorithmic details are given in section 5.5.

The effort of evaluating the determinant (although relatively minimal) may be spared if another prediction error cost function, namely the trace (rather than determinant) of the residual covariance matrix is chosen.

Caines and Ljung (1976) have considered the properties of a large class of prediction error estimators. They consider the general system $y_t = f(y^{t-1}, u^t, \theta) + e_t$ 5.35

where, in their notation

$$y^t = \{y_i \; ; \; i \leq t-1\} \; , \quad u^t = \{u_i \; ; \; i \leq t\} \qquad 5.36$$

θ is a vector of the system parameters

and e_t is a zero mean process of independent random variables and constitutes the innovations process of y_t. Their study produces the following results.

a) If Σ is a positive definite matrix, ε_t is given by

$$\varepsilon(t,\hat{\theta}) = y_t - f(y^{t-1},u^t,\hat{\theta}) \qquad 5.37$$

and $\mathring{\theta}$ parametrizes the true system eqn. 5.35 and is unique, then

$$\text{tr } E\left[\Sigma\varepsilon(t,\hat{\theta})\varepsilon^T(t,\hat{\theta})\right] \geq \text{tr } E\left[\Sigma e_t e_t^T\right] \qquad 5.38$$

and the equality holds only if $\hat{\theta} = \mathring{\theta}$. Introducing assumptions of ergodicity ensures that the matrix

$$\frac{1}{N}\sum_{t=1}^{N} \varepsilon_t\varepsilon_t^T$$

converges to $E(\varepsilon_t\varepsilon_t^T)$ as $N \to \infty$ 5.39

b) Under certain mild conditions, generally fulfilled (see Caines and Ljung 1976) the parameter estimates $\hat{\theta}_N$ are asymptotically normally distributed in the sense that:

$$\sqrt{N}\,(\hat{\theta}_N - \mathring{\theta}) \sim N(0,P) \qquad 5.40$$

where covariance matrix P is given by

$$P = \{E(Z^T\Sigma\, Z)\}^{-1}\{E(Z^T\Sigma\Lambda\Sigma Z)\}\{E(Z^T\Sigma Z)\}^{-1} \qquad 5.41$$

and where $$z = \frac{d}{d\theta}\,\varepsilon(t,\theta)\Big|_{\theta=\dot{\theta}} \qquad 5.42$$

Λ = covariance matrix of the innovations process

c) The information matrix based on N observations is

$$F_N = N\,E(Z^T\Lambda^{-1}Z) \qquad 5.43$$

From (b) and (c) it is evident that if $\Sigma = \Lambda^{-1}$, P becomes

$$P = E\,(Z^T\Lambda^{-1}Z) \qquad 5.44$$

and minimization of the special cost function $\mathrm{tr}(\Lambda^{-1}\sum_{i=1}^{N}\varepsilon_i\varepsilon_i^T)$ leads to efficient estimates. Λ is however not generally known, and the estimates obtained using the trace cost function will not generally be efficient. The advantage of the maximum likelihood cost function is that it will guarantee efficient estimates.

The important point is that the cost function

$$\mathrm{tr}\left(\frac{1}{N}\sum_{t=1}^{N}\varepsilon_t\varepsilon_t^T\right) \qquad 5.45$$

does lead to consistent estimates (from (a) above) if $\dot{\theta}$ is unique. Thus the trace cost function may be considered to be a useful alternative to the maximum likelihood cost function.

It is important to note that when the model is over parametrized the concept of consistency must be modified. The cost functions only ensure that asymtotically

$$\{\hat{A}_o+\hat{A}(z^{-1})\}^{-1}\hat{B}(z^{-1}) \to \{A_o+A(z^{-1})\}^{-1}B(z^{-1}) \qquad 5.46$$

and

$$\{\hat{A}_o+\hat{A}(z^{-1})\}^{-1}\{1+\hat{c}(z^{-1})\} \to \{A_o+A(z^{-1})\}^{-1}\{1+c(z^{-1})\} \qquad 5.47$$

in the notation of the representation of eqn 5.12 (where the driving noise has uncorrelated elements). Thus, it is only 'impulse responses' to input signals u_t and v_t that are asymptotically correctly estimated. This is consistent with the initial expectations for this algorithm.

Having discussed both system parametrization and choice of cost function, the extension of the basic RHYP algorithm is straightforward and is presented below.

5.4 The MVHYP Algorithm

In order to define the notation, let the difference equation

$$\{I+\hat{\hat{A}}(z^{-1})\}y_t = z^{-k}\hat{\hat{B}}(z^{-1})u_t + \{1+\hat{\hat{c}}(z^{-1})\}\varepsilon_t \qquad 5.48$$

model the system representation of eqn. 5.17, and

$$\{\hat{K}+\hat{A}(z^{-1})\}y_t = z^{-k}\hat{B}(z^{-1})u_t + \{1+\hat{c}(z^{-1})\}v_t \qquad 5.49$$

model the system representation of eqn. 5.14. The orders of the estimated polynomials are the same as the system polynomials.

The algorithm, which depends on the choice of cost function, is based on the arguments of the previous section and is:

(1) Select quantization intervals $q_j, j=1,\ldots,n_{\hat{c}}$ and find all allowable polynomials $1+\hat{c}_i(z^{-1})$, $i=1,\ldots,M$ according to Procedure A of Chapter 4.

(2) Generate M sets of filtered data, for $t=1,2,\ldots,N$

$$y_{f_t}^i = y_t/\{1+\hat{c}_i(z^{-1})\}$$
$$u_{f_t}^i = u_t/\{1+\hat{c}_i(z^{-1})\} \qquad i=1,2,..,M \qquad 5.50$$

(3) For each i, i=1,2,...,M hypothesize that $1+\hat{c}_i(z^{-1})$ is in fact the 'optimal' estimate (for the chosen cost function) $1+\hat{c}_o(z^{-1})$.(For the determinantal cost function $1+\hat{c}_o(z^{-1})$ is assumed to be the maximum likelihood estimate of $1+c(z^{-1})$). Then obtain estimates $\hat{\hat{A}}(z^{-1})$ and $\hat{\hat{B}}(z^{-1})$ by minimizing

$$V_N^{i(j,j)} = \sum_{t=1}^{N} (\varepsilon_t^{i(j)})^2 \qquad j=1,2,..,p \qquad 5.51$$

where $\varepsilon_t^{i(j)}$ is the jth element of the residual vector ε_t^i under the ith hypothesis at time t. The estimates are then optimal under the hypothesis.

4(a) If the overall cost function is to be the Trace of the residual covariance matrix, compute

$$V_{N_{tr}}^i = \sum_{i=1}^{p} V_N^{i(j,j)} \quad \text{for} \quad i=1,2,..,M \qquad 5.52$$

and record the minimizing hypothesis number i*. Then the minimizing estimates are $\hat{\hat{A}}_{i*}(z^{-1})$ and $\hat{\hat{B}}_{i*}(z^{-1})$.

4(b) If the overall cost function is to be the Determinant of the residual covariance matrix (yielding maximum likelihood estimates) compute

$$V_N^{i(j,\ell)} \qquad j=1,2,..,p \ , \ \ell=1,2,..,j$$

where $V_N^{i(j,\ell)}$ is the (j,ℓ) th element of the residual covariance matrix

$$V_N^i = \sum_{t=1}^{N} \varepsilon_t^i \, \varepsilon_t^{iT} \qquad 5.53$$

Evaluate $$V^i_{N_{det}} = |V^i_N| \qquad 5.54$$

and record the minimizing hypothesis number i*. Then the minimizing estimates $\hat{\hat{A}}_{i*}(z^{-1})$ and $\hat{\hat{B}}_{i*}(z^{-1})$ are the required maximum likelihood estimates.

4(c) An alternative to 4(b) arises if the representation with a diagonal residual covariance matrix is required. Compute $\hat{K}_i$ such that:

$$v^i_t = \hat{K}_i \varepsilon^i_t \qquad 5.55$$

and $$\sum_{t=1}^{N} v^i_t(j)\, v^i_t(\ell) = 0 \quad \text{for } j \neq \ell \ , \ i=1,2,..,M$$

and hence compute

$$V^i_{N_{det}} = \prod_{j=1}^{p} \bar{V}^i_N(j) \quad \text{for} \quad i=1,2,..,M \qquad 5.56$$

where

$$\bar{V}^i_N(j) = \sum_{t=1}^{N} (v^i_t(j))^2 \qquad j=1,2,..,p$$

Select the optimum hypothesis number i* as in 4(b).

The computational details are given in section 5.5.

5.5 Recursive Computation

Re-writing the model equations 5.48 and 5.49 for the ith hypothesis gives

$$\{I+\hat{\hat{A}}_i(z^{-1})\}y^i_{f_t} = z^{-k}\hat{\hat{B}}_i(z^{-1})u^i_{f_t} + \varepsilon^i_t \qquad 5.57$$

and $$[\hat{K}_i + \hat{A}_i(z^{-1})]\, y^i_{f_t} = z^{-k}\, \hat{B}_i(z^{-1}) u^i_{f_t} + v^i_t \qquad 5.58$$

respectively. In the remainder of this section, the subscript f and superscript i indicating filtered data and hypothesis number are omitted to simplify notation. It is also helpful to re-write eqn. 5.57 in the form

$$y^{(i)}_t = x^T_t \hat{\theta}_{i_t} + \varepsilon^{(i)}_t \quad , \quad i=1,2,\ldots,p \qquad 5.59$$

where superscript (i) indicates the element number of the vector,

$$x_t = \{-y^{(1)}_{t-1},..,-y^{(1)}_{t-n_a},..,-y^{(p)}_{t-1},..,-y^{(p)}_{t-n_a},$$
$$u^{(1)}_{t-k-1},..,u^{(1)}_{t-k-n_b},..,u^{(p)}_{t-k-1},..,u^{(p)}_{t-k-n_b}\}^T \quad , \qquad 5.60$$

$\hat{\theta}_{i_t}$ is the vector of parameters for the ith output equation and n_a and n_b are the orders of polynomials $\hat{\hat{A}}(z^{-1})$ and $\hat{\hat{B}}(z^{-1})$ respectively.

The estimation of $\hat{\theta}_{i_t}$, i=1,...,p by minimizing the sum of squares of the residuals for each output equation (see eqn. 5.51 of the algorithm above) is therefore extremely simply achieved using recursive least squares. Note also that the 'covariance matrix'

$$P_t = (X^T_t X_t)^{-1}$$

where $$X_t = (x_t,\ldots x_1)^T \qquad 5.61$$

and factor $$\gamma_t = (1 + x^T_t P_t x_t)^{-1}$$

remain the same for $\hat{\theta}_{i_t}$, i=1,...,p. Thus only ONE covariance matrix need be updated. Also the elements of the residual covariance matrix (required in steps 3 and 4(b) of the algorithm) may be updated recursively in a manner similar to that used in the RHYP algorithm (eqn. 4.24) as follows:

$$V_t^{(j,\ell)} = V_{t-1}^{(j,\ell)} + \gamma_t^{-1}\, \varepsilon_t^j\, \varepsilon_t^\ell \qquad 5.62$$

where $\varepsilon_t^j = y_t^j - x_t^T\, \hat{\theta}_{j_t}$

The off diagonal terms ($j \neq \ell$) need only be evaluated when using the determinantal cost function. Thus the minimization of both cost functions is based on simple recursions at each sampling time, for each hypothesis.

The parametrization requiring orthogonalization of the residual vector demands further processing. The method described below shows how the parameter estimates $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$ may be evaluated with relatively little effort from data already computed. The diagonal covariance matrix may also be computed and thus evaluation of the determinant is very easy indeed.

The method requires the estimation of matrix $\hat{K}$ such that

$$\sum_{t=1}^{N} v_t^{(i)} v_t^{(j)} = 0 \text{ for } i \neq j.$$

Note also that $v_t = \hat{K}\, \varepsilon_t$ 5.63

or $$v_t^{(i)} = \varepsilon_t^{(i)} + \sum_{\ell=1}^{i-1} \hat{K}_{i\ell} \varepsilon_t^{(\ell)}$$

where $\hat{K}_{i\ell}$ is the $\{i,\ell\}$th element of the unit triangular matrix $\hat{K}$.

First re-write eqn. 5.58 in a manner similar to eqn. 5.59, i.e.

$$y_t^{(i)} = x_t^T \hat{\phi}_i - \sum_{j=1}^{i-1} \hat{K}_{ij} y_t^{(j)} + v_t^{(i)} \qquad i=1,2,\ldots,p \qquad 5.64$$

where $\hat{\phi}_i$ is the vector of parameters for the ith output equation and

$$\hat{\phi}_i = \hat{\theta}_i + \sum_{\ell=1}^{i-1} \hat{\theta}_\ell \hat{K}_{i\ell} \quad ; \quad \hat{\phi}_1 = \hat{\theta}_1 \qquad 5.65$$

If $$Y_t^{(i)} = \{y_t^{(i)}, y_{t-1}^{(i)}, \ldots, y_1^{(i)}\}^T \qquad 5.66$$

$$V_t^{(i)} = \{v_t^{(i)}, \ldots, v_1^{(i)}\}^T \qquad 5.67$$

and $$E_t^{(i)} = \{\varepsilon_t^{(i)}, \ldots, \varepsilon_1^{(i)}\}^T \qquad 5.68$$

then from eqn.5.64

$$Y_t^{(i)} = X_t \hat{\phi}_i - \sum_{j=1}^{i-1} \hat{K}_{ij} Y_t^{(j)} + V_t^{(i)} \qquad i=1,2,..,p \qquad 5.69$$

and from eqn. 5.63

$$V_t^{(i)} = E_t^{(i)} + \sum_{j=1}^{i-1} \hat{K}_{ij} E_t^{(j)} \qquad i=1,2,..,p \qquad 5.70$$

Now if $\hat{\phi}_i$ and $\hat{K}_{ij}$, i=1,...,p , j=1,...,i-1 are the least squares estimates, then clearly the required orthogonality condition on the residual vector v_t is met. However direct computation of these least squares estimates would require updating a different covariance matrix for each output equation (eqn. 5.69), and is computationally undesirable. The method derived here enables existing information to be utilized. Premultiplying eqn. 5.69 by $V_t^{(m)^T}$, m < i, and rearranging yields

$$V_t^{(m)^T} V_t^{(i)} = V_t^{(m)^T} Y_t^{(i)} + \hat{K}_{im} V_t^{(m)^T} Y_t^{(m)} + \sum_{\substack{j=1 \\ j\neq m}}^{i-1} \hat{K}_{ij} V_t^{(m)^T} Y_t^{(j)} \qquad 5.71$$

since from the properties of least squares

$$E_t^{(i)T} X_t = 0 \qquad \text{for } i=1,2,\ldots,p \tag{5.72}$$

and observing eqn. 5.70

$$V_t^{(m)T} X_t = 0 \qquad \text{for } m=1,2,\ldots,p \tag{5.73}$$

Now $\hat{K}$ is to be selected so that $V_t^{(m)T} V_t^{(i)} = 0 \quad i \neq m$.
Substituting into eqn. 5.71 and observing that for $\hat{K}$ to satisfy the orthogonality conditions for least squares estimation

$$V_t^{(m)T} Y_t^{(j)} = 0 \quad \text{for } j < m \tag{5.74}$$

and that

$$V_t^{(m)T} Y_t^{(m)} = V_t^{(m)T} V_t^{(m)} \tag{5.75}$$

it is possible to solve for $\hat{K}_{im}$ thus

$$\hat{K}_{im} = \frac{-V_t^{(m)T} Y_t^{(i)} - \sum_{j=m+1}^{i-1} \hat{K}_{ij} V_t^{(m)T} Y_t^{(j)}}{V_t^{(m)T} V_t^{(m)}} \tag{5.76}$$

The terms $V_t^{(i)T} Y_t^{(j)}$ and $V_t^{(m)T} V_t^{(m)}$ must be available.
Now from eqn. 5.70 and noting that the properties of least squares ensure that

$$E_t^{(i)T} Y_t^{(j)} = E_t^{(i)T} E_t^{(j)} \tag{5.77}$$

it is evident that, by post-multiplying the transpose of eqn. 5.70 by $Y_t^{(j)}$,

$$V_t^{(m)T} Y_t^{(j)} = E_t^{(m)T} E_t^{(j)} + \sum_{\ell=1}^{m-1} \hat{K}_{m\ell} E_t^{(\ell)T} E_t^{(j)} \qquad 5.78$$

and since the terms $E_t^{(\ell)T} E_t^{(j)}$ may be computed recursively (see eqn. 5.62) the above equation is easily computed.

Also, $V_t^{(m)T} V_t^{(m)}$ can be computed from the known residual error covariance matrix $\sum_{i=1}^{t} \varepsilon_i \varepsilon_i^T$ by operating on it as follows:

$$V_t^{(m)T} V_t^{(m)} = \{\underbrace{K_{m1},..,K_{m,m-1}\ 1\ 0..0}_{p}\}\{\sum_{i=1}^{t} \varepsilon_i \varepsilon_i^T\}\begin{bmatrix} K_{m1} \\ \vdots \\ K_{m,m-1} \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad 5.79$$

The order in which the various terms is computed is important, as $\hat{K}_{im}$ is itself a function of $\hat{K}_{ij}$, j=m+1,...,i-1, (m<i), and $\hat{K}_{m\ell}$, ℓ=1,...,m-1. The suggested computational sequence is therefore:

(1) Compute parameter covariance matrix P_t and

$$\gamma_t = (1 + x_t^T P_{t-1} x)^{-1}$$

as in the normal recursive least squares algorithm (see eqns. 4.22 and 4.23 for analagous operations) and hence compute $\hat{\theta}_i$, i=1,...p, and residual covariance matrix $\sum_{i=1}^{t} \varepsilon_i \varepsilon_i^T$.

For i=2:

(2) Compute $V_t^{(m)^T} Y_t^{(j)}$ from eqn. 5.78 m=1,...,i-1,

j=m+1,...,i

where this computation is a function of

$\hat{K}_{m\ell}$,m=1,..,i-1, ℓ=1,..,m-1, already computed.

(3) Now, from eqn. 5.79 evaluate $V_t^{(m)T} V_t^{(m)}$, m=1,...,i-1

where $V_t^{(m)^T} V_t^{(m)}$ is a function of

$\hat{K}_{m\ell}$, m=1,...,i-1, ℓ=1,...,m-1 already computed and the residual covariance matrix already available.

(4) Compute $\hat{K}_{im}$, m=i-1,...,1 from eqn. 5.76 (i.e. starting with $\hat{K}_{i,i-1}$) where $\hat{K}_{im}$ is a function of evaluated covariances and

$$\hat{K}_{ij} \ , \ j = m+1,\ldots,i-1 \ .$$

(5) If i<p, increment i and repeat from (2).

Having computed the $\hat{K}$ matrix, $\hat{\phi}_i$, i=1,...,p can be evaluated from eqn. 5.65 and the orthogonalized residual sequence is

$$v_t = \hat{K} \, \varepsilon_t \qquad 5.80$$

The determinant of the residual covariance matrix might also be evaluated more easily from

$$\left| \sum_{i=1}^{t} \varepsilon_i \varepsilon_i^T \right| = \prod_{i=1}^{p} V_t^{(i)^T} V_t^{(i)} \qquad 5.81$$

5.6 Applications of the Algorithm

The algorithm is intended to be used interactively in a manner precisely analogous to that described for RHYP (section 4.3). Similarly, it is quite feasible to implement the algorithm non-recursively

using for example the correlation, and covariance methods described for the SISO algorithm. No such implementation is presented here as the extension is quite obvious.

The major difficulty in applying MVHYP is one of dimensionality. Whilst hypotheses on second order or even third order $1+c(z^{-1})$ polynomials are amenable to interactive analysis, higher order problems require extremely tedious analysis of the cost function, and it is in such instances that a more automated (hill-climbing) approach may be advisable. (This was also outlined for RHYP). However, the concept of reducing the non-linear search to one over only the parameters of the scalar $1+c(z^{-1})$ polynomial is still extremely valuable, as it allows estimates of the $A(z^{-1})$ and $B(z^{-1})$ polynomials to be obtained by linear least squares.

By modelling the system on the basis of a truncated (under-parametrized) $I+C(z^{-1})$ polynomial, the order of the hypothesized polynomials may be reduced. Although the method cannot lead to optimal estimates of the impulse response transfer functions, it certainly will lead to a substantial improvement in results over ordinary least squares. An example of such an approximation is given in the following section.

5.7 Simulation Examples

5.7.1 A Simple Illustrative Example

The 2-input, 2-output system described by the difference equation

$$(I+A_1 z^{-1})y_t = B_1 u_{t-1} + (I+C_1 z^{-1})e_t \qquad 5.82$$

in which

$$A_1 = \begin{bmatrix} -0.9 & 0.5 \\ -0.5 & -0.2 \end{bmatrix} \qquad B_1 = I$$

$$C_1 = \begin{bmatrix} -0.2 & -0.4 \\ 0.2 & -0.8 \end{bmatrix}$$

the driving noise vector e_t has statistics

$$E(e_t) = 0$$

$$E(e_t e_t^T) = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1.09 \end{bmatrix}, \quad E(e_t e_{t+\ell}^T) = 0 \ , \ \ell \neq 0$$

and the measurable (zero mean) input sequence u_t has covariance

$$E(u_t u_t^T) = I$$

was used to generate 300 pairs of input-output data (y_t, u_t).

Following eqn. 5.17, eqn. 5.82 may be reformulated by pre-multiplying

by adj $(I+C_1 z^{-1}) = I + \begin{bmatrix} -0.8 & 0.4 \\ -0.2 & -0.2 \end{bmatrix}$, giving

$$(I+\bar{A}_1 z^{-1}+\bar{A}_2 z^{-2})y_t = (\bar{B}_1 z^{-1}+\bar{B}_2 z^{-2})u_t + |I+C_1 z^{-1}|\ e_t \qquad 5.83$$

where

$$\bar{A}_1 = \begin{bmatrix} -1.7 & 0.9 \\ -0.7 & -0.4 \end{bmatrix} \qquad \bar{A}_2 = \begin{bmatrix} 0.52 & -0.48 \\ 0.28 & -0.06 \end{bmatrix}$$

$$\bar{B}_1 = I \qquad \bar{B}_2 = \begin{bmatrix} -0.8 & 0.4 \\ -0.2 & -0.2 \end{bmatrix}$$

and $$|I+C_1 z^{-1}| = 1-z^{-1}+0.24z^{-2} = 1+c_1 z^{-1}+c_2 z^{-2}$$

Using MVHYP in precisely the same interactive manner as described for RHYP, the quantisation intervals q_1 and q_2 were both finally set to 0.05 in the region

$$\begin{aligned} -1.15 &< c_1 < -0.85 \\ 0 &< c_2 < 0.3 \end{aligned} \qquad 5.84$$

enclosing the minimum of the cost function based on the trace of the residual covariance matrix, $tr(\frac{1}{N}\sum_{t=1}^{300}\varepsilon_t\varepsilon_t^T)$. The parameter estimates obtained are shown in Fig. 5.1 and demonstrate the typically swift convergence of the $1+c(z^{-1})$ polynomial estimates. The final estimates are:

$$\begin{aligned} \hat{A}_1 &= \begin{bmatrix} -1.678 & 0.967 \\ -0.747 & -0.324 \end{bmatrix} \qquad \hat{A}_2 = \begin{bmatrix} 0.455 & -0.495 \\ 0.272 & -0.115 \end{bmatrix} \\ \hat{B}_1 &= \begin{bmatrix} 1.036 & 0.0534 \\ 0.0177 & 1.1057 \end{bmatrix} \qquad \hat{B}_2 = \begin{bmatrix} -0.829 & 0.469 \\ -0.3111 & -0.2017 \end{bmatrix} \\ 1+\hat{c}_1 z^{-1}+\hat{c}_2 z^{-2} &= 1-0.95z^{-1}+0.2z^{-2} \end{aligned} \qquad 5.85$$

These are reasonable estimates of the true values considering the limited data. The covariances of the two residual sequences are:

$$\frac{1}{N}\sum_{t=1}^{300}(\varepsilon_t^{(1)})^2 = 0.9718 \qquad 5.86$$

$$\frac{1}{N}\sum_{t=1}^{300}(\varepsilon_t^{(2)})^2 = 0.9529 \qquad 5.87$$

Eqn. 5.83 is the system description given two correlated noise vector elements $e_t^{(1)}$ and $e_t^{(2)}$. Now, vector e_t may be generated from noise v_t with uncorrelated elements as follows:

$$e_t = \begin{bmatrix} 1 & 0 \\ 0.3 & 1 \end{bmatrix} v_t \qquad 5.88$$

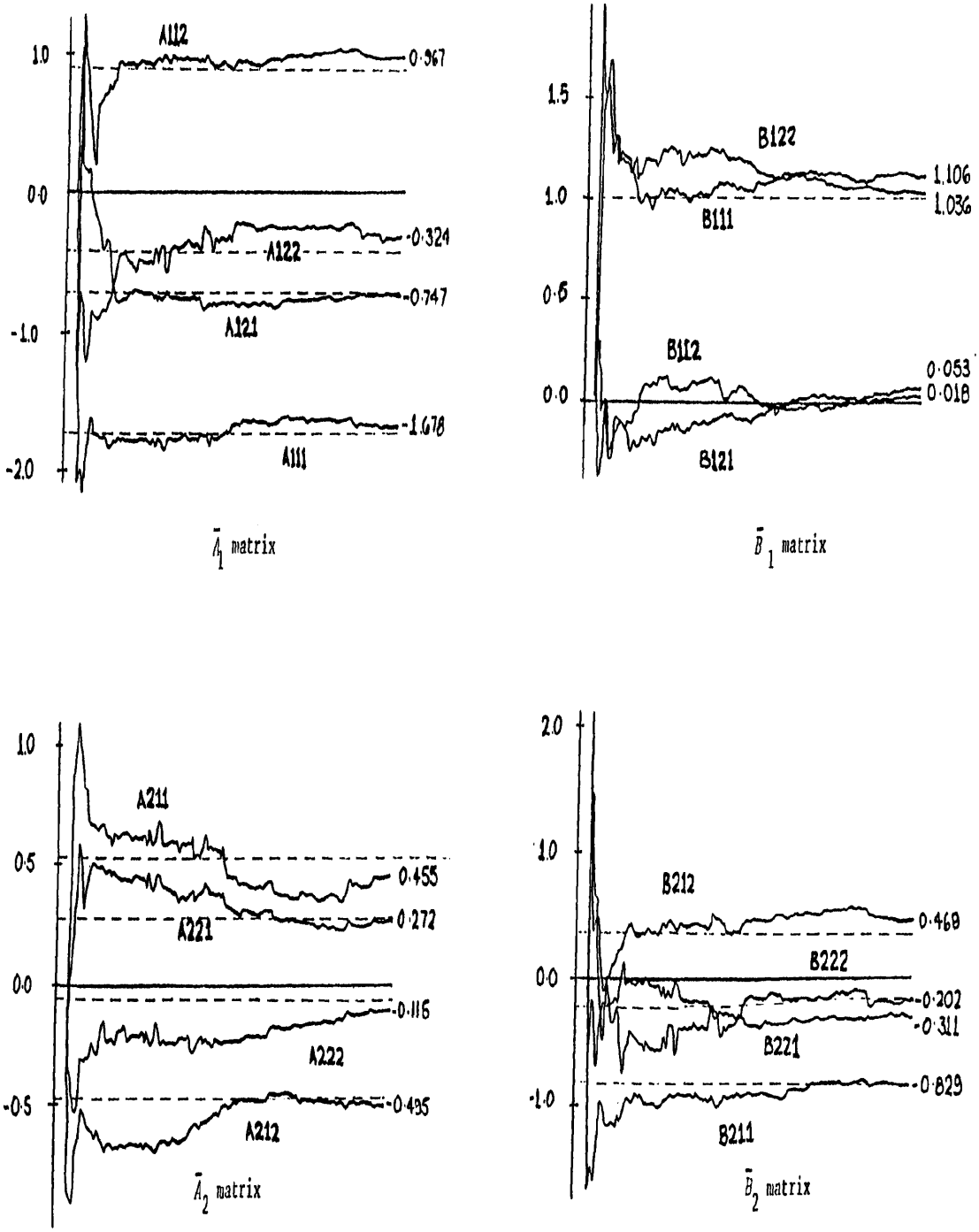


Fig. 5.1(a) Time evolution of the $\bar{A}$ and $\bar{B}$ parameter estimates

Key: Alij = (i,j)th element of Al

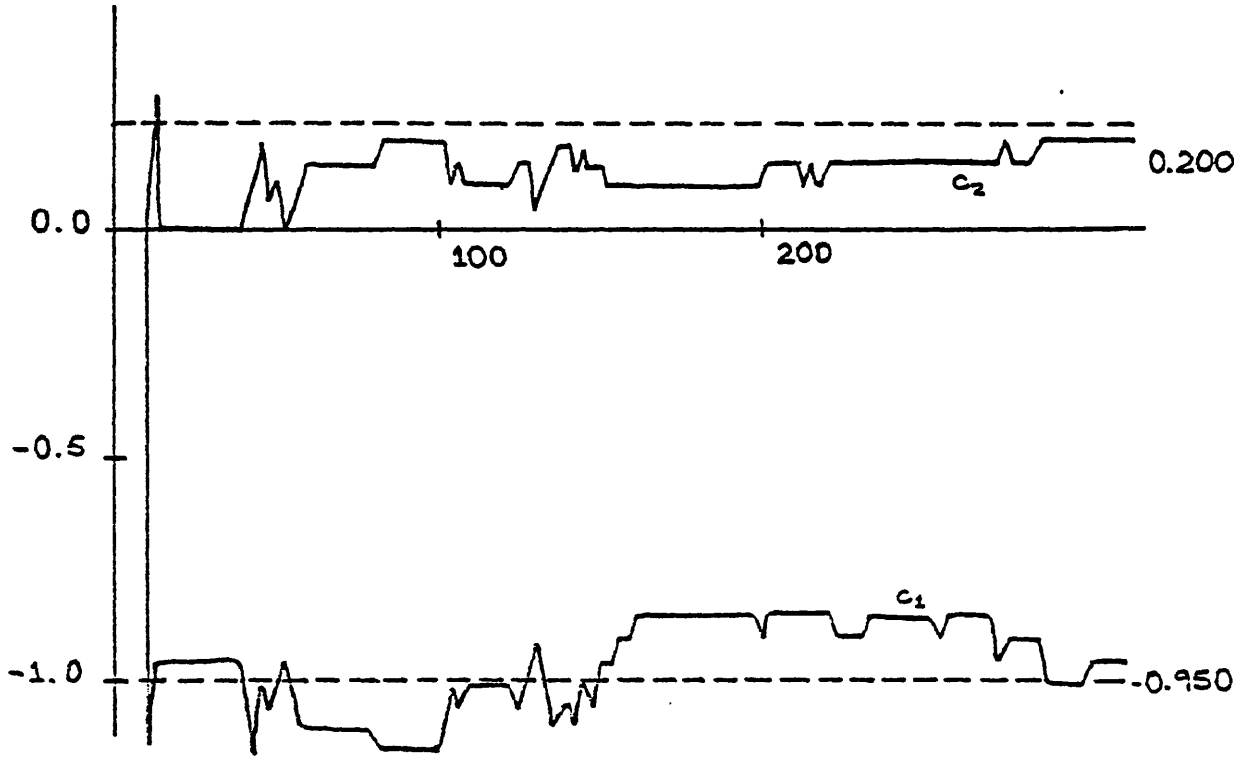


Fig. 5.1b Time evolution of the 'c' parameter estimates (hypotheses)

where

$$E(v_t) = 0$$

$$E(v_t v_t^T) = I$$

$$E(v_t v_{t+\ell}^T) = \phi \ , \quad \ell \neq 0$$

or $$e_t = \bar{K} v_t \qquad 5.89$$

in the notation of eqn. 5.5.

Thus, the system of eqn. 5.82 is represented equivalently by

$$(A_o + A_1 z^{-1} + A_2 z^{-2}) y_t = (B_1 z^{-1} + B_2 z^{-2}) u_t + (1 + c_1 z^{-1} + c_2 z^{-2}) v_t \qquad 5.90$$

as suggested by eqn. 5.12 where

$$A_o = \bar{K}^{-1} = K = \begin{bmatrix} 1 & 0 \\ -0.3 & 1 \end{bmatrix}$$

$$A_i = \bar{K}^{-1} \bar{A}_i \qquad i = 1,2$$

$$B_i = \bar{K}^{-1} \bar{B}_i \qquad i = 1,2$$

Using the method described in section 5.4, the estimate $\hat{K}$ of $K = \bar{K}^{-1}$ was found to be:

$$\hat{K} = \begin{bmatrix} 1 & 0 \\ -0.273 & 1 \end{bmatrix} \qquad 5.91$$

and thus the estimates of the parameters of eqn. 5.88 may be found directly from those given in eqn. 5.85.

When the maximum likelihood estimates were computed, i.e. the cost function used was the determinant of the residual covariance matrix, the optimal hypothesis turned out to be

$$1+\hat{c}_1 z^{-1}+\hat{c}_2 z^{-2} = 1-z^{-1}+0.2z^{-2} \qquad 5.92$$

This is a neighbouring hypothesis to that selected by using the trace cost function, the difference being one quantum in the $\hat{c}_1$ parameter estimate. The final parameter estimates were:

$$\hat{\bar{A}} = \begin{bmatrix} -1.729 & 0.9571 \\ -0.7344 & -0.3774 \end{bmatrix} \quad \hat{\bar{A}} = \begin{bmatrix} 0.5095 & -0.5213 \\ 0.2953 & -0.1064 \end{bmatrix}$$
$$\hat{\bar{B}} = \begin{bmatrix} 1.037 & 0.0529 \\ -0.1871 & 1.1098 \end{bmatrix} \quad \hat{\bar{B}} = \begin{bmatrix} -0.8813 & 0.4519 \\ -0.2906 & -0.2726 \end{bmatrix} \qquad 5.93$$

and the covariances of the two residual sequences are:

$$\frac{1}{N}\sum_{t-1}^{300} (\varepsilon_t^{(1)})^2 = 0.9771 \qquad 5.94$$

$$\frac{1}{N}\sum_{t-1}^{300} (\varepsilon_t^{(2)})^2 = 0.9487 \qquad 5.95$$

The effect of the determinantal cost function is to permit a slight increase in the first residual covariance in exchange for a decrease in that for residual two. The difference is minute.

The residual covariances turn out to be less than the theoretical driving noise covariances only due to the small sample inaccuracy of the noise generator used.

5.7.2 Transforming to a Standard Representation

This example shows how the results obtained for the non-standard parametrization used in MVHYP may be used to produce a model in a more conventional representation. The method follows the procedure outlined in section 5.2 (eqns. 5.18 - 5.20). The above example is continued, with a view to obtaining a model in the form of eqn. 5.82.

Using the parameter estimates in eqn. 5.85, the residual vector sequence

$$\varepsilon_t = (I+\hat{\bar{A}}_1 z^{-1}+\hat{\bar{A}}_2 z^{-2})y_{f_t} - (\hat{\bar{B}}_1 z^{-1}+\hat{\bar{B}}_2 z^{-1})u_{f_t} \qquad 5.96$$

was computed where y_{f_t} and u_{f_t} are the input and output sequences y_t and u_t filtered by the optimum hypothesis $1+\hat{c}_1 z^{-1}+\hat{c}_2 z^{-2}$ in the usual fashion. Now, ε_t is assumed to be an estimate of the driving noise e_t, and the matrices A_1, B_1 and C_1 in eqn. 5.82 are estimated by minimizing $V_w=\text{tr}\sum_{t=1}^{300} w_t w_t^T$ in the model:

$$(I+\hat{A}_1 z^{-1})y_t = \hat{B}_1 z^{-1}u_t + (I+\hat{C}_1 z^{-1})\varepsilon_t + w_t \qquad 5.97$$

The resulting estimates are shown in Fig. 5.2 and the final values are:

$$\hat{A}_1 = \begin{bmatrix} -0.8823 & 0.4867 \\ -0.4872 & -0.2051 \end{bmatrix} \quad \hat{B}_1 = \begin{bmatrix} 1.0235 & 0.05427 \\ 0.02391 & 1.0986 \end{bmatrix}$$

$$\hat{C}_1 = \begin{bmatrix} -0.1540 & -0.4808 \\ 0.2599 & -0.8308 \end{bmatrix} \qquad 5.98$$

These are remarkably good estimates considering the short data record. The 'efficiency' of the transformation may be monitored by observing the cost V_w. Ideally this should be zero if in fact there

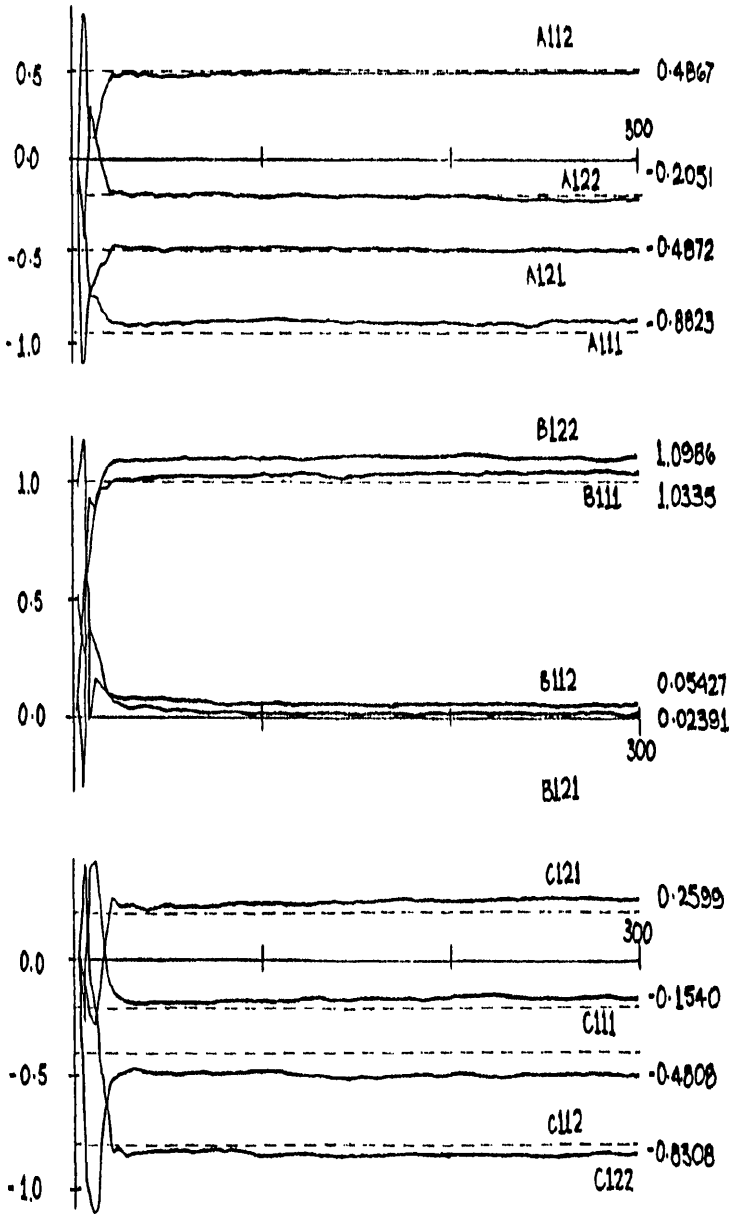


Fig. 5.2 Parameter estimates on transforming back to the standard representation

is a factor adj $\{I+\hat{C}(z^{-1})\}$ of the $(I+\hat{\bar{A}}(z^{-1}))$ and $\hat{\bar{B}}(z^{-1})$ polynomials, and the estimate $1+\hat{c}(z^{-1})$ is the determinant of $I+\hat{C}(z^{-1})$. In this example, the value of the cost function is 0.0527 and this extremely low value indicates that the transformation to the new representation is almost deterministic and exact. This is further corroborated by the swift convergence of the parameter estimates in Fig. 5.2.

