

AD-A072 147

CALIFORNIA UNIV LOS ANGELES SCHOOL OF ENGINEERING A--ETC F/G 12/2
MAXIMUM LIKELIHOOD IDENTIFICATION OF LINEAR DISCRETE STOCHASTIC--ETC(U)
JUL 78 A J GLASSMAN, C T LEONDES F33615-77-C-3013

UNCLASSIFIED

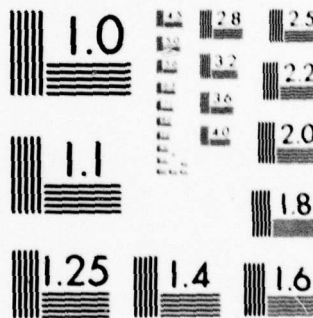
AFFDL-TR-78-84

NL

1 OF 3

AD
A072147





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFFDL-TR-78-84

B.S. 2

AD A072147

LEVEL

II

**MAXIMUM LIKELIHOOD IDENTIFICATION OF
LINEAR DISCRETE STOCHASTIC SYSTEMS**

UNIVERSITY OF CALIFORNIA, LOS ANGELES
SCHOOL OF ENGINEERING AND APPLIED SCIENCE
7620 BOELTER HALL, UCLA
LOS ANGELES, CALIFORNIA 90024

July 1978

DDC
RECEIVED
AUG 1 1978
RECEIVED

TECHNICAL REPORT AFFDL-TR-78-84
Final Report for Period 1977-1978

Approved for public release; distribution unlimited.

DDC FILE COPY

AIR FORCE FLIGHT DYNAMICS LABORATORY
AIR FORCE WRIGHT AERONAUTICAL LABORATORIES
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433

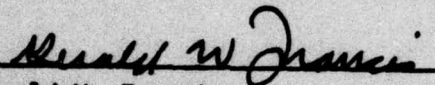
79 07 30 053

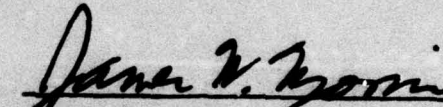
NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report has been reviewed by the Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.


Gerald W. Francis


JAMES W. MORRIS, Chief
Control Systems Development Branch
Flight Control Division

FOR THE COMMANDER


MORRIS A. OSTGAARD
Assistant for Research and
Technology
Flight Control Division

"If your address has changed, if you wish to be removed from our mailing list, or if the addressee is no longer employed by your organization please notify AEEDL/EGI, N-PAFB, OH 45433 to help us maintain a current mailing list".

Copies of this report should not be returned unless return is required by security considerations, contractual obligations, or notice on a specific document.

19 REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFFDL TR-78-84	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) MAXIMUM LIKELIHOOD IDENTIFICATION OF LINEAR DISCRETE STOCHASTIC SYSTEMS	5. TYPE OF REPORT & PERIOD COVERED 5/77-5/78	
7. AUTHOR(s) A. J. Glassman (C. T. Leondes Principal Investigator)	8. CONTRACT OR GRANT NUMBER(s) F33615-77-C-3013	
9. PERFORMING ORGANIZATION NAME AND ADDRESS School of Engineering and Applied Science University of California, Los Angeles Los Angeles, CA 90024	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2307-03-02	
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Air Force Systems Command AFFDL/ FGL WPAFB, OH 45433	12. REPORT DATE Jul 1978	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Office of Naval Research 1030 E. Green Street Pasadena, CA 91106 Attention: Mr. Perry Beilke	13. NUMBER OF PAGES 203	
15. SECURITY CLASS. (of this report) Unclassified		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) Final rept. for May 77-May 78		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Maximum Likelihood Identification, Stochastic System, Flight Control Systems, Identification of Parameters, Estimation Theory		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) The method of maximum likelihood is applied to the identification of parameters in systems described by linear difference equations. The equations are assumed to be completely known except for the state variable coefficients, i.e., the state transition matrix, and, in certain situations, the initial conditions. The estimates are based on known normal operating input and on output measurements corrupted by additive Gaussian noise. Maximum likelihood estimators of the parameters are developed for the following four cases:		

next page

mt

initial condition known, initial condition unknown parameter, initial condition unknown random variable, and an equivalent equation-error model configuration. Finite sample and asymptotic properties of the estimators as well as computational aspects are investigated. The study is oriented toward real time applications. Application of maximum likelihood to the above four cases differs from the classical situation in statistics because the measurements are not identically distributed or are not independent or both. The resulting estimates are roots of cumbersome nonlinear equations. First, the likelihood equations of each of the four cases are developed. Generally, they are found to be expressible as polynomials in the unknown parameters. The degree and complexity of the polynomial likelihood equations grow with the number of samples upon which the estimates are based. The theorems on minimal sufficient statistics by Dynkin shown that this characteristic is unavoidable. Some finite sample properties of the likelihood equation root corresponding to the maximum likelihood estimate for the scalar model are sought through averaging or using limiting conditions. As the measurement noise variance goes to zero, the maximum likelihood estimate is seen to approach the true value of the parameter. This conclusion is extended for stable autonomous systems by showing that the true parameter value is the only stable root, i.e., the only root in $(-1,1)$. Simulation results indicate this extension holds for forced systems also. A related result but without the above assumptions is also proven. In all four cases, on the average the true parameter value is a root of the corresponding likelihood equation. Also, root distributions as found from Monte Carlo simulations are displayed. Consistency is shown when the initial condition is known and the system is stable and is inferred for the unknown initial condition cases. If the system is stable and autonomous, the proof for consistency fails because of a loss of uniqueness in the limit. Numerical solutions for the estimates depend on such factors as root distribution, likelihood equation sensitivity to coefficient perturbations, and shape of the likelihood function derivative. Simulation results for stable systems show that at least up to moderate noise levels, while the likelihood equation has multiple roots, generally none of the other roots appear to lie in the neighborhood of the maximum likelihood estimate, and the likelihood function derivative is relatively smooth and insensitive to noise in the neighborhood of the maximum likelihood estimate. To overcome continued increase in the number of computations and in the amount of storage required, two approximations are proposed. In one, the likelihood equation polynomials are truncated and averaged coefficients are used. In the other, a curve fitting scheme is presented which exploits a recursive aspect of the likelihood function derivative. The former in its simplest form has only two real roots, one in the stable region and the other outside. Under certain restrictions it is shown to yield a consistent estimate.

PREFACE

Flight vehicle parameter analysis and synthesis requires---indeed, demands---most effective vehicle parameter determination techniques for many reasons including vehicle parameter confirmation. This report presents some of the most powerful results developed to date.

Accession For		☒ ☐ ☐
NTIS GRA&I		
DDC TAB		
Unannounced Justification		
By _____		
Distribution/		
Availability Codes		
Dist	Avail and/or special	
A		

TABLE OF CONTENTS

Section	Page
I INTRODUCTION	1
II BACKGROUND	5
2.1 Parameter identification in classical statistics.	5
2.2 Parameter identification in control systems	12
2.3 Identifiability	25
III THE LIKELIHOOD EQUATIONS	27
3.1 Introduction	27
3.2 Problem statement	29
3.3 The necessary condition - x_0 known.	31
3.4 The necessary condition - x_0 unknown parameter.	34
3.5 The necessary condition - distribution of x_0 known.	37
3.6 The necessary condition - difference equation error approach.	42
3.7 Plant noise	52
3.8 Minimal sufficient statistics	56
IV PROPERTIES OF THE IDENTIFIERS AND THEIR APPROXIMATIONS	61
4.1 Introduction	61
4.2 Finite sample characteristics	61
4.3 Large sample characteristics.	81

LIST OF ILLUSTRATIONS

Figure		Page
1	Equation Error Configuration for Identification	16
2	Model-Plant Error Configuration for Identification.	16
3	Basic Model	30
4	Comparison of ML, Least Squares, Average Coefficient, and 3-Point Fit Estimation for x_0 Known Case ($a_0=-0.5$, $\sigma^2=0.01$).	104
5	Comparison of ML, Least Squares, Average Coefficient, and 3-Point Fit Estimation for x_0 Known Case ($a_0=0.75$, $\sigma^2=0.01$).	105
6	Comparison of ML, Least Squares, Average Coefficient, and 3-Point Fit Estimation for x_0 Known Case ($a_0=-0.5$, $\sigma^2=0.4356$)	106
7	Cost Function of the MLE in the x_0 Known Case for 3, 5, 10, 15, and 20 samples ($a_0=-0.5$, $\sigma^2=0.01$)	107
8	Derivative Function of the MLE in the x_0 Known Case for 3, 5, 10, 15, and 20 samples ($a_0=-0.5$, $\sigma^2=0.01$)	108
9	Derivative Function fo the MLE in the x_0 Known Case for 3, 5, 10, 15, and 20 samples ($a_0=0.75$, $\sigma^2=0.01$)	109
10	Derivative Function of the MLE in the x_0 Known Case for 3, 5, 10, 15, and 20 samples ($a_0=-0.5$, $\sigma^2=1.0$)	110
11	No Noise Derivative Function of the MLE in the x_0 Known Case for 3, 5, 10, 15, and 20 samples ($a_0=-0.5$, $\sigma^2=0$)	111
12	Frequency Distribution for MLE of a_0 in x_0 Known Case ($a_0=0.75$, $\sigma^2=0.01$)	112
13	Comparison of ML, Least Squares, Average Coefficient, and 3-Point Fit Estimation for x_0 Unknown Parameter ($a_0=-0.5$, $\sigma^2=0.01$)	113
14	Derivative Function of the MLE in the x_0 Unknown Parameter Case for 3, 5, 10, 15, and 20 samples ($a_0=0.75$, $\sigma^2=0.01$)	114
15	Derivative Function of the MLE in the x_0 Unknown Parameter Case for 3, 5, 10, 15, and 20 samples ($a_0=-0.5$, $\sigma^2=1.0$)	115

LIST OF ILLUSTRATIONS (continued)

Figure		Page
16	No Noise Derivative Function of the MLE in the x_0 Unknown Parameter Case for 3, 5, 10, 15 and 20 samples ($a_0=-0.5$, $\sigma^2=0$)	116
17	Frequency Distribution for MLE of a_0 in x_0 Unknown Parameter Case ($a_0=0.75$, $\sigma^2=0.01$)	117
18	Comparison of ML, Least Squares, Average Coefficient, and 3-Point Fit Estimation for x_0 Random Variable ($a_0=-0.5$, $\sigma^2=0.01$)	118
19	Derivative Function of the MLE in the x_0 Random Variable Case for 3, 5, 10, 15, and 20 samples ($a_0=0.75$, $\sigma^2=0.01$)	119
20	Derivative Function of the MLE in the x_0 Random Variable Case for 3, 5, 10, 15, and 20 samples ($a_0=-0.5$, $\sigma^2=1.0$)	120
21	No Noise Derivative Function of the MLE in the x_0 Random Variable Case for 3, 5, 10, 15, and 20 samples ($a_0=-0.5$, $\sigma^2=0$)	121
22	Frequency Distribution for MLE of a_0 in x_0 Random Variable Case ($a_0=0.75$, $\sigma^2=0.01$)	122
23	Comparison of ML, Least Squares, Average Coefficient, and 3-Point Fit Estimation for Differencing Approach ($a_0=-0.5$, $\sigma^2=0.01$)	123
24	Derivative Function fo the MLE for the Differencing Approach for 3, 5, 10, 15, and 20 samples ($a_0=0.75$, $\sigma^2=0.01$)	124
25	No Noise Derivative Function of the MLE in the Differencing Approach for 3, 5, 10, 15, and 20 samples ($a_0=-0.5$, $\sigma^2=0$)	125
26	No Noise Derivative Function of the MLE in the Differencing Approach for 3, 5, 10, 15, and 20 samples ($a_0=0.75$, $\sigma^2=0$)	126
27	Frequency Distribution for MLE of a_0 in Differencing Approach ($a_0=0.75$, $\sigma^2=0.01$)	127
28	Cost Function Along A_0 in x_0 Known Case for 3, 5, 10, 15, and 20 samples.	131
29	Norm of Derivative Function Along A_0 in x_0 Known Case for 3, 5, 10, and 15 samples.	132

LIST OF ILLUSTRATIONS (continued)

Figure		Page
30	Directional Derivative Along A_0 in x_0 Known Case for 3, 5, 10 and 15 samples	133
31	Cost Function Along A_0 in x_0 Unknown Parameter Case for true x_0 and 3, 5, 9, 13, and 18 samples	134
32	Norm of the Derivative Function Along A_0 in x_0 Unknown Parameter Case for $x_0=\hat{x}_0$ and 3, 5, 9, 13, and 18 samples.	135
33	Directional Derivative Along A_0 in x_0 Unknown Parameter Case for $x_0=\hat{x}_0$ and 3, 5, 9, 13, and 18 samples.	136

SECTION I

INTRODUCTION

The experimental determination of the numerical values of parameters in an otherwise completely known mathematical model of a system based on measurements of quantities which are functions of the parameters is generally known as parameter identification. When the measurements are subject to random inaccuracies (noise), the identification problem can no longer be trivial. Parameter identification using noisy measurements is the problem of estimating a parameter θ , whose true value is θ_0 , from a sample $(x_1, \dots, x_n)$ assumed to have been drawn from a population having a distribution function of specified functional form $F(x; \theta)$ but where θ is unknown and $\theta, \theta_0 \in \Theta$, the set of admissible values of θ .

The identification problem arises in the development of mathematical models of systems. Frequently, physical laws or established empirical relationships exist from which the functional form of the model can be determined. However, physical, engineering, or economic limitations can prevent the direct measurement of certain aspects of the system required to completely satisfy the model.

Because the mathematical model is a common tool for analysis in many fields, the identification problem is similarly widespread. Specific examples occur in economics, biology, geology and engineering. The need for identification in engineering often comes about as part of a larger problem - optimum or adaptive automatic control of systems subject to stringent performance requirements. This situation can be found in such areas as industrial process control and control of high performance aircraft and aerospace vehicles.

The characteristic of identification in automatic control applications that tends to make it distinct from those in other areas is the relatively short time in which the identification must be accomplished to be useful. The identification problem generally takes the form of completing the description of the relationships between the input states of the plant and its output states. The parameters to be identified usually are the coefficients of equations (difference, differential, or partial differential) of the plant model taken as linear with constant or slowly varying coefficients. Noisy output measurements are assumed, but because inputs to the plant can often be generated with considerably less uncertainty than exists in the measurement of the output states, input signals frequently are taken as known. The identification is carried out with normal operating input or with no more than minor perturbations to the input. (If no restrictions on input exist, then conceivably the identification problem could be made trivial by adjusting the input so that the output signal swamps the measurement noise.)

The number of techniques available for identification of parameters is large. Among the various possibilities that are statistical in nature, maximum likelihood is often considered as a standard of comparison largely because of the desirable large sample properties it typically has. To use this method sufficient information must be available to determine the functional form of the distribution of the measurements. Considering the joint probability distribution of the measurements as a function of the unknown parameter θ , the maximum

likelihood estimate of θ is that value of $\theta \in \Theta$ for which the function is a maximum. The estimate is usually found by seeking the roots of the derivative of that function (known as the likelihood function).

The literature on identification in control systems is fairly extensive, but only a relatively small portion is devoted to maximum likelihood estimation (probably, because of the fact that with the exception of certain special cases, the evaluation of the estimate can be a difficult numerical problem). One important aspect in the application of maximum likelihood estimation to the identification of parameters in dynamic systems about which there appears to be little written is the effect that various levels of information on the initial conditions of the system have on the form, properties, and ease of solution of the estimator. The primary purpose of this study is to investigate these effects. The basic model used was a linear constant coefficient difference equation plant whose output measurements were corrupted with additive gaussian noise.

Chapter 2 discusses in considerable depth the identification problem as it arises in modeling dynamic systems and presents an extended review of the pertinent literature.

In Chapter 3 the maximum likelihood estimators are developed for the basic model under each of three assumptions on the nature of the initial condition - known, unknown parameter, and unknown random variable with known gaussian distribution. A fourth situation involving correlated noise and based on an equivalent form of the basic model is also treated. It represents an extension of a much earlier work and treats the initial measurement as a known deterministic initial condition.

The estimators are given in the form of likelihood functions for scalar and, with one exception, for matrix parameters in the case without plant noise and for only scalar parameters with plant noise. Because the likelihood equations grow in complexity with the number of measurements, the question of existence of minimal sufficient statistics which would overcome this problem was examined.

In Chapter 4 an analytical investigation of the properties of the estimators of Chapter 3 is made. The characteristics of the roots of the likelihood equations and the number of stable roots for finite samples are investigated. Large sample properties of the estimates are established. Averaging approximations to the maximum likelihood estimate are proposed. Their finite sample and large sample properties are also discussed.

In Chapter 5 the numerical aspects of the estimators developed in Chapter 3 are treated. Results for evolution of the estimates as the number of samples increases, examples of the functional behavior of the derivatives of the joint distributions of the measurements and histograms for root distribution based on Monte Carlo simulations are given. Numerical evaluation of the roots of the likelihood equation and a recursive curve fitting approximation are considered. The averaging approximation and the curve fitting approximation are simulated and compared to maximum likelihood and least squares.

In Chapter 6 the summary of results and conclusions are presented.

SECTION II

BACKGROUND

Systems identification is concerned with the experimental determination of a mathematical model to characterize a system through the use of measured input-output data. Frequently, this problem appears in a form where the only unknown aspects are the numerical values of various parameters in an otherwise completely defined mathematical model. This situation is referred to as parameter identification (or parameter estimation).

Early methods of parameter identification in control systems tended to be ad hoc in nature. Later, the already highly developed identification techniques in statistics, and in particular, maximum likelihood estimation, were adapted to applications in control systems problems. Thus, a complete view of the development of parameter estimation in control systems requires an appreciation for the relevant contributions in the field of statistics as well as in control systems.

2.1 PARAMETER IDENTIFICATION IN CLASSICAL STATISTICS

The problem of parameter estimation in classical statistics deals with obtaining a best estimate, in some statistical sense, of a parameter vector θ on the basis of measurements y_i which are in error. In general, the model takes the form:

$$y_i = f_i(\theta) + e_i \quad (2.1)$$

(The term "regression" is usually reserved for estimation with this model.)

2.1.1 MODEL DEVELOPMENT

The basic thread that runs through the branch of statistics that moved to the point where it had essentially direct application in control systems was the development of models of stochastic systems, especially in time series analysis. According to Parzen [1961] and Wold [1954], a series of advances in the modeling of stochastic systems was made starting in the 1920's. Yule in 1927 developed the scheme of linear autoregression by modeling observations x_t as linear combinations of previous observations plus noise, i.e.,

$$x_t = a_1 x_{t-1} + \dots + a_m x_{t-m} + e_t \quad (2.2)$$

where m is the order of the autoregressive scheme and the sequence $\{e_t\}$ consists of independent identically distributed random variables. Also in 1927, Slutsky introduced the notion of a moving average scheme in which observations x_t are assumed to be generated by a shifting linear combination of members of an independent identically distributed sequence of random variables $\{\eta_i\}$, i.e.,

$$x_t = a_0 \eta_t + a_1 \eta_{t-1} + \dots + a_m \eta_{t-m} \quad (2.3)$$

The theory of discrete, random, stationary processes emerges with the work of Kintchine during 1932-1934. Wold in 1938 combines the work of the above to show that moving average schemes and autoregressive schemes are special cases in the theory of stationary random processes. Finally, combining moving average and autoregression schemes yields a model that is closely related to the modern linear stochastic control theory model:

$$x_{i+1} = Ax_i + Bu_i + w_i$$

$$y_i = Hx_i + \eta_i \quad (2.4)$$

where the measurements are y_i , and the $\{\eta_i\}$ and $\{\omega_i\}$ are noise sequences.

2.1.2 CLASSICAL METHODS OF PARAMETER IDENTIFICATION

A number of statistical schemes for identification exist. Intuitively, a generally desirable property of an estimator $\hat{\theta}$ compared to any other estimator $\tilde{\theta}$ would be minimum mean square error in an admissible set,

$$E(\hat{\theta} - \theta)^2 \leq E(\tilde{\theta} - \theta)^2 \quad (2.5)$$

or minimum variance if $\hat{\theta}$ and $\tilde{\theta}$ are unbiased,

$$V(\hat{\theta}) \leq V(\tilde{\theta}) \quad (2.6)$$

where E is the expectation operator and V is the variance operator. Frequently, identification problems in control systems are approached by directly seeking a scheme with one of the above properties. In most of the remaining cases where statistical estimates are sought, one of the classical statistical methods is chosen. The three most popular methods appear to be Gauss-Markov, maximum likelihood, and Bayes.

The Gauss-Markov estimation technique applies to the linear model,

$$y = H\theta + e \quad (2.7)$$

where the expectation $E(e) = 0$, $E(ee^T) = R > 0$, and H is assumed to have maximal rank. Minimization of the cost function,

$$c = e^T R^{-1} e \quad (2.8)$$

leads to the Gauss-Markov estimate,

$$\hat{\theta} = (H^T R^{-1} H)^{-1} H^T R^{-1} y \quad (2.9)$$

If $R = \sigma^2 I$, σ^2 a scalar constant, then $\hat{\theta}$ is known as the least squares estimate. The Gauss-Markov estimate is the minimum variance unbiased linear estimate (Rao (1965)).

Note that the Gauss-Markov estimate requires knowledge of the first two moments of the probability distribution of the error e . A least squares estimate by virtue of its definition can be used with no second moment information. The term "least squares" is often used to describe more general minimum mean square error estimations than its strict definition would encompass.

An interesting bit of history is related to the development of the method of maximum likelihood estimation. Undoubtedly, the earliest major milestone in parameter estimation occurs in the works of Gauss where he places the method of least squares on a rather firm foundation. Curiously enough, Gauss apparently used the principle of maximum likelihood to accomplish this but later rejected the principle as a meaningful approach to estimation in its own right. Edgeworth [1908] in a translation of a letter from Gauss to Bessel in 1839 reveals Gauss stating,

"...That the metaphysic employed in my Theoria Motus Corp. Coel. to justify the method of least squares has been subsequently allowed by me to drop has chiefly occurred for a reason that I have myself not mentioned publicly. The fact is, I cannot but think it in every way less important to ascertain that value of an unknown magnitude the probability of which is greatest-which probability is nevertheless infinitely small-rather than that value by employing which we render the Expectation of detriment a minimum...."

However, this stigma on maximum likelihood estimation was finally overcome in 1922 by Fisher (Berkson [1956]).

The method of maximum likelihood estimation requires that the probability density of the error (or noise) be known at least within some constants, e.g., mean or variance. Using the equations of the

model and the distribution of the noise, the likelihood function L can in principle be determined where

$$L = p(y_1, \dots, y_n, \theta) \quad (2.10)$$

For a given set of measurements $y_1, \dots, y_n$, the value of the unknown parameters θ which maximizes L is the maximum likelihood estimate of θ .

According to Rao [1965], [1952], and Finney [1968], maximum likelihood (ML) estimates, $\hat{\theta}_L$, have several desirable properties under a wide variety of situations. Among them are:

1. The ML estimate is consistent, $\hat{\theta}_L \xrightarrow{P} \theta$ or $\hat{\theta}_L \xrightarrow{a.s.} \theta$.
2. The ML estimate is asymptotically efficient, i.e., among consistent estimators it has minimum variance in the limit.
3. For large samples, the distribution of $\hat{\theta}_L$ becomes normal.
4. If L possesses a sufficient estimator for θ , then $\hat{\theta}_L$ is sufficient.

Berkson [1956] points out that the nice properties of ML estimates occur in the limit. This asymptotic information is not necessarily useful in any practical situation. Except in those special cases where the least squares estimate and ML estimate are identical, e.g., with normal distributions, little has been said about finite sample properties of ML estimates.

For Bayesian estimation, some *a priori* information on the probability densities of the parameters θ in addition to the noise densities is required. The *a posteriori* density $p(\theta/y)$ is found from Bayes' Rule,

$$p(\theta/y) = p(y/\theta)p(\theta)/p(y) \quad (2.11)$$

Various types of estimates can be obtained from this density (Ho and Lee[1964] and Stear [1970]). The minimum variance unbiased estimate

of θ is the conditional mean of $p(\theta/y)$. Often this is difficult to find, and instead the mode of $p(\theta/y)$ is used as an estimate of θ . The latter estimate is sometimes known as a *posteriori* maximum likelihood.

2.1.3 MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS IN DIFFERENCE

EQUATIONS

Prior to the development of modern control theory, i.e., prior to about 1960, the literature contains relatively few examples of general applications of maximum likelihood estimation for identification of parameters in difference equations. Two of the better known and more significant contributions are briefly reviewed below.

Koopmans [1937] in a monograph on linear regression investigates maximum likelihood estimation of regression coefficients. In state variable notation, the model in his more general case is of the form:

$$\begin{aligned}x_{i+1} &= Ax_i + u_i \quad i = 0, 1, \dots, N \\ y_i &= x_i + \eta_i\end{aligned} \tag{2.12}$$

where the $\{\eta_i\}$ is a sequence of zero mean independent normal random vectors with covariance R ,

$$R = \sigma^2 [r_{ij}]$$

and where,

$$u_i^T = (0, \dots, 0, c) \quad (c = \text{scalar constant})$$

and A is a companion matrix, i.e.,

$$A = \begin{bmatrix} 0 & I \\ \underline{a}^T \end{bmatrix}$$

Using maximum likelihood he estimates σ^2 , c , x_i , and $\underline{a}$. (An explicit expression for the estimate of $\underline{a}$ generally cannot be given because to

find the estimate an eigenvalue problem must be solved.) In addition, he looks into the asymptotic properties of the estimates and the significance of R being singular.

Mann and Wald [1943] treat a related problem. Their model in the scalar case has the form:

$$\begin{aligned} x_t &= \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \dots + \alpha_p x_{t-p} + \alpha_0 + \epsilon_t \\ y_t &= x_t \end{aligned} \quad (2.13)$$

where the regression equation is assumed to be stable and the $\{\epsilon_t\}$ are independent normally distributed random variables with zero mean and variance σ^2 . The maximum likelihood estimates of $\alpha_1, \dots, \alpha_p, \alpha_0, \sigma^2$ are shown to be the solution of a set of linear algebraic equations. They prove the estimates are consistent and asymptotically normal.

The second half of the paper deals with the more general case of several equations in several variables, i.e.,

$$\sum_{j=1}^r \sum_{k=0}^{p_{ij}} \alpha_{ijk} x_{j,t-k} + \alpha_i = \epsilon_{it} \quad i = 1, \dots, r \quad (2.14)$$

or in state variable form with A_{ij} an $n \times n$ companion matrix and assuming the matrix $\begin{bmatrix} \alpha_{i,j,0} \end{bmatrix}$ is non-singular

$$x_{i+1,j} = \sum_{k=1}^n A_{jk} x_{ik} + a_j + \epsilon_{i+1,j} \quad j=1,2,\dots,r \quad (2.15)$$

where:

$$a_j^T = (0, \dots, 0, \tilde{a}_j), \quad \epsilon_{ij}^T = (0, \dots, 0, \tilde{\epsilon}_{ji}), \quad n = \max_{i,j} p_{ij}$$

A development similar to the scalar case, but considerably more complex, is given for this case.

2.2 PARAMETER IDENTIFICATION IN CONTROL SYSTEMS

Parameter identification in the terminology of stochastic control systems generally refers to the estimation of the parameters p in the otherwise known functional relations (or their discrete equivalent):

$$\begin{aligned}\dot{x} &= g(x, u(t), p, \xi(t), t) \\ y &= h(x, p, t) + \eta(t)\end{aligned}\tag{2.16}$$

where η and ξ are unknown random variables and u , t , and the measurement y are known. In the linear case, a typical form would be:

$$\begin{aligned}\dot{x} &= Ax + Bu + \xi \\ y &= Hx + \eta\end{aligned}\tag{2.17}$$

where the elements of p would be the elements of the matrix A and, in addition, could include the elements of B and H .

Occasionally, the parameter vector p also includes unknown parameters in the description of the processes η and ξ . State estimation, estimating $x(t)$, is in a sense also identification but is treated as a separate problem in control systems except in those situations where identification is most efficiently achieved when carried out jointly with state estimation.

The literature on identification in control systems has been quite extensive and varied. Significant numbers of publications began to appear in the middle and late 1950's and have continued to the present. The early applications naturally tended to be *ad hoc* in approaches and typically were concerned with estimating the impulse response of linear systems. By the early 1960's, approaches employing more powerful statistical techniques were appearing with greater frequency. Within a few years, most of the applicable tools of the statistician seem to have been borrowed by the controls engineer.

Before launching into a review of the literature one would hope to be able to identify categories into which to group the many contributions. There seems to be no satisfactory way to do this. However, for this discussion two major groupings serve as a guide - those methods which are based on least squares or minimum square error criterion and those which are statistical in nature. Within these classifications, the literature can be grouped by specific techniques. There are other characteristics which could be used such as (1) the model type - linear or nonlinear, continuous or discrete, single or multiple input/output, constant coefficient or time varying or (2) the quantity being identified - difference or differential equation coefficients, Laplace or Z-transform coefficients, or the entire impulse response or (3) restrictions such as real time estimation or use of only normal operating input.

2.2.1 MINIMUM SQUARE ERROR METHODS

Among the popular early methods and ones which continue to receive attention are numerical deconvolution and related impulse response approximation methods. These techniques are not truly parameter identification methods in the sense of the definition given earlier. Their objective is to determine some best values for undetermined parameters in combinations of functions which are intended to approximate the input/output characteristics of the actual system. (In numerical deconvolution these parameters are discrete time values of the impulse response.)

In this regard Zabusky (1956) works with the convolution equation of a linear continuous system

$$x(t) = \int_0^{\infty} h(\tau)u(t-\tau)d\tau \quad (2.18)$$

and seeks the value of the system's impulse response $h(t)$ at discrete points in time. He approximates the impulse response by products of exponentials and polynomials with undetermined parameters. The parameters are fixed by minimizing ϵ where

$$\epsilon = \int_0^T [x(t) - x_c(t)]^2 dt \quad (2.19)$$

and $x_c(t)$ is the output of the approximate model. The two systems are subject to identical inputs.

Goodman and Reswick [1956] perform a similar investigation but use delay lines and recognizing the noise problems with convolution equations, base the deconvolution on the correlation equation

$$\phi_{xu} = \int_{-\infty}^t h(t-\tau)\phi_{uu}(\tau)d\tau \quad (2.20)$$

Goodman [1957] extends his previous year's work to multiple input/output systems. Taylor series is used with the convolution integral by Braun [1959]. Orthogonal filters are used by Elkind, et al. [1963], Eykhoff [1963], and Kekre and Glenski [1968] but with different approaches.

The model reference technique is similar to those already discussed except that this method is used when the true system transfer function is known but for some or all the coefficients. Again, the system and the model are driven by the same input. The free parameters in the model are adjusted to minimize some measure of the differences in their outputs, generally the integral of the squares of the difference

Probably the most frequently referenced paper in this area is the one by Margolis and Leondes [1959]. They use the integral of the square of the output error and its derivatives as their cost function and by a gradient method drive the coefficients in the model to minimize the cost. Surber [1963] presents a relatively comprehensive investigation of the model reference method. Hsia and Vimolvanich [1969] apply the technique to the tracking of variable parameters in a linear control system. They adjust the model parameters by differential equations and explicitly account for measurement noise.

There is another group of methods that is perhaps best described as least squares methods. Included here is Turin [1957] who estimates the impulse response of a system by designing a filter whose input is the output of the system and whose output is the estimate. The design criterion is the integral of the square of the difference between the true and estimated responses. (Techniques like that of Turin which give continuous real time estimation of an unknown impulse response often are referred to as "matched-filter identification" (Gibson [1963]).) King [1967] proposes an off-line gradient solution to the problem of identifying system parameters subject to the cost function, the integral of the square of the difference in true and measured output. A similar situation but with a discrete model is treated by Aoki [1967a] where, in addition, the feasibility of estimating only part of the A matrix when the A matrix contains some known elements is considered. Dolbin [1969] also deals with the discrete control problem but has unknown parameters in A , B , and H matrices. He develops a gradient type solution.

The equation error model approach, Figure 2-1a, refers to the least squares regression problem where the system is represented by a difference equation

$$x_i + a_1 x_{i-1} + \dots + a_p x_{i-p} = b_1 u_{i-1} + \dots + b_p u_{i-p} \quad (2.21)$$

and the coefficients a_k, b_k are found by introducing measured (as opposed to true) input-output data into the equation and minimizing the resulting error. By contrast, the model-plant error approach, Figure 2-1b, seeks to minimize the difference in measured plant output and model output for the same input through adjustment of a_k and b_k .

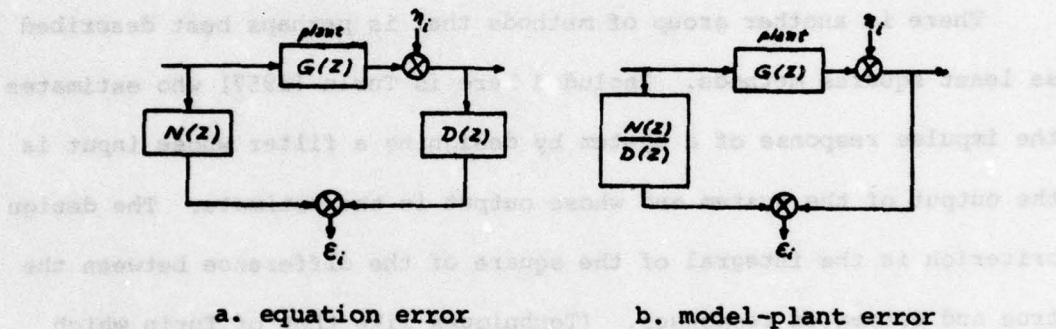


Figure 2-1

The equation error model problem led to a series of papers. Kalman [1958] treated the equation error problem without noise. He sought the coefficients in N and D , both of which are polynomials in z , in order to minimize the sum of ϵ_i^2 . (See Figure 2-1.) This was accomplished by Gauss-Siedel iteration on a set of equations involving weighted correlation functions. Later, Steiglitz and McBride [1965] solve the model-plant error problem by repeated application of the equation error method and prefiltering. Lion [1967] improves on Steiglitz and McBride by introducing filters prior to $N(z)$ and $D(z)$

in the equation error scheme and uses a gradient on ϵ^2 to derive the polynomial coefficients. (A special version of this solution was given earlier by Weygandt and Puri [1961].) The Steiglitz and McBride solution was improved upon by Schultz [1968] by applying quasilinearization to the model-plant error case.

Lendaris [1962] and Weygandt and Puri [1966] present relatively complex methods for finding the coefficients of the plant's transfer function by using Z-transforms. Solutions of sets of algebraic equations and roots of polynomials are required to do this. Because a differencing scheme on the system output is used in this method, the estimates are likely to be sensitive to noise. Neither of the approaches incorporate any smoothing, but they could be extended to do so. Hoppe [1965] has a related method but incorporates integration for smoothing.

2.2.2 STATISTICALLY ORIENTED METHODS

Cross-correlation and cross-spectral methods can be used to determine the impulse response of linear time invariant systems. By observing system input and output and forming auto- and cross-correlations, the system impulse response can be found from the correlation equation (2.20) assuming all stochastic aspects are stationary and independent. Goodman and Reswick [1956] and Goodman [1957] have already been mentioned as early examples. Later, Levin [1960] demonstrates that a relation exists between optimal least squares estimates of the impulse response and correlation methods, citing the Weiner Hopf equation as the link. He also develops a finite numerical deconvolution scheme which uses a discrete version of the correlation equation

and sample auto- and cross-correlations to yield a least squares multi-point fit to the system impulse response. The effect on optimality of the estimates when short operating records are used is investigated by Kerr [1961]. A comprehensive discussion on correlation techniques may be found in Akaike [1967].

The instrumental variable techniques are useful with linear regression problems (equation error models) in the determination of least squares estimates of the regression coefficients. The instrumental variables Z are defined as an additional set of observations with the following correlation properties (Goldberger [1964]),

$$\text{plim}_{N \rightarrow \infty} \frac{1}{N} Z_N^T V_N(e) = 0 \quad (2.22)$$

$$\text{plim}_{N \rightarrow \infty} \frac{1}{N} Z_N^T V_N(x) = P > 0 \quad (2.23)$$

The instrumental variable estimate $\hat{a}_N$ of the vector a of unknown autoregression coefficients is defined as (Wong and Polak [1967]):

$$\hat{a}_N = (Z_N^T V_N(y))^{-1} Z_N^T u_p \quad (2.24)$$

where: $a^T = (a_1, \dots, a_p)$

Z_N = appropriately dimensioned matrix
of instrumental variables

$$V_N(x) = (x_1, \dots, x_p)$$

$$x_r^T = (x_r, \dots, x_{N+r-1})$$

$$u_p^T = (u_p, \dots, u_{N+p-1})$$

$$y_i = x_i + e_i$$

$$a^T x_r = u_{N+r-1}$$

The main advantage of instrumental variables is that they always yield

consistent estimates. Reiersol [1941] is generally credited with being first to use the method. A detailed description of the method can be found in Sargan [1958], and an informative summary is given by Goldberger [1964].

Joseph, et al. [1961] use the input to their linear discrete system as the instrumental variable when applying correlation techniques to establish an unbiased estimator of the Z-transform of the system's plant. Wong and Polak discuss the application of the instrumental variables to estimation of coefficients of a linear autoregression with noisy state measurements, the properties of instrumental variables, and some computationally efficient approximations. Since some freedom exists in the choice of the instrumental variable, Wong and Polak explore the existence of optimal sequences of these variables.

An approach to parameter estimation that commonly occurs is to augment the system state vector with the unknown parameter vector p by introducing additional state equations to describe the dynamics of the parameters. These equations typically have the form,

$$\dot{p} = \alpha p + \beta \quad (2.25)$$

where β is either a deterministic or a random variable but often both α and β are zero.

This formulation even with a linear system leads to a nonlinear arrangement when augmented because some of the terms in the system equations by definition become products of state variables. In this situation the usual approach is to estimate simultaneously the original state variables and the parameters. State vector estimation with a

nonlinear model is known as nonlinear filtering and typically requires solution of multipoint boundary value problems to determine the estimates.

A relatively early work in nonlinear filtering that appears to have become a basic reference for much of the work in the area is Bryson and Frazier [1962]. The model used is basically the general nonlinear one, Equation (2.16), given earlier. The objective is to find the minimizing state variable function $x(t)$ (which for the identification problem would have included the parameters $p(t)$) for the cost function J_t where,

$$J_t = \frac{1}{2} (x(t_0) - \mu)' p_0^{-1} (x(t_0) - \mu) + \frac{1}{2} \int_{t_0}^t (\tilde{\xi}' R^{-1} \tilde{\xi} + \tilde{\eta}' Q^{-1} \tilde{\eta}) dt' \quad (2.26)$$

subject to the constraints :

$$\dot{x} - g(x, u(t), t, \xi) = 0 \quad (2.27)$$

$$h(y, x, \eta) = 0$$

and where $\mu = E(x(t_0))$, $\tilde{\xi} = \xi - E(\xi)$, and $\tilde{\eta} = \eta - E(\eta)$.

The minimizing $x(t)$ is found by the method of steepest descent.

Although Bryson and Frazier did not explicitly concern themselves with parameter estimation, their formulation could have incorporated that task. In fact, most nonlinear estimation schemes could handle parameter estimation. However, for the most part, no papers were selected for the following discussion which did not mention at least some connection with identification. The schemes presented divide into two groups - those that are patterned directly after Bryson and Frazier but use some other modern control theory solution technique and those that depart along the way by linearizing. These papers serve

to illustrate the variety of ways the nonlinear filtering problem can be attacked.

A number of solutions of the first category use quasilinearization instead of steepest descent. Kumar and Sridhar [1964] having measurements at a number of discrete points solve the estimation problem as a multipoint boundary value problem using $\dot{p} = 0$. Lavi and Strauss [1965] show that if the total number of measurements (boundary points) does not exceed the total number of free variables, the solution may not be unique. They use excess measurements and perform a least squares solution as do Kumar [1965] and Lee [1968a].

Detchmندی and Shridhar [1965], on the other hand, treat the problem with noise on input and output by deriving the Hamiltonian, the canonical equations, and then solving by invariant imbedding. Lee [1968b] derives a least squares version of Detchmندی and Shridhar's solution and applies it to a chemical reactor problem. Cox [1964] gives a Bayesian approach with a dynamic programming solution; and a Hamilton-Jacobi route is used by Mortensen [1968].

As a result of the work of Kalman [1960], [1961] which led to the computationally desirable recursive linear state estimation equations, and the expanding role of digital computers, aspirations for Kalman type solution to the nonlinear filtering problem were heightened. However, generally this Kalman filter characteristic can be achieved only by linearizing the nonlinear problems about some nominal (except in very special cases, Farison [1967]). Examples of recursive solutions by linearization are Kopp and Orford [1963] and Budin [1969].

The method known as stochastic approximation is one which has considerable appeal from a computational point of view because of its recursive nature. It resembles the recursive linearized solution to the nonlinear filtering problem in that it too represents a linearization and the gain (or relaxation factor) generally depends on the error covariance. It differs by computing only the unknown parameters and not the entire augmented state vector. The recursion equation has the general form (Balakrishnan and Peterka [1969]),

$$\hat{a}_{N+1} = \hat{a}_N + \rho_N \nabla_a Q(\hat{a}_N) \quad (2.28)$$

where ρ_N is a predetermined relaxation factor and $Q(\cdot)$ is the equation error at the N th stage with a taken as $\hat{a}_N$. The basis for convergence of this technique rests heavily upon the proofs in Dvoretzky [1956].

Ho and Whalen [1963] and Ho and Lee [1965] develop stochastic approximation solutions for the linear discrete model and show convergence of their estimates. Sakrison [1967] treats the continuous time problem with the equation error type model and develops an algorithm for identifying system Laplace transform coefficients. Saridis and Stein [1968] extend the work of Ho and Lee.

In spite of convergence claims of the above and others, Balakrishnan and Peterka state that this method has fallen short of expectations apparently with slow convergence being a major difficulty. Albert and Gardner [1967] give a comprehensive discussion of stochastic approximation.

When appropriate statistics on the parameters to be identified are available, Bayesian estimation techniques can be used. Unfortunately, using this additional information tends to result in estimators of greater complexity than found by other methods that require less

information. Examples of Bayesian identification can be found in Sawaragi and Katayama [1967], Aoki [1967], and Kroy and Stubberud [1967].

In the situation where noise statistics are available, maximum likelihood methods may be employed. (For ML estimates of the usual *a priori* type, parameter statistics are not used.) Frequent claims are made in the literature that particular solutions are maximum likelihood estimates. Often these are least squares estimates made in a situation where maximum likelihood gives an identical estimate, e.g., with gaussian noise. The papers described below are not members of that category.

The problem of Koopmans [1937] is adapted to control systems by Levin [1964]. The model used was a discrete single-input, single-output equation error type with input and output measurement noise. To achieve independence among the measurements as in Koopmans, Levin has to stack his measurements. This means that parameter estimates can be updated only after each new stack has been accumulated. However, using the stacked measurements he arrives at the eigenvalue problem of Koopmans. The estimate is shown to correspond to a least square hyperplane fit to the data. Properties of the estimates and estimation with overlapping stacks of measurements are discussed.

Astrom and Bolin [1965] estimate the coefficients in the shift operators a , b , c , and the scalar λ in the system Z transfer function

$$a^*(Z^{-1})y(t) = b^*(Z^{-1})x(t) + \lambda c^*(Z^{-1})e(t) \quad (2.29)$$

where $a(Z) = 1 + a_1Z + \dots + a_NZ^N$, etc.

and the $\{e(t)\}$ is a sequence of normal independent random variables

To maximize the likelihood function so that the estimates can be found, a Newton-Raphson algorithm was developed which made use of symmetry among the partials in order to reduce the number of computations. Astrom [1967] applies this solution to the control of a paper-making machine.

Smith and Hilton [1965] review the characteristics of the least squares solution of the error equation model and Levin's generalized least squares eigenvector solution. In 1967 they presented the results of a numerical comparison of the two methods. They found that the bias magnitudes of the least squares estimates generally were greater than those of the eigenvector method, but their standard deviations were smaller. Also, using overlapping vectors in the eigenvector method substantially reduced the variance of the estimates.

Rogers and Steiglitz [1967] approach identification in the model-error formulation by passing the output error through a whitening filter whose coefficients are estimated along with those in the model. An approximate Newton-Raphson algorithm is implemented to find the estimates.

Smith [1968] explores the problems of recovering the Laplace transform of the system transfer function after forming a sample data estimate using Levin's eigenvector method. Mayne [1966] presents various on-line algorithms for particular regression problems but finds that in the more general case, the method of Astrom and Bohlin performs best. Kashyap [1970] extends the work of Astrom and Bohlin to include vector state and input variables but without the moving average input. The model includes correlated plant noise and uncorrelated

output measurement noise. He develops algorithms for estimates of the system coefficients as well as the plant noise correlation matrices.

2.3 IDENTIFIABILITY

The problem of under what conditions can a meaningful estimate of a parameter be obtained as well as the whole question of input selection clearly are of interest when developing parameter identification techniques. Identifiability refers to the ability to excite all the modes of a system and being able to observe the results of the excitation. Input selection deals with input signal design to best facilitate identification. (If the identification scheme is restricted to the use of normal operating inputs, then optimal input selection is not of any direct interest.) Astrom and Bohlin [1965], Currie [1968], and Staley [1968] pursue identifiability. Turin [1957], Gagliardi [1967], and Staley deal with input selection.

SECTION III

THE LIKELIHOOD EQUATIONS

3.1 INTRODUCTION

The mathematical development of a maximum likelihood estimator for the identification of unknown parameters in the mathematical model of a system can be viewed as a three step operation once the model is completely defined (except for the unknown parameters). Employing the usual terminology, e.g., see Cramér [1966], first the likelihood function L must be determined. The likelihood function is the density function of the measurements considered as a function of the unknown parameter. It can be viewed as a family of probability density functions $p(y, \theta)$ on the samples (or measurements) $y_1, \dots, y_n$ of the system output indexed by the unknown parameter vector θ which lies in some set Θ .

In most of the literature on maximum likelihood estimation, the samples are assumed to be independent and identically distributed resulting in:

$$L = p(y_1, \dots, y_n, \theta) = p(y_1, \theta) p(y_2, \theta) \dots p(y_n, \theta) \quad (3.1)$$

However, in the following discussions not all these conditions are present, and less restrictive definitions will be required. For independent but not identically distributed samples:

$$L = p_1(y_1, \theta) p_2(y_2, \theta) \dots p_n(y_n, \theta) \quad (3.2)$$

and for samples which are neither independent nor identically distributed:

$$L = p(y_1, y_2, \dots, y_n, \theta) \quad (3.3)$$

The unknown parameters θ in all but one of the cases to be

considered are either the scalar or matrix coefficients of the difference equations in the models investigated. In that exception, θ includes the initial conditions as well as the coefficients, but explicit estimation of the initial conditions can be eliminated from the estimator.

The second step is to find the necessary condition for the maximization of the likelihood function over θ for a set of measurements $y_1, \dots, y_n$. The necessary condition is known as the likelihood equation when defined as:

$$\frac{d \log L}{d\theta} = 0 \quad (3.4)$$

Of course, the logarithm and the derivative must exist, and the maximizing $\theta \in \Theta$ must not be a boundary point of the set Θ . (Because gaussian densities and $\theta \in R$, the real line, will be assumed, these conditions will be satisfied.) Unfortunately, when θ is in fact a matrix, finding the likelihood equation may not be straightforward if only matrix operations are used.

As will be shown, the likelihood equations for the situations treated here can quite naturally be expressed as finite polynomials or, more precisely, as sums of finite polynomials in the unknown parameters. The desirability of expressing the necessary conditions in that form was based on the fact that there exists an extensive body of knowledge on the properties of roots of polynomials. On the other hand, there are well-known problems associated with the numerical solution of roots of polynomials which must be faced.

The third step is the solution of the likelihood equation for the appropriate roots. Normally, this can only be done numerically.

Since sufficient conditions for the maximization of the likelihood function will not be sought, it will be assumed that other means will be available to determine which real root is the desired one if multiple real roots exist. (The solution of the equations is the topic of Chapter 5.)

3.2 PROBLEM STATEMENT

The basic mathematical model used for this study is the usual linear discrete control system model but without plant noise. This model can correspond to a local approximation to a more complex nonlinear system. The measurement noise is treated as additive and assumed to be white on the basis that in practice its bandwidth is frequently found to be much wider than that of the plant.

The model is represented as follows:

$$\begin{aligned} \text{Plant: } x_{i+1} &= Ax_i + Bu_i \\ \text{Measurement: } y_i &= Hx_i + \eta_i \quad i = 0, 1, \dots, N \end{aligned} \quad (3.5)$$

where:

x_i = n -dimensional state vector

y_i = m -dimensional measurement vector

u_i = r -dimensional input vector

η_i = m -dimensional measurement noise vector

with $\eta_i \sim \mathcal{N}(0, R)$, $E[\eta_i \eta_j^T] = R \delta_{ij}$, $R > 0$

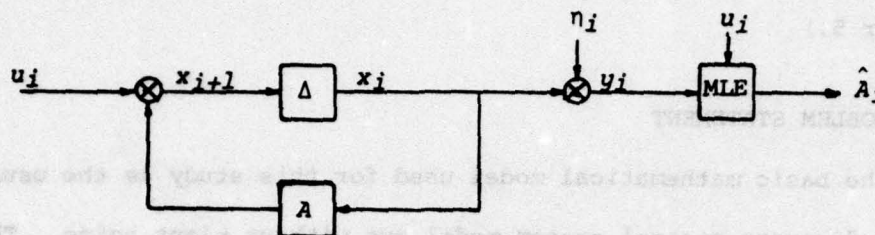
A = $n \times n$ constant matrix

B = $n \times r$ constant matrix

H = $m \times n$ constant matrix

The following are known: the dimension of all the quantities, the matrices B and H , and at time t_N , the variables u_i and y_i ,

$i = 0, 1, \dots, N$. The unknowns are the matrix A and the variables x_i and η_i , $i = 0, 1, \dots, N$. (The statistical properties of η_i are assumed to be known.) The objective is to estimate A using the output measurement $\{y_i\}$ and the input sequence $\{u_i\}$. (See Figure 3-1.)



Basic Model

Figure 3-1

The matrix A is taken to be general except in certain analyses, when noted, it is assumed to be stable. The model is assumed to be observable (in the deterministic sense), i.e., the $n \times mn$ matrix $[H^T, A^T H^T, \dots, (A^T)^{n-1} H^T]$ has rank n . Further restrictions such as canonical forms*, e.g., see Lee[1964], are not considered here.

The type of estimator of A that is sought is one which is based on maximum likelihood but operates in real time, requires a fixed and minimal amount of data storage and computation, and can function with only normal operating input. Separation** of identification from other

* Though going to an equivalent companion matrix system, for example, reduces the number of parameters to be identified, it is not clear that attempting to recover the A matrix from an equivalent system will be any less difficult than directly estimating A . While in some situations knowledge of the equivalent system might be sufficient, availability of the A matrix for filtering and control is often required.

** A separation theorem for the filtering aspect of the control problem does exist, Joseph and Tou [1961], but, apparently, with the exception of only a few special cases, Horowitz and Grammatikos [1970], one has not been found for identification, Sawaragi and Katayama [1967].

aspects, in terms of the total control problem, is assumed to be allowed.

The nature of the *a priori* information on the initial condition x_0 significantly affects the form of the estimator of λ . Three possibilities are examined: x_0 known, x_0 unknown parameter, and x_0 unknown random variable with known gaussian distribution. A fourth case which is related to the first two but based on a different approach is also discussed. In the terminology established in Chapter 2, the first three cases are similar to the model-plant error formulation while the fourth is similar to the equation error formulation.

3.3 THE NECESSARY CONDITION - x_0 KNOWN

In this case the initial condition x_0 is assumed to be known. The maximum likelihood (ML) estimate corresponds to finding a "best" fit of the solution of the model equations to the system output measurements. The likelihood equation for the scalar model will be developed first, followed by the vector-matrix case.

