AN ABSTRACT OF THE THESIS OF

Donald L. Stevens, Jr. for the degree of Doctor of Philosophy

in Statistics presented on March 8, 1979.

Title: SMALL SAMPLE PARAMETER ESTIMATION FOR FORCED DISCRETE LINEAR DYNAMIC MODELS

Abstract approved: Redacted for Privacy
Professor W. S. Overton

The problem of estimating the parameters of a forced discrete linear dynamic model is considered. The system model is conceptualized to include the value of the initial state as a parameter. The forces driving the system are partitioned into accessible and inaccessible inputs. Accessible inputs are those that are measured; inaccessible inputs are all others, including random disturbances.

Maximum likelihood and mean upper likelihood estimators are derived. The mean upper likelihood estimator is a variant of the mean likelihood estimator and apparently has more favorable small sample properties than does the maximum likelihood estimator. A computational algorithm that does not require the inversion or storage of large matrices is developed.

The estimators and the algorithm are derived for models having an arbitrary number of inputs and a single output. The extension to a two output system is illustrated; further extension to an arbitrary number

of outputs follows trivially.

The techniques were developed for the analysis of possibly unique realizations of a process. The assumption that the inaccessible input is a stationary process is necessary only over the period of observation. Freedom from the more general usual assumptions was made possible by treatment of the initial state as a parameter. The derived estimation technique should be particularly suitable for the analysis of observational data.

Simulation studies are used to compare the estimators and assess their properties. The mean upper likelihood estimator has consistently smaller mean square error than does the maximum likelihood estimator.

An example application is presented, representing a unique realization of a dynamic system. The problems associated with determination of concurrence of a hypothetical "system change" with a temporally identified event are examined, and associated problems of inference of causality based on observational data are discussed with respect to the example.

Small Sample Parameter Estimation for Forced Discrete Linear Dynamic Models

by

Donald L. Stevens, Jr.

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Completed March 8, 1979

Commencement June 1979

APPROVED:

Redacted for Privacy

Professor of Statistics
in charge of major

Redacted for Privacy

Chairman of Department of Statistics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented March 8, 1979

Typed by Mary Lou Lomon for Donald L. Stevens, Jr.

Table of Contents

List of Figures

List of Tables

SMALL SAMPLE PARAMETER ESTIMATION FOR FORCED DISCRETE LINEAR DYNAMIC MODELS

I. INTRODUCTION

One of the first levels of sophistication in modeling a dynamic system is the construction of a linear dynamic model. Such models have found extensive application in such areas as ecosystem studies, economics, radio-biology, pharmacokinetics, and engineering. The models can be derived from often reasonable elementary assumptions, for instance, the assumption that transfer rates are proportional to levels.

Even though it is becoming generally recognized that many interesting systems exhibit non-linear behavior, the linear dynamic model is nevertheless a useful tool for preliminary study of a system. The behavioral properties of linear system models have been extensively studied and are well characterized (De Russo, et al., 1965; Freeman, 1965; Meditch, 1969). The model may be an adequate first approximation, and the attempt to fit the model to data can provide valuable insight. Moreover, if a non-linear system is near equilibrium and remains so during the period of observation, the linear system approximation may be adequate. In addition, there are many instances in which a sub-process can be represented by a linear model even if the entire system cannot.

The modeling process does not end with the selection of the structural form of the model. Parameter values must be identified or estimated. In some instances the parameters of a linear dynamic model can be directly related to measurable physical quantities. Fluxes or flow rates are often important system parameters, and in some instances can be measured. Generally, such measurements are difficult or impossible to make and the parameters of the dynamic linear model must be estimated from observations of the inputs and outputs of the system, or from the time-sampled observations of the state variables.

Given the wide range of application of linear dynamics models, it is not surprising that the above problem has been addressed by several distinct disciplines. Each discipline has developed its own peculiar approach, emphasis, assumptions, model forms, and terminology. Four main types or classifications of models can be distinguished. These are signal flow models, econometric models, compartment models, and time series models. Roughly these model types correspond to the disciplines of engineering, economics, the life sciences, and statistics.

- Signal Flow Models

Linear systems arise naturally in many branches of engineering, and particularly in electrical engineering. Much of the work on characterizing linear system behavior has been done by electrical engineers and has appeared first in the electrical engineering literature. A popular form is the 'signal flow system' leading to a specific paradigm which has greatly influenced linear system theory. Signal

flow graphs or diagrams provide a simple means of obtaining transfer functions (briefly, a transfer function is the ratio of the Laplace transforms of an output and an input). Corresponding to the representation of systems by transfer functions has been an emphasis in the "engineering approach" on means of obtaining the transfer function.

Another characteristic of the engineering approach is the extensive use of state variable models. Again the signal flow diagram provides a vehicle for obtaining state equations. The state variable model is a powerful tool for the investigation of the behavior of linear systems.

- Econometric Models

Econometric models are often characterized by having few observations on many related variables measured with error. In addition many economic systems are dynamic: their history influences their future. This has long been recognized in models that include lagged values of both exogeneous and endogenous variables. The "econometric approach" to the estimation problem engendered by dynamic systems appears to have evolved from the concern over correlation between variables used as predictors and the error term. Such techniques as 'two-stage least square' and 'instrumental variables', which were developed to handle 'error-in-variables bias', are applied to the estimation problem for dynamic systems.