5.7.3 MVHYP using an Underparametrized Model

The following discrete time system is now considered:

$$(I+A_1z^{-1}+A_2z^{-2})y_t = (B_1z^{-1}+B_2z^{-2})u_t + (I+C_1z^{-1}+C_2z^{-2})e_t \qquad 5.99$$

where

$$A_1 = \begin{bmatrix} -1.5 & 0.2 \\ -0.3 & -1.1 \end{bmatrix} \qquad A_2 = \begin{bmatrix} 0.6 & -0.1 \\ 0.1 & 0.3 \end{bmatrix}$$

$$B_1 = \begin{bmatrix} 1 & 0.3 \\ -0.5 & 1 \end{bmatrix} \qquad B_2 = \begin{bmatrix} 0.5 & 0.2 \\ -0.4 & 0.4 \end{bmatrix}$$

$$C_1 = \begin{bmatrix} -0.7 & 0 \\ 0.3 & -1.3 \end{bmatrix} \qquad C_2 = \begin{bmatrix} 0.1 & 0 \\ 0.2 & 0.5 \end{bmatrix}$$

u_t is a measurable zero mean white input with covariance

$$E(u_t u_t^T) = I \qquad 5.100$$

and e_t is a white disturbance with statistics

$$E(e_t) = 0 \qquad E(e_t e_t^T) = 0.25\ I \qquad 5.101$$

The object of this example is to show that a prediction model notably better than a least squares model can be found by assuming only a

first order $I+C(z^{-1})$ polynomial (i.e. that $C_2 = 0$) and using MVHYP to estimate the parameters of the special MVHYP method parametrization

$$(I+\hat{\hat{A}}_1 z^{-1}+\hat{\hat{A}}_2 z^{-2}+\hat{\hat{A}}_3 z^{-3})y_t = (\hat{\hat{B}}_1 z^{-1}+\hat{\hat{B}}_2 z^{-2}+\hat{\hat{B}}_3 z^{-3})u_t+(1+\hat{c}_1 z^{-1}+\hat{c}_2 z^{-2})I\varepsilon_t \qquad 5.102$$

300 samples of input-output data were available, and the trace cost function was used.

The prediction error variances are compared with those achieved when ordinary least squares is used to estimate the parameters of the model:

$$(I+\hat{A}_1 z^{-1}+\hat{A}_2 z^{-2})y_t = (\hat{B}_1 z^{-1}+\hat{B}_2 z^{-2})u_t + \varepsilon_t \qquad 5.103$$

Again the trace of the residual covariance matrix is minimized.

Using MVHYP, the minimum of the cost function was found to be satisfactorily bounded by:

$$\begin{aligned} -1.45 \;&<\; \hat{c}_1 \;<\; -1.05 \\ 0.35 \;&<\; \hat{c}_2 \;<\; 0.6 \end{aligned} \qquad 5.104$$

where a quantization of 0.05 was chosen for both $\hat{c}_1$ and $\hat{c}$. The sample variances of the residuals were found to be

$$\begin{aligned} \text{Var}\ (\varepsilon_t^{(1)}) &= 0.244 \\ \text{Var}\ (\varepsilon_t^{(2)}) &= 0.352 \end{aligned} \qquad 5.105$$

as against the variances

$$\begin{aligned} \text{Var}\ (\varepsilon_t^{(1)}) &= 0.33 \\ \text{Var}\ (\varepsilon_t^{(2)}) &= 0.69 \end{aligned} \qquad 5.106$$

obtained for the least squares model of eqn. 5.103. The MVHYP estimator clearly produces a better predictor. The comparison may be somewhat unfair in that 3rd order $I+\hat{\bar{A}}(z^{-1})$ and $\hat{\bar{B}}(z^{-1})$ polynomials have been used for MVHYP, whereas only 2nd order polynomials have been used for the least squares comparison. Using 3rd order polynomials the least squares residual variances are reduced to

$$\begin{aligned} \text{Var } (\varepsilon_t^{(1)}) &= 0.267 \\ \text{Var } (\varepsilon_t^{(2)}) &= 0.414 \end{aligned} \qquad 5.107$$

which are still significantly higher than the MVHYP method results.

The theoretical lower bounds on the prediction error variances are 0.25 for both $\text{Var}(\varepsilon_t^{(1)})$ and $\text{Var } (\varepsilon_t^{(2)})$. The characteristics of the noise generator are of course not ideal, but the variance of $\varepsilon_t^{(1)}$ given in eqn. 5.105 is clearly very nearly optimum as is borne out in Fig. 5.3 in which $\varepsilon_t^{(1)}$ is superimposed on the true driving noise $e_t^{(1)}$. Fig. 5.3 also shows that $\varepsilon_t^{(2)}$ is a fair estimate of $e_t^{(2)}$.

5.8 Conclusion

This chapter has demonstrated how the concepts of parameter estimation by hypothesis testing and the principles of computer interaction embodied in the RHYP algorithm may be extended to the multivariable case. Two prediction error cost functions may be minimized, namely the trace and the determinant of the residual covariance matrix, using the computationally simple least squares method. However, the method is complicated by the necessity to use a

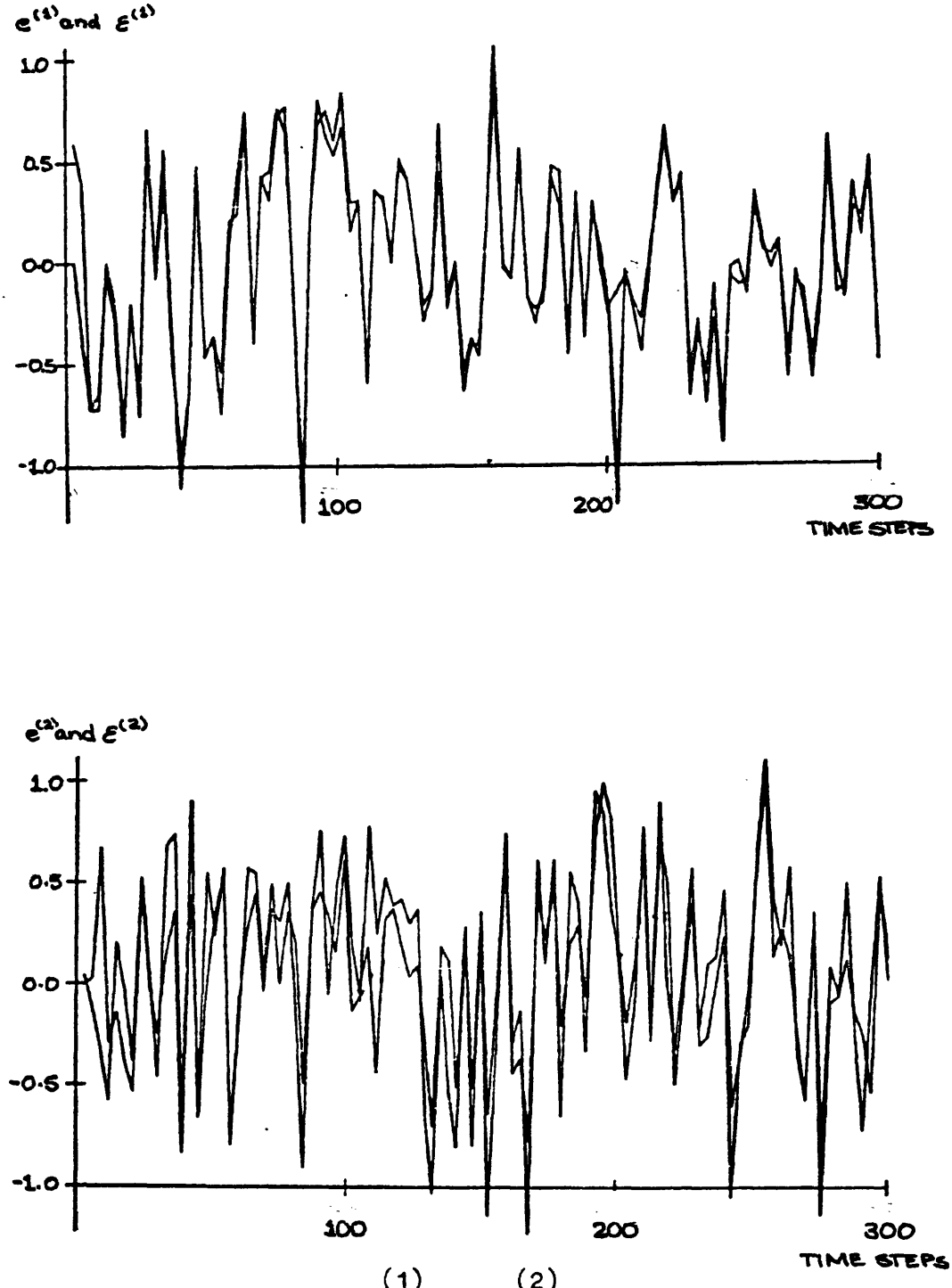


Fig. 5.3 System noise ($e^{(1)}$ and $e^{(2)}$) superimposed on the residual sequences ($\varepsilon^{(1)}$ and $\varepsilon^{(2)}$)

special, non-minimal representation, and soon becomes unwieldy both with increasing number of outputs and with order of the noise colouration matrix polynomial $I+C(z^{-1})$. It may therefore be necessary to embed the algorithm in a hill-climbing technique in which the hill-climber iterates upon the noise colouration polynomial hypothesis. The resulting algorithm is nevertheless a considerable simplification of the direct approach in which the parameters of the $I+A(z^{-1})$ and $B(z^{-1})$ polynomials are also included in the search.

The examples illustrate the characteristically swift convergence properties of the hypothesis testing approach, and also show that it is possible to easily transform the parametrization to a more conventional and concise one. Finally, it is shown that an approximation involving an under-parametrized estimate of $I+C(z^{-1})$ still offers significant improvement in performance over ordinary least squares estimation.

It is therefore hoped that the algorithm may prove useful in certain practical applications where the requirement for a minimal, optimal solution, may give way to this more pragmatic approach which nevertheless provides sensible and reliable system models.

CHAPTER 6

CONTROL STRATEGIES FOR SELF TUNING CONTROLLERS

6.1 Introduction

The advent of fast, powerful, and relatively inexpensive microcomputers has made it possible to offer digital controllers as a cost-effective alternative to conventional analogue types. The flexibility of a computer implementation makes it easier to introduce more complex control strategies, and largely remove constraints on controller order so that high order controllers may be implemented with little difficulty. In many systems, digital control is not only attractive for its own sake but because the controller can be embedded in a software structure which facilitates system monitoring and permits 'intelligent' decisions to be made in the areas of fault detection, fault correction and safety.

Classical control design techniques (for analogue controllers) are today well-established. They are largely based on a system description which may be measured reasonably easily, namely frequency response. In particular, the Bode Plot, Nyquist diagram and Nichols chart forms have proved (and still prove) extremely valuable. (see for example classic texts such as D'Azzo and Houpis,1966, Shinners, 1972 etc.)

Whilst digital controllers can be designed from models parametrized in the continuous s-domain, or as continuous time state space models, it makes sense to consider a direct approach

in which design proceeds from a discrete time system model. The ARMAX model in particular is an extremely attractive such representation and its parameters may be estimated by a variety of methods, for example, the RHYP and MVHYP techniques presented earlier in this thesis. Such analysis has the advantage over frequency response methods that it can be effected from normal operating records. Clearly a good model of the system is fundamental to the controller design procedure. The reader is referred to the standard works of Tou (1959) Kuo (1970), Ragazzini and Franklin (1958) and Saucedo and Schiring (1968) for an exposition of general design techniques. The linear quadratic problem, and linear stochastic control theory for discrete time systems is considered in works by, for example Astrom (1970) and Bryson and Ho (1969).

Whilst acknowledging the usefulness of this 'standard' approach to digital controller design, the following chapters deviate from this path, and instead explore the possibilities of Self Tuning Control. This class of controllers exploits the available computing power to the full by blending the concept of on-line estimation with that of on-line controller design. The attractions and background to this 'learning algorithm' approach have already been covered in Chapter 2. The purpose of this chapter is to bridge the gap between the conventional and self tuning approaches to digital control by reviewing in more detail those controllers which lend themselves to self tuning, and indeed, to extend this class by introducing the multivariable detuned minimum variance and pole-shifting regulators. The self tuning controllers are not discussed at this stage. These will be covered in Chapter 7. Thus, the basic assumption now is that an ARMAX model of the system to be controlled is available, the

parameters of which might perhaps have been estimated using a maximum likelihood or other estimator, and the controller design is carried out off-line.

Two distinct approaches both based on output feedback are considered. The first consists of methods based on optimality criteria and includes the minimum variance regulator (Astrom and Wittenmark, 1973; Borisson, 1975) presented for the multivariable system case, and the generalized minimum variance controller (Clarke and Gawthrop, 1975; Gawthrop,1977). The latter strategy is treated only for the single input single output case, and 'generalises' the minimum variance control by seeking to minimize not only the output variance, but a weighted sum of output and input variances. This method is not developed further here, but is a valuable technique which is included only for completeness. The multivariable minimum variance regulator is however extended by introducing the multivariable detuned minimum variance regulator in which the placement of closed loop poles (with fairly severe restrictions) is possible. The SISO version of this regulator is due to Edmunds (1976).

The second approach presented for multivariable systems, is novel for the self tuning application, and permits the specification of closed loop system poles. The method is developed from the SISO system work of Edmunds (1976).

In all cases, the plant, assumed to be controllable and observable, is represented in the ARMAX form

$$\{I+A(z^{-1})\}y_t = z^{-k}B(z^{-1})u_t + \{I+C(z^{-1})\}e_t \qquad 6.1$$

where y_t and u_t are p-vector output and input respectively and

e_t is a zero mean white noise p-vector with covariance Q. The polynomial matrices $A(z^{-1})$, $B(z^{-1})$ and $C(z^{-1})$ follow previously defined conventions (see for example eqn.5.2). Furthermore, $I+A(z^{-1})$ and $B(z^{-1})$ are relatively left prime and the zeros of $|I+C(z^{-1})|$ lie within the z-plane unit disc.

6.2 Minimum Variance Regulators

The minimum variance regulator seeks to minimize the cost function $V = E(y_{t+k+1}^T y_{t+k+1})$ 6.2

The cost function is a special case of the linear quadratic strategy, the criterion being a quadratic function of the output only, and does not require the solution of a Riccati equation. The solution to this problem for SISO systems (formulated using the ARMAX model) is well known and has been discussed by, for example, Astrom (1970). The cost function 6.2 is minimized with respect to the control input u_t, which may be a function only of known data, present and past outputs $y_t, y_{t-1}, \ldots,$ and past inputs $u_{t-1}, u_{t-2}, \ldots,$ all of which are known at time t.

The solution for the multivariable case has been documented by Borisson (1975). The fundamental difference between the multivariable and single variable analysis stems from the fact that matrices do not commute. The derivation will not be given here, but emerges as a special case of the solution to the 'Detuned Minimum Variance' regulator discussed in section 6.3. Therefore, it suffices at this stage to state the control law:

$$\{I+\tilde{M}(z^{-1})\}\{zB(z^{-1})\}u_t = G(z^{-1})y_t \qquad 6.3$$

where, as an exception to the usual polynomial definition (since $G_o \neq 0$)

$$G(z^{-1}) = G_o + G_1 z^{-1} + \ldots + G_{n_g} z^{-n_g} \qquad 6.4$$

and $$\{I+C(z^{-1})\} = \{I+A(z^{-1})\}\{I+M(z^{-1})\} - z^{-k-1}\tilde{G}(z^{-1}) \qquad 6.5$$

where $$\{I+\tilde{M}(z^{-1})\}\{\tilde{G}(z^{-1})\} = \{G(z^{-1})\}\{I+M(z^{-1})\} \qquad 6.6$$

$$\det\{I+\tilde{M}(z^{-1})\} = \det\{I+M(z^{-1})\} \qquad 6.7$$

and $M(z^{-1})$ follows the standard polynomial convention employed and is of order k. $\tilde{G}(z^{-1})$ is similar in form to $G(z^{-1})$ and is of order n_a-1.

It can be shown that polynomial matrices $\tilde{M}(z^{-1})$ and $G(z^{-1})$ always exist, but are not unique (Borisson, 1975; Wolovich,1974). It is often possible to specify that the orders of $\tilde{M}(z^{-1})$ and $G(z^{-1})$ should be equal to those of $M(z^{-1})$ and $\tilde{G}(z^{-1})$ respectively. Under this constraint, eqn. 6.6 implies the solution of the following set of linear simultaneous equations:

$$G_o = \tilde{G}_o$$

$$\begin{array}{l} \\ k+n_g \text{ block rows} \end{array} \overbrace{\begin{bmatrix} I & & & -\tilde{G}_o^T & & \\ M_1^T & I & & -\tilde{G}_1^T & & -\tilde{G}_o^T \\ & M_1^T & & \vdots & & \vdots \\ \vdots & \vdots & & -\tilde{G}_{n_g}^T & & \\ M_k^T & & & & & \\ & M_k^T & & & & -\tilde{G}_{n_g}^T \end{bmatrix}}^{\begin{array}{c} n_g \text{ block} \\ \text{columns} \end{array} \quad \overbrace{\qquad\qquad}^{\begin{array}{c} k \text{ block} \\ \text{columns} \end{array}}} \begin{bmatrix} G_1^T \\ \vdots \\ G_{n_g}^T \\ \tilde{M}_1^T \\ \vdots \\ \tilde{M}_k^T \end{bmatrix} = \begin{bmatrix} \tilde{G}_1^T \\ \vdots \\ \tilde{G}_{n_g}^T \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} M_1^T \\ \vdots \\ M_k^T \\ 0 \\ \vdots \\ 0 \end{bmatrix} G_o^T \qquad 6.8$$

This clearly requires that the $(k+n_g)p \times (k+n_g)p$ matrix on the left hand side of eqn. 6.8 be non-singular. As in general the initial model

parameters are obtained by estimation it is unlikely that a rank deficiency will exist in practice. Of course, for SISO systems eqns. 6.6 and 6.7 are not required as then

$$\tilde{M}(z^{-1}) = M(z^{-1}) \quad ; \quad \tilde{G}(z^{-1}) = G(z^{-1})$$

Under minimum variance control (eqn. 6.3) the system output is given by

$$y_t = \{I+M(z^{-1})\}e_t \qquad 6.9$$

There are a number of other factors to bear in mind when using minimum variance control namely:

i) In order to solve for the input u_t in eqn. 6.3, B_1 (the first coefficient matrix of polynomial $B(z^{-1})$) must in general be non-singular. There are exceptions to this rule, and an example is given in section 6.6.4.

ii) The determinant of $B(z^{-1})$ should be inverse stable. Unstable zeros of $\det(B(z^{-1}))$, corresponding to o.d. zeros, lead to unstable modes of the closed loop system which can only be cancelled out when $B(z^{-1})$ in the control law is identified exactly. Therefore, in any practical situation, if $\det\{B(z^{-1})\}$ is not inverse stable, the closed loop system will be unstable.

iii) The roots of $\det\{I+C(z^{-1})\}$ must lie within the stability region.

The minimum variance regulator has been found on occasions to result in rather large control excursions. These are sometimes undesirable as they lead to saturation effects either during digital to analogue conversion, or in the plant itself, and contribute to system wear. A simple modification is to reformulate the minimum variance problem making provision for the specification of closed

loop poles. Edmunds (1976) has suggested the technique for SISO systems and the extension for multivariable systems is discussed below.

6.3 Detuned Minimum Variance

The aim of this regulator is to preserve the zeros of the closed loop transfer function obtained under minimum variance regulation but to introduce closed loop poles determined by the zeros of the determinant of a specified matrix polynomial $I+z^{-k}T(z^{-1})$. The desired closed loop system is thus

$$\{I+z^{-k}T(z^{-1})\}y_t = \{I+M(z^{-1})\}e_t \qquad 6.10$$

where $T(z^{-1}) = T_1 z^{-1} + \ldots + T_{n_t} z^{-n_t}$

and $\{I+M(z^{-1})\}$ is that given in eqn. 6.9.

The system description (eqn. 6.1) must first be manipulated into a suitable form. Following eqn. 6.5, let

$$\{I+C(z^{-1})\} = \{I+A(z^{-1})\}\{I+M(z^{-1})\} - z^{-k-1}\tilde{G}(z^{-1}) \qquad 6.11$$

where $M(z^{-1})$ is of order k

and define $\tilde{M}(z^{-1})$ and $G(z^{-1})$ according to:

$$\{I+\tilde{M}(z^{-1})\}\tilde{G}(z^{-1}) = G(z^{-1})\{I+M(z^{-1})\} \qquad 6.12$$

where $\det\{I+M(z^{-1})\} = \det\{I+\tilde{M}(z^{-1})\}$

and $\det\{G(z^{-1})\} = \det\{\tilde{G}(z^{-1})\}$ (c.f. eqn. 6.6)

Now, define

$$I+\tilde{C}(z^{-1}) = \{I+\tilde{M}(z^{-1})\}\{I+A(z^{-1})\} - z^{-k-1}G(z^{-1}) \qquad 6.13$$

By postmultiplying eqn. 6.13 by $\{I+M(z^{-1})\}$, premultiplying eqn. 6.11 by $\{I+\tilde{M}(z^{-1})\}$ and using eqn. 6.12 it is evident that

$$\{I+\tilde{M}(z^{-1})\}\{I+C(z^{-1})\} = \{I+\tilde{C}(z^{-1})\}\{I+M(z^{-1})\} \qquad 6.14$$

and since $\det\{I+\tilde{M}(z^{-1})\} = \det\{I+M(z^{-1})\}$

it follows that $\det\{I+C(z^{-1})\} = \det\{I+\tilde{C}(z^{-1})\}$ 6.15

Postmultiplying eqn. 6.13 by $I+z^{-k}T(z^{-1})$ gives

$$\{I+\tilde{C}(z^{-1})\}\{I+z^{-k}T(z^{-1})\} = \{I+\tilde{M}(z^{-1})\}\{I+A(z^{-1})\} - z^{-k-1}L(z^{-1}) \qquad 6.16$$

where $L(z^{-1}) = G(z^{-1}) - z\{I+\tilde{C}(z^{-1})\}\{T(z^{-1})\}$

$= L_o+L_1z^{-1}+ \ldots + L_{n_\zeta}z^{-n_\zeta}$

Now, by premultiplying the system eqn. 6.1 by $I+\tilde{M}(z^{-1})$, substituting eqn. 6.16, and using relationship 6.14, it follows that

$$\{I+\tilde{C}(z^{-1})\}[\{I+z^{-k}T(z^{-1})\}y_t - \{I+M(z^{-1})\}e_t]$$
$$= z^{-k-1}[\{I+\tilde{M}(z^{-1})\}\{zB(z^{-1})\}u_t - L(z^{-1})y_t] \qquad 6.17$$

If it is assumed that $\det[I+C(z^{-1})]$ (and hence $\det[I+\tilde{C}(z^{-1})]$) is inverse stable, it is clear from eqn. 6.17 that by choosing the control law

$$\{I+\tilde{M}(z^{-1})\}\{zB(z^{-1})\}u_t = L(z^{-1})y_t \qquad 6.18$$

the closed loop system response will be

$$\{I+z^{-k}T(z^{-1})\}y_t = \{I+M(z^{-1})\}e_t \qquad 6.19$$

as desired. The same conditions, given in (i), (ii), (iii) of section 6.2 of course apply.

By setting $T(z^{-1}) = 0$, it follows that $L(z^{-1}) = G(z^{-1})$, and

the control law reduces to ordinary minimum variance. From eqn. 6.17 it is also possible to demonstrate why this law minimizes the system output variance. Re-writing eqn. 6.17,

$$y_{t+k+1} = \{I+M(z^{-1})\}e_{t+k+1} + f\{y_t, y_{t-1}, \ldots, u_t, u_{t-1}, \ldots\}$$

$$= \quad y^a \quad + \quad y^b \qquad 6.20$$

(where f(.) indicates some function of the past data)

Now, since the signal $y^a = \{I+M(z^{-1})\}e_{t+k+1}$ is uncorrelated with data at, and prior to time t,

$$E\{y^T_{t+k+1}y_{t+k+1}\} = E\{(y^a)^T y^a\} + E\{(y^b)^T y^b\} \qquad 6.21$$

Since the choice of u_t cannot affect the term $E\{(y^a)^T y^a\}$, the cost function is minimized simply by setting $y^b = 0$.

The effect of introducing the auto-regressive terms defined by $\{I+z^{-k}T(z^{-1})\}$ is best illustrated by a simple SISO example.

Example 1:

Let the system be described by

$$(1+a_1z^{-1})y_t = b_1u_{t-1} + e_t \qquad 6.22$$

where e_t is a zero mean white driving noise with variance σ^2.
Then the minimum variance regulator is given by

$$u_t = \frac{a_1}{b_1} y_t \ , \quad b_1 \neq 0 \qquad 6.23$$

The detuned minimum variance regulator for $1+T(z^{-1}) = 1+t_1z^{-1}$ is

$$u_t = \frac{a_1-t_1}{b_1} y_t \ , \quad b_1 \neq 0 \qquad 6.24$$

If the regulator gain is denoted g, so that $u_t = gy_t$, then by substituting into eqn. 6.22, the closed loop response is

$$(1 + (a_1-b_1g)z^{-1})y_t = e_t \qquad 6.25$$

The minimum variance regulator eliminates the closed loop pole at $z = -(a_1-b_1g)$. The detuned minimum variance regulator moves the pole to $z = -t_1$, and if t_1 is suitably chosen, will have reduced gain. The output variance however, increases from σ^2 to $\frac{\sigma^2}{1-t_1^2}$.

Note that inappropriate choice of $T(z^{-1})$ can increase the regulator gain.

The following example illustrates the calculation of minimum variance and detuned minimum variance control laws for a multivariable system.

Example 2

Consider the multivariable system described by

$$\{I+A_1z^{-1}+A_2z^{-2}\}y_t = z^{-k}\{B_1z^{-1}+B_2z^{-2}\}u_t + \{I+C_1z^{-1}\}e_t \qquad 6.26$$

where $k = 1$

$$A_1 = \begin{bmatrix} 0.8 & -2.9 \\ 0.6 & -1.4 \end{bmatrix} \qquad A_2 = \begin{bmatrix} 0.2 & -1.75 \\ 0.2 & -0.95 \end{bmatrix}$$

$$B_1 = I \qquad B_2 = \begin{bmatrix} -0.4 & 0.6 \\ 0.4 & 0.6 \end{bmatrix}$$

$$C_1 = \begin{bmatrix} -0.25 & 0 \\ 0 & -0.45 \end{bmatrix}$$

Then, from eqn. 6.11,

$$M_1 = C_1 - A_1 = \begin{bmatrix} -1.05 & 2.9 \\ -0.6 & 0.95 \end{bmatrix} \qquad 6.27$$

$$\tilde{G}_o = [A_2 + A_1 M_1] = \begin{bmatrix} 1.1 & -2.185 \\ 0.41 & -0.54 \end{bmatrix}$$

$$\tilde{G}_1 = [A_2 \; M_1] = \begin{bmatrix} 0.84 & -1.0825 \\ 0.36 & -0.3225 \end{bmatrix} \qquad 6.28$$

Now, from eqn. 6.8, $G_o = \tilde{G}_o$

$$\begin{bmatrix} I & -\tilde{G}_o^T \\ M_1^T & -\tilde{G}_1^T \end{bmatrix} \begin{bmatrix} G_1^T \\ \tilde{M}_1^T \end{bmatrix} = \begin{bmatrix} \tilde{G}_1^T - M_1^T G_o^T \\ 0 \end{bmatrix} \qquad 6.29$$

Hence $$\tilde{M}_1 = \begin{bmatrix} -1.6113 & 3.6485 \\ -0.871 & 1.5113 \end{bmatrix}$$

and $$G_1 = \begin{bmatrix} 0.4074 & -0.6462 \\ 0.1289 & 0.08847 \end{bmatrix}$$

Then the minimum variance control law becomes:

$$z\{I + \tilde{M}(z^{-1})\}B(z^{-1})u_t = G(z^{-1})y_t$$

$$\text{or} \quad \{I + \begin{bmatrix} -2.0113 & 4.2484 \\ -0.471 & 2.111 \end{bmatrix} z^{-1} + \begin{bmatrix} 2.1038 & 1.222 \\ 0.9529 & 0.3842 \end{bmatrix} z^{-2} \}u_t$$

$$= \{ \begin{bmatrix} 1.1 & -2.185 \\ 0.41 & -0.54 \end{bmatrix} + \begin{bmatrix} 0.4074 & -0.6462 \\ 0.1289 & 0.08847 \end{bmatrix} z^{-1} \}y_t \qquad 6.30$$

The detuned minimum variance law is found by a simple modification of the above procedure. Polynomial matrix $G(z^{-1})$ is replaced by $L(z^{-1}) = G(z^{-1}) - \{I + \tilde{C}(z^{-1})\}\{zT(z^{-1})\}$ 6.31

where $I+\tilde{C}(z^{-1}) = \{I+\tilde{M}(z^{-1})\}\{I+A(z^{-1})\} - z^{-k-1}G(z^{-1})$

$$= I + \begin{bmatrix} -0.8113 & 0.7485 \\ -0.271 & 0.1113 \end{bmatrix} z^{-1} \qquad 6.32$$

Thus if $I+z^{-1}T(z^{-1})$ is chosen to be

$$I + z^{-1}\begin{bmatrix} -0.7 & 0 \\ 0 & -0.7 \end{bmatrix} z^{-1} \qquad 6.33$$

$L(z^{-1})$ becomes:

$$L(z^{-1}) = \begin{bmatrix} 1.8 & -2.185 \\ 0.41 & 0.16 \end{bmatrix} + \begin{bmatrix} -0.05951 & -0.12225 \\ -0.0608 & 0.16638 \end{bmatrix} z^{-1} \qquad 6.34$$

and the above polynomial replaces that on the right hand side of eqn. 6.30 to give a detuned minimum variance regulator.

6.4 The Optimal Predictor

It is appropriate at this point to mention the close relationship that necessarily exists between the optimal minimum variance control strategy and optimal prediction. Borisson (1975) has shown that the optimal k+1-step ahead predictor of y_{t+k+1} given data $\{y_t, y_{t-1}, y_{t-2}, \ldots, u_t, u_{t-1}, \ldots\}$ denoted by $\hat{y}_{t+k+1/t}$ is given by

$$\hat{y}_{t+k+1/t} = y_{t+k+1} - \{I+M(z^{-1})\}e_{t+k+1} \qquad 6.35$$

and for $T(z^{-1}) = 0$ satisfies the identity (from eqn. 6.17)

$$\{I+\tilde{C}(z^{-1})\}\hat{y}_{t+k+1/t} = \{I+M(z^{-1})\}\{zB(z^{-1})\}u_t - L(z^{-1})y_t \qquad 6.36$$

where the definitions previously assigned to all the above variables are preserved. Clearly the prediction error $\{I+M(z^{-1})\}e_{t+k+1}$ is

precisely equal to the minimum variance regulation error. The action of the minimum variance regulator is to set the k+1-step ahead predictor of y_{t+k+1} to zero.

6.5 Generalized Minimum Variance

The minimum variance strategy is sometimes found to be unsuitable in two ways:

i) Control inputs required to attain optimal control may become too vigorous

ii) Non-minimum phase systems, that is systems for which the zeros of $\det\{B(z^{-1})\}$ lie outside the z-domain unit disc, cannot be controlled.

The detuned minimum variance approach is to some extent useful in overcoming the first disadvantage. The second difficulty is more prevalent than might initially be expected. This is because systems which may be minimum phase in the continuous s-domain may transform to non-minimum phase systems in the z-domain. Furthermore, pure time delays which are a non-integer multiple of the sampling period (for example computational time delays) can readily lead to non-minimum phase systems. It is possible to use a sub-optimal variant of the minimum variance regulator to control such systems. Such a controller was described by Astrom and Wittenmark (1974) for SISO systems and requires the factorization of the $B(z^{-1})$ polynomial and the solution of a polynomial identity to derive the controller parameters. Its complexity is a disadvantage.

Recognizing the disadvantages of ordinary minimum variance regulation, Clarke and his co-workers (Clarke and Gawthrop, 1975; Gawthrop,1977) have developed another, simple, approach which allows control effort to be adjusted, and which in some instances can cater for non-minimum phase systems. In addition, it provides a means of specifying the closed loop transfer function and permits the introduction of set points. The fundamental facts are summarized here. Since a SISO system is being considered, the polynomials (e.g. $A(z^{-1})$, $B(z^{-1})$ and $C(z^{-1})$) and signals (e.g. y_t, u_t and e_t) are scalar in this section.

An auxiliary system with output ϕ_t is defined by

$$\phi_{t+k+1} = Py_{t+k+1} + Qu_t - Rw_t \qquad 6.37$$

where P, Q, and R are rational transfer functions in the backward shift operator z^{-1}. w_t is the desired value or setpoint of y_t. From eqn. 6.37 and the ARMAX system eqn. 6.1, written below in scalar form as

$$\{1+A(z^{-1})\}y_t = z^{-k}B(z^{-1})u_t + \{1+C(z^{-1})\}e_t \qquad 6.38$$

the auxiliary output can be expressed as

$$\phi_{t+k+1} = \{\frac{P\{zB(z^{-1})\}}{1+A(z^{-1})} + Q\}u_t + \frac{P\{1+C(z^{-1})\}}{1+A(z^{-1})}e_{t+k+1} - Rw_t \qquad 6.39$$

which, apart from the term $(-Rw_t)$ is of the same form as eqn. 6.38

The control strategy is to minimize the auxiliary system output variance, $E(\phi^2_{t+k+1})$. As is to be expected, the controller minimizing the cost function can be found in much the same way as the regular minimum variance problem discussed previously. In fact, the optimal k+1-step ahead predictor of ϕ_{t+k+1} is given by

$$\hat{\phi}_{t+k+1/t} = \phi_{t+k+1} - (1+M(z^{-1}))e_{t+k+1} \qquad 6.39$$

satisfying

$$\{1+C(z^{-1})\}\hat{\phi}_{t+k+1/t} = \frac{1}{P_D}G(z^{-1})y_t + \{1+M(z^{-1})\}\{zB(z^{-1})\}u_t + \{1+C(z^{-1})\}\{Qu_t - Rw_t\} \qquad 6.40$$

where $$\{1+C(z^{-1})\}P_N = \{1+M(z^{-1})\}\{1+A(z^{-1})\}P_D + z^{-k-1}G(z^{-1}) \qquad 6.41$$

$$G(z^{-1}) = G_o + G_1 z^{-1} + \ldots + G_{n_g} z^{-n_g} \quad ; \quad n_g = n_a + n_{p_D} - 1$$

$$P = \frac{P_N}{P_D}$$

and $M(z^{-1})$ is a kth order polynomial.
The associated prediction error is $\{1+M(z^{-1})\}e_t$.

As for the minimum variance regulator, the optimal controller is that which sets the k+1-step ahead predictor to zero at each instant. If $1+C(z^{-1})$ has all its roots within the stability region, the required control law is

$$\left[\{1+M(z^{-1})\}\{zB(z^{-1})\} + \{1+C(z^{-1})\}Q\right]u_t = \{1+C(z^{-1})\}Rw_t - \frac{1}{P_D}G(z^{-1})y_t \qquad 6.42$$

resulting in the auxiliary system response

$$\phi_t = \{1+M(z^{-1})\}e_t \qquad 6.43$$

There are numerous interpretations of this control strategy. When Q=0, the system response is

$$y_{t+k+1} = \frac{1}{P}\{1+M(z^{-1})\}e_{t+k+1} + \frac{R}{P}w_t \qquad 6.44$$

Therefore, the output can be made to follow a given model $\frac{R}{P} w_t$ plus a closed loop disturbance term $\frac{1}{P}\{1+M(z^{-1})\}e_t$, subject to the available control effort. When P=1 and R=0, the controller reduces to the minimum variance regulator. This strategy works satisfactorily when the open loop system is minimum phase. However, as the controller poles are used to cancel system zeros, this model following scheme fails for non-minimum phase systems.

Non-minimum phase systems may be controlled by introducing a suitably chosen input weighting transfer function Q in the auxiliary system. The closed loop system then becomes:

$$y_t = \frac{z^{-k}B(z^{-1})P_D Q_D R}{\{zB(z^{-1})\}P_N Q_D + P_D\{1+A(z^{-1})\}Q_N} w_t$$

$$+ \frac{[\{1+M(z^{-1})\}\{zB(z^{-1})\}Q_D + \{1+C(z^{-1})\}Q_N]P_D}{\{zB(z^{-1})\}P_N Q_D + P_D\{1+A(z^{-1})\}Q_N} e_t \qquad 6.45$$

Cancellation of $B(z^{-1})$ is now avoided, but $Q = \frac{Q_N}{Q_D}$ must be chosen so that the closed loop system is in fact stable.

In its simplest form, when P=1, R=1, and Q=λ, where λ is a positive constant, the controller may be interpreted as minimizing the cost function

$$E[(y_{t+k+1} - w_t)^2 + \lambda u_t^2] \qquad 6.46$$

λ may thus be viewed as a weighting factor in the cost function, weighting the control effort at the expense of output variance.