For the analysis of the scalar problem, notational changes are made in the model as given in Equation (3.5). The scalar model is written as:

$$\begin{aligned} x_{i+1} &= ax_i + bu_i \\ y_i &= hx_i + \eta_i \quad i = 0, 1, \dots, N \end{aligned} \quad (3.6)$$

where: $\eta_i \sim \mathcal{N}(0, \sigma^2)$

Since the η_i are gaussian and independent, the distribution of y_i follows directly from (3.6) as:

$$y_i \sim \mathcal{N}(ha^i x_0 + hb \sum_{k=1}^i a^{i-k} u_{k-1}, \sigma^2) \quad i = 1, 2, \dots, N \quad (3.7)$$

The maximum likelihood estimate of a , denoted by $\hat{a}$, is the one which maximizes $p(y_1, \dots, y_N; a)$ or, equivalently, the one which minimizes:

$$Q_N(a) = \sum_{i=1}^N (y_i - a^i h x_0 - h b \sum_{k=1}^i a^{i-k} u_{k-1})^2 \quad (3.8)$$

As expected, the estimate will be a form of least squares. It is independent of the noise variance and the first sample y_0 (since x_0 is known).

Forming $\frac{dQ_N(a)}{da} = 0$ and rearranging terms give the necessary condition as:

$$\begin{aligned} & x_0 \sum_{i=1}^N i a^{i-1} y_i + b \sum_{i=1}^N \sum_{j=1}^i (i-j) a^{i-j-1} u_{j-1} y_i \\ & - h x_0^2 \sum_{i=1}^N i a^{2i-1} - h b x_0 \sum_{i=1}^N \sum_{j=1}^i (2i-j) a^{2i-j-1} u_{j-1} \\ & - h b^2 \sum_{i=1}^N \sum_{j=1}^i \sum_{k=1}^i (i-k) a^{2i-j-k-1} u_{j-1} u_{k-1} = 0 \end{aligned} \quad (3.9)$$

where $\hat{a}$ is the appropriate value of " a " which satisfies (3.9). For the autonomous case ($u_i \equiv 0$), the necessary condition reduces to:

$$\sum_{i=1}^N i a^{i-1} y_i - h x_0 \sum_{i=1}^N i a^{2i-1} = 0 \quad (3.10)$$

When A is an $n \times n$ matrix, the density of y_i is:

$$y_i \sim \mathcal{N}(H A^i x_0 + H \sum_{j=1}^i A^{i-j} B u_{j-1}, R) \quad i = 1, 2, \dots, N \quad (3.11)$$

and the cost function for the maximum likelihood estimate becomes:

$$Q_N(A) = \sum_{i=1}^N \left\| y_i - H A^i x_0 - H \sum_{j=1}^i A^{i-j} B u_{j-1} \right\|_{R^{-1}}^2 \quad (3.12)$$

Taking differentials at a stationary point gives:

$$\Delta Q_N(A) = 2 \sum_{i=1}^N [(y_i - HA^i x_0 - H \sum_{j=1}^i A^{i-j} B u_{j-1})^T R^{-1} (-H(\Delta A^i) x_0 - H \sum_{k=1}^i (\Delta A^{i-k}) B u_{k-1})] = 0 \quad (3.13)$$

where

$$\Delta A^p = \sum_{r=1}^p A^{p-1} (\Delta A) A^{p-r}$$

By rewriting (3.13) as an equation in the traces of matrices and using the commutivity of matrices under the trace operation, an equation in the trace of the product of two matrices of the following form can be derived:

$$Q_N(a) = \text{tr} [(D)(\Delta A)] = 0 \quad (3.14)$$

Since the differential matrix ΔA is arbitrary, then D equals zero, and the necessary condition can be expressed as:

$$\begin{aligned} D = & \sum_{i=1}^N \sum_{p=1}^i A^{i-p} x_0 y_i^T R^{-1} H A^{p-1} \\ & - \sum_{i=1}^N \sum_{p=1}^i A^{i-p} x_0 x_0^T (A^T)^i H^T R^{-1} H A^{p-1} \\ & - \sum_{i=1}^N \sum_{j=1}^i \sum_{p=1}^i A^{i-p} u_{j-1}^T B^T (A^T)^{i-j} H^T R^{-1} H A^{p-1} \\ & + \sum_{i=2}^N \sum_{k=1}^{i-1} \sum_{p=1}^{i-k} A^{i-k-p} B u_{k-1} y_i^T R^{-1} H A^{p-1} \\ & - \sum_{i=2}^N \sum_{k=1}^{i-1} \sum_{p=1}^{i-k} A^{i-k-p} B u_{k-1} x_0^T (A^T)^i H^T R^{-1} H A^{p-1} \\ & - \sum_{i=2}^N \sum_{j=1}^i \sum_{k=1}^{i-1} \sum_{p=1}^{i-k} A^{i-k-p} B u_{k-1} u_{j-1}^T B^T (A^T)^{i-j} H^T R^{-1} H A^{p-1} = 0 \quad (3.15) \end{aligned}$$

where $\hat{A}$ is the appropriate value of A which satisfies (3.15).

If the model were autonomous, i.e., $u_1 \equiv 0$, then (3.15) undergoes considerable simplification:

$$\sum_{i=1}^N \sum_{p=1}^i A^{i-p} x_0 (y_i - HA^i x_0)^T R^{-1} H A^{p-1} = 0 \quad (3.16)$$

(Note that if x_0 is zero in the autonomous case, then the root $\hat{A}$ of (3.16) is indeterminate.)

3.4 THE NECESSARY CONDITION - x_0 UNKNOWN PARAMETER

Now the initial condition x_0 is taken to be an unknown parameter of the system in the same sense as the coefficient A . The geometric interpretation of the ML estimate is basically the same as the x_0 known case with the exception that the initial point is free to participate in the optimization in the present case. As with x_0 known, but accounting for the free initial condition, the distribution of the i th measurement for the scalar model is:

$$y_i \sim \mathcal{N}(ha^i x_0 + hb \sum_{k=0}^{i-1} a^k u_{i-k-1}, \sigma^2) \quad i = 1, 2, \dots, N \quad (3.17)$$

$$y_0 \sim \mathcal{N}(hx_0, \sigma^2)$$

Once again, since the measurements are independent and gaussian, the maximum likelihood estimate is a least squares estimate, i.e., minimize:

$$Q_N(a, x_0) = \sum_{i=1}^N (y_i - ha^i x_0 - hb \sum_{k=1}^i a^{i-k} u_{k-1})^2 + (y_0 - hx_0)^2 \quad (3.18)$$

The necessary conditions on a and x_0 become:

$$\begin{aligned} \frac{\partial Q_N}{\partial x_0} &= 2 \sum_{i=1}^N (y_i - ha^i x_0 - hb \sum_{k=1}^i a^{i-k} u_{k-1}) (-ha^i) + 2(y_0 - hx_0) (-h) \\ &= 0 \end{aligned} \quad (3.19)$$

or:

$$\hat{x}_{0N} = \left[\sum_{i=0}^N y_i a^i - hb \sum_{i=1}^N \sum_{j=1}^i a^{2i-j} u_{j-1} \right] / \left(h \sum_{i=0}^N a^{2i} \right) \quad (3.20)$$

Forming $\frac{\partial Q_N}{\partial a}$ leads to the same condition as obtained with x_0

known, i.e.,

$$\begin{aligned} x_0 \sum_{i=1}^N i a^{i-1} y_i + b \sum_{i=1}^N \sum_{j=1}^i (i-j) a^{i-j-1} u_{j-1} y_i \\ - h x_0^2 \sum_{i=1}^N i a^{2i-1} - h b x_0 \sum_{i=1}^N \sum_{j=1}^i (2i-j) a^{2i-j-1} u_{j-1} \\ - h b^2 \sum_{i=1}^N \sum_{j=1}^i \sum_{k=1}^i (i-k) a^{2i-j-k-1} u_{j-1} u_{k-1} = 0 \end{aligned} \quad (3.21)$$

(Note that the noise variance does not appear in the necessary conditions in this case nor when x_0 is known.)

Introduction of (3.20) into (3.21) gives an expression for the stationary points of Q_N for the parameter a as a function of the measurements:

$$\begin{aligned} & \left\{ \sum_{t=0}^N \sum_{i=0}^N \sum_{j=0}^N (i-t) a^{2t+i+j-1} y_i y_j \right. \\ & - \sum_{s=0}^N \sum_{t=0}^N \sum_{i=1}^N \sum_{j=1}^i (t-2s+2i-j) a^{2(i+s)+t-j-1} y_t u_{j-1} h b \\ & + h b \sum_{s=0}^N \sum_{t=0}^N \sum_{i=1}^N \sum_{j=1}^i (i-j) a^{2(s+t)+i-j-1} u_{j-1} y_i \\ & + h^2 b^2 \sum_{i=0}^N \sum_{p=1}^N \sum_{r=1}^N \sum_{s=1}^p \sum_{t=1}^r (2p-s-i) a^{2(i+p+r)-s-t-1} u_{s-1} u_{t-1} \\ & \left. - h^2 b^2 \sum_{i=0}^N \sum_{p=0}^N \sum_{r=1}^N \sum_{s=1}^r \sum_{t=1}^r (r-t) a^{2(i+p+r)-s-t-1} u_{s-1} u_{t-1} \right\} \\ & = 0 \end{aligned} \quad (3.22)$$

For the autonomous case, (3.20) reduces to:

$$\hat{x}_{0N} = \left(\sum_{i=0}^N y_i a^i \right) / \left(h \sum_{i=0}^N a^{2i} \right) \quad (3.23)$$

and (3.22), assuming $\hat{x}_{0N}$ not equal to zero, reduces to:

$$\sum_{i=0}^N \sum_{j=0}^N (j-i) a^{2i+j-1} y_j = 0 \quad (3.24)$$

For the vector-matrix case only the autonomous version is treated because it yields considerably simpler equations than the nonautonomous one yet illustrates all the basic steps required to derive the latter. The density of the output measurements y_i for the autonomous model is:

$$y_i \sim \mathcal{N}(HA^i x_0, R) \quad i = 0, 1, \dots, N \quad (3.25)$$

and since the $\{y_i\}$ are independent, the cost function for the maximum likelihood estimate becomes:

$$Q_N(A, x_0) = \sum_{i=0}^N \|y_i - HA^i x_0\|_{R^{-1}}^2 \quad (3.26)$$

Taking the differentials of x_0 at a stationary point of Q_N gives:

$$\Delta_{x_0} [Q_N(A, x_0)] = 2 \sum_{i=0}^N [-HA^i (\Delta x_0)]^T R^{-1} [y_i - HA^i x_0] = 0 \quad (3.27)$$

or,
$$\text{tr} \left[\sum_{i=0}^N R^{-1} (y_i - HA^i x_0) (\Delta x_0)^T (A^T)^i H^T \right] = 0$$

or,
$$\sum_{i=0}^N (A^T)^i H^T R^{-1} y_i - \sum_{i=0}^N (A^T)^i H^T R^{-1} HA^i x_0 = 0 \quad (3.28)$$

Let
$$\phi_N \triangleq \sum_{i=0}^N (A^T)^i H^T R^{-1} HA^i \quad (3.29)$$

Since all models were assumed to be observable and R^{-1} is symmetric and positive definite, ϕ_N^{-1} exists for $N \geq n$. Then $\hat{x}_{0N}$ can be expressed

as:

$$\hat{x}_{0N} = \phi_N^{-1} \sum_{i=0}^N (A^T)^i H^T R^{-1} y_i \quad (3.30)$$

Taking differentials of $\hat{A}$ at a stationary point of Q_N gives:

$$\Delta_A [Q_N(A, x_0)] = 2 \sum_{i=0}^N [-H(\Delta A^i) x_0]^T R^{-1} [y_i - H A^i x_0] = 0 \quad (3.31)$$

or
$$\text{tr} \left[\sum_{i=1}^N \sum_{p=1}^i R^{-1} (y_i - H A^i x_0) x_0^T (A^T)^{i-p} (\Delta A)^T (A^T)^{p-1} H^T \right] = 0$$

or
$$\sum_{i=1}^N \sum_{p=1}^i (A^T)^{p-1} H^T R^{-1} y_i x_0^T (A^T)^{i-p} - \sum_{i=1}^N \sum_{p=1}^i (A^T)^{p-1} H^T R^{-1} H A^i x_0 x_0^T (A^T)^{i-p} = 0 \quad (3.32)$$

Introducing (3.30) into (3.32) with $x_0 = \hat{x}_{0N}$ gives the necessary condition for $\hat{A}$ as:

$$\begin{aligned} & \sum_{i=1}^N \sum_{j=0}^N \sum_{p=1}^i (A^T)^{p-1} H^T R^{-1} y_i y_j^T R^{-1} H A^j \phi_N^{-1} (A^T)^{i-p} \\ & - \sum_{i=1}^N \sum_{j=0}^N \sum_{k=0}^N \sum_{p=1}^i (A^T)^{p-1} H^T R^{-1} H A^i \phi_N^{-1} (A^T)^j H^T R^{-1} y_i y_k^T R^{-1} H A^k \phi_N^{-1} (A^T)^{i-p} \\ & = 0 \end{aligned} \quad (3.33)$$

3.5 THE NECESSARY CONDITION - DISTRIBUTION OF x_0 KNOWN

In this section the initial condition x_0 is assumed to be an unknown random variable whose density is known. Again the nature of the estimator, roughly speaking, is to seek an $\hat{A}$ which results in some best fit of model output to measured output. While this case is intermediate to the previous two cases with respect to the amount of information on x_0 assumed available, the polynomial form of the likelihood equations is substantially more difficult to obtain.

For the scalar case, take x_0 distributed as:

$$x_0 \sim \eta(\bar{x}_0, \epsilon^2) \quad (3.34)$$

and x_0 independent of $\{\eta_i\}$. Because the random variables $x_0, \eta_0, \eta_1, \dots, \eta_N$ are independent and gaussian, and their joint distribution is:

$$p(\eta^*) = (2\pi)^{-\frac{N+2}{2}} |R_1|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\eta^* - \bar{x}^*)^T R_1^{-1} (\eta^* - \bar{x}^*)\right] \quad (3.35)$$

where:

$$\eta^{*T} = (x_0, \eta_0, \eta_1, \dots, \eta_N)$$

$$\bar{x}^{*T} = (\bar{x}_0, 0, \dots, 0)$$

$$R_1 = \text{diag}(\epsilon^2, \sigma^2, \dots, \sigma^2)$$

From the scalar model equations (3.6), the output measurements are seen to be related to the noise and initial conditions as follows:

$$y = D\eta^* + hbu^* \quad (3.36)$$

where:

$$y^T = (y_0, y_1, \dots, y_N)$$

$$D = [h\underline{a} \mid I]$$

$$\underline{a}^T = (1, a, a^2, \dots, a^N)$$

$$u^{*T} = (0, u_0, u_1 + au_0, \dots, u_{N-1} + \dots + a^{N-1}u_0)$$

$$I = \text{identity matrix}$$

By applying the theorem on the linear transformation of jointly gaussian random variables (e.g., see Anderson [1958, p.26]) to (3.35) with transformation (3.36), the measurement density becomes,

$$p(y; a) = (2\pi)^{-\frac{N+1}{2}} |R_2|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} \|y - h\underline{x}_0 \underline{a} - hbu^*\|^2_{R_2^{-1}}\right] \quad (3.37)$$

where:

$$R_2 = DR_1D^T$$

From (3.35), (3.36), and (3.37):

$$R_2 = \sigma^2 I + h^2 \epsilon^2 \underline{a} \underline{a}^T \quad (3.38)$$

where:

$$R_2 = (N+1) \times (N+1) \text{ matrix.}$$

(Note that the y_i are now not independent. Further, (3.37) indicates that the ML estimate will not be a least squares estimate, nor even a Gauss-Markov estimate.)

It can be shown (see Appendix A) that

$$|R_2| = (\sigma^2)^N (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a}) \quad (3.39)$$

$$R_2^{-1} = \frac{1}{\sigma^2} [I - (h^2 \epsilon^2 / (h^2 \epsilon^2 \underline{a}^T \underline{a} + \sigma^2)) \underline{a} \underline{a}^T] \quad (3.40)$$

The cost function can now be expressed in the following form:

$$Q_N(a) = \sigma^2 (\log |R_2| + ||y - h\bar{x}_{0a} - hbu^*||_{R_2^{-1}}^2) \quad (3.41)$$

Equivalently, let the likelihood function L be defined as:

$$L \triangleq p(y;a) \quad (3.42)$$

Since $p(y;a)$ is smooth and greater than zero for all y and a , and the logarithm is monotonic, the stationary points of L satisfy:

$$\begin{aligned} \frac{d \log L}{da} &= -\frac{1}{2} |R_2|^{-1} \frac{d}{da} |R_2| + (y - h\bar{x}_{0a} - hbu^*)^T R_2^{-1} (h\bar{x}_0 \frac{da}{da} + hb \frac{du^*}{da}) \\ &\quad - \frac{1}{2} (y - h\bar{x}_{0a} - hbu^*)^T \frac{dR_2^{-1}}{da} (y - h\bar{x}_{0a} - hbu^*) = 0 \end{aligned} \quad (3.43)$$

Multiplying through (3.43) by $[-\sigma^4 (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a})^2]$ gives:

$$\begin{aligned} &\{\sigma^2 (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a}) h^2 \epsilon^2 \underline{a}^T \underline{a} - (y - h\bar{x}_{0a} - hbu^*)^T [(\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a})^2 I \\ &\quad - (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a}) (h^2 \epsilon^2 \underline{a} \underline{a}^T)] (h\bar{x}_0 \underline{a} + hbu^*) \\ &\quad + (y - h\bar{x}_{0a} - hbu^*)^T [(h^4 \epsilon^4 \underline{a}^T \underline{a}) (\underline{a} \underline{a}^T) - (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a}) (h^2 \epsilon^2) (\underline{a} \underline{a}^T)] \\ &\quad (y - h\bar{x}_{0a} - hbu^*)\} = 0 \end{aligned} \quad (3.44)$$

where:

$$\underline{a}_a = \frac{d}{da} a \text{ and } u_a^* = \frac{d}{da} u^*$$

Dividing through by $h^2 \epsilon^2$, expanding, and regrouping by the various combinations of inner products results in the following necessary condition for $\hat{a}$:

$$\begin{aligned} & \{ [\sigma^2 (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a}) + h^2 \bar{x}_0^2 \Psi \sigma^2] \underline{a}^T \underline{a}_a \\ & + (y^T \underline{a}) [2(\underline{a}^T \underline{a}_a) (h \bar{x}_0) \sigma^2 + (\underline{a}^T u_a^*) h b (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a}) + (y^T \underline{a}) h^2 \epsilon^2 (\underline{a}^T \underline{a}_a) \\ & - 2(\underline{a}^T u_a^*) h b h^2 \epsilon^2 (\underline{a}^T \underline{a}_a) + (\underline{a}^T u_a^*) h b (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a})] \\ & + (y^T \underline{a}_a) [(\underline{a}^T u_a^*) h b (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a}) - (y^T \underline{a}) (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a}) - h \bar{x}_0 \Psi \sigma^2 \\ & - \sigma^2 h \bar{x}_0 \underline{a}^T \underline{a}] \\ & - (y^T u_a^*) h b h^2 \epsilon^2 (\underline{a}^T \underline{a} + \Psi)^2 \\ & + (\underline{a}^T u_a^*) [(\underline{a}^T u_a^*) (\underline{a}^T \underline{a}_a) h^2 b^2 h^2 \epsilon^2 - (\underline{a}^T u_a^*) h^2 b^2 (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a}) \\ & - 2 h \bar{x}_0 h b \sigma^2 (\underline{a}^T \underline{a}_a)] \\ & + (\underline{a}^T u_a^*) [h \bar{x}_0 h b (\Psi \sigma^2 + \sigma^2 \underline{a}^T \underline{a}_a)] \\ & + (\underline{a}^T u_a^*) [h b h \bar{x}_0 \sigma^2 (\underline{a}^T \underline{a} + \Psi) - (\underline{a}^T u_a^*) h^2 b^2 (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a})] \\ & + (u_a^T u_a^*) [h^2 b^2 h^2 \epsilon^2 (\underline{a}^T \underline{a} + \Psi)^2] \} = 0 \end{aligned} \quad (3.45)$$

where:

$$\Psi = \frac{\sigma^2}{h^2 \epsilon^2}$$

Rewriting the above using summations instead of inner products allows many of the terms to combine (though the resulting expression appears less compact and less efficient computationally):

$$\begin{aligned} & \{ [\sigma^2 (\sigma^2 + h^2 \bar{x}_0^2 \Psi) + \sigma^2 h^2 \epsilon^2 (\sum_{i=0}^N a^{2i})] (\sum_{i=0}^N i a^{2i-1}) \\ & + \sigma^2 h \bar{x}_0 [\sum_{i=0}^N \sum_{j=0}^N (2i - j) a^{2i+j-1} y_j] - h \bar{x}_0 \Psi \sigma^2 \sum_{i=0}^N i a^{i-1} y_i \end{aligned}$$

$$\begin{aligned}
& + h b h^2 \epsilon^2 \left[\sum_{i=0}^N \sum_{j=0}^N \sum_{p=1}^N \sum_{r=1}^p (2p - 2i + j - r) a^{2(p+i)-r+j-1} y_j u_{r-1} \right] \\
& + h b \sigma^2 \left[\sum_{i=0}^N \sum_{p=1}^N \sum_{r=1}^p (2p - r + i) a^{2p-r+i-1} y_i u_{r-1} \right] \\
& + h^2 \epsilon^2 \left[\sum_{i=0}^N \sum_{j=0}^N \sum_{k=0}^N (i - k) a^{2i+j+k-1} y_j y_k \right] - \sigma^2 \sum_{i=0}^N \sum_{j=0}^N i a^{i+j-1} y_i y_j \\
& + h \bar{x}_0 h b \sigma^2 \left[\sum_{i=0}^N \sum_{p=1}^N \sum_{r=1}^p (2p - r - 2i) a^{2(p+i)-r-1} u_{r-1} \right] \\
& + h \bar{x}_0 h b \Psi \sigma^2 \left[\sum_{p=1}^N \sum_{r=1}^p (2p - r) a^{2p-r-1} u_{r-1} \right] \\
& - h b h^2 \epsilon^2 \left(\Psi + \sum_{i=0}^N a^{2i} \right)^2 \left(\sum_{p=1}^N \sum_{r=1}^p (p - r) a^{p-r-1} u_{r-1} u_{p-1} \right) \\
& + h^2 b^2 h^2 \epsilon^2 \left[\sum_{i=0}^N \sum_{p=1}^N \sum_{r=1}^p \sum_{s=1}^N \sum_{t=1}^s (i - 2s + t) a^{2(p+i+s)-r-t-1} u_{r-1} u_{t-1} \right] \\
& + h^2 b^2 \sigma^2 \left[\sum_{p=1}^N \sum_{r=1}^p \sum_{s=1}^N \sum_{t=1}^s (t - 2s) a^{2(p+s)-r-t-1} u_{r-1} u_{t-1} \right] \\
& + [h^4 b^2 \epsilon^2 \left(\Psi + \sum_{i=0}^N a^{2i} \right)^2 \left(\sum_{p=1}^N \sum_{r=1}^p \sum_{s=1}^p (p - s) a^{2p-r-s-1} u_{r-1} u_{s-1} \right)] \\
& = 0
\end{aligned} \tag{3.46}$$

For the autonomous case (3.45) reduces to:

$$\begin{aligned}
& \{ [\sigma^2 (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a}) + \Psi \sigma^2 h^2 \bar{x}_0^2] \underline{a}^T \underline{a} + 2 \sigma^2 h \bar{x}_0 (\underline{a}^T \underline{a}) (y^T \underline{a}) \\
& - [\Psi \sigma^2 h \bar{x}_0 + \sigma^2 h \bar{x}_0 (\underline{a}^T \underline{a})] (y^T \underline{a}_a) + (h^2 \epsilon^2 \underline{a}^T \underline{a}_a) (y^T \underline{a})^2 \\
& - (\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a}) (y^T \underline{a}) (y^T \underline{a}_a) \} = 0
\end{aligned} \tag{3.47}$$

and (3.46) reduces to:

$$\begin{aligned}
& \{ \sigma^2 (\sigma^2 + h^2 \bar{x}_0^2 \psi) \sum_{i=0}^N i a^{2i-1} + \sigma^2 h^2 \epsilon^2 \sum_{i=0}^N \sum_{j=0}^N i a^{2(i+j)-1} \\
& + \sigma^2 h^2 \bar{x}_0 \sum_{i=0}^N \sum_{j=0}^N (2i-j) a^{2i+j-1} y_j - \sigma^2 \sum_{i=0}^N \sum_{j=0}^N j a^{i+j-1} y_i y_j \\
& + h^2 \epsilon^2 \sum_{i=0}^N \sum_{j=0}^N \sum_{k=0}^N (i-k) a^{2i+j+k-1} y_j y_k - \sigma^2 \psi h \bar{x}_0 \sum_{i=0}^N i a^{i-1} y_i \} \\
& = 0
\end{aligned} \tag{3.48}$$

The situation where A is an unknown $n \times n$ matrix presents some difficulties. Arriving at a density for the vector measurements analogous to (3.37) for the scalar case is straightforward enough. Unfortunately, no useful expressions for the determinant and the inverse of the covariance could be found. Since this precludes reducing the likelihood equation to the desired polynomial form at this time, the matrix case will not be developed here.

3.6 THE NECESSARY CONDITION - DIFFERENCE EQUATION ERROR APPROACH

The general class of parameter identification problems treated in the previous three sections could have been approached somewhat differently. Instead of seeking the parameter values which gave a best fit of the model output to the measured output, parameter values could be selected to minimize the error that results when system input and output measurements are introduced into the model equation.

To implement the latter approach, referred to as the "differencing approach" in the following discussions, the model Equations (3.5) (or (3.6) in the scalar case) must be rearranged. The assumption that H^{-1} exists is made to facilitate this (which, of course, means that now $m = n$.) Working with Equation (3.5), the rearranged model is

$$y_{i+1} = Hx_{i+1} + \eta_{i+1} \quad (3.49)$$

$$= HAx_i + HBu_i + \eta_{i+1}$$

or

$$y_{i+1} = Cy_i + HBu_i + \zeta_i \quad i = 0, 1, \dots, N \quad (3.50)$$

where:

$$C = HAH^{-1}$$

$$\zeta_i = \eta_{i+1} - C\eta_i$$

(In the scalar case, C becomes a and HB becomes hb .)

The equivalent system model, Equation (3.50), has two important differences relative to the original formulation of the model. In this new system, the state variables y_i in the difference equation are known as opposed to the x_i in the previous system model which were unknown. In addition, the noise variable, though still zero mean, is now correlated and acts as part of the input.

Just as was demonstrated with the original formulation of the model, the form of the ML estimator based on the equivalent model strongly depends upon the assumptions made about the initial conditions. In fact, for each of the three initial condition situations treated earlier, the equivalent model leads to the same likelihood equations and thus the same estimators as found previously - a none too surprising result if the models are one-to-one. To see this, look at the case where x_0 is an unknown parameter. From Equations (3.50) and (3.5), where with no loss of generality take $u_i \equiv 0$:

$$y_0 = Hx_0 + \eta_0$$

$$y_1 = Cy_0 + \zeta_0$$

$$\vdots$$

$$y_N = Cy_{N-1} + \zeta_{N-1} \quad (3.51)$$

Using Equation (3.50):

$$\zeta^* \sim \mathcal{N}(0, R^*) \quad (3.52)$$

where:

$$\zeta^{*\top} = (\eta_0, \zeta_0, \zeta_1, \dots, \zeta_{N-1}) \quad (3.53)$$

and

$$R^* = \begin{pmatrix} R & -RC^\top & & \\ -CR & R+CRC^\top & -RC^\top & \\ & \ddots & \ddots & \\ & & -CR & R+CRC^\top \end{pmatrix} \quad (3.54)$$

$$= \begin{pmatrix} I & & \\ & O & \\ -C & & \\ & \ddots & \\ O & -C & I \end{pmatrix} \begin{pmatrix} R & & \\ & O & \\ & & R \end{pmatrix} \begin{pmatrix} I & -C^\top & \\ & O & \\ & & I \end{pmatrix} \quad (3.55)$$

From Equation (3.51), the Jacobian J is:

$$J = \left| \frac{\partial y}{\partial \zeta^*} \right| = 1 \quad (3.56)$$

Equation (3.51) can be rewritten as:

$$\begin{pmatrix} I & & \\ & O & \\ -C & & \\ & \ddots & \\ O & -C & I \end{pmatrix} \begin{pmatrix} y_0 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{pmatrix} = \begin{pmatrix} Hx_0 + \eta_0 \\ \zeta_0 \\ \cdot \\ \cdot \\ \zeta_{N-1} \end{pmatrix} \quad (3.57)$$

Using Equations (3.57), (3.56), and (3.54) along with (3.51)

leads directly to the likelihood function (3.26) except for the mean

of the joint distribution as given in Equation (3.25). However, recognizing that:

$$\begin{bmatrix} I & & & \\ -C & 0 & & \\ & & \ddots & \\ 0 & & & -C & I \end{bmatrix}^{-1} = \begin{bmatrix} I & & & \\ C & 0 & & \\ \vdots & & \ddots & \\ C^N & \dots & & C & I \end{bmatrix} \quad (3.58)$$

and

$$C^i = (HAH^{-1})^i = HA^iH^{-1} \quad (3.59)$$

the mean is easily established, and the equivalence is shown.

There exists another interesting assumption on the initial conditions for the alternate model, Equation (3.50), that can be made. In this case, the initial condition is considered to be y_0 and is assumed known, as it obviously is since it is a measurement, and a deterministic constant. While this assumption can be applied to the model of Equation (3.50), it is inconsistent with the underlying model of Equation (3.5) which says that y_0 is a random variable. It would appear that this discrepancy has the effect of assigning an improper weight to the first error - an effect which could be expected to have diminishing influence as the number of samples increases. Treating y_0 as a known deterministic constant in the alternate model gives good results experimentally for systems that correspond to the original model. Investigation of this formulation, irrespective of whether or not it directly applies to the original problem, is of interest. (See Mann and Wald [1943] and Levin [1964] and the related discussion in Chapter 2.)

The likelihood function for the alternate model taking y_0 as a known constant is found as follows. The basic set of equations for this problem is generated by (3.50) indexed by $i = 0, 1, \dots, N-1$. For this system, the joint distribution of $\zeta_0, \zeta_1, \dots, \zeta_{N-1}$ is easily shown to be:

$$\zeta \sim \eta(0, R_z) \quad (3.60)$$

where:

$$\zeta^* = (\zeta_0, \zeta_1, \dots, \zeta_{N-1})$$

and

$$R_z = \begin{bmatrix} R+CR C^T & -RC^T & & & \\ -CR & R+CR C^T & -RC^T & & \\ & \ddots & \ddots & \ddots & \\ 0 & & & -RC^T & \\ & & & -CR & R+CR C^T \end{bmatrix} \quad (3.61)$$

Since the Jacobian J_1 for Equation (3.50) equals one where:

$$J_1 = \left| \frac{\partial \tilde{y}}{\partial \zeta} \right| \quad (3.62)$$

and

$$\tilde{y}^* = (y_1, y_2, \dots, y_N) \quad (3.63)$$

then:

$$L_z = (2\pi)^{-\frac{nN}{2}} |R_z|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\tilde{y}^* - f)^T R_z^{-1} (\tilde{y}^* - f)] \quad (3.64)$$

where:

$y_i = n$ -vector,

$$f = \begin{bmatrix} HBu_0 \\ . \\ . \\ . \\ HBu_{N-1} \end{bmatrix}, \quad \tilde{y}^* = [(y_1 - Cy_0), \dots, (y_N - Cy_{N-1})] \quad (3.65)$$

The characteristics of the likelihood function L_z are considerably

different from those of the previous sections. The means of those densities in all three cases were expressed in terms of powers of the unknown matrix A . In L_z , the mean is not a function of A , but A does enter linearly in the set of differences of the y_i 's. On the other hand, the covariances of the previous densities were of the form $\sigma^2 I$ or $I \otimes R$, where $\otimes$ denotes the Kronecker product, while in L_z the covariance has tri-diagonal form with elements which are of the form A or $1 + A^2$.

Because R_z is not diagonal (in the sense that the earlier covariances of the form $I \otimes R$ were diagonal), crudely speaking, the estimate for A is found by fitting a hyperplane to the measurements in the non-trivial norm of R_z^{-1} . This characteristic is interesting relative to the solutions of the previous sections. There the residuals were weighted equally. That this can be undesirable is easily seen by considering the scalar autonomous model with $|a| < 1$. Then early measurements have more useful information than later ones, and thus the later ones should be weighted less than the early ones. The likelihood function L_z will give uneven weighting to the residuals. Whether or not this will yield any better estimates than those of the previous sections is not immediately clear since the nature of the residuals is different, and the arrangement of the weighting is not obvious.

Because of the complexity of this approach, first the scalar autonomous case will be developed. The covariance matrix R_z reduces to the $N \times N$ matrix:

$$Q = \sigma^2 \begin{pmatrix} 1+a^2 & -a & 0 \\ -a & 1+a^2 & -a \\ 0 & -a & 1+a^2 \end{pmatrix} \quad (3.66)$$

It can be shown (see Appendix B) that:

$$|Q| = (\sigma^2)^N \sum_{i=0}^N a^{2i} \quad (3.67)$$

and

$$Q^{-1} = (r_{ij}^{-1}) \quad (3.68)$$

where:

$$r_{ij}^{-1} = \left(\sum_{k=0}^{i-1} a^{2k} \right) a^{j-i} \left(\sum_{t=0}^{N-j} a^{2t} \right) / \left(\sum_{p=0}^N a^{2p} \right) \quad (j \geq i)$$

Also, in Equation (3.64), for the scalar autonomous case:

$$n = 1, f = 0, \text{ and } C = a \quad (3.69)$$

Forming $\frac{d \log L_z}{da} = 0$ gives after considerable manipulation (see

Appendix C):

$$V_N + Q_N' = 0 \quad (3.70)$$

where:

$$V_N = -2\sigma^2 \sum_{p=0}^N \sum_{q=0}^N q a^{2(p+q)-1} \quad (3.71)$$

$$= -2\sigma^2 \sum_{k=1}^{2N} \left[\sum_{i=0}^k i + \sum_{\substack{j=k-N \\ k > N}}^N j \right] a^{2k-1}$$

$$Q_N' = \sum_{k=0}^{2(2N-1)} [y^T \left(\sum_{i=0}^N (k+1-4i) S_{k+1-N-2i} J \right) y] a^k \quad (3.72)$$

where:

$$J = \begin{pmatrix} & & & 1 \\ & 0 & & \\ & & \ddots & \\ 1 & & & 0 \end{pmatrix}, \text{ the unit Hankel matrix}$$

$$S_p = \begin{cases} (S_1)^p & p > 0 \\ I & p = 0 \\ (S_1^T)^p & p < 0 \end{cases}$$

$$S_1 = \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 1 & \\ & & & \ddots \\ 1 & & & & 0 \end{pmatrix}$$

As a result of differencing the measurements, it would appear that information about the initial condition is lost, and consequently, this case is similar to one where x_0 was an unknown parameter. In fact, Q_N' above is equivalent to the first term of the necessary condition for x_0 unknown parameter case (Equation 3.22), i.e.,

$$Q_N' = 2 \sum_{t=0}^N \sum_{i=0}^N \sum_{j=0}^N (i-t) a^{2t+i+j-1} y_i y_j \quad (3.73)$$

For the scalar plant with a forcing function, a necessary condition in polynomial form can also be found. However, introduction of the forcing function destroys much of the symmetry characteristic of the autonomous case and as a result precludes the extensive reduction in complexity the autonomous expression for Q_N' can undergo (see Appendix C). Because of the similarity of this case to the x_0 unknown parameter case, Q_N' is probably equivalent to the necessary condition for the latter case (Equation 3.22), but this equivalence has not yet been shown.

The necessary conditions for the vector-matrix case are somewhat more difficult to establish. First take the log of L_z :

$$\log L_z(A) = -\frac{nN}{2} \log 2\pi - \frac{1}{2} \log |R_z| - \frac{1}{2} (\tilde{y}^* - f)^T R_z^{-1} (\tilde{y}^* - f) \quad (3.74)$$

Taking differentials at a stationary point:

$$\begin{aligned} \Delta(\log L_z(A)) &= -\frac{1}{2} \text{tr}[R_z^{-1} \Delta(R_z)] - \Delta(\tilde{y}^*)^T R_z^{-1} (\tilde{y}^* - f) \\ &\quad - \frac{1}{2} (\tilde{y}^* - f)^T \Delta(R_z^{-1}) (\tilde{y}^* - f) = 0 \end{aligned} \quad (3.75)$$

where:

$$\Delta(\tilde{y}^*)^T = - (y_0, \dots, y_{N-1}) (I \otimes \Delta^T) = - \tilde{y}_0^T (I \otimes \Delta^T)$$

$$\tilde{y}_0^T = - (y_0^T, \dots, y_{N-1}^T)$$

$$\Delta(R_z) = \begin{pmatrix} \Delta R A^T + A R \Delta^T & -R \Delta^T & 0 \\ -\Delta R & & \\ 0 & & -R \Delta^T \\ & -\Delta R & \Delta R A^T + A R \Delta^T \end{pmatrix}$$

$$\Delta(R_z^{-1}) = -R_z^{-1} (\Delta(R_z)) R_z^{-1}$$

But $\Delta(R_z)$ can be rewritten as:

$$\Delta(R_z) = \begin{pmatrix} \Delta R & & 0 \\ & & \\ 0 & & \Delta R \end{pmatrix} \begin{pmatrix} A^T & & 0 \\ -I & & \\ 0 & & -I \end{pmatrix} + \begin{pmatrix} A & -I & 0 \\ & & -I \\ 0 & & A \end{pmatrix} \begin{pmatrix} R \Delta^T & & 0 \\ & & \\ 0 & & R \Delta^T \end{pmatrix} \quad (3.76)$$

or

$$\Delta(R_z) = (I_N \otimes \Delta^T) [(I_N \otimes A^T) - S_n] + [(I_N \otimes A^T) - S_n^T] (I_N \otimes \Delta^T) \quad (3.77)$$

where:

The likelihood equations with any nontrivial form for Δ would

$I_N = N \times N$ identity matrix

Using the above and the properties of the trace, Equation (3.74) can be written as:

$$\begin{aligned} \text{tr}\{[(I_N \otimes A^* - S_N)R_Z^{-1} - (I \otimes R^{-1})\tilde{y}_0(\tilde{y}^* - f)^*R_Z^{-1} \\ - (I_N \otimes A^* - S_N)R_Z^{-1}(\tilde{y}^* - f)(\tilde{y}^* - f)^*R_Z^{-1}] \\ (I_N \otimes \Delta')\} = 0 \end{aligned} \quad (3.78)$$

Definition: Let $\mathcal{J}$ be an $(nN) \times (nN)$ matrix partitioned into N^2 equal submatrices of size $n \times n$. Using the same scheme as associated with a matrix having scalar elements, denote the ij th submatrix of $\mathcal{J}$ by G_{ij} . Define the generalized trace operation:

$$\overline{\text{TR}}(\mathcal{J}) = \sum_{i=1}^N G_{ii} \quad (3.79)$$

Since Δ' is an arbitrary matrix:

$$\begin{aligned} \overline{\text{TR}}\{[(I_N \otimes A^* - S_N) - (I \otimes R^{-1})\tilde{y}_0(\tilde{y}^* - f)^* \\ - (I_N \otimes A^* - S_N)R_Z^{-1}(\tilde{y}^* - f)(\tilde{y}^* - f)^*R_Z^{-1}]\} = 0 \end{aligned} \quad (3.80)$$

The usefulness of the necessary condition might be enhanced if an explicit form for R^{-1} could be found. However, without some restrictive assumptions on the structure of the matrices, finding an explicit inverse appears to be difficult. For example, when $R = \sigma^2 I$ and A is normal, i.e., $AA^* = A^*A$, then the inverse can be found. However, restrictions on the structure of the A matrix invalidate the likelihood equations whose derivations are based on completely general variations of A at the stationary points. It is not clear that rederivation of

the likelihood equations with any such canonical form for A would result in any benefits.

Notice that to derive the equivalent system model (Equation (3.50)), H^{-1} had to exist. In the more general case, H is taken as an $m \times n$ matrix with $m \leq n$. If $m < n$, the most obvious approach is to use a pseudo-inverse form for H which is suited for this problem.

Another approach that appears promising is use of the observer of Luenberger [1964]. With this scheme, the measurement equation can be augmented so that it can be inverted directly for plant state x_i . There are a number of difficulties associated with this technique, not the least of which is the necessity for the solution of a Lyapunov equation for the matrix required to augment the H matrix.

3.7 PLANT NOISE

Consideration of plant noise introduces additional complexities in the task of finding polynomial type likelihood functions. Typically, the effect of plant noise is to add a term to the covariance matrix for the system without the plant noise. Finding the inverse of the covariance becomes the problem of finding the resolvent of a matrix much as already occurred in a simpler form in the case of x_0 with known distribution.

The scalar versions of the four cases covered in the previous sections will be considered. The basic model (3.6) with plant noise becomes:

$$\begin{aligned} x_{i+1} &= ax_i + bu_i + \xi_i \\ y_i &= hx_i + \eta_i \end{aligned} \tag{3.81}$$

The following assumptions are made on the properties of the noise:

$\{\xi_i\}$ independent and $\{\xi_i\} \sim \mathcal{N}(0, \beta^2)$

$\{\eta_i\}$ independent and independent of $\{\xi_i\}$; $\eta_i \sim \mathcal{N}(0, \sigma^2)$

For convenience, the likelihood equations will be presented in a pre-polynomial form. The polynomials may be found merely by expanding the equations. (The derivations of the likelihood equations are given in Appendix D).

a. Known initial condition x_0

The likelihood equation is:

$$0 = |R_y|^{-1} \left(\frac{d}{da} |R_y| \right) + \{ (y_N - hx_0 a_N - hb \phi u_N)^T \left(\frac{d}{da} R_y^{-1} \right) - 2[hx_0 \left(\frac{d}{da} a_N \right) + hb \left(\frac{d}{da} \phi \right) u_N]^T R_y^{-1} \} (y_N - hx_0 a_N - hb \phi u_N) \quad (3.82)$$

where:

$$y_N^T = (y_1, \dots, y_N) \quad (3.83)$$

$$a_N^T = (a, a^2, \dots, a^N) \quad (3.84)$$

$$u_N^T = (u_0, \dots, u_{N-1}) \quad (3.85)$$

$$\phi = \begin{pmatrix} 1 & & & & & \\ & a & & & & \\ & & a^2 & & & \\ & & & \ddots & & \\ & & & & a^2 & a \\ & & & & & 1 \end{pmatrix} \quad (3.86)$$

$$J_p = (\sigma^2)^p \prod_{k=1}^p \left(\frac{h^2 \beta^2}{\sigma^2} + 1 + a^2 - 2a \cos \frac{k\pi}{p+1} \right), p \geq 1 \quad (3.87)$$

$$\text{and } J_0 = 1$$

$$L_t = (\sigma^2 + h^2 \beta^2) J_{t-1} - a^2 \sigma^4 J_{t-2}, t \geq 3 \quad (3.88)$$

and

$$L_0 = 1$$

$$L_1 = \sigma^2 + h^2 \beta^2$$

$$L_2 = \sigma^4 + \sigma^2 \beta^2 h^2 (2 + a^2) + \beta^4 h^4$$

$$|R_y| = L_N \quad (3.89)$$

$$R_y^{-1} = (\Phi^T)^{-1} R^{-1} \Phi^{-1} \quad (3.90)$$

$$\Phi^{-1} = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ -a & & \ddots & \\ & 0 & & \ddots & \\ & & -a & & 1 \end{pmatrix} \quad (3.91)$$

$$R^{-1} = (r_{ij}^{-1}) \quad (3.92)$$

$$r_{ij}^{-1} = (\sigma^2 a)^{j-i} (L_{i-1}) (J_{N-j}) / (L_N) \quad , j \geq 1 \quad (3.93)$$

b. Initial condition x_0 an unknown parameter

This case is identical to the one in part (a) above but with the addition of an equation for $\hat{x}_{0N}$. Referring to the definitions above and Equation (3.19):

$$\frac{\partial L}{\partial x_0} = (\underline{y}_N - h x_0 \underline{a}_N - h b \Phi \underline{u}_N)^T R_y^{-1} (\underline{a}_N) \sigma^2 + (y_0 - h x_0) = 0 \quad (3.94)$$

or:

$$\hat{x}_{0N} = [(\underline{y}_N - h b \Phi \underline{u}_N)^T R_y^{-1} \underline{a}_N \sigma^2 - y_0] / [(1 + \sigma^2 \underline{a}_N^T R_y^{-1} \underline{a}_N) h] \quad (3.95)$$

c. Known distribution of initial condition x_0

Assume:

$$x_0 \text{ independent of } \{\xi_i\} \text{ and } \{\eta_i\}; x_0 \sim \mathcal{N}(x_0, \epsilon^2)$$

The form of the solution is similar to that where x_0 is known.

The likelihood equation is:

$$0 = |\tilde{R}_y|^{-1} \left(\frac{d}{da} |\tilde{R}_y| \right) + \{(\tilde{\underline{y}}_N - h \tilde{x}_0 \tilde{\underline{a}}_N - h b \Phi \tilde{\underline{u}}_N)^T \left(\frac{d}{da} \tilde{R}_y^{-1} \right)\}$$

$$- 2[h\bar{x}_0 \left(\frac{d}{da} \bar{a}_N\right) + hb \left(\frac{d}{da} \phi_0\right) u_N]^T \bar{R}_y^{-1} \} (\bar{y}_N - h\bar{x}_0 \bar{a}_N - hb \phi_0 u_N) \quad (3.96)$$

where:

$$\bar{y}_N^T = (y_0, y_1, y_2, \dots, y_N) \quad (3.97)$$

$$\bar{a}_N^T = (1, a, a^2, \dots, a^N) \quad (3.98)$$

$$\phi_0 = \begin{pmatrix} 0 & \dots & 0 \\ & \phi & \end{pmatrix} \quad (3.99)$$

$$\tilde{J}_p = \prod_{k=1}^p [\beta^2 + \sigma^2(1 + a^2) - 2a\sigma^2 \cos \frac{\pi k}{p+1}] \quad (3.100)$$

and $\tilde{J}_0 = 1$

(Note that the number of samples is $N + 1$.)

$$\begin{aligned} \tilde{L}_t &= [(\sigma^2 + h^2 \epsilon^2)(\beta^2 + \sigma^2) + \sigma^2 h^2 \epsilon^2 a^2] \tilde{J}_{t-2} \\ &\quad - a^2 \sigma^4 \tilde{J}_{t-3}, \quad t \geq 4 \end{aligned} \quad (3.101)$$

and

$$\begin{aligned} \tilde{L}_0 &= 1 \\ \tilde{L}_1 &= \sigma^2 + h^2 \epsilon^2 \\ \tilde{L}_2 &= (\sigma^2 + h^2 \epsilon^2)(\beta^2 + \sigma^2) + h^2 \epsilon^2 a^2 \sigma^2 \\ \tilde{L}_3 &= (\sigma^2 + h^2 \epsilon^2)[(\beta^2 + \sigma^2)^2 + \sigma^2 \beta^2 a^2] \\ &\quad + \sigma^2 h^2 \epsilon^2 a^2 [\beta^2 + \sigma^2(1 + a^2)] \end{aligned}$$

$$|\tilde{R}_y| = \tilde{L}_{N+1} \quad (3.102)$$

$$\tilde{R}_y^{-1} = (\phi_1^T)^{-1} \tilde{R}^{-1} \phi_1^{-1} \quad (3.103)$$

$$\phi_1^{-1} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \phi^{-1} & \\ 0 & & & \end{pmatrix} \quad (3.104)$$

$$\tilde{R}^{-1} = (\tilde{r}_{ij}^{-1}) \quad (3.105)$$

$$\bar{r}_{ij}^{-1} = (\sigma^2 a)^{j-1} (\bar{L}_{i-1}) (K_{N+1-j}) / (\bar{L}_{N+1}), j \geq 1 \quad (3.106)$$

$$K_{N+1-j} = \begin{cases} (h^2 \epsilon^2 a^2 + \beta^2 + \sigma^2) \bar{J}_{N-1} - a^2 \sigma^4 \bar{J}_{N-2}, & j=1 \\ \bar{J}_{N+1-j}, & 2 \leq j \leq N+1 \end{cases} \quad (3.107)$$

d. Differencing approach

The likelihood equation is:

$$0 = |\bar{R}_y|^{-1} \left(\frac{d}{da} |\bar{R}_y| \right) + \{ (\underline{y}_N - a \underline{y}_{N-1} - h b \underline{u}_N)^T \left(\frac{d}{da} \bar{R}_y^{-1} \right) - 2 (\underline{y}_{N-1})^T \bar{R}_y^{-1} \} (\underline{y}_N - a \underline{y}_{N-1} - h b \underline{u}_N) \quad (3.108)$$

where:

$$(\underline{y}_{N-1})^T = (y_0, y_1, \dots, y_{N-1})$$

$$\bar{J}_p = \prod_{k=1}^p [h^2 \beta^2 + \sigma^2 (1 + a^2) - 2a\sigma^2 \cos \frac{\pi k}{p+1}], p \geq 1$$

and

$$\bar{J}_0 = 1 \quad (3.110)$$

$$|\bar{R}_y| = \bar{J}_N \quad (3.111)$$

$$\bar{R}_y^{-1} = (\bar{r}_{ij}^{-1}) \quad (3.112)$$

$$\bar{r}_{ij}^{-1} = (\sigma^2 a)^{j-1} (\bar{J}_{i-1}) (\bar{J}_{N-j}) / \bar{J}_N, j \geq 1 \quad (3.113)$$

3.8 MINIMAL SUFFICIENT STATISTICS

When the likelihood equations are derived, the question of what might be their simplest form inevitably arises. Concern for simplicity is heightened as the number of samples increases because computational effort can rapidly increase with more samples.

The problem can be viewed from two aspects. One is purely algebraic manipulation to reduce complexity. Any approach in this area is basically *ad hoc*. The second, which in a sense is a special case

of the first, is concerned with condensing the information in the set of samples into a smaller set which contains an equivalent amount of information about the unknown parameter. This second area, which is based on the theory of sufficient statistics, has formal structure and is the more important of the two because through sufficient statistics, the amount of computation can be stabilized as the number of samples increases. Establishing the existence of (non-trivial) sufficient statistics for the four cases investigated in this chapter is clearly of interest.

For scalar variables, Dynkin [1951], on whom most of the following discussion will be based, defines a sufficient statistic in the following way. Let $\{p(x, \theta) : \theta \in \Theta\}$ be a family of probability densities denoted by Γ , defined on the set D in the m -dimensional space R^m . The function $\chi(x)$, defined in D and with values in some set T , is called a sufficient statistic in the domain D for the family Γ , if the probability densities $p(x, \theta)$ may be factored into the form:

$$p(x, \theta) = v[\chi(x), \theta]w(x) \quad (x \in D, \theta \in \Theta) \quad (3.114)$$

(For a more rigorous definition see Rao [1965, p.110].)

Then, for example, if the samples $x_1, \dots, x_N$ are independent and identically distributed, the existence of a sufficient statistic χ would allow the following factorization*:

$$\prod_{i=1}^N p(x_i, \theta) = v[\chi(x_1, \dots, x_N), \theta]w(x_1, \dots, x_N) \quad (3.115)$$

(Of course, from a computational point of view what is desired is that

* The statistic $\chi = (x_1, \dots, x_N)$ is sufficient and is known as the trivial statistic.

χ have a recursive form such as:

$$\chi(x_1, \dots, x_N) = F[\chi(x_1, \dots, x_{N-1}), x_N] \quad (3.116)$$

where F is some reasonable function. Otherwise nothing is likely to be gained.)

If the sufficient statistic for a family Γ is not unique, then the question of which one is most desirable arises. Since sufficient statistics in a sense partition the sample space, a possible characterization of the most desirable one is that it impose the coarsest partition on the sample space. Pursuing the approach more formally, Dynkin says let $\chi_1(x)$ and $\chi_2(x)$ be defined in D . Then χ_1 is dependent on χ_2 if it follows that $\chi_2(x') = \chi_2(x'')$ implies $\chi_1(x') = \chi_1(x'')$. This gives a partial ordering among the sufficient statistics. The statistic $\chi(x)$ is called a necessary statistic for the family Γ in the domain D if it is dependent on every sufficient statistic. A statistic which is both necessary and sufficient is minimal sufficient.

In order to test for minimal sufficient statistics two theorems of Dynkin, as corrected by Brown [1964], for scalar, independent identically distributed samples are useful. The first theorem (Theorem 2) considers the linear space $L(\Gamma, D)$ generated by constants and the functions $g_x(\theta)$ for any $\theta \in \Theta$ where

$$g_x(\theta) = \log p(x, \theta) - \log p(x, \theta_0) \quad (3.117)$$

and θ_0 some fixed element in Θ . If the functions $1, \phi_1(x), \dots, \phi_r(x)$ are a basis in $L(\Gamma, D)$, then for $N \geq r$ the system of functions:

$$\chi_i(x_1, x_2, \dots, x_N) = \phi_i(x_1) + \dots + \phi_i(x_N) \quad i = 1, \dots, r \quad (3.118)$$

is shown to be a minimal sufficient statistic for the sample of size N .