- Compartment Models

Another conceptualization that often leads to a linear representation is the compartment model. A compartment model is useful when compartments or "pools" and some quantity that "flows" between the compartments can be identified. For instance an ecosystem might be modeled by taking trophic levels as compartments and identifying the flow of biomass between trophic levels.

The construction of a linear compartment model usually goes through the following steps:

- Identification of the compartments.
- Identification of flow pathways (of material, energy, biomass, etc.) between compartments.
- Writing down difference (or differential) equations describing the dynamics, which make the parameters explicit.

In many applications of compartment models the output is represented as a sum of exponential terms. Graphical methods and non-linear regression have been used to estimate parameters.

These models also lead naturally to a state variable representation. In a first order system the state variables represent the content of the compartments. In a higher order discrete system other state variables represent past values of compartment contents.

- Time Series Models

The time series approach is generally used in situations where

large amounts of data are available. Moreover the focus of the approach is on the noise in the system; much of time series work is concerned with methods for characterization of noise structure under the assumption that deterministic structure is absent or can be easily removed. The methods in use can be divided into two main areas: time domain methods and frequency domain methods. The most popular time domain methods involve the representation of a process as a mixed autoregressive moving average (ARMA) model. The frequency domain methods are essentially non-parametric. They can be thought of as means of obtaining system transfer functions without estimating parameters.

In all of the approaches mentioned above, the underlying mathematical models are nearly identical, differing mostly in the specification of noise structure. However, the essential similarity of the models is often masked by the terminology used to describe them. The economist's simultaneous equation model with lagged endogeneous variables is the engineer's forced linear system model and the ecologist's driven compartment model. The ARMA model of time series is also a linear system model, although a forcing or control function is usually not included.

Linear system terminology and state variable format will be used in the following development. This format has a natural interpretation for all of the approaches giving rise to these models. It is also a form that is easily obtained from the differential or difference equations that are often used to describe the dynamics of a linear system. This is a common underlying general linear system structure.

The intent behind this research was to develop an estimation pro-

cedure as a part of a paradigm for modeling environmental systems. The data are envisioned as being collected at discrete, regularly spaced time intervals. Hence a development of the model in terms of difference equations was natural. There is no real limitation, since a natural correspondence exists between constant coefficient linear difference and differential equations.

The modeling effort is conceived of as taking place in a modified "black box" environment, and as being a part of a paradigm for the study of an environmental system. The effort is not merely to fit equations to data; rather, the effort is to understand the phenomena that are being modeled. The structure of the model is expected to mimic the structure of the real system. The model is intended to be realistic and not just an empirical representation of behavior. The model is regarded as one step of an iterative process in which current knowledge and theory is tied to data in order to both test the theory and give direction to future data collection efforts.

It is not assumed that the system being modeled is fully understood. Rather, the explicit assumption is made that the external qualities of the system have been only partially identified. The system may be driven by an input that has not been accounted for in the observation protocol that was used to collect the data. It may also be that an input cannot be measured, even if one were clever enough to identify it and acknowledge its importance. For instance the introduction of a probe without destroying the system may not be technically feasible. Inputs to the system that have been identified in the theory and are

capable of being measured will be called accessible; all other inputs to the system will be called inaccessible. Note that inaccessible inputs includes inputs that statisticians would call the error term, and an engineer would call noise. It may also include other inputs, the discovery of which may well be an important research activity.

The state variable representation of a p^{th} order discrete linear system is

$$z(k+1) = \Phi z(k) + Bu(k) + \Theta w(k) \quad . \tag{1.1}$$

Here z(k) is the pxl state vector at time k, u(k) and w(k) are vectors of inputs at time k, Φ is a pxp state transition matrix, B is a pxr accessible input transition matrix, and Θ is a pxq inaccessible input transition matrix. The vector u(k) is regarded as an accessible input and w(k) as an inaccessible input.

o Approaches in the Literature

A variety of methods for estimating the parameters of (1.1) have been proposed. Since much of the development of systems theory has been carried out by electrical engineers, it is not surprising that some implicit assumptions of the engineering approach to system identification are that the system is available for experimentation and that data collection is easy and inexpensive. Thus, the outputs can be observed for inputs that have been carefully chosen.

The easy and often very quick availability of data has made on-line techniques attractive. A number of on-line methods have been pro-

posed (Prasad and Sinha, 1977, Hsia, 1976, Desai and Oldenburger, 1969, Steiglitz and McBride, 1965). The extended Kalman filter (Sage and Melsa, 1971) is an on-line technique in which unknown parameters are estimated as part of an augmented state variable.

Stochastic approximation is also essentially an on-line technique. Saridis and Stein (1968a) consider several stochastic approximation algorithms for identifying parameters. Briefly a stochastic approximation algorithm is based on expressing a parameter estimate as a function of a previous estimate and additional data. The authors use a model similar to (1.1), with Φ having the special form

$$\Phi = \left|\begin{array}{c|c} 0 & I \\ \hline \multicolumn{2}{c}{\phi'} \end{array}\right|$$

where $\phi' = (\phi_1, \phi_2, \ldots, \phi_p)$. B is taken as a pxl matrix, Θ is assumed to be known, and the noise properties are assumed to be known. In addition it is assumed that both the input and output are observed with error, so observation equations of the form

$$y(k) = (1,0,0,\ldots,0)'z(k) + \eta(k)$$
$$m(k) = u(k) + \nu(k)$$

are included, where $\{\eta(k)\}$ and $\{\nu(k)\}$ are white noise.