6.6 The Pole Shifting Regulator

The methods discussed so far have centred on variants of optimal control (or regulation) schemes. The pole-shifting technique represents an entirely different approach to the special class of design rules considered in this chapter, and has its roots in classical control rather than optimal design. It has three distinct advantages over the minimum variance approach in that:

i) It can regulate non-minimum phase systems without exception

ii) The control signal is generally less vigorous

iii) Pure time delays may, under certain conditions, vary between loops (i.e. B_1 need not be non-singular). This is significant as the minimum variance controller requirement that all loops must in general have the same time delay is not always fulfilled by real systems.

Against this must be weighed the greater complexity in computing the controller parameters and the disadvantage of 'non-optimality'.

The advantage of pole placement over the generalized minimum variance rule (for which only a SISO version has as yet been published) is the ease with which non-minimum phase systems may be controlled. There is, for example, no need to carefully select parameters such as the 'P' and 'Q' polynomials in eqn. 6.37 to guarantee stability for non-minimum phase systems.

The essential feature of the pole-shifting law then is that the poles of the closed loop system under regulation may be specified by the designer. The closed loop zeros are decided by the poles of the feedback law and are not subject to selection. This restriction

however, allows the complexity of the design rule computations to be kept to a minimum.

6.6.1 The Design Rule

The regulator design procedure for multivariable systems defined by eqn. 6.1 is now presented. The orders of polynomials $A(z^{-1})$ and $B(z^{-1})$ are as usual n_a and n_b, but it is noted that normally $n_a = n_b = n$.

Introduce a control law of the form

$$u_t = G(z^{-1})\left[I+F(z^{-1})\right]^{-1} y_t \qquad 6.47$$

where

$$G(z^{-1}) = G_o + G_1 z^{-1} + \ldots + G_{n_g} z^{-n_g}$$
$$F(z^{-1}) = F_1 z^{-1} + \ldots + F_{n_f} z^{-n_f} \qquad 6.48$$

(Note that $G(z^{-1})$ deviates slightly from the general polynomial definition in that a coefficient matrix G_o is added). The coefficient matrices G_i, i=0,1, .. ,n_g and F_i, i=1,2, .. ,n_f are of dimension pxp. The orders n_f and n_g are given by

$$n_g = n_a - 1$$
$$n_f = n_b + k - 1 \qquad 6.49$$

Substituting eqn. 6.47 into eqn. 6.1 gives:

$$y_t = \{I+F(z^{-1})\}\{I+R(z^{-1})\}^{-1}\{I+C(z^{-1})\}e_t \qquad 6.50$$

where $$I+R(z^{-1}) = \{I+A(z^{-1})\}\{I+F(z^{-1})\} - z^{-k}B(z^{-1})G(z^{-1}) \qquad 6.51$$

Now if $F(z^{-1})$ and $G(z^{-1})$ are chosen so that

$$I+R(z^{-1}) = \{I+C(z^{-1})\}\{I+T(z^{-1})\} \qquad 6.52$$

where the order n_t of $T(z^{-1})$ is governed by the inequality

$$n_t \leq n_a + n_b + k - 1 - n_c \qquad 6.53$$

the closed loop system response is given by

$$y_t = \{I+F(z^{-1})\}\{I+T(z^{-1})\}^{-1} e_t \qquad 6.54$$

The poles of this system may therefore be specified by selecting the polynomial matrix $I+T(z^{-1})$ with the required determinant. It is argued that $I+T(z^{-1})$ and $I+F(z^{-1})$ will almost always be relatively prime, and that the case when this is not true would be the exception rather than the rule. Therefore 'pole-zero cancellation' is not expected to occur.

The evaluation of the controller parameters reduces quite simply to the solution of a set of linear equations. From eqns. 6.51 and 6.52,

$$\{I+A(z^{-1})\}\{I+F(z^{-1})\} - z^{-k}B(z^{-1})G(z^{-1}) = \{I+C(z^{-1})\}\{I+T(z^{-1})\}$$
$$= \{I+R(z^{-1})\} \qquad 6.55$$

where $n_r = n_t + n_c$. Eqn. 6.55 may be rewritten in the form:

$$\begin{bmatrix} I & & & 0 & \\ A_1 & I & & \vdots & \\ \vdots & A_1 & & 0 & \\ A_{n_a} & \vdots & I & -B_1 & \\ & A_{n_a} & A_1 & \vdots & -B_1 \\ & & \vdots & -B_{n_b} & 0 \\ & & A_{n_a} & 0 & \vdots \\ & & & \vdots & -B_{n_b} \\ & & & 0 & \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_{n_f} \\ G_o \\ G_1 \\ \vdots \\ G_{n_g} \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ \\ \\ \\ \\ R_{n_r} \end{bmatrix} - \begin{bmatrix} A_1 \\ \vdots \\ A_{n_a} \\ 0 \\ \vdots \\ \\ \\ 0 \end{bmatrix}$$

$(n_b+k-1)p$ $\quad$ $(n_a)p$ $\qquad$ 6.56

The existence of a solution to the design equation 6.55 is guaranteed when the $(n_b+n_a+k-1)p$ square transformation matrix in eqn. 6.57 is non-singular. It follows that the non-singularity of this matrix is a fundamental condition for the pole-shifting law. Since $I+A(z^{-1})$ and $B(z^{-1})$ are assumed to be relatively left prime and the system is assumed to be both controllable and observable it is claimed that in practical applications the transformation matrix will generally be non-singular.

6.6.2 Implementation of the Regulator Equation

At this stage the control law is specified in the form

$$u_t = G(z^{-1})\{I+F(z^{-1})\}^{-1}y_t \qquad 6.57$$

Two ways of implementing this law are considered. The first is to note that

$$\{I+F(z^{-1})\}^{-1} = \frac{1}{|I+F(z^{-1})|}\text{adj}\{I+F(z^{-1})\} \qquad 6.58$$

and on substitution of eqn. 6.58 into the controller equation, eqn. 6.57 becomes

$$|I+F(z^{-1})|u_t = G(z^{-1})\text{adj}\{I+F(z^{-1})\}y_t \qquad 6.59$$

from which the control input u_t can be easily computed.

For the second method it is necessary to introduce an assumption on the form of the regulator. Let

$$F^*(z) = z^{n_f}\{I+F(z^{-1})\} \qquad 6.60$$

$$G^*(z) = z^{n_g}G(z^{-1}) \qquad 6.61$$

Assumption: The regulator has $F^*(z)$ and $G^*(z)$ relatively right prime with all observability and controllability indices equal to n_f and $n_a = n_b = n$.

Then it follows that (Wolovich, 1974) the regulator equation 6.57 may be written in the form

$$u_t = \{I+\tilde{F}(z^{-1})\}^{-1}\,\tilde{G}(z^{-1})y_t \qquad 6.62$$

where $$\tilde{n_f} = n_f \qquad 6.63$$

and $$\tilde{n_g} = n_g$$

That is, $$G(z^{-1})\{I+F(z^{-1})\}^{-1} = \{I+\tilde{F}(z^{-1})\}^{-1}\,\tilde{G}(z^{-1})$$

or $$\{I+\tilde{F}(z^{-1})\}G(z^{-1}) = \tilde{G}(z^{-1})\{I+F(z^{-1})\} \qquad 6.64$$

Rewriting the eqn. 6.64 gives the set of simultaneous equations

$$\tilde{G}_o = G_o$$

$$\begin{bmatrix} G_o^T & \bigcirc & -I & \bigcirc \\ G_1^T & \ddots & -F_1^T & \ddots \\ \vdots & G_o^T & \ddots & -I \\ G_{n_g}^T & \vdots & -F_{n_f}^T & \vdots \\ \ddots & \vdots & \ddots & \vdots \\ \bigcirc & G_{n_g}^T & \bigcirc & -F_{n_f}^T \end{bmatrix} \begin{bmatrix} \tilde{F}_1^T \\ \vdots \\ \tilde{F}_{n_f}^T \\ \tilde{G}_1^T \\ \vdots \\ \tilde{G}_{n_g}^T \end{bmatrix} = \begin{bmatrix} F_1^T\tilde{G}_o^T \\ \vdots \\ F_{n_f}^T\tilde{G}_o^T \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} G_1^T \\ \vdots \\ G_{n_g}^T \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad 6.65$$

The solution of this set of $(n_f+n_g)p$ simultaneous equations yields the coefficients $\tilde{F}_i$, $i=1,2,\ldots,n_f$ and $\tilde{G}_i$, $i=1,2,\ldots,n_g$. The control law then becomes

$$\{I+\tilde{F}(z^{-1})\}u_t = \tilde{G}(z^{-1})y_t \qquad 6.66$$

From this equation it is easy to compute the controller output u_t. The author prefers this method to the first as it produces a regulator equation consistent in form with the minimum variance regulator.

It is not generally necessary for the condition $n_a = n_b = n$ in the Assumption to hold, so long as the transformation matrix in eqn. 6.65 remains non-singular.

6.6.3 SISO System Simplifications

The pole-shifting law for SISO systems permits some simplification. In particular, the closed loop system output (see eqn. 6.54) reduces to

$$y_t = \frac{1+F(z^{-1})}{1+T(z^{-1})}\ e_t \qquad 6.67$$

where the polynomial coefficients, e_t and y_t are now scalars. Assuming no pole-zero cancellation occurs, the denominator of the closed loop transfer function may be specified exactly. Furthermore the control law may be evaluated directly from the $F(z^{-1})$ and $G(z^{-1})$ polynomials as scalar polynomials commute and therefore

$$1+F(z^{-1}) = 1+\tilde{F}(z^{-1})$$
$$G(z^{-1}) = \tilde{G}(z^{-1})$$
$$\text{and}\quad \{1+F(z^{-1})\}u_t = G(z^{-1})y_t \qquad 6.68$$

6.6.4 Illustrative Examples

The advantage of the pole-shifting algorithm over the minimum variance approach for non-minimum phase systems is fairly obvious. However, pole-shifting has another important practical advantage over the multivariable minimum variance regulator (Borisson, 1975)

in that the system pure time delays in each loop need not necessarily be the same, that is, B_1, the coefficient of z^{-1} in polynomial $B(z^{-1})$ need not necessarily be non-singular.

Borisson gives an example of a special case in which the minimum variance regulator is used to control such a system. Consider the system

$$y_t + \begin{bmatrix} a_1 & a_2 \\ 0 & 0 \end{bmatrix} y_{t-1} + \begin{bmatrix} a_3 & a_4 \\ a_5 & a_6 \end{bmatrix} y_{t-2} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} u_{t-1} + \begin{bmatrix} b_1 & b_2 \\ 0 & 1 \end{bmatrix} u_{t-2} + e_t \qquad 6.69$$

The minimum variance strategy is

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} u_t + \begin{bmatrix} b_1 & b_2 \\ 0 & 1 \end{bmatrix} u_{t-1} = \begin{bmatrix} a_1 & a_2 \\ 0 & 0 \end{bmatrix} y_t + \begin{bmatrix} a_3 & a_4 \\ a_5 & a_6 \end{bmatrix} y_{t-1} \qquad 6.70$$

The reason an admissible control law results in this case is that the coefficient matrix of y_t has zero entries in positions (2,1) and (2,2) and there is a zero entry in position (2,1) of the coefficient matrix of u_t. Hence element $u_t^{(2)}$ of vector $u_t = (u_t^{(1)}, u_t^{(2)})^T$ is found as an admissible function of y_t and causality is preserved. This however is a very simple and special example, and in general the minimum variance solution when B_1 is singular leads to techniques involving the solution of a Riccati equation.

Solutions to control problems of this type are treated more simply using the more general pole-shifting rule. This is an important feature, as real systems in which loop pure time delays differ are not uncommon. The only condition that must be satisfied is the non-singularity of the transformation matrix of eqn. 6.55.

An example is presented where loop time delays differ and in addition the system is non-minimum phase (i.e. the zeros of $|B(z^{-1})|$ lie outside the unit disc in the z-plane).

Example: Consider the multivariable system

$$\{I+A_1z^{-1}+A_2z^{-2}\}y_t = \{B_1z^{-1}+B_2z^{-2}\}u_t + \{I+C_1z^{-1}\}e_t \qquad 6.71$$

where

$$A_1 = \begin{bmatrix} -1.4 & -0.2 \\ -0.1 & -0.9 \end{bmatrix} \quad A_2 = \begin{bmatrix} 0.48 & 0.1 \\ 0 & 0.2 \end{bmatrix}$$

$$B_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad B_2 = \begin{bmatrix} 1.5 & 1 \\ 0 & 1 \end{bmatrix}$$

$$C_1 = \begin{bmatrix} -0.5 & 0 \\ 0.1 & -0.3 \end{bmatrix}$$

and where the closed loop poles are to be placed at z = 0.5 and z = 0.4. A suitable choice of $I+T(z^{-1})$ is therefore

$$I+T(z^{-1}) = I + \begin{bmatrix} -0.5 & 0 \\ 0 & -0.4 \end{bmatrix} z^{-1} \qquad 6.72$$

Solving the pole-shifting equation 6.55 yields

$$F_1 = \begin{bmatrix} 0.3 & 0.1517 \\ 0.2 & 0.2 \end{bmatrix}$$

$$G_o = \begin{bmatrix} -0.1008 & -0.04829 \\ -0.1599 & -0.1152 \end{bmatrix}$$

$$G_1 = \begin{bmatrix} 0.08239 & 0.035214 \\ 0.04 & 0.04 \end{bmatrix}$$

and invoking the transformation of eqn. 6.64

$$\tilde{F}_1 = \begin{bmatrix} 0.21315 & 0.20335 \\ 0.15477 & 0.286 \end{bmatrix} \qquad 6.74$$

$$\tilde{G}_o = G_o \qquad \tilde{G}_1 = \begin{bmatrix} 0.06821 & 0.02646 \\ 0.04953 & 0.04688 \end{bmatrix}$$

giving the control law:

$$u_t = -F_1 u_{t-1} + G_o y_t + G_1 y_{t-1} \qquad 6.75$$

The reader is referred to the example in section 7.3.4.3 where this same problem is solved by direct self-tuning.

The closed loop system output is:

$$y_t = \{I+F_1 z^{-1}\}\{I+T_1 z^{-1}\}^{-1} e_t \qquad 6.76$$

$$= \frac{1}{(1-0.5z^{-1})(1-0.4z^{-1})} \begin{bmatrix} (1+0.3z^{-1})(1-0.4z^{-1}) & 0.1517z^{-1}(1-0.5z^{-1}) \\ 0.2z^{-1}(1-0.4z^{-1}) & (1-0.2z^{-1})(1-0.5z^{-1}) \end{bmatrix} e_t$$

Now it is clear that although the stability of the closed loop system has been ensured by the choice of poles, the system has not (and will not in general) be decoupled. This may be viewed as a disadvantage of the method. The only instance in which decoupling can be guaranteed is for the very simple case when $n_b = 1$, $k = 0$. Then $I+F(z^{-1}) = I$, and $I+T(z^{-1})$ can be chosen to be diagonal.

Another possible disadvantage of the pole-shifting method is that the output variance is not minimized. However, the generally robust nature of the method in that it handles non-minimum phase systems and

time delay differences between loops, plus the benefit of normally modest control excursions must weigh in its favour.

6.7 Regulating to a Reference Value

The regulators discussed can be extended quite simply to enable them to regulate to a constant reference value. The method is discussed by Wittenmark (1973) for his minimum variance regulator but is equally well applicable to the pole-shifting regulator.

Let y_r be the desired reference level (set point). Define a reference input u_r such that

$$\{I+A(z^{-1})\}y_{r_t} = z^{-k}B(z^{-1})u_{r_t} \qquad 6.77$$

Subtracting eqn. 6.77 from eqn. 6.1 gives

$$\{I+A(z^{-1})\}(y_t-y_{r_t}) = z^{-k}B(z^{-1})(u_t-u_{r_t}) + \{I+C(z^{-1})\}e_t \qquad 6.78$$

Let

$$\begin{aligned} \Delta y_t &= y_t - y_{r_t} \\ \Delta u_t &= u_t - u_{r_t} \end{aligned} \qquad 6.79$$

Then the pole-shifting equation for the system of eqn. 6.78 is

$$\Delta u_t = G(z^{-1})\{I+F(z^{-1})\}^{-1}\,\Delta y_t \qquad 6.80$$

or

$$u_t = u_{r_t} + G(z^{-1})\{I+F(z^{-1})\}^{-1}\,\Delta y_t \qquad 6.81$$

giving a closed loop response

$$y_t = y_{r_t} + \{I+T(z^{-1})\}\{I+F(z^{-1})\}^{-1}e_t \qquad 6.82$$

Unfortunately eqn. 6.79 presupposes knowledge of u_{r_t} which is a function of the future reference value $y_{r_{t+k+1}}$. Furthermore it

requires the cancellation of the possibly non-minimcm phase $B(z^{-1})$ polynomial. However, if y_{r_t} is a constant, then $y_{r_{t+k+1}} = y_{r_t}$ and for steady state control u_{r_t} may be chosen to be

$$u_{r_t} = B^{-1}(1)\{I+A(1)\}y_{r_t} \qquad 6.83$$

An even simpler solution results when the open loop system contains at least one integrator. In that case, the steady state value of u_r is zero.

A technique frequently used (especially when the self-tuning controller is used, as discussed in Chapter 7) is to introduce a digital integrator in cascade with the system. The new system so formed is then subjected to the appropriate regulator law. The system block diagram is shown in Fig. 6.1.

Although the introduction of an integrator allows u_{r_t} to be set to zero for correct steady state performance, it usually does not promote good transient performance when the input reference is changed sharply. This point is taken further in the next section.

6.8 A Feedforward Structure for Servo Following

Unless the reference input is constant or varying very 'slowly' the cascade integrator method described above is not entirely satisfactory. The reason for this will become apparent shortly.

The feedforward structure shown in Fig. 6.2 provides a more general vehicle for the inclusion of set points. The method is not unusual in industrial control systems and has been described by Wellstead and Zanker (1978a) for SISO systems.

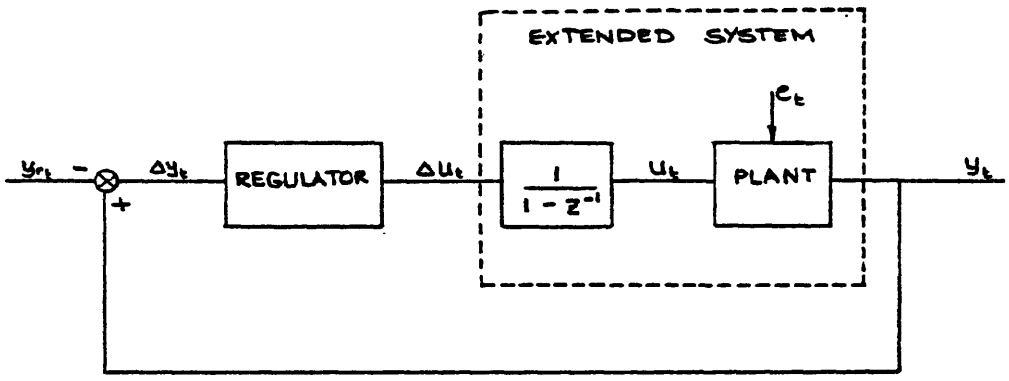


Fig. 6.1 Introduction of a digital integrator for regulating to a reference value

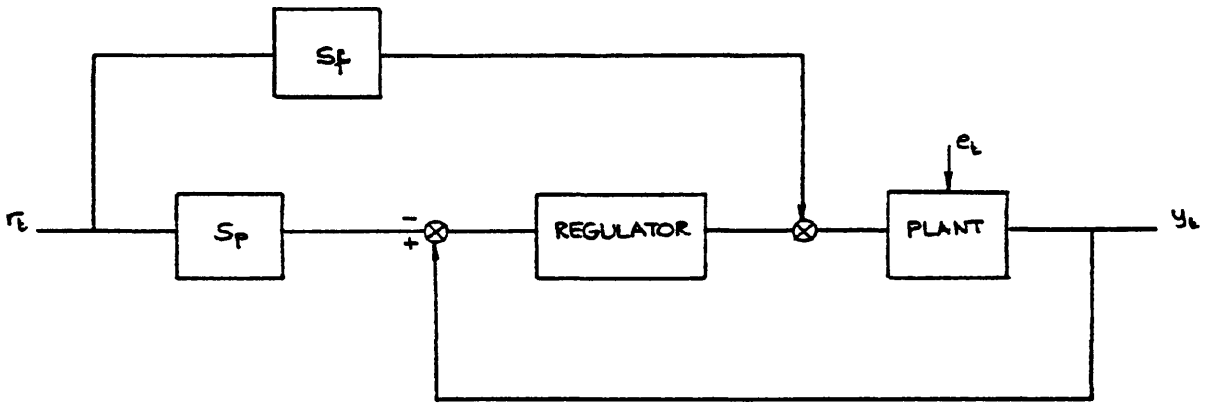


Fig. 6.2 A feedforward and cascade structure for servo following

In the following, it is assumed that the pole-shifting regulator is employed. Referring to Fig. 6.2, S_f and S_p are feedforward and servo-compensators, and r_t is the reference input. It is easy to verify that the closed loop system is described by the equation:

$$y_t = z^{-k}\{I+F(z^{-1})\}\{I+T(z^{-1})\}^{-1}\{I+C(z^{-1})\}^{-1}B(z^{-1}) \; * \; \{S_f-G(z^{-1})\{I+F(z^{-1})\}^{-1}S_p\}r_t + \{I+F(z^{-1})\}\{I+T(z^{-1})\}^{-1}e_t \qquad 6.84$$

Various choices of S_p and S_f lead to different closed loop responses. A few options are considered.

a) $S_p = 1$, $S_f = 0$

Assume that the open loop system contains an integrator, and for simplicity consider a SISO system. Then the structure is precisely that discussed in the previous section for $u_r = 0$, and the closed loop response is:

$$y_t = \frac{-z^{-k}B(z^{-1})G(z^{-1})}{(1+T(z^{-1}))(1+C(z^{-1})} r_t + \frac{1+F(z^{-1})}{1+T(z^{-1})} e_t \qquad 6.85$$

The closed loop poles in the transfer function between r_t and y_t are not only the specified poles at zeros of $1+T(z^{-1})$, but also exist at the zeros of $1+C(z^{-1})$. The latter are of course only 'absent' in the transfer function betwwen e_t and y_t due to pole-zero cancellation (see eqn. 6.50), but are stable poles (see eqn. 6.1). More significantly, zeros are contributed to the transfer function between r_t and y_t, and it is these that often lead to excessive overshoot. An example of this difficulty is given in Fig. 6.3 where the system considered is

$$(1-z^{-1})(1-0.8z^{-1})y_t = (z^{-1}+0.5z^{-2})u_t + (1-0.4z^{-1})e_t \qquad 6.86$$

and $1+T(z^{-1}) = 1$, giving the pole-shifting control

$$(1+0.3846z^{-1})u_t = (-1.0154+0.6154z^{-1})y_t \qquad 6.87$$

The input r_t is a square wave, and the response y_t is shown in the figure.

b) $S_p = z^{-k}B(z^{-1})$, $S_f = 1+A(z^{-1})$

Again, a SISO system under pole-shifting control is considered. The closed loop response is:

$$y_t = z^{-k}B(z^{-1})r_t + \frac{1+F(z^{-1})}{1+T(z^{-1})}\, e_t \qquad 6.88$$

This is an extremely useful formulation, as it does not cancel out $B(z^{-1})$ which may have zeros outside the stability region. However the extension to multivariable systems is not straightforward due to difficulties introduced by matrices not obeying the Commutative Law.

c) $S_p = 1$, $S_f = B(z^{-1})^{-1}(I+A(z^{-1}))$

This method is only applicable when $|B(z^{-1})|$ and $|I+A(z^{-1})|$ have all their zeros within the stability region. The closed loop response is:

$$y_t = z^{-k}r_t + \{I+F(z^{-1})\}\{I+T(z^{-1})\}^{-1}\, e_t$$

Clearly this is the ideal response, but direct inversion of the system can never be exact in practice.

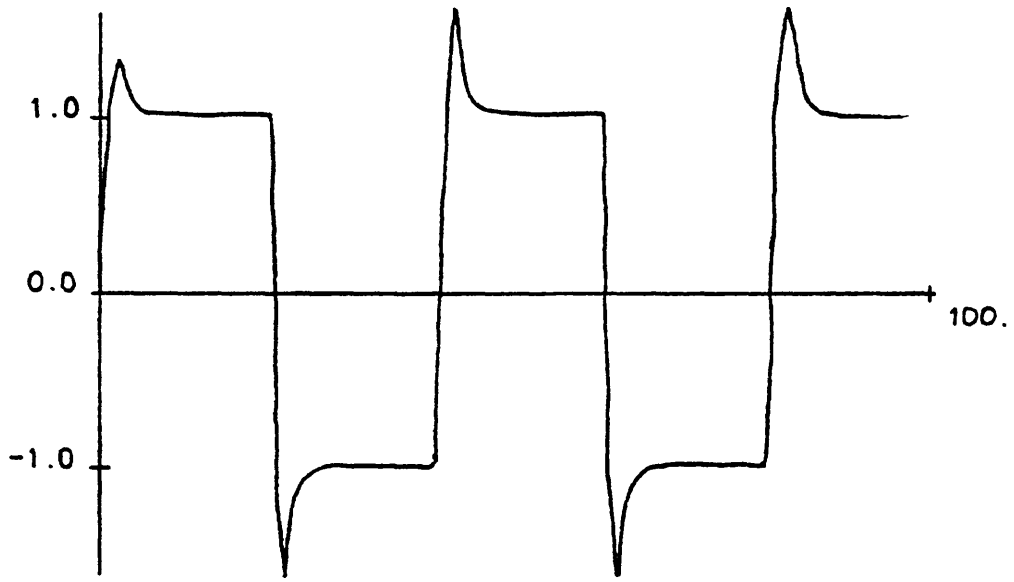


Fig. 6.3 Example of overshoot when relying on a system (or inserted) integrator to extend a regulator for servo following

6.9 Conclusion

This chapter has provided the background necessary for the following chapter by discussing various control strategies that are amenable to self-tuning. The basic multivariable minimum variance regulator and the generalized minimum variance controller (for SISO systems) have been reviewed. In addition, the multivariable extensions to the detuned minimum variance and pole-shifting laws have been presented and are new strategies offered for self-tuning. The advantages of the pole-shifting law in particular, with its ability to deal generally with non-minimum phase systems and many multivariable systems in which pure time delays differ between loops, have been discussed and illustrated.

Finally, techniques and difficulties encountered in extending regulator laws to deal with constant and time varying reference inputs have been considered.

CHAPTER 7

SELF TUNING REGULATORS

7.1 Introduction

It has been suggested in Chapter 6 that the design of a digital controller usually involves a two-stage process. The first stage is the modelling of the plant to be controlled about some operating point under the assumption that it is a finite order, linear, time invariant system. The ARMAX representation serves as a convenient form for this model, and the values of its parameters may be assigned either by physical modelling, or increasingly commonly now, as the result of parameter estimation. Stage one is therefore the system identification phase. Stage two involves the derivation of a control law to meet a chosen design objective for the closed loop system performance.

Clearly, a more elegant solution emerges if the operations of Identification and Control can be combined into a single on-line process. The concept is certainly not new. Chapter 2 has highlighted some of the varied approaches to this problem, one of which is 'Self Tuning'. Self tuning controllers are charaterized by being based on output rather than state feedback, and a stochastic ARMAX plant description rather than a deterministic state space model.

The basic structure of a self tuning regulator is shown in Fig. 7.1. A recursive least squares estimator is typically used to estimate the parameters of a suitable plant model which is generally in ARMAX form. The parameter estimates are updated at each sampling instant and are passed to a controller design algorithm which synthesizes controller

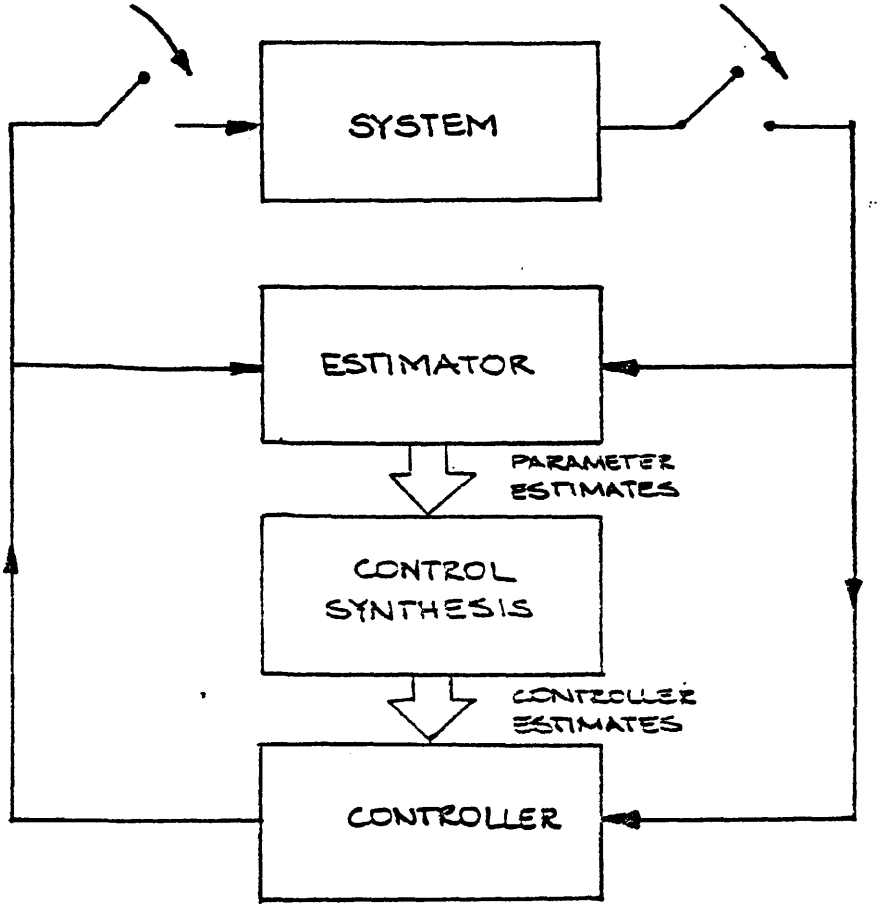


Fig. 7.1 Structure of a Self-Tuning Controller

parameters according to some pre-specified design rule. Updated controller parameters are now inserted into the regulator equations and used to compute the next input to the system.

Self tuning regulators have advantages over the conventional two stage off-line identification and fixed law controller design method which go beyond that of elegance. They possess a property referred to here as the 'Self Tuning Property' which enables a simple least squares estimator to be used in estimating the plant model parameters instead of a complicated non-linear technique (or for example that presented in Chapters 4 and 5) which normally would be required to estimate the parameters of the noise colouration polynomial $C(z^{-1})$ of the ARMAX model (see eqn. 6.1). This fundamental property is shown to hold for two new self tuners in this chapter. Furthermore, the process of simultaneous on-line estimation and control makes it possible for the self tuning regulator to adapt to slow changes in the plant dynamics and in the case of the pole-shifting self tuner, even changes in system pure time delays. The implication of this feature in an industrial environment is that once set up, self tuning regulators are self maintaining and do not require periodic adjustments to allow for minor changes in the plant characteristics.

The control strategies on which self tuning regulators and controllers are based have been described in Chapter 6. Most self tuners are designed to meet some optimality criterion. For example, Astrom and Wittenmark (1973) seek to minimize the output variance ('Minimum Variance' Self Tuner). Their SISO system work was extended to multivariable systems by Borisson (1975). Clarke and Gawthrop

(Clarke and Gawthrop 1975; Gawthrop 1977) in their SISO system controller minimize the variance of a generalized output. Edmunds (1976) deviated from this approach in his early work on the detuned minimum variance regulator and particularly the pole shifting regulator which is not based on an optimality criterion but rather is endebted more to the philosophies of classical control theory.

This chapter is divided into two main sections. The first is, for completeness, dedicated to a brief review of the minimum variance and generalized minimum variance self tuning controllers. The second section is devoted to new self tuning regulators, namely the multivariable detuned minimum variance and multivariable pole shifting regulators based on the SISO concepts of Edmunds (1976). Examples of computer simulations using these techniques are given.

7.2 A Review of Self Tuners based on Optimality Criteria

This section reviews the Minimum Variance Self Tuning Regulator (Borisson 1975) and the Generalized Minimum Variance Self Tuning Controller (Clarke and Gawthrop 1975; Gawthrop 1977). The strategies are based directly upon those discussed in sections 6.2 and 6.5. The ARMAX system description given in eqn. 6.1 is assumed to hold. No proofs are given, but the salient features of each technique are presented. The reader is referred to the references named above for detailed development of theory and proofs. Further references to related work are given in Chapter 2.

7.2.1 The Self Tuning Minimum Variance Regulator

This regulator requires that the plant be modelled as:

$$\{I+z^{-k}\hat{A}(z^{-1})\}\, y_t = z^{-k}\hat{B}(z^{-1})u_t + \varepsilon_t \qquad 7.1$$

where $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$ are polynomial matrices of the form

$$\hat{X}(z^{-1}) = \hat{X}_1 z^{-1} + \ldots\ldots\ldots + \hat{X}_{n_{\hat{x}}} z^{-n_{\hat{x}}}$$

u_t is a p-vector of inputs, y_t is a p-vector of outputs, and ε_t is the p-vector residual when $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$ are estimated using a recursive least squares estimator. The orders of the polynomials are defined in terms of n_g and $n_{\tilde{m}}$, the orders of polynomials $G(z^{-1})$ and $\tilde{M}(z^{-1})$ in eqn. 6.3 and are:

$$n_{\hat{a}} \geq n_g + 1 \qquad 7.2$$

$$n_{\hat{b}} \geq n_{\tilde{m}} + n$$

where $n = n_a = n_b$

The regulator law chosen is then simply:

$$\hat{B}(z^{-1})u_{t+1} = \hat{A}(z^{-1})y_{t+1} \qquad 7.3$$

where $\hat{B}_1$ is assumed to be non-singular. It is important to note that under the present assumptions eqn. 7.3 may not be the minimum variance strategy. In fact, even when further assumptions are made, in particular that the parameter estimates have converged and that the closed loop system is such that the output is ergodic (in the second moments), Borisson (1975) shows that the minimum variance strategy is only a <u>possible</u> outcome. Only when the process disturbances are white is the minimum variance strategy the only possible resulting strategy.

Notice also that in general the orders n_g and $n_{\tilde{m}}$ required to specify $n_{\hat{a}}$ and $n_{\hat{b}}$ are not known and require knowledge of polynomials $A(z^{-1})$, $B(z^{-1})$ and $C(z^{-1})$ of the ARMAX system model (eqn. 6.1). For SISO systems however it is always true that

$$n_g = n_a - 1 \qquad 7.4$$
$$n_{\tilde{m}} = k$$

Furthermore, for SISO systems Wittenmark (1973) has shown that the minimum variance strategy is the only possible outcome for a suitably configured minimum variance self tuning regulator.

In fact certain assumptions can be introduced which have enabled Borisson to prove a similarly useful result for the multivariable regulator.

Let $$\hat{A}^{\times}(z) = z^{n_{\hat{a}}-1}(z\hat{A}(z^{-1})) \qquad 7.5$$

$$\hat{B}^{\times}(z) = z^{n_{\hat{b}}-1}(zB(z^{-1})) \qquad 7.6$$

i.e. $\hat{A}^{\times}(z) = \hat{A}_1 z^{n_{\hat{a}}-1} + \ldots\ldots + \hat{A}_{n_{\hat{a}}}$

and similarly for $\hat{B}^{\times}(z)$.

Then the important additional assumptions required are:

1. The controlled process has a minimum variance regulator with all observability and controllability indices equal to n+k-1.
2. The limiting regulator has $\hat{A}^{\times}(z)$ and $\hat{B}^{\times}(z)$ given by eqns. 7.5 and 7.6 respectively left prime with all observability and controllability indices equal to n+k-1.
3. The closed loop system has the maximum observability index not higher than 2n+k-1.

Also, let $n_{\hat{a}} = n = n_a$ 7.7

and $n_{\hat{b}} = n + k$ 7.8

implying that $n_{\tilde{m}} = n_m = k$.

Under these assumptions the minimum variance strategy is the only possible resulting strategy.

The implication of this is that the residual sequence ε_t converges to a kth order moving average of the system noise e_t and the system output is

$$y_t = [I + M(z^{-1})]e_t = \varepsilon_t \qquad 7.9$$

where $M(z^{-1})$ is the same as in eqn. 6.9, that is, the response is the same as would be expected had the regulator been designed according to the rule given in Chapter 6. Thus, in spite of a non-zero noise colouration polynomial $C(z^{-1})$ (see the general ARMAX system equation 6.1) it has been possible to design the regulator from a model whose parameters were estimated using simple linear least squares. This 'self tuning property' whereby the parameters converge to the desired regulator law without the need to explicitly model the noise colouration is extremely useful and is a common feature of the self tuning regulator and controller methods.

The required assumptions are not stringent and are almost always true in practice. Just as its SISO system counterpart, the multivariable minimum variance self tuner has the advantage of computational simplicity. The regulator parameters are identified directly in the model eqn. 7.1. In general however, it is necessary to invert $\hat{B}_1$ to compute the control input. As in the single variable case, it is possible to fix the value of $\hat{B}_1$, the most convenient

value being the identity matrix. However, in fixing B_1 the following condition must be satisfied: (Borisson 1975)

$$|\mu^{i}_{(I - B_1\hat{B}_1^{-1})}| < 1 \qquad i = 1,2,\ldots,p \qquad 7.10$$

where $\mu^{i}_{(I - B_1\hat{B}_1^{-1})}$ are the eigenvalues of the matrix $(I-B_1\hat{B}_1^{-1})$.

The SISO system condition is:

$$|1 - B_1\hat{B}_1^{-1}| < 1 \qquad 7.11$$

Note that B_1 and $\hat{B}_1$ are the coefficient matrices of z^{-1} in polynomials $B(z^{-1})$ (in system eqn. 6.1) and $\hat{B}(z^{-1})$ (in model eqn. 7.1) respectively.