In the second theorem (Theorem 3a) Dynkin shows that if the probability density of the sample has the form

$$p(x, \theta) = \exp \left\{ \sum_{i=1}^r C_i(\theta) \phi_i(x) + C_0(\theta) + \phi_0(x) \right\} \quad (3.119)$$

then the ϕ_i here correspond to those of the previous theorem (Theorem 2) and thus form a minimal sufficient statistic when summed as in (3.118).

However, Dynkin's results do not apply directly to the four initial condition situations in the previous sections because identically distributed samples are assumed for the above two theorems. The case of independent samples which are not necessarily identically distributed is treated by Zhuravlev [1963]. A theorem based on the above theorems of Dynkin is presented which results in the desired generalization. In this case each $p_j(x, \theta)$ of the form given in (3.119) has associated with it the sets of functions $\{C_i(\theta)\}_j$ and $\{\phi_i(x)\}_j$ where $1 \leq j \leq k \leq N$, the number of samples. The minimal sufficient statistics are found by forming linear combinations of various sets $\{\phi_i\}_j$ after examining the amount of dependency in spaces generated by all possible combinations of the sets $\{C_i\}_j$. Non-trivial minimal sufficient statistics result only if the dimension of the space generated by $\{C_i\}_1, \{C_i\}_2, \dots, \{C_i\}_k$ is less than k .

Applying Zhuravlev to the known x_0 and x_0 unknown parameter cases shows no non-trivial sufficient statistics exist. In the x_0 known case the i th sample is distributed as:

$$p_i(y_i, a) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left[-\frac{1}{2\sigma^2} (y_i - a^i x_0 - \sum_{j=1}^i a^{i-j} u_{j-1})^2 \right] \quad (3.120)$$

The $\{C_k\}$ for the i th sample are $a^i, a^{i-1}, \dots, a, 1$. Clearly, for $i = 1, \dots, N$ the dimension of $\{C_k\}_1, \dots, \{C_k\}_N$ is not less than N . The

same conclusion holds when x_0 is an unknown parameter. For the other two cases, none of the above theorems apply because the samples are neither identically distributed nor independent. By factoring the covariance matrix and transforming the original samples with the factors, a transformed set of samples which are independent can be found. Nothing is gained by this approach because the transformed samples are unknown quantities since they depend on the parameter a . The existence of non-trivial sufficient statistics for these cases is unlikely because of the sample dependence. (The literature offers little for investigation of the vector sample versions of the four cases. Some related work was done by Barndorff-Nielsen and Pedersen [1968].) The conclusion from all of this is that the degree of the polynomials which define the necessary conditions for the ML estimate will increase without bound as the number of samples increases without bound.

SECTION IV

PROPERTIES OF THE IDENTIFIERS AND THEIR APPROXIMATIONS

4.1 INTRODUCTION

The properties of the parameter estimate $\hat{a}$ and, in particular, the characteristics of the roots of the likelihood equations for each of the four cases treated in Chapter 3 need to be considered in order that some evaluation of the practicality of these estimates can be made. Both the degree to which the ML estimate $\hat{a}$ can be expected to approximate the true value of a and the extent of the effort required to determine $\hat{a}$ are of interest. (Discussion of the latter is primarily the topic of the next chapter.)

The investigation of $\hat{a}$ is developed in two parts - finite sample properties and large sample or asymptotic properties. To this end, a number of questions could be posed such as the bias, consistency, efficiency, asymptotic distribution, and uniqueness of $\hat{a}$ as well as the number of real roots, if any, of the likelihood equation and their sensitivity to the measurements. While furnishing answers to these questions, as well as related ones, might be desirable, in general this tends to be difficult to accomplish. Some of these questions plus possible approximations to the ML estimate are explored below, but for simplicity, generally only the scalar versions of the four cases in Chapter 3 are investigated.

4.2 FINITE SAMPLE CHARACTERISTICS

When the number of samples is finite, purely deterministic analysis of the likelihood equations is of only minimal value. Short of

forming confidence limits for the properties of interest, limiting forms and averaging appear most advantageous for finite sample analysis.

For this analysis, let a_0 and x_{00} be the true values of a and x_0 , the parameters to be identified, $\hat{a}$ and $\hat{x}_0$ be their ML estimates, as previously, and a and x_0 be points in some subset of the real line. Then a measurement y_i can be expressed as:

$$y_i = ha_0^i x_{00} + hb \sum_{j=1}^i a_0^{i-j} u_{j-1} + \eta_i \quad (i = 1, 2, \dots, N) \quad (4.1)$$

except for the differencing model in which case y_i becomes:

$$y_i = a_0^i y_0 + \sum_{j=1}^i [a_0^{i-j} (hbu_{j-1} + \zeta_{j-1})] \quad (i = 1, 2, \dots, N) \quad (4.2)$$

4.2.1 LIMITING ESTIMATE FOR ZERO NOISE

One of the simplest questions to answer about the likelihood equation is what happens to $\hat{a}$ and $\hat{x}_0$ as σ^2 or, equivalently, as the noise measurement goes to zero. When the initial condition x_{00} is known, introducing (4.1) with $\eta_i = 0$ into (3.9) with the above notational changes gives:

$$\begin{aligned} D_N(a) = & h x_{00}^2 \sum_{i=1}^N i a_0^i a^{i-1} - h x_{00}^2 \sum_{i=1}^N i a^{2i-1} \\ & + h b x_{00} \sum_{i=1}^N \sum_{j=1}^i i a_0^{i-j} a^{i-1} u_{j-1} \\ & + h b x_{00} \sum_{i=1}^N \sum_{j=1}^i (i-j) a_0^i a^{i-j-1} u_{j-1} \\ & - h b x_{00} \sum_{i=1}^N \sum_{j=1}^i (2i-j) a^{2i-j-1} u_{j-1} \end{aligned}$$

$$\begin{aligned}
& + hb^2 \sum_{i=1}^N \sum_{j=1}^i \sum_{k=1}^i (i-j) a_0^{i-k} a^{i-j-1} u_{j-1} u_{k-1} \\
& - hb^2 \sum_{i=1}^N \sum_{j=1}^i \sum_{k=1}^i (i-k) a^{2i-j-k} u_{j-1} u_{k-1}
\end{aligned} \tag{4.3}$$

Clearly, $D_N(a_0) = 0$, but a_0 may not be the only root. However, if $\eta_i = 0$, then from Equation (3.8) $Q_N(a_0) = 0$ and $\hat{a} = a_0$ (uniquely, unless $x_{00} = 0$ and $u_i \equiv 0$). (Of course, to show the above, the $Q_N(a)$ equation could have been appealed to directly, but then the details of what happens in $D_N(a)$ as $\sigma^2 \rightarrow 0$ would not have been illustrated.)

When x_0 is an unknown parameter, the limiting condition can be found by introducing (4.1) with $\eta_i = 0$ into (3.22). Equivalently, working with (3.20):

$$\begin{aligned}
\hat{x}_0 = & (h \sum_{i=0}^N a_0^i a^i x_{00} + hb \sum_{i=1}^N \sum_{j=1}^i a_0^{i-j} a^i u_{j-1} \\
& - hb \sum_{i=1}^N \sum_{j=1}^i a^{2i-j} u_{j-1}) / (h \sum_{i=0}^N a^{2i})
\end{aligned} \tag{4.4}$$

Setting a equal to a_0 gives $\hat{x}_0 = x_{00}$.

From the discussion on the x_0 known case and the above result, a_0 is seen to be a root of the likelihood equation for the x_0 unknown case. Thus $\hat{a} = a_0$.

In the case where x_0 is an unknown random variable with known gaussian distribution, introduce (4.1) into (3.46) and let η_i go to zero. (To be consistent in taking the limit, σ^2 must also go to zero.) This gives:

$$D_N(a) = h^4 \epsilon^2 x_{00}^2 \sum_{i=0}^N \sum_{j=0}^N \sum_{k=0}^N (i-k) a^{2i+j+k-1} a_0^{j+k}$$

$$\begin{aligned}
& + h^4 \epsilon^2 b x_{00} \left[\sum_{i=0}^N \sum_{j=0}^N \sum_{p=1}^N \sum_{r=1}^p (2p-2i+j-r) a^{2(p+i)+j-r-1} a_0^j u_{r-1} \right. \\
& + \sum_{i=0}^N \sum_{j=0}^N \sum_{p=1}^N \sum_{r=1}^p (2i-j-p) a^{2i+j+p-1} a_0^{j+p-r} u_{r-1} \\
& - \sum_{i=0}^N \sum_{j=0}^N \sum_{p=1}^N \sum_{r=1}^p (p-r) a^{2(i+j)+p-r-1} a_0^p u_{r-1} \left. \right] \\
& + h^4 b^2 \epsilon^2 \left[\sum_{i=0}^N \sum_{j=1}^N \sum_{p=1}^N \sum_{r=1}^j \sum_{s=1}^p (2p-2i+j-s) a^{2(i+p)+j-s-1} a_0^{j-r} u_{r-1} u_{s-1} \right. \\
& + \sum_{i=0}^N \sum_{j=1}^N \sum_{p=1}^N \sum_{r=1}^j \sum_{s=1}^p (i-p) a^{2i+j+p-1} a_0^{j+p-r-s} u_{r-1} u_{s-1} \\
& + \sum_{i=0}^N \sum_{j=1}^N \sum_{p=1}^N \sum_{r=1}^j \sum_{s=1}^p (i-2p+s) a^{2(i+j+p)-r-s-1} u_{r-1} u_{s-1} \\
& - \sum_{i=0}^N \sum_{j=0}^N \sum_{p=1}^N \sum_{r=1}^p \sum_{s=1}^p (p-r) a^{2(i+j)+p-r-1} a_0^{p-s} u_{r-1} u_{s-1} \\
& + \sum_{i=0}^N \sum_{j=0}^N \sum_{p=1}^N \sum_{r=1}^p \sum_{s=1}^p (p-s) a^{2(i+j+p)-r-s-1} u_{r-1} u_{s-1} \left. \right] \quad (4.5)
\end{aligned}$$

By symmetry, $D_N(a_0) = 0$.

Note that the covariance inverse (3.40) is positive definite for $\sigma^2 > 0$ and $\sigma^2 \underline{a}^T R_Z^{-1} \underline{a} \rightarrow 0$ as $\sigma^2 \rightarrow 0$. Consider the cost function $Q_N(a)$, Equation (3.41). Then, if $\sigma^2 = \eta_i = 0$, $Q_N(a)$ is a minimum when $a = a_0$. Thus, provided $\sigma^2 = 0$, once again, as $\eta_i \rightarrow 0$, $\hat{a} \rightarrow a_0$.

The same conclusion holds for the differencing approach. As in the previous case, σ^2 must be set to zero. Once this is done, the equations are identical to those in the x_0 unknown parameter case for which the limit has been shown.

4.2.2 ESTIMATES FROM AVERAGED LIKELIHOOD EQUATIONS

Another finite sample property is the parameter value which on the average is a root of the likelihood equation. In other words, if L is the likelihood function and

$$r(a) = \frac{d}{da} \log L(a) \quad (4.6)$$

then which a , if any, results in the expectation $E(r) = 0$. (This property is obviously closely related to the one for zero noise.)

Let $y \in Y$ be the N samples from one of the four models previously considered, where the samples could be vector quantities and the system either autonomous or not. The parameter vector $\theta = (\theta_1, \dots, \theta_p)$ is taken as appropriate for the model, e.g., (a) or (a, x_0) or the elements of the A matrix, etc. Let θ_0 and $\hat{\theta}$ be the true value and estimate of θ , respectively, where $\theta_0, \hat{\theta} \in \mathbb{M}$. The likelihood function L is defined as previously, $L = p(y, \theta)$, and let

$$r_i(\theta) = \frac{\partial}{\partial \theta_i} \log L \quad (4.7)$$

Assume $\log L$ and r_i , $i = 1, \dots, p$ are continuous on $Y \otimes \mathbb{M}$ and that $|\log L|$ and $|r_i|$ are bounded $\forall y \in Y, \theta \in \mathbb{M}$ by functions on Y which are integrable over Y . Then the following theorem can be stated:

Theorem 4.1: For N samples, on the average, the true parameter value is the maximum likelihood estimate of the parameter.

Proof:

$$\begin{aligned} E[r_i(\theta)] &= \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial \theta_i} \log p(y, \theta) \right] p(y, \theta_0) dy \\ &= \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial \theta_i} p(y, \theta) \right] \frac{p(y, \theta_0)}{p(y, \theta)} dy \end{aligned} \quad (4.8)$$

If $\theta = \theta_0$,

$$E[r_i(\theta_0)] = \left[\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta_i} p(y, \theta) dy \right] \bigg|_{\theta_0} \quad (4.9)$$

By the above assumptions,

$$E[r_i(\theta_0)] = \left[\frac{\partial}{\partial \theta_i} \int_{-\infty}^{\infty} p(y, \theta) dy \right] \bigg|_{\theta_0} = 0 \quad (4.10)$$

This conclusion unfortunately does not directly answer the question of bias of $\hat{\theta}$. Showing that on the average the true value of θ is a root of the likelihood equation does not necessarily mean that the average of the root $\hat{\theta}$ is θ_0 .

4.2.3 THE STABLE ROOTS OF THE LIKELIHOOD EQUATIONS

The question of the number of zeros of a real-valued polynomial in an interval of the real line is at best difficult to answer in general but when, in addition, the coefficients of the polynomial are random variables, as is the case with the likelihood equations, general statements with much practical value are rare. Kac [1943, 1959] investigated the average number of real roots of a real n th degree polynomial whose coefficients are independent identically distributed normal random variables. The results are in the form of complicated integrals. The conclusions indicate that the density of the root distribution peaks at ± 1 , and the average number of roots within the interval $(-1, 1)$ is the same as the average number outside.

As might well be expected, more may be said about the nature of the roots of the likelihood equations if these equations are investigated directly rather than through the general case considered by Kac. For simplicity, the discussion will be limited to autonomous models.

The analysis of each of the four cases has the same pattern. First, the number of roots of the underlying deterministic portion of those terms in the likelihood equation with random coefficients is investigated in the open interval $(-1,1)$. Then assessments of the effects of the remaining purely deterministic terms, if any, and the purely random segment of the random terms are attempted.

When the initial condition x_{00} is known, the likelihood equation for the autonomous plant is given by (3.10) and (3.6):

$$\begin{aligned} \sum_{i=1}^N ia^{i-1}y_i - hx_{00} \sum_{i=1}^N ia^{2i-1} \\ = hx_{00} \left[\sum_{i=1}^N ia^{i-1}a_0^i - \sum_{i=1}^N ia^{2i-1} \right] + \sum_{i=1}^N ia^{i-1}\eta_i \\ = 0 \end{aligned} \quad (4.11)$$

The plant is assumed to be stable, i.e., $-1 < a_0 < 1$.

The number of roots in the open interval $(-1,1)$ for the deterministic portion of (4.11) (the bracketed quantity) will be treated first. In the interest of clarity, some lemmas are presented prior to the main theorem on the roots.

Lemma 4.1: Let N be a positive integer, and let k be a real number where $0 < k < 1$. Then,

$$\frac{N+1}{N} \frac{k}{1+k^N} < 1 \quad (4.12)$$

Proof:

When $N = 1$,

$$\frac{N+1}{N} \frac{k}{1+k^N} = 2k/(1+k) < 1 \quad (4.13)$$

Let $N \geq 2$. Find $\max_k \frac{k}{1+k^N}$.

$$\frac{d}{dk} \left(\frac{k}{1+k^N} \right) = \frac{1-(N-1)k^N}{(1+k^N)^2} = 0$$

or,

$$k_0 = \left(\frac{1}{N-1} \right)^{1/N} \quad (4.14)$$

Clearly, k_0 is where the maximum in $(0,1)$ occurs. Then,

$$\frac{N+1}{N} \frac{k_0}{1+k_0^N} = \frac{N^2-1}{N^2} \left(\frac{1}{N-1} \right)^{1/N} < 1 \quad (4.15)$$

Lemma 4.2: Let N be a positive integer and let k, a_0 be real numbers where $0 < a_0, k < 1$. Then,

$$N(1 + k^N) - a_0^2(N+1)k(1 - k^{N+1}) > 0 \quad (4.16)$$

Proof:

$$\begin{aligned} & N(1 + k^N) - a_0^2(N+1)k(1 - k^{N+1}) \\ &= N(1 + k^N) \left[1 - a_0^2(1 - k^{N+1}) \frac{N+1}{N} \frac{k}{1+k^N} \right] \\ &> 0 \quad (\text{by Lemma 4.1}) \end{aligned} \quad (4.17)$$

Establishing the sign of the deterministic term of (4.11) when $0 < a_0 < 1$ and $-a_0 < a < 0$ is complicated by the fact that the sign of the i th term of the summations is either positive or negative depending on whether or not i is even or odd. The next lemma deals with this situation.

Lemma 4.3: If N is a positive integer and the real numbers a_0 and a are such that $0 < a_0 < 1$ and $-a_0 < a < 0$, then

$$\sum_{i=1}^N i a^{i-1} a_0^i > \sum_{i=1}^N i a^{2i-1} \quad (4.18)$$

Proof:

Make the following definitions -

$$A_N = \sum_{i=1}^N i a^{i-1} a_0^i \text{ and } B_N = \sum_{i=1}^N i a^{2i-1} \quad (4.19)$$

Inductive proof for N even:

At $N = 2$,

$$\begin{aligned} A_2 - B_2 &= a_0 + 2aa_0^2 - a - 2a^3 \\ &> -2a + 2a(a_0^2 - a^2) \\ &> 0 \end{aligned} \quad (4.20)$$

Assume $A_{N-1} > B_{N-1}$, $N-1$ even. Then,

$$\begin{aligned} A_{N+1} - B_{N+1} &= A_{N-1} - B_{N-1} + Na^{N-1}a_0^N + (N+1)a^Na_0^{N+1} \\ &\quad - Na^{2N-1} - (N+1)a^{2N+1} \\ &> Na^{N-1}a_0^N + (N+1)a^Na_0^{N+1} - Na^{2N-1} \\ &\quad - (N+1)a^{2N+1} \end{aligned} \quad (4.21)$$

$$\text{Let } a = -ka_0, \quad 0 < k < 1 \quad (4.22)$$

Then,

$$\begin{aligned} A_{N+1} - B_{N+1} &> Nk^{N-1}a_0^{2N-1} - (N+1)k^Na_0^{2N+1} + Nk^{2N-1}a_0^{2N-1} \\ &\quad + (N+1)k^{2N+1}a_0^{2N+1} \\ &= k^{N-1}a_0^{2N-1}[N - (N+1)ka_0^2 + Nk^N \\ &\quad + (N+1)k^{N+2}a_0^2] \\ &= k^{N-1}a_0^{2N-1}[N(1+k^N) - (N+1)a_0^2k(1-k^{N+1})] \\ &> 0 \quad (\text{by Lemma 4.2}) \end{aligned} \quad (4.23)$$

Now, odd N :

$$\text{At } N = 1, A_1 - B_1 = a_0 - a > 0 \quad (4.24)$$

Let N be odd, and using the above conclusions for when N was even,

$$\begin{aligned} A_N - B_N &= A_{N-1} - B_{N-1} + Na^{N-1}a_0^N - Na^{2N-1} \\ &> Na^{N-1}a_0^N - Na^{2N-1} \\ &= Nk^{N-1}a_0^{2N-1}(1+k^N) > 0 \end{aligned} \quad (4.25)$$

With the above lemmas, the main theorem follows easily.

Theorem 4.2: Assume the initial condition x_{00} is known and the plant is scalar, stable, and autonomous. Then in the limit as the measurement noise η_i goes to zero, the likelihood equation for the unknown parameter a_0 has only one stable root. Furthermore, that root is a_0 .

Proof:

If h or x_{00} (in Equation (4.11)) is zero, no conclusion on root distribution can be made. Assume h and x_{00} are not zero. Consider the sign of $A_N - B_N$, N a positive integer, for $a \in (-1, 1)$ where

$$A_N = \sum_{i=1}^N i a^{i-1} a_0^i \text{ and } B_N = \sum_{i=1}^N i a^{2i-1}$$

If a_0 is zero, the conclusion is immediate.

Take a_0 positive, i.e., $a_0 \in (0, 1)$. Then,

1. $a_0 < a < 1$

Compare the j th terms of A_N and B_N , $1 \leq j \leq N$

Since $j a^{j-1} a_0^j < j a^{2j-1}$, $A_N < B_N$.

2. $a_0 = a$, then $A_N = B_N$.

3. $0 < a < a_0$

Since $j a^{j-1} a_0^j > j a^{2j-1}$, $A_N > B_N$.

4. $a = 0$

Since $a_0 > 0$, $A_N > B_N$

5. $-a_0 < a < 0$

$A_N > B_N$ by Lemma 4.3

6. $-1 < a \leq -a_0$

$j a^{j-1} a_0^j > j a^{2j-1}$, j odd

$j a^{j-1} a_0^j \geq j a^{2j-1}$, j even

Since summations on j in A_N and B_N go from 1 to N , $A_N > B_N$.

The above follow similarly when $a_0 \in (-1, 0)$.

The effect on the above conclusions due to the polynomial noise term in (4.11), $\sum_{i=1}^N ia^{i-1}\eta_i$, is not very clear as indicated by the earlier discussion of Kac's work. Perhaps the most important characteristic of the term which can easily be determined is its expectation. Since the η_i are zero mean and independent, the expectation is zero, and thus on the average the conclusions in the above theorem hold for the likelihood equation. However, for a given realization, the noise term equals η_1 when $a = 0$. Except for that possible bump in the neighborhood of $a = 0$, simulation results indicate that the polynomial should be smooth on $(-1, 1)$ for finite N . While the number of roots of the noise polynomial is $N-1$, one would not expect all of them to be real, nor all the real ones to be in $(-1, 1)$. For small variance relative to x_{00} and a_0 , the noise term probably only has the effect of biasing the (deterministic) root of the likelihood equation in $(-1, 1)$.

When the initial condition x_{00} is an unknown parameter x_0 , the likelihood equation for the scalar autonomous plant is given by (3.6) and (3.24):

$$\begin{aligned} & \sum_{i=0}^N \sum_{j=0}^N (j-i)a^{2i+j-1}y_j \\ &= hx_{00} \left[\sum_{i=0}^N \sum_{j=0}^N ja^{2i+j-1}a_0^j - \sum_{i=0}^N \sum_{j=0}^N ia^{2i+j-1}a_0^j \right] \\ &+ \sum_{i=0}^N \sum_{j=0}^N (j-i)a^{2i+j-1}\eta_i = 0 \end{aligned} \quad (4.26)$$

The number of roots in the open interval $(-1, 1)$ for the deterministic

portion of (4.26) (the bracketed quantity) is investigated below.

Again, preliminary to the main theorem, some lemmas are presented.

Lemma 4.4: Let z be a real number where $0 < z < 1$ and let n be an integer where $n > 2$. Then,

$$z^n - nz + n - 1 > 0 \quad (\text{Beckenbach and Bellman [1961]}) \quad (4.27)$$

Proof:

$$z^n - nz + n - 1 = (z - 1)(z^{n-1} + z^{n-2} + \dots + z - n + 1) \quad (4.28)$$

$$\text{Since } z^{n-1} + \dots + z < n - 1 \quad (4.29)$$

$$z^n - nz + n - 1 > 0 \quad (4.30)$$

Lemma 4.5: Let a_0 and k be real numbers where $a_0, k \in (0, 1)$, and let M be an integer where $M > 2$. Then,

$$\frac{1-k^3 a_0^4}{k a_0^2} \frac{M-1}{M} \frac{1+k^{M-1}}{1-k^M} + k \frac{M-2}{M} \frac{1-k^{M-2}}{1-k^M} > 1 \quad (4.31)$$

Proof:

$$\begin{aligned} & \frac{1-k^3 a_0^4}{k a_0^2} \frac{M-1}{M} \frac{1+k^{M-1}}{1-k^M} + k \frac{M-2}{M} \frac{1-k^{M-2}}{1-k^M} \\ & > \frac{1-k^3}{k} \frac{M-1}{M} \frac{1+k^{M-1}}{1-k^M} + k \frac{M-2}{M} \frac{1-k^{M-2}}{1-k^M} \end{aligned} \quad (4.32)$$

Placing the right side of (4.32) over a common denominator and subtracting the denominator from the combined numerators gives:

$$(1-k^3)(M-1)(1+k^{M-1}) + k^2(M-2)(1-k^{M-2}) - kM(1-k^M) \quad (4.33)$$

$$\begin{aligned} & = (M-1) - Mk + (M-2)k^2 - (M-1)k^3 + (M-1)k^{M-1} - (M-2)k^M \\ & \quad + Mk^{M+1} - (M-1)k^{M+2} \end{aligned}$$

$$\begin{aligned} & = [(M-1) - Mk + k^{M-1}] + [(M-2)k^2 - (M-1)k^3 + k^{M+1}] \\ & \quad + (M-2)k^{M-1}(1-k) + (M-1)k^{M+1}(1-k) \end{aligned} \quad (4.34)$$

$$> [k^{M-1} - (M-1)k + (M-2)] + 1 - k + k^2[k^{M-1} - (M-1)k + M - 2] \quad (4.35)$$

> 0 (by Lemma 4.4)

$$\text{Also, } kM(1 - k^N) > 0 \quad (4.36)$$

Lemma 4.6: Let a and a_0 be real numbers where $-1 < -a_0 \leq a < 0$.

Let N and j be even integers where $N \geq 2$ and $0 \leq j \leq N-1$.

Then:

$$\begin{aligned} d = & a_0^{j+1} a^{N+2j} [a(N-j-1)(a_0^{N-j-1} - a^{N-j-1}) + (N-j-2)(a_0^{N-j-2} - a^{N-j-2})] \\ & + a_0^j a^{N+2j-2} [a(N-j)(a_0^{N-j} - a^{N-j}) + (N-j-1)(a_0^{N-j-1} - a^{N-j-1})] \\ & > 0 \end{aligned} \quad (4.37)$$

Proof:

Let $a = -ka_0$, $0 < k \leq 1$. Then,

$$\begin{aligned} d = & a^{N+2j-2} a_0^{N-1} [-ka_0^2(N-j)(1-k^{N-j}) + (1-k^3 a_0^4)(N-j-1)(1+k^{N-j-1}) \\ & + k^2 a_0^2(N-j-2)(1-k^{N-j-2})] \end{aligned} \quad (4.38)$$

> 0 when $k = 1$, i.e., $a = -a_0$

For $-a < a < 0$,

$$\begin{aligned} d = & a^{N+2j-2} a_0^{N+1} k(N-j)(1-k^{N-j}) \left[\frac{1-k^3 a_0^4}{ka_0^2} \frac{N-j-1}{N-j} \frac{1+k^{N-j-1}}{1-k^{N-j}} \right. \\ & \left. + k \frac{N-j-2}{N-j} \frac{1-k^{N-j-2}}{1-k^{N-j}} - 1 \right] \end{aligned}$$

> 0 (by Lemma 4.5)

Referring to (4.26), make the following definitions for the subsequent discussions:

$$A_N' \triangleq \sum_{i=0}^N \sum_{j=0}^N j a^{2i+j-1} a_0^j \quad (4.39)$$

$$B_N' \triangleq \sum_{i=0}^N \sum_{j=0}^N i a^{2i+j-1} a_0^j \quad (4.40)$$

$$dA_N' \triangleq A_N' - A_{N-1}' = \sum_{i=0}^N i a^{2N+i-1} a_0^i + \sum_{i=0}^{N-1} N a^{2i+N-1} a_0^N \quad (4.41)$$

$$dB_N' \triangleq B_N' - B_{N-1}' = \sum_{i=0}^N Na^{2N+i-1}a_0^i + \sum_{i=0}^{N-1} ia^{2i+N-1}a_0^N \quad (4.42)$$

Then j th term of $dA_N' - dB_N'$ is:

$$\begin{aligned} & ja^{2N+j-1}a_0^j + Na^{2j+N-1}a_0^N - Na^{2N+j-1}a_0^j - ja^{2j+N-1}a_0^N \\ &= (N-j)a_0^j a^{N+2j-1}(a_0^{N-j} - a^{N-j}) \end{aligned} \quad (4.43)$$

Establishing the sign of the deterministic part of (4.26) when $0 < a_0 < 1$ and $-a_0 < a < 0$ is even more difficult than was the case when x_{00} was known. Now the signs of the terms depend on both N and a summing index. Note that for N even, $dA_N' - dB_N'$ is negative for all j . However, when N is even the j th term of $(dA_N' - dB_N') + (dA_{N-1}' - dB_{N-1}')$, i.e.,

$$a_0^j a^{N+2j-2} [a(N-j)(a_0^{N-j} - a^{N-j}) + (N-j-1)(a_0^{N-j-1} - a^{N-j-1})] \quad (4.44)$$

is positive when j is even, but if j is odd, it can be negative.

The next lemma deals with this problem through showing that for N even, the combined negative terms at N and $N+1$ are dominated by the combined positive terms.

Lemma 4.7: If N is a positive integer and the real numbers a_0 and a are such that $0 < a_0 < 1$ and $-a_0 \leq a < 0$, then,

$$A_N' > B_N' \quad (4.45)$$

Proof:

Inductive proof for N even:

At $N = 2$,

$$A_2' - B_2' = (a_0 - a)[1 + 2a(a_0 + a) + a_0 a^3] \quad (4.46)$$

Let $a = -ka_0$, $0 < k \leq 1$. Then,

$$1 + 2a(a_0 + a) + a_0 a^3 = 1 - 2a_0^2 k(1-k) - a_0^4 k^3$$

$$\begin{aligned}
&> 1 - 3a_0k + 3a_0^2k^2 - a_0^3k^3 \\
&= (1-a_0k)^3 \\
&> 0
\end{aligned} \tag{4.47}$$

Thus, $A_2' - B_2' > 0$

Assume $A_{N-2}' > B_{N-2}'$, $N \geq 4$ and even. From (4.41) and (4.42),

$$A_N' = A_{N-2}' + dA_N' + dA_{N-1}' \tag{4.48}$$

$$B_N' = B_{N-2}' + dB_N' + dB_{N-1}' \tag{4.49}$$

Consider the sum of the j th and $j+1$ st terms of $(dA_N' - dB_N') + (dA_{N-1}' - dB_{N-1}')$ when j is even and $0 \leq j \leq N-2$. From (4.44), this sum becomes:

$$\begin{aligned}
&a_0^{j+1}a^{N+2j}[a(N-j-1)(a_0^{N-j-1}-a^{N-j-1}) + (N-j-2)(a_0^{N-j-2}-a^{N-j-2})] \\
&+ a_0^ja^{N+2j-2}[a(N-j)(a_0^{N-j}-a^{N-j}) + (N-j-1)(a_0^{N-j-1}-a^{N-j-1})] \tag{4.50} \\
&> 0 \quad (\text{by Lemma 4.6})
\end{aligned}$$

Also, when $j = N$, $dA_N' - dB_N' = 0$.

Therefore, for N even and ≥ 4 ,

$$dA_N' + dA_{N-1}' > dB_N' + dB_{N-1}' \tag{4.51}$$

or, for N even and ≥ 2 ,

$$A_N' > B_N' \tag{4.52}$$

Now, take N odd.

At $N = 1$,

$$A_1' - B_1' = a_0 - a > 0 \tag{4.53}$$

For odd $N > 1$, the j th term of $dA_N' - dB_N'$ from (4.43) becomes,

$$(N-j)a_0^ja^{N+2j-1}(a_0^{N-j}-a^{N-j}) \geq 0 \quad \forall j \tag{4.54}$$

For j even, the inequality in (4.54) is strict. Thus, for odd N ,

$$dA_N' - dB_N' > 0 \tag{4.55}$$

From (4.52) and (4.55) for odd $N > 1$,

$$A_N' - B_N' = A_{N-1}' - B_{N-1}' + dA_N' - dB_N' > 0 \quad (4.56)$$

Lemma 4.8: If N is a positive integer and the real numbers a_0 and a are such that $0 < a_0 < 1$ and $-1 < a < -a_0$, then, $A_N' > B_N'$.

Proof:

Inductive proof for N even:

At $N = 2$, from (4.46),

$$A_2' - B_2' = (a_0 - a)[1 + 2a(a_0 + a) + a_0 a^3] \quad (4.57)$$

Let $b = -a$, $a_0 = kb$, $0 < k < 1$. Then

$$\begin{aligned} [1 + 2a(a_0 + a) + a_0 a^3] &= 1 - 2b^2(k-1) - kb^4 \\ &= (1 - kb^4) + 2b^2(1-k) > 0 \end{aligned} \quad (4.58)$$

Or, $A_2' > B_2'$

Assume $A_{N-2}' > B_{N-2}'$, $N \geq 4$ and even.

Consider the sum of the j th and $j+1$ st terms of

$(dA_N' - dB_N') + (dA_{N-1}' - dB_{N-1}')$ where j is even and

$$\begin{aligned} 0 \leq j \leq N-2. \text{ From (4.44) this sum may be expressed as,} \\ a_0^{j+1} a^{N+2j} [a(N-j-1)(a_0^{N-j-1} - a^{N-j-1}) + (N-j-2)(a_0^{N-j-2} - a^{N-j-2})] \\ + a_0^j a^{N+2j-2} [a(N-j)(a_0^{N-j} - a^{N-j}) + (N-j-1)(a_0^{N-j-1} - a^{N-j-1})] \\ = a_0^j a^{N+2j-2} [(1+a^3 a_0)(N-j-1)(a_0^{N-j-1} - a^{N-j-1}) + a(N-j)(a_0^{N-j} - a^{N-j}) \\ + a^2 a_0(N-j-2)(a_0^{N-j-2} - a^{N-j-2})] \\ > a_0^j a^{N+2j-2} [a(N-j)(a_0^{N-j} - a^{N-j}) + a^2 a_0(N-j-2)(a_0^{N-j-2} - a^{N-j-2})] \\ = a_0^j a^{N+2j-2} [b^{N-j+1}(N-j)(1-k^{N-j}) - kb^{N-j+1}(N-j-2)(1-k^{N-j-2})] \\ > 0 \end{aligned} \quad (4.59)$$

When $j = N$, $dA_N' - dB_N' = 0$

Therefore, for N even and ≥ 4 ,

$$dA_N' + dA_{N-1}' > dB_N' + dB_{N-1}' \quad (4.60)$$

Or, for N even and ≥ 2 ,

$$A_N' > B_N'$$

Take N odd.

At $N = 1$, $A_1' > B_1'$ by (4.53)

For odd $N > 1$, consider the j th and $j+1$ st terms of $dA_N' - dB_N'$,

j even:

$$\begin{aligned} & (N-j)a_0^j a^{N+2j-1} (a_0^{N-j} - a^{N-j}) + (N-j-1)a_0^{j+1} a^{N+2j+1} (a_0^{N-j-1} - a^{N-j-1}) \\ & = a_0^j a^{N+2j-1} [(N-j)b^{N-j}(1+k^{N-j}) - k(N-j-1)b^{N-j+2}(1-k^{N-j-1})] \\ & > 0 \end{aligned} \quad (4.61)$$

Thus $A_N' > B_N'$.

Theorem 4.3: Assume the initial condition x_{00} is an unknown parameter and the plant is scalar, stable, and autonomous. Then in the limit as the measurement noise η_i goes to zero, the likelihood equation for the unknown parameter a_0 has only one stable root and that root is a_0 .

Proof:

If h or x_{00} (in Equation (4.26)) is zero, no conclusion on root distribution can be made. Assume that h and x_{00} are not zero.

Consider the sign of $A_N' - B_N'$, N positive integer, for $a \in (-1, 1)$ where:

$$A_N' = \sum_{i=0}^N \sum_{j=0}^N j a^{2i+j-1} a_0^j \quad (4.62)$$

$$B_N' = \sum_{i=0}^N \sum_{j=0}^N i a^{2i+j-1} a_0^j \quad (4.63)$$

If a_0 is zero, the conclusion is immediate. Take $a_0 \in (0, 1)$.

$$1. a_0 < a < 1$$

$$\text{From (4.53), } A_1' - B_1' = a_0 - a < 0 \quad (4.64)$$

From (4.43), for $N > 1$,

$$dA_N' - dB_N' = (N-j)a_0^j a^{N+2j-1} (a_0^{N-j} - a^{N-j}) < 0, 0 \leq j \leq N-1 \quad (4.65)$$

and equals zero when $j = N$. Thus

$$A_N' < B_N'$$

$$2. a_0 = a$$

$$\text{From (4.62) and (4.63), } A_N' - B_N' = 0$$

$$3. 0 < a < a_0$$

$$\text{From (4.64), } A_1' - B_1' > 0$$

From (4.43), for $N > 1$

$$dA_N' - dB_N' = (N-j)a_0^j a^{N+2j-1} (a_0^{N-j} - a^{N-j}) > 0, 0 \leq j \leq N-1 \quad (4.66)$$

and equals zero when $j = N$. Then

$$A_N' > B_N'$$

$$4. a = 0$$

From (4.62), (4.63),

$$A_N' - B_N' = a_0 > 0$$

$$5. -a_0 \leq a < 0$$

By Lemma 4.7, $A_N' > B_N'$

$$6. -1 < a < -a_0$$

By Lemma 4.8, $A_N' > B_N'$

The above follow similarly if $-1 < a_0 < 0$. I

Little more can be said about the random polynomial term of the likelihood equation (4.26) than could be said when the x_{00} known case was discussed. Again, the expectation of the random polynomial is zero.

When the initial condition x_{00} is an unknown random variable, the

likelihood equation for the autonomous plant is given by (3.48):

$$\begin{aligned}
 & \sigma^2 [(\sigma^2 + h^2 \bar{x}_0^2 \psi) \sum_{i=0}^N i a^{2i-1} + h^2 \epsilon^2 \sum_{i=0}^N \sum_{j=0}^N i a^{2(i+j)-1} \\
 & + h^2 \bar{x}_0 \sum_{i=0}^N \sum_{j=0}^N (2i-j) a^{2i+j-1} y_j - \sum_{i=0}^N \sum_{j=0}^N j a^{i+j-1} y_i y_j \\
 & - h \bar{x}_0 \psi \sum_{i=0}^N i a^{i-1} y_i] - (h^2 \epsilon^2 \sum_{i=0}^N \sum_{k=0}^N (k-i) a^{2i+k-1} y_k) \left(\sum_{j=0}^N a^j y_j \right) \\
 & = 0
 \end{aligned} \tag{4.67}$$

The last term of (4.67) corresponds to the x_{00} unknown parameter case likelihood equation (4.26). If in the limit as η_i goes to zero, its variance σ^2 is assumed to go to zero, then the conclusions for the deterministic portion of the x_{00} unknown parameter likelihood equation hold for the deterministic portion of (4.67). However, how the σ^2 terms affect the roots of the likelihood equation when σ^2 is not zero is not clear.

The scalar autonomous differencing approach likelihood equation is given by Equations (3.70), (3.71), (3.73), and (3.50):

$$\begin{aligned}
 & -2\sigma^2 \sum_{i=0}^N \sum_{j=0}^N j a^{2(i+j)-1} + \left(\sum_{i=0}^N \sum_{j=0}^N (j-i) a^{2i+j-1} y_j \right) \left(\sum_{k=0}^N a^k y_k \right) \\
 & = -2\sigma^2 \sum_{i=0}^N \sum_{j=0}^N j a^{2(i+j)-1} \\
 & + \left[\sum_{i=0}^N \sum_{j=0}^N (j-i) a^{2i+j-1} (a_0^j y_0 + \sum_{p=1}^j a_0^{j-p} \zeta_{p-1}) \right] \\
 & \left[\sum_{k=0}^N a^k (a_0^k y_0 + \sum_{p=1}^k a_0^{k-p} \zeta_{p-1}) \right] \\
 & = 0
 \end{aligned} \tag{4.68}$$

where:

$$\sum_{p=1}^r a_0^{r-p} \zeta_{p-1} \triangleq 0$$

when $r = 0$ and $\zeta_i = \eta_{i+1} - a_0 \eta_i$.

If $\sigma^2 \rightarrow 0$ as $\zeta_i \rightarrow 0$, then the conclusions for the deterministic portion of the x_{00} unknown parameter case hold for this case if $y_0 \neq 0$.

When $\sigma^2 \neq 0$ and the random terms are considered, the deterministic conclusions are again obscured. In this case, the expectation of the random terms is not zero. Furthermore, the term $\sum_{k=0}^N a^k y_k$ can have zeros when the random terms are considered. Neglecting the σ^2 term, these zeros of the multiplicative random term become zeros of the likelihood equation in addition to the stable deterministic root however modified by the additive random term.

The σ^2 term can be written as:

$$-2\sigma^2 \sum_{i=0}^N \sum_{j=0}^N j a^{2(i+j)-1} = -2\sigma^2 \left(\sum_{i=0}^N a^{2i} \right) \left(\sum_{j=0}^N j a^{2j-1} \right) \quad (4.69)$$

If $N \geq 1$, this term has its only root at $a = 0$. Its effect for finite N and small σ^2 is to bias the stable deterministic root toward zero. Also, since this term is bounded on $(-1, 1)$, as N increases, the product of N^{-1} and this term diminishes to zero.

The behavior of the roots of the likelihood equations in the interval $-1 < a < 1$ for each of the four scalar cases when the forcing function is not identically zero is less obvious. In an earlier section, the fact that as the measurement noise goes to zero, $\hat{a}$ approaches a_0 was established for forced plants in all cases except the differencing approach (because only the autonomous version was developed here).

For these same cases without the limiting condition on the noise, Equations (3.9), (3.22), and (3.46) indicate by inspection that after $N+1$ samples, the likelihood equations are polynomials of odd degree if some $u_i \neq 0$ (not including u_{N-1}) whether or not $x_{00} = 0$. Thus the likelihood equations for forced plants can be expected to have at least one real root.

4.3 LARGE SAMPLE CHARACTERISTICS

One of the most important and desirable large sample characteristics of an estimator is consistency, convergence to the true parameter value. Proofs of consistency of maximum likelihood estimators are common in the literature. The assumptions on which the proofs are based may vary, but the instance is relatively rare when the assumption of independent identically distributed samples is not included. Unfortunately, the samples for the case when x_0 is known or x_0 is unknown parameter are not identically distributed. In the x_0 unknown random variable and the differencing approach cases, the samples are not even independent.

Kendall and Stuart [1961, p.60] present a brief general discussion of maximum likelihood estimation when the samples are independent but not identically distributed. They point out that in this situation it is no longer necessarily true that ML estimators are consistent and give examples to illustrate this. In fact, for certain situations the ML estimator may not be meaningful. Thus ML estimators in non-standard situations must be considered individually.

When the initial condition x_0 is known and the plant is scalar, the distribution of the i th sample from (3.7) is:

$$p_i(y, a) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (y - ha^i x_0 - hb \sum_{j=1}^i a^{i-j} u_{j-1})^2\right] \quad i = 1, 2, \dots \quad (4.70)$$

Assume h , b , and x_0 not zero and u_k not identically zero. (If instead, $x_0 = 0$, neglect p_1 and assume $u_0 \neq 0$ for the following development.)

Then $p_i(y; a_1) = p_i(y; a_2)$ for a.e. y only if $a_1 = a_2$. Taking the limit of p_i on i gives:

$$\lim_{i \rightarrow \infty} p_i(y; a) = p(y; a) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (y - f(a, u_0, u_1, \dots))^2\right] \quad (4.71)$$

The function f exists and is continuous on the interior of its region of convergence, $(-1, 1)$, if the u_i are assumed to be uniformly bounded, i.e., $|u_i| \leq M < \infty$, and $a \in (-1, 1)$. Since f is not a constant, $p(y; a_1) = p(y; a_2)$ for a.e. y only if $a_1 = a_2$.

Let the subset of the real line $[-1, 1]$ be denoted by $\mathcal{Q}$, and assume a_0 , the true value of the unknown parameter, is an interior point of $\mathcal{Q}$. Since $\mathcal{Q}$ is compact and p_i is continuous on $\mathcal{Q}$, there exists a maximum likelihood estimator of a_0 based on N samples. Denote this estimator by $\hat{a}_N$.

Let

$$g_i(y, a) = \log[p_i(y, a)/p_i(y, a_0)] \quad (4.72)$$

and

$$g(y, a) = \log[p(y, a)/p(y, a_0)] \quad (4.73)$$

Both g_i and g are integrable for all a interior to $\mathcal{Q}$ in the sense that their expectations exist.

The following theorem is an extension of one by Jennrich [1970]† for independent identically distributed random variables.

† Class notes in Classical Statistics, Department of Mathematics, University of California, Los Angeles.

Theorem 4.3: Under the above assumptions,

$$\hat{a}_N \rightarrow a_0 \quad \text{a.s.}$$

Proof:

Since the logarithm is a strictly convex function, by Jensen's inequality,

$$\int g_i(y, a) p_i(y, a_0) d\mu(y) \leq \log \int p_i(y, a) d\mu(y) = 0 \quad (4.74)$$

with equality holding only if $a = a_0$. Let B be a neighborhood of a . Then,

$$\sup_{\alpha \in B} g_i(y, \alpha) + g_i(y, a) \text{ as } B \rightarrow a. \quad (4.75)$$

Define the expectation operator E_i as

$$E_i(\cdot) = \int (\cdot) p_i(y, a_0) d\mu(y) \quad (4.76)$$

By the monotone convergence theorem

$$E_i[\sup_{\alpha \in B} g_i(y, \alpha)] + E_i[g_i(y, a)] \text{ as } B \rightarrow a \quad (4.77)$$

and therefore there exists a B such that

$$E_i[\sup_{\alpha \in B} g_i(y, \alpha)] < 0 \quad (4.78)$$

whenever $a \neq a_0$. Let $D \subset Q$ be a neighborhood of a_0 , and let

$D^C = Q - D$. Because Q is compact, the complement D^C can be covered by a finite number of B neighborhoods, and thus

$$E_i[\sup_{\alpha \in D^C} g_i(y, \alpha)] < 0 \quad (4.79)$$

Similarly,

$$E[\sup_{\alpha \in D^C} g(y, \alpha)] < 0 \quad (4.80)$$

The variance of $g_i(y, a)$ depends on the first, second and fourth moments of y which are bounded above. Then by the strong law of large numbers, for a.e. set of samples $\{y_i\}$ from $p_i(y, a_0)$, $i = 1, \dots, N$,

$$\begin{aligned} \sup_{\alpha \in D^C} \frac{1}{N} \sum_{i=1}^N g_i(y_i, \alpha) &\leq \frac{1}{N} \sum_{i=1}^N \sup_{\alpha \in D^C} g_i(y_i, \alpha) \\ &\rightarrow \frac{1}{N} \sum_{i=1}^N E_i[\sup_{\alpha \in D^C} g_i(y, \alpha)] < 0 \end{aligned} \quad (4.81)$$

Choose such a sample and let $\hat{a}_N = \hat{a}_N(y_1, \dots, y_N)$. By definition of the ML estimate,

$$\frac{1}{N} \sum_{i=1}^N g_i(y_i, \hat{a}_N) = \frac{1}{N} \log \frac{p_1(y_1, \hat{a}_N) \dots p_N(y_N, \hat{a}_N)}{p_1(y_1, a_0) \dots p_N(y_N, a_0)} \geq 0 \quad (4.82)$$

Thus, $\hat{a}_N \in D$ for sufficiently large N . Since D is arbitrary, $\hat{a}_N \rightarrow a_0$ a.s. I

When the system is autonomous, the above proof for consistency does not hold. Referring to the limit in (4.71), $a^i x_0 \rightarrow 0$ as $i \rightarrow \infty$ for a an interior point of $\mathcal{Q}$. The uniqueness of the density p_i with respect to a is lost in the limit.

If x_0 is an unknown parameter, the above proof must be reworked with the unknown parameter as a vector instead of a scalar. This appears to be a natural extension of the theorem. Using a different approach, Aoki and Yue [1970] have shown consistency for this case.

The above theorem can also be used to show consistency when x_0 is an unknown random variable. The proof follows through directly when the densities for this situation are conditioned on x_0 . Since consistency exists for a.e. x_0 , then $\hat{a}_N \rightarrow a_0$ a.s.

In the final case, the differencing approach, the samples are not independent. There does not appear to be any simple technique to get around this problem as there was when x_0 was an unknown random variable. Wald [1948] and Aoki and Yue, however, do consider the problem of

For vector samples and parameters most of the above should go through with perhaps some additional algebra. Aoki and Yue treat the companion matrix case when x_0 is an unknown parameter. Mann and Wald [1943] develop the companion matrix case for the differencing approach but with independent samples.

4.4 APPROXIMATIONS

A characteristic common to the ML estimators in all four cases considered is that all the samples must be saved to be able to evaluate the estimate $\hat{a}_N$, and as the number of samples N increases, the amount of computation involved in this evaluation increases. This situation is inconsistent with the requirement of real time identification. The possibility of condensing the data through sufficient statistics was eliminated earlier. Exact algebraic factoring appears hopeless. Approaches to approximating the inverse of the covariance matrix in the differencing approach are given by Cochrane and Orcutt [1949], Hannan [1960, p.47] and Anderson [1963]. None of these appear to be very satisfactory.

The approximations with most appeal involve some form of truncation of the likelihood equation polynomial. The simplest approach of this nature is to truncate the polynomials after some arbitrary number of terms. However, this limits the number of samples that can be used to compute the estimate, and as a result, new data beyond some point will not be used. Forgetting for the moment how to accomodate initial condition information, for systems whose parameters are in fact slowly varying with time, the truncated polynomial could be made to undergo a

continual shift in indices so that old data is dropped off as new data comes in. If, however, use of all the data is desirable as would be the case for constant parameters, some sort of averaging scheme can be used with the truncation. This latter approach is pursued in what follows.

There are two obvious types of averaging modifications that could be made to the shifting polynomial scheme just described. One would be to define a new estimate as the running average of the estimates from the shifting polynomials. Because in certain situations this estimate tends to have an (infinite variance) Cauchy distribution, it does not appear to be as useful as an alternative scheme which keeps a running average of each of the coefficients of the shifting polynomials. The latter scheme bases estimates on the truncated polynomial evaluated using averaged coefficients.

Both the x_0 known and the x_0 unknown random variable cases use initial condition information in the ML estimate. Use of this information in either shifting polynomial scheme generates another growing polynomial required to shift the origin thus nullifying the computational advantage gained by truncating. The coefficient averaging scheme for the x_0 unknown parameter case, which is more or less a steady state version of the other two, will be assumed to apply to all three cases.