The algorithms are based on minimizing the objective function

$$E\left[(y(k+1) - \hat{\phi}'z(k) - \hat{B}'u(k) - \Theta w(k))^2\right] ,$$

where $E[\cdot]$ is the expectation operator, which leads to an estimator of the form

$$\begin{pmatrix}\hat{\phi}\\ \hat{\beta}\end{pmatrix} = [E(YY')]^{-1} E[Yy]$$

where $Y = (y(k), y(k+1),\ldots,y(k+p-1), m(k),\ldots,m(k+p-1))'$

which can be written as

$$\hat{\phi} = \phi + [E(YY')]^{-1} E[Y\xi]$$

where ξ is a composite error term.

The authors note that the estimates that minimize the objective function are biased, and they present several algorithms that correct for bias. The first assumes all noise characteristics are known. Two additional algorithms relax that assumption but require repeated runs of the system's transient response. For each run 2p measurements of the output and input are collected. The algorithms assume the initial states to be drawn from an unknown but stationary probability distribution.

Sairdis and Stein (1968b) present another stochastic approximation algorithm that does not require knowledge of the noise statistics. In this case the accessible input {u(k)} is assumed to be a sequence of independent, zero mean random variables, and the system is assumed to be in a stationary condition.

In each of the above algorithms convergence in mean square has been obtained. In the procedures that do not require knowledge of the noise covariance, no provision is made for estimating the covariance. Con-

vergence seems to be rather slow. An updated estimate of the parameter vector is obtained once every p or 2p measurements, and in the examples given, from 100 to 10000 updates were used.

Although the above techniques do not appear to be useful for the problem at hand, they might be useful in updating estimates as additional data are collected.

Kashyap (1970) has presented a derivation of the maximum likelihood estimators of the parameters of a discrete stationary linear system. The model is equivalent to (1.1), but stated in difference equation format as

$$Z(i) + \sum_{j=1}^{n} A_j Z(i-j) = \sum_{j=1}^{n} C_j U(i-j) + W(i) .$$

The characteristics of W are specified by the covariance matrices F_j, $j = 0,1,\ldots,n$ where

$$F_j = \begin{cases} E[W(i)\ W(i-j)'] & , \quad j = 0,1,\ldots,n \\ 0 & , \quad j > n \end{cases}$$

A set of coefficient matrices B_j, $j = 1,\ldots,n$ are defined implicitly by

$$F_i = \sum_{j=0}^{n-i} \left((B_j\ R_e\ B'_{i+j} - A_j\ A'_{i+j}) \right) , \quad i = 0, \quad , \quad n \quad ,$$

where $R_e = \mathrm{Cov}(Z(i))$. Estimates of A_j, B_j, and C_j, $j = 1,\ldots,n$ are obtained by minimizing

$$\det\left(\sum_{i=1}^{N} e(i)\ e'(i)/N \right)$$

where N is the sample size and the e(i) are defined by

$$e(i) + \sum_{j=1}^{n} B_j\, e(i-j) = Z(i) + \sum_{j=1}^{n} (A_j\, Z(i-j) - C_j\, U(i-j)).$$

The result is essentially a large sample result, since some expressions were replaced by limiting values. The method also is computationally difficult, requiring the solution of a constrained optimization problem.

Bellman et al. (1965) present a method for parameter identification based on the numerical inversion of Laplace transforms. The system is observed at time points that correspond to the ordinates of numerical quadrature formulas. The sum of squared differences of observation and model in the transformed domain is minimized. No provision is made for the presence of noise in the system.

Stoica and Söderström (1977) develop an iterative algorithm to estimate the parameters of (1.1). At each stage of the iteration estimates of Φ and B are obtained by least squares using observations that have been whitened through filters obtained in the previous stages. The filter for the j^{th} stage is obtained by minimizing

$$\sum_i ((1+\theta F)^{-1}\, d_j(i))^2$$

where F is the forward shift operator defined by $F(d(i)) = d(i+1)$, and $d_j(i) = Z(i) - \hat{Z}_j(i)$ is the i^{th} residual at the j^{th} stage of the iteration.

The approaches in the econometric literature seem to place great stress on the consistency of an estimator. Although consistency is certainly a desirable property, it in itself is not a sufficient criterion. For instance minimum mean square error might be a more relevant criterion for the small samples that are common in econometrics. However, the estimators of linear system parameters commonly in use in econometrics have been shown simply to be consistent.

Griliches (1967) recommends two stage least squares for estimating the parameters of a model of the form

$$z_t = \alpha z_{t-1} + \beta u_t + \varepsilon_t$$

when the $\{\varepsilon_t\}$ process is autocorrelated. In this estimation technique, z_t is regressed on u_t, u_{t-1}, ..., u_{t-k}, where k is chosen large enough to obtain a "reasonable" estimator $\hat{z}_t(u)$. Then α and β are estimated using least squares and the model

$$z_t = \alpha \hat{z}_{t-1}(u) + \beta u_t \quad .$$

Johnston (1972, pp. 316-320) discusses the estimation problems associated with dynamic linear models. He suggests use of a two stage least squares procedure that is best explained by example. Let the model be

$$Z_t = \beta_0 + \beta_1 Z_{t-1} + \beta_2 U_t + W_t$$

where $W_t = \rho W_{t-1} + \varepsilon_t$, and $\{\varepsilon_t\}$ is a white noise sequence. Then

$$\rho Z_{t-1} = \rho\beta_0 + \rho\beta_1 Z_{t-2} + \rho\beta_2 U_{t-1} + \rho W_{t-1}$$