The properties and limitations of the self tuning minimum variance regulator are of course similar to those given for minimum variance regulation in Chapter 6. In particular, systems in which the zeros of $|B(z^{-1})|$ lie outside the stability region ('non minimum phase' systems) are not amenable to this type of regulation. Even when the basic system is minimum phase, the effect of computational time delays which are a fraction of the sampling period ('non integer' time delays) introduced in computing the control input can make the system appear non-minimum phase. Such computational time delays may be treated in one of two ways:

a) Increasing the value of k (the system pure time delay expressed in multiples of the sampling period) by 1 in the system model, eqn. 7.1.

b) Extending the $\hat{B}(z^{-1})$ polynomial by one term.

The second method is preferable so long as the system does not become non-minimum phase as an extra delay of one sampling period is then not introduced into the loop.

Notice also that the time delay k should never be modelled smaller than in the actual system. If it is, $\hat{B}_1$ may become singular and the algorithm will then go unstable.

The minimum variance regulator requires that all the system loops have the same pure time delay.

7.2.2 The Generalized Minimum Variance Controller

This controller is presently defined only for SISO systems. As such, the underlying system is assumed to behave according to the ARMAX description of eqn. 6.1 with the number of inputs and outputs p, set to 1, i.e. as in eqn. 6.38. Define also an auxiliary system with output ϕ_t as given in eqn. 6.37 such that

$$\phi_{t+k+1} = Py_{t+k+1} + Qu_t - Rw_t \qquad 7.12$$

where P, Q and R are rational transfer functions in the backward shift operator acting on the system output, input and set point respectively. Also define

$$P = \frac{P_N}{P_D} \qquad 7.13$$

$$Q = \frac{Q_N}{Q_D} \qquad 7.14$$

$$R = \frac{R_N}{R_D} \qquad 7.15$$

where P_N and P_D are the numerator and denominator respectively of P, and similar relationships hold for Q and R.

The controller requires that the auxiliary system be modelled as:

$$\phi_t = z^{-k}G(z^{-1})y_{f_t} + z^{-k}F(z^{-1})u_{f_t} - z^{-k}H(z^{-1})w_{ff_t} + \varepsilon_t \qquad 7.16$$

where
$$y_{f_t} = (1/P_D)y_t \quad 7.17$$
$$u_{f_t} = (1/Q_D)u_t \quad 7.18$$
$$w_{ff_t} = (R_N/R_D)w_t = Rw_t \quad 7.19$$

and $F(z^{-1})$, $G(z^{-1})$ and $H(z^{-1})$ are polynomials of the form

$$X(z^{-1}) = X_1 z^{-1} + \ldots\ldots + X_{n_x} z^{-n_x} \quad 7.20$$

with
$$n_g = n + n_{P_D} - k \quad 7.21$$
$$n_f = \max(k+n+n_{q_D}, n_{q_N}+n) \quad 7.22$$
$$n_h = n + 1 \quad 7.23$$

and where $n=n_a=n_b$ and k are defined in the ARMAX plant model (eqn. 6.38). ε_t is the residual when $F(z^{-1})$, $G(z^{-1})$ and $H(z^{-1})$ are estimated using a recursive least squares estimator.

The required self tuning control law is then:

$$G(z^{-1})y_{f_t} + F(z^{-1})u_{f_t} - H(z^{-1})w_{ff_t} = 0 \quad 7.24$$

When the algorithm has converged it can be shown that the residual ε_t tends to the optimal prediction error in the k+1 step ahead predictor of ϕ_t, i.e.

$$\varepsilon_t = [1+M(z^{-1})]e_t \quad 7.25$$

where $M(z^{-1})$ is defined in eqn. 6.41. Thus the auxiliary system output becomes

$$\phi_t = [1+M(z^{-1})]e_t \quad 7.26$$

which is precisely the result achieved by the off line design of the generalized minimum variance controller given in section 6.5. The choice of the parameters P, Q and R to give various closed loop

system characteristics may be made in the same way as described in section 6.5.

The self tuning version of the generalized minimum variance controller has the same advantages and disadvantages of this strategy when used in off line controller design. One significant point however is that the method may not in general be used to control non minimum phase systems except by suitable choice of the P and Q transfer functions. This choice requires knowledge of the system parameters which although clearly known when designing controllers in the conventional off-line manner may not be available when self tuning. It is after all one of the objects of a self tuning control system that precise knowledge of system parameters is not required and that the controller should tolerate and adapt to variations in the system characteristics.

However, the method is relatively straightforward to implement and requires little more work than the minimum variance controller to partially specify the closed loop transfer function. As such it is an important and valuable technique.

7.3 New Multivariable Self Tuning Regulators

This section discusses two new multivariable self tuning regulators, the detuned minimum variance regulator and the multivariable pole shifting regulator. The features of these regulators when designed off line have already been discussed in sections 6.3 and 6.6. Both regulators require a similar proof and therefore a generalized 'Self Tuning Lemma' is first given. Thereafter each regulator is discussed separately and simulation examples are given.

7.3.1 Generalized Self Tuning Lemma

Let the ARMAX system behaving according to eqn. 6.1 be modelled by the difference equation

$$\left[I + z^{-k_m}\hat{A}(z^{-1})\right] y_t = z^{-k_m}\hat{B}(z^{-1})u_t + \varepsilon_t \qquad 7.27$$

where k_m is an integer constant selected according to the self tuning configuration used and $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$ are pXp matrix polynomials of the form

$$X(z^{-1}) = X_1 z^{-1} + \ldots\ldots\ldots + X_{n_x} z^{-n_x} \qquad 7.28$$

y_t and u_t are the system output and input p-vectors respectively, and ε_t is the residual p-vector when $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$ are estimated using a recursive least squares estimator, i.e. so as to minimize $\sum_{j=1}^{t} \varepsilon_j^{(i)2}$, i = 1,2,...,p where $\varepsilon_t = \left[\varepsilon_t^{(1)},\ldots,\varepsilon_t^{(p)}\right]^T$.

Let $$n_{\hat{a}} = n_a \qquad 7.29$$

$$n_{\hat{b}} = n_b + k \qquad 7.30$$

where $n_b = n_a$

and select the control input u_t so that

$$u_t = G(z^{-1})\left[I + F(z^{-1})\right]^{-1} y_t \qquad 7.31$$

where $$G(z^{-1}) = G_o + G_1 z^{-1} + \ldots\ldots + G_{n_g} z^{-n_g} \qquad 7.32$$

$$I + F(z^{-1}) = I + F_1 z^{-1} + \ldots\ldots + F_{n_f} z^{-n_f} \qquad 7.33$$

$$n_g = n_{\hat{a}} - 1 \qquad 7.34$$

and $$n_f = n_{\hat{b}} - 1 \qquad 7.35$$

$F(z^{-1})$ and $G(z^{-1})$ are chosen to satisfy the identity

$$I+K(z^{-1}) = [I+z^{-k_m}\hat{A}(z^{-1})][I+F(z^{-1})] - z^{-k_m}\hat{B}(z^{-1})G(z^{-1}) \quad 7.36$$

where $K(z^{-1})$ of order n_k is defined to be of the form given in eqn. 7.28, and is specified according to the regulator law desired.

Define also:

$$I+L(z^{-1}) = [I+A(z^{-1})][I+F(z^{-1})] - z^{-k}B(z^{-1})G(z^{-1}) \quad 7.37$$

where $L(z^{-1})$ follows the form given in eqn. 7.28 and is of order

$$n_\ell \leq n_a + n_b + k - 1 \quad 7.38$$

Furthermore, assume the following conditions are met:

C1: A regulator designed off line corresponding to the desired self tuning regulator exists and has all observability and controllability indices equal to n_f and $G(z^{-1})$ and $I+F(z^{-1})$ are relatively right prime.

C2: $[I+L(z^{-1})][I+K(z^{-1})]^{-1}$ may be represented as $[I+\tilde{K}(z^{-1})]^{-1}[I+\tilde{L}(z^{-1})]$ where $n_{\tilde{k}} \leq n_k$, $n_{\tilde{\ell}} \leq n_\ell$.

C3: $n_k \leq n_a+n_b+k_m-n_c-1$ where n_c is defined as the order of $C(z^{-1})$ (see eqn. 6.1).

C4: The model parameters converge and the output of the closed loop system is ergodic.

Then the closed loop system output converges to

$$y_t = [I+F(z^{-1})][I+K(z^{-1})]^{-1}[I+\Phi(z^{-1})]\ e_t \quad 7.39$$

where $\Phi(z^{-1})$ is a matrix polynomial of the form shown in eqn. 7.28 of

order $n_\phi = k_m$ and the residual ε_t is a k_mth order moving average of the system driving noise e_t, viz.

$$\varepsilon_t = [I + \Phi(z^{-1})]\, e_t \qquad 7.40$$

7.3.1.1 Comments on the Conditions

1. Condition C1 is a condition that will be met for a wide range of systems. It is used in the proof for the self tuning minimum variance regulator (Borisson 1975) and ensures that the regulator equation 7.31 can be re-written as:

$$u_t = [I+\tilde{F}(z^{-1})]^{-1}\, \tilde{G}(z^{-1})\, y_t \qquad 7.41$$

where $I+\tilde{F}(z^{-1})$ and $\tilde{G}(z^{-1})$ are relatively left prime and have orders n_f and n_g respectively (Wolovich 1974). $\tilde{F}(z^{-1})$ follows the form given in eqn. 7.28 and $\tilde{G}(z^{-1})$ is analogous to the form of $G(z^{-1})$.

2. Let $I+L(z^{-1})$ and $I+K(z^{-1})$ be factored so that

$$[I+L(z^{-1})][I+K(z^{-1})]^{-1} = [I+L^*(z^{-1})][I+K^*(z^{-1})]^{-1} \qquad 7.42$$

where $I+L^*(z^{-1})$ and $I+K^*(z^{-1})$ are now relatively right prime and of order $n_{\ell*} \leq n_\ell$, $n_{k*} \leq n_k$. Then condition C2 requires that the matrix

$$L = \begin{bmatrix} I & & -I & \\ K_1^{*T} & & -L_1^{*T} & \\ & I & & -I \\ & K_1^{*T} & & -L_1^{*T} \\ K_{n_{k*}}^{*T} & & -L_{n_{\ell*}}^{*T} & \\ & K_{n_{k*}}^{*T} & & -L_{n_{\ell*}}^{*T} \end{bmatrix} \qquad 7.43$$

$n_{\ell*}$ block columns $\quad$ n_{k*} block columns

is non-singular. This matrix is obtained by writing the required relationship

$$[I+\tilde{K}(z^{-1})][I+L^{*}(z^{-1})] = [I+\tilde{L}(z^{-1})][I+K^{*}(z^{-1})]$$

equating coefficients, and forming the equations necessary to solve for $\tilde{L}(z^{-1})$ and $\tilde{K}(z^{-1})$. Now $I+L(z^{-1})$ is a function of estimated regulator parameters $I+F(z^{-1})$ and $G(z^{-1})$. $I+K(z^{-1})$ varies according to the type of self tuner and may be deterministic or a function of an estimated quantity. The presence of stochastic elements in L will, it is claimed, in general ensure that the matrix remains non-singular. In the case of SISO systems condition C2 is trivially fulfilled.

Conditions C2 and C3 are used together in the proof to limit the order of the autoregressive and moving average elements in the ARMA system relating the residual ε_t to the system driving noise e_t. Bearing in mind that $y_t=\varepsilon_t$ for the minimum variance self tuner the conditions have a similar purpose to that imposed by Borisson for the minimum variance self tuner proof. There, the closed loop system may have maximum observability index not higher than 2n+k-1.

7.3.1.2 Proof

From the properties of least squares estimation the following results hold at time t:

$$\frac{1}{t}\sum_{j=1}^{t} y_{j-i}\varepsilon^{T}_{j+k_m} = 0 \qquad i=1,2,\ldots,n_{\hat{a}} \qquad 7.44$$

$$\frac{1}{t}\sum_{j=1}^{t} u_{j-i}\varepsilon^{T}_{j+k_m} = 0 \qquad i=1,2,\ldots,n_{\hat{b}} \qquad 7.45$$

where the system is assumed to start at rest and $y_t=u_t=0$ for $t \leq 0$. Now, by assumption, since the parameters converge and by virtue of the assumptions of ergodicity (C4) replace the time averages as $t\to\infty$ of eqns. 7.44 and 7.45 with ensemble averages and hence

$$E\,(y_{t-i}\varepsilon^T_{t+k_m}) = R_{y\varepsilon}(k_m+i) = 0, \quad i=1,2,\ldots,n_{\hat{a}} \qquad 7.46$$

$$E\,(u_{t-i}\varepsilon^T_{t+k_m}) = R_{u\varepsilon}(k_m+i) = 0, \quad i=1,2,\ldots,n_{\hat{b}} \qquad 7.47$$

where E(.) represents the expectation operator.

Define an auxiliary p-vector w_t such that

$$w_t = \left[I+F(z^{-1})\right]^{-1} y_t \qquad 7.48$$

Then from eqns. 7.48 and 7.31

$$y_t = \left[I+F(z^{-1})\right] w_t \qquad 7.49$$

$$u_t = G(z^{-1})\, w_t \qquad 7.50$$

and the following $(n_{\hat{a}}+n_{\hat{b}}-1)p$ independent equations may be written:

$$\begin{matrix} n_{\hat{a}}p \\ \\ \\ (n_{\hat{b}}-1)p \\ \\ \\ \end{matrix} \left\{\begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix}\right. \underbrace{\begin{bmatrix} I & F_1 & F_2 & \cdots & F_{n_f} & \bigcirc \\ & \ddots & & & & \ddots \\ \bigcirc & & I & F_1 & \cdots & F_{n_f} \\ G_o & G_1 & \cdots & G_{n_g} & & \bigcirc \\ & \ddots & & & & \ddots \\ \bigcirc & & G_o & G_1 & \cdots & G_{n_g} \end{bmatrix}}_{\text{Matrix } Q} \begin{bmatrix} w_{t-1} \\ w_{t-2} \\ \vdots \\ \\ \\ w_{t-m} \end{bmatrix} = \begin{bmatrix} y_{t-1} \\ \vdots \\ y_{t-n_{\hat{a}}} \\ u_{t-1} \\ \vdots \\ u_{t-n_{\hat{b}}+1} \end{bmatrix} \qquad 7.51$$

where $m = n_{\hat{a}}+n_{\hat{b}}-1$

Matrix Q has full rank.

It follows that, postmultiplying both sides of eqn. 7.51 by $\varepsilon^T_{t+k_m}$, taking expectations and noting eqns. 7.46 and 7.47

$$Q\left[R^T_{w\varepsilon}(k_m+1), \ldots\ldots, R^T_{w\varepsilon}(k_m+m) \right]^T = \underline{0} \qquad 7.52$$

where $\underline{0}$ is the null vector, and that the solution to eqn. 7.52 is given by

$$R_{w\varepsilon}(k_m+\tau) = 0 , \quad \tau = 1,2,\ldots,m \qquad 7.53$$

By substituting the control law (eqn. 7.31) into the model equation 7.27 and observing eqns. 7.49 and 7.36 it is easy to establish that

$$\left[I+K(z^{-1})\right] w_t = \varepsilon_t \qquad 7.54$$

Also, it can be shown quite easily by substituting the control law into system eqn. 6.1 and observing eqns. 7.37 and 7.49 that

$$\left[I+L(z^{-1})\right] w_t = \left[I+C(z^{-1})\right] e_t \qquad 7.55$$

Together with condition C2, equation 7.54 and 7.55 allow the residual and system driving noise e_t to be related as follows:

$$\left[I+\tilde{L}(z^{-1})\right] \varepsilon_t = \left[I+\tilde{K}(z^{-1})\right]\left[I+C(z^{-1})\right] e_t \qquad 7.56$$

where $n_{\tilde{l}} \leq n_a+n_b+k-1 = m \qquad 7.57$

Now define

$$\begin{aligned} I+S(z^{-1}) &= \left[I+\tilde{K}(z^{-1})\right]\left[I+C(z^{-1})\right] \qquad 7.58 \\ &= I + S_1 z^{-1} + \ldots\ldots + S_{n_s} z^{-n_s} \end{aligned}$$

where $n_s = n_{\tilde{k}} + n_c$

Transposing eqn. 7.56, premultiplying by w_{t-m-k_m-1} and taking expectations gives:

$$R_{w\varepsilon}(m+k_m+1) + R_{w\varepsilon}(m+k_m)\tilde{L}_1^T + \ldots + R_{w\varepsilon}(m+k_m+1-n_{\tilde{l}})\tilde{L}_{n_{\tilde{l}}}^T$$

$$= R_{we}(m+k_m+1) + R_{we}(m+k_m)S_1^T + \ldots + R_{we}(m+k_m+1-n_s)S_{n_s}^T \qquad 7.59$$

From eqn. 7.53 the left hand side of eqn. 7.59 reduces to $R_{w\varepsilon}(m+k_m+1)$ and the right hand side vanishes if

$$n_s \leq m+k_m \qquad 7.60$$

since, from eqns. 7.49 and 7.50, w_t is a function only of present and past values of y_t and u_t which are by definition uncorrelated with future values of the driving noise e_t. Eqn. 7.60 is satisfied by condition C3. Thus

$$R_{w\varepsilon}(m+k_m+1) = 0 \qquad 7.61$$

Similarly, postmultiplying the transpose of eqn. 7.56 by w_{t-m-k_m-2}, w_{t-m-k_m-3}, etc., and taking expectations it follows that

$$R_{w\varepsilon}(k_m+\tau) = 0 \qquad \tau > 0 \qquad 7.62$$

From eqns. 7.54 and 7.62 it is easy to deduce that

$$R_{\varepsilon\varepsilon}(k_m+\tau) = 0 \qquad \tau > 0 \qquad 7.63$$

i.e. ε_t is a k_mth order moving average of a white noise. However, from eqn. 7.56 for example, ε_t may be represented as a linear combination of all past noise inputs e_t,

$$\varepsilon_t = e_t + \sum_{i=1}^{\infty} \phi_i e_{t-i} \qquad 7.64$$

where ϕ_i, i=1,2,...,∞ are pXp coefficient matrices. From this

relationship and the correlation property of eqn. 7.63 it follows that ε_t is a k_mth order moving average of e_t

$$\varepsilon_t = [I+\Phi(z^{-1})]\, e_t \qquad 7.65$$

where $\Phi(z^{-1}) = \Phi_1 z^{-1} + ... + \Phi_{k_m} z^{-n_{k_m}}$

Substituting eqns. 7.65 and 7.48 into 7.54 the closed loop system output is given by

$$y_t = [I+F(z^{-1})][I+K(z^{-1})]^{-1}[I+\Phi(z^{-1})]\, e_t \qquad 7.66$$

where $\Phi(z^{-1})$ is defined as above, and hence the required result is obtained.

7.3.2 Multivariable Self Tuning Detuned Minimum Variance Regulator

The SISO system detuned minimum variance self tuning regulator has often proved to be more useful in practical situations than ordinary minimum variance (e.g. Wellstead and Zanker 1978) largely because the additional freedom to adjust, albeit with restrictions, closed loop system poles can be used to obtain a smoother control. This observation has encouraged the development of the multivariable version of the self tuning rule. The design of the detuned minimum variance regulator has been discussed in section 6.3, and it is shown here that the self tuning rule entails a very simple modification to the basic self tuning minimum variance law already discussed in section 7.2.1. The proof follows from the Generalized Self Tuning Lemma as is shown below.

Let the plant, defined by the ARMAX system in eqn. 6.1, be

modelled as in eqn. 7.27 with k_m set to k, that is

$$\left[I + z^{-k}\hat{A}(z^{-1})\right] y_t = z^{-k}\hat{B}(z^{-1})\, u_t + \varepsilon_t \qquad 7.67$$

where the previous definitions of variables hold. The objective is to produce a closed loop response

$$\left[I + z^{-k}T(z^{-1})\right] y_t = \left[I + M(z^{-1})\right] e_t \qquad 7.68$$

where the variables follow the definitions given in section 6.3. The regulator is to take the form

$$\left[I + \tilde{F}(z^{-1})\right] u_t = \tilde{G}(z^{-1})\, y_t \qquad 7.69$$

as defined in eqn. 7.41.

The result of the lemma indicates that $I+K(z^{-1})$ should be chosen as follows:

$$I+K(z^{-1}) = \left[I+z^{-k}T(z^{-1})\right]\left[I+F(z^{-1})\right] \qquad 7.70$$

where $n_t \leq n_a - n_c$ to satisfy condition C3 of the Lemma. Substituting eqn. 7.70 into eqn. 7.36 gives:

$$\left[I+z^{-k}\hat{A}(z^{-1})\right]\left[I+F(z^{-1})\right] - z^{-k}\hat{B}(z^{-1})G(z^{-1})$$
$$= \left[I+z^{-k}T(z^{-1})\right]\left[I+F(z^{-1})\right] \qquad 7.71$$

Postmultiplying by $\left[I+F(z^{-1})\right]^{-1}$ and invoking the equivalence from eqn. 7.41, viz.

$$\left[I+\tilde{F}(z^{-1})\right]^{-1}\tilde{G}(z^{-1}) = G(z^{-1})\left[I+F(z^{-1})\right]^{-1} \qquad 7.72$$

leads to:

$$I+z^{-k}\hat{A}(z^{-1}) - z^{-k}\hat{B}(z^{-1})\left[I+\tilde{F}(z^{-1})\right]^{-1}\tilde{G}(z^{-1}) = I+z^{-k}T(z^{-1}) \qquad 7.73$$

The equation is satisfied by the solution

$$I+\tilde{F}(z^{-1}) = z\hat{B}_1^{-1}\hat{B}(z^{-1}) \qquad 7.74$$

$$\tilde{G}(z^{-1}) = z\hat{B}_1^{-1}[\hat{A}(z^{-1})-T(z^{-1})]$$

leading to the self tuning control law

$$\hat{B}(z^{-1})u_{t+1} = [\hat{A}(z^{-1}) - T(z^{-1})]\, y_{t+1} \qquad 7.75$$

B_1 and $\hat{B}_1$ (the first coefficient matrices of polynomials $B(z^{-1})$ and $\hat{B}(z^{-1})$) must be non-singular, as in the ordinary minimum variance self tuner. Then, if the Lemma conditions are satisfied the closed loop system output will become

$$[I+z^{-k}T(z^{-1})]y_t = [I+\Phi(z^{-1})]\, e_t \qquad 7.76$$

which is similar to the required output in eqn. 7.68.

It remains to show that the kth order matrix polynomial $I+\Phi(z^{-1})$ is in fact $I+M(z^{-1})$. Substitute control law eqn. 7.69 into the system eqn. 6.1 yielding:

$$\{I+A(z^{-1}) - z^{-k}B(z^{-1})[I+\tilde{F}(z^{-1})]^{-1}\tilde{G}(z^{-1})\}y_t = \{I+C(z^{-1})\}e_t \qquad 7.77$$

Now define a polynomial matrix $I+P(z^{-1})$ of order k such that

$$[I+P(z^{-1})][I+A(z^{-1})] = I+z^{-k}A'(z^{-1}) \qquad 7.78$$

namely, $P(z^{-1})$ satisfies the equation

$$\begin{bmatrix} I & & & \\ A_1^T & I & & \\ \vdots & \ddots & \ddots & \\ A_{k-1}^T & A_{k-2}^T & \cdots\cdots & I \end{bmatrix} \begin{bmatrix} P_1^T \\ \vdots \\ P_k^T \end{bmatrix} = \begin{bmatrix} -A_1^T \\ \vdots \\ -A_k^T \end{bmatrix} \qquad 7.79$$

and $A'(z^{-1})$ has order n_a. Substituting the closed loop system response given in eqn. 7.76 for y_t in eqn 7.77 and premultiplying by $I+P(z^{-1})$ gives:

$$\{I+z^{-k}[A'(z^{-1})-[I+P(z^{-1})]B(z^{-1})(I+\tilde{F}(z^{-1}))^{-1}\tilde{G}(z^{-1})]\} *$$

$$\{I+z^{-k}T(z^{-1})\}^{-1}\{I+\Phi(z^{-1})\}e_t = \{I+P(z^{-1})\}\{I+C(z^{-1})\}e_t \qquad 7.80$$

Using the fact that e_t is a full rank process, pre-multiplying eqn. 7.80 further by $I+z^{-k}T(z^{-1})$ and equating the first k+1 polynomial coefficients on both sides of the equation shows that

$$I+\Phi(z^{-1}) = {}_k\{[I+z^{-k}T(z^{-1})][I+P(z^{-1})][I+C(z^{-1})]\} \qquad 7.81$$

where ${}_k\{\ .\ \}$ indicates the first k+1 terms of polynomial (.).

From eqn. 6.5, $I+M(z^{-1})$ is defined by

$$I+C(z^{-1}) = [I+A(z^{-1})][I+M(z^{-1})] - z^{-k-1}\tilde{G}_c(z^{-1}) \qquad 7.82$$

where $\tilde{G}_c(z^{-1}) = \tilde{G}_{c_o} + \tilde{G}_{c_1}z^{-1} + \ldots + \tilde{G}_{c_{n_{gc}}}z^{-n_{gc}}$ 7.83

Therefore

$$\{I+z^{-k}T(z^{-1})\}\{I+P(z^{-1})\}\{I+C(z^{-1})\} = I+M(z^{-1})+z^{-k-1}N(z^{-1}) \qquad 7.84$$

where $N(z^{-1}) = N_o + N_1z^{-1} + \ldots + N_{n_n}z^{-n_n}$

(i.e. $N(z^{-1})$ does not affect the coefficients of z^{-1}, z^{-2},.., z^{-k} on the RHS of eqn. 7.84) and substituting into eqn. 7.81 it is shown that

$$I+\Phi(z^{-1}) = I+M(z^{-1}) \qquad 7.85$$

as required. Then the self tuned closed loop system is described by

$$\{I+z^{-k}T(z^{-1})\}y_t = \{I+M(z^{-1})\}\ e_t \qquad 7.86$$

which is precisely the result obtained in the off line design case where polynomials $A(z^{-1})$, $B(z^{-1})$ and $C(z^{-1})$ are all assumed to be known. In the self tuning version only $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$ need to be estimated and the property of obtaining the correct closed loop response in this manner (and without the knowledge of the noise colouration) is referred to as the 'Self Tuning Property'. It follows from the Lemma that the residual sequence is defined by

$$\varepsilon_t = \{I+M(z^{-1})\}\, e_t \qquad 7.87$$

The proof of the detuned minimum variance self tuning regulator is therefore a straight-forward development from the Lemma.

The algorithm is summarized as follows:

1. Select the estimation model according to eqn. 7.67 and estimate the parameters using a recursive least squares algorithm.
2. Generate the regulator output using the law of eqn. 7.75 and $n_t \leqslant n_a - n_c$.
3. Then if the conditions of the Lemma are satisfied the closed loop system output is given by eqn. 7.86.

The algorithm clearly reduces to ordinary minimum variance for $T(z^{-1})=0$. Restrictions on the $I+z^{-k}T(z^{-1})$ polynomial are rather severe, both in the order constraint on $T(z^{-1})$ and the k zero coefficients of z^{-1}, z^{-2}, ... , z^{-k}. However the algorithm is extremely simple to implement and is computationally fast. The comments made on minimum variance self tuning in section 7.2.1 concerning non-minimum phase systems, non-integer time delays and the requirement for equal loop time delays are of course also applicable here.

A simulation example is now presented to demonstrate the algorithm in operation.

Example: This is an extension of Example 2 discussed for off-line detuned minimum variance regulator design in section 6.3. It was shown there that the required regulator law is:

$$(I+F_1z^{-1}+F_2z^{-2})\ u_t = (G_o+G_1z^{-1})\ y_t \qquad 7.88$$

where

$$F_1 = \begin{bmatrix} -2.0113 & 4.2484 \\ -0.471 & 2.111 \end{bmatrix} \qquad F_2 = \begin{bmatrix} 2.1038 & 1.222 \\ 0.9529 & 0.3842 \end{bmatrix}$$

$$G_o = \begin{bmatrix} 1.8 & -2.185 \\ 0.41 & 0.16 \end{bmatrix} \qquad G_1 = \begin{bmatrix} -0.05951 & -0.12225 \\ -0.0608 & 0.16638 \end{bmatrix}$$

5000 steps were simulated. The initial estimator covariance matrix was chosen to be 1000I and the forgetting factor, λ, was initially set to 0.96 being subsequently adjusted at each step according to

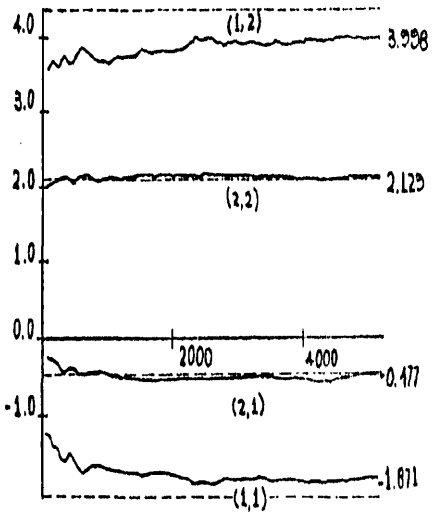
$$\lambda_{k+1} = 0.99\lambda_k + 0.01$$

so that the forgetting factor eventually tends to unity and 'forgets' itself. This method is used to aid in 'tuning in' at the start of a run.

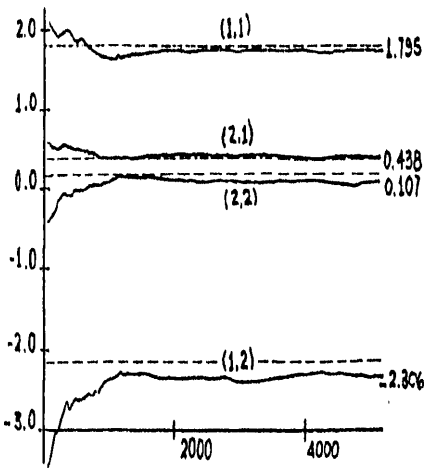
The regulator parameter estimates are plotted at intervals of 50 steps in Fig. 7.2. The values at the final (5000th) step are:

$$\hat{F}_1 = \begin{bmatrix} -1.871 & 3.998 \\ -0.477 & 2.129 \end{bmatrix} \qquad \hat{F}_2 \begin{bmatrix} 1.941 & 1.146 \\ 0.957 & 0.386 \end{bmatrix} \qquad 7.89$$

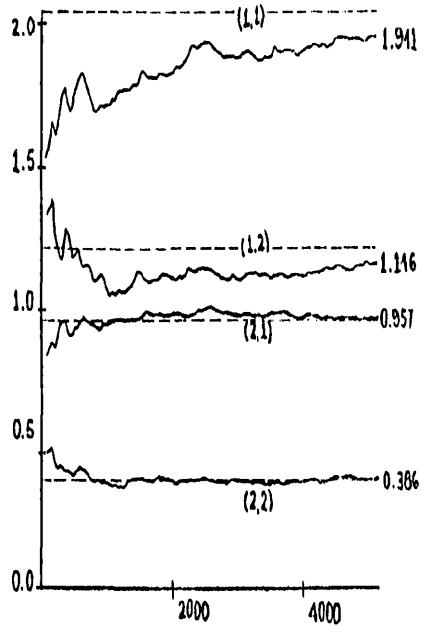
$$\hat{G}_o \begin{bmatrix} 1.795 & -2.306 \\ 0.438 & 0.107 \end{bmatrix} \qquad \hat{G}_1 \begin{bmatrix} -0.0528 & -0.309 \\ -0.0475 & 0.182 \end{bmatrix}$$



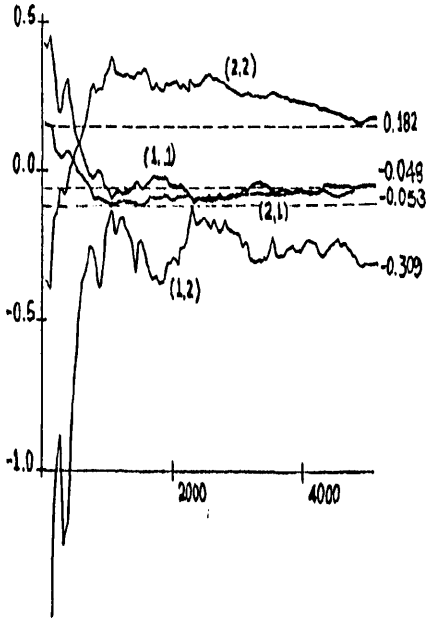
F_1 matrix



G_0 matrix



F_2 matrix



G_1 matrix

Fig. 7.2 Time evolution of regulator parameters

Examination of the parameter estimate plots shows that $\hat{G}_1$ has not converged fully yet, but the self tuned regulator is fairly close to the correct law considering the number of parameters that must be tuned.

7.3.3 Multivariable Self Tuning Pole Shifting Regulator

The self tuning regulator to be described in this section is based on the pole-shifting method detailed in section 6.6 where the features of this control strategy are also discussed. The proof of the algorithm follows naturally from the Self Tuning Lemma.

Let the plant, defined by the ARMAX system in eqn. 6.1, be modelled as in eqn. 7.27 with k_m set to 0, so that the estimation model becomes

$$[I+\hat{A}(z^{-1})]\ y_t = \hat{B}(z^{-1})\ u_t + \varepsilon_t \qquad 7.90$$

where the definition of the variables remains as given in the Lemma. The objective is to produce a closed loop system response

$$y_t = \{I+F(z^{-1})\}\{I+T(z^{-1})\}^{-1}\ e_t \qquad 7.91$$

as defined in section 6.6, using a regulator law of the form

$$u_t = G(z^{-1})\{I+F(z^{-1})\}^{-1}\ y_t \qquad 7.92$$

as in eqn. 7.31 of the Lemma. From the results of the Lemma it is clear that the closed loop response in eqn. 7.91 will be obtained if polynomial $I+K(z^{-1})$ is specified as

$$I+K(z^{-1}) = I+T(z^{-1}) \qquad 7.93$$

(where $T(z^{-1})$ is open to the designers choice) and if the conditions

of the Lemma are met. In order to satisfy C3, n_t, the order of $I+T(z^{-1})$ must be chosen so that

$$n_t \leq n_a+n_b+k-n_c-1 \qquad 7.94$$

Eqns. 7.93 and 7.36 imply that the regulator parameters must be chosen so as to satisfy

$$I+T(z^{-1}) = \{I+\hat{A}(z^{-1})\}\{I+F(z^{-1})\} - \hat{B}(z^{-1})G(z^{-1}) \qquad 7.95$$

Unlike the detuned minimum variance self tuning regulator, the solution of this design equation is not trivial and is computed from the following set of linear equations derived by equating coefficients of z^{-1}, z^{-2}, ... , etc. in eqn. 7.95:

$$\underbrace{\begin{bmatrix} I & & -\hat{B}_1 & & \\ \hat{A}_1 & \ddots & & \ddots & -\hat{B}_1 \\ \vdots & & -\hat{B}_{n_{\hat{b}}} & & \vdots \\ \hat{A}_{n_{\hat{a}}} & \ddots & & & -\hat{B}_{n_{\hat{b}}} \end{bmatrix}}_{\substack{n_f \text{ block columns} \\ n_g+1 \text{ block columns} \\ \text{Matrix } \Psi}} \begin{bmatrix} F_1 \\ \vdots \\ F_{n_f} \\ G_o \\ \vdots \\ G_{n_g} \end{bmatrix} = \begin{bmatrix} T_1 \\ \vdots \\ T_{n_t} \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} \hat{A}_1 \\ \vdots \\ \hat{A}_{n_{\hat{a}}} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad 7.96$$

Since the elements of the transformation matrix Ψ on the left hand side of eqn. 7.96 are estimated values (and are therefore subject to noise) it is claimed that the matrix will generally be non-singular and a solution to eqn. 7.95 will exist.

Notice that the matrix polynomial $I+\Phi(z^{-1})$ in the Lemma reduces to the identity matrix, leading to the interesting result

$$\varepsilon_t = e_t \qquad 7.97$$

The poles of the closed loop system depend on $|I+T(z^{-1})|$ which may be specified by the designer. It is argued that the situation where $I+F(z^{-1})$ and $I+T(z^{-1})$ are not relatively right prime (thus causing pole-zero cancellation) will not in general occur, especially as $I+F(z^{-1})$ is determined from estimated parameters $\hat{A}(z^{-1})$, $\hat{B}(z^{-1})$. In the case of SISO systems, specification of $1+T(z^{-1})$ clearly defines the system closed loop transfer function denominator. The designer does not, however, have the freedom to adjust the system transfer function 'numerator polynomial' or system zeros.

The 'Self Tuning Property' for the pole shifting self tuning regulator may now be stated.

Self Tuning Property:

Let the plant be described by the linear difference equation given in eqn. 6.1 and be modelled by the difference equation of eqn. 7.90 and let the regulator law of eqn. 7.92 apply. Furthermore assume that the conditions attached to the Generalized Self Tuning Lemma hold. Then the on-line solution for the parameters $F(z^{-1})$ and $G(z^{-1})$ of the equation

$$\{I+\hat{A}(z^{-1})\}\{I+F(z^{-1})\} - \hat{B}(z^{-1})G(z^{-1}) = I+T(z^{-1}) \qquad 7.98$$

will yield the same control law and closed loop system response

$$y_t = \{I+F(z^{-1})\}\{I+T(z^{-1})\}^{-1} e_t \qquad 7.99$$

as the control law designed off-line by the solution of the pole-shifting design equation

$$\{I+A(z^{-1})\}\{I+F(z^{-1})\} - z^{-k}B(z^{-1})G(z^{-1})$$

$$= \{I+C(z^{-1})\}\{I+T(z^{-1})\} \qquad 7.100$$

7.3.3.1 A Justification of the Model Equation

The model equation 7.90 is interesting in that the system time delay k enters the equation only in the definition of the order of polynomial $B(z^{-1})$. Unlike the minimum variance regulators, the delay term z^{-k} (see eqn. 7.67) is absent. In order to demonstrate how such a model may in fact represent the plant eqn. 6.1 it is assumed that the self tuner has converged properly so that the system output and input are given by

$$y_t = \{I+F(z^{-1})\}\{I+T(z^{-1})\}^{-1}e_t \qquad 7.101$$

and

$$u_t = G(z^{-1})\{I+T(z^{-1})\}^{-1}e_t \qquad 7.102$$

Now, postmultiply the control design equation, eqn. 7.95, by $\{I+T(z^{-1})\}^{-1}e_t$ yielding

$$\{I+\hat{A}(z^{-1})\}\{I+F(z^{-1})\}\{I+T(z^{-1})\}^{-1}e_t$$

$$- \hat{B}(z^{-1})G(z^{-1})\{I+T(z^{-1})\}^{-1}e_t = e_t \qquad 7.103$$

Substituting 7.101 and 7.102 into 7.103, the chosen model form emerges:

$$\{I+\hat{A}(z^{-1})\}y_t = \hat{B}(z^{-1})u_t + e_t \qquad 7.104$$

During the tuning-in phase, e_t is of course replaced by the residual ε_t.