4.4.1 AVERAGE COEFFICIENT APPROXIMATIONS TO THE LIKELIHOOD EQUATIONS

The average coefficient approximation equation when x_0 is an unknown parameter can be developed from Equation (3.22). The number of samples, $(N+1)$ in Equation (3.22), at which to truncate the dependent observations.

polynomials is arbitrary, but at least two samples must be used over which to average. Because truncation after two samples yields the simplest result, the truncation will be taken at that point. Thus for two samples, the likelihood equation becomes:

$$y_0(y_1 - hbu_0) + [(y_1 - hbu_0)^2 - y_0^2]a - y_0(y_1 - hbu_0)a^2 = 0 \quad (4.83)$$

or with averaging over $N + 1$ samples gives the average coefficient expression for x_0 unknown parameter (as well as x_0 known and x_0 unknown random variable).

$$C_N' a^2 - D_N' a - C_N' = 0 \quad (4.84)$$

where:

$$C_N' = \frac{1}{N} \sum_{i=1}^N y_{i-1}(y_i - hbu_{i-1}) \quad (4.85)$$

$$D_N' = \frac{1}{N} \left(\sum_{i=1}^N (y_i - hbu_{i-1})^2 - \sum_{i=1}^N y_{i-1}^2 \right) \quad (4.86)$$

For the same situation but with the vector-valued autonomous system, Equations (3.29) and (3.33) give

$$\phi_1 = H^T R^{-1} H + A^T H^T R^{-1} H A \quad (4.87)$$

$$AD_1 = [H^T R^{-1} y_1 - H^T R^{-1} H A \phi_1^{-1} H^T R^{-1} y_0 - H^T R^{-1} H A \phi_1^{-1} A^T H^T R^{-1} y_1] \\ [y_0^T R^{-1} H + y_1^T R^{-1} H A] = 0 \quad (4.88)$$

or, averaging over $N + 1$ samples:

$$H^T R^{-1} \left(\frac{1}{N} \sum_{i=1}^N y_i y_{i-1}^T \right) R^{-1} H + H^T R^{-1} \left(\frac{1}{N} \sum_{i=1}^N y_i y_i^T \right) R^{-1} H A \\ - H^T R^{-1} H A \phi_1^{-1} H^T R^{-1} \left[\left(\frac{1}{N} \sum_{i=1}^N y_{i-1} y_{i-1}^T \right) R^{-1} H + \left(\frac{1}{N} \sum_{i=1}^N y_i y_{i-1}^T \right)^T R^{-1} H A \right] \\ - H^T R^{-1} H A \phi_1^{-1} A^T H^T R^{-1} \left[\left(\frac{1}{N} \sum_{i=1}^N y_i y_{i-1}^T \right) R^{-1} H + \left(\frac{1}{N} \sum_{i=1}^N y_i y_i^T \right) R^{-1} H A \right] = 0 \quad (4.89)$$

Similarly, for the differencing approach, (3.64) gives,

$$\sigma^2 a^3 + y_0(y_1 - hbu_0)a^2 + [\sigma^2 + y_0^2 - (y_1 - hbu_0)^2]a - y_0(y_1 - hbu_0) = 0 \quad (4.90)$$

or treating the shifted y_0 's as known initial conditions and averaging over $N + 1$ samples gives the average coefficient expression for the differencing approach:

$$\sigma^2 a^3 + C_N' a^2 + [\sigma^2 - D_N']a - C_N' = 0 \quad (4.91)$$

If the plant is vector-valued and $H = I$, the identity matrix, then from (3.80) the two sample average coefficient approximation becomes:

$$\begin{aligned} -AR + \frac{1}{N} \left(\sum_{i=1}^N (y_i - Ay_{i-1} - Bu_{i-1}) y_{i-1}^T \right) \\ + \frac{1}{N} \left(\sum_{i=1}^N (y_i - Ay_{i-1} - Bu_{i-1}) (y_i - Ay_{i-1} - Bu_{i-1})^T \right) (R + ARA^T)^{-1} AR = 0 \end{aligned} \quad (4.92)$$

4.4.2 PROPERTIES OF THE TWO-SAMPLE AVERAGE COEFFICIENT APPROXIMATIONS

The finite sample properties of the two-sample average coefficient approximations to the scalar plant likelihood equations will be investigated first. When the initial condition is an unknown parameter, the average coefficient equation (4.82) is a quadratic with roots:

$$(D_N' \pm \sqrt{(D_N')^2 + 4(C_N')^2}) / (2C_N') \quad (4.93)$$

Two conclusions are immediate. The two roots are always real. By the triangle inequality, one root lies in the closed interval $[-1, 1]$, and the other root lies outside the open interval $(-1, 1)$ unless $D_N' = 0$, an event which occurs with probability zero. In that case, the roots are ± 1 .

Earlier, in the case where x_0 is an unknown parameter, the ML

AD-A072 147

CALIFORNIA UNIV LOS ANGELES SCHOOL OF ENGINEERING A--ETC F/G 12/2
MAXIMUM LIKELIHOOD IDENTIFICATION OF LINEAR DISCRETE STOCHASTIC--ETC(U)
JUL 78 A J GLASSMAN, C T LEONDES F33615-77-C-3013

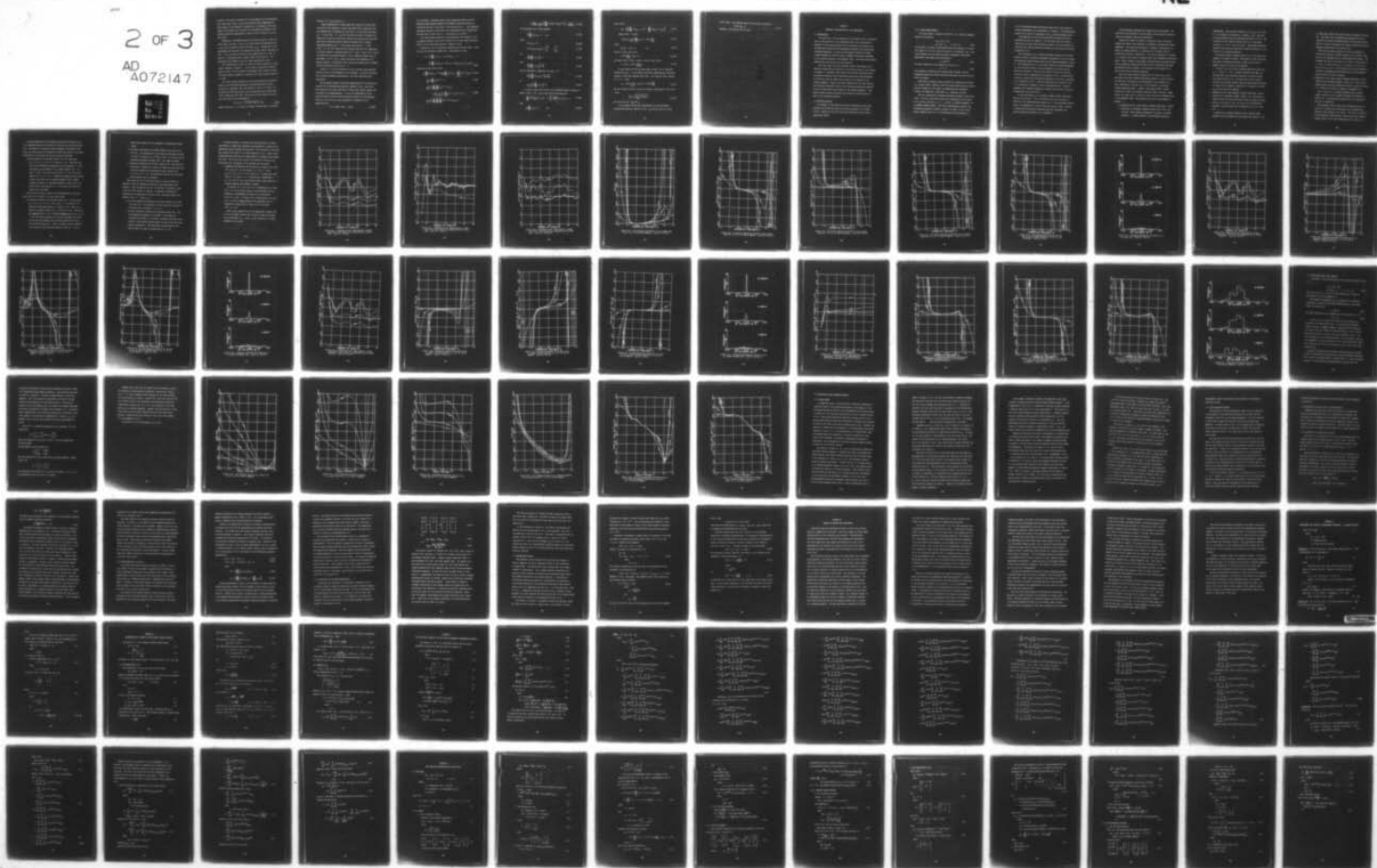
UNCLASSIFIED

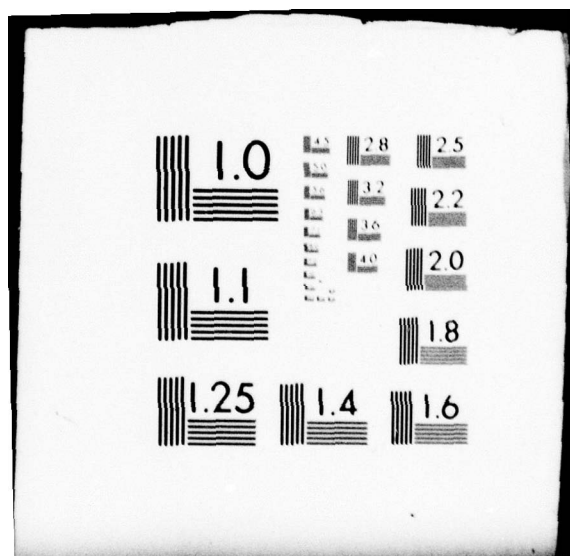
AFFDL-TR-78-84

NL

2 OF 3

AD
A072147





estimate $\hat{a}$ was shown to approach the true parameter a_0 as the measurement noise goes to zero. Since this property holds independent of which sample in the sequence of samples $\{y_i\}$ is denoted y_0 (the first sample of the string to be used for the two-sample ML estimate), the average coefficient approximation (4.84) for any N yields $\hat{a} = a_0$ for zero noise.

Furthermore, the expected value of the noise terms of (4.84) are easily seen to be zero. In C_N' , the noise terms are weighted sums of η_i and a sum of product terms of the form $\eta_i \eta_{i+1}$. Because the η_i are independent and zero mean, the expectation of each term in the sums is zero. In D_N' , the noise terms are also weighted sums of η_i and, in addition, a telescoping sum of square terms of the form η_i^2 . The telescoping sum reduces to $\eta_N^2 - \eta_0^2$, whose expectation is zero.

This approximation (4.84) is related to the ML estimator of Levin [1964] discussed in Chapter 2. If his result is applied as each new sample is made, in the so-called "overlapping" mode, instead of after collecting groups of samples as intended, his result becomes identical to (4.84).

The finite sample properties of the average coefficient approximation to the differencing approach likelihood equation are more difficult to establish. Now, the equation, (4.91), is a cubic. An indication of the root location can be obtained by first considering (4.91) without the $\sigma^2 a^3$ term. This portion of the equation has two real roots which can be expressed as:

$$[-(\sigma^2 - D_N') \pm \sqrt{(\sigma^2 - D_N')^2 + 4C_N'^2}] / (2C_N') \quad (4.94)$$

Again, unless $D_N' = \sigma^2$, one root is stable, and the other is unstable.

When $D_N' = \sigma^2$, the roots are ± 1 .

Closer examination of (4.94) shows that the pair of roots falls into one of two categories. Either the stable root is positive, and the unstable root is negative or vice versa. In the former case when $C_N' > 0$ (and $\sigma^2 - D_N' > 0$), the $\sigma^2 a^3$ term has the effect of biasing the stable root toward zero and either biasing the unstable root away from zero while introducing another negative unstable root or merely removing the unstable root. In the latter case when $C_N' > 0$ (and $\sigma^2 - D_N' < 0$), the $\sigma^2 a^3$ term moves the unstable positive root toward zero to the point where it becomes stable if $\sigma^2 > D_N'/2$. Also, if $\sigma^2 \leq D_N'/2$, the stable root is shifted toward -1 , and a negative unstable root is introduced. If $\sigma^2 > D_N'/2$, either the stable root disappears or it becomes unstable (and negative) and a still more negative root is added. Similar conclusions follow for $C_N' < 0$.

Unless $\sigma^2 \rightarrow 0$ as the noise goes to zero, the zero noise condition does not give the true parameter as the estimate. As was the case with the x_0 unknown parameter approximation, the noise terms have zero expectation.

The two-sample average coefficient approximation to the likelihood equation for the differencing approach, Equation (4.91), could have been derived in two other ways each of which gives further insight into the nature of the approximation. In one, the approximation (4.91) follows directly from the scalar autonomous version of the likelihood function (3.64) with the noise covariance R of Equation (3.66) approximated as:

$$R = \sigma^2 \text{diag} (1+a^2, \dots, 1+a^2) \quad (4.95)$$

In the second, a modified version of the likelihood function can be developed using grouped samples in the sense of Levin by basing the likelihood function on the model (3.50) with $i=0,2,4,\dots$. The resulting likelihood equation is then used in an overlapping mode by resubscripting such that in effect $i=0,1,\dots$ once again as in (3.50).

The large sample properties of the approximate ML estimate in the x_0 unknown parameter case can be inferred by the large sample characteristics of its "likelihood equation", Equation (4.84).

Assume the $\{u_i\}$ are uniformly bounded by $M \geq 0$ and that $|a_0| < 1$. Then the $\{x_i\}$ are uniformly bounded also. From Equation (4.85):

$$\begin{aligned} C_N' &= \frac{1}{N} \sum_{i=1}^N (hx_{i-1} + \eta_{i-1})(ha_0 x_{i-1} + \eta_i) \\ &= \frac{1}{N} \sum_{i=1}^N [h^2 a_0 x_{i-1}^2 + hx_{i-1} \eta_i + ha_0 \eta_{i-1} x_{i-1} + \eta_{i-1} \eta_i] \end{aligned} \quad (4.96)$$

Examining (4.96) term by term:

$$\frac{1}{N} \sum_{i=1}^N h^2 a_0 x_{i-1}^2 = h^2 a_0 \frac{1}{N} \sum_{i=1}^N (a_0^{i-1} x_0 + \sum_{j=1}^{i-1} a_0^{i-1-j} u_{j-1})^2 \quad (4.97)$$

(i>1)

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (a_0^2)^{i-1} = 0 \quad (4.98)$$

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{N} \left| \sum_{i=2}^N \sum_{j=1}^{i-1} a_0^{2(i-1)-j} u_{j-1} \right| \\ \leq \frac{M}{1-a_0} \lim_{N \rightarrow \infty} \frac{1}{N} \left| \sum_{i=2}^N [a_0^{i-1} - (a_0^2)^{i-1}] \right| = 0 \end{aligned} \quad (4.99)$$

$$\lim_{N \rightarrow \infty} \frac{1}{N} \left| \sum_{i=2}^N \sum_{j=1}^{i-1} \sum_{k=1}^{i-1} a_0^{2i-j-k-2} u_{j-1} u_{k-1} \right|$$

$$\leq \frac{N^2}{(1-a_0)^2} \lim_{N \rightarrow \infty} \frac{1}{N} \left| \sum_{j=2}^N (1-2a_0^{j-1} + (a_0^2)^{j-1}) \right| = \frac{N^2}{(1-a_0)^2} \quad (4.100)$$

By the strong law of large numbers,

$$\frac{1}{N} \sum_{i=1}^N h x_{i-1} \eta_i \rightarrow 0 \quad \text{a.s.} \quad (4.101)$$

Also,

$$E(\eta_i \eta_{i+1}) = 0 \quad (4.102)$$

$$E[(\eta_i \eta_{i+1})(\eta_j \eta_{j+1})] = \begin{cases} \sigma^4 & , i=j \\ 0 & , i \neq j \end{cases} \quad (4.103)$$

So,

$$E\left[\frac{1}{N} \sum_{i=1}^N \eta_i \eta_{i-1}\right] = 0 \quad (4.104)$$

and,

$$E\left[\frac{1}{N} \sum_{i=1}^N \eta_i \eta_{i-1}\right]^2 = \frac{\sigma^4}{N} \quad (4.105)$$

Then by Chebyshev's inequality, for any $\epsilon > 0$,

$$P\left(\left|\frac{1}{N} \sum_{i=1}^N \eta_i \eta_{i-1}\right| \geq \epsilon\right) \leq \frac{1}{N} \frac{\sigma^4}{\epsilon^2} \quad (4.106)$$

or

$$P \lim_{N \rightarrow \infty} \left(\frac{1}{N} \sum_{i=1}^N \eta_i \eta_{i-1}\right) = 0 \quad (4.107)$$

A stronger result than (4.107) can be shown by using a theorem of Révész [1968, p. 87]. By (4.102) and (4.103) and since,

$$\sum_{i=1}^{\infty} \frac{E(\eta_i^2 \eta_{i-1}^2)}{i^2} \log^2 i < \sigma^4 \int_1^{\infty} \frac{\log^2 x}{x^2} dx = \sigma^4 \Gamma(3) < \infty \quad (4.108)$$

then,

$$\frac{1}{N} \sum_{i=1}^N \eta_i \eta_{i-1} \rightarrow 0 \quad \text{a.s.} \quad (4.109)$$

From (4.86),

$$D_N' = \frac{1}{N} \left[\sum_{i=1}^N (h a_0 x_{i-1} + \eta_i)^2 - \sum_{i=1}^N (h x_{i-1} + \eta_{i-1})^2 \right] \quad (4.110)$$

Using (4.97) - (4.100),

$$h^2 (a_0^2 - 1) \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N x_{i-1}^2 = -h^2 M^2 \frac{1+a_0}{1-a_0} \quad (4.111)$$

Since,

$$\frac{1}{N} (\eta_N^2 - \eta_0^2) \rightarrow 0 \quad \text{a.s.} \quad (4.112)$$

then by (4.111) and (4.101),

$$-h^2 M^2 \frac{1+a_0}{1-a_0} \leq -|D_\infty'| \leq 0 \quad (4.113)$$

and from (4.98), (4.99), (4.100), (4.101), and (4.109),

$$0 \leq |C_\infty'| \leq h^2 M^2 \frac{a_0}{(1-a_0)^2} \quad (4.114)$$

At this stage all that has been shown is that if $\{u_i\}$ uniformly bounded and $|a_0| < 1$, the average coefficient approximation likelihood equation (4.84) does approach some limit. As a special case, consider $u_i \equiv M > 0$. Then

$$C_\infty' = h^2 M^2 \frac{a_0}{(1-a_0)^2} \quad \text{and} \quad D_\infty' = -h^2 M^2 \frac{1+a_0}{1-a_0} \quad (4.115)$$

The two limiting roots of (4.84) are found by introducing (4.115) into (4.93),

$$\frac{D_\infty' \pm \sqrt{(D_\infty')^2 + 4(C_\infty')^2}}{2C_\infty'} \quad (4.116)$$

The two roots are $-\frac{1}{a_0}$ and a_0 .

In the average coefficient approximation for the autonomous differencing approach, Equation (4.91), C_N' and D_N' both go to zero

in the limit. The limiting roots for this case are given by

$$\sigma^2(a^2+1)a = 0 \quad (4.117)$$

Therefore, the limiting root is zero.

SECTION V

NUMERICAL CONSIDERATIONS OF THE IDENTIFIERS

5.1 INTRODUCTION

The objectives of the discussions in this chapter are to illustrate some of the properties of the identifiers which were established by theorems in the previous chapter and to investigate the mechanics for the numerical evaluation of the parameter estimates from the nonlinear likelihood equations. To this end the computer simulation results, grouped according to the four initial condition categories considered in the preceding chapters, are presented first. The results and related computational implications are then explored.

The terms "cost" and "cost function" used in this chapter, with some minor modifications detailed in the following section, were defined in Chapter 3. Basically, they refer to the function formed by taking the natural logarithm of the likelihood function and then discarding the additive terms and common factors which do not depend on the unknown parameter. Terms of the type "derivative of the cost function" and "derivative function" refer to the function obtained by differentiating the cost function with respect to the unknown parameter. (The equation which results by setting that derivative function equal to zero is the likelihood equation.)

5.2 SIMULATION RESULTS

The figures in this section are based on computations and noise generated on the IBM 360, model 91 and were prepared on a Cal-Comp plotter. Results are given for both the scalar model and the multi-dimensional models.

5.2.1 SCALAR MODEL RESULTS

The scalar model is defined by Equation (3.6), which is repeated below for convenience.

$$\begin{aligned}x_{i+1} &= ax_i + bu_i \\y_i &= hx_i + \eta_i \quad i = 0, 1, \dots, N\end{aligned} \quad (5.1)$$

(In the text, in order to distinguish between the parameter a and the true value of a in (5.1), the true value of a is denoted by a_0 .) The noise sequence $\{\eta_i\}$ was taken as gaussian independent identically distributed, each member with distribution:

$$\eta_i \sim \mathcal{N}(0, \sigma^2) \quad (5.2)$$

The known coefficients were assumed to be unity, i.e.,

$$h = b = 1 \quad (5.3)$$

(For purposes of exercising the identification schemes, various assumptions about the initial conditions were made, but these did not affect the model.)

Some theory on optimal input selection for identification exists, e.g., Staley [1968]. However, since the normal operating input restriction was assumed for this study, no optimization was attempted. Instead, for simplicity a step input, $u_i = \text{constant}$, was used.

The four identification problems discussed in Chapter 3 were simulated. The first three differ strictly by assumptions on the nature of the initial condition x_0 , i.e., x_0 known, x_0 unknown parameter, or x_0 unknown random variable. The fourth identification problem, the differencing approach, actually is based on a model, (3.49), which differs somewhat from (5.1). By differencing the measurements y_i

in (5.1) the models become equivalent except that in the latter the initial measurement y_0 is considered as a constant. (The differencing approach most closely corresponds to the x_0 unknown parameter case.)

The equations for the four identification problems that were simulated are given in Chapter 3. For the case where the initial condition x_0 is known and the system is scalar, the cost function is given by (3.8), and the derivative function is given by the left side of Equation (3.9).

The case where x_0 is an unknown parameter requires a bit of discussion in order to maintain reasonable consistency in the terminology. The difficulty arises because there are two unknown parameters, x_0 and a . Since estimation of a is of primary interest the second parameter, x_0 , was eliminated through using $\hat{x}_0$ instead of x_0 in the cost and derivative functions. Thus the cost function is given by Equation (3.18) but with x_0 replaced by $\hat{x}_0$ of Equation (3.20). The derivative function is given by the left side of Equation (3.21) with x_0 replaced by $\hat{x}_0$ of Equation (3.20). (When the derivative is set equal to zero it is equivalent to Equation (3.22), the equation for a . Strictly speaking, the likelihood equation is neither of these but is the pair of Equations (3.19) and (3.21).)

For the case where x_0 is a gaussian random variable with known mean $\bar{x}_0$ and variance ϵ^2 , the cost function which was simulated is given by (3.41) normalized with respect to σ^2 . The derivative function used in the simulation is given by the left side of (3.44). However, in order that the derivative and cost functions correspond, the derivative must be divided by $(\sigma^2 + h^2 \epsilon^2 \underline{a}^T \underline{a})^2$.

The differencing approach cost function was not simulated. The left side of the Equation (3.70) was taken as the derivative function. (Only the autonomous version was simulated.)

Because the number of figures is relatively large, the figure numbers are coded to help identify the situation the associated figure represents. The numerical designation 1 through 4 corresponds to x_0 known, x_0 unknown parameter, x_0 unknown random variable, and the differencing approach, respectively. The letter designations a through j correspond to the various situations which were simulated as described below. Again, because of the number of figures and the varying amounts of new information introduced by them, not all the figures that were developed have been included. This accounts for what appear to be gaps in the literal numbering sequences.

The first three groups of figures are an illustration of the evolution of the MLE (maximum likelihood estimate) of a_0 for specific, but arbitrary realizations of the measurement noise sequence $\{\eta_i\}$. Also, comparison of the MLE of a_0 to the estimate of a_0 by other schemes (described later) - namely, least squares (LSQ), 3-point recursive fit (3PT), and average coefficient (AVC) are shown. These figures are shown first to unify the later ones which concentrate more on the immediate issue - the solution for $\hat{a}_N$ for a given number of samples.

a. Comparison of ML, least squares, average coefficient, 3-point recursive fit estimation schemes: Figures 5-1a, 5-2a, 5-3a, and 5-4a. These figures correspond to x_0 known, x_0 unknown parameter, x_0 random variable, and differencing approach,

respectively. For all four figures, $a_0 = -0.5$, $x_0 = -3.0$, and $u_1 \equiv 1.0$ except in 5-4a where $u_1 \equiv 0$ and $x_0 = 6.0$. The level of the measurement noise is relatively low with $\sigma^2 = 0.01$. The increment for the 3-point recursive fit is 0.1 with points located at $a = -0.65$, -0.55 , and -0.45 . The maximum likelihood estimates were found by *regula falsi* iterative solution for the roots of the likelihood equations.

The maximum number of samples shown is 30. Most curves were computed up through 60 samples. The behavior of the curves for the second 30 samples was similar to that of the first 30 samples, except for the first few samples.

In Figure 5-3a, the initial condition mean and variance are $\bar{x}_0 = -3.0$ and $\epsilon^2 = 0.02$. (Another simulation was made, not included here, with conditions identical to those of Figure 5-3a except that $\bar{x}_0 = 6.0$. The fact that the true initial condition and the mean initial condition were grossly mismatched in terms of the variance ϵ^2 resulted in a transient in the MLE of a_0 for the first few samples.)

b. Comparison of ML, least squares, average coefficient, 3-point recursive fit estimation schemes: Figure 5-1b. This group is computed under the same conditions as those in group (a) except that $a_0 = 0.75$ and $x_0 = \bar{x}_0 = 2.0$ ($x_0 = 0.6$ for differencing approach), and the 3-point fit was made at $a = 0.6$, 0.7 , and 0.8 .

Figures for x_0 unknown parameter and x_0 unknown random variable are not included, but they appear very similar to the

x_0 known case, Figure 5-1b, much as was the situation in group (a). The figure for the differencing approach also is not included but initially resembles Figure 5-1b and then settles down as in Figure 5-4a except the average coefficient and least squares solutions drift at a greater rate.

Both for this group and group (a) with the exception of differencing approach, another variation of the 3-point recursive fit was computed but not included among the figures. The increment for the fit was reset to 0.2 from 0.1 with points at $a = -0.8, -0.6$ and -0.4 for $a_0 = -0.5$ and at $a = 0.45, 0.65$, and 0.85 for $a_0 = 0.75$. This increase in the point separation resulted in estimates which differed from the true parameter value by about 10% after 30 samples.

c. Comparison of ML, least squares, average coefficient, 3-point recursive fit estimation schemes: Figure 5-1c. This group is computed under the same conditions as those in group (a) except $\sigma^2 = 0.4356$ and the number of samples is extended to 60.

Again, figures for x_0 unknown parameter and x_0 unknown random variable are not included but appear very similar to Figure 5-1c. The differencing approach was not simulated for this set of conditions.

The measurement noise variance was increased over that in group (a) to observe the effectiveness of the schemes when operating under moderate noise levels. The value of 0.4356 for σ^2 was selected to make the one- σ value of the noise (approximately) equal to $2/3$, the limiting value of x_i with $u_i \equiv 1$.

Of prime importance in the numerical solution for the MLE of a_0 is the expected shape of the derivative function and the root distribution. Preliminary to displaying various examples of derivative functions, the cost function is presented in order that source of the sharp fluctuations in the derivative curves be better understood.

d. *Cost functions for the MLE: Figure 5-1d.* For this group,

$a_0 = -0.5$, $x_0 = -3.0$, $\sigma^2 = 0.01$ and $u_i \equiv 1.0$. (The cost function for the differencing approach was not evaluated.) In the case where x_0 was assumed to be a random variable, $\bar{x}_0 = -3.0$ and $\epsilon^2 = 0.02$. The curves are given for 3, 5, 10, 15, and 20 samples and $-1.5 \leq a \leq 1.5$. Though not necessarily very similar overall, the curves for x_0 unknown parameter and x_0 random variable do exhibit the essential feature of Figure 5-1d, the oscillation at just beyond $a = 1$. The figures for these two cases are not included.

The derivative function curves for various parameter values and noise levels are included in the next three groups.

e. *Derivative functions for the MLE: Figure 5-1e.* For this group,

$a_0 = -0.5$, $x_0 = -3.0$, $\sigma^2 = 0.01$, and $u_i \equiv 1.0$ (except in the differencing approach where $u_i \equiv 0$ and $x_0 = 6.0$). When x_0 is taken as a random variable, $\bar{x}_0 = -3.0$ and $\epsilon^2 = 0.02$. The curves are computed for 3, 5, 10, 15, and 20 samples and $-1.5 \leq a \leq 1.5$.

Over the range displayed, for the scale employed the curves for the four ML estimators are virtually identical to the no-noise curves of group (h). (They, of course, are not identical as can easily be seen from the curves of group (a).) The x_0

known case, Figure 5-1e, is presented to illustrate the similarity.

f. *Derivative functions for the MLE: Figures 5-1f, 5-2f, 5-3f, and 5-4f.* The conditions for this group are the same as those of group (e) except $a_0 = 0.75$ and $x_0 = \bar{x}_0 = 2.0$ (or 0.6 for the differencing approach, Figure 5-3f). All cases are shown.

g. *Derivative functions for the MLE: Figures 5-1g, 5-2g, and 5-3g.*

This group is identical to group (e) except the noise level was raised by increasing the measurement noise variance from

$\sigma^2 = 0.01$ to $\sigma^2 = 1.0$. The derivative function for the differencing approach was not computed.

The curves for all of the above groups of course represent responses to only one possible realization of the measurement noise sequence. Just how typical they are is difficult to establish, especially for small numbers of samples. To provide a deterministic reference for the derivative function curves, limiting versions were computed where $\sigma^2 \rightarrow 0$ and $\eta_i \rightarrow 0$.

h. *No noise derivative functions for the MLE: Figures 5-1h, 5-2h, 5-3h, and 5-4h.* The conditions for this group are the same as for group (e) or (g) except $\sigma^2 = \eta_i = 0$.

i. *No noise derivative functions for the MLE: Figure 5-4i.* Only the differencing approach is presented. (The reason for including this figure is to provide a contrast with Figure 5-4f similar to what exists between groups (g) and (h) for the three other ML estimators.) The conditions in this group are the same as those in group (f) except that $\sigma^2 = \eta_i = 0$.

In another attempt to overcome the inconclusiveness of single realizations, a Monte Carlo approach was used wherein a random set of measurement noise sequences was generated. The simulation of the model was repeated using each of the noise sequences in turn. Based on the accumulated data from the set of experiments, $\hat{a}$ frequency distributions were made. With these, some insight into derivative function root distribution and convergence of the estimate can be derived.

j. Frequency distributions for the MLE of a_0 : Figures 5-1j, 5-2j, 5-3j, and 5-4j. For this group, $a_0 = 0.75$, $\sigma^2 = 0.01$, $x_0 = 2.0$, $\bar{x}_0 = 2.0$, $\epsilon^2 = 0.02$, and $u_1 \equiv 1.0$ (except for 5-4j where $x_0 = 0.6$ and $u_1 \equiv 0$). The low noise combination of $a_0 = 0.75$ and $\sigma^2 = 0.01$ was chosen to help insure rapid convergence primarily for economic reasons.

In Figures 5-1j, 5-2j, and 5-3j, $\hat{a}$ distributions are given for 3, 10, and 60 samples based on 100 experiments with 81 frequencies corresponding to steps in $\hat{a}$ of 0.002 where $0.675 \leq \hat{a} \leq 0.837$. In Figure 5-4j, $\hat{a}$ distributions are given for 3, 10, and 30 samples based on 50 experiments with 15 frequencies corresponding to steps in $\hat{a}$ of 0.04 where $0.45 \leq \hat{a} \leq 1.05$.

Since x_0 was constant for all the experiments, Figure 5-3j for the x_0 random variable case should be viewed as distributions conditioned on x_0 . Also, y_0 was simulated as a random variable for all cases.

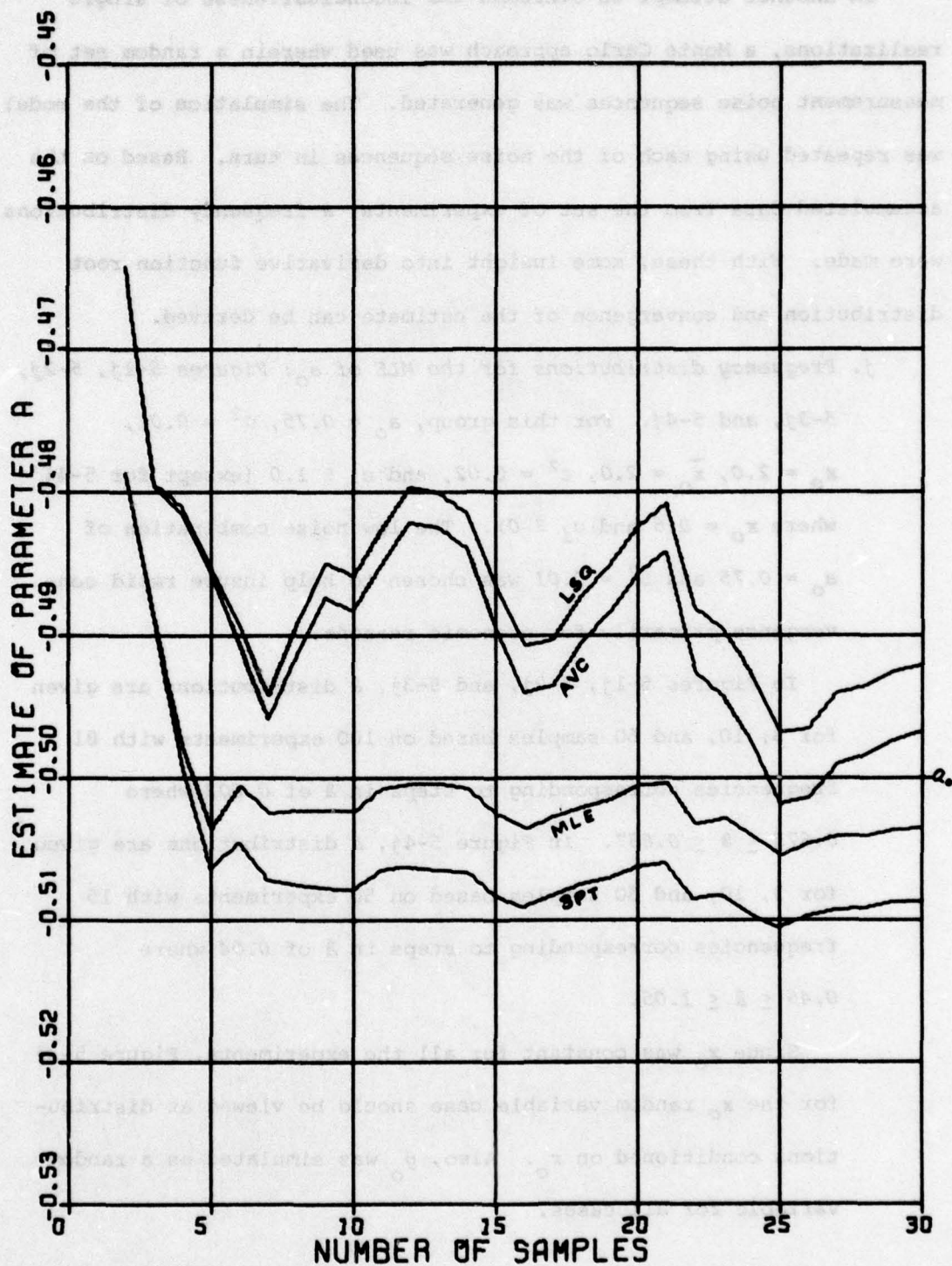


Figure 5-1a. Comparison of ML, least squares, average coefficient, and 3-point fit estimation for x_0 known case. ($a_0 = -0.5$, $\sigma^2 = 0.01$).

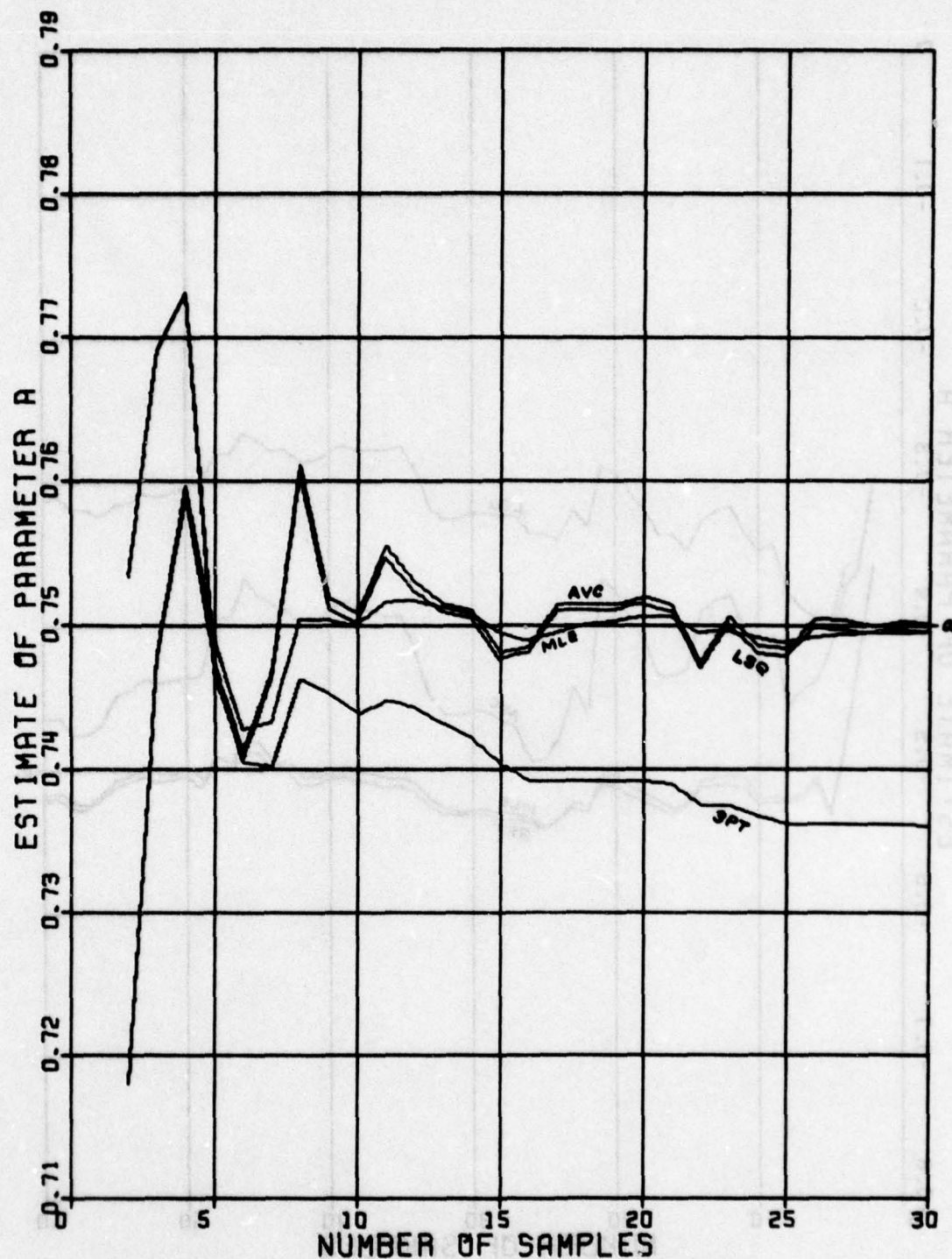


Figure 5-1b. Comparison of ML, least squares, average coefficient, and 3-point fit estimation for x_0 known case. ($a_0=0.75$, $\sigma^2=0.01$).

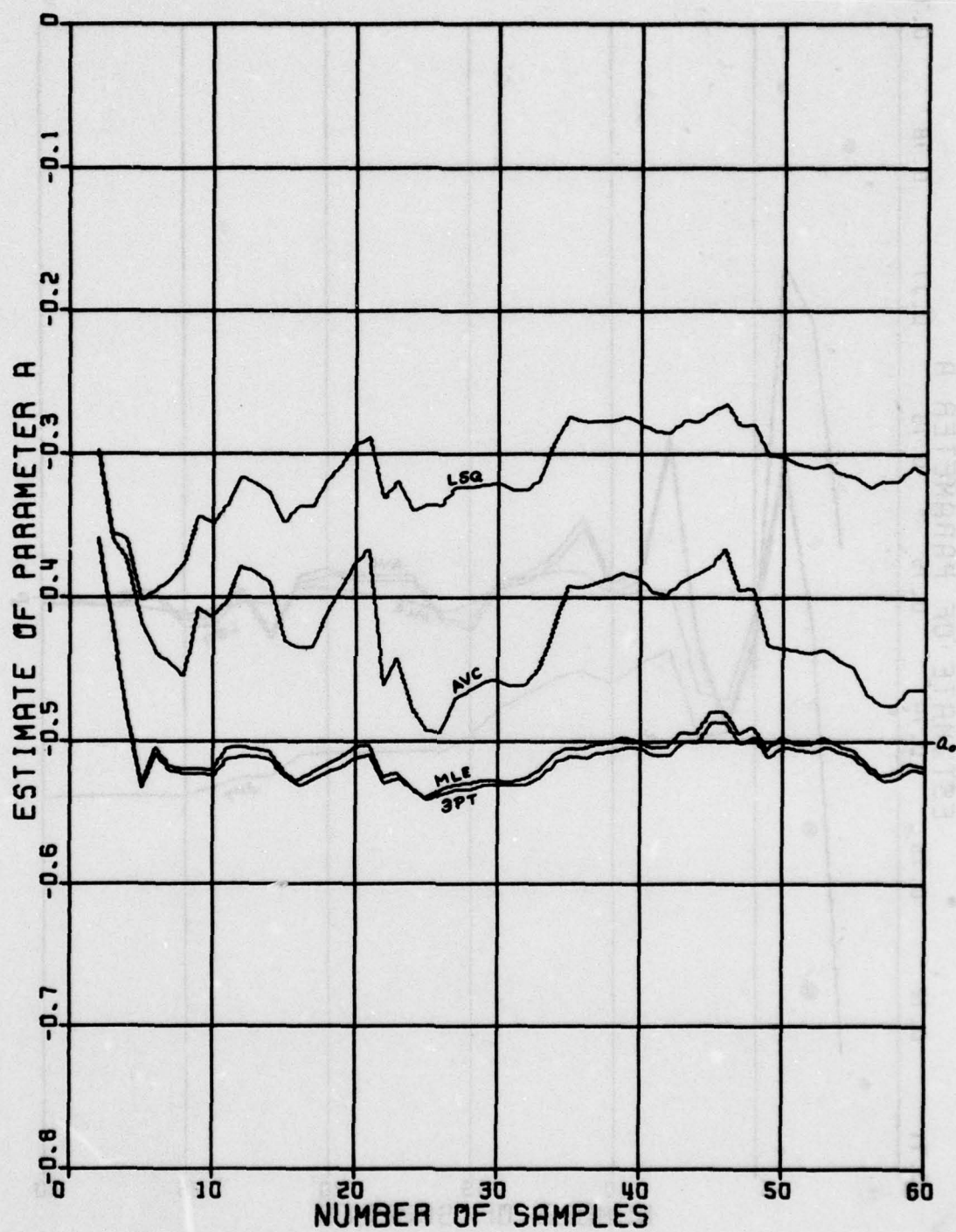


Figure 5-1c. Comparison of ML, least squares, average coefficient, and 3-point fit estimation for x_0 known case. ($a_0 = -0.5$, $\sigma^2 = 0.4356$).

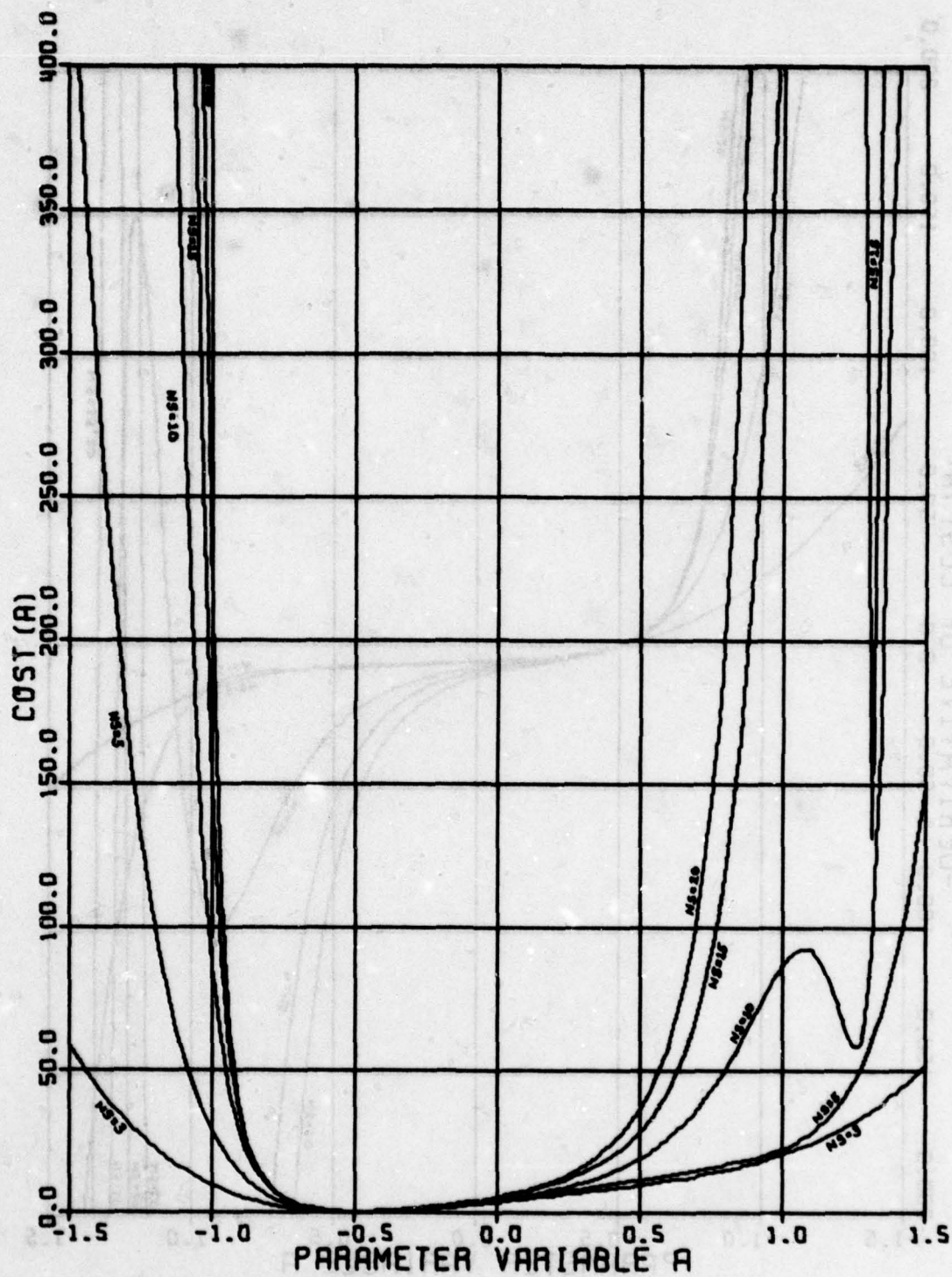


Figure 5-1d. Cost function of the MLE in the x_0 known case for 3, 5, 10, 15, and 20 samples. ($a_0 = -0.5$, $\sigma^2 = 0.01$).

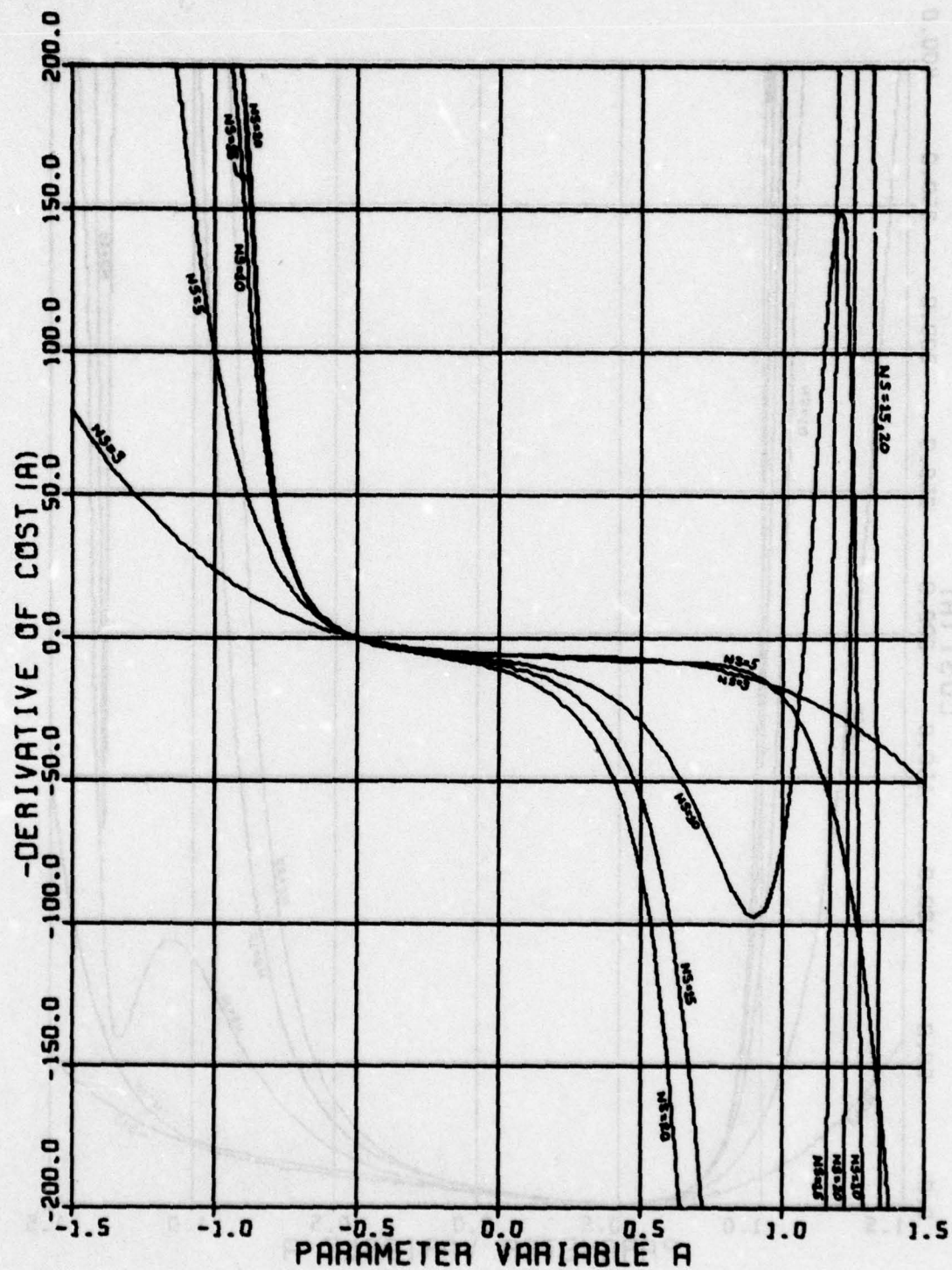


Figure 5-1e. Derivative function of the MLE in the x_0 known case for 3, 5, 10, 15, and 20 samples. ($a_0 = -0.5$, $\sigma^2 = 0.01$).

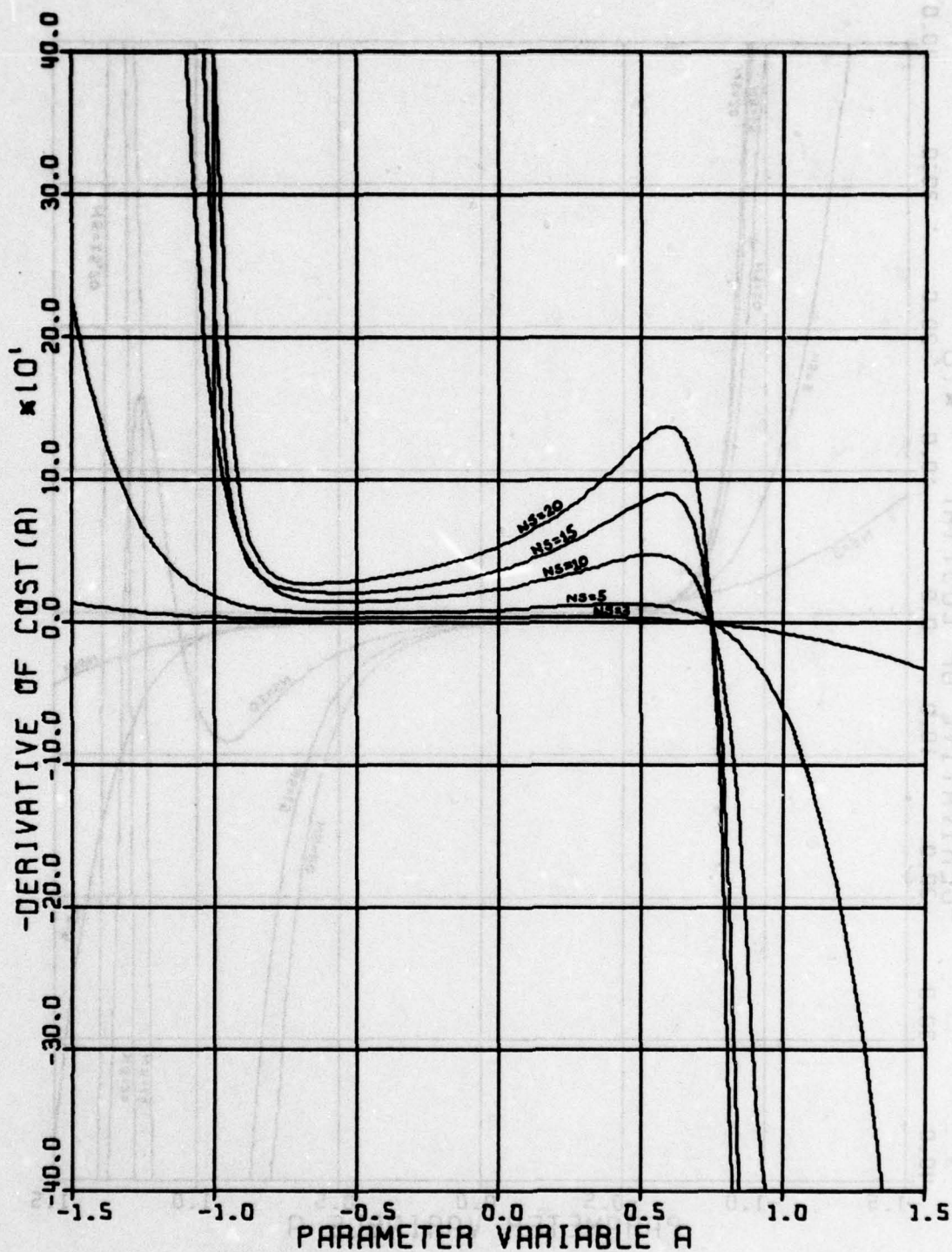


Figure 5-1f. Derivative function of the MLE in the x_0 known case for 3, 5, 10, 15, and 20 samples. ($a_0=0.75$, $\sigma^2=0.01$).