so that

$$Z_t = \beta_0(1-\rho) + (\beta_1+\rho)Z_{t-1} - \beta_1\rho Z_{t-2} + \beta_2 U_t$$

$$- \beta_2\rho U_{t-1} + \varepsilon_t \ .$$

The coefficients of the above equation are estimated using least squares and an estimate of ρ is obtained from

$$\hat{\rho} = \frac{\hat{\beta_2\rho}}{\hat{\beta}_2}$$

This estimate is used to compute

$$\tilde{Y}_t = Z_t - \hat{\rho}Z_{t-1} \quad \text{and} \quad \tilde{X}_t = U_t - \hat{\rho}U_{t-1} \ ,$$

and β_1 and β_2 are estimated by least squares from

$$\tilde{Y}_t = \beta_0(1-\hat{\rho}) + \beta_1\tilde{Y}_{t-1} + \beta_2\tilde{X}_t + \varepsilon_t \ .$$

Fair (1970) develops a procedure that is a mixture of instrumental variables and two stage least squares. The procedure requires a first

stage regression to obtain an instrument $\hat{z}$ for the second stage. The second stage regression is repeated in either a grid search or an iteration to locate a least squares estimate of the noise serial correlation parameter. A similar method is given by Dhrymes et al. (1974).

In general the methods commonly used in econometrics to estimate linear system parameters are limited to first order noise models and do not appear to be conveniently extended. Also the parameters are often estimated sequentially, not simultaneously. Even though such estimates may be consistent, they in general will not be minimum variance nor minimum mean square error.

Most of the techniques for parameter estimation that have appeared in conjunction with the use of compartment models have been relatively naive. Jacquez (1972) recommends several methods: (1) use smoothed estimates of derivatives and least squares to fit differential or difference equations directly, (2) integrate differential equations using Simpson's rule and fit to data, (3) solve the differential equations and fit using methods given in Berman (1962a, 1962b) for fitting sums of exponentials, (4) numerically "deconvolve" the integral equation

$$g(t) = \int_0^t f(\tau)\, h(t-\tau)d\tau$$

to obtain h(t), the system impulse response function, where f(t) is an input and g(t) is the system output.

Rescigno and Segre (1962) suggest using the method given by Levenstein (1960) or Prony's method (Weiss and McDonough, 1963, or

Dudley, 1977). These methods involve using Z or Laplace transforms to obtain algebraic equations in the unknown parameters and then using n observations to estimate n parameters.

The standard parametric time series approaches to estimating the parameters of an ARMA process assume that there is no forcing function or a constant forcing function, and the approaches do not appear to be easily adaptable to the present case. The most straight-forward approach (Jenkins & Watts, 1968, p. 189-190) replaces unobserved values by their unconditional expectations of zero. Then, an initial guess is made of the parameters and estimates are derived by minimizing the error sum of squares iteratively. This approach results in conditional maximum likelihood estimates. With a large sample the effect of conditioning should be slight.

Durbin (1960a,b) considers several similar models. Asymptotically efficient estimators are developed for autoregressive models, forced autoregressive models with uncorrelated noise, forced linear models with autoregressive noise, moving average models, and autoregressive models with moving-average noise. The model most similar to the one treated here is the autoregressive model with moving average errors. The efficient estimation algorithm that Durbin develops is an iterative method in which the autoregressive and moving average parameters are estimated alternately. The moving average estimator exploits the autoregressive representation of a moving average process. The estimates used for autoregressive parameters are ordinary least squares estimates.

An implicit assumption of Durbin's method is that sufficient data are available so that the autoregressive approximation can be made as accurate as desired. Thus, the method may not be suited to small sample situations. Moreover, some simulation results in Durbin (1960b) indicated that the moving average parameter estimates may be seriously biased.

The approach of Box and Jenkins (1970) exploits the fact that an unforced ARMA process can be described by either a forward or backward equation with the same parameters, both leading to the same covariance structure. The backward equation is used to obtain an approximate conditional expectation of Z(0) by replacing Z(T+1) with its unconditional expectation 0 and propagating it back to time 0 by using the observations and assumed values of parameters. The technique is called "backforecasting." Again, parameter estimates are obtained iteratively by minimizing the error sum of squares.

If a forcing function is present, however, these approaches are not easy to apply. The unconditional estimate of the state at time 0 or T+1 is no longer zero. It depends on past values of the input and the unknown parameters. The duality between the forward and backward descriptions of the system required by the backforecasting technique no longer exists. It is not evident how one would begin the recursion to obtain parameter estimates.

An approach that has been used (McMichael and Hunter, 1972) is to estimate the Φ and B of (1.1) as if $\Theta = I$ and w(k) were white noise.

The residuals are then modeled by an appropriate order ARMA model. This approach has the virtue of producing a small residual sum of squares with a few parameters. However, the estimates for the deterministic portion are still unweighted least squares estimates, not maximum likelihood estimates and not Gauss-Markov estimates. The approach would probably be improved if it were expanded to iterative form. If the model were given by

$$Z = X\beta + \varepsilon$$

where X consists of u and lagged values of Z, then the first estimate of β is

$$\hat{\beta}_1 = (X'X)^{-1}X'Z \quad .$$

The residuals $\tilde{Z}_1 = Z - \hat{Z}_1$ are then modeled by an ARMA process which provides an estimated covariance matrix Γ_1. This in turn leads to a new estimate of β given by

$$\hat{\beta}_2 = (X'\Gamma_1^{-1}X)^{-1}X'\Gamma_1^{-1}Z \quad ,$$

etc., until, hopefully, the solution converges.