If the control law satisfies the conditions on it (C1 of Lemma) then polynomials $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$ always exist. To see this, solve eqn. 7.95 for $\hat{A}(z^{-1})$ and $\hat{B}(z^{-1})$. One coefficient matrix may in general be specified arbitrarily. (The solution requires a matrix of the form of Q in eqn. 7.51 to be non-singular.)

Note that the pole shifting rule does not in general require $\hat{B}_1$ or B_1 to be non-singular (as is the case for minimum variance self tuners). This is important, as in the initial tuning-in phase $\hat{B}_1$ in the estimation model could easily become singular if the system pure time delay k is in fact non-zero. It is however necessary that the matrix Ψ in eqn. 7.96 be non-singular, as already discussed, and if initial parameter estimates for the estimation model are chosen with this in mind difficulties do not generally occur in practice.

7.3.3.2 Time Varying Time Delays

The structure of the estimation model equation is particularly suited to dealing with the problem of slowly varying system time delays, a complication not uncommon in real processes. As pointed out in the previous section, the system time delay enters the estimation model equation through the order of the $\hat{B}(z^{-1})$ polynomial. If $n_{\hat{B}} = n_b + k_{max}$ where k_{max} is the largest expected system pure time delay (in multiples of the sampling period) it is possible for the self tuning regulator to converge correctly. Some formal difficulties in the proof arise in that the model becomes 'over-parametrized' when the pure time delay is actually less than k_{max}. In particular, $I+F(z^{-1})$ will be over-parametrized so that the matrix Q in eqn. 7.51

may lose rank. In this case the self tuning property may be lost but this does not imply loss of control or that the system will go unstable. It merely means that the control law parameters will fail to converge exactly to the values required by theory. In fact the practical application of the self tuner is not generally affected as the notion of an underlying 'finite order' linear time-invariant system which leads to the formal complications is more often than not a mathematical fiction and a substantial simplification of the true system structure. Thus, formal difficulties associated with model order may be liberally interpreted.

It is a requirement that matrix Ψ in eqn. 7.96 should be non-singular so that it is possible to solve for the regulator parameters. This is generally fulfilled as estimation error will almost always ensure the non-singularity of Ψ.

Another feature of the pole shifting regulator mentioned in the previous section is that the coefficient matrix B_1 in $B(z^{-1})$ need not be non-singular. This is important as it effectively means that the pole shifter can cope with multivariable systems in which the loops have differing pure time delays. The minimum variance regulator cannot regulate such systems. Nor is it easy for minimum variance self tuning regulators to cope with time-varying time delays. Such variations will almost inevitably enter a region in which $B(z^{-1})$ is non-minimum phase and this is likely to trigger instability. An example of this behaviour (for a SISO system) is given in section 7.3.4.2.

A further useful feature of the self-tuning pole shifting

regulator is that, should the time delay exceed k_{max}, the control scheme need not necessarily become unstable. A minimum variance regulator would almost certainly be forced to go unstable.

The general robustness of the pole shifting self tuning regulator is clearly a decided advantage.

7.3.3.3 Implementation

The pole shifting regulator is computationally complex and in this respect is at a disadvantage compared with the minimum variance regulator. The regulator parameters $F(z^{-1})$ and $G(z^{-1})$ are computed by solving the set of $(n_f+n_g+1)p$ simultaneous linear equations given in eqn. 7.96. However, the form of the regulator law

$$u_t = G(z^{-1})\{I+F(z^{-1})\}^{-1}y_t \qquad 7.105$$

is not in general suitable for direct implementation in the multivariable case. For the single variable case there is no difficulty as the polynomials commute and the regulator input u_t can be computed directly from

$$u_t = -F(z^{-1})u_t + G(z^{-1})y_t \qquad 7.106$$

In the case of multivariable systems it is necessary to solve a further $(n_f+n_g)p$- simultaneous equations to derive the polynomials $\tilde{F}(z^{-1})$ and $\tilde{G}(z^{-1})$ as discussed in section 6.6.2 for the off-line design case. The equations that must be solved are those given in eqn. 6.65. Then the control law may be rewritten as:

$$u_t = -\tilde{F}(z^{-1})u_t + \tilde{G}(z^{-1})y_t \qquad 7.107$$

$\tilde{F}(z^{-1})$ and $\tilde{G}(z^{-1})$ are defined in eqn. 7.41.

A simplification occurs when $n_f=0$ as then

$$I+\tilde{F}(z^{-1}) = I \qquad 7.108$$

$$\tilde{G}(z^{-1}) = G(z^{-1})$$

and the transformation of eqn. 6.65 is unnecessary.

In summary, the complete algorithm involves the following steps:

1. At each sampling instant identify the parameters of the estimation model given in eqn. 7.90 using a recursive least squares estimator.
2. Compute $F(z^{-1})$ and $G(z^{-1})$ using eqn. 7.96.
3. Where necessary, compute $\tilde{F}(z^{-1})$ and $\tilde{G}(z^{-1})$ according to eqn.6.65.
4. Apply the regulator law of eqn. 7.107.
5. Return to 1.

Then the closed loop system output tends to

$$y_t = \{I+F(z^{-1})\}\{I+T(z^{-1})\}^{-1} e_t \qquad 7.109$$

where the closed loop system poles have been specified by a suitable choice of $I+T(z^{-1})$, $n_t \leq n_a+n_b+k-n_c-1$.

Typical computation times on DEC-10 and PDP 11/10 machines for a system with $n_a=2$, $n_b=2$, $k=0$, $p=2$ (i.e. 16 estimation parameters) are:

DEC-10	58ms/iteration
PDP 11/10	1 s/iteration

If n_b is reduced to 1 so that n_f=0, the transformation of point 3 in the algorithm is unnecessary and the computation time for the PDP 11/10 (which was without floating point hardware) reduces substantially to 400ms.

Convergence of the regulator parameters is usually slower for the multivariable version than the SISO system version. This is understandable as there are generally more parameters to tune.

7.3.4 Simulation Examples for the Pole-Shifting Self Tuning Regulator

In order to demonstrate the features and operation of the self tuning pole shifting regulator several simulation examples are now presented. All but one were obtained using discrete-time-system digital simulation. The following chapter is devoted to the practical application of self tuning and this algorithm in particular on hardware.

Unless stated to the contrary, the recursive least squares estimator in all the following digital simulation examples was initialized so that the covariance matrix was 100I and the forgetting factor λ was 0.96. Furthermore, the forgetting factor evolved so that

$$\lambda_{k+1} = 0.99\lambda_k + 0.01 \qquad 7.110$$

This technique aids initial tuning, and allows the forgetting factor to gradually increase to unity.

7.3.4.1 Non-Minimum Phase system with time varying time delay

The following SISO system was simulated:

$$(1-1.3z^{-1}+0.4z^{-2})y_t = z^{-k}(z^{-1}+1.5z^{-2})u_t + (1-0.65z^{-1}+0.12z^{-2})e_t \qquad 7.111$$

where e_t is a zero mean white noise process with variance 0.1. The input u_t was generated using the pole-shifting self-tuner law, designed so that

$$1 + T(z^{-1}) = 1 \qquad 7.112$$

5000 steps were simulated. During the first 2499 steps, k (the system pure time delay) was set to 0, but for the rest of the run it was switched to unity. Note that the system is non-minimum phase and that it is therefore not amenable to minimum variance regulation. The example illustrates the behaviour of a pole-shifting regulator under conditions of a time-varying pure time delay and a non-minimum phase system. The estimation model was:

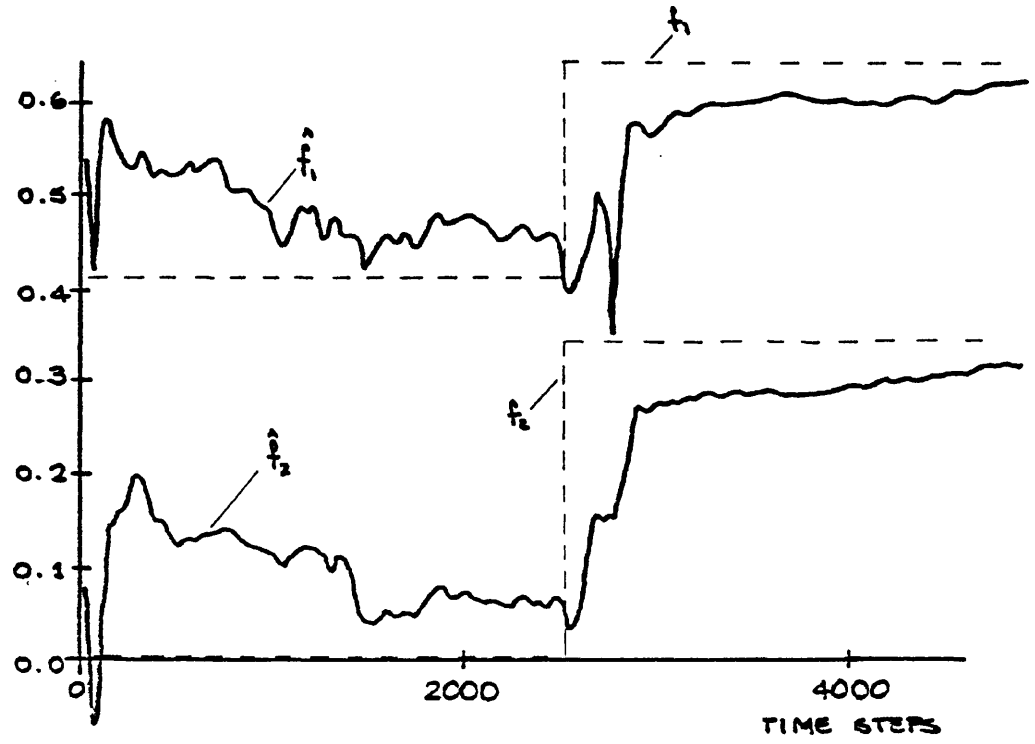
$$(1+\hat{a}_1z^{-1}+\hat{a}_2z^{-2})y_t = (\hat{b}_1z^{-1}+\hat{b}_2z^{-2}+\hat{b}_3z^{-3})u_t + \varepsilon_t \qquad 7.113$$

where the $\hat{B}(z^{-1})$ polynomial has 3 terms to allow for both pure time delays (k=0,k=1). The control law then has the following structure:

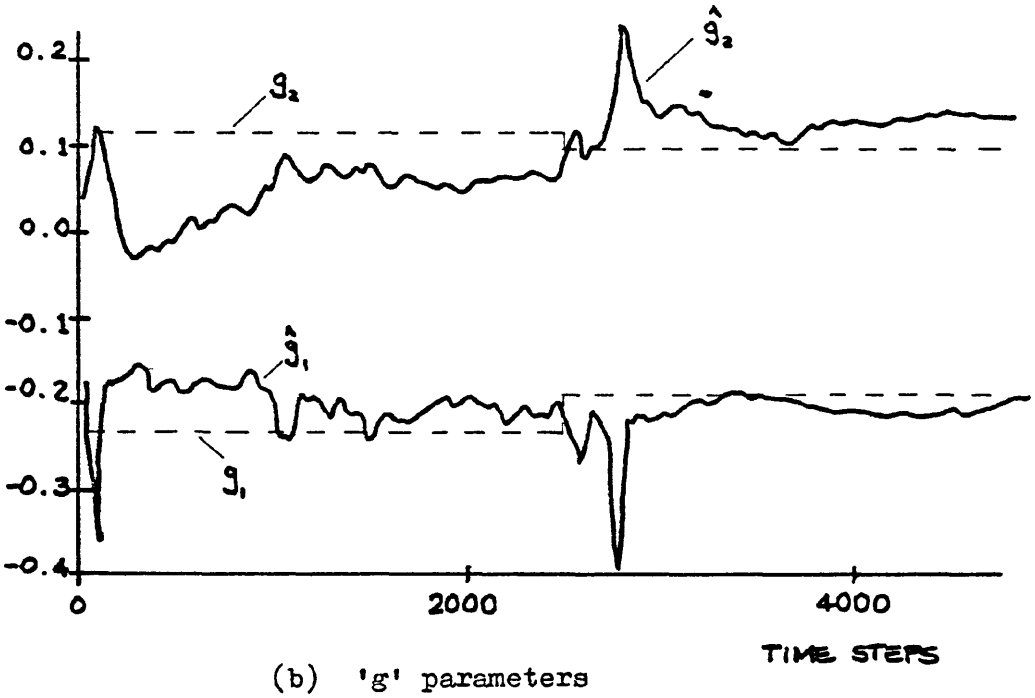
$$(1+f_1z^{-1}+f_2z^{-2})u_t = (g_o+g_1z^{-1})y_t \qquad 7.114$$

where f_2 should tend to zero when k=0.

Figures 7.3a and 7.3b show the time evolution of the control law parameters together with the correct values to which they should ideally converge. The f parameters in particular show a marked



(a) 'f' parameters



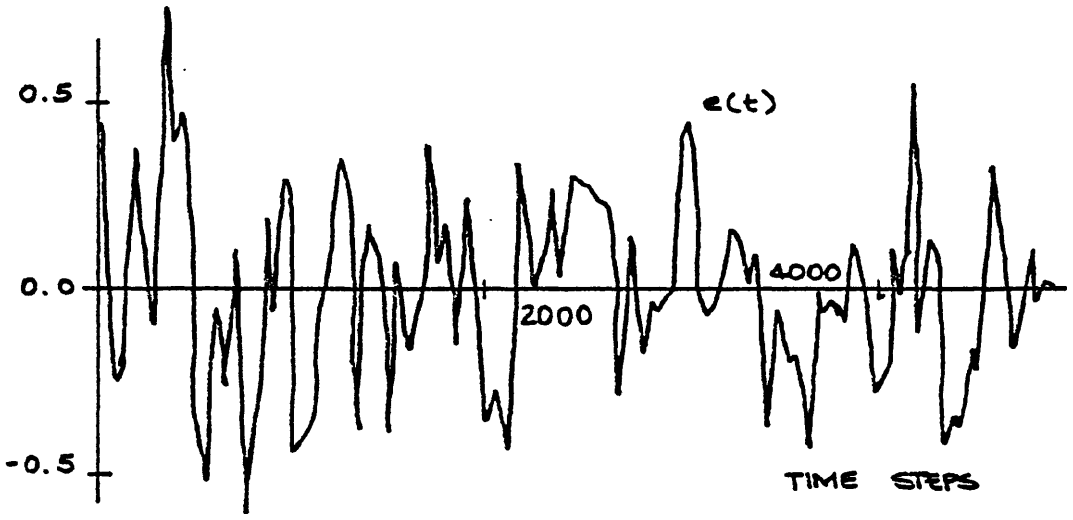
(b) 'g' parameters

Fig. 7.3 Time evolution of regulator parameters
(System time delay changes at step 2500)

change after the 2500 th step, following the change in time delay. The unusually high variability of the parameters during the first 2500 steps is a result of overparametrization which impedes convergence. In a real system where the system order is not exact this is less likely to occur.

Figure 7.3c shows the system driving noise e_t and the difference between e_t and the residual sequence ε_t. When the self tuner has converged correctly, $\varepsilon_t = e_t$. Before the change in time delay the difference is small. At the point where the time delay changes, the sequences initially differ, but soon begin to converge again (as shown by the difference sequence) indicating that the control law is retuning to take account of the change in time delay.

The input and output sequences shown in Figures 7.3d and 7.3e verify that there is no sign of instability as the time delay changes. The asymptotic value of the output variance is 0.1173 and 0.1545 for k=0 and k=1 respectively, compared with 'minimum variance' figures of 0.1 and 0.14225. However, due to the non-minimum phase characteristics of the system, the minimum variance regulator would in practice lead to an unstable system. Furthermore, even if the system had been minimum-phase, it is highly improbable that the minimum variance regulator would have tolerated a change in time delay without, at the very least, some sign of instability. The superior performance of the pole shifting self tuner over the minimum variance self tuner in this respect is demonstrated briefly in the following example using analogue simulation.



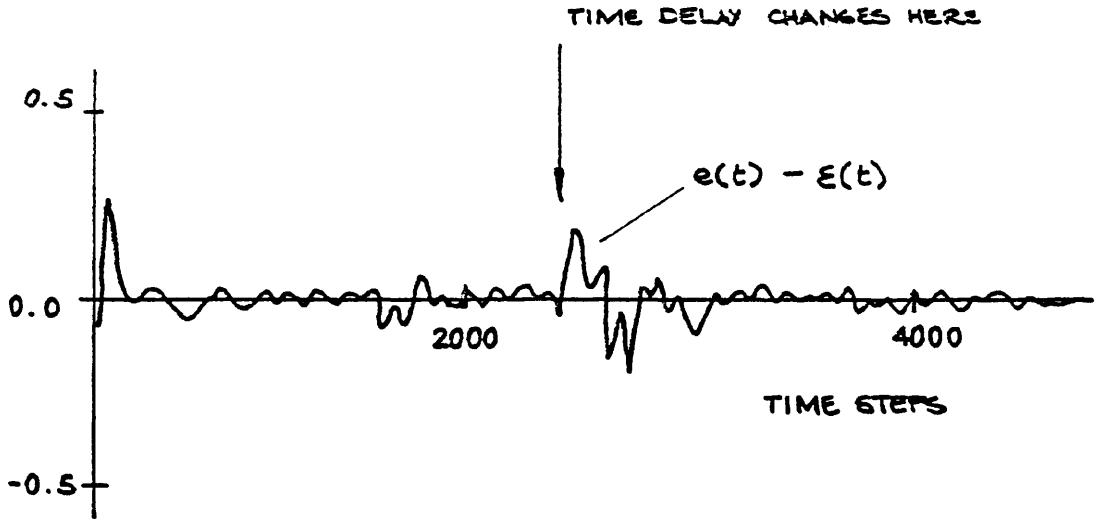


Fig. 7.3(c) System noise e(t) and
difference sequence e(t) - residual sequence $\mathcal{E}(t)$

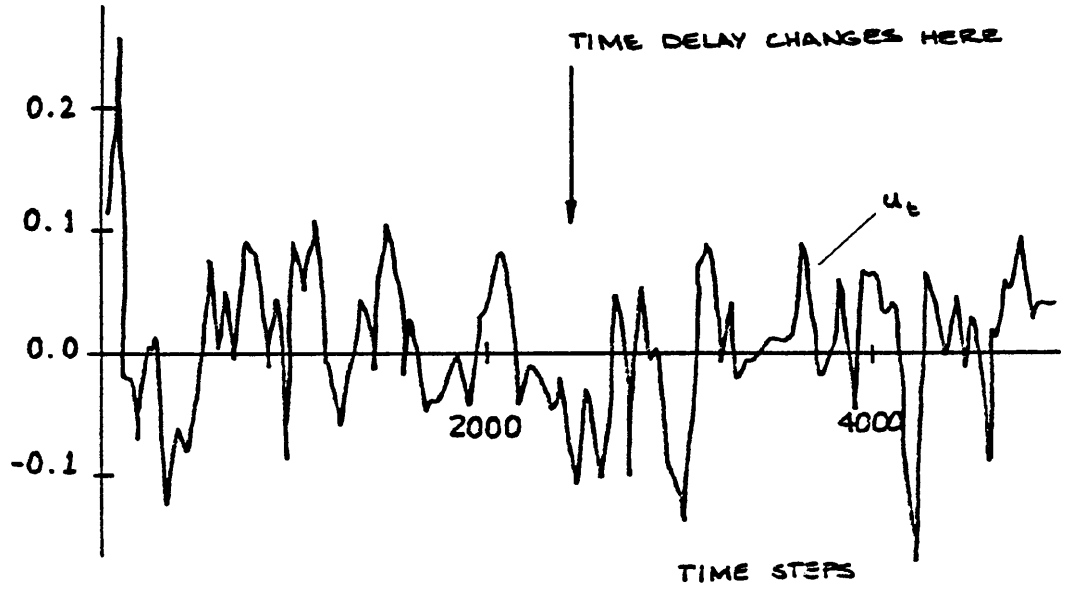


Fig. 7.3 (d) Control Signal

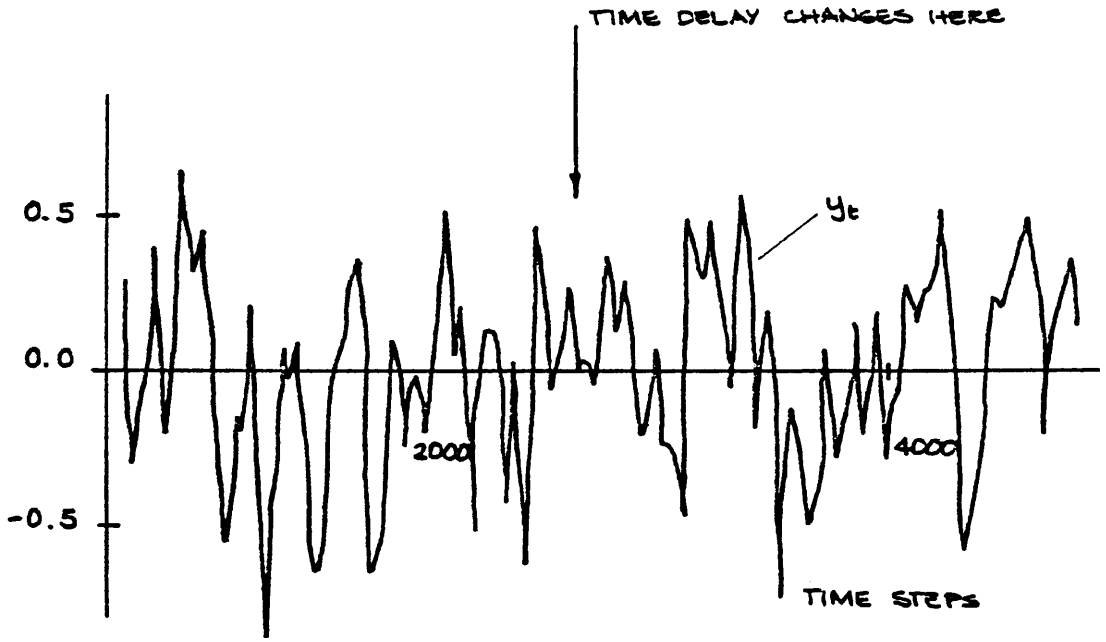


Fig. 7.3 (e) System Output

7.3.4.2 Pole-shifting and Minimum Variance Self Tuning for a Variable Time Delay system

An analogue system defined by

$$H_1(s) = \frac{-0.5e^{-s\tau}}{s-0.25} \qquad 7.115$$

was used to demonstrate the self tuning pole shifting control of an open-loop unstable system with variable time delay. The sampling interval was 2 seconds and the time delay τ was varied from 0.5 to 3.5 times the sampling interval in steps of 0.5. The response of the system to a sinusoidal reference input is shown in Figure 7.4a together with the control signal and time delay k (as a multiple of the sampling period.) Note that the controller rapidly adjusts itself to the new delay.

It was not possible to compare minimum variance control directly as the self tuning minimum variance controller failed to converge. The analogue system was therefore changed to

$$H_2(s) = \frac{-2.5e^{-s\tau}}{s} \qquad 7.116$$

and, when the time delay was over-estimated by the minimum variance self tuning controller, stable control could be achieved for time delays of integer multiples of the sampling period (0.4s). The result is shown in Figure 7.4b for values of time delay k (as a multiple of the sampling period) varying from 1 to 3. Note however that on transitions of the time delay control is partially lost in the intervening non-minimum phase region. Figure 7.4c demonstrates the sensitivity of the minimum variance self tuning controller to non-minimum phase system behaviour. A change of 5% in the time delay brought a discrete-time zero onto the unit disc and caused the self tuner to go unstable.

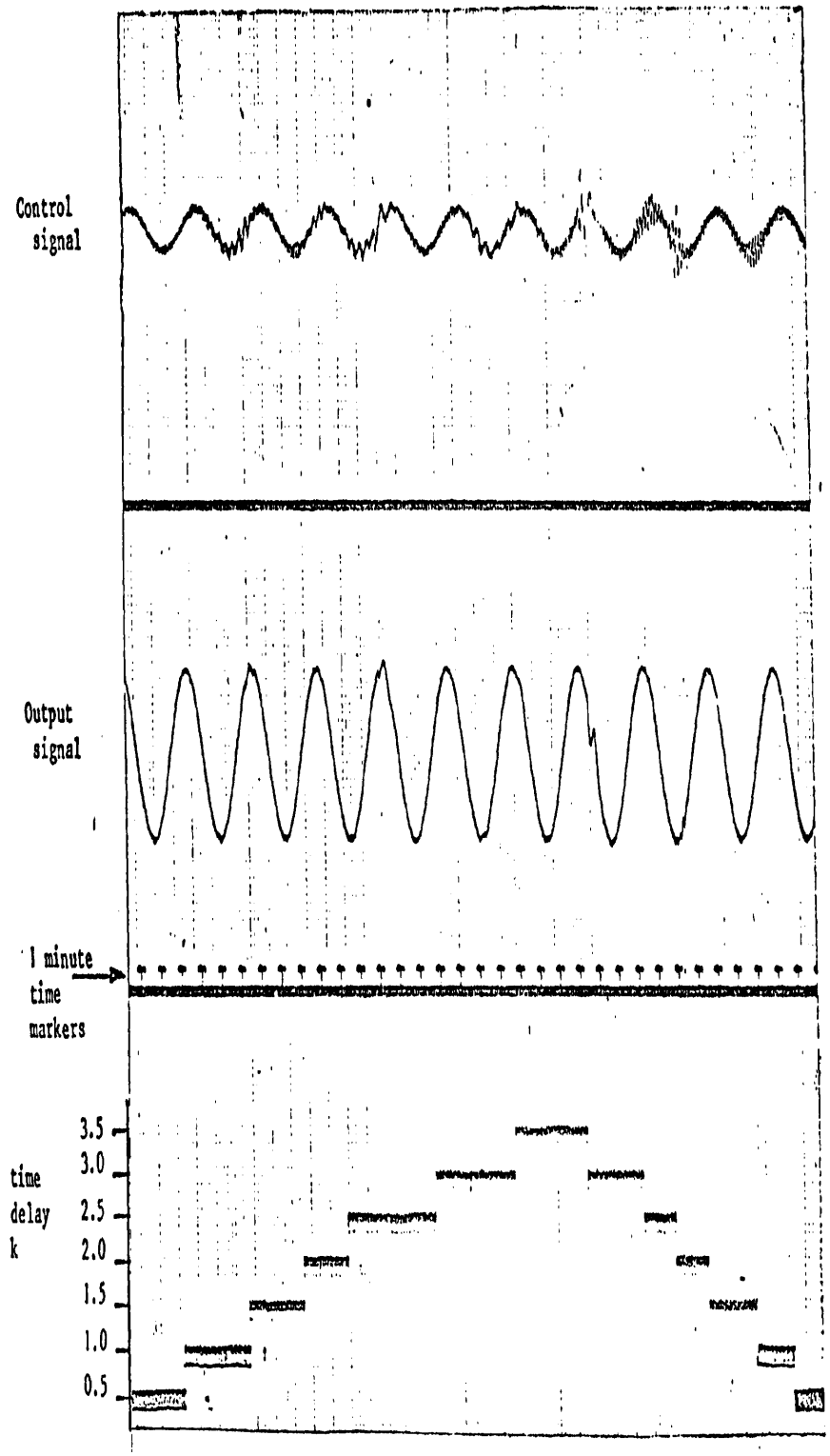


Fig. 7.4(a) Showing the control, output and time delay variations for the open-loop unstable system $H_1(s)$ under pole-assignment self-tuning. The system is tracking a sinusoidal reference input.

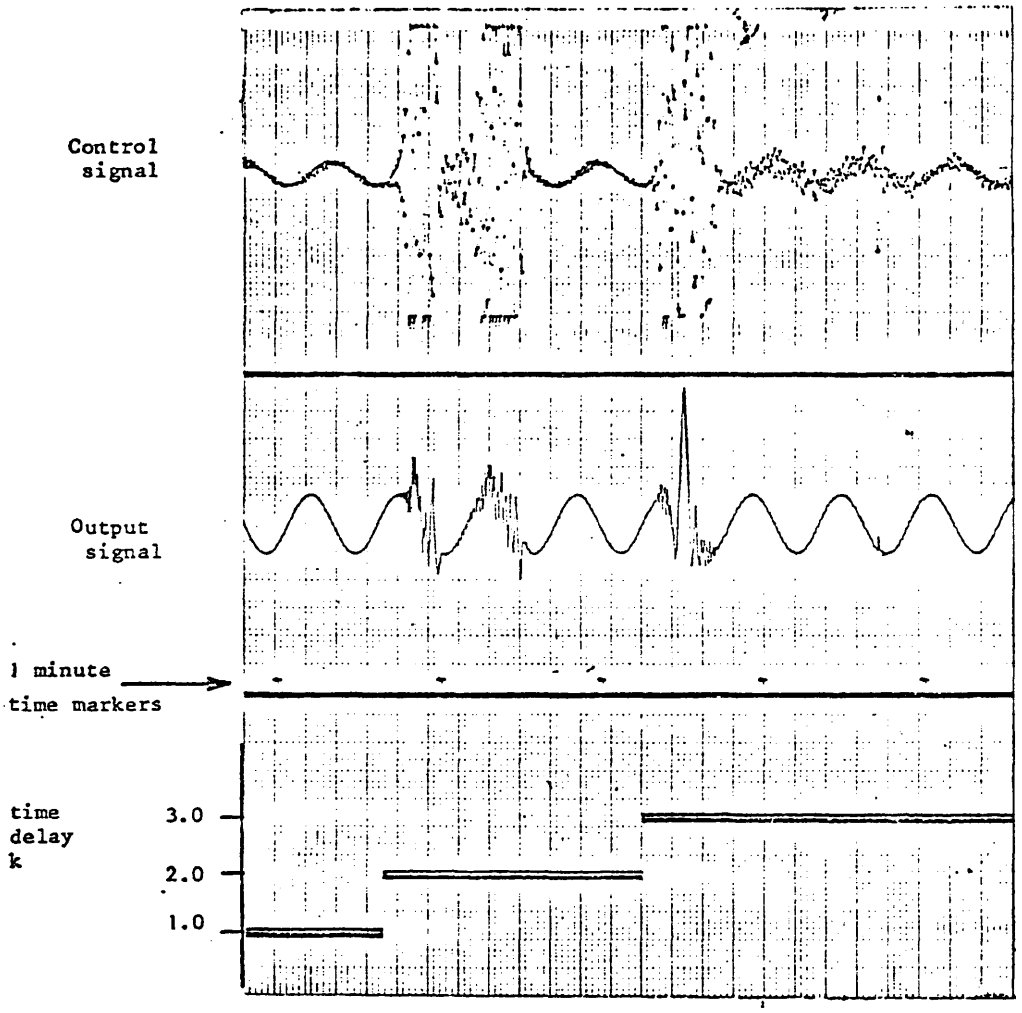


Fig. 7.4(b) Showing the control, output and time delay variations for the integrator $H_2(s)$ under minimum variance self-tuning.
Note that because the delay changes by integer multiples of the sample rate, the minimum variance self-tuner can manage, although the control during changes is poor; compare with Fig. 7.4(a)

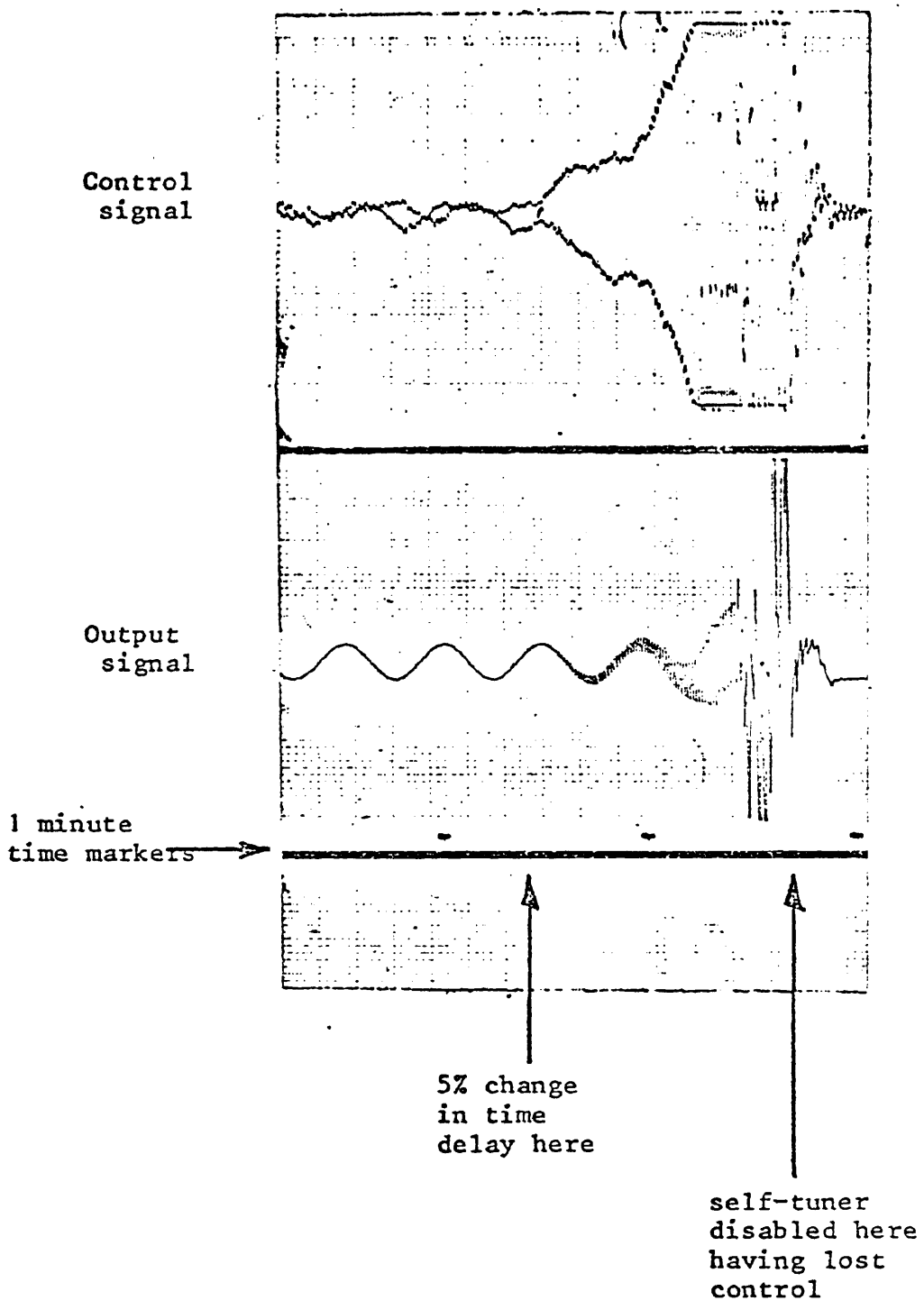


Fig. 7.4(c) Showing the control and output signals for the integrator $H_2(s)$ under minimum variance self-tuning. Note the sensitivity to small changes in time delay, where a 5% change causes instability by inducing non-minimum phase behaviour.

7.3.4.3 Multivariable Pole-shifting Self Tuning for differing loop-Time-Delay Systems

This is an extension of the example for off-line multivariable pole shifting regulator design in section 6.6.4. It is recalled that B_1 in this example was singular, output 2 having in effect a pure time delay of 1 sampling period as against 0 for output 1. Also, a zero of $|B(z^{-1})|$ lies outside the stability region. The example therefore would not be amenable to minimum variance regulation.