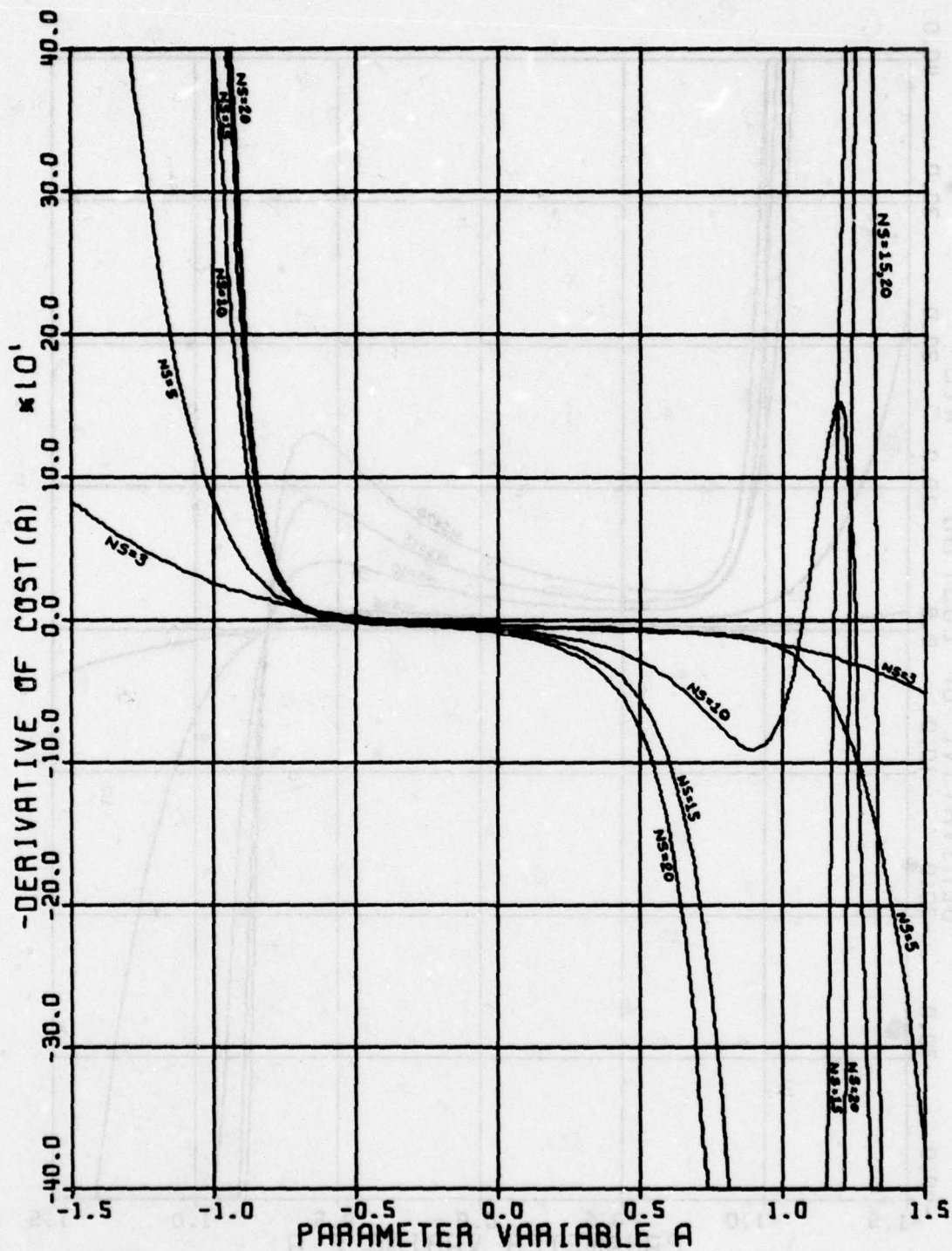


Figure 5-1g. Derivative function of the MLE in the x_0 known case for 3, 5, 10, 15, and 20 samples. ($a_0 = -0.5$, $\sigma^2 = 1.0$).

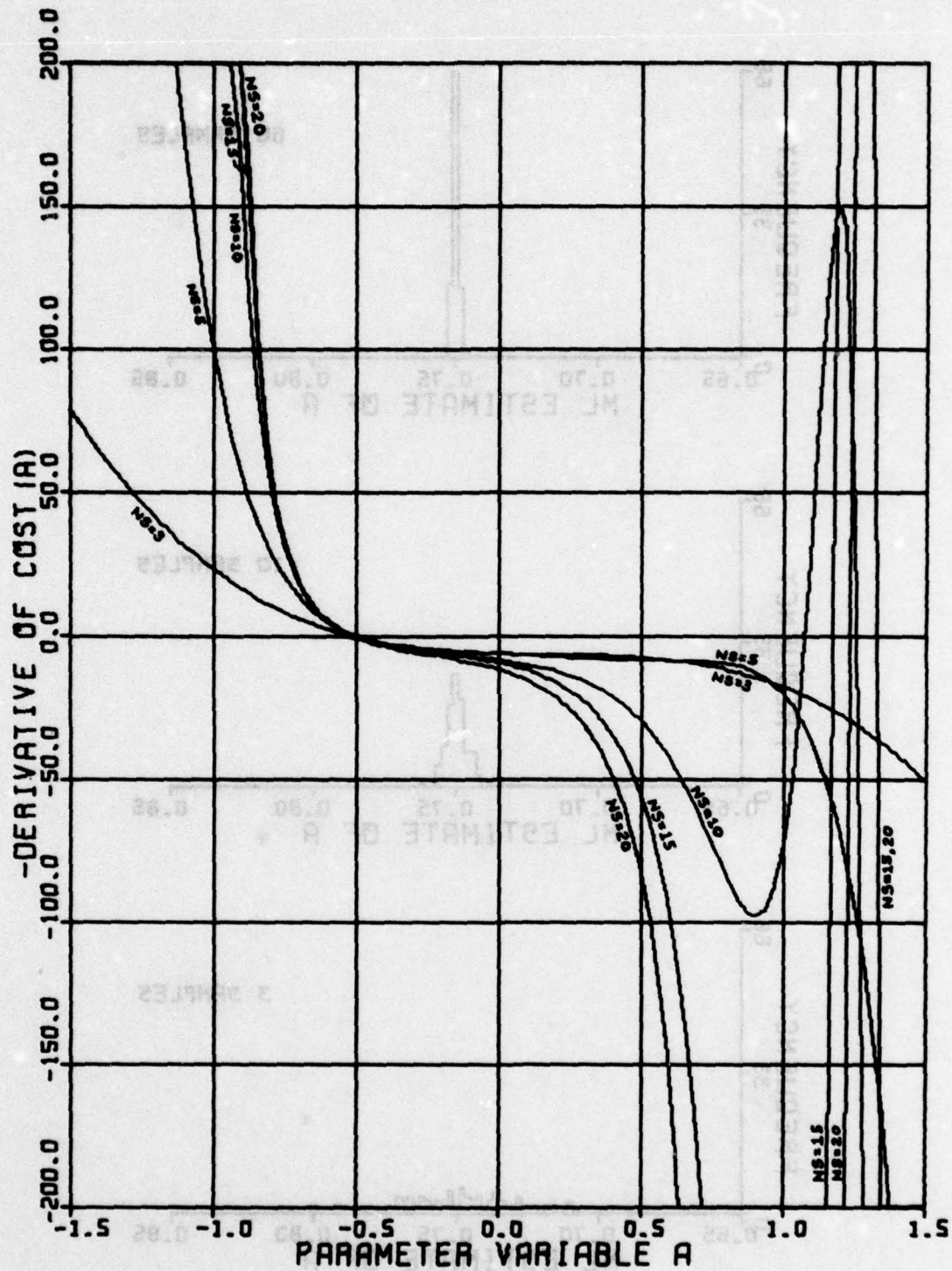


Figure 5-1h. No noise derivative function of the MLE in the x_0 known case for 3, 5, 10, 15, and 20 samples. ($a_0 = -0.5$, $\sigma^2 = 0$).

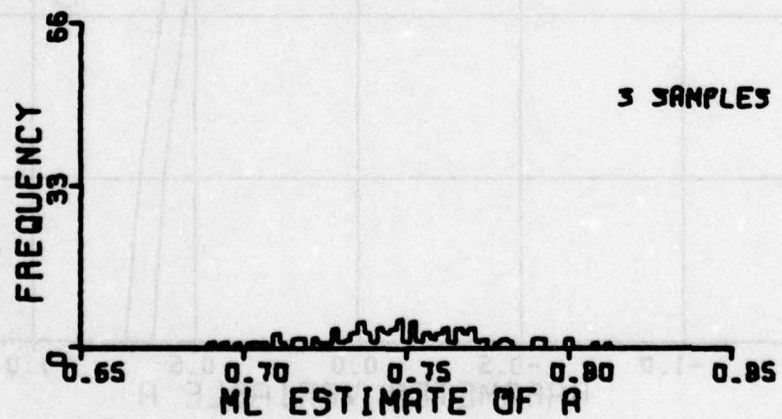
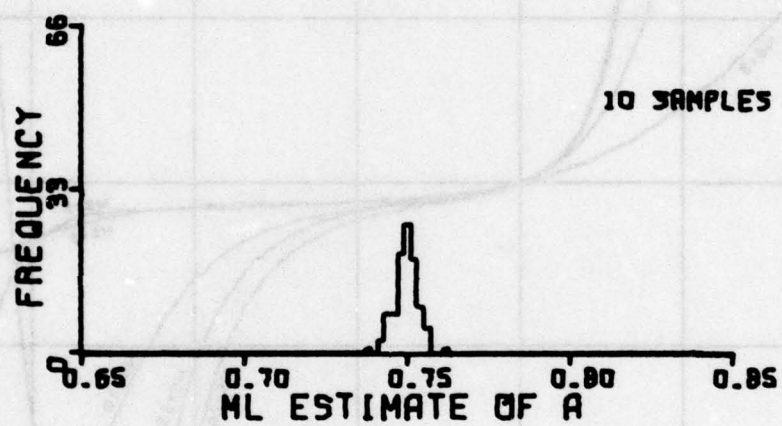
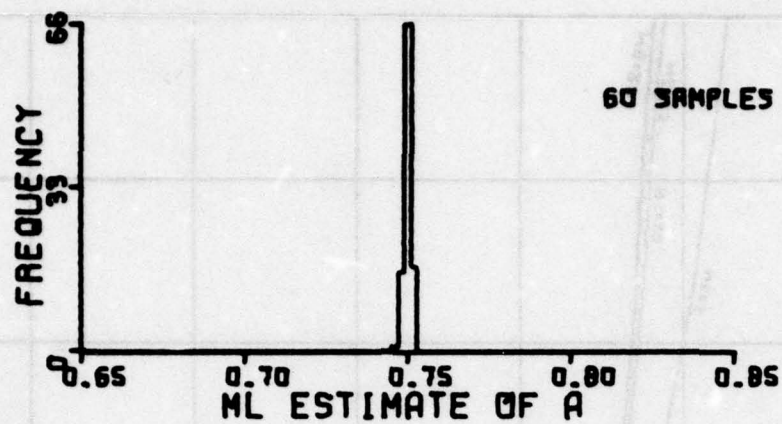


Figure 5-1j. Frequency distribution for MLE of a_0 in x_0 known case. ($a_0=0.75$, $\sigma^2=0.01$).

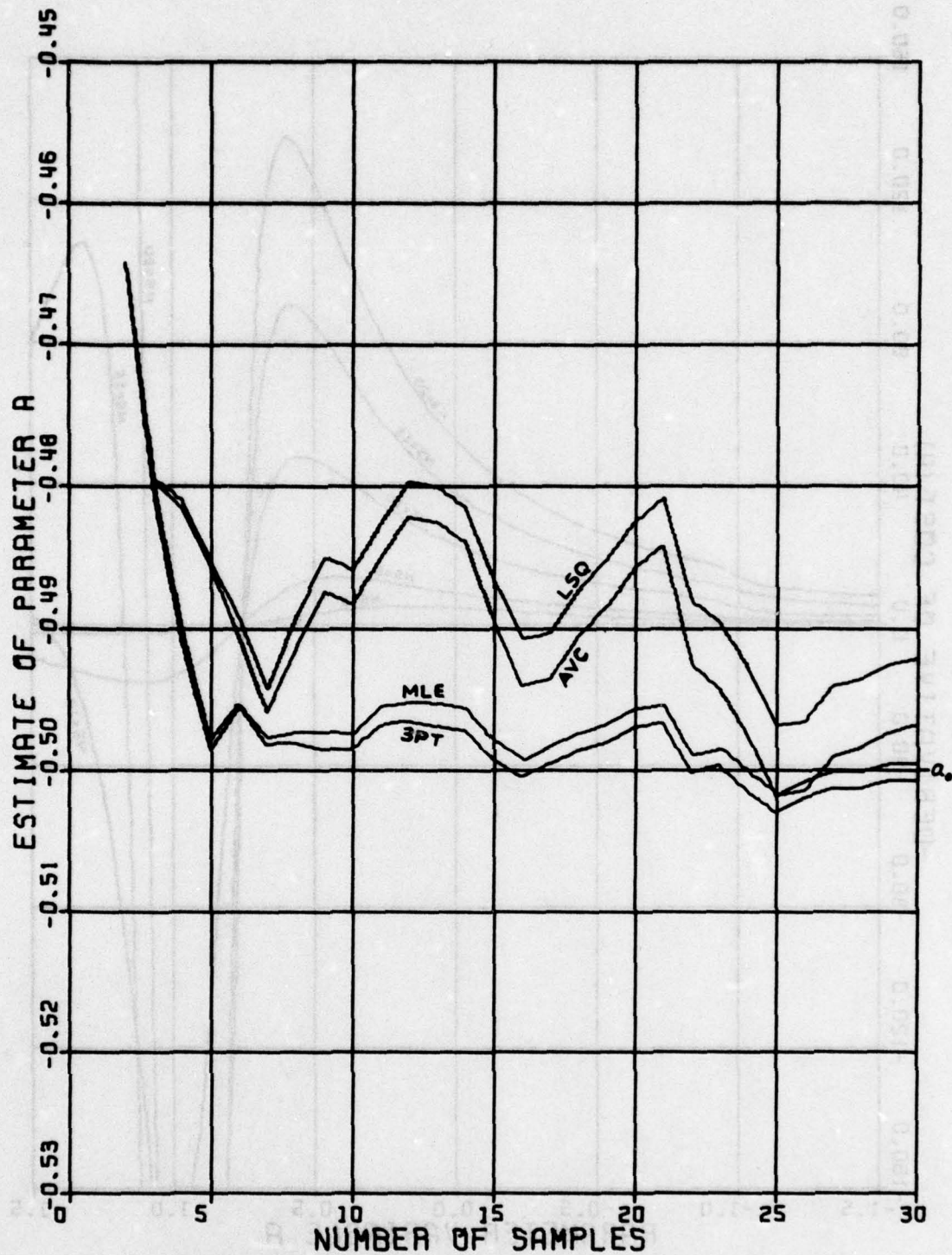


Figure 5-2a. Comparison of ML, least squares, average coefficient, and 3-point fit estimation for x_0 unknown parameter. ($a_0 = -0.5$, $\sigma^2 = 0.01$).

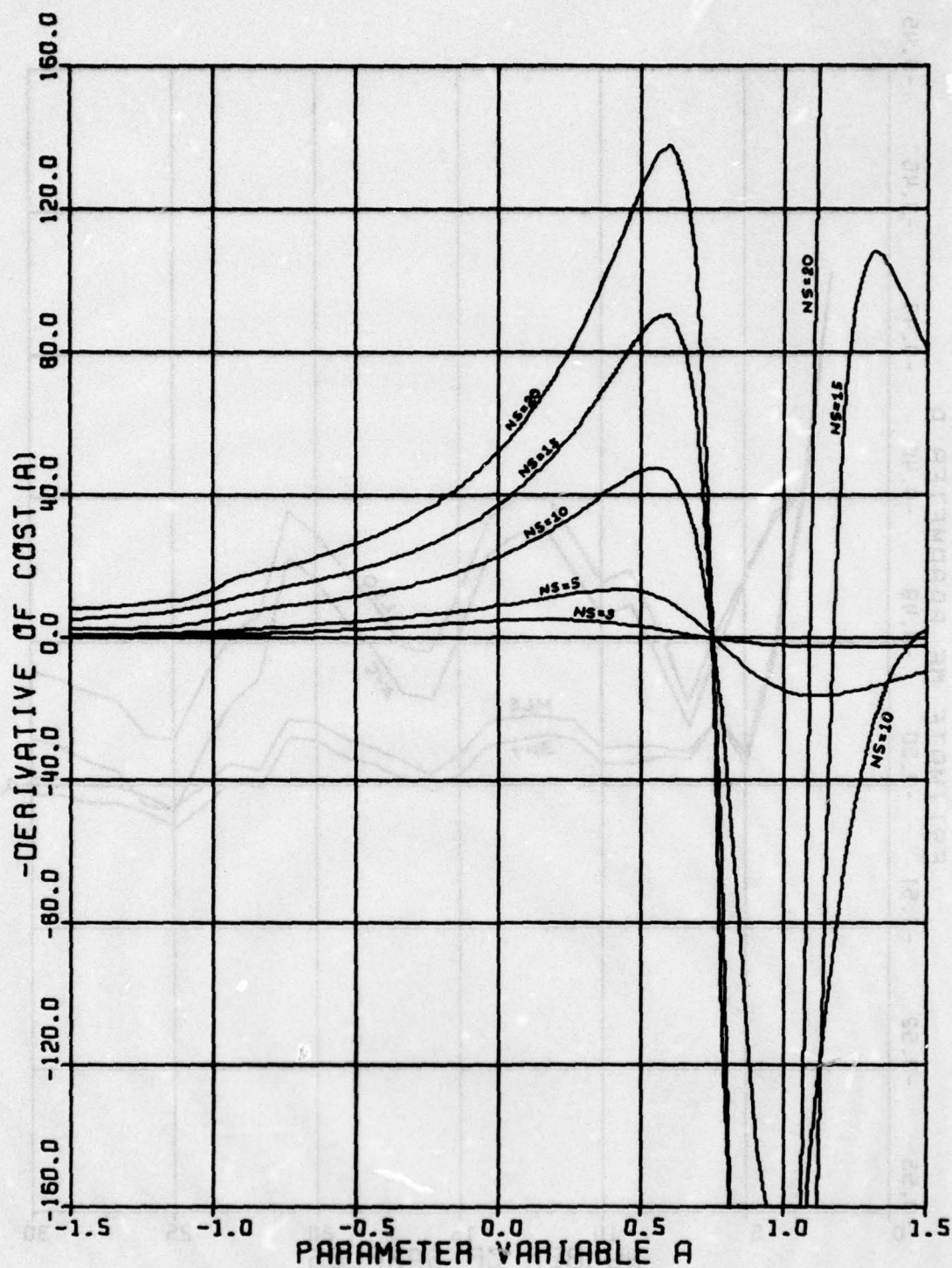


Figure 5-2f. Derivative function of the MLE in the x_0 unknown parameter case for 3, 5, 10, 15, and 20 samples. ($a_0=0.75$, $\sigma^2=0.01$).

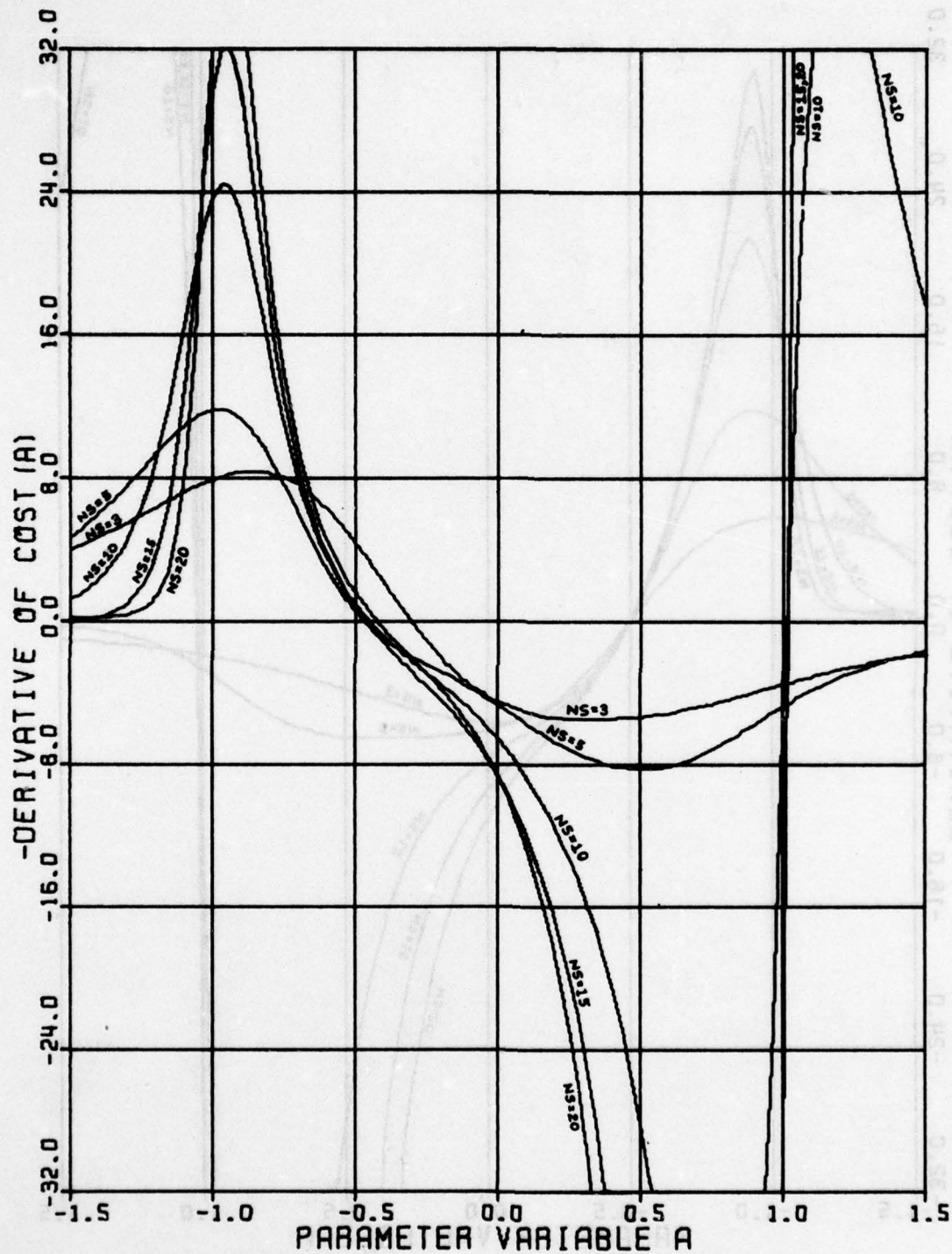


Figure 5-2g. Derivative function of the MLE in the x_0 unknown parameter case for 3, 5, 10, 15, and 20 samples. ($a_0 = -0.5$, $\sigma^2 = 1.0$).

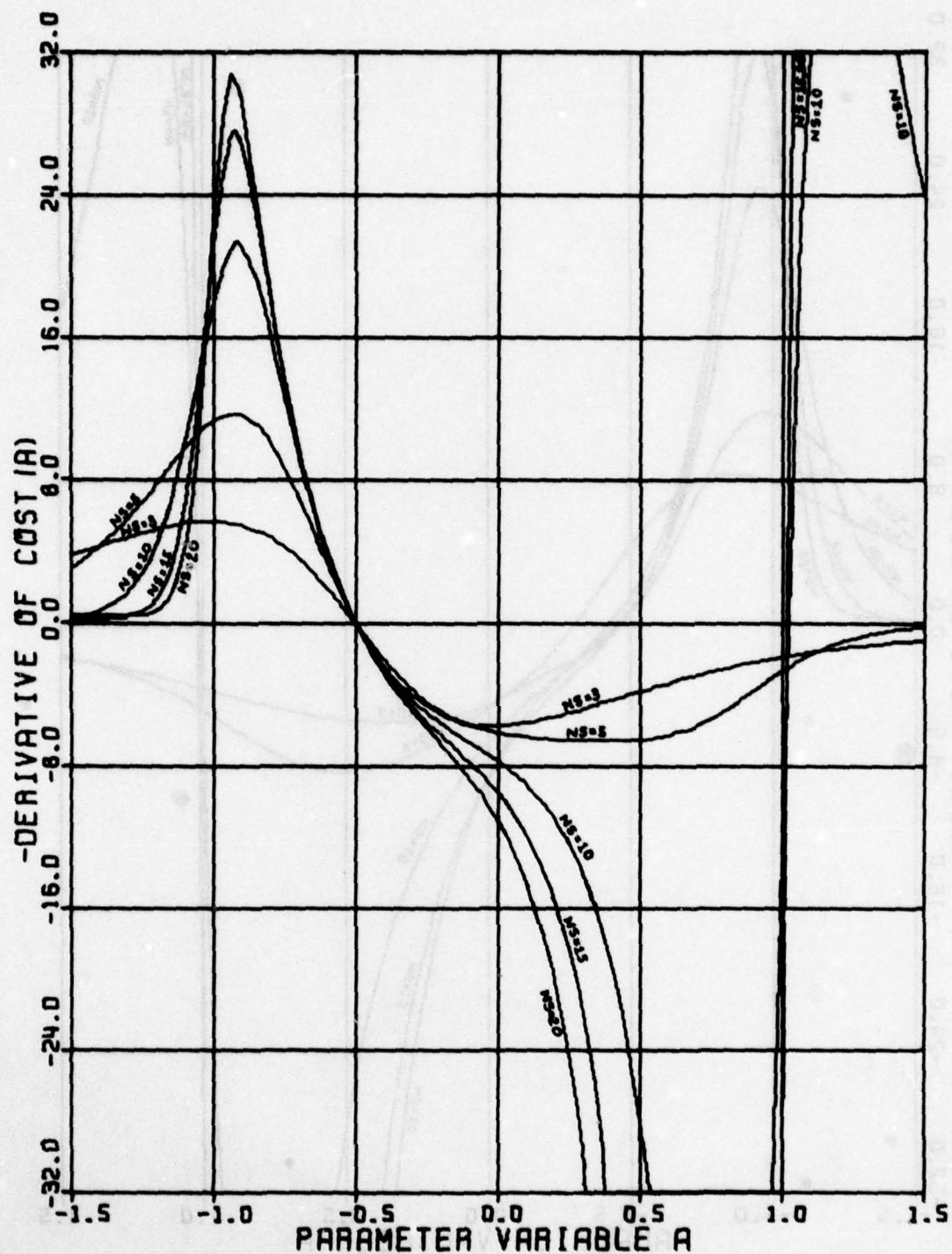


Figure 5-2h. No noise derivative function of the MLE in the x_0 unknown parameter case for 3, 5, 10, 15, and 20 samples. ($a_0 = -0.5$, $\sigma^2 = 0$).

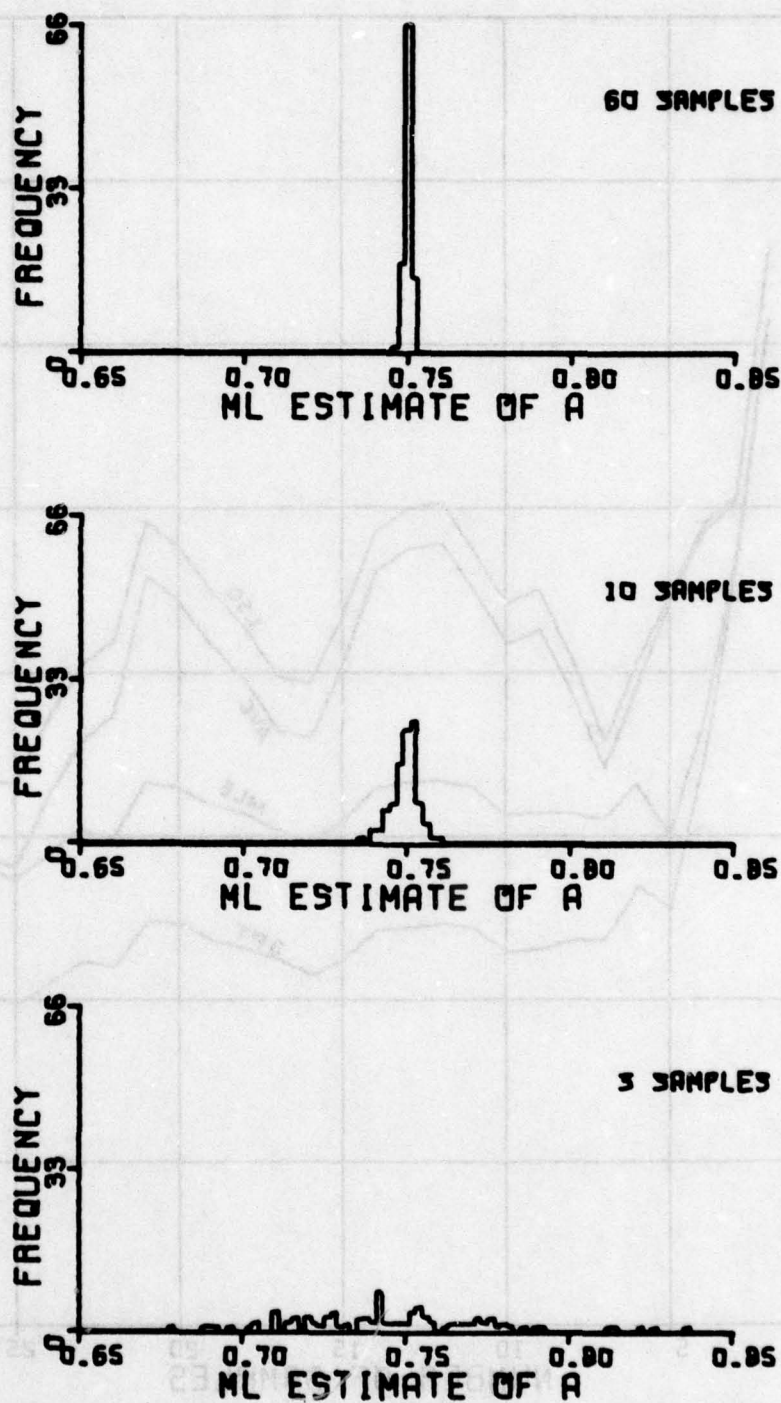


Figure 5-2j. Frequency distribution for MLE of a_0 in x_0 unknown parameter case. ($a_0=0.75$, $\sigma^2=0.01$).

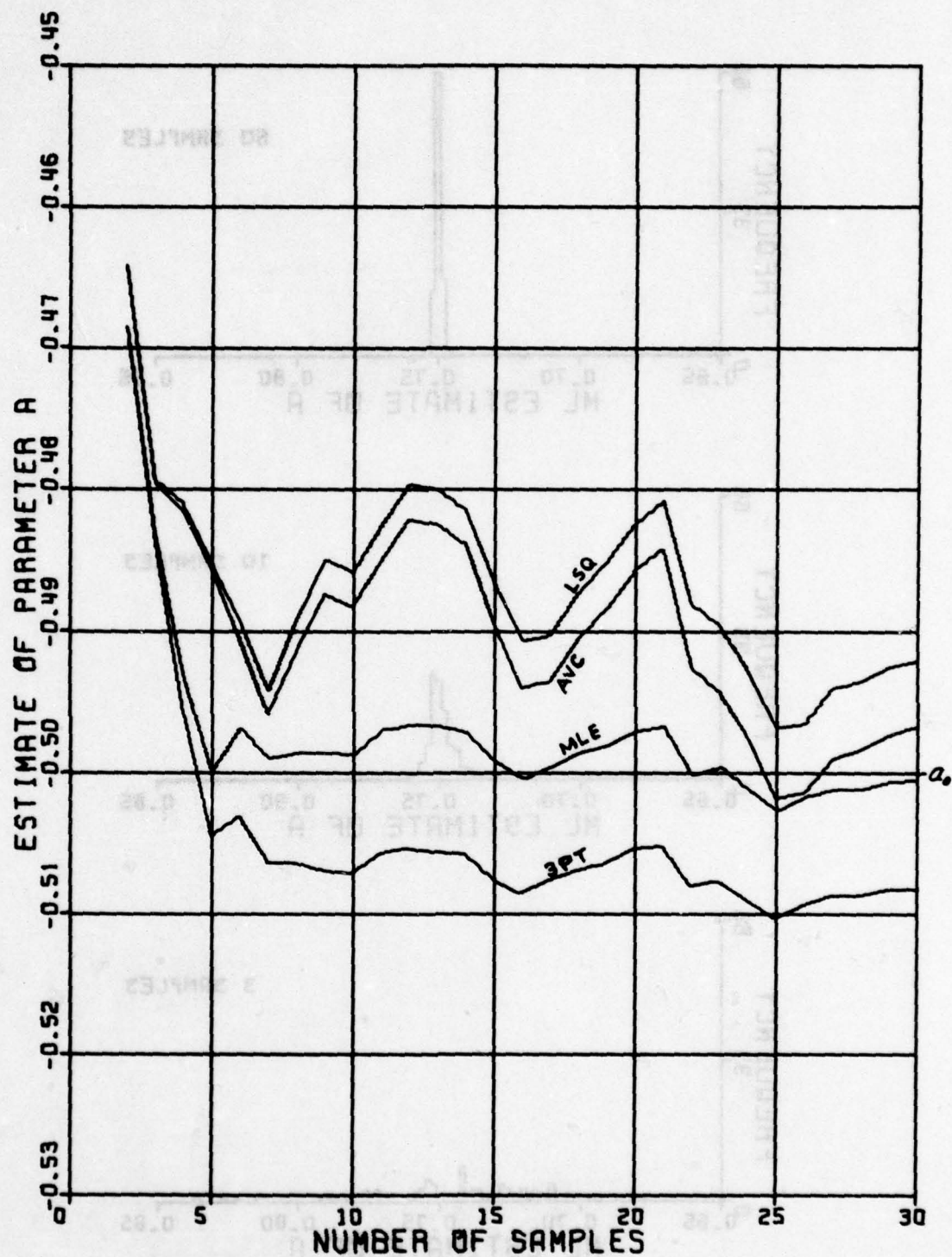


Figure 5-3a. Comparison of ML, least squares, average coefficient, and 3-point fit estimation for x_0 random variable. ($a_0 = -0.5$, $\sigma^2 = 0.01$).

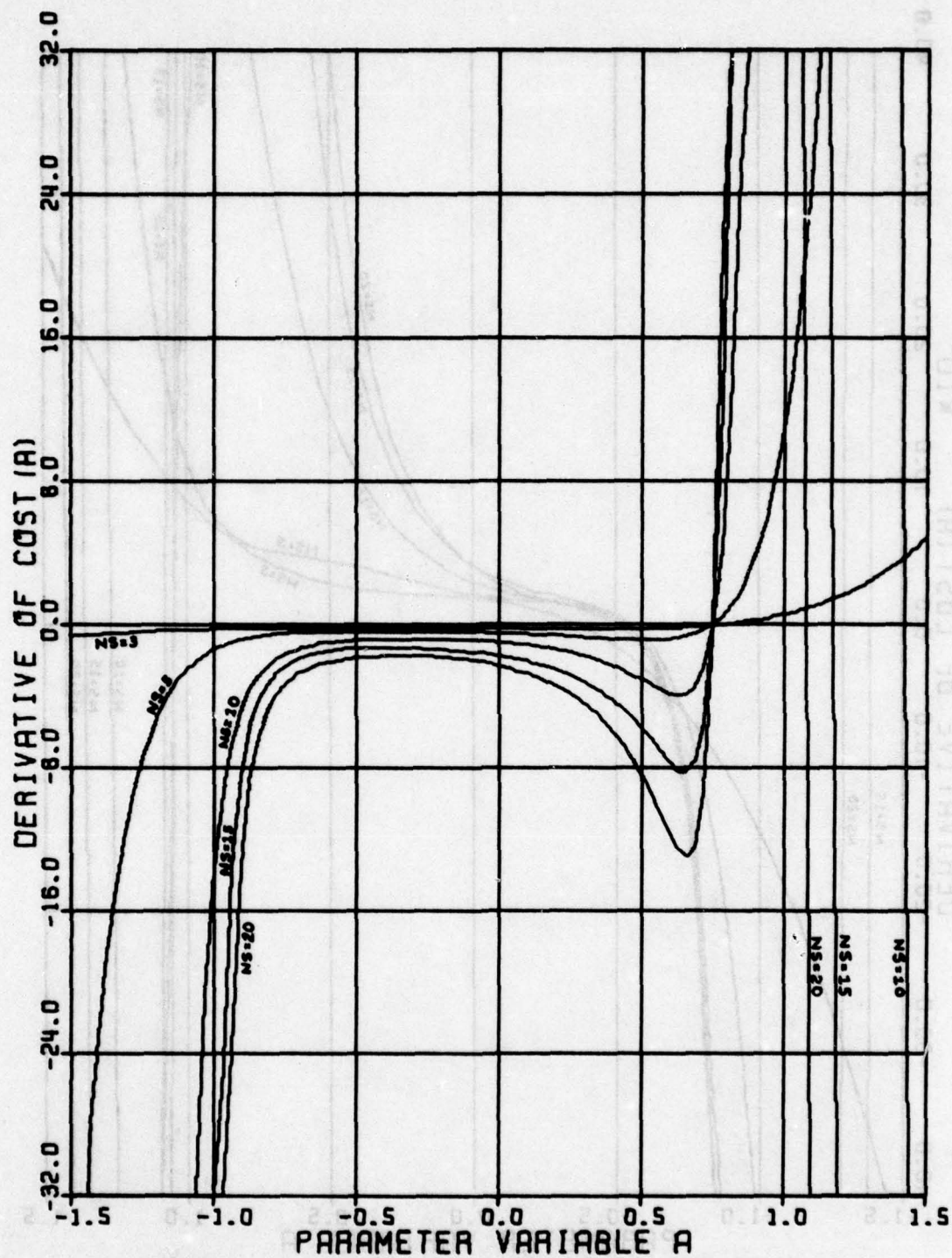


Figure 5-3f. Derivative function of the MLE in the x_0 random variable case for 3, 5, 10, 15, and 20 samples. ($a_0=0.75$, $\sigma^2=0.01$).

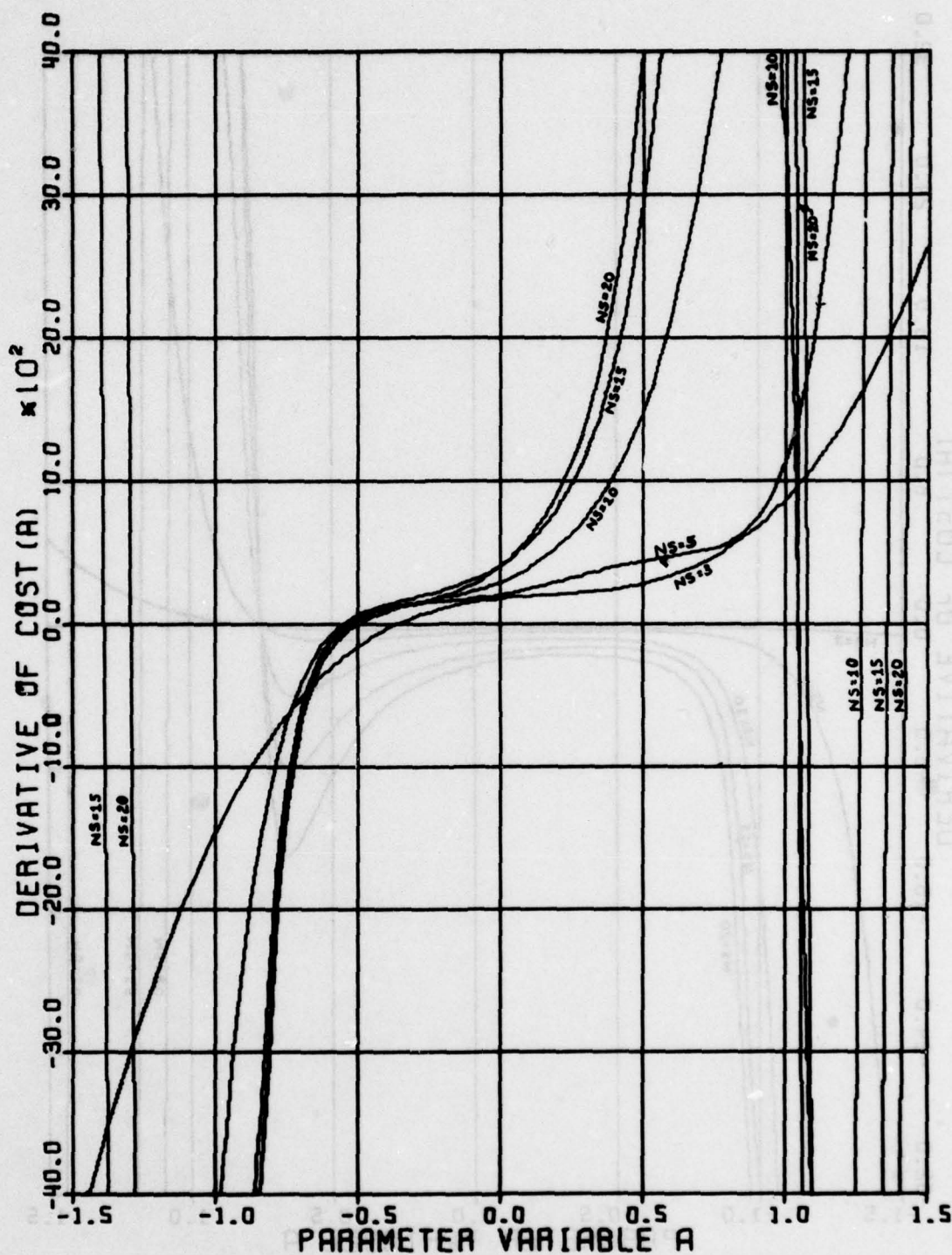


Figure 5-3g. Derivative function of the MLE in the x_0 random variable case for 3, 5, 10, 15, and 20 samples. ($a_0 = -0.5$, $\sigma^2 = 1.0$).

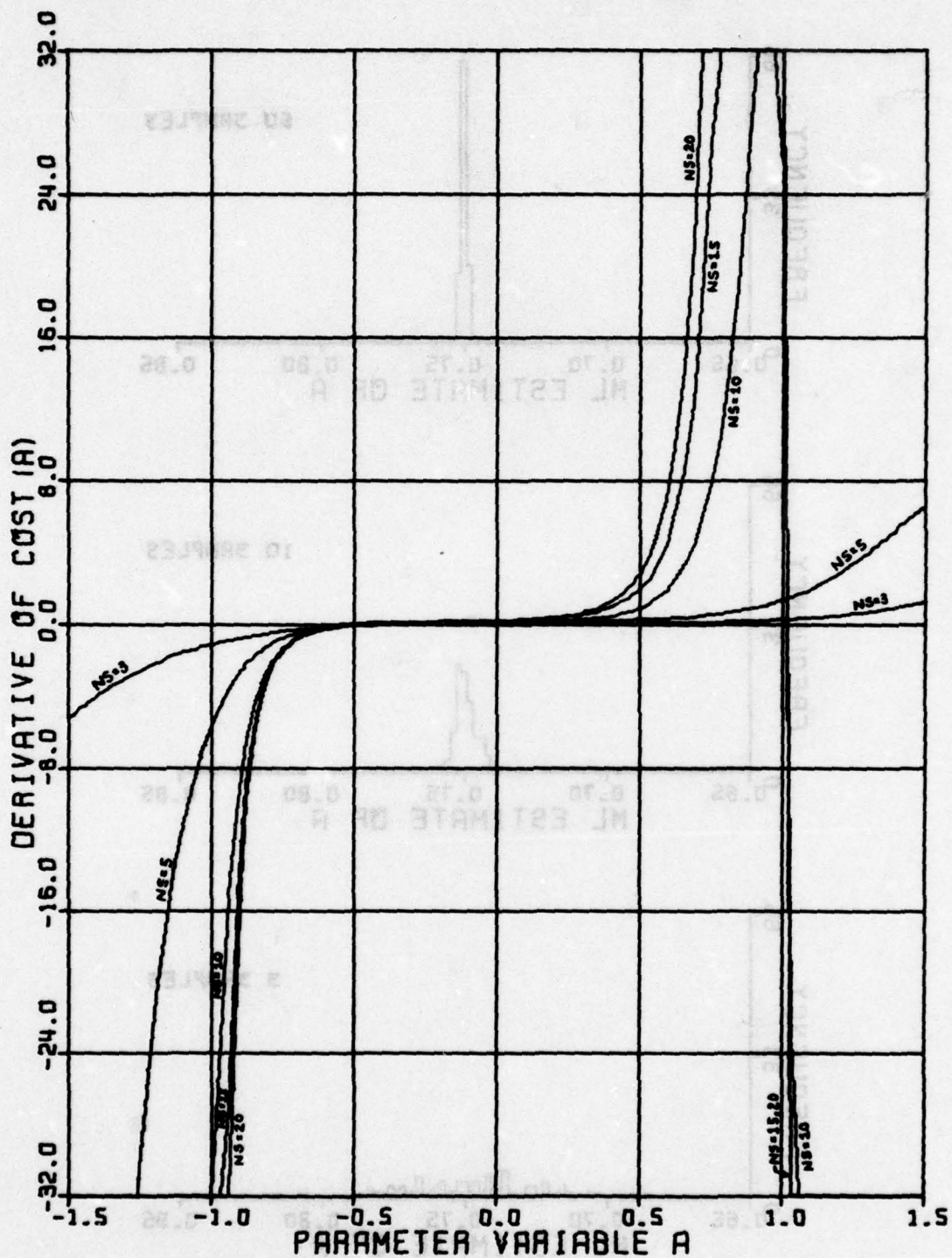


Figure 5-3h. No noise derivative function of the MLE in the x_0 random variable case for 3, 5, 10, 15, and 20 samples. ($a_0 = -0.5$, $\sigma^2 = 0$).

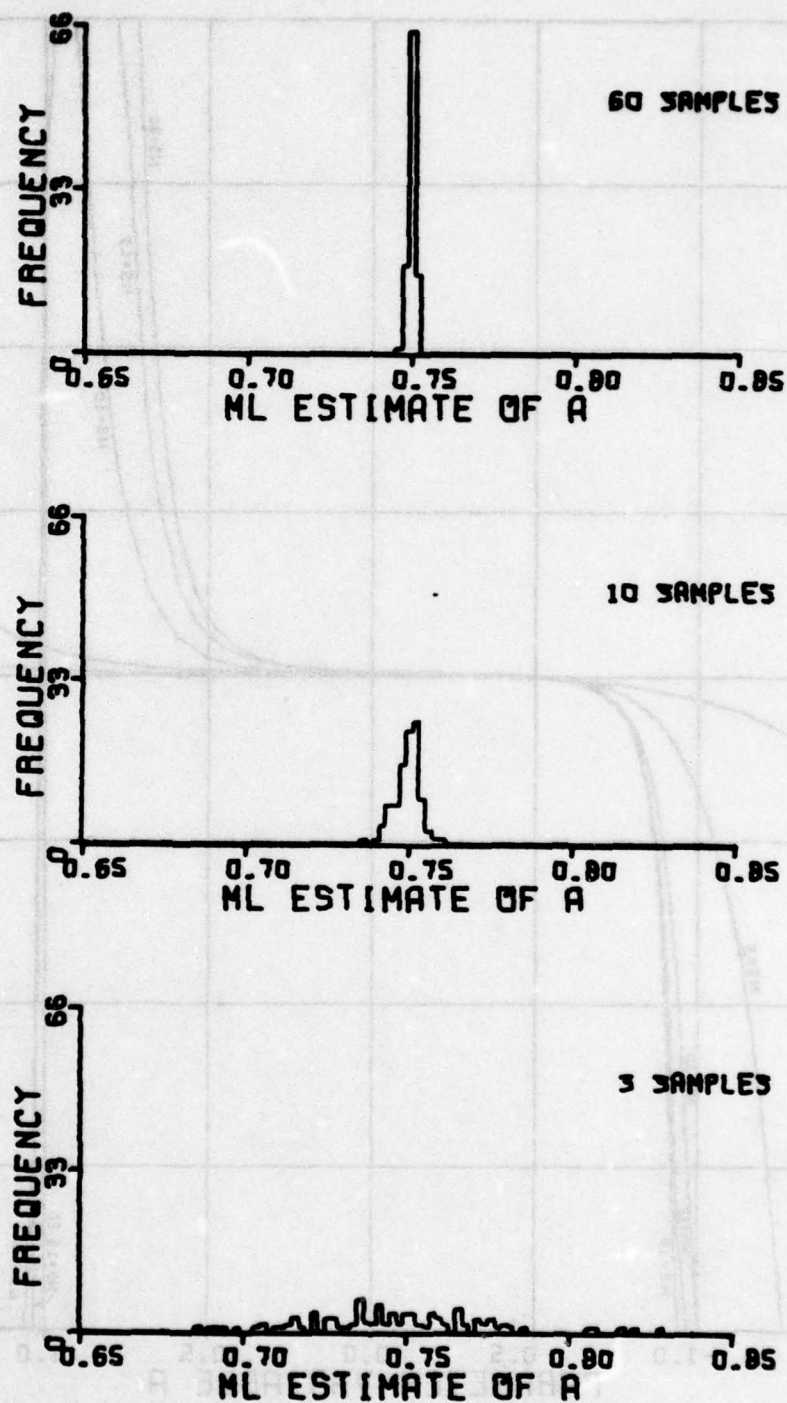


Figure 5-3j. Frequency distribution for MLE of a_0 in x_0 random variable case. ($a_0=0.75$, $\sigma^2=0.01$).