A shortcoming of the above algorithm is the required inversion of the Γ matrix. It is possible to write down an expression for the elements γ_{ij} of the matrix; however, no exact expression has been obtained for the inverse of the covariance matrix of a general ARMA process. Siddiqui (1958) has obtained an explicit form for the inverse of a pure AR process, Shaman (1972) has given techniques for construct-

ing the inverse for a pure MA process, and Shaman (1975) has provided an approximate inverse for a general ARMA process.

The above estimation method indicate two reasons why maximum likelihood estimates of the parameters of a forced dynamic linear model are not in common use: starting values are not available and the covariance matrix is difficult to invert. The current research has been directed towards developing an estimation procedure that takes explicit account of the starting values and avoids the need to compute the inverse of a large matrix.

To place the proposed answer in the proper context, a digression is in order. The situations where we envision the estimation scheme being applied are modeling efforts aimed at understanding ecological systems. Many such systems tend to have long memories. The interval between successive observations is often months or years rather than seconds. Moreover, the real system is almost surely non-linear so that the constant coefficient linear model should, at the very least, be a time varying linear model. Certainly a time invariant linear model is not adequate over the range of behavior that the system can exhibit. However, provided that the system is not too far from equilibrium, that the time variation in the coefficients is not too great, and provided that we do not attempt to use the model for too long a time, the time invariant linear model can still provide valuable insight into the structure of the real system.

Still the assumption of a stationary system over a long period of time (long enough to acquire sufficient observations to make classical

time series approaches feasible) is difficult to defend. Within the narrow context required to define the systems in order to study them, the systems are almost surely not sufficiently stationary over an appreciable length of time.

Moreover, the parameter estimates obtained from the observations of a particular system at a particular time are, strictly speaking, not portable to another system at another time unless the two systems are regarded as different realizations of the same stochastic process. In fact, this is seldom if ever true for the kinds of systems that we intend to model. In the perspective of current ecosystem theory, the observations are of essentially unique realizations.

But even if one treats the different realizations as deriving from the same stochastic process, it is permissible to treat the initial conditions for each realization as parameters of that realization. The other parameters would then be constant for the process, and the initial condition parameters would be variables among realizations of the process. Whether or not the observations are considered a unique realization or a member of a family generated by a process, this orientation is useful in addressing the parameter estimation question.

Therefore the approach taken here is to parameterize the initial conditions, and to estimate them along with the other model parameters. The model and the estimation procedures will be developed in the next chapter. Two sets of estimators are derived. The maximum likelihood estimator (MLE) was first derived, along with an algorithm for obtaining the estimate that was computationally feasible. However, examin-

ation of the likelihood function indicated problems. Hence another estimator, the mean upper likelihood estimator (MULE), similar to the mean likelihood estimator (MELE) suggested by Barnard (1959), is also developed.

Barnard's approach to inference is quite interesting. It is based on the 'likelihood principle' advocated by Barnard (1949, 1962) and Birnbaum (1962). A concise discussion of the likelihood principle and the inferential techniques based on it can be found in Jenkins and Watts (1968). One of the precepts of 'likelihood inference' is that the inference drawn from a sample should be a summarization of the entire likelihood function. The maximum likelihood estimator is criticized in this regard because it can be misleading, particularly for bounded parameter spaces.

In the estimation problem being considered in the thesis, the parameter space for the Φ and Θ parameters is bounded. Moreover, the MLE's of the Θ parameters showed a distressing tendency to be near or on the boundary of the parameter space. In these circumstances the mean likelihood estimator should be superior to the MLE (Jenkins and Watts, 1968).

The MLE's of the Φ and B parameters appeared to be quite good, provided the Θ parameters were estimated accurately. The MLE's of the Φ and B parameters are essentially weighted least squares estimators. These estimators should have all the optimal properties of both MLE and Gauss-Markov estimators, provided the weights (which are functions of the Θ parameters) are reasonably accurate.

II. FORCED ARMA MODEL

A. Derivation of State Equations

If (1.1) were treated in complete generality, then $p(p + q + r)$ parameters would be required. However, it is often possible to eliminate some of these because of the model structure. For instance, the cannonical form of the Φ - matrix introduced below for a single output system requires only p (as opposed to p^2) parameters. Secondly it may be known that some elements of the input transition matrices are zero. Although such restrictions on the input transition matrices are necessary during application of the algorithm, they needlessly complicate the notation during the derivation and will be ignored. The dimensions of the B and Θ matrices will be made conformal with the Φ matrix by filling with zero's if necessary. To minimize notational labor, the estimation algorithm will be developed for a single output, single input p^{th} order difference equation. The extension to a multiple input system is easy, and a generalization to a multiple output system will be developed in Chapter 6.