The estimation model used for the self-tuning version was:

$$(I+\hat{A}_1 z^{-1}+\hat{A}_2 z^{-2})y_t = (\hat{B}_1 z^{-1}+\hat{B}_2 z^{-2})u_t + \varepsilon_t \qquad 7.117$$

leading to a control law structure:

$$(I+\hat{F}_1 z^{-1})u_t = (\hat{G}_o+\hat{G}_1 z^{-1})y_t \qquad 7.118$$

The pole shifting self tuning regulator was run for 3000 steps. The time evolution of the control parameters is plotted in Figure 7.5 and the final values are shown below together with the theoretical values to which they should converge which were computed in section 6.6.4:

$$\hat{F}_1 = \begin{bmatrix} 0.275 & 0.213 \\ 0.139 & 0.292 \end{bmatrix} \qquad F_1 = \begin{bmatrix} 0.213 & 0.203 \\ 0.155 & 0.286 \end{bmatrix}$$

$$\hat{G}_o = \begin{bmatrix} -0.107 & -0.0582 \\ -0.153 & -0.106 \end{bmatrix} \qquad G_o = \begin{bmatrix} -0.101 & -0.0483 \\ -0.16 & -0.115 \end{bmatrix}$$

$$\hat{G}_1 = \begin{bmatrix} 0.0707 & 0.0382 \\ 0.0459 & 0.0332 \end{bmatrix} \qquad G_1 = \begin{bmatrix} 0.0682 & 0.0265 \\ 0.0495 & 0.0469 \end{bmatrix}$$

7.119

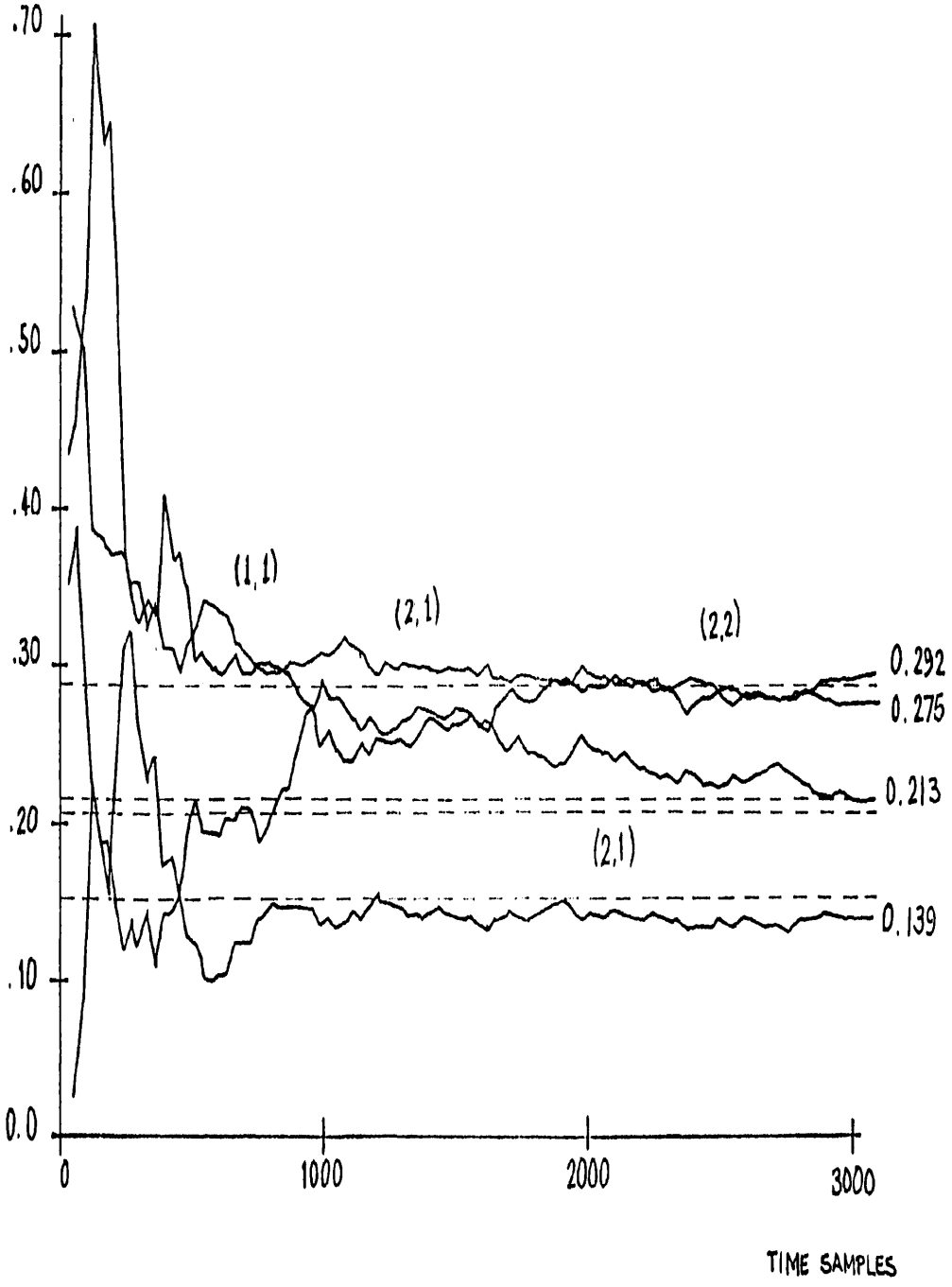


Fig. 7.5(a) Time evolution of regulator parameters, F_1 matrix

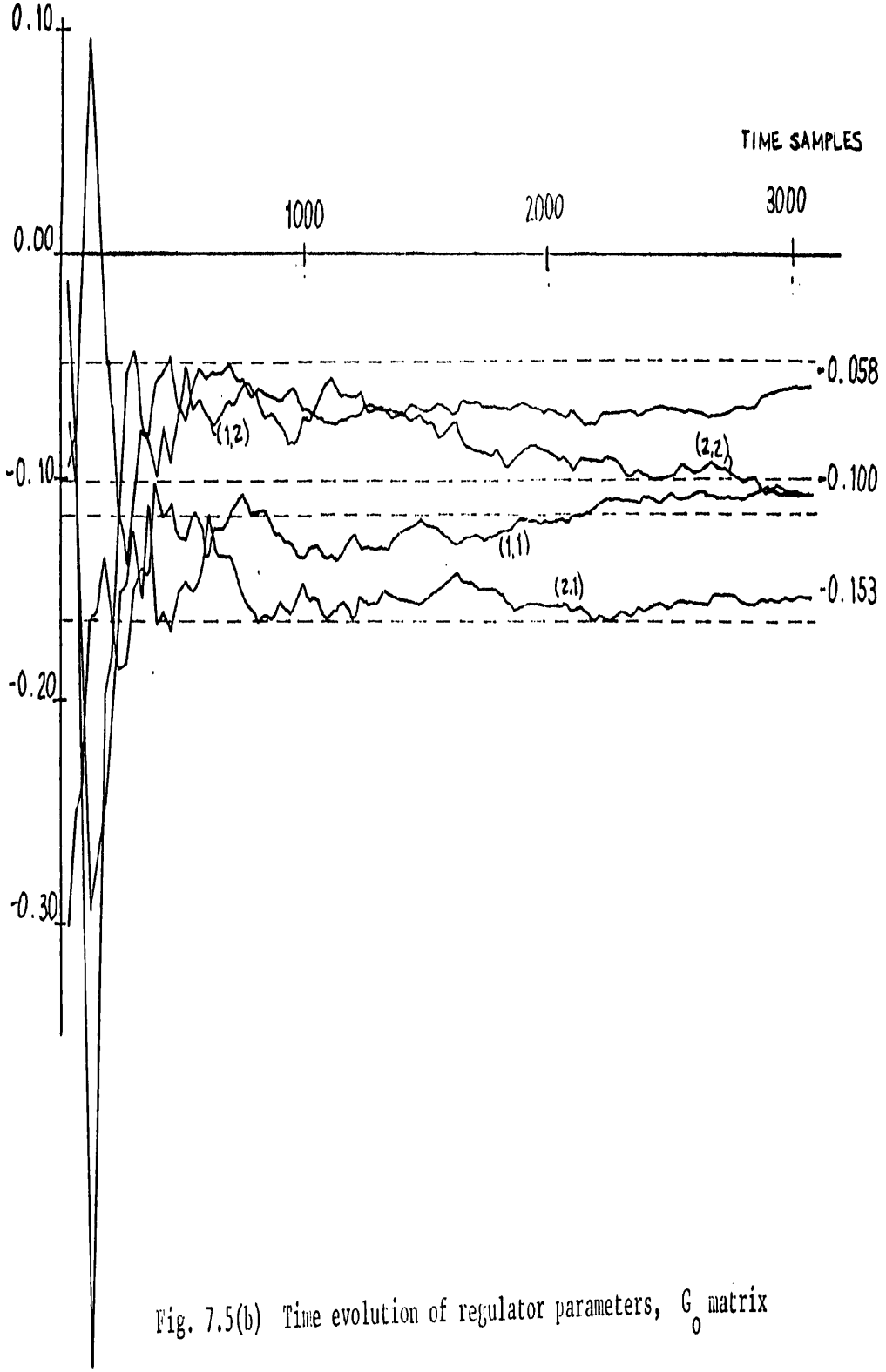


Fig. 7.5(b) Time evolution of regulator parameters, G_0 matrix

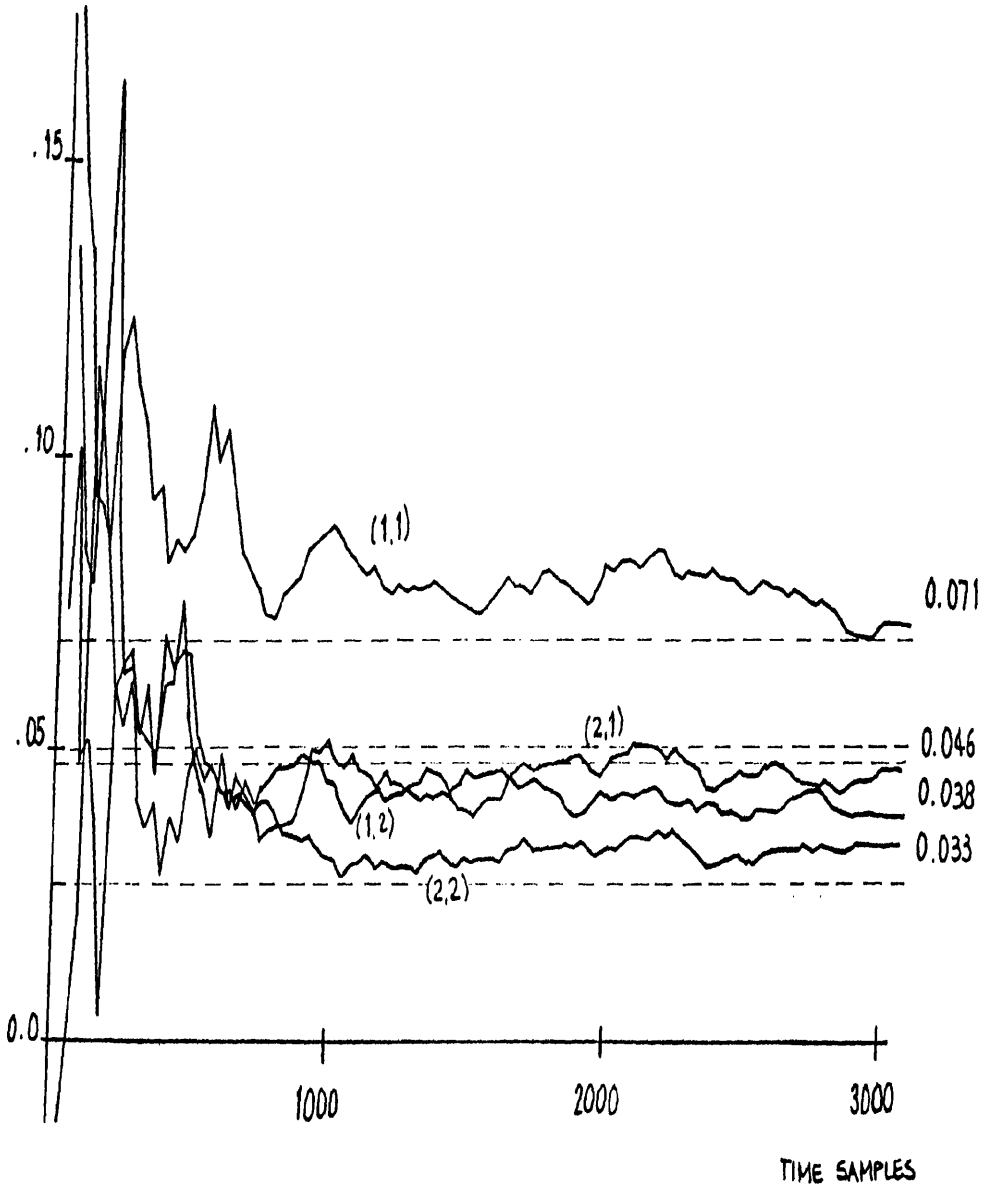


Fig. 7.5(c) Time evolution of regulator parameters, G_1 matrix

The estimated parameters have not converged exactly, but are close to the desired values. The pole-shifting regulator clearly operates well in a situation where minimum variance regulation would fail.

7.3.4.4 Multivariable Pole-Shifting Regulation Example

Finally a second example of a multivariable system under self tuning pole-shifting regulation is given. The system is described by:

$$(I+A_1z^{-1}+A_2z^{-2})y_t = (B_1z^{-1}+B_2z^{-2})u_t + (I+C_1z^{-1})e_t \qquad 7.120$$

where:
$$A_1 = \begin{bmatrix} 0.8 & -2.9 \\ 0.4 & -1.4 \end{bmatrix} \qquad A_2 = \begin{bmatrix} 0.2 & -1.75 \\ 0.2 & -0.95 \end{bmatrix}$$

$$B_1 = I \qquad B_2 = \begin{bmatrix} -0.4 & 0.6 \\ 0.4 & 0.6 \end{bmatrix}$$

$$C_1 = \begin{bmatrix} -0.25 & 0 \\ 0 & -0.45 \end{bmatrix}$$

and the statistics of the white noise process e_t are

$$E(e_t) = 0$$
$$E(e_te_t^T) = 0.1I$$

was simulated for 4000 steps under a self tuning pole-shifting regulator. The regulator was designed to set

$$I + T(z^{-1}) = I + \begin{bmatrix} -0.1 & 0 \\ 0 & -0.3 \end{bmatrix} z^{-1} \qquad 7.121$$

The estimation model

$$(I+\hat{A}_1z^{-1}+\hat{A}_2z^{-2})y_t = (\hat{B}_1z^{-1}+\hat{B}_2z^{-2})u_t + \varepsilon_t \qquad 7.122$$

was used, leading to a regulator having the structure

$$(I+\hat{F}_1 z^{-1})u_t = (\hat{G}_o+\hat{G}_1 z^{-1})y_t \qquad 7.123$$

The time evolution of the regulator parameter estimates are shown in Figure 7.6a. The final estimated regulator coefficient matrices $(\hat{F}_1,\hat{G}_o,\hat{G}_1)$ at step 4000 are compared with the values computed using the off-line design equation (F_1,G_o,G_1) below:

$$\hat{F}_1 = \begin{bmatrix} -1.601 & -3.068 \\ 0.137 & -0.228 \end{bmatrix} \qquad F_1 = \begin{bmatrix} -1.687 & -3.398 \\ 0.09881 & -0.3307 \end{bmatrix}$$

$$\hat{G}_o = \begin{bmatrix} -0.519 & -1.555 \\ 0.038 & -0.874 \end{bmatrix} \qquad G_o = \begin{bmatrix} -0.6427 & -1.6438 \\ 0.00972 & -0.875 \end{bmatrix}$$

$$\hat{G}_1 = \begin{bmatrix} -1.025 & 5.348 \\ -0.078 & 0.665 \end{bmatrix} \qquad G_1 = \begin{bmatrix} -1.1328 & 5.959 \\ -0.11023 & 0.8428 \end{bmatrix}$$

7.124

When the self tuner has converged correctly, the residual sequence ε_t equals the system driving noise e_t. Figure 7.6b shows the elements of these two sequences superimposed, i.e. $e_t^{(1)}$ and $\varepsilon_t^{(1)}$ are superimposed and $e_t^{(2)}$ and $\varepsilon_t^{(2)}$ are superimposed, where $e_t = (e_t^{(1)}, e_t^{(2)})^T$. The convergence of the algorithm is demonstrated by the convergence of the ε_t and e_t traces as time proceeds.

7.4 Servo Following

With the exception of the generalized minimum variance controller the algorithms presented in this chapter are all intended as regulators. When it is necessary to control to a reference value it is often

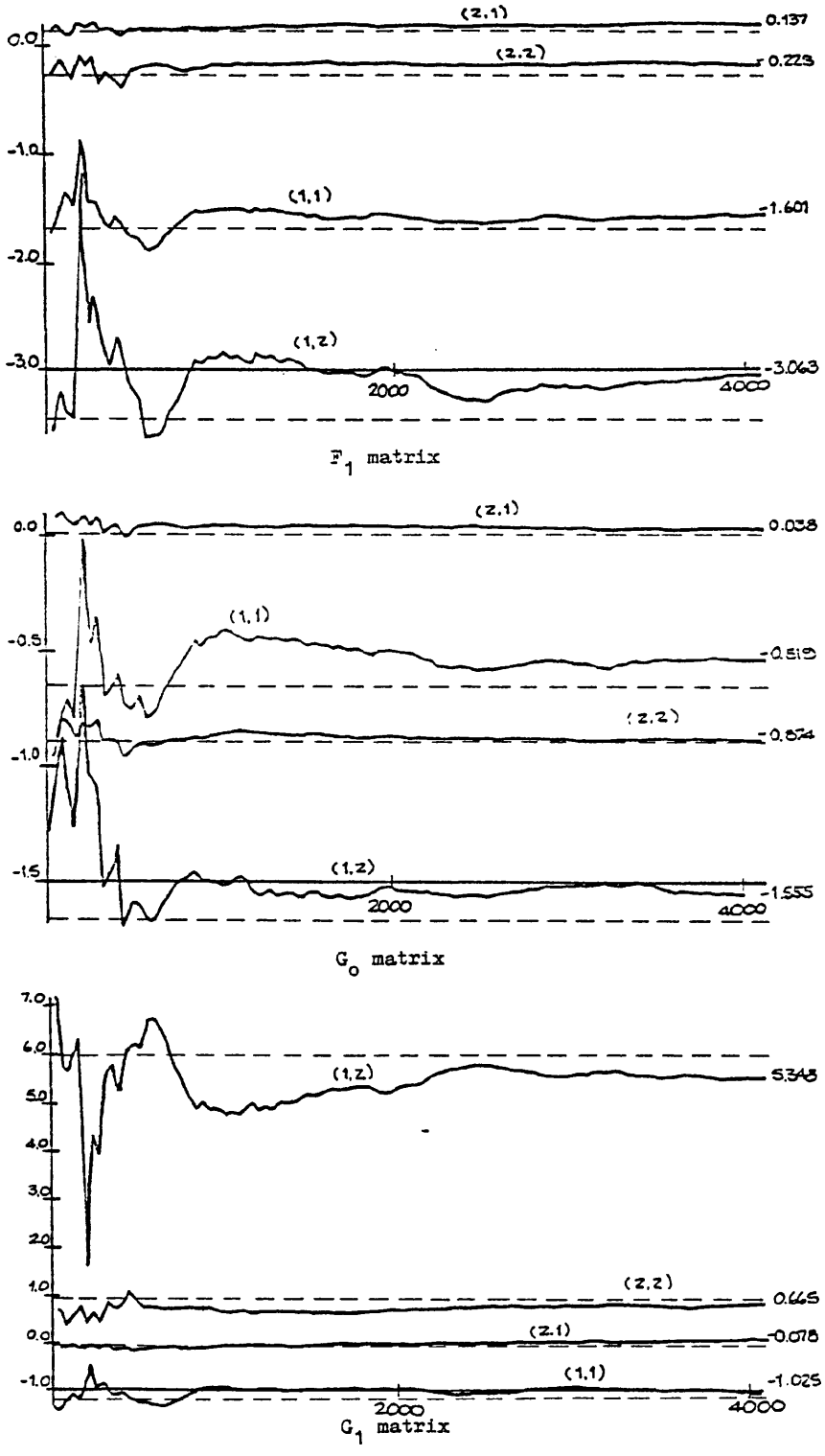
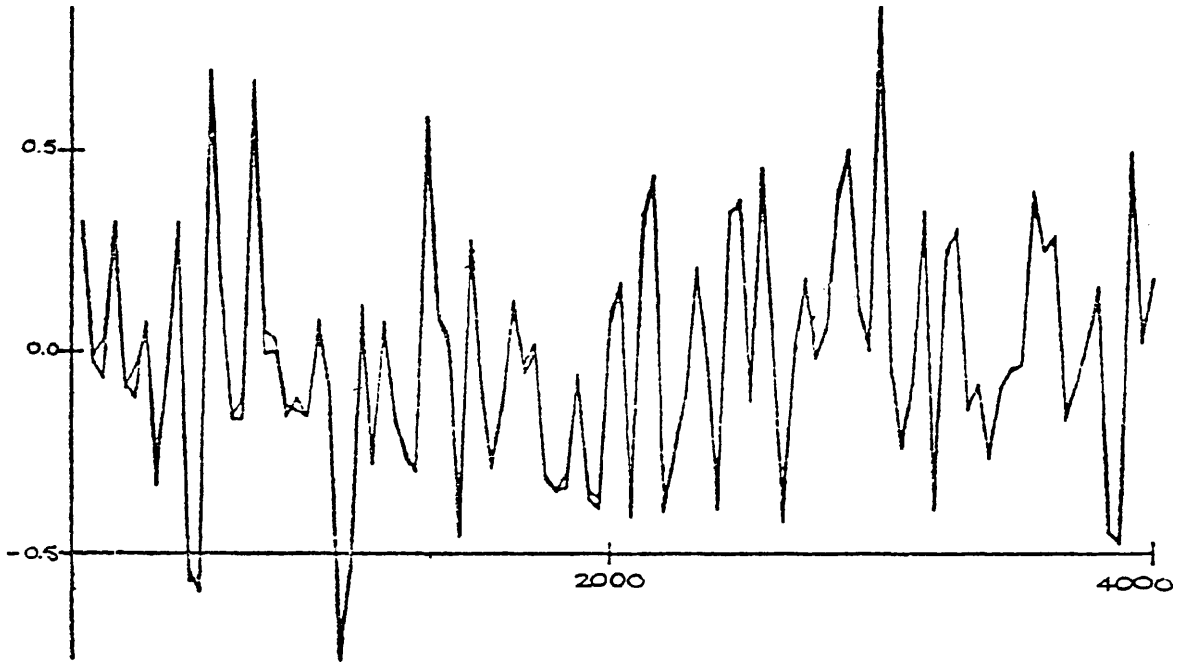
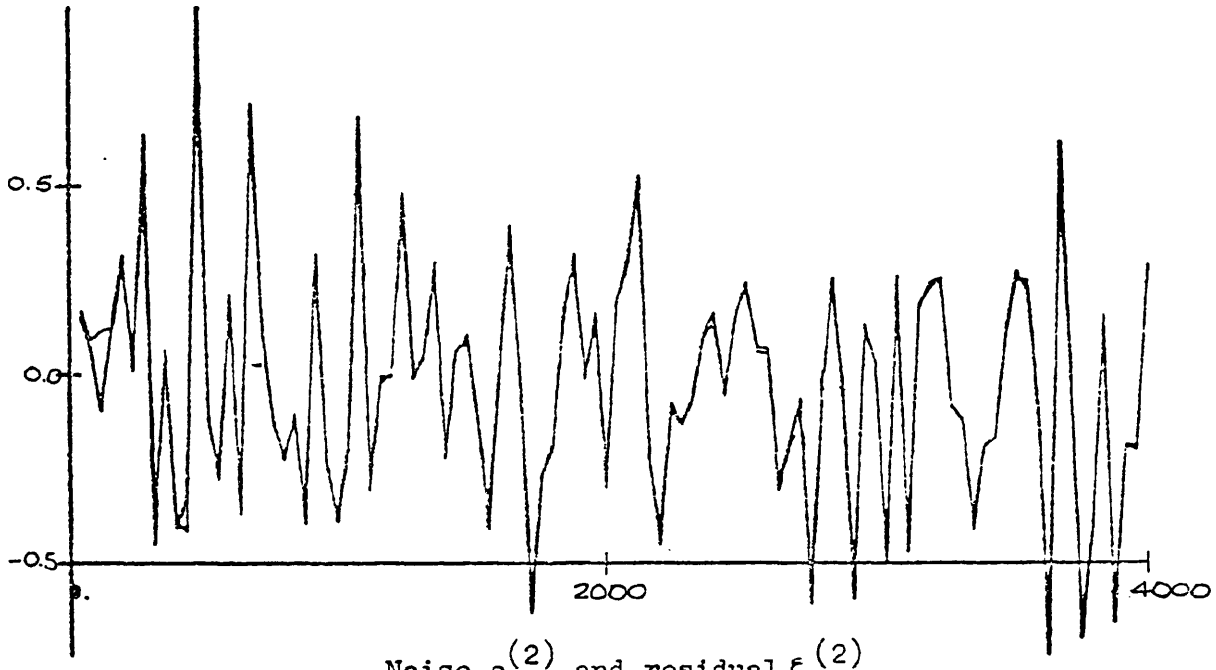


Fig. 7.6(a) Time evolution of regulator parameters



Noise $e_t^{(1)}$ and residual $\varepsilon_t^{(1)}$



Noise $e_t^{(2)}$ and residual $\varepsilon_t^{(2)}$

Fig. 7.6(b) Estimation residuals superimposed on system noise

perfectly satisfactory to introduce a digital integrator in the loop in precisely the same way as was suggested in section 6.7.

Some of the feedforward structures for servo following discussed in section 6.8 may also prove to be useful in a self tuning context. Their applicability is however the subject of on-going research. In particular, the parameters of the series and feedforward compensators S_p and S_f must be estimated, and this will add to the computational burden.

It has been found that when set-point changes are relatively 'smooth' (i.e. for example a ramp rather than a step) the addition of an integrator (if not already present in the system) in the loop will normally give satisfactory performance. The self tuning structure required is depicted in Figure 7.7. In particular, the incremental input Δu and error signal $\Delta y = y - y_r$ (where y_r is the set point) are fed to the estimator and controller. The input to the plant is formed by integrating the incremental input so that

$$u_t = u_{t-1} + \Delta u_t$$

The configuration is extremely simple. However, when large and rapid changes in set point are applied to the system the method is not entirely satisfactory and effects such as those depicted in Figure 6.3 may occur.

7.5 Conclusion

This chapter has discussed a number of self tuning strategies based on the control schemes presented in Chapter 6. Its contribution has been to introduce two new multivariable self tuning regulators,

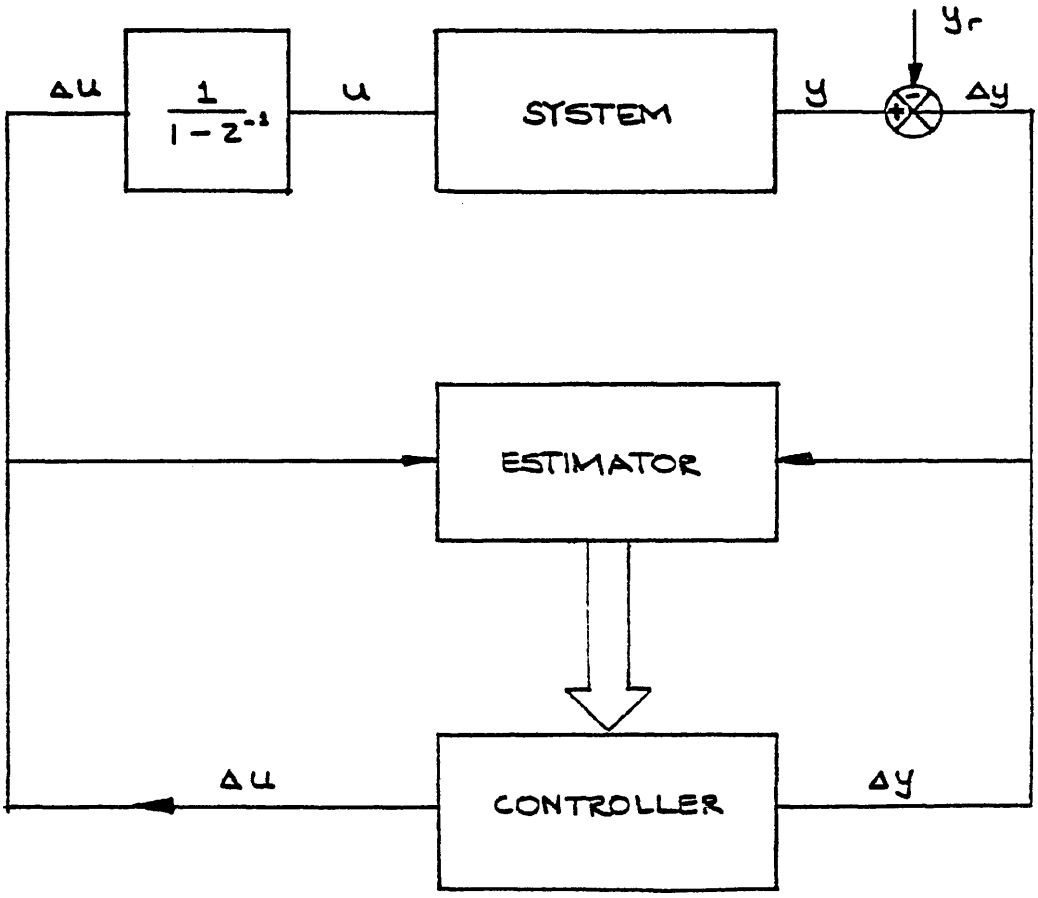


Fig. 7.7 Self-Tuning controller structure
for set point following

demonstrating in particular that an alternative approach to optimal self tuning regulation exists in the classical control concept of pole placement. The advantages of detuned minimum variance and pole assignment have been discussed in Chapter 6, but it is appropriate to emphasize the benefits particularly of the pole shifting regulator in a self tuning context. The robust nature of the algorithm in that it can handle non-minimum phase systems, multivariable systems in which loops have different pure time delays, systems with variable time delays and the fact that it is less sensitive to incorrect estimates of system time delay is especially valuable in a self tuning situation where it is expected that the precise details of system characteristics are not always available. While the generalized minimum variance self tuning controller (for SISO systems) does allow non-minimum phase systems to be controlled it requires a greater degree of knowledge about system characteristics than the pole-shifting algorithm in order to select the cost function weighting polynomials appropriately.Furthermore, these weighting polynomials may have to be changed if the system characteristics change and the algorithm is therefore not always robust enough to self-tune time varying systems. It is relevant too that optimal self-tuners (and also the detuned minimum variance self tuner) attempt to cancel system dynamics which is a much less robust approach than modifying them as in the pole-shifting regulator.

Robustness is the essential advantage of the pole assignment regulator, and it is this feature that has been highlighted by choosing in the examples situations where the pole-shifting regulator successfully copes with difficult problems, all of which are not uncommon in practical cases.

CHAPTER 8

IMPLEMENTATION OF SELF-TUNING CONTROLLERS

8.1 Introduction

The theoretical aspects of several self-tuning strategies have been discussed in Chapter 7. Although the application of the algorithm is fairly straightforward, no discussion of self-tuning control would be complete without also examining the practicalities of its engineering application. This chapter therefore addresses itself to two aspects of self-tuning implementation. The first part is devoted to a brief, general discussion of self-tuning in practice, and the second part reports on the self-tuning control of a real process using the pole-shifting method.

8.2 General Aspects of the Implementation of Self-Tuning Controllers.

This section covers a number of important practical problems in implementing self-tuning controllers. These include the initialization phase, the role of the 'forgetting factor', dealing with saturation effects and time-varying systems, incorporating set point following when a regulator law is used, and coping with measurement noise. The following notes are by no means exhaustive, but are designed to provide basic practical guidelines drawn from the author's experimental experience.

8.2.1 Getting the Algorithm Started.

Before self-tuning can be attempted, the engineer must have a reasonably good feel for certain system parameters. These include system order, system pure time delays, and approximate values of system time constants. It is generally advisable to select the lowest system order that will still account for the dominant system dynamics as this reduces the number of parameters to be estimated, minimizes computation time and aids convergence rate. The system should however never be grossly underparametrized.

It is often useful to estimate the parameters of the system from a short record of input-output data (or in real time) using a recursive least squares algorithm. Although the parameters will almost certainly be biased from their correct values, the estimates still provide a useful initial condition for setting up the self-tuning controller estimation model. Where this is not feasible, another useful starting point is to assume that the system outputs are all decoupled, and each system output transfer function is modelled as n_a cascade digital integrators.

For the minimum variance regulator, the initial estimation model would thus be

$$\{I+P(z^{-1})\}\text{diag}\{(1-z^{-1})^{n_a}\}\, y_t = z^{-k}\{I+P(z^{-1})\}u_t + \varepsilon_t \qquad 8.1$$

where $I+P(z^{-1})$ is chosen so that the coefficients of $z^{-1}, \ldots, z^{-k}$ are zero in the polynomial multiplier of y_t (ie $P(z^{-1})$ is selected as in eqn 7.79).

For the pole-shifting model the initial estimation model would be

$$\text{diag}\{(1-z^{-1})^{n_a}\}y_t = \{Iz^{-1}+\hat{B}_2 z^{-2}+ \ldots +\hat{B}_{n_b} z^{-n_b}\}\, u_t + \varepsilon_t \qquad 8.2$$

where $\hat{B}_i$, i=2,3, ... ,n_b are chosen with small element values to ensure that the transformation to the controller parameters is defined. The effect of an assumed system of cascaded integrators is to produce fairly vigorous initial control action which generates good estimation data.

The recursive least squares estimator itself requires initialization, in particular, the covariance matrix and the 'forgetting factor'. If the model parameters have been obtained as the result of estimation and are thus likely to be reasonably good the covariance matrix may be initialized to the identity matrix, whereas with poorer starting conditions a value of 100I would be more appropriate. To assist in initial tuning, the 'forgetting factor' (λ) may initially be set as low as 0.96 and then allowed to rise towards unity. A suggested rule is:

$$\lambda_{k+1} = 0.99\lambda_k + 0.01$$

$$\lambda_o = 0.96$$

The forgetting factor plays an important role in self tuning and this is discussed later.

Sampling rates are related to the fastest significant system modes. Typically a sampling interval of half the time constant of this mode is a good starting point. Sampling rates should not be

chosen too fast as this can lead to excessively active control.

8.2.2 Saturation Effects

Most processes will suffer from saturation limitations of one form or another. Unless these are taken into account it is almost certain that the estimator will yield degraded or invalid results with a concommitant degradation in control. The self-tuner may even lose control altogether. Precautions designed to protect the estimator should therefore be taken, as follows. The saturation limits should, where ever possible, be referred to the input of the system and upper and lower control limits set in the self-tuning controller software. The control signal is then subjected to these limits, and where the demanded input is modified so as to satisfy the limits, the modification must be applied to estimator and control law data vectors as well. This ensures that the estimator always operates on input data actually applied to the process. Where an integrator has been inserted for set point following both the integrator output (u_t) and the incremental control input (Δu_t) must be modified.

8.2.3 Time Varying Systems and the Choice of Forgetting Factor

Although for simplicity, the theory of self-tuning controllers is developed under the assumption of linear, time-invariant systems, the true value of self-tuning control is of course in its application to non-linear, time-varying plants. Self-tuning control works well in such systems provided that the plant dynamics vary

'slowly' relative to the self-tuning controller's rate of adaptation.

The adaptability of a self-tuning controller is determined largely by the value of the estimator 'forgetting factor'. For a linear time-invariant system the forgetting factor must be unity for the algorithm to converge. However, if the controller is to adapt quickly, its value must be reduced. The reduction of the forgetting factor effectively weights the most recent data by scaling down the information matrix of past data before adding the contribution of the latest 'information'. Although the principle of this approach is sound, extreme caution must be exercised when manipulating the forgetting factor. Unless new information is captured at each iteration (i.e. non-zero incremental input data and output errors) a non-unity forgetting factor will successively reduce the value of the information matrix thus increasing the magnitude of entries in the covariance matrix P (proportional to the inverse information matrix) used in the least squares estimator, or even, due to dominant numerical errors, allow it to become negative definite. Ultimately, all 'information' will be lost so that when a system parameter change does occur (leading to a large error in the estimation update equation) the estimator gain which depends on P will be unrealistically high, and the whole algorithm can become unstable.

Intuitively, the forgetting factor is set correctly when (given that the covariance matrix is in a satisfactory state) an 'information balance' exists between information being lost through forgetting and information being gained from new incoming data.

Let the estimator parameter vector $\hat{\theta}_i$ for output i be ordered as follows:

$$\hat{\theta}_i = (\underbrace{a_{i11},..,a_{iln_a},..,a_{ip1},..,a_{ipn_a}}_{p \text{ groups}},\underbrace{b_{i11},..,b_{iln_b},..,b_{ip1},..,b_{ipn_b}}_{p \text{ groups}})^T \quad 8.3$$

so that, for example, the estimation model is:

$$y_t^{(i)} = (y_{t-1}^{(1)},..,y_{t-n_a}^{(1)},..,y_{t-1}^{(p)},..,y_{t-n_a}^{(p)},u_{t-1}^{(1)},..,u_{t-n_b}^{(1)},..,u_{t-1}^{(p)},..,u_{t-n_b}^{(p)})\hat{\theta}_i + \varepsilon_t \quad 8.4$$

where u_t and y_t represent the system input and output p-vectors.

Then the diagonal of the covariance matrix is asymptotically proportional to the diagonal of the matrix $E(\hat{\theta}_i\hat{\theta}_i^T)$ and may be divided into 2p sections, as is $\hat{\theta}_i$.

One way of monitoring the rather qualitative concept of information balance is to periodically check the 2p sub-traces of the partitioned covariance matrix. During the tuning-in phase the forgetting factor may be chosen as suggested in section 8.2.1. Once the self-tuner has tuned in and is operating satisfactorily, the forgetting factor is chosen so that the sub-traces remain more or less constant. The forgetting factor should be increased if the traces increase (indicating a loss of information) and may be decreased otherwise. A typical value to which the forgetting factor might be set is 0.995. This method of choosing the forgetting factor is intended merely as a guide. It may also prove useful to automate the procedure.

8.2.4 Set Point Handling.

It has been suggested that the introduction of a digital integrator in the control loop will facilitate set-point handling by regulators. However, as demonstrated in section 6.8 (Fig. 6.3) transient response to rapid set point changes can be poor.

Most systems have inherent slew rate limitations and it is as well to bear this in mind when introducing set point changes. An extremely useful technique is to apply software rate limits to set-point signals, consistent with the response of the system. This will improve transient performance and minimize overshoot.

The introduction of closed loop poles, as is possible in the detuned minimum variance and pole-shifting regulators can also aid transient response, although care should be taken not to slow down the system response unnecessarily.

8.2.5 Measurement Noise.

Although the self-tuning regulators are designed to cope with system noise, the overall performance will naturally improve if measurement noise is minimized and accurate measurements of the system outputs are available. Measurement noise may be introduced due to a noisy transducer, or even quantization noise in the analogue to digital conversion. In order to minimize noise of this type it is wise to take a number of samples in rapid succession at each sampling instant, and to average the result.

When computation time is already short this may not be possible. The author has found that averaging over 10 samples can dramatically improve performance.

8.3 An Application of the Self-Tuning Multivariable Pole-Shifting Regulator.

The rest of this chapter is devoted to the application of the self-tuning multivariable pole-shifting regulator to a real system. The process is a hydraulic one comprising two water tanks arranged side by side and coupled hydraulically by an orifice in the partitioning wall of the tanks. Each tank has an outlet tap, a pump supplying water, and a water height sensor represented schematically in Fig. 8.1 and Fig. 8.2.

The control objective was to independently control the water level in tanks 1 and 2. The linearized dynamics about any one operating point are fairly straightforward to calculate, and the analysis is given in Appendix 1. However the dynamics vary according to operating point and the control problem is further complicated by the non-linear characteristic of the pump voltage-flow characteristic. The pumps also exhibited saturation and deadband effects.

Thus the system is effectively non-linear, has dynamics which vary with operating point, and as such is a suitable candidate for self-tuning control.