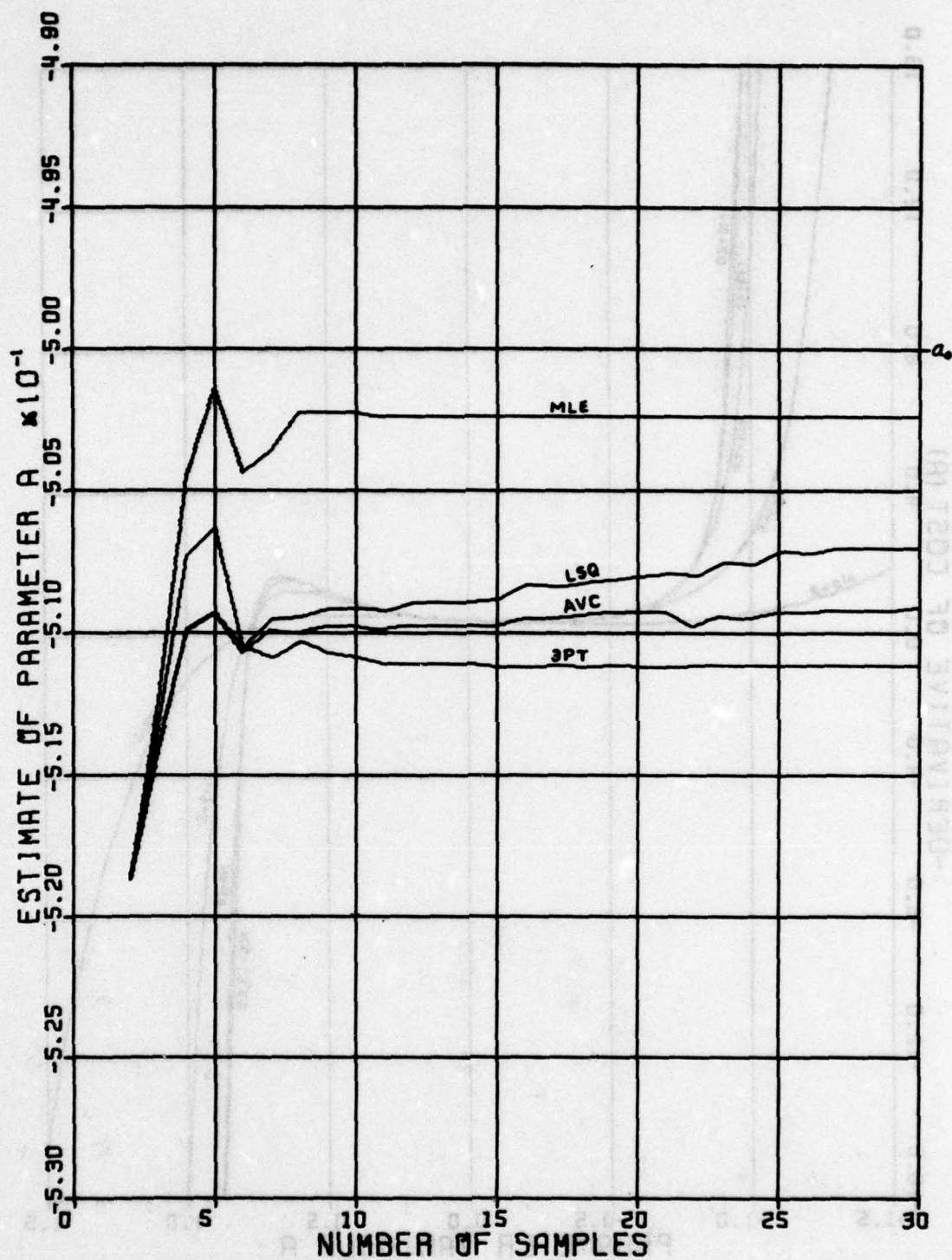


Figure 5-4a. Comparison of ML, least squares, average coefficient, and 3-point fit estimation for differencing approach. ($a_0 = -0.5$, $\sigma^2 = 0.01$).

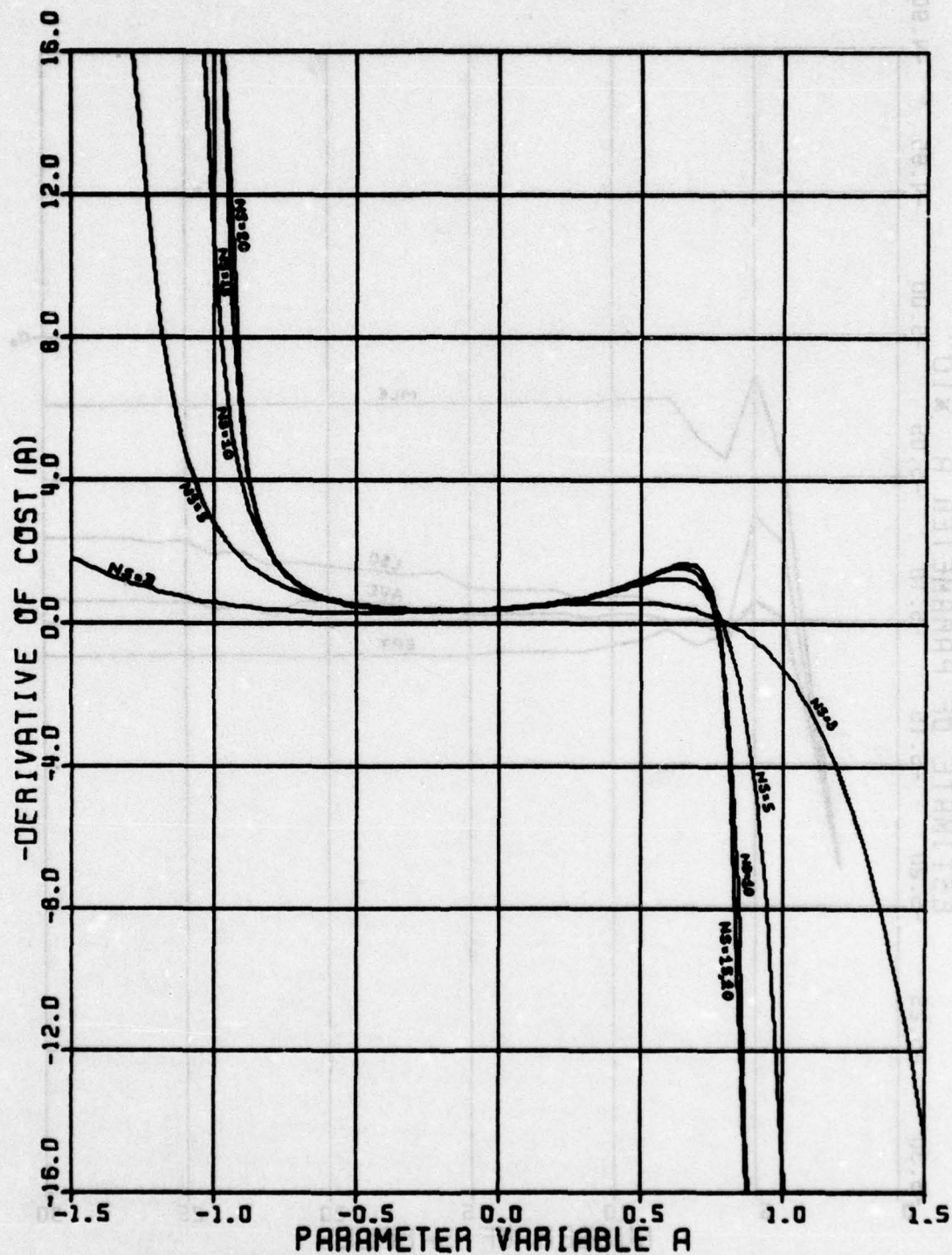


Figure 5-4f. Derivative function of the MLE for the differencing approach for 3, 5, 10, 15, and 20 samples. ($a_0=0.75$, $\sigma^2=0.01$).

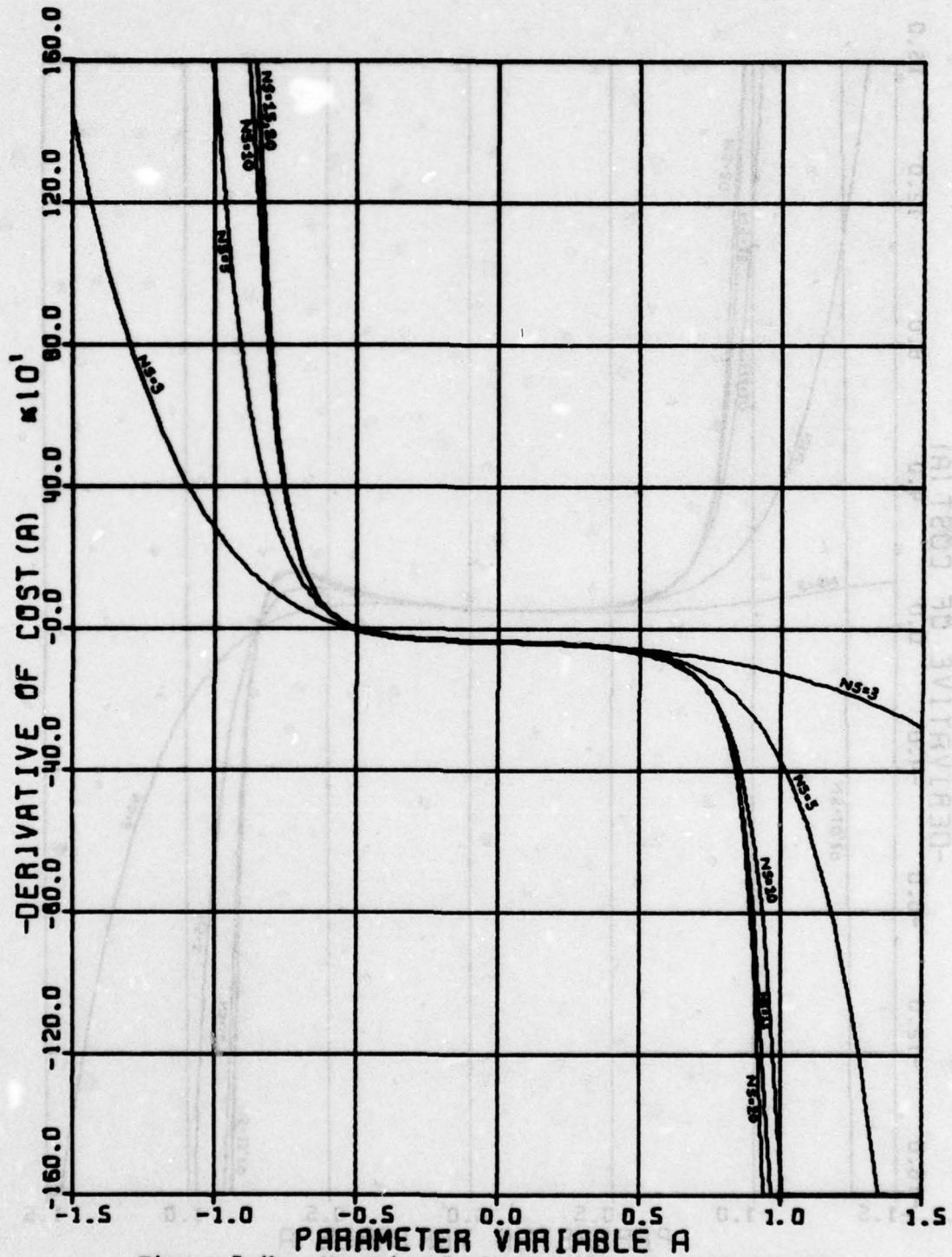


Figure 5-4h. No noise derivative function of the MLE in the differencing approach for 3, 5, 10, 15, and 20 samples. ($a_0 = -0.5$, $\sigma^2 = 0$).

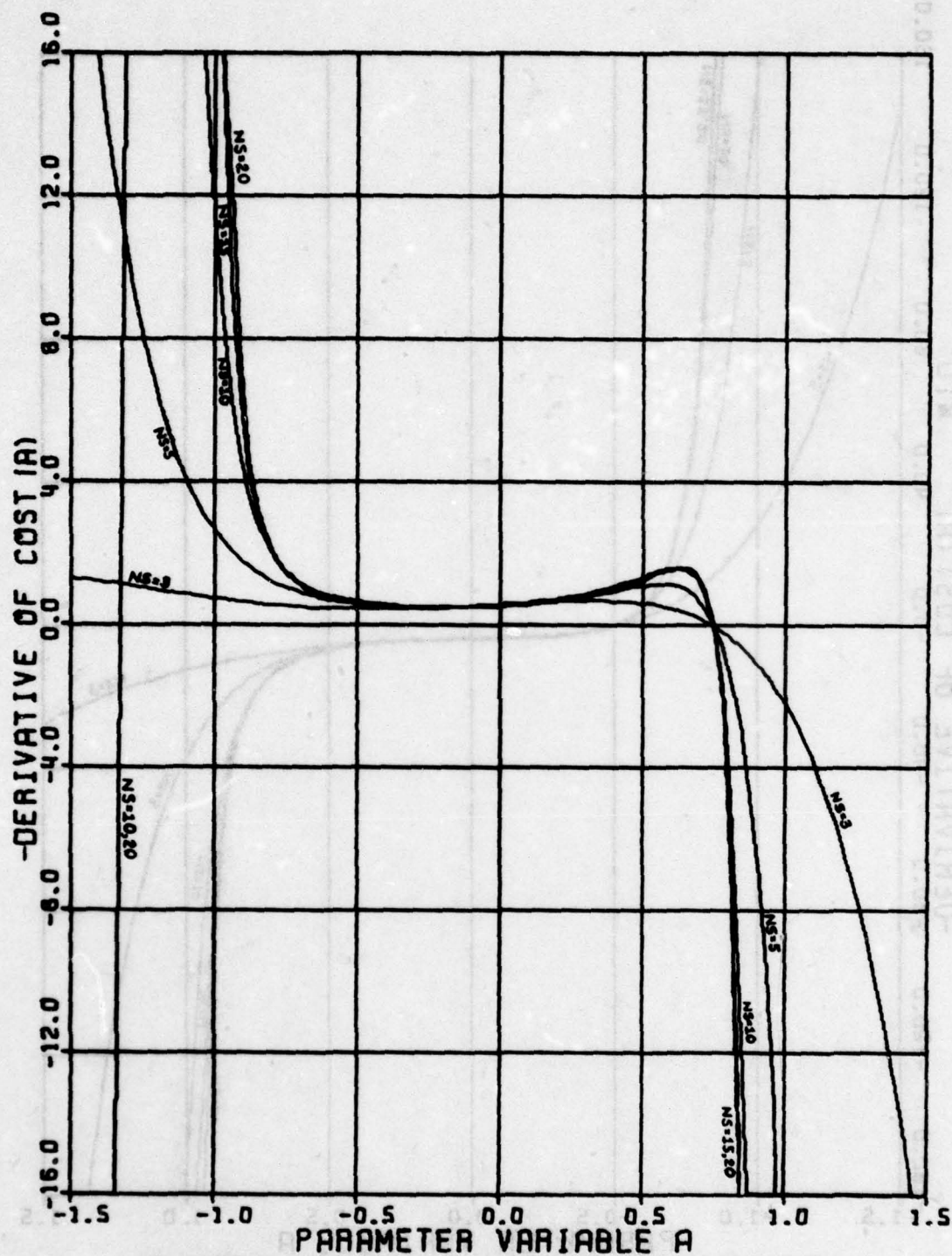


Figure 5-4i. No noise derivative function of the MLE in the differencing approach for 3, 5, 10, 15, and 20 samples. ($a_0=0.75$, $\sigma^2=0$).

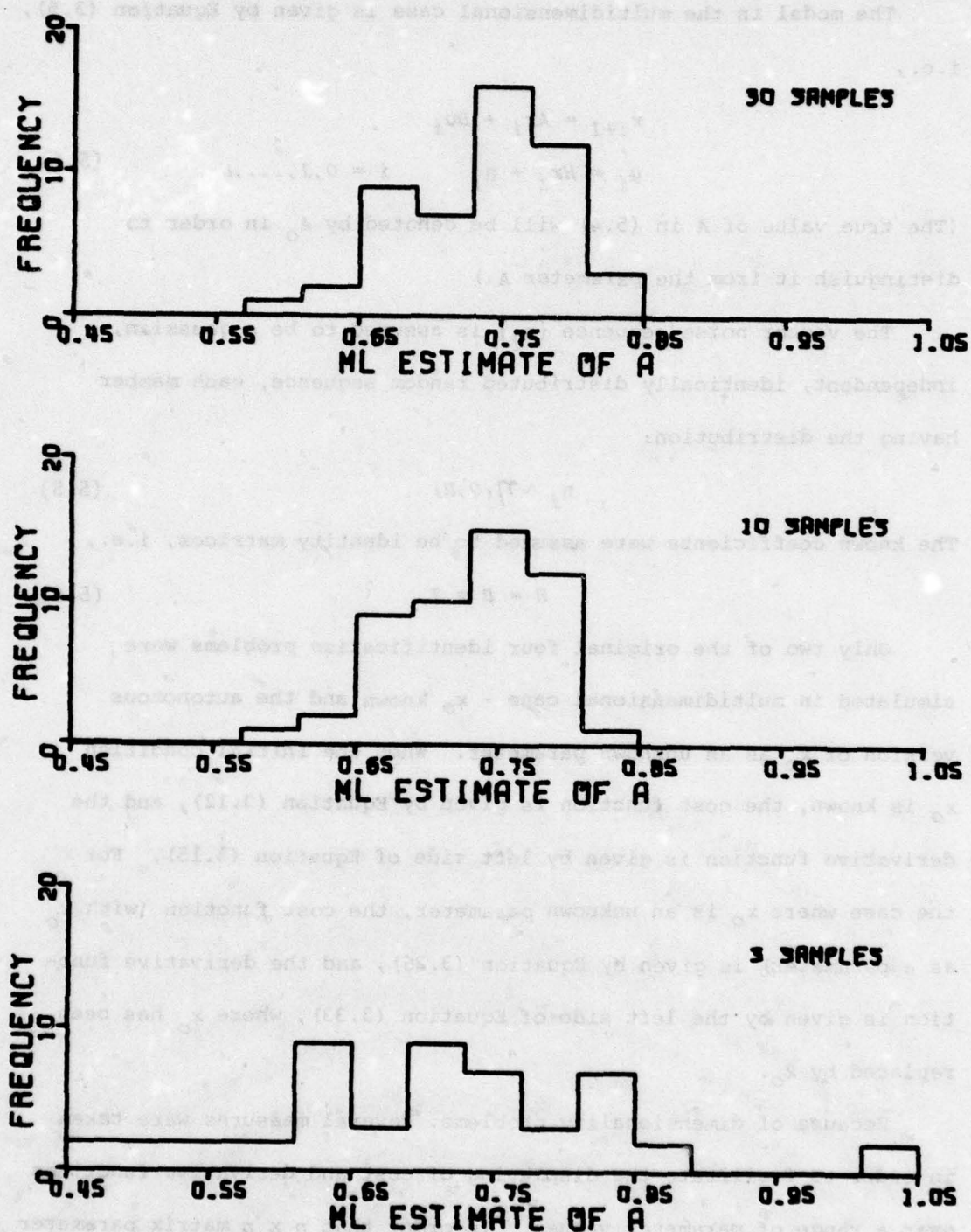


Figure 5-4j. Frequency distribution for MLE of a_0 in differencing approach. ($a_0=0.75$, $\sigma^2=0.01$).

5.2.2 MULTIDIMENSIONAL MODEL RESULTS

The model in the multidimensional case is given by Equation (3.5), i.e.,

$$\begin{aligned}x_{i+1} &= Ax_i + Bu_i \\y_i &= Hx_i + \eta_i \quad i = 0, 1, \dots, N\end{aligned}\tag{5.4}$$

(The true value of A in (5.4) will be denoted by A_0 in order to distinguish it from the parameter A .)

The vector noise sequence $\{\eta_i\}$ is assumed to be a gaussian, independent, identically distributed random sequence, each member having the distribution:

$$\eta_i \sim \mathcal{N}(0, R)\tag{5.5}$$

The known coefficients were assumed to be identity matrices, i.e.,

$$H = B = I\tag{5.6}$$

Only two of the original four identification problems were simulated in multidimensional case - x_0 known and the autonomous version of x_0 as an unknown parameter. When the initial condition x_0 is known, the cost function is given by Equation (3.12), and the derivative function is given by left side of Equation (3.15). For the case where x_0 is an unknown parameter, the cost function (with x_0 as a parameter) is given by Equation (3.26), and the derivative function is given by the left side of Equation (3.33), where x_0 has been replaced by $\hat{x}_0$.

Because of dimensionality problems, several measures were taken in order to facilitate the displaying of cost and derivative functions over a range of parameter values. In order that $n \times n$ matrix parameter argument A of the functions be represented by a scalar, the cost

functions and derivative functions were evaluated only along a vector in n^2 dimensional space passing through A_0 . The derivative functions are also $n \times n$ matrices. Two scalar representations of these were computed. One is the derivative norm which is the sum of the magnitudes of the elements of the derivative matrix. (This norm is related to one more commonly used, the element with largest magnitude. The latter is bounded by the former and the former $\pm n^2$.) The second scalar representation, the directional derivative, corresponds to the inner product of the derivative represented as a vector, i.e., the gradient vector, and the true parameter matrix A_0 represented as a vector.

Only the 2×2 dimensioned possibility was simulated. For the simulations:

$$A_0 = \begin{pmatrix} 0.1 & 0.3 \\ -0.6 & 1.0 \end{pmatrix} \text{ and } x_0 = \begin{pmatrix} 2.0 \\ 1.5 \end{pmatrix}$$

where the eigenvalues of A_0 are 0.4 and 0.7, and the eigenvectors are $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

The measurement noise covariance is,

$$R = \begin{pmatrix} 0.01 & 0.005 \\ 0.005 & 0.025 \end{pmatrix}$$

The input sequence for the x_0 known and x_0 unknown parameter, respectively, are,

$$u_i \equiv \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \equiv \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

The computations were made for $A = cA_0$ over the range $-1.5 \leq c \leq 1.5$, the approximate range over which A is stable.

Figures 5-5a, 5-5b, and 5-5c display the cost function, norm of the derivative, and directional derivative, respectively, along A_0 for 3, 5, 10, and 15 samples (and 20 samples for the cost function) in the case where x_0 is known. Figures 5-6a, 5-6b, and 5-6c give the cost function, norm of the derivative, and directional derivative, respectively, along A_0 for 3, 5, 9, 13, and 18 samples in the case where x_0 is an unknown parameter. However, the cost function in this latter case is not entirely consistent with the derivative plots because in computing the cost curves the true x_0 was used instead of $\hat{x}_0$ (see Equation (3.30)).

The figures for the multiparameter case follow.

$$A_0 = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.1 \\ 0.1 & 0.1 \end{bmatrix} \text{ and } x_0 = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}$$

where the eigenvalues of A_0 are 0.4 and 0.7, and the eigenvectors

$$v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } v_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

The measurement noise covariance is

$$R = \begin{bmatrix} 0.001 & 0.001 \\ 0.001 & 0.001 \end{bmatrix}$$

The input sequence for the x_0 known and x_0 unknown parameter, respectively, are

$$x_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } \hat{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

The computations were made for $A = 0.1$ over the range $-1.5 \leq x \leq 1.5$.

The approximate range over which $\hat{x}_0$ is stable.

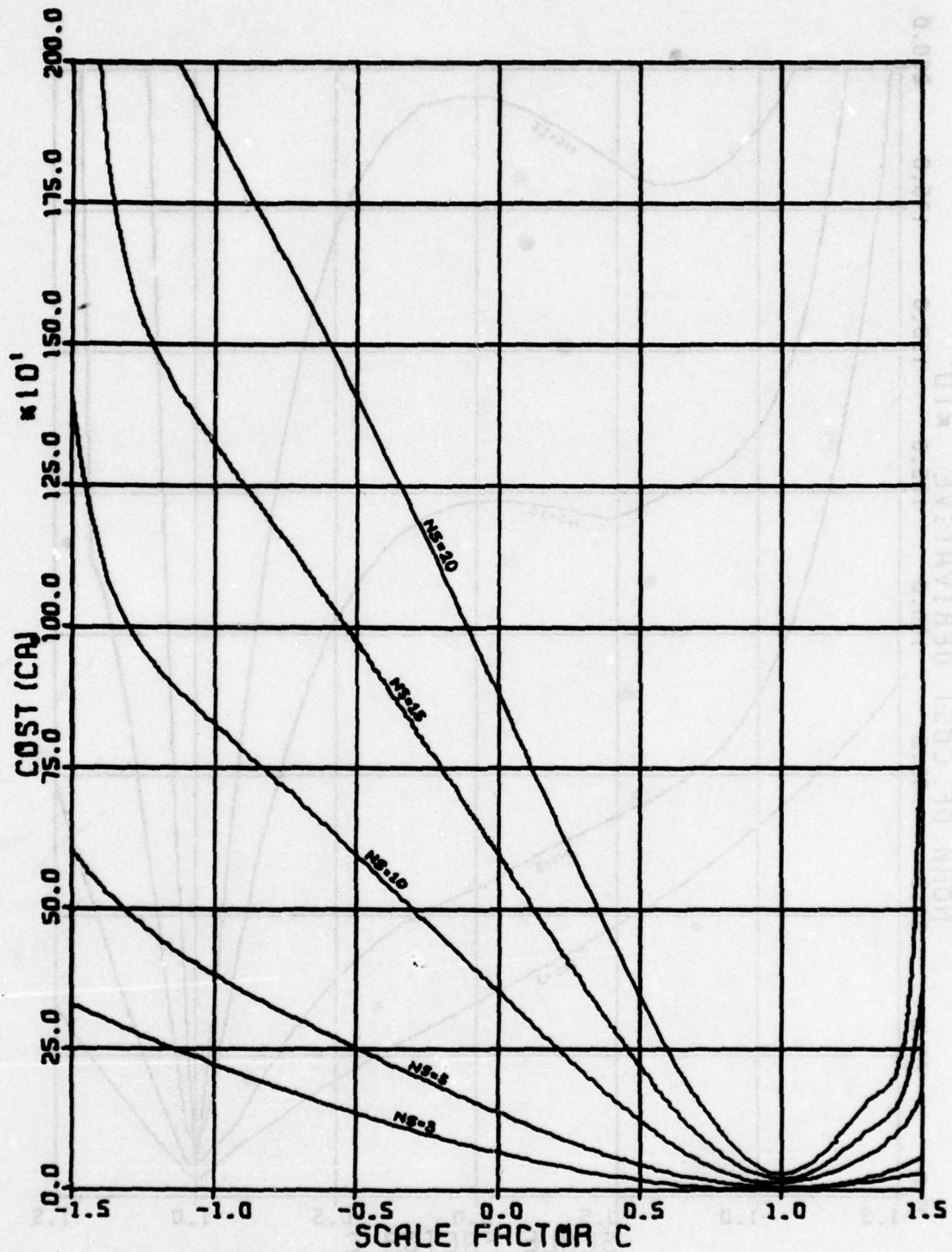


Figure 5-5a. Cost function along A_0 in x_0 known case for 3, 5, 10, 15, and 20 samples.

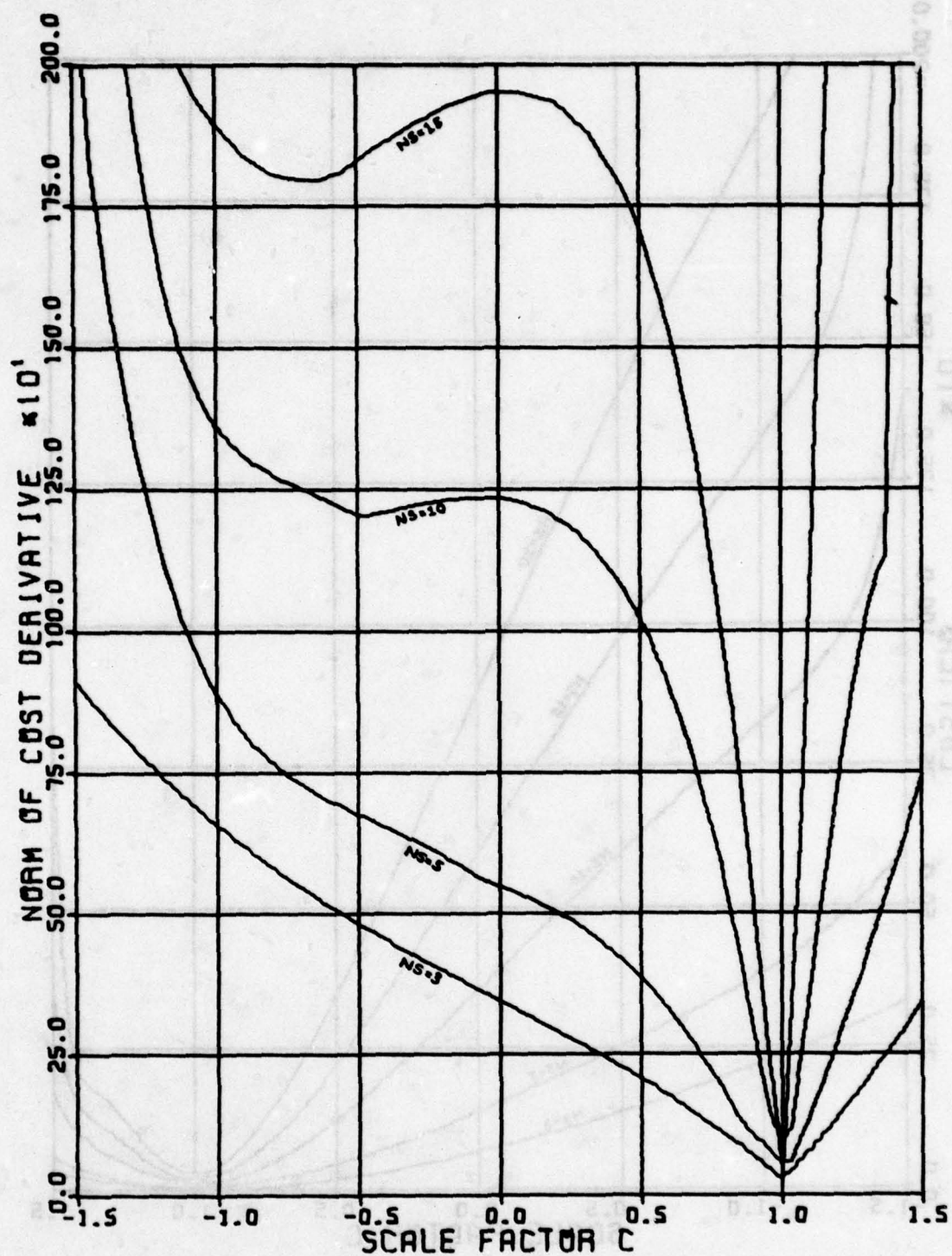


Figure 5-5b. Norm of derivative function along A_0 in x_0 known case for 3, 5, 10, and 15 samples.

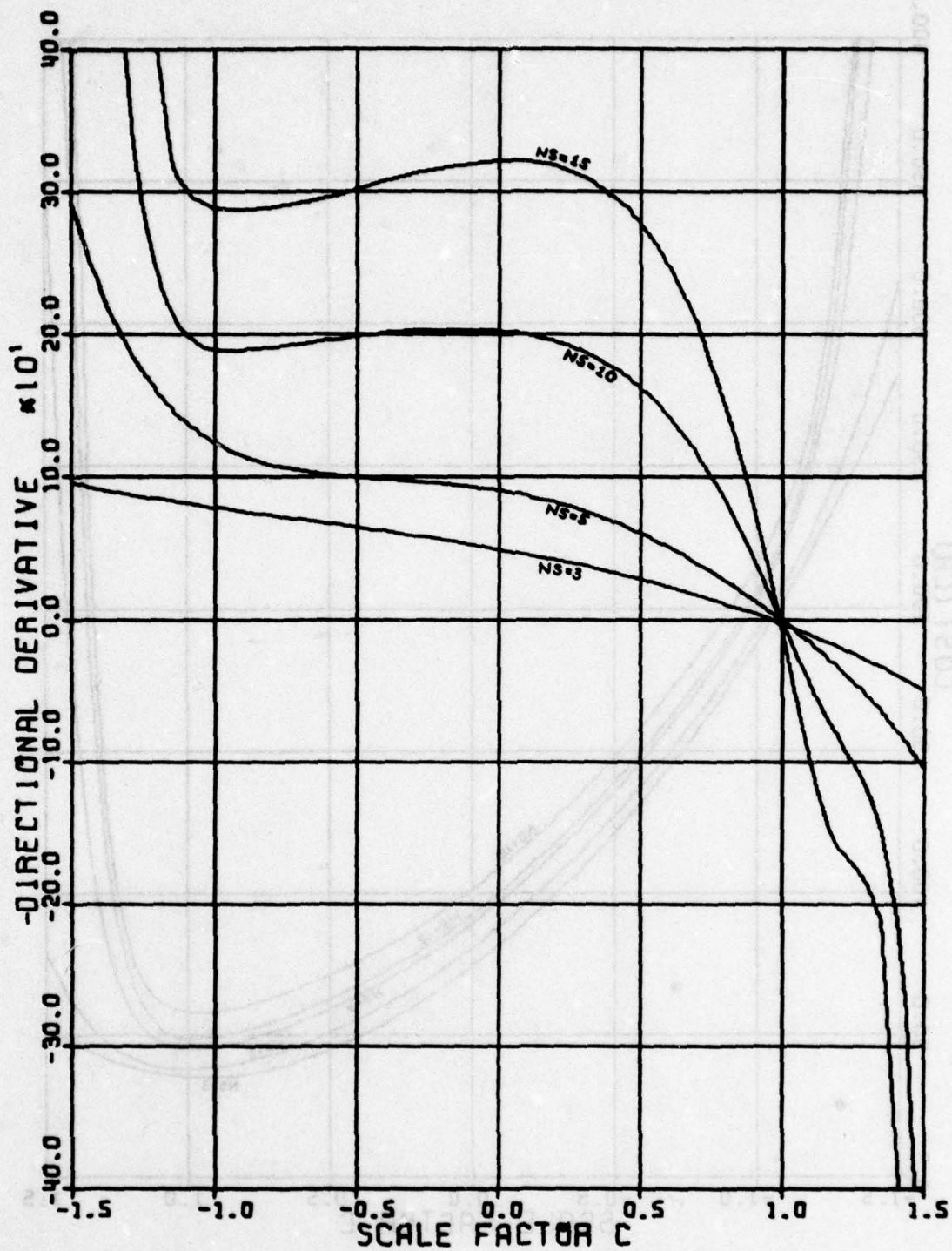


Figure 5-5c. Directional derivative along A_0 in x_0 known case for 3, 5, 10, and 15 samples.

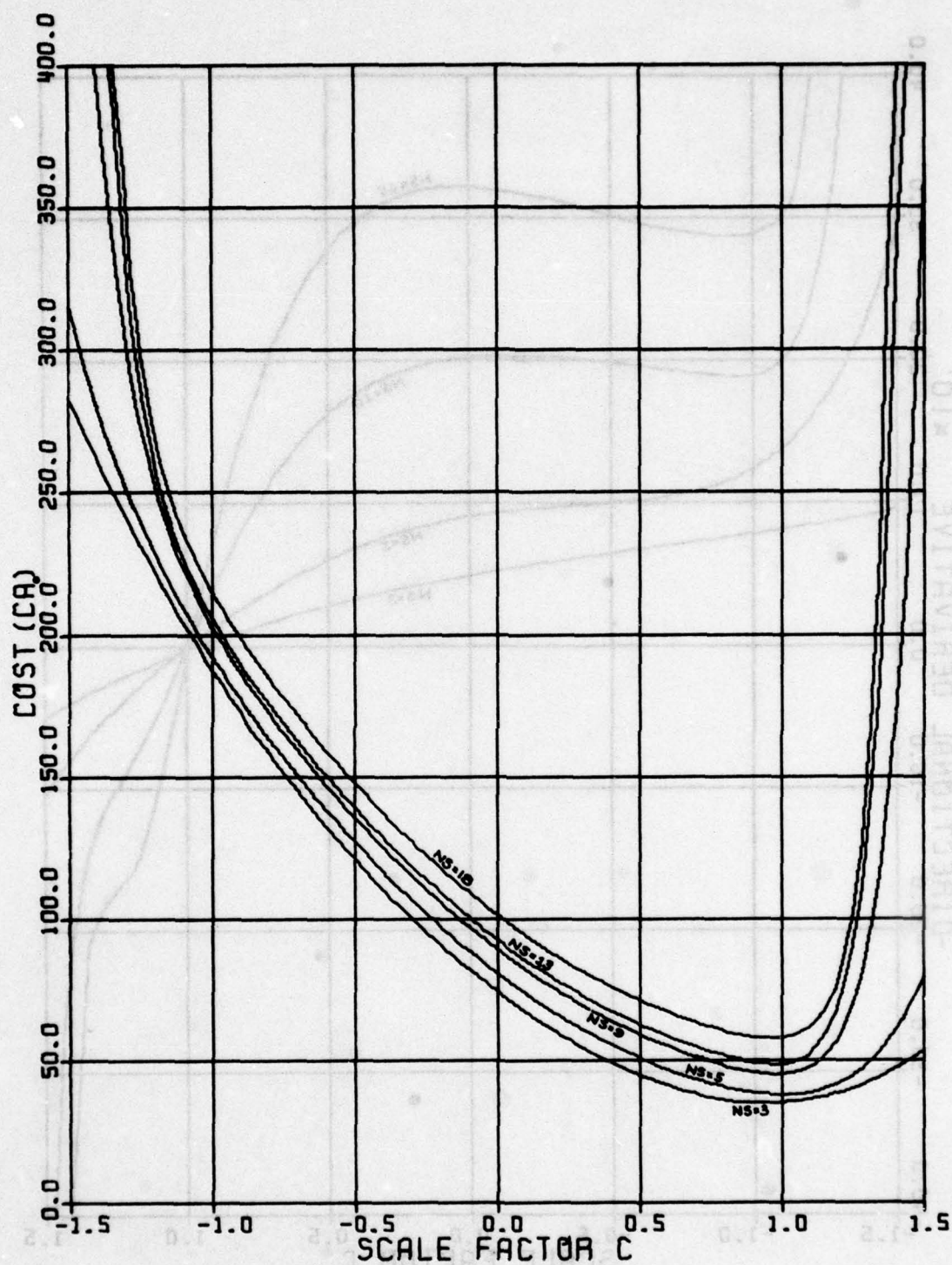


Figure 5-6a. Cost function along A_0 in x_0 unknown parameter case for true x_0 and 3, 5, 9, 13, and 18 samples.

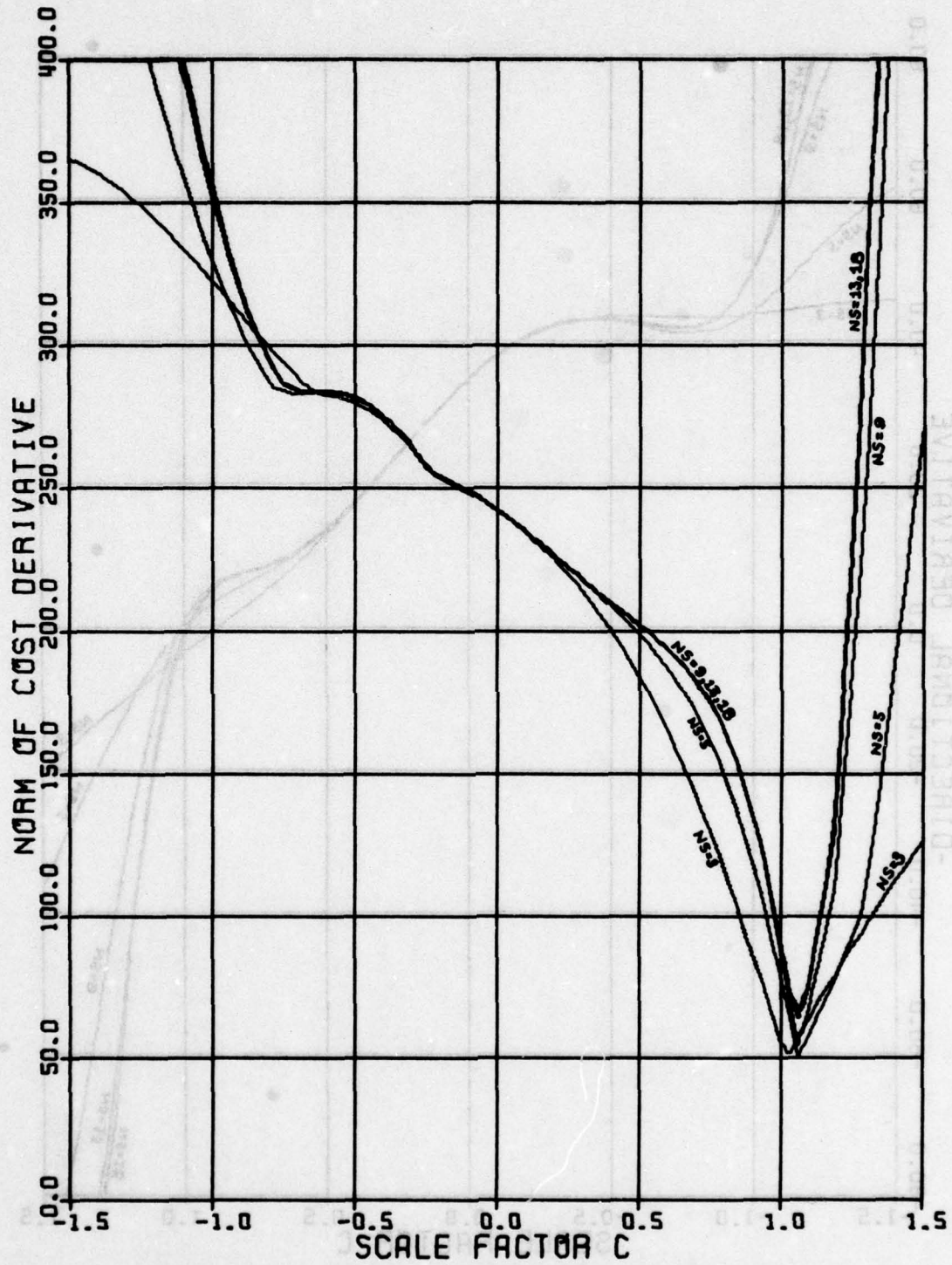


Figure 5-6b. Norm of the derivative function along λ_0 in x_0 unknown parameter case for $x_0 = \hat{x}_0$ and 3, 5, 9, 13, and 18 samples.

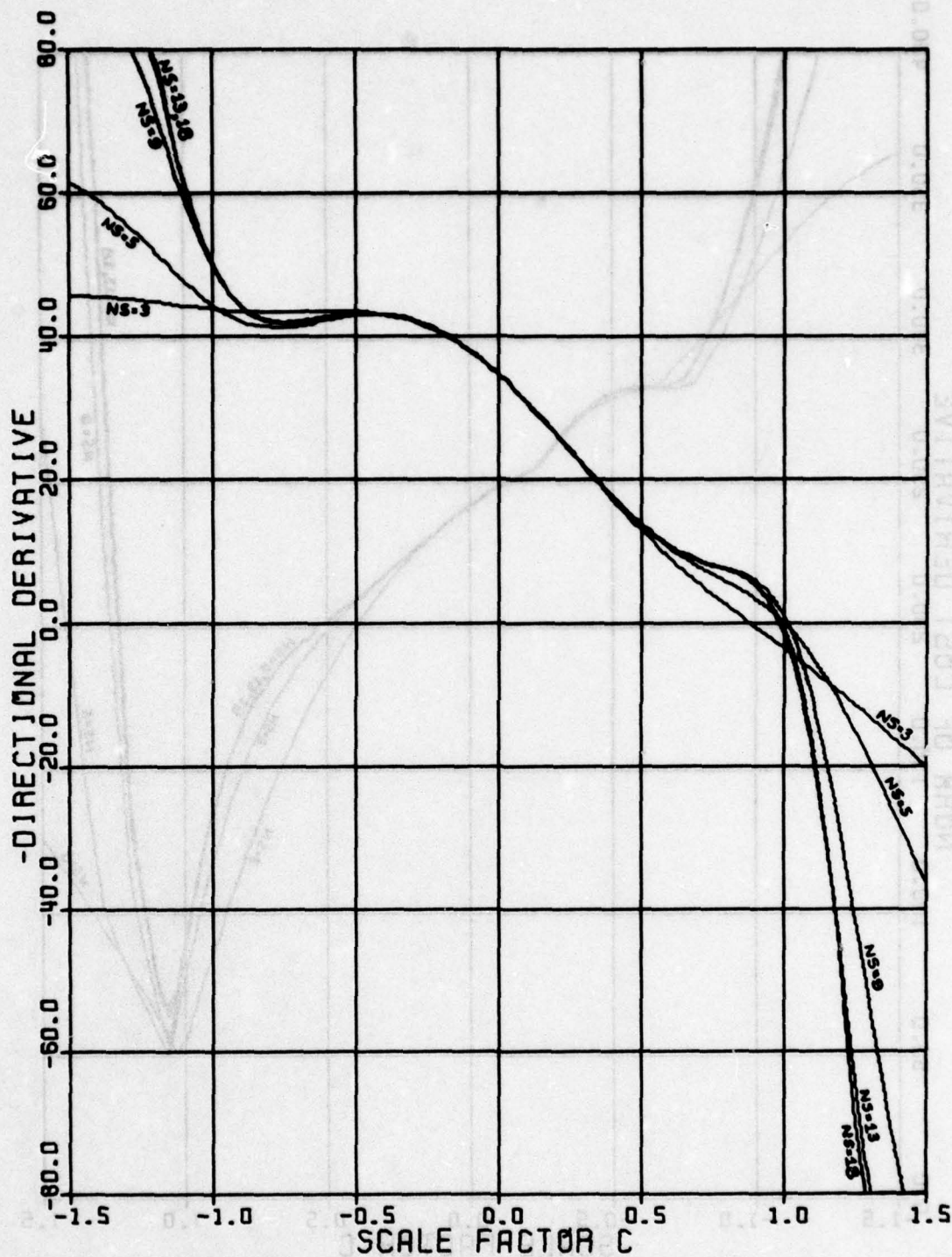


Figure 5-6c. Directional derivative along A_0
in x_0 unknown parameter case for $x_0 = \hat{x}_0$
and 3, 5, 9, 13, and 18 samples.

5.3 DISCUSSION OF MLE SIMULATION RESULTS

5.3.1 SCALAR MODEL

An important factor in the selection of numerical techniques to locate the roots of the likelihood equations is the behavior of the associated derivative function in the neighborhood of the desired root or better yet, over some area within which the root is nearly certain to lie. The no-noise derivative functions of groups (h) and (i) of the previous section provide some deterministic information on behavior but only in the limit as the noise variance goes to zero. However, the question of sensitivity of the behavior of the (polynomial) derivative functions to noise, i.e., to perturbations in its coefficients, must be answered before the usefulness of no-noise information in this regard can be assessed. Lacking well established conclusions on sensitivity, the investigation of derivative function behavior cannot be limited to the no-noise results.

The figures in groups (e), (f), and (g) provide some insight into derivative function behavior. (Because this study is concerned with stable systems, the figures display essentially only the stable range of the parameter a .) When $a, a_0 \in (-1, 1)$, the derivative function generally appears very smooth. In fact, for $a_0 = -0.5$, except when x_0 is an unknown parameter, Figures 5-1e, 5-1g, 5-3g, 5-4g indicate that the derivative function is basically monotone. (This probably would also have been the case when x_0 is an unknown parameter had Equation (3.22) been used in the simulation instead of (3.21) and (3.20).) The smoothness continues for parameter values somewhat less than -1 , but for those somewhat greater than $+1$, sharp fluctuations begin to

appear in groups (e), (f), and (g) as the number of samples increases. The derivative functions for the differencing approach, Figures 5-4f and 5-4h, do not exhibit the fluctuations near $a = 1$. (This may be related to the fact that the model is autonomous for the differencing approach simulations while for the three other cases, $u_i \equiv 1.0$.) The derivative fluctuations result from the dips that occur in the cost function just beyond $a = 1$ as illustrated by Figure 5-1e. (No attempt has been made to uncover any physical explanation for the dips.)

Comparing group (f) in which $a_0 = 0.75$ to group (g) (or group (e)) in which $a_0 = -0.5$, demonstrates that at least locally, the behavior of the derivative functions is significantly affected by a_0 . In group (f) the curves exhibit a bump adjacent to the root while for $a_0 = -0.5$ except for the case where x_0 is an unknown parameter, the curves tend to be monotone in form. This characteristic greatly detracts from the general applicability of Newton type numerical methods for root evaluations for stable a_0 .

The effect of noise level is less than might have been expected. As pointed out in the previous section, for relatively low noise level ($\sigma^2 = 0.01$) and the scales selected, the resulting curves (group (f)) are practically indistinguishable from the corresponding no-noise curves of group (h). (This conclusion is less accurate for the differencing approach when $a_0 = 0.75$ as shown by Figures 5-4f and 5-4i, and, of course, is inaccurate on a magnified scale as demonstrated by the curves of group (a).) Even for moderate noise levels ($\sigma^2 = 1.0$, $x_\infty = 2/3$), group (g) indicates the derivative functions undergo only relatively minor changes for stable a . (However, the spread of $\hat{a}$ does appear to change noticeably.)

As the number of samples increases, the magnitude of the slope generally increases with both roots and peaks becoming sharper. In the neighborhood of the root $\hat{a}$, in particular, the slope increase indicates a lowering sensitivity of $\hat{a}$ to noise perturbations. (See Section 5.5.) While this trend probably continues indefinitely, it likewise probably does not continue without bound for stable systems and $a \in (-1,1)$ because the derivative functions are polynomials in a . Since the measurement noise has a gaussian distribution, the y_i and thus the slope at $\hat{a}$ cannot be bounded in a deterministic sense. Still, the slope is most likely bounded in probability if not with probability one.

Because the derivative functions are polynomials, the possibility of multiple real roots must be faced. The question of multiple roots takes on added importance as the number of samples increases because the degree of the derivative function polynomial grows with the number of measurements. The figures of groups (e), (f), and (g) verify that multiple real roots can occur, but in none of the figures does more than one root fall in the interval $(-1,1)$. Furthermore, in the curves of groups (a), (b), and (c), in the derivative functions curves of groups (e), (f), and (g) and in the frequency histograms of group (j), $\hat{a}$ is always stable (except for one instance with only three samples in the differencing approach frequency distribution of Figure 5-4j). There, of course, is nothing preventing $\hat{a}$ from taking on unstable values if a_0 is stable. All that can be concluded from the above is that at least up to moderate noise levels, the spread in $\hat{a}$ distribution is relatively small.

The no-noise derivative functions curves of Figures 5-1h, 5-2h, 5-3h, 5-4h, and 5-4i provide a reference for gauging the effect of measurement noise on the shape of the derivative functions. Also, they demonstrate that in the limit as noise level goes to zero, $\hat{a} \rightarrow a_0$ and $\hat{a}$ is the only stable root. Both these situations were predicted by the theoretical discussions of Chapter 4 except that the proof of the latter condition was limited to autonomous models of which only Figures 5-4h and 5-4i are examples.

The $\hat{a}$ curves of groups (a), (b), and (c) are examples of the evolution of the MLE of a_0 as the number of samples increases. Each group was computed at a different noise level. (In both group (a) and group (b), $\sigma^2 = 0.01$, but in group (a), $x_\infty = 2/3$, and in group (b), $x_\infty = 4$.) Once again, the data lead to the conclusion that low and moderate noise levels present no difficulty for the ML estimation schemes.

Figures 5-1j, 5-2j, 5-3j, and 5-4j give insight into the distribution and consistency of $\hat{a}$. As expected from theory, the MLE under the conditions for Figures 5-1j, 5-2j, and 5-3j has all appearances of being consistent and gaussian in the limit. On the other hand, consistency of $\hat{a}$ in Figure 5-4j is not obvious although considerable convergence occurred in going from 3 to 10 samples. The problem here is that the model for this figure is autonomous. Although improvement in the estimate can be expected as the number of samples grows, consistency cannot be shown for the autonomous case. In fact, in any practical situation, improved estimates will probably sooner or later be supplanted by degrading estimates as the signal content in the

measurements tends to fall below the word length of the computer providing the MLE.

5.3.2 MULTIDIMENSIONAL MODEL

The figures for the multidimensional model give some indication of the behavior of the derivative functions for x_0 known and the autonomous version of x_0 as an unknown parameter. The cost functions in both cases, Figures 5-5a and 5-6a, are relatively smooth and symmetric - at least along a line in 2×2 A-space passing through A_0 . The curves for the norm of the derivative, Figures 5-5b and 5-6b, though reasonably smooth in the range shown, take a sharp plunge to their minima.

The curves for the directional derivatives, Figure 5-5c and 5-6c behave less dramatically than those for the norm of the derivative and have the more familiar form of a scalar parameter quadratic cost function derivative. Figure 5-5c corresponds directly to the cost function of Figure 5-5a, but Figure 5-6c corresponds only in an approximate sense because Figure 5-6a is based on the true x_0 while Figure 5-6c is based on $\hat{x}_0$. Note that while the directional derivative curves pass through zero, the normed derivative curves do not, because there in general is non-zero gradient vector orthogonal to the A_0 direction when the directional derivative passes through zero.

In all four derivative figures, " $\hat{c}$ " is obvious and close to $c = 1$ (or " c_0 "). However, $\hat{c}$ is unique for the range shown in the directional derivative figures, but there are local minima in normed derivative figures. Thus, for numerical evaluation of $\hat{A}$ the combination of moving the solution along directional derivative type curves but

measuring convergence by some norm of the derivative function appears to have merit.

5.4 CALCULATION OF THE MLE AND ITS APPROXIMATIONS

Expressing the derivative functions in the form of polynomials does have the advantages that if the coefficients are bounded, the functions will be bounded and will possess all derivatives. Furthermore, a wealth of literature exists on the properties, analysis, and solution of (scalar) polynomial equations. Also well known is the fact that the roots of a polynomial may be difficult to locate numerically.