The forward shift operator F is defined by $F\, z(k) = z(k+1)$. The definition can be extended to operators of the form

$$\Phi_p(F) = \phi_1 F^{p-1} + \phi_2 F^{p-2} + \ldots + \phi_p$$

where

$$\Phi_p(F)z(k) = \phi_1 z(k+p-1) + \phi_2 z(k+p-2) + \ldots + \phi_p z(k) \ .$$

The difference equation form of the model is then

$$z(k+p) = \Phi_p(F)z(k) + B_p(F)u(k) + \Theta_p(F)w(k) . \qquad (2.1)$$

The inaccessible input w(k) will be modeled as an AR process of the form

$$\Lambda_p(F)w(k) = \alpha(k) , \qquad (2.2)$$

where $\alpha(k)$ is a zero mean Gaussian white noise process. By multiplying both sides of (2.1) by $\Lambda(F)$ and substituting (2.2), one gets

$$\Lambda_p(F)z(k+p) = \Lambda_p(F)\,\Phi_p(F)z(k) + \Lambda_p(F)B_p(F)u(k) + \Theta_p(F)\alpha(k) .$$

Since the product of two difference operators, e.g., $\Lambda_p(F)\,\Phi_p(F)$, is just a higher order difference operator, e.g., $\Lambda^*_{2p}(F)$, the last equation can be written as

$$z(k+2p) = \Phi^*_{2p}(F)z(k) + B^*_{2p}(F)u(k) + \Theta_p(F)\alpha(k) \qquad (2.3)$$

which has the same form as (2.1) except the noise process is now uncorrelated.

By the foregoing argument, it is seen that the introduction of the AR structure of the noise process w(k) increases the order of the system model, but does not extend its generality. Hence, the model will be

$$F^p z(k) = \Phi_p(F)z(k) + B_p(F)u(k) + \Theta_p(F)\alpha(k) . \qquad (2.4)$$

Without loss of generality one can take $\theta_1 = 1$ and $\alpha(k) \simeq N(0,\sigma^2)$. The coefficients of $\Theta_p(F)$ will be renumbered so that θ_1 is the first non-unitary coefficient.

When the parameters are referred to as groups, those in Φ_p, Θ_p, and B_p will be termed autoregressive (AR), moving average (MA) and regression parameters, respectively.

A set of cannonical state equations representing (2.4) is (Chan, et al., 1972)

$$\begin{bmatrix} x_1(k+1) \\ \vdots \\ x_p(k+1) \end{bmatrix} = \begin{bmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & & \\ \vdots & \vdots & & & 1 \\ \phi_p & 0 & & & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ \vdots \\ x_p(k) \end{bmatrix} + \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} u(k)$$

$$+ \begin{bmatrix} \theta_1 \\ \vdots \\ \theta p \end{bmatrix} \alpha(k) \qquad (2.5)$$

or, in vector-matrix notation

$$X(k+1) = AX(k) + BU(k) + C\alpha(k) \qquad (2.6)$$

and the observation equation

$$z(k) = x_1(k) . \qquad (2.7)$$

The cannonical form given by (2.5) holds only if the roots of the

fundamental equation

$$1 + \phi_1 F + \phi_2 F^2 + \ldots + \phi_p F^p$$

are distinct. It will henceforth be assumed that the roots are indeed distinct. The following theorem implies that this assumption is not a serious limitation.

Theorem 2.1: Let $z_k = a_0 r_1^k + a_1 (r_1+\varepsilon)^k$ and $w_k = b_0 r_1^k + b_1 k r_1^k$ for $k = 0,2,3,\ldots$, and suppose $z_0 = w_0$, $z_1 = w_1$. Then $\lim_{\varepsilon \to 0} z_k = w_k$, $k > 1$.

Proof: From the initial conditions, it follows that $b_0 = z_0$, $b_1 = (z_1 - z_0 r_1) r_1$ and $a_0 = z_0 - a_1$, $a_1 = (z_1 - z_0 r_1)/\varepsilon$. Thus, $a_1 = (r_1 b_1)/\varepsilon$ and $a_0 = b_0 - (r_1 b_1/\varepsilon)$. Hence,

$$\begin{aligned}
\lim_{\varepsilon \to 0} (z_k - w_k) &= \lim_{\varepsilon \to 0} (a_0 r_1^k + a_1 (r_1+\varepsilon)^k - b_0 r_1^k - b_1 k r_1^k) \\
&= \lim_{\varepsilon \to 0} \left(- \frac{r_1 b_1 r_1^k}{\varepsilon} + \frac{r_1 b_1 (r_1+\varepsilon)^k}{\varepsilon} - b_1 k r_1^k \right) \\
&= r_1 b_1 \lim_{\varepsilon \to 0} \left(\frac{(r_1+\varepsilon)^k - r_1^k}{\varepsilon} \right) - b_1 k r_1^k \\
&= r_1 b_1 \frac{dr_1^k}{dr_1} - b_1 k r_1^k \\
&= 0
\end{aligned}$$

Theorem 2.1 implies that the output of a system with repeated roots in the fundamental equation can be matched within arbitrary

closeness by a system whose fundamental equation has distinct roots.

In the problem to be solved, there are available observations $z(1), z(2), \ldots, z(T)$ and $u(0), u(1), u(2), \ldots, u(T-1)$. The parameters $\phi_1, \ldots, \phi_p$; $\beta_1, \ldots, \beta_p$; $\theta_1, \ldots, \theta_p$ and σ^2 are unknown and to be estimated from the observations. Unfortunately, these parameters and the observations that can be collected do not completely describe the behavior of the system during the period of observation. A description of the state of the system at the time observation began is missing. This is clear from the convolution form of the solution to the state equations:

$$X(k+1) = A^{k+1}X(0) + \sum_{m=0}^{k} A^{k-m}(BU(m) + C\alpha(m)). \qquad (2.8)$$

In the next section a parametric representation of the initial conditions will be provided.