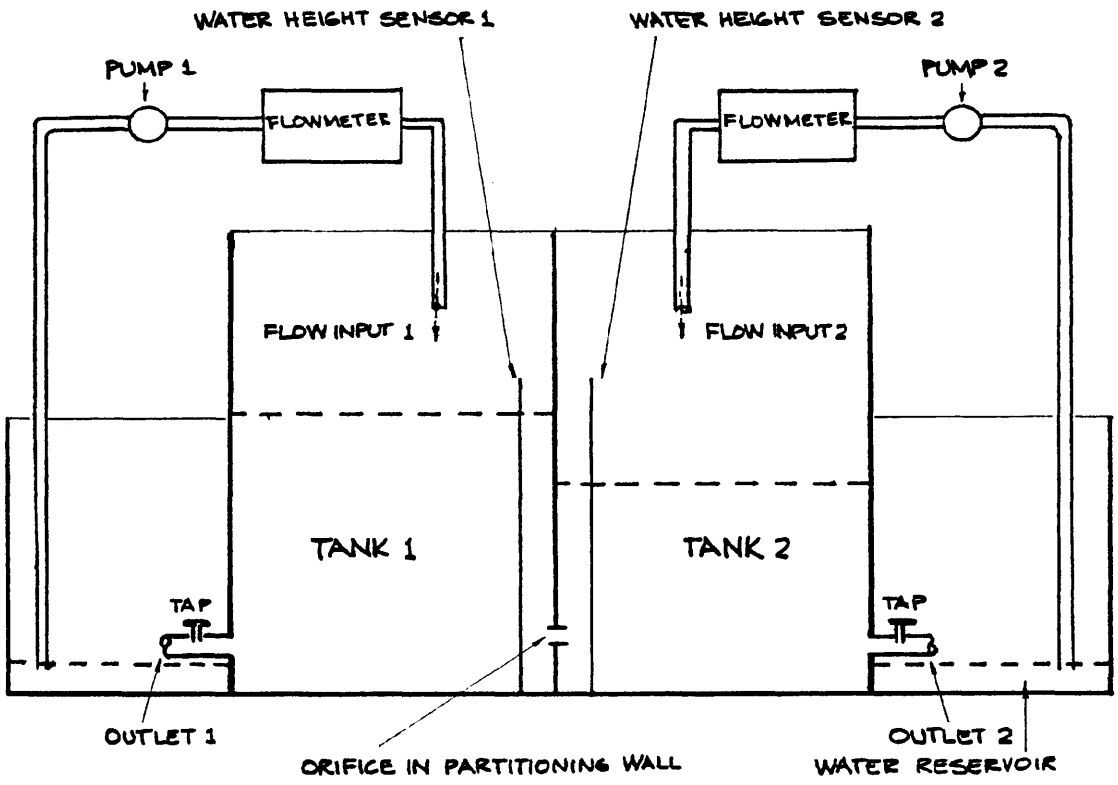


Fig. 8.1 Coupled Tank System

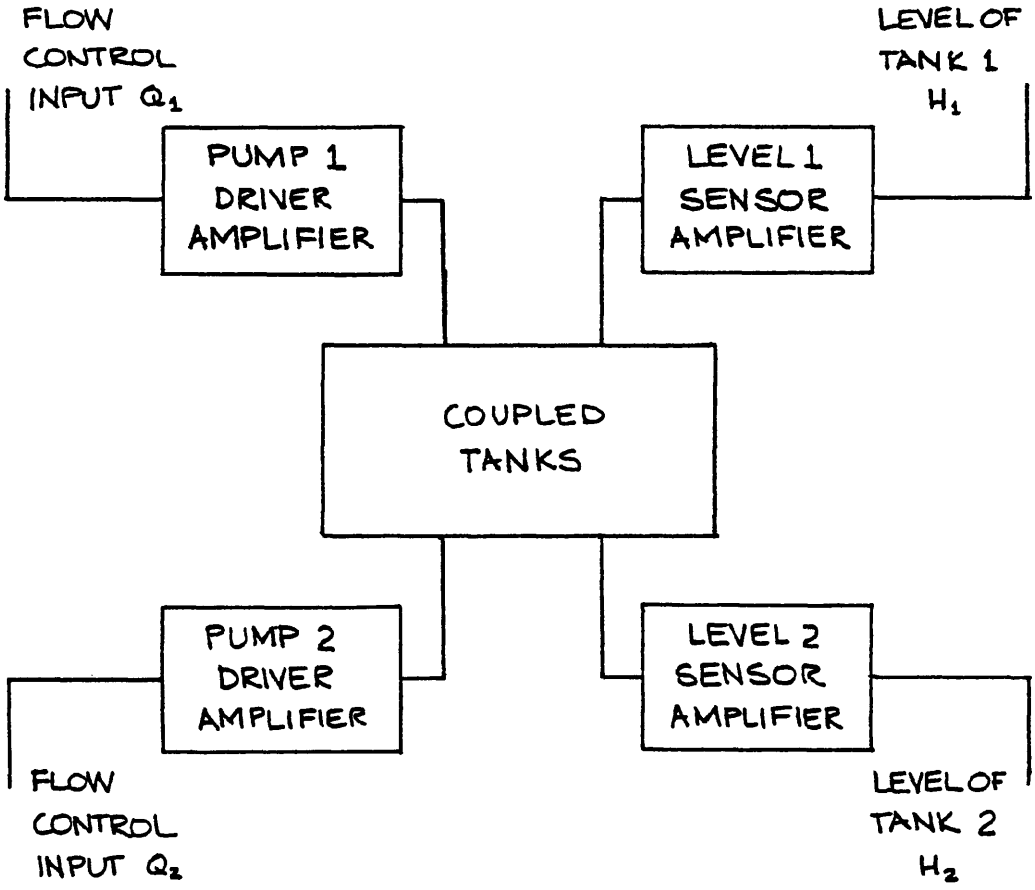


Fig. 8.2 'Black box' diagram of coupled tank system

8.3.1 Coupled Tank Dynamics.

It can be shown (Appendix 1) that the variation of the system time constants T_1 and T_2 can be approximated very crudely by:

$$T_1 \propto \sqrt{\Delta H} \qquad 8.5$$

$$T_2 \propto \sqrt{H - 1.2} \qquad 8.6$$

where ΔH is the difference in level between the tanks, and H is the approximate level of the tanks. The approximation

$$H \simeq H_1 \simeq H_2$$

is valid because the limited pump and drain capacity of the system dictates that the steady state level differences between the tanks cannot exceed approximately 2cm which is generally very much less than H_1 or H_2. Thus, fairly small changes in the difference between the tank levels can lead to quite substantial fluctuations in T_1, whereas T_2 will be more dependent on changes in actual tank levels. A table of sample values of T_1 and T_2 is quoted from Appendix 1:

	H_1(cm)	H_2(cm)	T_1(s)	T_2(s)
1.	8	7	9.65	119
2.	13	12	9.93	159
3.	22	21	10	215
4.	13	11	13.56	156
5.	13	12.5	3.55	162

Entries 1, 2, and 3 show how T_2 increases as the tank levels increase whilst a constant head difference of 1 cm is maintained between the two tanks. Modifying this head difference results in substantial changes in T_1 (entries 4 and 5). In fact, over the operating range used in the control experiments factor of 2 changes in T_2 and factor of 4 changes in T_1 were not uncommon.

8.3.2 The Pump Characteristic.

A typical voltage-flow characteristic for the type of pump used is given in Fig. 8.3. Note the deadband (region A) and saturation (region B) effects as well as the non-linear characteristic. To accommodate the saturation and deadband, software saturation limits on the pump voltages (ie system control input) were set at V_l and V_u volts for the lower and upper limits respectively. These limits varied, depending on the pump and power amplifier used to drive the pump, and the limits were not necessarily the same for each input. It is important to set the limits fairly accurately so that the real system inputs are used in the estimator.

8.3.3 The Tank Level Measurement.

The depth transducers had a non-linear characteristic, a typical example of which is given in Fig. 8.4. Although it would have been feasible to determine H, the tank level, from the voltage reading, the fact that the characteristic was not constant but tended to vary due to the sensor amplifier drift made it simpler to use the voltage output directly as the control output. This introduced a further non-

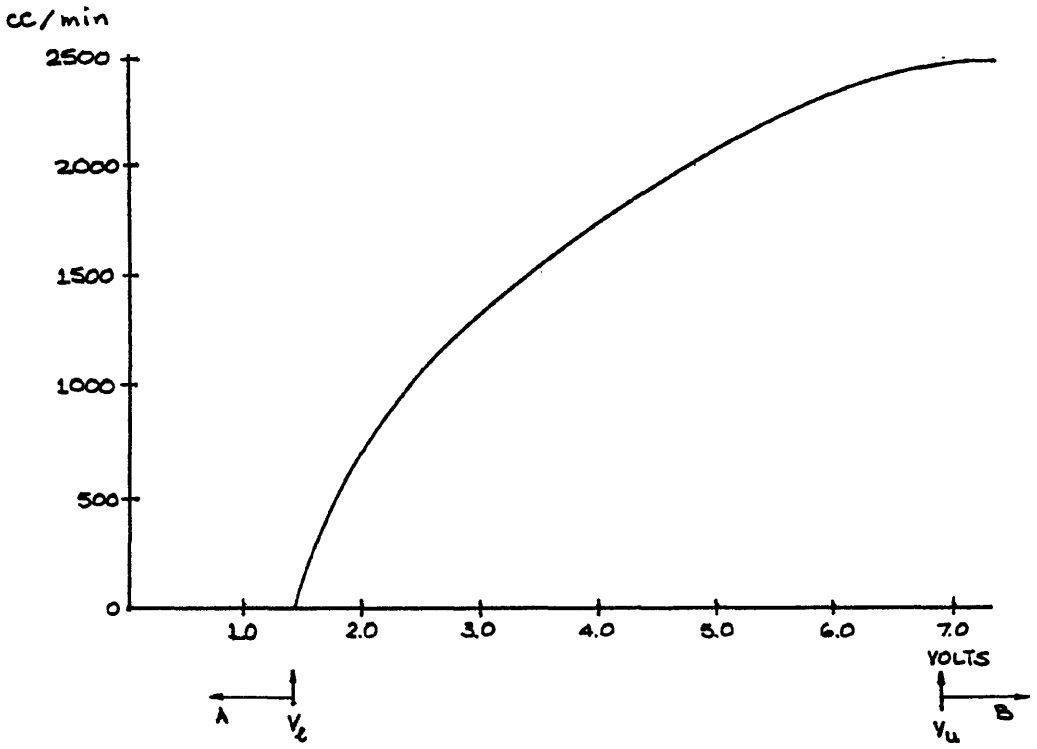


Fig. 8.3 Pump Characteristic

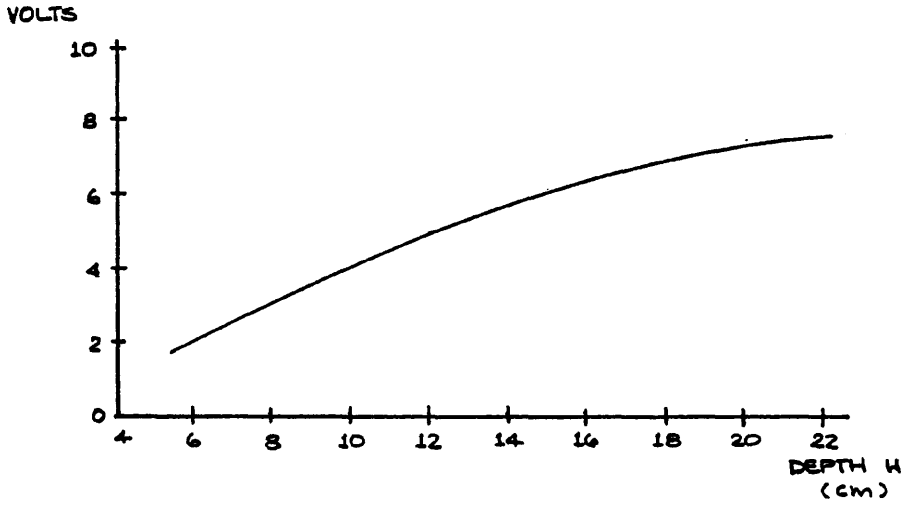


Fig. 8.4 Depth Transducer Characteristic

linearity into the system, to which however the self-tuner readily adjusted. The demand tank levels were therefore also specified in 'volts' rather than directly in height. To minimize transducer noise, 10 measurement samples were taken in rapid succession at each sampling instant and averaged.

8.3.4 The Self Tuner Configuration

The s-domain model (see Appendix 1) indicates that a z-transform model with $n_a=n_b=1$, $k=0$, in eqn. 6.1 would be appropriate. However, as the system output was required to follow a set-point, an integrator was inserted in series with each input. Set-point rate limiting was also introduced. Let

H_i = measured depth of tank i in volts

H_{d_i} = demanded (rate limited) depth of tank i in volts

$\Delta H_i = H_i - H_{d_i}$

Q_i = flow rate control signal for pump i in volts

ΔQ_i = incremental flow rate control signal in volts

A schematic of the system configuration is given in Fig. 8.5.

The estimation model used had the structure:

$$(I + \hat{A}_1 z^{-1} + \hat{A}_2 z^{-2}) \begin{pmatrix} \Delta H_1 \\ \Delta H_2 \end{pmatrix} = (\hat{B}_1 z^{-1}) \begin{pmatrix} \Delta Q_1 \\ \Delta Q_2 \end{pmatrix} + \varepsilon_t \qquad 8.7$$

where the symbols retain their usual meaning. Note that polynomial $I+\hat{A}(z^{-1})$ has order $n_{\hat{a}}=2$ instead of 1 to allow for the digital integration.

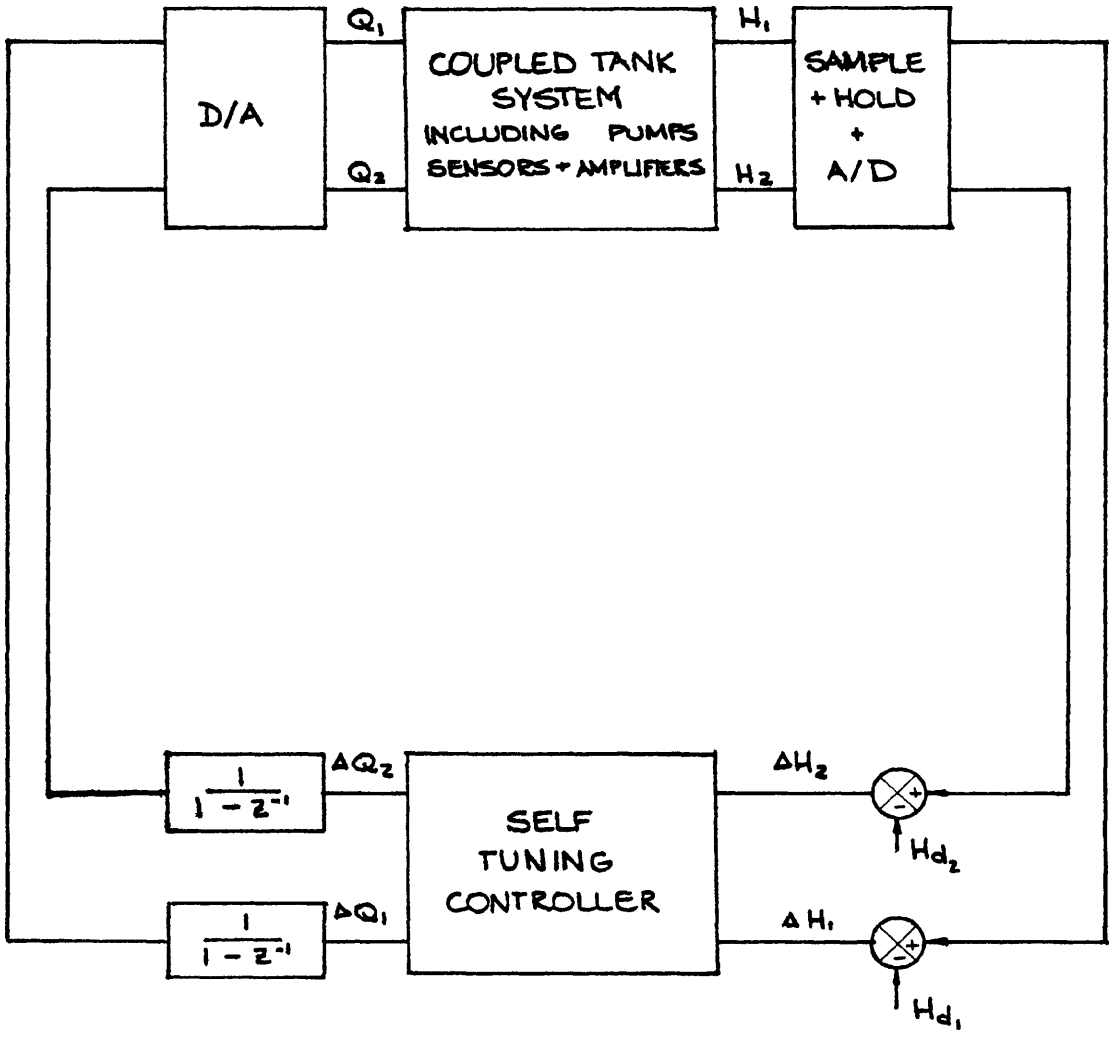


Fig. 8.5 Self-Tuning Controller Configuration

A sampling period of 5 seconds was chosen and although this might seem rather long for the smallest values of time constant T_1, it was shown experimentally to be satisfactory. The total computation time for one iteration on the PDP 11/10 computer used was 400ms and the computational delay thereby introduced was a small enough fraction of the sampling period not to have to extend the $\hat{B}_1(z^{-1})$ polynomial.

The self-tuner was initialized by the choice of model described in eqn. 8.2 and the controller was allowed to run for a while. The resulting parameter estimates were stored and used in subsequent runs as initial conditions.

The choice of $I+T(z^{-1})$ polynomial,

$$I + T_1 z^{-1} = I - 0.6Iz^{-1} \qquad 8.8$$

corresponding to two time constants of 9.8s was found to achieve good results.

8.3.5 Example Results

Example 1: This example demonstrates tuning-in behaviour. The level of tank 1 was required to follow a positive and negative going square wave of amplitude 1cm and period 500s centred about a depth of 13cm whilst tank 2 was to be maintained at a constant depth of 13cm. A constant forgetting factor of 0.99 was used with an initial covariance matrix of 10I (since, as discussed above, reasonable initial parameter estimates were available). No set point rate limiting was used. Software saturation limits were however imposed on the pump flowrate inputs. In particular, the lower limit for pump 1 was 0.4V.

Fig. 8.6 shows the system output responses and the control signal to pump 1. Notice how during the first period of the square wave the response is poor. However, it soon improves and it is particularly interesting to see how the interaction effects on output 2 due to set point changes on tank 1 die down as the controller tunes in. Note also that the control input to pump 1 becomes smoother as tuning proceeds. As tank 1 drains from the 14cm to the 12 cm level the input to pump 1 falls to the lower saturation limit of 0.4V, switching the pump off.

This example is also used to illustrate another possible use of self-tuning control. The self-tuning algorithm may be used to design a controller that gives satisfactory response and the controller parameters may then be fixed to produce a constant parameter discrete time controller. However, the same computational time delay that occurred during self-tuning must be retained. Fig. 8.7 illustrates the system output responses and control input 1 for 1 cycle of the square wave set point when the controller parameters are set to the final values obtained at the end of the self-tuning run. The response is an improvement on the self-tuning response, indicating that the set-point changes caused detuning of the self-tuner. This is to be expected considering the simplified approach to incorporating the set-point.

Example 2: This example demonstrates the system response to a sequence of desired tank levels. The set-point inputs were rate limited to 0.5 V/iteration (i.e. 0.1 V/sec) and the pump control voltage software saturation limits were set to:

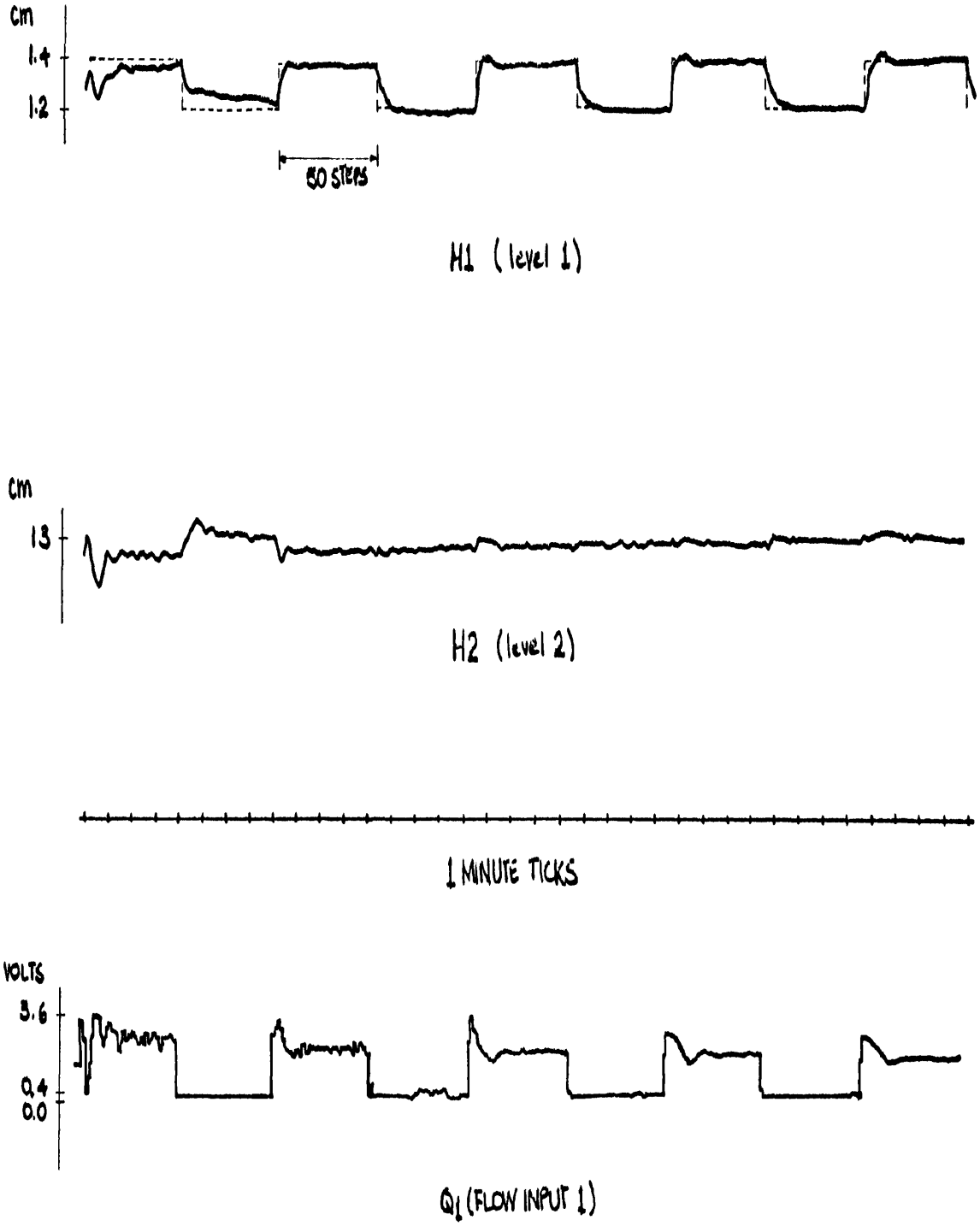


Fig. 8.6 System output responses and flow input 1 for a square wave demand on tank level 1

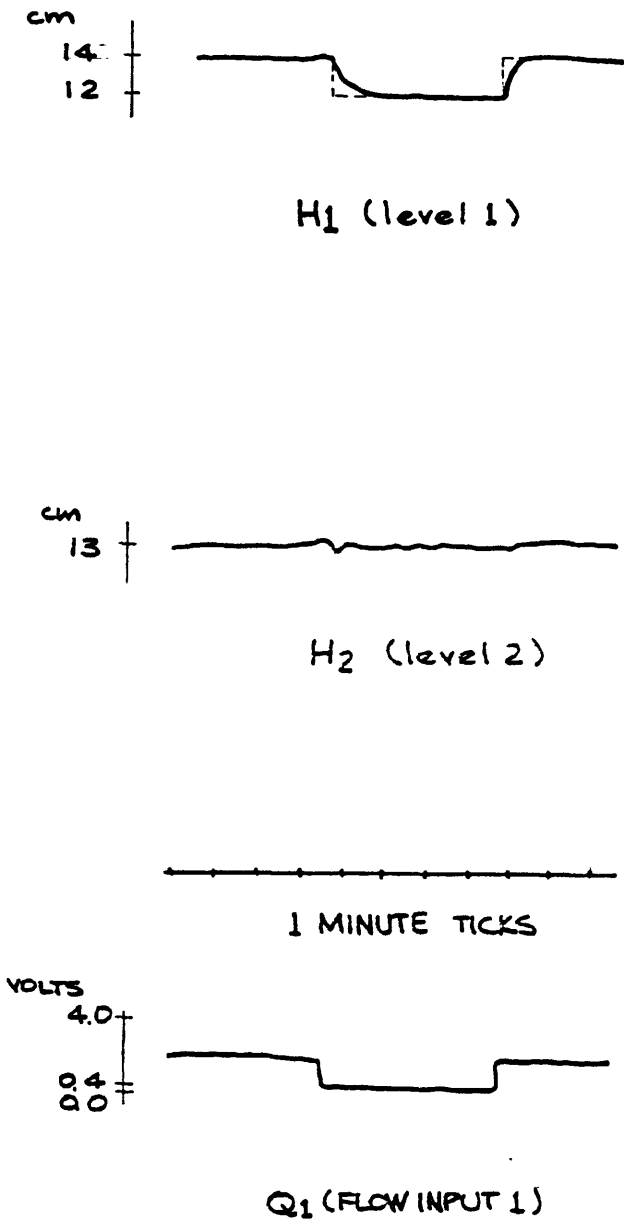


Fig. 8.7 System response when self-tuning control law in Fig. 8.6 is fixed

Pump No.	Lower Limit (V) (pump off)	Upper Limit (V) (pump full on)
1	1.2V	4V
2	4.6V	7V

The system was started with empty tanks. The estimator covariance was initialized to 10I and the forgetting factor λ_k adjusted as follows:

$$\lambda_{k+1} = \begin{cases} 0.99\lambda_k + 0.01 & \lambda_k < 0.99 \\ 0.99 & \text{otherwise} \end{cases}$$

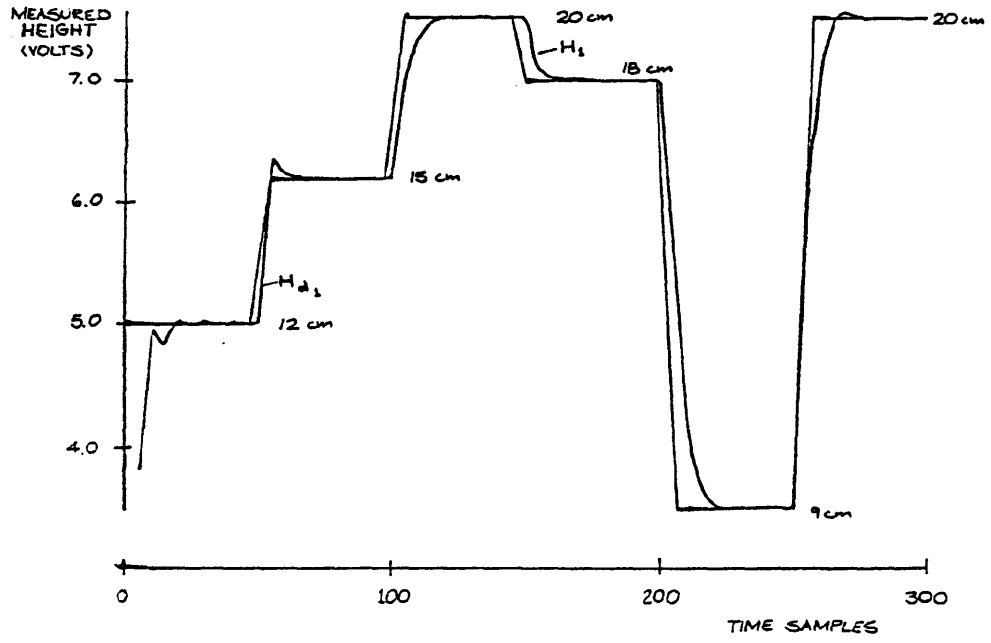
$$\lambda_o = 0.97$$

Fig. 8.8 shows the set point and the resulting system output. The pump control inputs are shown in Fig. 8.9. The controller tunes in rapidly. Although output 1 appears to rise very slowly after the set point change at step 100 it must be noted that in fact the pump is full on. Similarly at step 200 when tank 1 is required to drain, pump 1 is switched off and the slow response is due to the limited draining capacity.

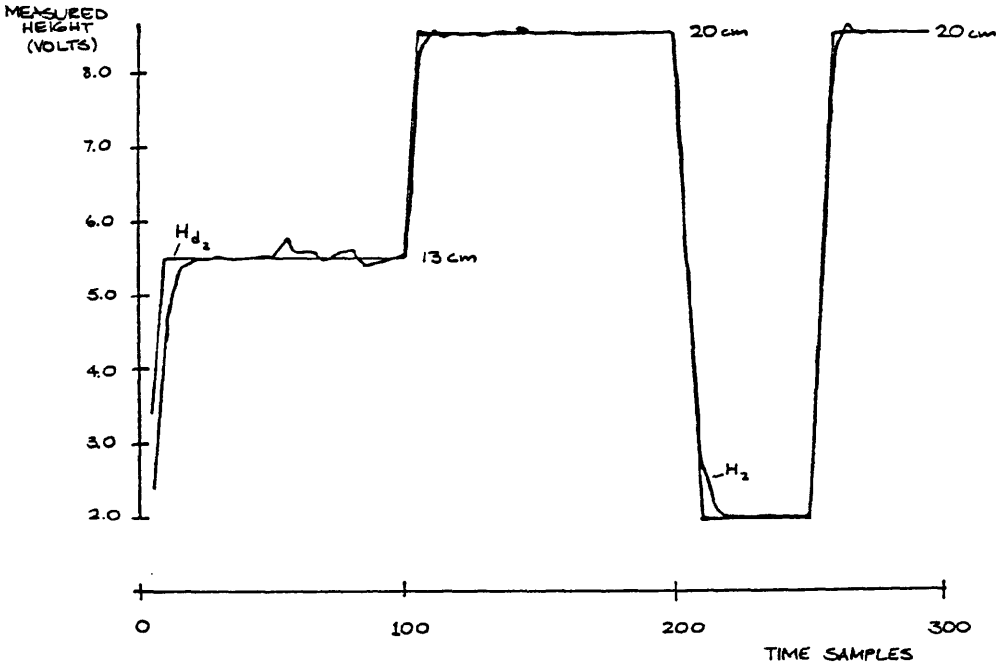
Fig. 8.10 shows the time evolution of the parameters of the control law:

$$\begin{bmatrix} \Delta Q_1 \\ \Delta Q_2 \end{bmatrix}_t = \begin{bmatrix} g_{110} & g_{120} \\ g_{210} & g_{220} \end{bmatrix} \begin{bmatrix} \Delta H_1 \\ \Delta H_2 \end{bmatrix}_t + \begin{bmatrix} g_{111} & g_{121} \\ g_{211} & g_{221} \end{bmatrix} \begin{bmatrix} \Delta H_1 \\ \Delta H_2 \end{bmatrix}_{t-1}$$

It is particularly impressive to observe that the control adapts at each set point change to the changing system dynamics.

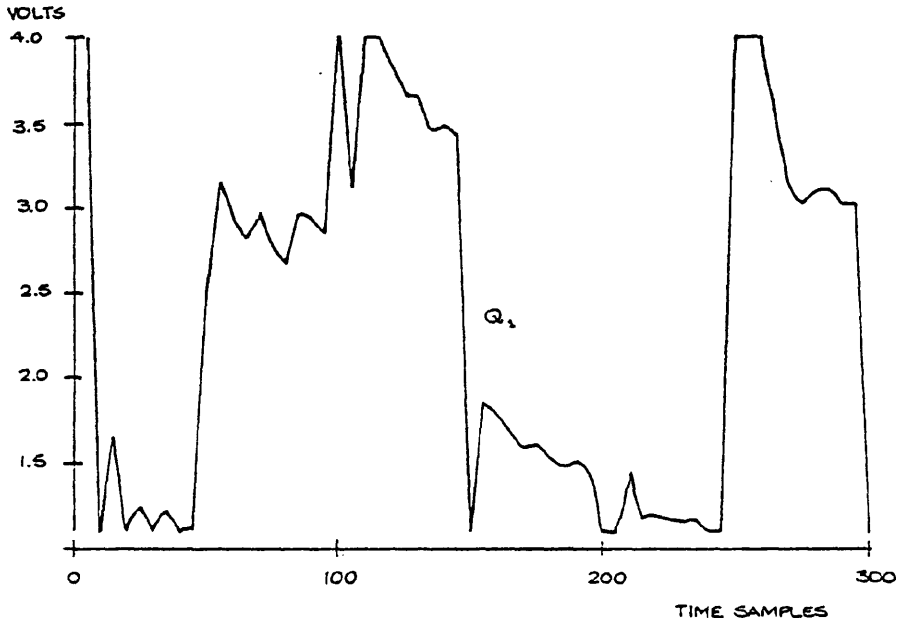


Level 1

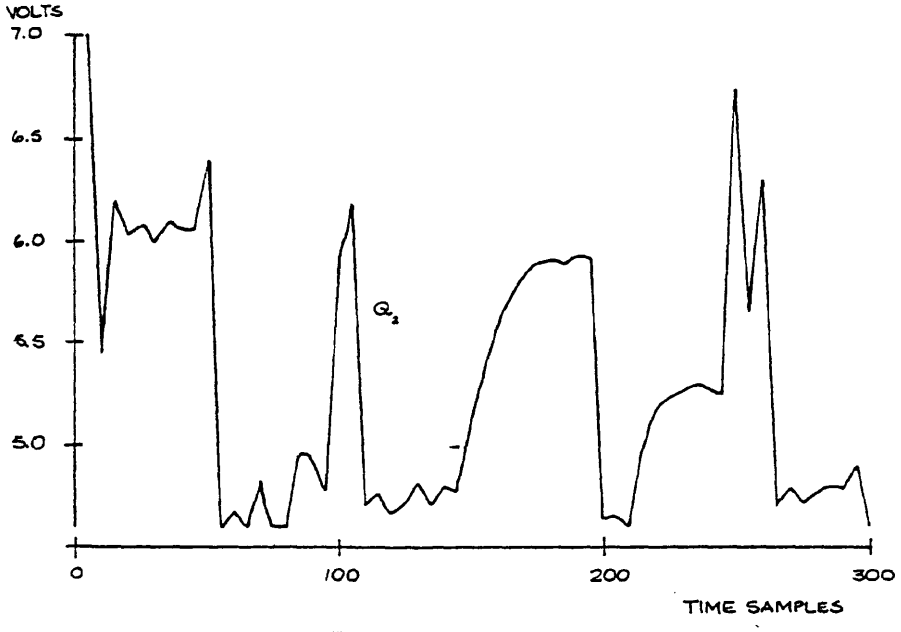


Level 2

Fig. 8.8 Tank level setpoints and system responses
N.B. Voltage-height characteristic is non-linear

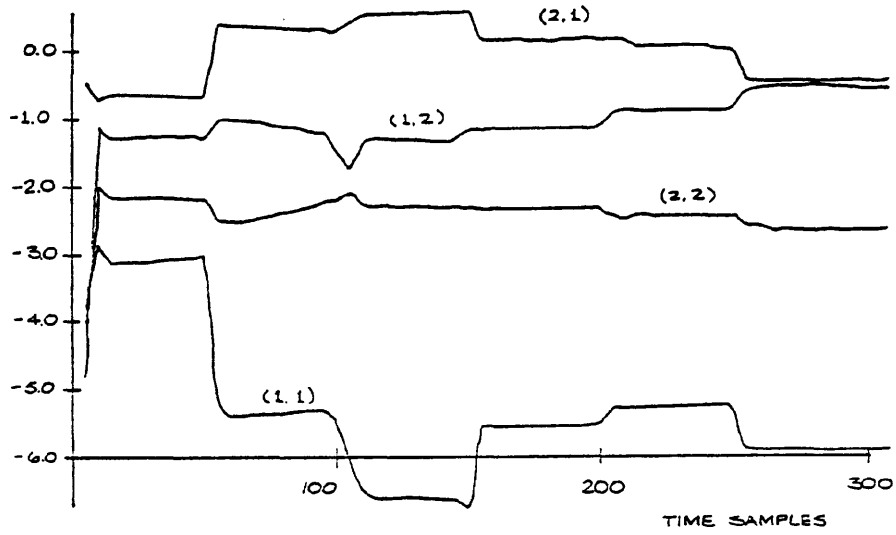


Input 1

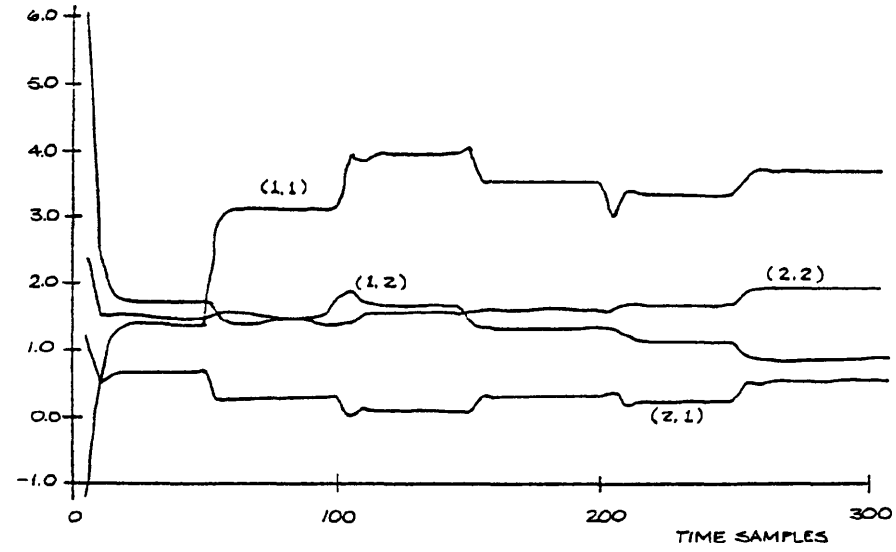


Input 2

Fig. 8.9 Control Inputs



G_0 matrix



G_1 matrix

Fig. 8.10 Time evolution of the controller parameters

8.4 Conclusion

This chapter has presented some of the more practical aspects of self-tuning, in particular demonstrating the successful control of a real process using the multivariable pole-shifting self-tuning regulator. The results given here are only examples from a series of successful runs and are encouraging evidence that self-tuning is a viable tool in the control of non-trivial processes.

CHAPTER 9

CONCLUSION

Each chapter contains its own concluding section, and the purpose of this chapter is to highlight some of the more important features of the results and propose areas for future work.