5.4.1 SOLUTION OF THE LIKELIHOOD EQUATION

The roots of a polynomial may be well defined mathematically, but a useful definition for numerical work is not as clear. The most common definition and the one used in the simulation studies is that any value α is a root of the likelihood equation $D_N(a) = 0$ if $|D_N(\alpha)| < \epsilon$, $\epsilon > 0$.

Having established a definition for a root of the polynomial, the next step is to select some technique to find the root. The classical numerical technique for solution of likelihood equations is Fisher's 'scoring for parameters', the 'score' being the derivative function evaluated at the latest estimate of the root (Rao [1965, p. 302]). The process is basically a Newton type iteration and has the following form for a sample size of n and parameter θ :

$$\hat{\theta}_{i+1} = \hat{\theta}_i + \frac{\partial \log L}{\partial \theta} \bigg|_{\hat{\theta}_i} / \{nI(\hat{\theta}_i)\} \quad (5.7)$$

where the information $I(\theta)$ is defined as

$$I(\theta) = E \left(- \frac{\partial^2 \log L}{\partial \theta^2} \middle| \theta \right) \quad (5.8)$$

The basis for this method is that generally in the classical considerations of maximum likelihood estimation

$$\frac{1}{n} \frac{\partial^2 \log L}{\partial \theta^2} \xrightarrow{\text{a.s.}} - I(\theta) \quad (5.9)$$

The numerical analysis literature offers a variety of methods to determine roots of polynomials, e.g., see Busk and Svejgaard [1962], many of which could be more desirable in specific situations, if not in most situations as contended by some, than scoring for parameters. The methods can be considered as members of one of two groups, direct or iterative. The direct methods are recursive and make no use of any initial estimate of the roots. Generally, for reasonably behaved functions for which there is at least some rough information on root locations, the iterative methods are more effective. Barnett [1966] compares Newton-Raphson (method of tangents), fixed derivative Newton, scoring for parameters, and *regula falsi* (method of chords) methods for the solution of likelihood equations with multiple roots. He concludes that *regula falsi* is most easily controlled and most reliable for seeking out the desired roots and in addition locates only roots which correspond to maxima or minima, as the case may be. Jennrich and Sampson [1968] review steepest descent, Newton-Raphson, and Gauss-Newton as applied to non-linear least squares estimation. They state that of the three, Gauss-Newton, an iteration method which applies standard linear regression to a linearized version of the non-linear least squares problem, is most popular because it specifies the step size for the iteration and does not require second derivatives. (For the studies

reported in this chapter, *regula falsi* appeared most appropriate and was the only technique used.)

One other aspect of the numerical solutions should at least be mentioned. The theory on convergence of most standard numerical root solving methods is reasonably well developed. However, since the roots of the likelihood equation and the sequence of approximations resulting from the process of numerical solution of the roots are random variables, the usual convergence criteria must be reconsidered from a statistical point of view. Large sample (stochastic) convergence for Newton-Raphson and scoring for parameters is shown by Kale [1961] for solution of the (classical) scalar likelihood equation and by Kale [1962] for the (classical) multiparameter likelihood equation. Jennrich [1969] shows large sample convergence of the Gauss-Newton method applied to non-linear least squares.

5.4.2 APPROXIMATIONS TO THE MLE

The maximum likelihood estimators developed in Chapter 3 grow in total number of terms as the number of samples grows and require the entire measurement sequence and input sequence to be stored. These characteristics could preclude real time application of the estimators, particularly if a completely updated MLE is demanded after each new sample. Obviously, given enough samples, any computer would eventually become clogged to the point where further evaluations of the estimate, real time or not, would be impractical.

The iterative root-finding numerical methods typically require evaluation of the likelihood equation (and, in some cases, also the derivative of the likelihood equation) for each iteration. The least

expensive evaluation of a general polynomial is given by Horner's method (Lyusternick, et al. [1965, p. 10]). For an n th degree polynomial, n additions and n multiplications are required.

Clearly, the computational constraints dictated by many practical situations can be met only by approximating the maximum likelihood estimate. This problem and some possible responses have been given limited discussion in Chapter 4. The average coefficient method was proposed as an analytical approximation to the MLE. The equations for the scalar parameter estimates are repeated below. Equation (5.10) (see Equation (4.84)) is the two-sample approximation for the x_0 known, x_0 unknown parameter, and x_0 unknown random variable cases, and Equation (5.11) (see Equation (4.91)) corresponds to the differencing approach approximation.

$$C_N' a^2 - D_N' a - C_N' = 0 \quad (5.10)$$

$$\sigma^2 a^3 + C_N' a^2 + (\sigma^2 - D_N') a - C_N' = 0 \quad (5.11)$$

where

$$C_N' = \frac{1}{N} \sum_{i=1}^N y_{i-1} (y_i - h b u_{i-1}) \quad (5.12)$$

$$D_N' = \frac{1}{N} \left[\sum_{i=1}^N (y_i - h b u_{i-1})^2 - \sum_{i=1}^N y_{i-1}^2 \right] \quad (5.13)$$

Another approximation, which is more numerical in nature than the average coefficient method, that also appears to have merit is a form of curve fitting which exploits a recursive aspect of the likelihood equation. Through curve fitting, the MLE can be in theory approximated to any degree desired whereas the extent to which the average coefficient estimate approximates the maximum likelihood estimate is relatively

unclear. The method consists of initially selecting several specific values of the parameter variable a such that the area in which $\hat{a}$ is expected to lie is spanned and, after each new sample, recursively evaluating the derivative function at these points. The appropriate root of a curve fitted through the updated points on the derivative function is taken as the approximation to $\hat{a}$.

This recursive curve fitting method results in a substantial reduction of computations and storage - a reduction by a factor of approximately N , the total number of samples at the time of computation. There are some problems associated with this method which tend to offset its computational advantage. The total number of points which must be computed depends on the size of the region in which $\hat{a}$ is expected to lie and the precision to which $\hat{a}$ must be known. As $\hat{a}$ begins to stabilize after several samples, moving the points to improve the approximation would be desirable. There appears to be no way to move the points without making some approximations. Also, curve fitting can introduce extraneous roots or, conversely, could result in a curve which has no real roots in the region of $\hat{a}$.

5.4.3 SIMULATION OF THE MLE APPROXIMATIONS

To provide some indication of performance, simulations of the two-sample average coefficient method and a 3-point recursive parabolic fit to the derivative functions were made. (The fit was arranged with a_0 midway between the second and third points to give a worst case when a_0 is spanned by the points.) For comparison, the results are presented along with the MLE and a least squares estimate, derived as follows. From Equation (3.50):

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = a \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{pmatrix} + hb \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-1} \end{pmatrix} + \begin{pmatrix} \zeta_0 \\ \zeta_1 \\ \vdots \\ \zeta_{N-1} \end{pmatrix} \quad (5.14)$$

or,

$$y_N = ay_{N-1} + hbu_{N-1} + \zeta_{N-1} \quad (5.15)$$

or,

$$\hat{a}_{LSQ} = \frac{y_{N-1}^T (y_N - hbu_{N-1})}{y_{N-1}^T y_{N-1}} \quad (5.16)$$

The curves of group (a) (Figures 5-1a, 5-2a, 5-3a, 5-4a), group (b) (Figure 5-1b), and group (c) (Figure 5-1c) display the evolution of $\hat{a}$ for maximum likelihood (MLE), 2-sample average coefficient (AVC), 3-point recursive fit (3PT), and least squares (LSQ). The results indicate that for low noise levels (group (b)), both least squares and average coefficient do well compared to MLE for all but the autonomous differencing approach. As the noise level increases (group (a) and group (c), respectively) the amount by which they are in error increases substantially relative to the MLE, though average coefficient performed noticeably better than least squares. (Since theory in Chapter 4 indicated consistency for the average coefficient method under the conditions of the simulation, it should have performed reasonably well.) On the other hand, for the autonomous differencing approach, Figure 5-4a, for example, least squares and average coefficient drift toward zero. (This drift of the average coefficient $\hat{a}$ was expected because the limiting root was shown to be zero.)

The 3-point recursive fit follows the MLE in group (a) and (c) but drifts away in group (b). The drift in group (b) is most likely due to the peak in the derivative curves near a_0 for group (b) (see Figure 5-2f).

In the description of group (b), the effect of increasing the separation of the points from 0.1 to 0.2 was mentioned. The error increased from a few per cent for 0.1 separation (as observed from the curves) to about 10% at 30 samples for 0.2. Thus, for 3 points separated by 0.1 and *a priori* knowledge of a_0 to within ± 0.1 , the figures show that the $\hat{a}$ approximation is good to nearly two significant figures whereas for the 0.2 situation barely more than one significant figure is obtained.

5.5 COMPUTATIONAL ERRORS

The numerical error resulting from the finite word length of digital computers can have a substantial effect on the reliability of computations. The problem can arise at either end, so to speak, roundoff or cancellation. (For a double precision accumulator and single precision storage with floating point operations these events are mutually exclusive for any single operation.) The problem of differencing nearly equal quantities (cancellation) was observed to occur during evaluation of the derivative functions at some distance from a_0 or A_0 , but generally only when $|a| > 1$ in the scalar case. Also for $|a| > 1$, exponential overflow can easily occur. Neither of these problems is of much interest here however since most of the computational effort would be confined to the neighborhood of the root $\hat{a}$.

The problem of roundoff is discussed in detail by Wilkinson [1963]. For double precision addition, normalization, and rounding to single

precision for storage, he shows floating point addition $x_1 + x_2$ yields $(x_1 + x_2)(1 + \epsilon_1)$, $|\epsilon_1| \leq \beta^{1-t}$. (β is the base for the computer's arithmetic and t is the number of digits in the single precision mantissa.) Similarly, the floating point multiplication $x_1 x_2$ yields $x_1 x_2(1 + \epsilon_2)$, $|\epsilon_2| \leq \beta^{1-t}$.

Following a development of Adams [1967], an estimate of the round-off error in polynomial evaluation can be made. Let $\sigma = |\epsilon_1|$ and $\pi = |\epsilon_2|$. Consider the polynomial

$$l(a) = c_0 + c_1 a + \dots + c_n a^n$$

Horner's recurrence for computing $l(x)$ is

$$b_0 = c_n$$

$$b_k = x b_{k-1} + c_{n-k}, \quad k = 1, \dots, n \quad (5.17)$$

where

$$l(x) = b_n$$

The roundoff accumulation on the k th step in the evaluation of the polynomial can be described as

$$(|b_k| + e_k) = [x(|b_{k-1}| + e_{k-1}) + |c_{n-k}|] (1 + \sigma) \quad (5.18)$$

Expanding (5.18), rearranging, and dropping higher order terms gives the following error recursion,

$$\delta_k = |x| \delta_{k-1} + |b_k| \quad (5.19)$$

where

$$\delta_k = \frac{e_k + |b_k| \pi}{\pi + \sigma}$$

and

$$\delta_0 = \frac{|c_n| \pi}{\pi + \sigma}$$

Let $l_0(x)$ be the true value of the polynomial and $l(x)$ be the computed

value. Then

$$|l_0(x) - l(x)| \leq (\pi + \sigma) \delta_n - |b_n| \pi$$

(Note that the coefficients $c_i = c_i(y_0, \dots, y_N, u_0, \dots, u_{N-1})$ must also be computed and consequently are in error).

Wilkinson demonstrates that root locations can be extremely sensitive to coefficient perturbations. As a measure of the sensitivity of the roots he develops what he calls the 'condition of a polynomial'. Briefly, Wilkinson proceeds as follows to arrive at the condition.

Let $\hat{a}$ be a root of $l(a)$. Consider the zero of $l(a) + \epsilon g(a)$ where

$$g(a) = g_0 + g_1 a + \dots + g_n a^n \quad (5.20)$$

By the theory on series reversion, the change in root location can be bounded for sufficiently small ϵ as

$$\left| \hat{a}(\epsilon) - \hat{a} + \epsilon \frac{g(\hat{a})}{l'(\hat{a})} \right| < k\epsilon^2 \quad (5.21)$$

where

$$l' = \frac{d}{da} l(a)$$

or,

$$\hat{a}(\epsilon) - \hat{a} \sim \epsilon \frac{g(\hat{a})}{l'(\hat{a})}, \quad \epsilon \rightarrow 0 \quad (5.22)$$

As expected, the crucial factor in root sensitivity is the slope at the root. In the earlier discussions, the fact that the slope of the derivative functions at $\hat{a}$ increases as the number of samples increases was pointed out.

SECTION VI

SUMMARY OF RESULTS AND CONCLUSIONS

Likelihood equations expressing the MLE for each of four initial condition assumptions were derived. The finite sample and large sample characteristics of the estimators were examined. Computationally efficient approximations to the MLE were proposed and investigated. The finite sample and large sample characteristics as well as the approximations are discussed for models without plant noise whereas likelihood equations are presented for both models with and without plant noise.

The likelihood equations based on the scalar model can be expressed as polynomials in the unknown parameter a for each of the four initial condition situations considered. For the vector-matrix model (without plant noise) the likelihood equations for initial condition x_0 known and x_0 unknown parameter are again polynomials but now are polynomials in the matrix A with matrix coefficients. The character of the polynomials varies considerably with initial condition assumptions. One particularly interesting observation in this regard is that the MLE without plant noise when x_0 is known or is an unknown parameter does not explicitly depend on the variance of the measurement noise. Also, the degree and complexity of the polynomials increase with the number of samples upon which the estimate is based. This expansion appears to be unavoidable because by the work of Dynkin no sufficient statistic other than the trivial one (all the samples) exists when x_0 is known or an unknown parameter. The same conclusion is expected to hold for

the other two initial condition cases since in those situations the samples are neither independent nor identically distributed.

For finite samples the scalar MLE in all four cases approaches the true parameter value as the noise goes to zero. A similar but stronger result is given by Theorem 4.1 which says that on the average the true parameter value is a root of the likelihood equation. The converse of that, the average of the roots of the likelihood equation corresponds to the true value of the parameter, was not shown nor is it clear that such is the case. Also, in the limit as the measurement noise goes to zero, the MLE for stable scalar autonomous models is the only stable root of the likelihood equations and, as already mentioned, is equal to the true parameter value. The simulations indicate that this conclusion may be nearly true even for non-zero forcing functions and up to moderate noise levels if the true parameter value is stable and not in the neighborhood of ± 1 , but extraneous roots do occur outside of $(-1,1)$.

Concern for possible problems that could be encountered in the numerical solution of the polynomial likelihood equations is lessened by evidence from the simulations that for stable models the derivatives of the likelihood functions in the interval $(-1,1)$ are relatively smooth and rather insensitive to perturbations from noise up to moderate levels. A development by Wilkinson shows that an important factor in the sensitivity of roots of polynomials with respect to perturbations in their coefficients is the magnitude of the slope of the polynomial at the roots. In the simulation results, the slope at the root corresponding to the MLE was observed to increase as the number of

samples increased. The shape of the derivatives of the likelihood functions varies considerably with true parameter value and can have peaks in the neighborhood of the roots. In the light of this and discussions by Barnett, *regula falsi* makes a good choice for a numerical method to compute the roots of the scalar likelihood equations.

The MLE settles down after the first few samples in the simulations. Any improvement after that comes about rather slowly if at all. When x_0 is an unknown random variable, even a relatively large difference between the actual initial condition and the mean initial condition appears to have no significant effect except for a transient during the first couple of estimates. The ML estimators generally performed well on an absolute scale as well as relative to least squares and any approximate ML estimators. Though the form of the estimator depends on the initial condition assumption, they appear to perform similarly for the same set of measurements.

The cost and derivative functions for matrix parameters appear relatively smooth over the region of interest in the situations which were simulated. Along a vector through the true matrix in n^2 A-matrix space, the roots (or minima, as the case may be) of the derivative functions occurred in the neighborhood of the true value of A when the noise level was relatively low.

The only large sample property investigated was consistency. The MLE when x_0 is known was shown to be consistent for stable scalar models. The same arguments for consistency appear to hold also when x_0 is an unknown parameter and when x_0 is an unknown random variable. Because of lack of uniqueness in the limit, consistency for autonomous

models was not shown. Though improvements in the MLE can be expected initially for stable autonomous systems, continued improvement that may well be theoretically possible for increasing but still finite numbers of samples cannot be expected to occur because of the finite word length of any processing computer. The Monte Carlo simulation results tend to substantiate the above conclusions.

To overcome the growth in complexity of the MLE computed with increasing numbers of samples, two approximations to the ML estimator are proposed—average coefficient and recursive curve fitting. The average coefficient approximation can be based on any length string of samples but only the two-sample form was considered. This approximation scheme applies in both scalar and matrix situations, but only to the differencing approach and x_0 unknown parameter case. Since there is no simple way to account for the initial condition information in the cases where x_0 is known or x_0 is an unknown random variable, the unknown parameter approximation was assumed to serve for these cases also. Both the two-sample approximations are related to other approximations and estimators in the literature.

The estimates in both average coefficient approximations approach the true parameter value as the noise goes to zero, and the expectation of their noise terms is zero. The one for x_0 unknown parameter always has two real roots, one in $[-1,1]$ and the second in $(-1,1)^c$, and gives a consistent estimate if the input to the system is a constant. In the one for the differencing approach, the root locations are roughly similar if the noise level is low. For this case the large sample root is zero if the model is autonomous. The performance of these methods in the simulations lies between least squares and ML.

The curve fitting method can be based on any number of points but the greater the number of points the less the computational advantage becomes. (The same is true for string length in the average coefficient methods.) For three points separated by 0.1 which span the true parameter value, the simulation of this approximation yielded an estimate good to nearly two significant figures. This approximation performs well if the derivative of the likelihood function is smooth near the desired root, but as presented can be used only with scalar likelihood equations.

A number of questions remain only partially answered. Evidence on the usefulness of initial condition information was not very conclusive. The numerical aspects of the solution of the matrix likelihood equations as well as the properties of the solutions were only touched upon. Finite sample root distributions were not firmly established.

There are many possible extensions to this study including increasing the unknown parameters to include the H and B matrices or parameters in the noise distribution, input measurement noise, time varying coefficients, and A matrices some of whose elements are known. As a special case of the latter, further studies of the companion matrix form of A should be considered since the multiparameter identification problem is often posed in this form.

APPENDIX A

DETERMINANT AND INVERSE OF MEASUREMENT COVARIANCE - x_0 RANDOM VARIABLE

$$\text{Let } R = \alpha I + \beta \underline{a} \underline{a}^T \quad \text{A.1}$$

where:

$$\underline{a}^T = (1, a, \dots, a^N)$$

I = identity matrix

α , β , and a are real numbers

Theorem A.1: Let R be the $(N+1) \times (N+1)$ matrix defined by (A.1). Then its determinant can be expressed as

$$|R| = \alpha^N \left(\alpha + \beta \sum_{i=0}^N a^{2i} \right) \quad \text{A.2}$$

Proof:

Since the rank of $\underline{a}$ is one, the rank A , where $A \triangleq \underline{a} \underline{a}^T$, cannot be greater than one. Clearly, the rank of A is one.

Then if

$$\beta = 0, |R| = \alpha^N, \text{ and if } \alpha = 0, |R| = 0.$$

Assume α and β not zero. $\underline{a}$ is a non-trivial eigenvector of A , i.e.,

$$A \underline{a} = (\underline{a}^T \underline{a}) \underline{a} \quad \text{A.3}$$

Because A is symmetric, there exists an orthogonal matrix M such that $A = M \Lambda M^T$ where $\Lambda = \text{diag}(\underline{a}^T \underline{a}, 0, \dots, 0)$. Then

$$|M(\alpha I + \beta \Lambda)M^T| = \alpha^N (\alpha + \beta \underline{a}^T \underline{a}) \quad \text{A.4}$$

Theorem A.2: Let R be the $(N+1) \times (N+1)$ matrix defined in (A.1). Then its inverse can be expressed as

$$R^{-1} = \frac{1}{\alpha} \left[I - \frac{\beta}{\alpha + \beta \underline{a}^T \underline{a}} \underline{a} \underline{a}^T \right] \quad \text{A.5}$$

Proof:

If $\alpha = 0$, R^{-1} does not exist, and if $\beta = 0$, $R^{-1} = \alpha^{-1}I$.

Assume α and β not zero. Then by (A.3) and the symmetry of A , there exists an orthogonal matrix M such that

$$M^T R M = \alpha I + \beta \text{diag}(\lambda, 0, \dots, 0) \quad \text{A.6}$$

where

$$\lambda = \underline{a}^T \underline{a}$$

The inverse of (A.6) is

$$\begin{aligned} M^T R^{-1} M &= \text{diag}[(\alpha + \beta\lambda)^{-1}, \alpha^{-1}, \dots, \alpha^{-1}] \\ &= \alpha^{-1}I + \text{diag}(\nu, 0, \dots, 0) \end{aligned} \quad \text{A.7}$$

where

$$\nu = -\beta\lambda\alpha^{-1}/(\alpha + \beta\lambda)$$

Since $MM^T = M^T M = I$, M must have the form

$$M = \begin{pmatrix} \frac{\underline{a}}{\sqrt{\lambda}} & M^* \end{pmatrix} \quad \text{A.8}$$

Then,

$$R^{-1} = \alpha^{-1}I + (\nu/\sqrt{\lambda})\Psi M^T \quad \text{A.9}$$

where,

$$\Psi = \begin{pmatrix} \underline{a} & 0 \end{pmatrix}$$

or,

$$\begin{aligned} R^{-1} &= \alpha^{-1}I + \frac{\nu}{\lambda} \underline{a} \underline{a}^T \\ &= \alpha^{-1} \left[I - \frac{\beta}{\alpha + \beta \underline{a}^T \underline{a}} \underline{a} \underline{a}^T \right] \end{aligned} \quad \text{A.10} \quad \blacksquare$$

APPENDIX B

DETERMINANT AND INVERSE OF TRI-DIAGONAL TOEPLITZ MATRIX

Let T_N be an $N \times N$ tri-diagonal Toeplitz matrix where

$$T_N = \alpha I + \beta I_0 \quad \text{B.1}$$

$$I_0 = \begin{pmatrix} 0 & 1 & & 0 \\ 1 & & & \\ & & & \\ & 0 & & 1 & 0 \end{pmatrix} \quad \text{B.2}$$

and where I is the identity matrix. The coefficients α and β are real.

B.1 THE DETERMINANT OF T_N

The eigenvalues of I_0 are the roots of $\Delta_N(\lambda)$ where

$$\Delta_N = \det(\lambda I - I_0) \quad \text{B.3}$$

Based on Grenander and Szego [1958, §5.3], Δ_N of (B.3) can be expressed as a recursion from which the roots can be found.

$$\Delta_N = \lambda \Delta_{N-1} - \Delta_{N-2}, \quad N = 3, 4, \dots \quad \text{B.4}$$

where

$$\Delta_1(\lambda) = \lambda$$

$$\Delta_2(\lambda) = \lambda^2 - 1$$

Solving the difference equation:

$$z^2 - \lambda z + 1 = 0 \quad \text{B.5}$$

$$\text{or, } z_{1,2} = \frac{1}{2}(\lambda \pm \sqrt{\lambda^2 - 4}) \quad \text{B.6}$$

$$\text{or, } \Delta_N = c_1 z_1^N + c_2 z_2^N \quad \text{B.7}$$

The coefficients c_1 and c_2 are difficult to evaluate when Δ_N is expressed in the form of (B.7). The following change of variable helps overcome this. Take a such that,

$$\lambda = \frac{1+a^2}{a} \quad \text{B.8}$$

Then the roots in (B.6) become,

$$z_1, z_2 = a, a^{-1} \quad \text{B.9}$$

The solution to (B.4) in terms of a is

$$\Delta_N = C_1 a^N + C_2 a^{-N} \quad \text{B.10}$$

The coefficients are found from the initial conditions:

$$\Delta_1 = (1+a^2)/a = C_1 a + C_2/a$$

$$\Delta_2 = [(1+2a^2+a^4)/a^2] - 1$$

$$= (1+a^2+a^4)/a^2 = C_1 a^2 + C_2/a^2$$

or,

$$C_1 = -a^2/(1-a^2) \quad \text{B.11}$$

$$C_2 = 1/(1-a^2) \quad \text{(B.12)}$$

From (B.10),

$$\Delta_N = \frac{1-a^{2N+2}}{(1-a^2)a^N} \quad \text{B.13}$$

The roots of Δ_N in terms of a are roots of unity. Noting that $a = \pm 1$ is not a root of Δ_N , the roots are

$$a_k = e^{i\frac{2\pi k}{2N+2}} \quad k = 1, \dots, N, N+2, \dots, 2N+1 \quad \text{B.14}$$

or in terms of λ ,

$$\begin{aligned} \lambda_k &= e^{i\frac{2\pi k}{2N+2}} + e^{-i\frac{2\pi k}{2N+2}} \\ &= 2 \cos \frac{2\pi k}{2N+2} \quad k = 1, \dots, N, N+2, \dots, 2N+1 \end{aligned} \quad \text{B.15}$$

Since T_N is $N \times N$, there are N eigenvalues. In (B.15), roots for $N+2 \leq k \leq 2N+1$ duplicate those for $1 \leq k \leq N$. Therefore, the eigenvalues of (B.2) are

$$\lambda_k = 2 \cos \frac{\pi k}{N+1} \quad k = 1, 2, \dots, N \quad \text{B.16}$$

Because I_0 is real and symmetric, there exists a similarity transformation to diagonalize I_0 . Thus,

$$|T_N| = \prod_{k=1}^N (\alpha + 2\beta \cos \frac{k\pi}{N+1}) \quad \text{B.17}$$

As a special case, let $\alpha = \sigma^2(1+r^2)$ and $\beta = -\sigma^2 r$. Then from B.13 taking $r = -a$,

$$|T_N| = (-\sigma^2 r)^N \frac{1-r^{2N+2}}{(1-r^2)(-r)^N} = (\sigma^2)^N (1+r^2+\dots+r^{2N}) \quad \text{B.18}$$

(The result in (B.18) could have been obtained somewhat more directly by factoring T_N into the form $G_N G_N^T$.)

B.2 INVERSE OF T_N

Let $T_N = [t_{ij}]$ and $T_N^{-1} = [\tau_{ij}]$. Since T_N is symmetric, τ_{ij} need be determined only for $j \geq i$.

The cofactor of t_{ij} , $j \geq i$, has the form:

$$\begin{vmatrix} T_{i-1} & B_1 & \\ \hline & D_{j-i} & B_2 \\ 0 & & \hline & 0 & T_{N-j} \end{vmatrix}$$

where D_{j-i} is a $(j-i) \times (j-i)$ upper triangular matrix with β along the diagonal and T_s is of the form (B.1). Then

$$\begin{aligned} \tau_{ij} &= (-1)^{i+j} |T_{i-1}| |D_{j-i}| |T_{N-j}| / |T_N| \\ &= (-\beta)^{j-i} |T_{i-1}| |T_{N-j}| / |T_N| \quad j \geq i \end{aligned} \quad \text{B.19}$$

where

$$T_0 = 1$$

As a special case, let $\alpha = \sigma^2(1+r^2)$ and $\beta = -\sigma^2 r$. Then for $j \geq i$,

$$\tau_{ij} = \frac{1}{\sigma^2} \left(\sum_{s=0}^{i-1} \sum_{t=0}^{N-j} r^{2(s+t)+j-i} \right) / \left(\sum_{v=0}^N r^{2v} \right) \quad \text{B.20}$$

APPENDIX C

THE LIKELIHOOD EQUATION FOR THE SCALAR AUTONOMOUS DIFFERENCING APPROACH

From Equation (3.64), the likelihood function for the scalar autonomous differencing approach after $N+1$ samples is

$$L_x = (2\pi)^{N/2} |R|^{-1/2} \exp \left[-\frac{1}{2} \tilde{y}^*{}^T R^{-1} \tilde{y}^* \right] \quad C.1$$

where:

$$\tilde{y}^* = [(y_1 - ay_0), \dots, (y_N - ay_{N-1})] \quad C.2$$

$$R = \sigma^2 \begin{bmatrix} 1+a^2 & -a & & 0 \\ -a & 1+a^2 & -a & \\ & -a & 1+a^2 & -a \\ 0 & & -a & 1+a^2 \end{bmatrix} \quad C.3$$

$$\text{Let } \tilde{y}^* = y_N^1 - ay_N^0 \quad C.4$$

where:

$$(y_N^1)^T = (y_1, \dots, y_N) \quad C.5$$

$$(y_N^0)^T = (y_0, \dots, y_{N-1}) \quad C.6$$

Forming $\frac{d \log L_x}{da} = 0$ gives:

$$\begin{aligned} & -|R|^{-1} \left(\frac{d}{da} |R| \right) + 2(y_N^0)^T R^{-1} (y_N^1 - ay_N^0) \\ & - (y_N^1 - ay_N^0)^T \left(\frac{d}{da} R^{-1} \right) (y_N^1 - ay_N^0) = 0 \end{aligned} \quad C.7$$

From (3.67):

$$|R| = (\sigma^2)^N \sum_{i=0}^N a^{2i} \triangleq (\sigma^2)^N |R_2| \quad C.8$$

$$R^{-1} \triangleq \alpha R_2 \quad C.9$$

where: R_2 is the adjoint of $\frac{1}{\sigma^2} R$

$$\alpha = (\sigma^2 |R_1|)^{-1} \quad \text{C.10}$$

$$\frac{d}{da} |R| = (\sigma^2)^N \frac{d}{da} |R_1| \quad \text{C.11}$$

$$\frac{d}{da} R^{-1} = \frac{d\alpha}{da} R_2^{-1} + \alpha \frac{d}{da} R_2^{-1} \quad \text{C.12}$$

where

$$\frac{d}{da} \alpha = -(\sigma^2 |R_1|^2)^{-1} \frac{d}{da} |R_1| \quad \text{C.13}$$

From (3.68)

$$R_2^{-1} = (r_{2ij}^{-1}) \quad \text{C.14}$$

where

$$r_{2ij}^{-1} = \sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i-1}, \quad j \geq i \quad \text{C.15}$$

$$\frac{d}{da} |R_1| = 2 \sum_{p=0}^N p a^{2p-1} \quad \text{C.16}$$

$$\frac{d}{da} r_{2ij}^{-1} = \sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(s+k)+j-i-1] a^{2(s+k)+j-i-2} \quad \text{C.17}$$

Multiplying through (C.7) by $\sigma^2 |R|^2 / (\sigma^2)^{2N}$ gives:

$$V_N + Q_N = 0 \quad \text{C.18}$$

where

$$V_N = -\sigma^2 |R_1| \frac{d}{da} |R_1| \quad \text{C.19}$$

$$\begin{aligned} Q_N = & -(y_N^1)^T \left[-\left(\frac{d}{da} |R_1| \right) R_2^{-1} + |R_1| \frac{d}{da} R_2^{-1} \right] y_N^1 \\ & + 2(y_N^0)^T [R_1 (R_2^{-1}) - a \left(\frac{d}{da} |R_1| \right) R_2^{-1} + a |R_1| \frac{d}{da} R_2^{-1}] y_N^1 \\ & - a(y_N^0)^T [2 |R_1| (R_2^{-1}) - a \left(\frac{d}{da} |R_1| \right) R_2^{-1} + a |R_1| \frac{d}{da} R_2^{-1}] y_N^0 \end{aligned} \quad \text{(C.20)}$$

The reduced form of (C.20) is relatively simple. At this time

no simple way has been uncovered that yields (3.73) from (C.20). An

inductive proof will be given instead after the following lengthy

Lemma is established.

Lemma Let $\Delta_{N-1} = Q_N - Q_{N-1}$

C.21

Then,

$$\begin{aligned} \Delta_{N-1} = & 2 \sum_{k=0}^{N-1} (N-k) a^{2(N+k)-1} y_N^2 \\ & + 2 \sum_{i=0}^{N-1} \sum_{p=0}^N (N+i-2p) a^{N+i+2p-1} y_i y_N \\ & - \sum_{r=0}^{N-1} \sum_{j=0}^{N-1} (2N-j-r) a^{2N+j+r-1} y_r y_j \end{aligned} \quad \text{C.22}$$

Proof:

From (C.20) and the preceding definitions:

$$\begin{aligned} Q_N = & \left(\sum_{p=0}^N 2pa^{2p-1} \right) \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} a^{2(s+k)} \right) y_i^2 \right] \\ & + 2 \left(\sum_{p=0}^N 2pa^{2p-1} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} \right) y_i y_j \right] \\ & - \left(\sum_{p=0}^N a^{2p} \right) \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} [2(s+k)] a^{2(s+k)-1} \right) y_i^2 \right] \\ & - 2 \left(\sum_{p=0}^N a^{2p} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(s+k)+j-i] a^{2(s+k)+j-i-1} \right) y_i y_j \right] \\ & + 2 \left(\sum_{p=0}^N a^{2p} \right) \left[\sum_{i=1}^N \sum_{j=i}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} \right) y_{i-1} y_j \right] \\ & + 2 \left(\sum_{p=0}^N a^{2p} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} \right) y_i y_{j-1} \right] \\ & - 2a \left(\sum_{p=0}^N 2pa^{2p-1} \right) \left[\sum_{i=1}^N \sum_{j=i}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} \right) y_{i-1} y_j \right] \\ & - 2a \left(\sum_{p=0}^N 2pa^{2p-1} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} \right) y_i y_{j-1} \right] \end{aligned}$$

$$\begin{aligned}
& + 2a \left(\sum_{p=0}^N a^{2p} \right) \left[\sum_{i=1}^N \sum_{j=i}^N \sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(s+k)+j-i] a^{2(s+k)+j-i-1} y_{i-1} y_j \right] \\
& + 2a \left(\sum_{p=0}^N a^{2p} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(s+k)+j-i] a^{2(s+k)+j-i-1} y_i y_{j-1} \right) \right] \\
& - 2a \left(\sum_{p=0}^N a^{2p} \right) \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} a^{2(s+k)} y_{i-1}^2 \right) \right] \\
& - 4a \left(\sum_{p=0}^N a^{2p} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} y_{i-1} y_{j-1} \right) \right] \\
& + a^2 \left(\sum_{p=0}^N 2pa^{2p-1} \right) \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} a^{2(s+k)} y_{i-1}^2 \right) \right] \\
& + 2a^2 \left(\sum_{p=0}^N 2pa^{2p-1} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} y_{i-1} y_{j-1} \right) \right] \\
& - a^2 \left(\sum_{p=0}^N a^{2p} \right) \left[\sum_{i=1}^N \sum_{k=0}^{i-1} \sum_{s=0}^{N-i} [2(s+k)] a^{2(s+k)-1} y_{i-1}^2 \right] \\
& - 2a^2 \left(\sum_{p=0}^N a^{2p} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(s+k)+j-i] a^{2(s+k)+j-i-1} y_{i-1} y_{j-1} \right) \right]
\end{aligned}$$

(C.23)

Forming $Q_N - Q_{N-1}$ and accounting for the fact that the inverses (C.14) hold only for $j \geq i$ gives:

$$\Delta_{N-1} = Q_N - Q_{N-1}$$

$$\begin{aligned}
& = 2Na^{2N-1} \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} a^{2(s+k)} y_i^2 \right) \right] \\
& + \left(\sum_{p=0}^{N-1} 2pa^{2p-1} \right) \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} a^{2(N-i+k)} y_i^2 \right) \right] \\
& + 4Na^{2N-1} \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} y_i y_j \right) \right]
\end{aligned}$$

$$\begin{aligned}
& + 2 \left(\sum_{p=0}^{N-1} 2pa^{2p-1} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} a^{2(N+k)-j-i} \right) y_i y_j \right] \\
& - a^{2N} \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} [2(s+k)] a^{2(s+k)-1} \right) y_i^2 \right] \\
& - \left(\sum_{p=0}^{N-1} a^{2p} \right) \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} [2(N-i+k)] a^{2(N-i+k)-1} \right) y_i^2 \right] \\
& - 2a^{2N} \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(s+k)+j-i] a^{2(s+k)+j-i-1} \right) y_i y_j \right] \\
& - 2 \left(\sum_{p=0}^{N-1} a^{2p} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} [2(N+k)-j-i] a^{2(N+k)-j-i-1} \right) y_i y_j \right] \\
& + 2a^{2N} \left[\sum_{i=1}^N \sum_{j=i}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} \right) y_{i-1} y_j \right] \\
& + 2 \left(\sum_{p=0}^{N-1} a^{2p} \right) \left[\sum_{i=1}^N \sum_{j=i}^N \left(\sum_{k=0}^{i-1} a^{2(N+k)-j-1} \right) y_{i-1} y_j \right] \\
& + 2a^{2N} \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} \right) y_i y_{j-1} \right] \\
& + 2 \left(\sum_{p=0}^{N-1} a^{2p} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} a^{2(N+k)-j-1} \right) y_i y_{j-1} \right] \\
& - 4a^{2N} \left[\sum_{i=1}^N \sum_{j=i}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} \right) y_{i-1} y_j \right] \\
& - 2 \left(\sum_{p=0}^{N-1} 2pa^{2p} \right) \left[\sum_{i=1}^N \sum_{j=i}^N \left(\sum_{k=0}^{i-1} a^{2(N+k)-j-1} \right) y_{i-1} y_j \right] \\
& - 4Na^{2N} \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} \right) y_i y_{j-1} \right] \\
& - 2 \left(\sum_{p=0}^{N-1} 2pa^{2p} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} a^{2(N+k)-j-1} \right) y_i y_{j-1} \right]
\end{aligned}$$

$$\begin{aligned}
& + 2a^{2N+1} \left[\sum_{i=1}^N \sum_{j=i}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(s+k)+j-i] a^{2(s+k)+j-i-1} \right) y_{i-1} y_j \right] \\
& + 2 \left(\sum_{p=0}^{N-1} a^{2p+1} \right) \left[\sum_{i=1}^N \sum_{j=i}^N \left(\sum_{k=0}^{i-1} [2(N+k)-j-i] a^{2(N+k)-j-i-1} \right) y_{i-1} y_j \right] \\
& + 2a^{2N+1} \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(s+k)+j-i] a^{2(s+k)+j-i-1} \right) y_i y_{j-1} \right] \\
& + 2 \left(\sum_{p=0}^{N-1} a^{2p+1} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} [2(N+k)-j-i] a^{2(N+k)-j-i-1} \right) y_i y_{j-1} \right] \\
& - 2a^{2N+1} \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} a^{2(s+k)} \right) y_{i-1}^2 \right] \\
& - 2 \left(\sum_{p=0}^{N-1} a^{2p+1} \right) \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} a^{2(N-i+k)} \right) y_{i-1}^2 \right] \\
& - 4a^{2N+1} \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} \right) y_{i-1} y_{j-1} \right] \\
& - 4 \left(\sum_{p=0}^{N-1} a^{2p+1} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} a^{2(N+k)-j-i} \right) y_{i-1} y_{j-1} \right] \\
& + 2Na^{2N+1} \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} a^{2(s+k)} \right) y_{i-1}^2 \right] \\
& + \left(\sum_{p=0}^{N-1} 2pa^{2p+1} \right) \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} a^{2(N-i+k)} \right) y_{i-1}^2 \right] \\
& + 4Na^{2N+1} \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} a^{2(s+k)+j-i} \right) y_{i-1} y_{j-1} \right] \\
& + 2 \left(\sum_{p=0}^{N-1} 2pa^{2p+1} \right) \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} a^{2(N+k)-j-i} \right) y_{i-1} y_{j-1} \right] \\
& - a^{2N+2} \left[\sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} [2(s+k)] a^{2(s+k)-1} \right) y_{i-1}^2 \right]
\end{aligned}$$

$$\begin{aligned}
& - \left(\sum_{p=0}^{N-1} a^{2p+2} \right) \left\{ \sum_{i=1}^N \left(\sum_{k=0}^{i-1} [2(N-i+k)] a^{2(N-i+k)-1} \right) y_{i-1}^2 \right\} \\
& - 2a^{2N+2} \left\{ \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(s+k)+j-i] a^{2(s+k)+j-i-1} \right) y_{i-1} y_{j-1} \right\} \\
& - 2 \left(\sum_{p=0}^{N-1} a^{2p+2} \right) \left\{ \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} [2(N+k)-j-i] a^{2(N+k)-j-i-1} \right) y_{i-1} y_{j-1} \right\}
\end{aligned}$$

C.24

Considering the 32 terms of (C.24) and combining them in the following groups - (1,5), (2,6), (3,7), (4,8), (9,13,17), (10,14,18), (11,15,19), (12,16,20), (21,25,29), (22,26,30), (23,27,31), and (24,28,32) gives:

$$\begin{aligned}
\Delta_{N-1} &= \sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} 2[N-s-k] a^{2(N+s+k)-1} \right) y_i^2 \\
&- \sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{p=0}^{N-1} [2(N-i+k-p)] a^{2(N-i+k+p)-1} \right) y_i^2 \\
&+ 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(N-s-k)-j+i] a^{2(N+s+k)+j-i-1} \right) y_i y_j \\
&- 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{p=0}^{N-1} [2(N+k-p)-j-i] a^{2(N+k+p)-j-i-1} \right) y_i y_j \\
&- 2 \sum_{i=1}^N \sum_{j=i}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(N-s-k)-j+i-1] a^{2(N+s+k)+j-i-1} \right) y_{i-1} y_j \\
&+ 2 \sum_{i=1}^N \sum_{j=i}^N \left(\sum_{k=0}^{i-1} \sum_{p=0}^{N-1} [2(N+k-p)-j-i+1] a^{2(N+k+p)-j-i-1} \right) y_{i-1} y_j \\
&- 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(N-s-k)-j+i-1] a^{2(N+s+k)+j-i-1} \right) y_i y_{j-1}
\end{aligned}$$

$$\begin{aligned}
& + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{p=0}^{N-1} [2(N+k-p)-j-i+1] a^{2(N+k+p)-j-i} \right) y_i y_{j-1} \\
& + 2 \sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} [N-s-k-1] a^{2(N+s+k)+1} \right) y_{i-1}^2 \\
& - 2 \sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{p=0}^{N-1} [N-i+k-p+1] a^{2(N-i+k+p)+1} \right) y_{i-1}^2 \\
& + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-j} [2(N-s-k-1)-j+i] a^{2(N+s+k)+j-i+1} \right) y_{i-1} y_{j-1} \\
& - 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{k=0}^{i-1} \sum_{p=0}^{N-1} [2(N+k-p+1)-j-i] a^{2(N+k+p)-j-i+1} \right) y_{i-1} y_{j-1}
\end{aligned}$$

C.25

Combining terms 3 and 5, 4 and 6, 7 and 11, 8 and 12 of

(C.25) gives:

$$\begin{aligned}
\Delta_{N-1} = & \sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} [2(N-s-k)] a^{2(N+s+k)-1} \right) y_i^2 \\
& - \sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{p=0}^{N-1} [2(N-i+k-p)] a^{2(N-i+k+p)-1} \right) y_i^2 \\
& + 2 \sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} [N-s-k-1] a^{2(N+s+k)+1} \right) y_{i-1}^2 \\
& - 2 \sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{p=0}^{N-1} [N-i+k-p+1] a^{2(N-i+k+p)+1} \right) y_{i-1}^2 \\
& - 2 \sum_{i=0}^{N-1} \sum_{j=i+1}^N \left(\sum_{s=0}^{N-j} [2(N-s)-j-i] a^{2(N+s)+j+i-1} \right) y_i y_j \\
& + 2 \sum_{i=0}^{N-1} \sum_{j=i+1}^N \left(\sum_{p=0}^{N-1} [2(N-p)-j+i] a^{2(N+p)-j+i-1} \right) y_i y_j
\end{aligned}$$

$$\begin{aligned}
& + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{s=0}^{N-j} [2(N-s)-j-i] a^{2(N+s)+j+i-1} \right) y_{i-1} y_{j-1} \\
& - 2 \sum_{i=2}^N \left(\sum_{k=0}^{i-2} \sum_{s=0}^{N-i} [2(N-s-k-1)] a^{2(N+s+k)+1} \right) y_{i-1}^2 \\
& + 2 \sum_{i=2}^N \left(\sum_{k=0}^{i-2} \sum_{p=0}^{N-1} [2(N+k-p-i+1)] a^{2(N+k+p-i)+1} \right) y_{i-1}^2 \\
& - 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{p=0}^{N-1} [2(N-p)-j+i] a^{2(N+p)-j+i-1} \right) y_{i-1} y_{j-1} \quad C.26
\end{aligned}$$

Combining terms 3 and 8, 6 and 10, and 2, 4, and 9 in (C.26) gives:

$$\begin{aligned}
\Delta_{N-1} = & \sum_{i=1}^N \left(\sum_{k=0}^{i-1} \sum_{s=0}^{N-i} [2(N-s-k)] a^{2(N+s+k)-1} \right) y_i^2 \\
& - 2 \sum_{i=0}^{N-1} \sum_{j=i+1}^N \left(\sum_{s=0}^{N-j} [2(N-s)-j-i] a^{2(N+s)+j+i-1} \right) y_i y_j \\
& + 2 \sum_{i=1}^N \left(\sum_{s=0}^{N-i} [N-s-i] a^{2(N+s+i)-1} \right) y_{i-1}^2 \\
& - 2 \sum_{i=2}^N \left(\sum_{k=0}^{i-2} \sum_{s=0}^{N-i} [N-s-k-1] a^{2(N+s+k)+1} \right) y_{i-1}^2 \\
& + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left(\sum_{s=0}^{N-j} [2(N-s)-j-i] a^{2(N+s)+j+i-1} \right) y_{i-1} y_{j-1} \\
& + 2 \sum_{i=0}^{N-1} \sum_{p=0}^{N-1} [N-2p+i] a^{N+2p+i-1} y_i y_N \\
& - 2 \sum_{i=0}^{N-1} \sum_{p=0}^{N-1} [N-p] a^{2(N+p)-1} y_i^2 \quad C.27
\end{aligned}$$

Combining terms 1 and 4 and 2 and 5 in (C.27) gives:

$$\begin{aligned}
\Delta_{N-1} = & 2 \sum_{i=1}^N \left(\sum_{s=0}^{N-1} [N-s-i] a^{2(N+s+1)-1} \right) y_{i-1}^2 \\
& + 2 \sum_{i=0}^{N-1} \left(\sum_{p=0}^{N-1} [N-2p+i] a^{N+2p+1-1} \right) y_i y_N \\
& - 2 \sum_{i=0}^{N-1} \left(\sum_{p=0}^{N-1} [N-p] a^{2(N+p)-1} \right) y_i^2 \\
& + 2 \sum_{i=1}^N \left(\sum_{k=0}^{i-1} [N-k] a^{2(N+k)-1} \right) y_i^2 \\
& - 2 \sum_{i=1}^N \sum_{j=i}^N [2N-j-i+1] a^{2(N-1)+j+i} y_{i-1} y_j
\end{aligned} \tag{C.28}$$

Rearranging and combining terms in (C.28) gives the desired result:

$$\begin{aligned}
\Delta_{N-1} = & 2 \sum_{k=0}^{N-1} [N-k] a^{2(N+k)-1} y_N^2 \\
& + 2 \sum_{i=0}^{N-1} \left(\sum_{p=0}^N [N-2p+i] a^{N+2p+1-1} \right) y_i y_N \\
& - \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} [2N-j-i] a^{2N+j+i-1} y_i y_j
\end{aligned} \tag{C.29}$$

Proposition Let Q_N be defined by Equation (C.20). Then Q_N can be written as:

$$Q_N = 2 \sum_{t=0}^N \sum_{i=0}^N \sum_{j=0}^N (i-t) a^{2t+i+j-1} y_i y_j \tag{C.30}$$

Proof:

By induction using (C.23), the expanded version of (C.20),

$$\begin{aligned}
Q_1 &= 2ay_1^2 + 2(1+a^2)y_0y_1 - 4a^2y_0y_1 - 2a(1+a^2)y_0^2 + 2a^3y_0^2 \\
&= 2y_0y_1 + (2y_1^2 - 2y_0^2)a - 2y_0y_1a^2
\end{aligned} \tag{C.31}$$

From (C.30),

$$Q_1 = 2(y_0 y_1 + a y_1^2 - a y_0^2 - a^2 y_0 y_1) \quad \text{C.32}$$

Again from (C.30),

$$Q_{N-1} = 2 \sum_{t=0}^{N-1} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i-t) a^{2t+i+j-1} y_i y_j \quad \text{C.33}$$

Assume (C.30) true for $N-1$. Then by the Lemma,

$$\begin{aligned} Q_{N-1} + \Delta_{N-1} &= 2 \sum_{t=0}^{N-1} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i-t) a^{2t+i+j-1} y_i y_j \\ &\quad + 2 \sum_{k=0}^{N-1} (N-k) a^{2(N+k)-1} y_N^2 \\ &\quad + 2 \sum_{i=0}^{N-1} \sum_{p=0}^N (N-2p+i) a^{N+2p+i-1} y_i y_N \\ &\quad - \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (2N-j-i) a^{2N+j+i-1} y_i y_j \end{aligned} \quad \text{C.34}$$

$$\begin{aligned} &= 2 \sum_{t=0}^{N-1} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i-t) a^{2t+i+j-1} y_i y_j \\ &\quad + 2 \sum_{t=0}^N \sum_{i=N}^N \sum_{j=N}^N (i-t) a^{2t+i+j-1} y_i y_j \\ &\quad + 2 \sum_{t=0}^N \sum_{i=0}^{N-1} \sum_{j=N}^N (i-t) a^{2t+i+j-1} y_i y_j \\ &\quad + 2 \sum_{t=0}^N \sum_{i=N}^N \sum_{j=0}^{N-1} (i-t) a^{2t+i+j-1} y_i y_j \\ &\quad + 2 \sum_{t=N}^N \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i-t) a^{2t+i+j-1} y_i y_j \end{aligned} \quad \text{C.35}$$

$$= 2 \sum_{t=0}^N \sum_{i=0}^N \sum_{j=0}^N (i-t) a^{2t+i+j-1} y_i y_j \quad \text{C.36}$$

Another version of Q_N (Equation 3.72) was developed. It is written in the standard form of a polynomial and as such was to have been used with known polynomial theory to yield information on root locations. The form of the coefficients of the polynomial, being quadratic forms in the measurements, is interesting. However, Q_N as expressed by (3.73) (or (C.30)) has more practical value computationally.

A proof by induction is outlined for this second version:

$$Q_N = \sum_{k=0}^{2(2N-1)} [y_N^T \left(\sum_{i=0}^N (k+1-4i) S_{k+1-N-2i} J \right) y_N] a^k \quad \text{C.37}$$

where:

$$y_N^T = (y_0, \dots, y_N)$$

S_P = shift matrix

J = unit Hankel matrix

$$\begin{aligned} Q_1 &= \sum_{k=0}^2 [y_1^T \left(\sum_{i=0}^1 (k+1-4i) S_{k-2i} J \right) y_1] a^k \\ &= 2y_0 y_1 + 2y_1^2 a - 2y_0^2 a - 2y_0 y_1 a^2 \end{aligned} \quad \text{C.38}$$

Assume (C.37) true for $N-1$. Then

$$\begin{aligned} Q_{N-1} &= \sum_{k=0}^{2(2N-3)} [y_{N-1}^T \left(\sum_{i=0}^{N-1} (k+1-4i) S_{k+2-N-2i} J \right) y_{N-1}] a^k \\ &= \sum_{k=0}^{2(2N-3)} [y_N^{*T} \left(\sum_{i=0}^{N-1} (k+1-4i) S_{k+1-N-2i} J \right) y_N^*] a^k \end{aligned} \quad \text{C.39}$$

where

$$y_N^{*T} = (y_0, y_1, \dots, y_{N-1}, 0)$$

Consider $Q_{N-1} + \Delta_{N-1}$.