B. Derivation of Likelihood Equation

A state variable representation of a system is not unique and the 'X' state variables introduced in Section IIA are not best suited for use in estimating the state at time 0. An attempt to use the convolution in (2.8) for estimation results in a complicated non-linear estimation problem. However, by examining the difference equation form of the system model, it can be seen that $z(k)$, $k > p$ depends on output or input prior to time 1 only through the values of $z(i)$, $i = 1,\ldots,p$. This information contained in the prior values of the output and accessible input can be embodied in the state at time 0 estimates. Thus, one is led to seek a set of state variables that appear explicitly only

in the first p expressions for $z(\cdot)$. As the process evolves from time 1, the explicit presence of these state variables disappears. A particular set that accomplishes this is the 'W' state variables defined implicitly by the following representation:

$$
\begin{aligned}
z(1) &= \phi_1 w_1(0) + \phi_2 w_2(0) + \ldots + \phi_p w_p(0) \\
&\quad + \beta_1 u(0) + \alpha(0) \\
z(2) &= \phi_1 z(1) + \phi_2 w_1(0) + \ldots + \phi_p w_{p-1}(0) \\
&\quad + \beta_1 u(1) + \beta_2 u(0) + \alpha(1) + \theta_1 \alpha(0) \\
&\vdots \\
z(p) &= \phi_1 z(p-1) + \phi_2 z(p-2) + \ldots + \phi_p w_1(0) \\
&\quad + \beta_1 u(p-1) + \beta_2 u(p-2) + \ldots + \beta_p u(0) \\
&\quad + \alpha(p-1) + \theta_1 \alpha(p-2) + \ldots + \theta_{p-1} \alpha(0) \\
z(p+1) &= \phi_1 z(p) + \phi_2 z(p-1) + \ldots + \phi_p z(1) \\
&\quad + \beta_1 u(p) + \beta_2 u(p-1) + \ldots + \beta_p u(1) \\
&\quad + \alpha(p) + \theta_1 \alpha(p-1) + \ldots + \theta_p \alpha(0) \\
&\vdots \\
z(k) &= \phi_1 z(k-1) + \phi_2 z(k-2) + \ldots + \phi_p z(k-p) \\
&\quad + \beta_1 u(k-1) + \beta_2 u(k-2) + \ldots + \beta_p u(k-p) \\
&\quad + \alpha(k-1) + \theta_1 \alpha(k-2) + \ldots + \theta_p \alpha(k-p-1), \quad k > p.
\end{aligned}
$$

It can be shown that the vector W(k) is a linear transformation of the vector X(k). The initial state W(0) will be termed a 'regression parameter' along with B_p. Note that by regarding the initial state as a parameter, the z(k) process is regarded as a non-stationary stochastic process. This turns out to have some significant advantages.

The transformation from X and W is a transformation of the state variables, i.e., a transformation of the system model. It is also convenient to make a transformation on the observations, i.e., a transformation of the statistical model. The new variable Y is defined as follows:

$$
\begin{aligned}
y(1) &= z(1) \\
y(2) &= z(2) - \phi_1 z(1) \\
&\vdots \\
y(p) &= z(p) - \phi_1 z(p-1) - \ldots - \phi_{p-1} z(1) \\
&\vdots \\
y(k) &= z(k) - \phi_1 z(k-1) - \ldots - \phi_p z(k-p), \quad k \geq p.
\end{aligned}
$$

This transformation can be written as

$$Y = MZ \quad . \tag{2.9}$$

M is a TxT matrix with 1's on the diagonal, $-\phi_1$'s on the first subdiagonal, $-\phi_2$'s on the second subdiagonal, etc.

The expected value of Y, say μ_y, is given by

$$\mu_{y(1)} = \phi_1 w_1 + \ldots + \phi_p w_p + \beta_1 u(0)$$

$$\mu_{y(2)} = \phi_2 w_1 + \ldots + \phi_p w_{p-1} + \beta_1 u(1) + \beta_2 u(0)$$

$$\vdots$$

$$\mu_{y(p)} = \phi_p w_1 + \beta_1 u(p-1) + \ldots + \beta_{p-1} u(1) + \beta_p u(0)$$

$$\vdots$$

$$\mu_{y(k)} = \beta_1 u(k-1) + \ldots + \beta_p u(k-p), \quad k > p$$

The covariance matrix of Y, say $\sigma^2\Sigma_T$, can be written as

$$\sigma^2\Sigma_T = \sigma^2 R'R \quad , \tag{2.10}$$

where

$$R = \sum_{i=0}^{p} \theta_i J_i \quad ,$$

$$\theta_0 = 1$$

and J_0 is the TxT identity matrix, J_1 is a TxT matrix with 1's on the first diagonal above the main diagonal, and $J_{i+1} = J_1 \cdot J_i$, $i \geq 1$. Since Σ_T is the product of the two triangular matrices with ones on the diagonal, $|\Sigma_T| = 1$. Moreover, M is also a triangular matrix with ones on the diagonal, so the covariance matrix of Z has unit determinant.