The research has provided new results and techniques in the related fields of system identification and self-tuning control. In the case of system identification, the contribution has been a fresh approach to parameter estimation, and maximum likelihood estimation in particular, in the form of algorithms based on hypothesis testing and strong analyst-computer interaction. Whilst the aim here has been the parametrization of ARMAX models from plant operating records for possible use in the off-line design of direct-digital-controllers, the work on self-tuning controllers goes further in that simultaneous identification and control are achieved. Two new multivariable self-tuning regulators, the detuned minimum variance and pole-shifting regulators have been described.

The first part of the thesis was concerned with system identification. Chapter 4 introduced the hypothesis testing approach for obtaining maximum likelihood parameter estimates in the case of SISO systems. The algorithm reduces the non-linear maximum likelihood estimation problem to a set of parallel linear least squares processes. Several variants of the algorithm were presented, the primary version being 'RHYP'. The alternative

'Correlation Method' and 'Covariance Method' approximations are designed to reduce computation. A number of simulation examples, including comparisons with well known approaches such as the 'Recursive Maximum Likelihood' (Soderstrom 1973) and 'Instrumental Variable' (Wong and Polak 1967) methods, testify to the advantages of the hypothesis testing scheme. These are its reliability (especially in the presence of short data records) and the interactive form of the algorithm which enables the analyst to visually assess the nature of the cost function. Thus the object of the estimation process becomes more than just seeking the minimum of the cost function, and diagnostics such as the time-evolution of the cost function and parameter sensitivity to changes in hypothesis are readily available. The algorithm is markedly more robust than approximate recursive methods when only short data records are available, and swift convergence of the noise colouration polynomial parameters is regularly observed. Its disadvantage is that the number of hypotheses grows exponentially with increasing order of this polynomial. Nevertheless as has been illustrated by simulation, even when for computational reasons the noise colouration polynomial is chosen to be under-parametrized, good estimates of the remaining polynomials can be obtained.

The extension of the hypothesis testing approach to cater for multivariable systems was discussed in Chapter 5. It was shown that both a determinantal (maximum likelihood) cost function and one based on the trace of the residual covariance matrix could be minimized. Both techniques led to consistent estimates of the system impulse

response models, though a minimal representation could not be obtained directly. The advantages and disadvantages of the multivariable algorithm, 'MVHYP', are of course similar to those of the single output version. However the algorithm becomes unwieldy both with increasing number of outputs and order of the noise colouration polynomial matrix. It has therefore been suggested that the algorithm might be embedded in a hill climber, where the selection of hypotheses is automated and the estimation problem is linear under a given hypothesis. Whilst some of the benefits of the approach (particularly with regard to information about the nature of the cost function) are lost, the method affords a substantial simplification of the direct approach in which all the model parameters would form part of the non-linear estimation search.

The hypothesis testing method is intended primarily as an off-line analysis tool. However, it is feasible that an automated hypothesis-selection algorithm (hill-climber) could be developed for on-line estimation in very slow systems. The algorithm's exceptionally good convergence properties could then be used to maximum effect with an on-line control synthesis algorithm. This approach to on-line identification and control is pursued later.

The second part of the thesis contributed two new multivariable self-tuning regulators, the detuned minimum variance and pole-shifting regulators. The off-line design principles were covered in Chapter 6, whilst the self-tuning properties of the algorithms were discussed in Chapter 7. Both regulators permit the specification of closed loop system poles, although the restrictions imposed differ in each case. In the case of the detuned minimum variance regulator this feature

may be used to obtain smaller and slower control excursions than for the conventional minimum variance regulator. However, whilst this regulator still derives from the optimality criteria, the pole-shifting regulator is more closely related to classical control objectives. The main advantage of the pole-shifting algorithm is that of robustness. It can regulate non-minimum phase systems without the need to specially select self-tuner parameters (as required by the generalized minimum variance method), multivariable systems in which the input-output relationships have differing pure time delays, and is less sensitive to incorrect modelling of the system pure time delay than other approaches. These features have been demonstrated in simulation examples presented in Chapter 7. The main disadvantage is that of computational complexity.

The proof of both the detuned minimum variance and pole-shifting regulator self-tuner properties was developed from a self-tuning lemma in Chapter 7. The minimum variance regulator naturally emerges as a special case of the detuned minimum variance regulator. A fundamental assumption of the proof is however that the estimated parameters converge. Now although simulation studies have indicated that this requirement is generally met, no convergence analysis for the new regulators has been given. This clearly is an area for future research. A further area which deserves urgent attention is the conversion of these regulation laws to full set-point handling controllers. Until such time the method of inserting a digital integrator in the loop to eliminate steady state error is a viable practical alternative.

Chapter 8 discussed some of the practical problems associated

with self-tuning, and discussed the implementation of a multivariable self-tuning pole-shifting controller on a hydraulic system. The good results achieved are evidence that the pole-shifting controller holds practical promise, especially since the system controlled featured non-linearities and widely varying dynamic characteristics.

In considering the future of self-tuning, it is however appropriate to reflect on the comparative merits of this approach, and that of a separate maximum likelihood algorithm operating on-line together with a control synthesis algorithm. Self-tuning is an elegant,computationally efficient method, based as it is on a simple linear least squares estimation algorithm. However it is in some ways a restrictive approach as, to date, the family of methods is fairly small, and in the absence of a generalized theory, the extension of this family is by no means trivial. In particular, it is necessary to prove a 'Self-tuning Property' for each new controller design method. Certainly in the case of SISO systems where an on-line version of the robust 'RHYP' algorithm could conceivably be used, a broader self-tuning approach of maximum likelihood estimation coupled with on-line controller synthesis could be valuable. The controller order, although selected so as to allow system identifiability, would than not need to be fixed as restrictively as in true self-tuning, and problems of controller convergence and set-point following greatly simplified. The system would be less computationally efficient, but using modern microprocessors, certainly viable for plants requiring sampling rates of the order of seconds. The use of an extremely reliable estimator like 'RHYP' would then

result in the possibility of a large, easily expandable family of high integrity controllers. In the case of multivariable systems, maximum likelihood estimation is further complicated, and it is difficult to justify this alternative to self-tuning.

It is the author's opinion that the future applications of estimation-based techniques in adaptive control systems will escalate enormously. It is therefore extremely important that the engineering aspects of self-tuning controller implementation are fully researched so that this valuable approach can make its full contribution to progress in this challenging field.

REFERENCES

1. Akaike, H., 1969 : 'Fitting autoregressive models for prediction', Ann. Inst. Statist. Math., 21, 243-247.

2. Akaike, H., 1972 : 'Information Theory and an extension of the Maximum Likelihood principle', 2nd International symp. on Information Theory, Ed. B.N. Petrov and F. Csaki, Akademiai Kiado, Budapest, 267.

3. Akaike, H., 1973 : 'Maximum likelihood identification of Gaussian autoregressive moving average models', Biometrika, 60, No. 2, 255-265.

4. Akaike, H., 1974a : 'A new look at the statistical model identification', IEEE Trans. Aut. Control, AC-19, No. 6, 716.

5. Akaike, H., 1974b : 'Information and statistical model building', Institut de la Vie World Conference 'Toward a Plan of Action for Mankind; Needs and Resources - Methods of Forecasting', Paris.

6. Alster, J. and Belanger, P.R., 1974 : 'A technique for dual adaptive control', Automatica, 10, 627.

7. Astrom, K.J. 1967 : 'On the achievable accuracy in identification problems', Proc. 1st IFAC Symp. on Identification, Prague.

8. Astrom, K.J. 1970 : 'Introduction to stochastic control theory', Academic Press, New York.

9. Astrom, K.J. and Bohlin, T., 1965 : 'Numerical identification of linear dynamic systems from normal operating records', IFAC Symp. Theory of Self-Adaptive Systems, Teddington, England.

10. Astrom, K.J., Borisson, U., Ljung, L., and Wittenmark, B., 1977: 'Theory and applications of self-tuning regulators', Automatica, 13, 457-476.

11. Astrom, K.J. and Soderstrom, T., 1974 : 'Uniqueness of the Maximum Likelihood Estimates of the parameters of an ARMA model', IEEE Trans. Aut. Control AC-19, No. 6, 769-773.

12. Astrom, K.J. and Wittenmark, B. 1971 : 'Problems of identification and control', J. of Math. Analysis and Applications, 34, 90-113.

13. Astrom K.J. and Wittenmark, B., 1973 : 'On self tuning regulators', Automatica, 9, 185-199.

14. Astrom K.J. and Wittenmark, B., 1974 : 'Analysis of a self tuning regulator for non minimum phase systems', IFAC Symp. on Stochastic Control, Budapest.

15. Bar Shalom, Y. and Tse, E., 1974 : 'Dual Effect, Certainty Equivalence and Separation in stochastic control', IEEE Trans. on Aut. Control, AC-19, 494.

16. Beck, M.B., 1977 : 'The identification and adaptive prediction of urban sewer flows', Int. J. Control, 25 No. 3, 425-440.

17. Bellman, R., 1961 : 'Adaptive Processes - A Guided Tour', Princeton University Press.

18. Bertram, J.E., 1956 : 'Factors in the design of digital controllers' Trans. AIEE, 75 Pt. 2, 151-159.

19. Borisson, U., 1975 : 'Self Tuning Regulators - Industrial applications and multivariable theory'. Report 7513, Division of Aut. Control, Lund Institute of Technology, Sweden.

20. Borisson, U. and Syding, R., 1974 : 'Self tuning control of an ore crusher'. Preprint IFAC Symp. on Stochastic Control, Budapest.

21. Borisson, U. and Wittenmark, B., 1974 : 'An industrial application of a self tuning regulator'. Preprint IFAC Symp. on Digital Computer Applications to Process Control, Zurich.

22. Bray, J., 1975 : 'Optimal control of a noisy economy with UK as an example', J. Roy. Stat. Soc., Series (A) General, 138, No. 3, 339-373.

23. Bryson, A.E. and Ho, Y.C., 1969 : 'Applied Optimal Control', Ginn and Co., Massachusetts.

24. Caines, P.E. and Chan, C.W., 1975 : 'Feedback between stationary stochastic processes', IEEE Trans. on Aut. Control, AC-20, No. 4, 498.

25. Caines, P.E. and Ljung, L., 1976 : 'Asymptotic normality and accuracy of prediction error estimators', Research Report 7602, Systems Control Group, Dept. of Elec. Eng., Univ. of Toronto.

26. Caines, P.E. and Rissanen, J., 1974 : ' Maximum likelihood estimation of parameters in multivariable Gaussian stochastic processes', IEEE Trans. on Inf. Theory, IT-20, 102.

27. Cegrell, T. and Hedqvist, T., 1973 : 'Successful adaptive control of paper machines'. Preprint, IFAC Symp. on Identification and System Parameter Estimation, The Hague.

28. Chan, C.W., Harris, C.J. and Wellstead, P.E., 1974 : 'An order testing criterion for mixed autoregressive moving average processes', Int. J. Control, 20, 817.

29. Chow, J.C., 1972 : 'On estimating the orders of an auto-regressive moving average process with uncertain observations', IEEE Trans. on Aut. Control, AC-17, No.5, 707-709.

30. Clarke, D.W., 1967 : 'Generalized least squares estimation of the parameters of a dynamical model', IFAC Symp. on Identification, Prague.

31. Clarke, D.W., Cope, S.N. and Gawthrop, P.J., 1975 : 'Feasibility study of the application of microprocessors to self tuning controllers', Report No. 1137/75, Engineering Lab., Oxford Univ.

32. Clarke, D.W. and Gawthrop, P.J., 1975 : 'Self Tuning Controller', Proc. IEE Control and Science, 122, No 9, 929.

33. D'Azzo, J.J. and Houpis, C.H., 1966 : 'Feedback control system analysis and synthesis', McGraw Hill, Tokyo.

34. Denham, M.J., 1974 : 'Canonical forms for the identification of multivariable linear systems', IEEE Trans. on Aut. Control, AC-19, No. 6, 646-655.

35. Deshpande, J.G., Upadhyay, T.N. and Lainiotis, D.G. , 1973 : 'Adaptive control of linear stochastic systems', Automatica, 9, 107.

36. Dickinson, B.W., Kailath, T. and Morf, M., 1974 : 'Canonical Matrix Fraction and State Space descriptions for deterministic and stochastic linear systems', IEEE Trans. on Aut. Control, AC-19, No. 6, 656-667.

37. Dunsmuir, W. and Hannan, E.J., 1976 : 'Vector linear time series models', Adv. Appl. Prob., 8, 339-364.

38. Durbin, J., 1959 : 'Efficient estimation of parameters in moving average models', Biometrika, 46, 306-316.

39. Durbin, J., 1961 : 'The fitting of time series models', Revue Inst. de Stat., 28, No. 3, 233-243.

40. Edmunds, J., 1976 : Ph.D. Thesis, UMIST, Manchester.

41. Edwards, A.W.F., 1972 : 'Likelihood', Cambridge University Press.

42. Eykhoff, P., 1974 : 'System Identification', Wiley Interscience, New York.

43. Farison, J.B., Graham, R.E. and Shelton, R.C., 1967 : 'Identification and control of linear discrete systems', IEEE Trans. on Aut. Control, AC-12, 438.

44. Feldbaum, A.A., 1960/61 : 'Dual Control Theory 1-4', Automn. Remote Control, 21, p.874, p.1033; 22, p.1, p.109

45. Florentine, J.J., 1962 : 'Optimal probing adaptive control of a simple Bayesian system', J. Electron. Control, 11, 165.

46. Gauss, K.F., 1809 : 'Theoria Motus Corporum Coelestium', in 'A Treasury of World Science', Ed. D.D. Runes Philosophical Library, New York, 1962.

47. Gawthrop, P.J., 1977: 'Some interpretations of the self tuning controller', Proc. IEE, 124, 10, 889.

48. Gawthrop, P.J., 1978: 'On the stability and convergence of self tuning algorithms', OUEL Report 1259/78, Oxford University.

49. Gertler, J. and Banfasz, C., 1974 : 'A recursive (on line) maximum likelihood identification method', IEEE Trans. on Aut. Control, AC-19, 816-820.

50. Glover, K., and Willems, J.C., 1974 : 'Parameterizations of linear dynamical systems: canonical forms and identifiability', IEEE Trans. on Aut. Control, AC-19, No. 640-645.

51. Goodwin, C.G., Zarrop, M.B. and Payne, R.L., 1974 : 'Coupled design of test signals, sampling intervals and filters for system identification', IEEE Proc. on Aut. Control, AC-19, No.6, 748.

52. Gustavsson, I., 1969: 'Parametric identification on multiple input, single output linear dynamic systems', Report 6907, Division of Aut. Control, Lund Institute of Technology, Sweden.

53. Hang, C.C. and Parks, P.C., 1973 : 'Comparative studies of model reference adaptive control systems', IEEE Trans. on Aut. Control, AC-18, No. 5 419.

54. Hannan, E.J., 1975 : 'The estimation of ARMA models', The Annals of Statistics, 3, No. 4, 975-981.

55. Hannan, E.J., 1976 : 'The identification and parametrization of ARMAX and state space forms', Econometrica, 44, No. 4, 713-723.

56. Hastings-James, R. and Sage, M.W., 1969 : 'Recursive Generalized least squares procedure for on-line identification of process parameters', Proc. IEE, 116, 2057-2062.

57. Hughes, D.J. and Jacobs, O.L.R., 1974 : 'Turn off, escape and probing in non linear stochastic control', Preprint IFAC Symp. on Stochastic Control, Budapest.

58. Isermann, R., 1973 : 'Identification of very noisy dynamic processes using models with few parameters', IFAC Symp. on Identification, The Hague.

59. Isermann, R., Baur, U., Bamberger, W., Kneppo, P. and Sieberg,H.: 'Comparison of six on-line identification and parameter estimation methods', Automatica, 10, 81-103.

60. Jacobs, O.L.R, and Langdon, S.M., 1970 : 'An optimal extremum control system', Automatica 6, 297.

61. Jensen, L. and Hansel, R. , 1974 : 'Computer control of an enthalpy exchanger', TFRT-3081, Dept. of Aut. Control, Lund Institute of Technology, Sweden.

62. Jury, E.I., 1964 : 'Theory and Application of the z-transform method', Wiley, New York.

63. Kailath, T., 1974 : 'A view of three decades of linear filtering theory', Trans. IEEE, IT-20, 145.

64. Kallstrom, C.G., Astrom, K.J., Thorell, N.E., Eriksson, J. and Sten, L., 1978 : 'Adaptive autopilots for steering of large tankers', IFAC Congress, Helsinki.

65. Kalman, R.E., 1958 : 'Design of a self optimizing control system', Trans. of the ASME, p.468.

66. Kalman, R.E., 1960 : 'A new approach to linear filtering and prediction problems', ASME Journal of Basic Engineering, 82D, 34-45.

67. Kashyap, R.L., 1970 : 'Maximum likelihood identification of stochastic linear systems', IEEE Trans. on Aut. Control, AC-15, No. 1, 25-34.

68. Kashyap, R.L., 1974 : 'Estimation of parameters in a partially whitened representation of a stochastic process', IEEE Proc. on Aut. Control, AC-19, No. 1, 13.

69. Kendal, M.G. and Stuart, A., 1961 : 'Advanced Theory of Statistics' Griffin, London.

70. Keviczky, L., Hetthessy, J., Hilger, M., and Kolostori, J., 1978: 'Self tuning adaptive control of a cement raw material blending', Automatica, 14, 525-532.

71. Ku, R. and Athans, M., 1973 : 'On the adaptive control of linear systems using the open-loop-feedback-optimal approach', IEEE Trans. on Aut. Control, AC-18, 489.

72. Kullback, S., 1959 : 'Information theory and statistics', Wiley New York.

73. Kuo, B.C., 1970 : 'Discrete data control systems', Prentice Hall, Englewood Cliffs, N.J.

74. Landau, I.D., 1974 : 'A survey of model reference adaptive techniques - theory and applications', Automatica, 10, 353-379.

75. Legendre, A.M., 1806: 'Nouvelles methodes pour la determination des orbits des cometes', Paris.

76. Levinson, N., 1947 : 'The Wiener RMS (root mean square) error criterion in filter design and prediction' J. Math. Phys., 25, 261-278.

77. Ljung, L., 1977 : 'Analysis of recursive stochastic algorithms' IEEE Trans. on Aut. Control, AC-22, No. 4 551.

78. Ljung L., Morf, M. and Falconer D., 1978 : 'Fast calculation of gain matrices for recursive estimation schemes', Int. J. Control, 27, No. 1, 1-19.

79. Ljung, L., Soderstrom, T. and Gustavsson, I., 1975 : 'Counter examples to general convergence of a commonly used recursive identification method', IEEE Trans. on Aut. Control, AC-20 No. 5, 643-652.

80. Ljung L., and Wittenmark, B., 1974 : 'Asymptotic properties of self tuning regulators', Report 7407, Dept. of Aut. Control, Lund Institute of Technology, Sweden.

81. Ljung, L. and Wittenmark, B., 1976 : 'On the stabilizing property of adaptive regulators', IFAC Symp. on Identification and System Parameter Estimation, Tbilisi.

82. Luxat, J.C. and Lees, L.H., 1973 : 'Suboptimal control of a class of nonlinear systems', Int. J. Control, 17, 965.

83. Mayne, D.Q. and Firoozan, F., 1977 : 'Linear estimation of ARMA systems, Parts 1 and 2', Publications 77/13 and 77/14, Dept. of Computing and Control, Imperial College of Science and Technology, London.

84. Mehra, R.K., 1974 : 'Optimal input signals for parameter estimation in dynamical systems - survey and new results', Proc IEEE on Aut. Control AC-19, No. 6, 753.

85. Morf, M., 1974 : Ph.D. dissertation, Stanford University, Stanford, California.

86. Morf, M., and Ljung, L., 1976 : 'Fast algorithms for recursive identification', Proc. Int. Conf. on Information Theory, Ronneby, Sweden.

87. Morf, M., Ljung, L., and Kailath, T., 1976 : 'Fast algorithms for recursive identification', Proc. IEEE Conf. on Decision and Control, Clearwater Beach, Florida.

88. Murphy, W.J., 1968 : 'Optimal stochastic control of discrete linear systems with unknown gain', IEEE Trans. on Aut. Control, 13, 338.

89. Nebeker, H.G., 1976 : 'A study of sequential hypothesis testing for system identification', UCLA, Ph.D. Thesis.

90. Panuska, V., 1968 : 'A stochastic approximation method for identification of linear systems using adaptive filtering', Proc. JACC, Ann Arbor, Michigan.

91. Panuska, V., 1969 : 'An adaptive least squares identification algorithm', Proc. 8th IEEE Symp. on Adaptive Processes.

92. Peterka, V., 1970 : 'Adaptive digital regulation of noisy systems', 2nd Prague IFAC Symp. on Identification and Process parameter estimation.

93. Peterka, V., 1972 : 'On steady state minimum variance control strategy', Kybernetika, 8, No. 3, 220.

94. Peterka, V., 1975 : 'A square root filter for real time multivariable regression', Kybernetika, 11, 53-67.

95. Prager, D.L., 1976: 'Application of Information Theory to Estimation and Identification', M.Sc. Dissertation, UMIST, Manchester.

96. Prager, D.L. and Wellstead, P.E., 1979 : 'Interactive recursive maximum likelihood estimation', 5th IFAC Symp. on Identification and System Parameter Estimation, Darmstadt.

97. Prager, D.L., and Wellstead, P.E., 1979a : 'Multivariable pole assignment self tuning regulators', Report 452, Control Systems Centre. (submitted for publication and under review)

98. Prager, D.L. and Wellstead, P.E., 1981 : 'Interactive maximum likelihood estimation', Int. J. Control, to appear.

99. Ragazzini, J.R. and Franklin, G.F., 1958 : 'Sampled data control systems', McGraw Hill.

100. Rissanen, J., 1974: 'Basis of invariants and canonical forms for linear dynamic systems', Automatica, 10, 175-182.

101. Risannen, J. and Ljung, L., 1976 : 'Estimation of optimal structures and parameters for linear systems', IFAC Symp. on Parameter Identification, Moscow.

102. Robins, A.J. and Wellstead, P.E., 1979 : 'An Appraisal of fast algorithms', 5th IFAC Symp. on Identification and system Parameter Estimation, Darmstadt.

103. Rosenbrock, H.H., 1960 : 'Automatic method for finding greatest or least value of function', The Computer Journal, 3, No. 3, 175-184.

104. Rosenbrock, H.H., 1969 : 'Design of multivariable control systems using inverse Nyquist array', Proc. IEE, 116, No. 11, 1929-1936.

105. Saradis, G.N. and Dao, T.K., 1972 : 'A learning approach to the parameter adaptive self-organizing problem', Automatica, 8, 589.

106. Saridis, G.N. and Stein, G., 1968 : 'Stochastic approximation algorithms for linear system identification' IEEE Trans. on Aut. Control, AC-13, 515-523 and 592-594.

107. Saucedo, R. and Schiring, E.E., 1968 : 'Introduction to continuous and digital control systems', Macmillan, New York.

108. Schwartz, S.C. and Steiglitz, K., 1971 : 'The identification and control of unknown linear discrete systems', Int. J. Control, 14, 43.

109. Sen, A. and Sinha, N.K., 1975 : 'A generalized pseudo-inverse algorithm for unbiased parameter estimation' Int. J. System Science, 6, 1103-1109.

110. Shinners, S.M., 1972 : 'Modern control system theory and applications', Addison-Wesley.

111. Sinha, A.K., 1977 : 'Minimum interaction minimum variance controller for multivariable systems', IEEE Trans. on Aut. COntrol, AC-22, 274.

112. Soderstrom, T., 1973 : 'An on-line algorithm for approximate maximum likelihood identification of linear dynamic systems', Report 7308, Lund Institute of Technology, Division of Automatic Control, Sweden.

113. Soderstrom, T., 1974a : 'Convergence properties of the generalized least squares identification method', Automatica, 10, 617-626.

114. Soderstrom, T., 1974b : 'Convergence of identification methods based on the instrumental variable approach', Automatica, 10, 685-688.

115. Soderstrom, T., 1977 : 'On model structure testing in system identification', Int. J. Control, 26, No. 1, 1-18.

116. Soderstrom, T., Gustavsson, I. and Ljung, L., 1974 : ' A comparative study of recursive identification methods', Report 7427, Dept. of Aut. Control, Lund Institute of Technology, Sweden.

117. Soderstrom, T., Gustavsson, I. and Ljung, L., 1975 : 'Identifiability conditions for linear systems operating in closed loop', Int. J. Control, 21, No. 2, 243-255.

118. Soderstrom, T., Ljung, L. and Gustavsson, I., 1978 : 'A theoretical analysis of recursive identification methods', Automatica, 14, 231-244.

119. Tomanek, R.L., 1976 : 'Sequential system identification by multiple hypothesis testing', Ph.D. Thesis, UCLA.

120. Tou, J.T., 1959 : 'Digital and sampled data control systems', McGraw Hill, New York.

121. Tse, E. and Athans, M., 1972 : 'Adaptive stochastic control for a class of linear systems', IEEE Trans. on Aut. Control, AC-18, 109.

122. Tse, E. and Bar Shalom, Y., 1973 : 'An actively adaptive control for linear systems with random parameters via the dual control approach', IEEE Trans. on Aut. Control, AC-18, 109.

123. Tse, E., Bar Shalom, Y. and Meier, L., 1973 : 'Wide sense adaptive dual control for non-linear stochastic systems', IEEE Trans. on Aut. Control, AC-18, 98.

124. Unbehauen, H. and Gohring, B., 1974 : 'Application of different statistical tests for the determination of the most accurate order of the model in parameter estimation', Automatica, 10, 233.

125. Van den Boom, A.J.W. and Van den Enden, A.W.M., 1974 : 'Determination of the order of process and noise dynamics', Automatica, 10, 233.

126. Van Trees, H.L., 1968: 'Detection, Estimation and Modulation Theory', John Wiley and Sons, New York.

127. Wald, A., 1945 : 'Sequential tests of statistical hypotheses' Annals of Math. Stat., 16, 117-186.

128. Wellstead, P.E. and Edmunds, J.M., 1975 : 'Least squares identification of closed loop systems', Int. J. Control, 21, No. 4, 689-699.

129. Wellstead, P.E., Edmunds, J.M., Prager, D.L. and Zanker, P., 1979: 'Self tuning pole-zero assignment regulators' Int. J. Control, 30, No. 1, 1-26.

130. Wellstead, P.E., Edmunds, J.M., Prager, D.L. and Zanker, P., 1980: 'Classical and Optimal Self Tuning Control', Int. J. Control, 31, No. 5, 1003-1005.

131. Wellstead, P.E., Prager, D.L. and Zanker, P., 1979 : 'Pole-assignment self tuning regulator', Proc. IEE, 126, No. 8, 781-787.

132. Wellstead, P.E. and Zanker, P., 1978 : 'On self tuning diesel engine governors', Control Systems Centre Report 422, UMIST, Manchester.

133. Wellstead, P.E. and Zanker, P., 1978a : 'Servo self tuners', Control Systems Centre Report 419, UMIST, Manchester.

134. Wieslander, J. and Wittenmark, B., 1971 : 'An approach to adaptive control using real time identification', Automatica, 7, 211-217.

135. Wilks, S.S., 1962 : 'Mathematical statistics', Wiley, New York.

136. Wittenmark, B., 1973: 'A self tuning regulator', Report 7311, Dept. of Aut. Control, Lund Institute of Technology, Sweden.

137. Wittenmark, B., 1974: 'A self tuning predictor', IEEE Trans. on Aut. Control, AC-19, No. 6, 848.

138. Wittenmark, B., 1975: 'Stochastic adaptive control methods: A survey', Int. J. Control, 21, No. 5, 705-730.

139. Wolovich, W.A., 1974: 'Linear Multivariable Systems', Springer Verlag, New York.

140. Wong, K.Y. and Polak, E., 1967 : 'Identification of linear discrete time system using the instrumental variable method', IEEE Trans. on Aut. Control, AC-12, 707-718.

141. Woodside, C.M., 1971: 'Estimation of the order of linear systems', Automatica, 7, No. 6, 727-733.

142. Young, P.C., 1968 : 'The use of linear regression and related procedures for the identification of dynamic processes', Proc. 7th IEEE Symp. on Adaptive Processes, UCLA.

143. Young, P.C., 1970 : 'An extension to the instrumental variable method for identification of a noisy dynamic process', Technical note CN/70/1, Dept. of Engineering, Univ. of Cambridge.

144. Young, P.C., 1974 : 'Recursive approaches to time series analysis', Bulletin Inst. of Maths. and Applications, 10, 209-224.

145. Young, P.C., 1976 : 'Some observations on instrumental variable methods of time-series analysis', Int. J. Control, 23, No. 5, 593-612.

146. Young, P.C., Shellswell, S.H. and Neethling, C.G., 1971 : 'A recursive approach to time series analysis', Report CUED/B-Control/TR16, Dept. of Engineering, Univ. of Cambridge.

APPENDIX 1

ANALYSIS OF COUPLED TANK SYSTEM

Fig. A1 shows the coupled tank system using the following notation:

- Q_{i_1} - input flow from pump 1 to tank 1
- Q_{i_2} - input flow from pump 2 to tank 2
- Q_2 - outlet flow from tank 2
- Q_3 - outlet flow from tank 1
- Q_1 - flow from tank 1 to tank 2
- H_1 - head of water in tank 1
- H_2 - head of water in tank 2
- H_3 - centre-line height of outflow orifices
- a_1 - area of inter-tank orifice
- a_2 - area of outlet orifice of tank 2
- a_3 - area of outlet orifice of tank 1
- A - surface area of each tank
- g - gravitational constant, 980cm/s^2
- C_{d_j} - orifice coefficient corresponding to orifice of area a_j
- V_1 - volume of tank 1 = AH_1
- V_2 - volume of tank 2 = AH_2

Units: Dimensions in cm

Areas in cm^2

Volumes in cm^3

Flow in cm^3/s

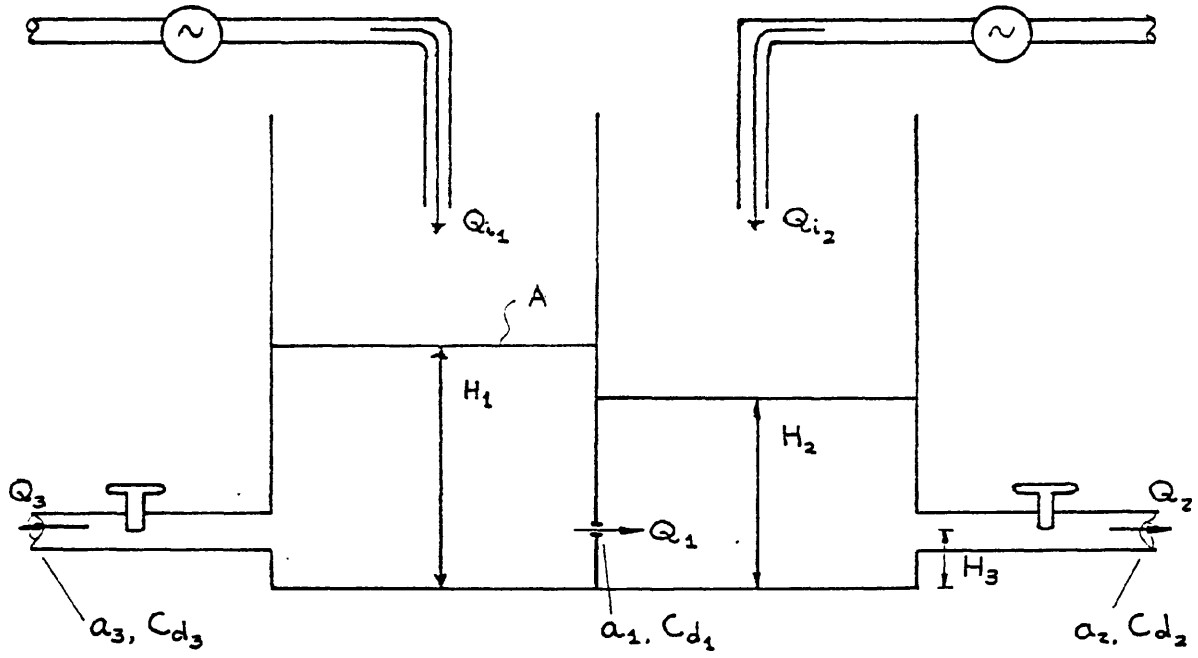


Fig. A.1 Coupled Tank System

The basic equations governing the system are:

$$\frac{dV_1}{dt} = A\frac{dH_1}{dt} = Q_{i_1} - Q_1 - Q_3 \qquad A1$$

$$\frac{dV_2}{dt} = A\frac{dH_2}{dt} = Q_{i_2} + Q_1 - Q_2 \qquad A2$$

Now, the flows Q_1, Q_2, and Q_3 can be expressed in terms of the respective water heads and orifice characteristics as follows:

$$Q_1 = C_{d_1} a_1 \sqrt{2g}\sqrt{H_1 - H_2} = 2k_1(H_1 - H_2) \qquad A3$$

$$Q_2 = C_{d_2} a_2 \sqrt{2g}\sqrt{H_2 - H_3} = 2k_2(H_2 - H_3) \qquad A4$$

$$Q_3 = C_{d_3} a_3 \sqrt{2g}\sqrt{H_1 - H_3} = 2k_3(H_1 - H_3) \qquad A5$$

where

$$k_1 = \tfrac{1}{2}(C_{d_1} a_1 \sqrt{2g})(\sqrt{H_1 - H_2})^{-1} \qquad A6$$

$$k_2 = \tfrac{1}{2}(C_{d_2} a_2 \sqrt{2g})(\sqrt{H_2 - H_3})^{-1} \qquad A7$$

and $$k_3 = \tfrac{1}{2}(C_{d_3} a_3 \sqrt{2g})(\sqrt{H_1 - H_3})^{-1} \qquad A8$$

In the steady state,

$$Q_{i_1} = Q_1 + Q_3 \qquad A9$$

and $$Q_{i_2} = Q_2 - Q_1 \qquad A10$$

Taking small variations q_{i_1} in Q_{i_1}, q_{i_2} in Q_{i_2}, q_1 in Q_1, q_2 in Q_2, q_3 in Q_3, h_1 in H_1, and h_2 in H_2 it is easy to show that

$$q_{i_1} - q_1 - q_3 = A\frac{dh_1}{dt} \qquad A11$$

$$q_{i_2} + q_1 - q_2 = A\frac{dh_2}{dt} \qquad A12$$

where $$q_1 = \frac{\partial Q_1}{\partial H_1} h_1 + \frac{\partial Q_1}{\partial H_2} h_2 = k_1(h_1 - h_2) \quad \text{A13}$$

$$q_2 = \frac{\partial Q_2}{\partial H_2} h_2 = k_2 h_2 \quad \text{A14}$$

and $$q_3 = \frac{\partial Q_3}{\partial H_1} h_1 = k_3 h_1 \quad \text{A15}$$

On taking Laplace Transforms, eqns. A11 and A12 may be assembled as follows:

$$\begin{bmatrix} s+(k_1+k_3)/A & -k_1/A \\ -k_1/A & s+(k_1+k_2)/A \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} 1/A & 0 \\ 0 & 1/A \end{bmatrix} \begin{bmatrix} q_{i_1} \\ q_{i_2} \end{bmatrix} \quad \text{A16}$$

Eqn. A16 describes the behaviour of the system about any operating point. The poles of the system are defined by the zeros of the determinant of the coefficient matrix of $(h_1 , h_2)^T$ in eqn. A16, i.e. by the zeros of:

$$s^2 + \frac{1}{A}(2k_1+k_2+k_3)\, s + \frac{1}{A^2}(k_1k_2+k_3k_1+k_3k_2) \quad \text{A17}$$

k_1, k_2 and k_3 are functions of the operating point and hence the poles of the multivariable system will change with the operating point.

Experiments on the rig showed that for $a_2=a_3=0.125\text{cm}^2$

$$\tfrac{1}{2}(C_d a_3 \sqrt{2g}) \simeq 2.2 \quad \text{A18}$$

(where $C_d = C_{d_2} = C_{d_3}$ as similar taps were used) and for $a_1=0.3\text{cm}^2$

$$\tfrac{1}{2}(C_{d_1} a_1 \sqrt{2g}) \simeq 5 \quad \text{A19}$$

Also, $A = 105\text{cm}^2$ and $H_3 = 1.2\text{cm}$. Thus from eqns. A6 - A8:

$$k_1 \simeq 5(\sqrt{H_1 - H_2})^{-1} \quad \text{A20}$$

$$k_2 \simeq 2.2(\sqrt{H_2 - H_3})^{-1} \quad \text{A21}$$

$$k_3 \simeq 2.2(\sqrt{H_1 - H_3})^{-1} \quad \text{A22}$$

Due to the limited pump and drain capacity of the system the actual possible steady state level differences between the two tanks could not exceed approximately 2cm. Since, generally,

$$H_1 >> 2\text{cm} \tag{A23}$$

$$H_2 >> 2\text{cm} \tag{A24}$$

and

$$\Delta H = |H_1 - H_2| < 2\text{cm} \tag{A25}$$

it is fair to set

$$H \simeq H_1 \simeq H_2$$

$$k \simeq k_2 \simeq k_3 = 2.2(\sqrt{H - 1.2})^{-1} \tag{A26}$$

Then the system time constants are approximately given by

$$T_1 = A/(2k_1+k) \tag{A27}$$

$$T_2 = A/k \tag{A28}$$

Now,

$$2k_1 >> k \tag{A29}$$

so T_1 is proportional mainly to $1/k_1$, or

$$T_1 \propto \sqrt{\Delta H} \tag{A30}$$

$$T_2 \propto \sqrt{H - 1.2} \tag{A31}$$

From eqn. A30 it is clear that fairly small changes in the difference between the tank levels can lead to quite substantial fluctuations in the time constant T_1, whereas T_2 will be more dependent on changes in the actual tank levels approximated by H. The following table gives sample values of these parameters, computed from equation A17.

	H_1 (cm)	H_2 (cm)	T_1 (s)	T_2 (s)
1.	8.0	7.0	9.65	119
2.	13.0	12.0	9.93	159
3.	22.0	21.0	10.0	215
4.	13.0	11.0	13.56	156
5.	13.0	12.5	3.55	162

Entries 1,2 and 3 illustrate how T_2 increases as the tank levels increase whilst a constant head difference of 1cm is maintained between the two tanks. Modifying the head difference results in substantial changes in T_1, as shown in entries 4 and 5.