Working with the first term of Δ_{N-1} :

$$\begin{aligned}
& 2 \sum_{k=0}^{N-1} (N-k) a^{2(N+k-1)} y_N^2 \\
&= \sum_{\substack{k=2N-1 \\ (k \text{ odd})}}^{4N-3} (4N-k-1) a^k y_N^2 \\
&= \sum_{k=2N-1}^{4N-3} (4N-k-1) \left(\sum_{i=0}^{N-1} S_{k+1-2i-2N} J \right) a^k y_N^2 \\
&= \sum_{k=0}^{2(2N-1)} \{ (0^T y_N) \left(\sum_{i=0}^N (k+1-4i) S_{k+1-2i-N} J \right) \begin{bmatrix} 0 \\ \vdots \\ y_N \end{bmatrix} \} a^k \quad \text{C.40}
\end{aligned}$$

Working with the second term of Δ_{N-1} :

$$\begin{aligned}
& 2 \sum_{i=0}^{N-1} \left(\sum_{j=0}^N (N+i-2j) a^{N+2j+i-1} \right) y_i y_N \\
&= \sum_{j=0}^{N-1} \left(\sum_{k=j+N-1}^{j+3N-1} (2N+2j-k-1) a^k y_i y_N \right) \\
&\text{where} \\
& k = j+N-1, j+N+1, j+N+3, \dots \\
&= 2 \sum_{k=0}^{4N-2} (0 \dots 0 y_N) \left(\sum_{i=0}^N (k+1-4i) S_{k+1-N-2i} J \right) \begin{bmatrix} y_0 \\ \vdots \\ y_{N-1} \\ 0 \end{bmatrix} a^k \quad \text{C.41}
\end{aligned}$$

Working with the third term of Δ_{N-1} :

$$\begin{aligned}
& - \sum_{p=0}^{N-1} \sum_{r=0}^{N-1} (2N-p-r) a^{2N+p+r-1} y_p y_r \\
&= \sum_{k=1}^{2N-1} [y_N^T (k-1-2N) S_{k-N} J y_N] a^{2N+k-2} \\
&= \sum_{k=0}^{4N-2} [y_N^{*T} (k+1-4N) S_{k-3N+1} J y_N^*] a^k \quad \text{C.42}
\end{aligned}$$

Combining (C.39) and (C.42) gives

$$\sum_{k=0}^{4N-2} [y_N^{*T} (\sum_{i=0}^N (k+1-4i) S_{k+1-N-2i} J) y_N^*] a^k \quad C.43$$

Combining (C.43), (C.41), and (C.42) gives

$$\begin{aligned} Q_{N-1} + \Delta_{N-1} &= \sum_{k=0}^{4N-2} [y_N^{*T} (\sum_{i=0}^N (k+1-4i) S_{k+1-N-2i} J) y_N^*] a^k \\ &= Q_N \end{aligned} \quad C.44$$

Returning to V_N (Equation (C.19)), and using (C.16) and (C.8),

$$\begin{aligned} V_N &= -\sigma^2 |R_1| \frac{d}{da} |R_1| \\ &= -2\sigma^2 \sum_{p=0}^N \sum_{q=0}^N q a^{2(p+q)-1} \end{aligned} \quad C.45$$

V_N can be expressed in standard polynomial form though in a somewhat awkward manner,

$$\begin{aligned} V &= -2\sigma^2 \sum_{i=0}^N \sum_{k=i}^{N+i} (k-i) a^{2k-1} \\ &= -2\sigma^2 \left[\sum_{k=1}^N \sum_{i=0}^k (k-i) a^{2k-1} + \sum_{k=N+1}^{2N} \sum_{i=k-N}^N (k-i) a^{2k-1} \right] \\ &= -2\sigma^2 \sum_{k=1}^{2N} \left[\sum_{\substack{t=0 \\ (k \leq N)}}^k t + \sum_{\substack{t=k-N \\ (k > N)}}^N t \right] a^{2k-1} \end{aligned} \quad C.46$$

APPENDIX D

THE LIKELIHOOD EQUATIONS WITH PLANT NOISE

D.1 THE MODEL

$$x_{i+1} = ax_i + bu_i + \xi_i$$

$$y_i = hx_i + \eta_i$$

$$i = 0, 1, \dots$$

D.1

where:

$$\{\xi_i\} \text{ independent and } \xi_i \sim \mathcal{N}(0, \beta^2)$$

$$\{\eta_i\} \text{ independent and independent of } \{\xi_i\}$$

$$\eta_i \sim \mathcal{N}(0, \sigma^2)$$

From (D.1),

$$y_i = ha^i x_0 + hb \sum_{j=1}^i a^{i-j} u_{j-1} + h \sum_{k=1}^i a^{i-k} \xi_{k-1} + \eta_i \quad i = 1, 2, \dots$$

$$y_0 = hx_0 + \eta_0$$

D.2

D.2 x_0 KNOWN

a. The likelihood function

Because of the assumed independence,

$$(\eta_1, \dots, \eta_N, \xi_0, \dots, \xi_{N-1}) \sim \mathcal{N}(0, R_{\eta\xi})$$

D.3

where:

$$R_{\eta\xi} = \begin{pmatrix} \sigma^2 I & 0 \\ 0 & \beta^2 I \end{pmatrix}$$

From the model (D.1) and Equation (D.2),

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = hx_0 \begin{pmatrix} a \\ \vdots \\ a^N \end{pmatrix} + hb \begin{pmatrix} u_0 \\ \vdots \\ u_{N-1} \end{pmatrix} + h \begin{pmatrix} \xi_0 \\ \vdots \\ \xi_{N-1} \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_N \end{pmatrix} \quad \text{D.4}$$

or with the obvious definitions,

$$y_N = hx_0 a_N + hb\phi u_N + h\phi \xi_N + \eta_N$$

D.5

where:

$$\phi = \begin{pmatrix} 1 & & & & 0 \\ a & & & & \\ a^2 & & & & \\ \vdots & & & & \\ a^{N-1} & \dots & a^2 & a & 1 \end{pmatrix}$$

From (D.3) and (D.5), the likelihood function is given as,

$$L = p(y_1, \dots, y_N; a) \\ = (2\pi)^{-N/2} |R_y|^{-1/2} \exp\left\{-\frac{1}{2} \left| y_N - hx_0 a_N - hb\phi u_N \right|^2_{R_y^{-1}} \right\}$$

D.6

where,

$$R_y = \phi_x R_{\eta} \phi_x^T$$

D.7

$$\phi_x = [I \mid h\phi]$$

D.8

b. The determinant of R_y

$$R_y = \phi_x R_{\eta} \phi_x^T = \sigma^2 I + h^2 \beta^2 \phi \phi^T$$

D.9

$$|R_y| = |\phi [\sigma^2 \phi^{-1} (\phi^T)^{-1} + h^2 \beta^2 I] \phi^T| \\ = |\sigma^2 \phi^{-1} (\phi^T)^{-1} + h^2 \beta^2 I|$$

D.10

Since,

$$\phi^{-1} = \begin{pmatrix} 1 & & & 0 \\ -a & & & \\ 0 & & -a & 1 \end{pmatrix}$$

D.11

then,

$$\phi^{-1} (\phi^T)^{-1} = \begin{pmatrix} 1 & -a & & 0 \\ -a & 1+a^2 & & \\ 0 & & -a & 1+a^2 \end{pmatrix}$$

D.12

Let the $N \times N$ matrices Ψ_N and A_N be defined as,

$$\Psi_N \triangleq \phi^{-1} (\phi^T)^{-1}$$

D.13

$$A_N \triangleq \begin{pmatrix} 1+a^2 & -a & & 0 \\ -a & & & \\ & & & \\ 0 & & -a & 1+a^2 \end{pmatrix} \quad \text{D.14}$$

Let L_t be the determinant of the $t \times t$ matrix of the form given by (D.9), i.e., $L_N = |R_t|$. Using Equation (B.17),

L_t may be expressed as:

$$\begin{aligned} LL_t &= (\sigma^2)^t |v^2 I + \Psi_t| \\ &= (\sigma^2)^t [(1+v^2) |v^2 I + A_{t-1}| - a^2 |v^2 I + A_{t-2}|] \\ &= (\sigma^2)^t \{ (1+v^2) \left[\prod_{k=1}^{t-1} (v^2 + 1 + a^2 - 2a \cos \frac{\pi k}{t}) \right] \right. \\ &\quad \left. - a^2 \prod_{k=1}^{t-2} (v^2 + 1 + a^2 - 2a \cos \frac{\pi k}{t-1}) \right\}, \quad t \geq 3 \end{aligned} \quad \text{D.15}$$

where,

$$v^2 = h^2 \beta^2 / \sigma^2$$

and,

$$L_2 = \sigma^2 + h^2 \beta^2$$

$$\begin{aligned} L_2 &= \det \{ \sigma^2 I + \beta^2 h^2 \begin{pmatrix} 1 & a \\ a & 1+a^2 \end{pmatrix} \} \\ &= \sigma^4 + \sigma^2 \beta^2 h^2 (2+a^2) + h^4 \beta^4 \end{aligned}$$

Consider the determinant J_p where

$$\begin{aligned} J_p &= |\sigma^2 (v^2 I + A_p)| \\ &= (\sigma^2)^p \prod_{k=1}^p (v^2 + 1 + a^2 - 2a \cos \frac{k\pi}{p+1}), \quad p \geq 1 \end{aligned} \quad \text{D.16}$$

and $J_0 = 1$.

Then (D.15) may be rewritten as

$$L_t = (\sigma^2 + h^2 \beta^2) J_{t-1} - a^2 \sigma^4 J_{t-2} \quad \text{D.17}$$

and

$$|R_y| = L_N \quad D.18$$

c. The inverse of R_y

From Equation (D.9)

$$R_y^{-1} = (\phi^T)^{-1} R^{-1} \phi^{-1} \quad D.19$$

where,

$$R^{-1} = (r_{ij}^{-1}) = [\sigma^2 \phi^{-1} (\phi^T)^{-1} + h^2 \beta^2 I] \quad D.20$$

By the same process that led to Equation (B.19),

$$r_{ij}^{-1} = (\sigma^2 a)^{j-i} (L_{i-1}) (J_{N-j}) / (L_N) \quad , \quad j \geq i \quad D.21$$

where,

$$L_0 \triangleq 1$$

$$r_{ij}^{-1} = r_{ji}^{-1}$$

d. The likelihood equation

Forming $\frac{d \log L}{da} = 0$ from (D.6) gives,

$$\begin{aligned} |R_y|^{-1} \left(\frac{d}{da} |R_y| \right) + \{ (\underline{y}_N - h x_0 \underline{a}_N - h b \phi \underline{u}_N)^T \left(\frac{d}{da} R_y^{-1} \right) \\ - 2 [h x_0 \left(\frac{d}{da} \underline{a}_N \right) + h b \left(\frac{d}{da} \phi \right) \underline{u}_N]^T R_y^{-1} \} (\underline{y}_N - h x_0 \underline{a}_N - h b \phi \underline{u}_N) \\ = 0 \end{aligned} \quad D.22$$

D.3 x_0 UNKNOWN PARAMETER

The likelihood equation for a is given by Equation (D.22) with x_0 replaced by $\hat{x}_0$. From (D.2),

$$\begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = h x_0 \begin{pmatrix} 1 \\ a \\ a^2 \\ \vdots \\ a^N \end{pmatrix} + h b \phi_0 \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-1} \end{pmatrix} + h \phi_0 \begin{pmatrix} \xi_0 \\ \xi_1 \\ \vdots \\ \xi_{N-1} \end{pmatrix} + \begin{pmatrix} \eta_0 \\ \eta_1 \\ \eta_2 \\ \vdots \\ \eta_N \end{pmatrix} \quad D.23$$

where:

$$\phi_0 = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \\ \phi & & \end{pmatrix}$$

Considering the set of random variables $(\eta_0, \eta_1, \dots, \eta_N, \xi_0, \dots, \xi_{N-1})$, the likelihood function can be written as:

$$L = (2\pi)^{-\frac{N+1}{2}} (\sigma^2)^{-\frac{1}{2}} |R_y|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \left[\left| \underline{y}_N - h x_0 \underline{a}_N - h b \phi \underline{u}_N \right|_{R_y}^2 + (\sigma^2)^{-1} (y_0 - h x_0)^2 \right]\right\} \quad D.24$$

Forming $\frac{\partial L}{\partial x_0}$ gives,

$$(\underline{y}_N - h x_0 \underline{a}_N - h b \phi \underline{u}_N)^T R_y^{-1} \underline{a}_N + (\sigma^2)^{-1} (y_0 - h x_0) = 0 \quad D.25$$

$$\text{or, } \hat{x}_0 = [(\underline{y}_N - h b \phi \underline{u}_N)^T R_y^{-1} \underline{a}_N \sigma^2 + y_0] / [h(1 + \underline{a}_N^T R_y^{-1} \underline{a}_N \sigma^2)] \quad D.26$$

D.4 x_0 UNKNOWN RANDOM VARIABLE

a. The likelihood function

Assume: $x_0 \sim \mathcal{N}(\bar{x}_0, \epsilon^2)$

and x_0 independent of $\{\xi_i\}$ and $\{\eta_i\}$

By independence,

$$(x_0, \eta_0, \eta_1, \dots, \eta_N, \xi_0, \xi_1, \dots, \xi_{N-1}) \sim \mathcal{N}(\bar{x}_0^*, R_{x\eta\xi}) \quad D.27$$

where:

$$(\bar{x}^*)^T = (\bar{x}_0, 0, \dots, 0)$$

$$R_{x\eta\xi} = \begin{pmatrix} \epsilon^2 & 0 & 0 \\ 0 & \sigma^2 I & 0 \\ 0 & 0 & \beta^2 I \end{pmatrix}$$

From (D.23) with the obvious definitions,

$$\tilde{y}_N = h x_0 \tilde{a}_N + h b \phi_0 \underline{u}_N + h \phi_0 \xi_N + \tilde{\eta}_N \quad D.28$$

Using (D.28) and (D.27), the likelihood function is,

$$\begin{aligned} L &= p(y_0, y_1, \dots, y_N; a) \\ &= (2\pi)^{-\frac{N+1}{2}} |\tilde{R}_y|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \left| \tilde{y}_N - h \bar{x}_0 \tilde{a}_N - h b \phi_0 \underline{u}_N \right|_{\tilde{R}_y}^2\right\} \end{aligned} \quad D.29$$

where,

$$\tilde{R}_y = T R_{x\eta\xi} T^T$$

$$T = [h \tilde{a}_N : I : \phi_0]$$

b. The determinant of $\tilde{R}_y$

From (D.29):

$$\tilde{R}_y = TR_{\eta\xi}T^T = h^2\epsilon^2\tilde{A}_N\tilde{A}_N^T + \sigma^2I + \beta^2\phi_0\phi_0^T \quad D.30$$

where,

$$\phi_0\phi_0^T = \left(\begin{array}{c|ccc} 0 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \phi\phi^T & \\ \vdots & & & \\ 0 & & & \end{array} \right)$$

Let

$$\phi_0 = \phi_1 - I_1$$

D.31

where,

$$\phi_1 = \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \phi & \\ \vdots & & & \\ 0 & & & \end{array} \right), \quad I_1 = \left(\begin{array}{c|ccc} 1 & & & 0 \\ \hline & & & \\ 0 & & & \\ & & 0 & \end{array} \right)$$

Note that

$$\begin{aligned} \phi_0\phi_0^T &= \phi_1\phi_1^T - \phi_1I_1^T - I_1\phi_1^T + I_1I_1^T \\ &= \phi_1\phi_1^T - I_1 \end{aligned} \quad D.32$$

Then,

$$\begin{aligned} \tilde{R}_y &= \phi_1 [h^2\epsilon^2\phi_1^{-1}\tilde{A}_N\tilde{A}_N^T(\phi_1^T)^{-1} + \sigma^2\phi_1^{-1}(\phi_1^T)^{-1} \\ &\quad + \beta^2I - \beta^2\phi_1^{-1}I_1(\phi_1^T)^{-1}] \phi_1^T \\ &= \phi_1 [h^2\epsilon^2\tilde{A} + \sigma^2\phi_1^{-1}(\phi_1^T)^{-1} + \beta^2(I - I_1)] \phi_1^T \end{aligned} \quad D.33$$

where:

$$\tilde{A} = \left(\begin{array}{cc|c} 1 & a & 0 \\ a & a^2 & \\ \hline & & 0 \end{array} \right)$$

Let $\tilde{L}_t$ be the determinant of the $t \times t$ matrix having the form of Equation (D.33). Expressing ϕ_1^{-1} in terms of (D.11) and expanding the $t \times t$ matrix gives (noting that $|\phi_1| = 1$):

$$\tilde{L}_t = \begin{bmatrix} h^2\epsilon^2 + \sigma^2 & ah^2\epsilon^2 & 0 & 0 \\ ah^2\epsilon^2 & a^2h^2\epsilon^2 + \beta^2 + \sigma^2 & -a\sigma^2 & 0 \\ 0 & -a\sigma^2 & \beta^2 + \sigma^2(1+a^2) & -a\sigma^2 \\ 0 & 0 & -a\sigma^2 & \beta^2 + \sigma^2(1+a^2) \end{bmatrix}$$

D.34

or,

$$\begin{aligned} \tilde{L}_t &= (h^2\epsilon^2 + \sigma^2) [(h^2\epsilon^2 a^2 + \beta^2 + \sigma^2) (|\beta^2 I + \sigma^2 A_{t-2}|) \\ &\quad - a^2 \sigma^4 (|\beta^2 I + \sigma^2 A_{t-3}|)] - h^4 \epsilon^4 a^2 (|\beta^2 I + \sigma^2 A_{t-2}|) \\ &= [h^2\epsilon^2 (\beta^2 + \sigma^2) + \sigma^2 (h^2\epsilon^2 a^2 + \beta^2 + \sigma^2)] (|\beta^2 I + \sigma^2 A_{t-2}|) \\ &\quad - a^2 \sigma^4 (|\beta^2 I + \sigma^2 A_{t-3}|) \end{aligned}$$

D.35

Using (B.17),

$$\tilde{L}_t = [(\sigma^2 + h^2\epsilon^2) (\beta^2 + \sigma^2) + \sigma^2 h^2\epsilon^2 a^2] \tilde{J}_{t-2} - a^2 \sigma^4 \tilde{J}_{t-3} \quad (t \geq 4) \quad D.36$$

where:

$$\begin{aligned} \tilde{L}_1 &= \sigma^2 + h^2\epsilon^2 \\ \tilde{L}_2 &= (\sigma^2 + h^2\epsilon^2) (\sigma^2 + \beta^2) + h^2\epsilon^2 a^2 \sigma^2 \\ \tilde{L}_3 &= (\sigma^2 + h^2\epsilon^2) [(\sigma^2 + \beta^2)^2 + \sigma^2 \beta^2 a^2] + h^2\epsilon^2 a^2 \sigma^2 [\beta^2 + \sigma^2(1+a^2)] \\ \tilde{J}_p &= \prod_{k=1}^p [\beta^2 + \sigma^2(1+a^2) - 2a\sigma^2 \cos \frac{\pi k}{p+1}] \end{aligned}$$

Thus,

$$|\tilde{R}_y| = \tilde{L}_{N+1} \quad D.37$$

c. The inverse of $\tilde{R}_y$

From (D.33)

$$\tilde{R}_y^{-1} = (\phi_2^T)^{-1} \tilde{R}^{-1} \phi_2^{-1} \quad \text{D.38}$$

where,

$$\tilde{R}^{-1} = (\tilde{r}_{jj}^{-1}) = [h^2 \epsilon^2 \tilde{A} + \sigma^2 \phi_2^{-1} (\phi_2^T)^{-1} + \beta^2 (I - I_1)]^{-1} \quad \text{D.39}$$

Following the same procedure to find (D.21) and (B.19),

$$\tilde{r}_{ji}^{-1} = \tilde{r}_{ij}^{-1} = (\sigma^2 a)^{j-i} (\tilde{L}_{i-1}) (K_{N+1-j}) / (\tilde{L}_{N+1}) \quad , j \geq i \quad \text{D.40}$$

where:

$$\begin{aligned} K_{N+1-j} &= \begin{cases} (h^2 \epsilon^2 a^2 + \beta^2 + \sigma^2) \tilde{J}_{N-1} - a^2 \sigma^4 \tilde{J}_{N-2} & , j = 1 \\ \tilde{J}_{N+1-j} & , 2 \leq j \leq N+1 \end{cases} \\ \tilde{J}_0 &= 1 \\ \tilde{L}_0 &= 1 \end{aligned}$$

d. The likelihood equation

From (D.29), forming $\frac{d \log L}{da} = 0$ gives,

$$\begin{aligned} |\tilde{R}_y|^{-1} \left(\frac{d}{da} |\tilde{R}_y| \right) + \{ (\tilde{y}_N - h\tilde{x}_0 \tilde{a}_N - hb\phi_0 u_N)^T \left(\frac{d}{da} \tilde{R}_y^{-1} \right) \\ - 2[h\tilde{x}_0 \left(\frac{d}{da} \tilde{a}_N \right) + hb \left(\frac{d}{da} \phi_0 \right) u_N]^T \tilde{R}_y^{-1} \} (\tilde{y}_N - h\tilde{x}_0 \tilde{a}_N - hb\phi_0 u_N) \\ = 0 \end{aligned} \quad \text{D.41}$$

D.5 THE DIFFERENCING APPROACH

a. The likelihood function

From (D.1), the equivalent model for this scheme is

$$y_{i+1} = ay_i + hbu_i + h\xi_i + (\eta_{i+1} - a\eta_i) \quad i = 0, 1, \dots \quad \text{D.42}$$

where: y_0 is a known constant

Stacking (D.42):

$$\begin{pmatrix} y_1 - ay_0 \\ y_2 - ay_1 \\ \vdots \\ y_N - ay_{N-1} \end{pmatrix} = hb \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-1} \end{pmatrix} + h \begin{pmatrix} \xi_0 \\ \xi_1 \\ \vdots \\ \xi_{N-1} \end{pmatrix} + \begin{pmatrix} \zeta_0 \\ \zeta_1 \\ \vdots \\ \zeta_{N-1} \end{pmatrix} \quad \text{D.43}$$

where $\zeta_i = \eta_{i+1} - a\eta_i$
or with the obvious definitions

$$y_{1,N} - ay_{0,N} = hbu_N + \xi_N + \zeta_N \quad D.44$$

Because of independence,

$$(\eta_0, \eta_1, \dots, \eta_N, \xi_0, \xi_1, \dots, \xi_{N-1}) \sim \eta(0, Q_{\eta\xi}) \quad D.45$$

where:

$$Q_{\eta\xi} = \begin{pmatrix} \sigma^2 I & 0 \\ 0 & \beta^2 I \end{pmatrix}$$

Also,

$$(\xi_0 + \zeta_0, \dots, \xi_{N-1} + \zeta_{N-1}) \sim \eta(0, R_{\xi\zeta}) \quad D.46$$

where:

$$R_{\xi\zeta} = \Psi Q_{\eta\xi} \Psi^T$$

$$\Psi = [\Omega : hI]$$

$$\Omega = \begin{pmatrix} -a & 1 & 0 \\ 0 & -a & 1 \end{pmatrix}$$

From (D.42), (D.43):

$$y_i = a^i y_0 + \sum_{j=1}^i a^{i-j} (hbu_{j-1} + h\xi_{j-1} + \zeta_{j-1}) \quad i = 1, 2, \dots \quad D.47$$

Since the Jacobian $|\frac{\partial y}{\partial \lambda}|$, $\lambda_i = h\xi_i + \zeta_i$, is one,

$$\begin{aligned} L &= p(y_1, y_2, \dots, y_N; a) \\ &= (2\pi)^{-N/2} |\bar{R}_y|^{-1/2} \exp\{-\frac{1}{2} |y_{1,N}^T - ay_{0,N} - hbu_N|^2_{\bar{R}_y^{-1}}\} \end{aligned} \quad D.48$$

where

$$\bar{R}_y = R_{\xi\zeta}$$

b. The determinant and inverse of $\bar{R}_y$

From (D.48) and (D.46),

$$\bar{R}_y = \sigma^2 \Omega \Omega^T + h^2 b^2 I \quad D.49$$

Then from (B.17) and (B.19)

$$\bar{J}_p = \prod_{k=1}^p [h^2 \beta^2 + \sigma^2 (1+a^2) - 2a\sigma^2 \cos \frac{\pi k}{p+1}] \quad \text{D.50}$$

$$\bar{R}_y^{-1} = (\bar{r}_{ij}^{-1}) \quad \text{D.51}$$

where:

$$\bar{r}_{ij}^{-1} = (\sigma^2 a)^{j-i} (\bar{J}_{i-1}) (\bar{J}_{N-j}) / (\bar{J}_N) \quad , \quad j \geq i \quad \text{D.52}$$

$$\bar{J}_0 = 1$$

$$|\bar{R}_y| = \bar{J}_N \quad \text{D.53}$$

c. The likelihood equation

From (D.48) forming $\frac{d \log L}{da} = 0$ gives,

$$\begin{aligned} & |\bar{R}_y|^{-1} \left(\frac{d}{da} |\bar{R}_y| \right) + \{ (y_{1,N} - ay_{0,N} - hby_N)^T \left(\frac{d}{da} \bar{R}_y^{-1} \right) \\ & - 2(y_{0,N})^T \bar{R}_y^{-1} \} (y_{1,N} - ay_{0,N} - hby_N) \\ & = 0 \end{aligned} \quad \text{D.54}$$

AD-A072 147

CALIFORNIA UNIV LOS ANGELES SCHOOL OF ENGINEERING A--ETC F/G 12/2
MAXIMUM LIKELIHOOD IDENTIFICATION OF LINEAR DISCRETE STOCHASTIC--ETC(U)
JUL 78 A J GLASSMAN, C T LEONDES F33615-77-C-3013

UNCLASSIFIED

AFFDL-TR-78-84

NL

3 OF 3

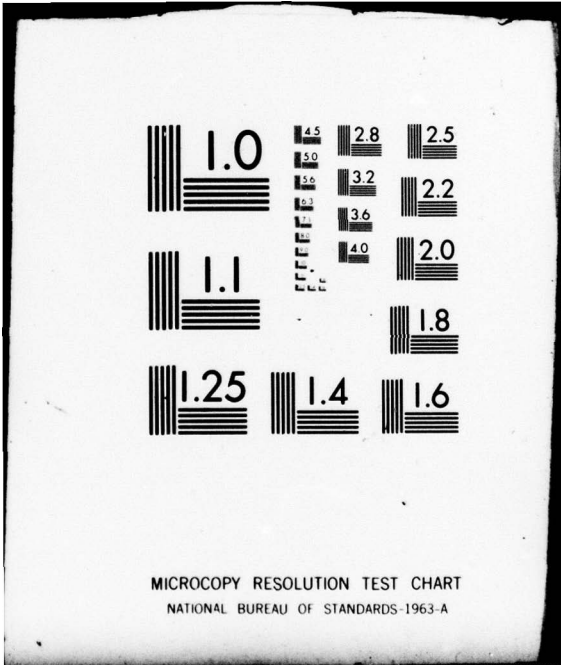
AD
A072147



END
DATE
FILMED
9-79

DDC





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

APPENDIX E

RECURSIONS FOR FIXED POINT CURVE FITTING

E.1 x_0 KNOWN

From Equation (3.9), the derivative function $D_N(a)$ at some point a_1 can be expressed as:

$$D_N(a_1) = \sum_{i=1}^N [(y_i - A_i - hbU_i^1)(iA_{i-1} + hb(U_i^2 - U_i^3))] \quad E.1$$

where:

$$A_i = a_1^i hx_0 \quad E.2$$

$$U_i^1 = \sum_{k=1}^i a_1^{i-k} u_{k-1} \quad E.3$$

$$U_i^2 = \sum_{r=1}^{i-1} ia_1^{i-r-1} u_{r-1} \quad (U_1^2 = 0) \quad E.4$$

$$U_i^3 = \sum_{s=1}^{i-1} sa_1^{i-s-1} u_{s-1} \quad (U_1^3 = 0) \quad E.5$$

The recursions are given below for $n \geq 1$.

$$D_n = D_{n-1} + (y_n - A_n - hbU_n^1)(nA_{n-1} + hb(U_n^2 - U_n^3)) \quad E.6$$

where $D_0 = 0$

$$A_n = a_1 A_{n-1} \quad E.7$$

where $A_1 = a_1 hx_0$

$$U_n^1 = a_1 U_{n-1}^1 + u_n \quad E.8$$

$$\text{where } U_1^1 = u_0 \quad E.9$$

$$U_n^2 = n\left(\frac{a_1}{n-1}\right) U_{n-1}^2 + u_{n-2} \quad E.9$$

where $U_1^2 = 0$

$$U_n^3 = a_1 U_{n-1}^3 + (n-1)u_{n-2} \quad E.10$$

where $U_1^3 = 0$

E.2 x_0 UNKNOWN PARAMETER

This case is similar to the previous except x_0 becomes $\hat{x}_{0N}$ which changes with each new sample. Rearranging (E.1) gives $D_N(a)$ at some point a_1 for this case.

$$D_N(a_1) = D_N^1 + (D_N^2 - D_N^3 - D_N^4 \hat{x}_{0N}) \hat{x}_{0N} \quad \text{E.11}$$

where:

$$D_N^1 = b \sum_{i=1}^N [(y_i - hbU_i^1)(U_i^2 - U_i^3)] \quad \text{E.12}$$

$$D_N^2 = \sum_{i=1}^N [(y_i - hbU_i^1)(ia_1^{i-1})] \quad \text{E.13}$$

$$D_N^3 = hb \sum_{i=1}^N [a_1^i (U_i^2 - U_i^3)] \quad \text{E.14}$$

$$D_N^4 = h \sum_{i=1}^N ia_1^{2i-1} \quad \text{E.15}$$

From Equation (3.20)

$$\hat{x}_{0N} = X_N / Z_N \quad \text{E.16}$$

where:

$$X_N = \sum_{i=0}^N y_i a_1^i - hb \sum_{i=1}^N \sum_{j=1}^i a_1^{2i-j} u_{j-1} \quad \text{E.17}$$

$$Z_N = h \sum_{i=0}^N a_1^{2i} \quad \text{E.18}$$

The recursions for $n \geq 1$ are given below.

$$A_n^1 = a_1 A_{n-1}^1 \quad \text{E.19}$$

where $A_0^1 = 1$

$$X_n = X_{n-1} + y_n A_n^1 - hb A_n^1 U_n^1 \quad \text{E.20}$$

where $X_0 = y_0$

$$z_n = z_{n-1} + h(A_n^1)^2 \quad \text{E.21}$$

$$\text{where } z_0 = h$$

$$D_n^1 = D_{n-1}^1 + b(y_n - hbU_n^1)(U_n^2 - U_n^3) \quad \text{E.22}$$

$$\text{where } D_1^1 = 0$$

$$D_n^2 = D_{n-1}^2 + (y_n - hbU_n^1)(nA_{n-1}^1) \quad \text{E.23}$$

$$\text{where } D_1^2 = y_1 - hbu_0$$

$$D_n^3 = D_{n-1}^3 + hbA_n^1(U_n^2 - U_n^3) \quad \text{E.24}$$

$$\text{where } D_1^3 = 0$$

$$D_n^4 = D_{n-1}^4 + hnA_n^1 A_{n-1}^1 \quad \text{E.25}$$

$$\text{where } D_1^4 = ha_1$$

E.3 x_0 UNKNOWN RANDOM VARIABLE

From Equation (3.45), the derivative function $D_N(a)$ at some point a_1 can be expressed as:

$$\begin{aligned} D_N(a_1) = & \{ (\sigma^2(\sigma^2 + h^2\epsilon^2 C_N^1) + h^2 \bar{x}_0^2 \sigma^2 \Psi) C_N^2 \\ & + C_N^6 (2C_N^2 h \bar{x}_0 \sigma^2 + C_N^3 hb(\sigma^2 + h^2\epsilon^2 C_N^1) + C_N^6 h^2 \epsilon^2 C_N^2 \\ & - 2C_N^4 h^3 b \epsilon^2 C_N^2 + C_N^5 hb(\sigma^2 + h^2\epsilon^2 C_N^1) \\ & + C_N^7 [C_N^4 hb(\sigma^2 + h^2\epsilon^2 C_N^1) - C_N^6 (\sigma^2 + h^2\epsilon^2 C_N^1) \\ & - h \bar{x}_0 \sigma^2 \Psi - \sigma^2 h \bar{x}_0 C_N^1] \\ & - C_N^8 [h^3 b \epsilon^2 (\Psi + C_N^1)^2] \\ & + C_N^4 [C_N^4 C_N^2 h^4 b^2 \epsilon^2 - C_N^3 h^2 b^2 (\sigma^2 + h^2\epsilon^2 C_N^1) \\ & - 2h^2 \bar{x}_0 b \sigma^2 C_N^2] \\ & + C_N^3 [h^2 \bar{x}_0 b \sigma^2 (\Psi + C_N^1)] \\ & + C_N^5 [h^2 \bar{x}_0 b \sigma^2 (\Psi + C_N^1) - C_N^4 h^2 b^2 (\sigma^2 + h^2\epsilon^2 C_N^1)] \\ & + C_N^9 [h^4 b^2 \epsilon^2 (\Psi + C_N^1)] \} \quad \text{E.26} \end{aligned}$$

where:

$$C_N^1 = z_N/h \quad \text{E.27}$$

$$C_N^2 = D_N^4/h \quad \text{E.28}$$

$$C_N^3 = D_N^3/hb \quad \text{E.29}$$

$$C_N^4 = \sum_{i=1}^N a_1^i u_i^1 \quad \text{E.30}$$

$$C_N^5 = \sum_{i=1}^N i a_1^{i-1} u_i^1 \quad \text{E.31}$$

$$C_N^6 = \sum_{i=0}^N a_1^i y_i \quad \text{E.32}$$

$$C_N^7 = \sum_{i=1}^N i a_1^{i-1} y_i \quad \text{E.33}$$

$$C_N^8 = \sum_{i=1}^N [y_i (u_i^2 - u_i^3)] \quad \text{E.34}$$

$$C_N^9 = \sum_{i=1}^N [u_i^1 (u_i^2 - u_i^3)] \quad \text{E.35}$$

$$\Psi = \frac{\sigma^2}{h^2 \epsilon^2} \quad \text{E.36}$$

The recursions for C_N^1 , C_N^2 , and C_N^3 are given by (E.21), (E.25), and (E.24), respectively. The recursions for the remaining terms are given below for $n \geq 1$.

$$C_n^4 = C_{n-1}^4 + A_n^1 u_n^1, \quad C_1^4 = a_1 u_0 \quad \text{E.37}$$

$$C_n^5 = C_{n-1}^5 + n A_{n-1}^1 u_n^1, \quad C_1^5 = u_0 \quad \text{E.38}$$

$$C_n^6 = C_{n-1}^6 + A_n^1 y_n, \quad C_1^6 = y_0 + a_1 y_1 \quad \text{E.39}$$

$$C_n^7 = C_{n-1}^7 + n A_{n-1}^1 y_n, \quad C_1^7 = y_1 \quad \text{E.40}$$

$$C_n^8 = C_{n-1}^8 + y_n (u_n^2 - u_n^3), \quad C_1^8 = 0 \quad \text{E.41}$$

$$C_n^9 = C_{n-1}^9 + u_n^1 (u_n^2 - u_n^3), \quad C_1^9 = 0 \quad \text{E.42}$$

E.4 DIFFERENCING APPROACH (AUTONOMOUS VERSION)

From Equations (3.70), (3.71), and (3.73), the derivative function $D_N(a)$ at some point a_1 can be expressed as:

$$D_N(a_1) = -\sigma^2 (F_N^1) (F_N^2) + \sum_{i=0}^N \{ (A_1^1)^2 (F_1^3) [(F_1^4) - 1(F_1^3)] \} \quad E.43$$

where:

$$F_n^1 = C_n^1 \quad E.44$$

$$F_n^2 = C_n^2 \quad E.45$$

$$F_n^3 = C_n^6 \quad E.46$$

$$F_n^4 = C_n^5 \quad E.47$$

The recursion becomes:

$$D_n = -\sigma^2 (F_n^1) (F_n^2) + F_n^5 \quad E.48$$

$$F_n^5 = F_{n-1}^5 + (A_n^1)^2 (F_n^3) [F_n^4 - nF_n^3], \quad F_0^5 = 0 \quad E.49$$

BIBLIOGRAPHY

- Adams, D., "A Stopping Criterion for Polynomial Root Finding," ACM Communications, 10, 655-8 (October 1967).
- Akaike, H., "Some Problems in the Application of the Cross-Spectral Method," Spectral Analysis of Time Series, B. Harris, ed., New York, Wiley, 1967. pp. 81-107.
- Albert, A.E. and L.A. Gardner, Stochastic Approximation and Nonlinear Regression, Cambridge, Mass., M.I.T. Press, 1967.
- Anderson, T.W., An Introduction to Multivariate Statistical Analysis, New York, Wiley, 1958.
- Anderson, T.W., "Determination of the Order of Dependence in Normally Distributed Time Series," Time Series Analysis, M. Rosenblatt, ed., New York, Wiley, 1963.
- Aoki, M., "On Identification of Constrained Dynamic Systems with High Dimensions," Allerton Conference on Circuit and System Theory; Proceedings, October 1967 (a), pp. 191-200.
- Aoki, M., Optimization of Stochastic Systems, New York, Academic Press, 1967 (b). ch. 3.
- Aoki, M. and P. Yue, "On Certain Convergence Questions in System Identification," SIAM J. Control, 8, 239-255 (May 1970).
- Astrom, K.J., "Computer Control of a Paper Machine - An Application of Linear Stochastic Control Theory," IBM Journal of Research and Development, 11, 389-405 (July 1967).
- Astrom, K.J. and T. Bohlin, "Numerical Identification of Linear Dynamic Systems from Normal Operating Records," IBM Development Division, Nordic Laboratory, Sweden, Technical Paper, TP 18.159, July 15, 1967. (Also, Proceedings of the Second IFAC Symposium on the Theory of Self-Adaptive Control Systems, September 14-17, 1965, Teddington, England).
- Balakrishnan, A.V. and V. Peterka, "Identification in Automatic Control Systems," Automatica, 5, 817-829 (November 1969).
- Barndorff-Nielsen, O. and K. Pedersen, "Sufficient Data Reduction and Exponential Families," Mathematica Scandinavica, 22(1), 197-202 (1968).
- Barnett, V., "Evaluation of the Maximum-Likelihood Estimator Where the Likelihood Function has Multiple Roots," Biometrika, 53, 151-165 (June 1966).

- Beckenbach, E. and R. Bellman, Inequalities, Berlin, Springer-Verlag, 1961.
- Berkson, J., "Estimation by Least Squares and by Maximum Likelihood," Third Berkeley Symposium of Mathematical Statistics and Probability, 1955, J. Neyman, ed., Berkeley, University of California Press, 1956.
- Braun, L., "On Adaptive Control Systems," IRE Trans. Automatic Control, AC-4, 30-42 (May 1959).
- Brown, L., "Sufficient Statistics in the Case of Independent Random Variables," Ann. Math. Stat., 35, 1456-1474 (December 1964).
- Bryson, A. and M. Frazier, "Smoothing for Linear and Nonlinear Dynamic Systems," Proceedings of the Optimum System Synthesis Conference, Wright Patterson Air Force Base, ASD-TDR-63-11, September 1962, pp. 25-34.
- Budin, M.A., "Estimator for Identification of Constant Linear Discrete Systems," IEEE Trans. Automatic Control, AC-12, 193-194 (April 1969).
- Busk, T. and B. Svejgaard, "Polynomial Equations," Selected Numerical Methods, C. Gram, ed., Copenhagen, Regnecentralen, 1962.
- Cochrane, D. and G. Orcutt, "Applications of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," J. Am. Stat. Assoc., 44, 32-61 (March 1949).
- Cox, H., "On the Estimation of State Variables and Parameters for Noisy Dynamic Systems," IEEE Trans. Automatic Control, AC-9, 5-12 (January 1964).
- Cramér, H., Mathematical Methods of Statistics, Princeton, Princeton University Press, 1946 (1966 printing).
- Currie, Malcolm George, Study and Application of Adaptive Systems, Ph.D. Dissertation in Engineering, University of California, Los Angeles, 1968.
- Detchmendy, D.M. and R. Sridhar, "Sequential Estimation of States and Parameters in Noisy Nonlinear Dynamical Systems," Joint Automatic Control Conference, Troy, New York, 1965.
- Dolbin, B.H., "A Differential Correction Method for the Identification of Airplane Parameters from Flight Test Data," Proc. National Electronics Conf., 25, 1969, pp. 90-94.
- Dvoretzky, A., "On Stochastic Approximation," Third Berkeley Symposium on Mathematical Statistics and Probability, 1955, 1, Berkeley, University of California Press, 1956. pp. 39-55.

- Dynkin, E.B., "Necessary and Sufficient Statistics for a Family of Probability Distributions," Select. Transl. Math. Stat. and Prob., 1, 17-40 (1961). (Translation of the original paper in Russian from 1951.)
- Edgeworth, F.Y., "On the Probable Errors of Frequency-Constants," J. Roy. Stat. Soc., 71, 381-397 (June 1908).
- Elkind, J.I., D.M. Green and E.A. Starr, "Application of Multiple Regression Analysis to Identification of Time-Varying Linear Dynamic Systems," IEEE Trans. Automatic Control, AC-8, 163-166 (April 1963).
- Eykhoff, P., "Some Fundamental Aspects of Process Parameter Estimation," IEEE Trans. Automatic Control, AC-8, 347-357 (October 1963).
- Farison, J.B., "Parameter Identification for a Class of Linear Discrete Systems," IEEE Trans. Automatic Control, AC-12, 109 (February 1967).
- Finney, D.J., Statistics for Mathematicians, Edinburgh, Oliver and Boyd, 1968.
- Gagliardi, R.M., "Input Selection for Parameter Identification in Discrete Systems," IEEE Trans. Automatic Control, AC-12, 597-599 (October 1967).
- Gibson, J.E., Nonlinear Automatic Control, New York, McGraw-Hill, 1963.
- Goldberger, A.S., Econometric Theory, New York, Wiley, 1964.
- Goodman, T.P., "Determination of the Characteristics of Multi-Input and Nonlinear Systems from Normal Operating Records," Trans. ASME, 79, 567-575 (April 1957).
- Goodman, T.P. and J.B. Reswick, "Determination of System Characteristics from Normal Operating Records," Trans. ASME, 78, 259-271 (February 1956).
- Grenander, U. and G. Szego, Toeplitz Forms and Their Applications, Berkeley, University of California Press, 1958.
- Hannan, E.J., Time Series Analysis, London, Methuen, 1960.
- Ho, Y.C. and R.C.K. Lee, "A Bayesian Approach to Problems in Stochastic Estimation and Control," IEEE Trans. Automatic Control, AC-9, 333-339 (October 1964).
- Ho, Y.C. and R.C.K. Lee, "Identification of Linear Dynamic Systems," Information and Control, 8, 93-110 (February 1965).

- Ho, Y.C. and B.H. Whalen, "An Approach to the Identification and Control of Linear Dynamic Systems with Unknown Parameters," IEEE Trans. Automatic Control, AC-8, 255-256 (July 1963).
- Hoppe, S.G., "A Least Squares Technique for the Identification of a Linear Time-Invariant Plant in a Sampled-Data System," IEEE Trans. Automatic Control, AC-10, 490-491 (October 1965).
- Horowitz, B.M. and A.J. Grammaticos, "A New Approach to Certain Problems of Identification and Control," IEEE Trans. Automatic Control, AC-15, 475-477 (August 1970).
- Hsia, T.C. and V. Vimolvanich, "An On-Line Technique for System Identification," IEEE Trans. Automatic Control, AC-14, 92-96 (February 1969).
- Jennrich, R. "Asymptotic Properties of Non-Linear Least Squares Estimators," Ann. Math. Stat., 40(2), 633-643 (1969).
- Jennrich, R. and P. Sampson, "Application of Stepwise Regression to Non-Linear Estimation," Technometrics, 10, 63-72 (February 1968).
- Joseph, P., J. Lewis and J. Tou, "Plant Identification in the Presence of Disturbances and Application to Digital Adaptive Systems," Trans. AIEE, pt. II, 80, 18-24 (March 1961).
- Joseph, P.D. and J.T. Tou, "On Linear Control Theory," Trans. AIEE, pt. II, 80, 193-196 (September 1961).
- Kac, M., "On the Average Number of Real Roots of a Random Algebraic Equation," Bulletin of the American Mathematical Society, 49, 314-320 (April 1943).
- Kac, M., Probability and Related Topics in Physical Sciences, New York, Interscience, 1959.
- Kale, B., "On the Solution of the Likelihood Equation by Iteration Processes," Biometrika, 48, 452-6 (December 1961).
- Kale, B., "On the Solution of Likelihood Equations by Iteration Processes. The Multiparametric Case," Biometrika, 49, 479-486 (December 1962).
- Kalman, R.E., "Design of a Self-Optimizing Control System," Trans. ASME, ser. D, (J. Basic Eng'g.), 80, 468-478 (February 1958).
- Kalman, R.E., "A New Approach to Linear Filtering and Prediction Problems," Trans. ASME, ser. D, (J. Basic Eng'g.), 82, 35-45 (March 1960).

- Kalman, R.E. and R.S. Bucy, "New Results in Linear Filtering and Prediction Problems," Trans. ASME, ser. D, (J. Basic Eng'g.), 83, 95-108 (March 1961).
- Kashyap, R.L., "Maximum Likelihood Identification of Stochastic Linear Systems," IEEE Trans. Automatic Control, AC-15, 25-34 (February 1970).
- Kekre, H.B. and G.S. Glenski, "Identification of Impulse Response of Linear Systems and Synthesis of System Model," Int'l J. Control, 7(4), 317-331 (1968).
- Kendall, M.G. and A. Stuart, The Advanced Theory of Statistics, New York, Hafner, 1961. Vol. 2.
- Kerr, R.B. and W.H. Surber, "Precision of Impulse-Response Identification Based on Short, Normal Operating Records," IRE Trans. Automatic Control, AC-6, 173-182 (May 1961).
- King, R.P., "Estimation of Parameters in Systems Defined by Differential Equations," South African J. Sci., 63(1), 91-96 (1967).
- Koopmans, T., Linear Regression Analysis of Economic Time Series, N.V. Haarlem, The Netherlands, De Erven F. Bohn, 1937.
- Kopp, R.E. and R.J. Orford, "Linear Regression Applied to System Identification for Adaptive Control Systems," AIAA J., 1, 2300-2306 (October 1963).
- Kroy, W.H. and A.R. Stubberud, "Identification Via Nonlinear Filtering," Int'l J. Control, 6(6), 499-522 (1967).
- Kumar, K.S.P., "Identification of Nonlinear, Nonstationary Processes," Proceedings IFAC Tokyo Symposium on Systems Design, 1965, pp. 237-243.
- Kumar, K.S.P. and R. Sridhar, "On the Identification of Control Systems by the Quasi-Linearization Method," IEEE Trans. Automatic Control, AC-9, 151-154 (April 1964).
- Lavi, A. and J.C. Strauss, "Parameter Identification in Continuous Dynamic Systems," IEEE International Convention Record, 1965, Pt. 6, pp. 49-61.
- Lee, E.S., "Quasilinearization and the Estimation of Parameters in Differential Equations," Industrial and Engineering Chemistry, Fundamentals, 7, 152-158 (February 1968) (a).
- Lee, E.S., "Invariant Imbedding, Nonlinear Filtering, and Parameter Estimation," Industrial and Engineering Chemistry, Fundamentals, 7, 164-171 (February 1968) (b).

- Lee, R.C.K., Optimal Estimation, Identification, and Control, Cambridge, Mass., M.I.T. Press, 1964.
- Lendaris, G.G., "The Identification of Linear Systems," Trans. AIEE, pt. II, 81, 231-242 (September 1962).
- Levin, M.J., "Optimum Estimation of Impulse Response in the Presence of Noise," IRE Trans. Circuit Theory, CT-7, 50-56 (March 1960).
- Levin, M.J., "Estimation of a System Pulse Transfer Function in the Presence of Noise," IEEE Trans. Automatic Control, AC-9, 229-235 (July 1964).
- Lion, P.M., "Rapid Identification of Linear and Nonlinear Systems," AIAA J., 5, 1835-1842 (October 1967).
- Luenberger, D.G., "Observing the State of a Linear System," IEEE Trans. on Military Electronics, 74-80 (April 1964).
- Lyusternick, L. et al., Computing Elementary Functions, Oxford, Pergamon Press, 1965.
- Mann, H.B. and A. Wald, "On the Statistical Treatment of Linear Stochastic Difference Equations," Econometrica, 11, 173-220 (July 1943).
- Margolis, M. and C.T. Leondes, "On the Philosophy of Adaptive Control for Plant Adaptive Systems," Proc. National Electronics Conf., 1959, pp. 27-33.
- Mayne, D.Q., "Parameter Estimation," Automatica, 3, 245-255 (January 1966).
- Mortensen, R.E., "Maximum-Likelihood Recursive Nonlinear Filtering," Journal of Optimization Theory and Applications, 2(6), 387-394 (1968).
- Parzen, E., "An Approach to Time Series Analysis," Ann. Math. Stat., 32, 951-989 (December 1961).
- Rao, C.R., Advanced Statistical Methods in Biometric Research, New York, Wiley, 1952.
- Rao, C.R., Linear Statistical Inference and Its Applications, New York, Wiley, 1965.
- Reiersol, O., "Confluence Analysis by Means of Lag Moments and Other Methods of Confluence Analysis," Econometrica, 9, (1), 1-23 (1941).
- Révész, P., The Laws of Large Numbers, New York, Academic Press, 1968.

- Rogers, A.E. and K. Steiglitz, "Maximum Likelihood Estimation of Rational Transfer Function Parameters," IEEE Trans. Automatic Control, AC-12, 594-597 (October 1967).
- Roy, R. and K.W. Jenkins, "Identification and Control of a Flexible Launch Vehicle," NASA Contractor Report CR-551 (August 1966).
- Sakrison, D.J., "The Use of Stochastic Approximation to Solve the System Identification Problem," IEEE Trans. Automatic Control, AC-12, 563-567 (October 1967).
- Sargan, J.D., "The Estimation of Econometric Relationships Using Instrumental Variables," Econometrica, 26, 393-415 (July 1958).
- Saridis, G.N. and G. Stein, "Stochastic Approximation Algorithms for Linear Discrete-Time System Identification," IEEE Trans. Automatic Control, AC-13, 515-523 (October 1968).
- Sawaragi, Y. and T. Katayama, "On the Parameter Estimation for Noisy Discrete-Time Nonlinear Dynamical Systems," Engineering Research Institute, Kyoto University, Technical Report 135, October 1967.
- Schultz, E.R., "Estimation of Pulse Transfer Function Parameters by Quasilinearization," IEEE Trans. Automatic Control, AC-13, 424-426 (August 1968).
- Smith, F.W., "System Laplace Transform Estimation from Sampled Data," IEEE Trans. Automatic Control, AC-13, 39-47 (February 1968).
- Smith, F.W. and W.B. Hilton, "Estimation of the Laplace Transfer Function from Sampled Data," Sylvania Electronic Systems, Mountain View, California, Technical Memo EDC-M874, November 30, 1965.
- Smith, F.W. and W.B. Hilton, "Monte Carlo Evaluation of Rational Transfer Function Parameters," IEEE Trans. Automatic Control, AC-12, 568-576 (October 1967).
- Staley, Robert Michael, Input Signal Synthesis in Identification Problems, Ph.D. Dissertation in Engineering, University of California, Los Angeles, 1968.
- Stear, E.B., "Estimation and Filtering Theory for Digital Computer Control Applications," Computer Control Workshop; JACC Proceedings, Denver, 1970.
- Steiglitz, K. and L.E. McBride, "A Technique for the Identification of Linear Systems," IEEE Trans. Automatic Control, AC-10, 461-464 (October 1965).
- Surber, W.H., "On the Identification of Time-Varying Systems," Identification Problems in Communication and Control Systems, Princeton University, March 1963.

Turin, G.L., "On the Estimation in the Presence of Noise of the Impulse Response of a Random, Linear Filter," IRE Trans. Information Theory, IT-3, 5-10 (March 1957).

Wald, A. "Asymptotic Properties of the Maximum Likelihood Estimate of an Unknown Parameter of a Discrete Stochastic Process," Ann. Math. Stat., 19, 40-46 (March 1948).

Weygandt, C.N. and N.N. Puri, "Transfer Function Tracking and Adaptive Control Systems," IRE Trans. Automatic Control, AC-6, 162-166 (May 1961).

Weygandt, C.N. and N.N. Puri, "Identification of a Linear System from Discrete Values of Input-Output Data," Allerton Conference on Circuit and System Theory; Proceedings, 4, 1966, pp. 903-912.

Wilkinson, J., Rounding Errors in Algebraic Processes, Englewood Cliffs, N.J., Prentice-Hall, 1963.

Wold, H., A Study in the Analysis of Stationary Time Series, Stockholm, Almqvist and Wiksell, 1954. Sec. 1.

Wong, K.Y. and E. Polak, "Identification of Linear Discrete Time Systems Using the Instrumental Variable Method," IEEE Trans. Automatic Control, AC-12, 707-718 (December 1967).

Zabusky, N.J., "A Numerical Method for Determining a System Impulse Response from the Transient Response to Arbitrary Inputs," IRE Trans. Automatic Control, AC-1, 40-55 (May 1956).

Zhuravlev, O.G., "Minimal Sufficiency Statistics for a Sequence of Independent Random Variables," Theory of Probability and its Applications, 8(2), 218-220 (1963).