An expression of the inverse of Σ_T will be needed. It will suffice to have an expression for R^{-1}. It can be shown that

$$R^{-1} = \sum_{t=0}^{T-1} \pi_t J_t \quad ,$$

where the π_i's are defined recursively by

$$\pi_0 = 1$$

$$\pi_i = -\sum_{t=1}^{k} \theta_t \pi_{k-t}, \quad k = \min(i,p), \; i \geq 1. \tag{2.11}$$

Then letting $\Sigma_T^{-1} = (s_{ij})$ it follows that

$$s_{ij} = \sum_{r=0}^{T-j} \pi_r \pi_{r+j-i}$$

for $1 \leq i \leq j \leq T$, and $s_{ji} = s_{ij}$. It was earlier assumed that the $\alpha(k)$ process was Gaussian white noise. It follows that both Y and Z have Gaussian distributions, and that the likelihood function is

$$L(\phi, \theta, w, \beta, \sigma^2/Z) = (2\pi\sigma^2)^{-T/2} \exp\left(\frac{-(Z-\mu_Z)'M'\Sigma_T^{-1}M(Z-\mu_Z)}{2\sigma^2}\right). \tag{2.12}$$

III. DERIVATION OF PARAMETER ESTIMATORS

A. Discussion

The problem was first approached as one of obtaining MLE's for all the parameters. As was noted in the introduction, the MLE's of the MA parameters (the moving average parameters defined in Section II A) were not well behaved for small samples. Hence a different estimator was sought.

The MA parameters are not of particular interest. The AR (auto-regressive) and regression parameters have an intrinsic interest; in many cases they have an immediate physical interpretation. The MA portion of the model must be included because "good" estimators of the parameters of primary interest require good estimators of the covariance structure of the observations. Further, reported experience indicates that models with feedback in the variables of interest tend to have feedback in the noise term and ARMA processes can accomodate a variety of stationary stochastic processes.

An additional property that would be desirable in the MA estimators is that they tend to stay away from boundaries. This reflects our bias towards the view that values near the invertibility boundary are not physically likely. Perhaps a Bayesian approach would prove to be fruitful, but it was not investigated. Note, however, that the approach that was adopted is equivalent to taking a uniform prior over the parameter space.

The mean likelihood estimator proposed by Barnard (1959) (and recommended by Jenkins and Watts (1968)) is defined as

$$\hat{\lambda}_i = \int_{\Lambda} \lambda_i \, L(\lambda|X)d\lambda \Big/ \int_{\Lambda} L(\lambda|X)d\lambda$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_K)$ is a parameter vector, $L(\lambda|X)$ is the likelihood function of λ given X, and Λ is the joint parameter space.

Application of this estimator in the present case quickly runs into difficulties. The integrals with respect to the AR and MA parameters cannot be evaluated analytically. It would be possible to perform the integration numerically, but relying on numerical integration for the general case would be cumbersome. But there is another approach that avoids numerical integration with respect to the AR parameters.

As will be shown below, the MLE of the AR, regression, and variance parameters can be readily computed as functions of the MA parameters. The estimators can be expressed in the explicit functional form

$$\hat{\phi} = h_1(\theta)$$
$$\hat{\beta} = h_2(\theta)$$

and

$$\hat{\sigma}^2 = h_3(\theta) \quad .$$

The joint likelihood function can be reduced to a function of only the MA parameters by replacing the other parameters by their MLE's, expressed in terms of the MA parameters. The resulting function will be termed the upper likelihood function (ULF):

$$\text{ULF}(\theta) = \max_{\phi,\beta,\sigma^2} L(\phi,\theta,\beta,\sigma^2|X) \ .$$

$$= L(h_1(\theta),\theta,h_2(\theta),\ h_3(\theta)|X)$$

The mean upper likelihood estimator (MULE) is defined analogously to the MELE:

$$\hat{\theta}_i = \int_{\Theta} \theta_i \ \text{ULF}(\theta)d\theta \Big/ \int_{\Theta} \text{ULF}(\theta)d\theta \ ,$$

where Θ is parameter space for θ. Once the estimate $\hat{\theta}$ of θ is obtained estimates of the other parameters are obtained by substituting $\hat{\theta}$ into the functions h_1, h_2 and h_3. If $\hat{\theta}$ were an MLE, then $\hat{\phi}$, $\hat{\beta}$, and $\hat{\sigma}^2$ would also be MLE's. Generally the MULE should have the same asymptotic behavior as the MLE, since large sample likelihood functions tend to be symmetric, unimodal functions.

B. Estimators for AR and Regression Parameters

Except for additive constants, the log likelihood is given by

$$\ell(\phi,\theta,\beta,w,\sigma|Z,U) = -T\,\ell n(\sigma) - (Z-\mu_z)'M'\Sigma_T^{-1}M(Z-\mu_z)/2\sigma^2 .$$

Maximizing with respect to ϕ, θ, β, and w is equivalent to minimizing

$$(Z-\mu_z)'M'\Sigma_T^{-1}M(Z-\mu_z) = (M^{-1}(Y-\mu_Y)'M'\Sigma^{-1}M(M^{-1}(Y-\mu_Y)) \qquad (3.1)$$

$$= (Y-\mu_Y)'\Sigma_T^{-1}(Y-\mu_Y) .$$

μ_Y can be expressed algebraically as

$$\mu_Y = G\rho , \qquad (3.2)$$

where

$$\rho = \begin{bmatrix} w_1 \\ \vdots \\ w_p \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad G = \begin{bmatrix} \phi_1 & \cdots & \phi_p & u(1) & 0 & \cdots & 0 \\ \phi_2 & \cdots & 0 & u(2) & u(1) & \cdots & 0 \\ \vdots & \iddots & \vdots & \vdots & \vdots & \iddots & \vdots \\ \phi_p & 0 \cdots & 0 & u(p) & u(p-1) & \cdots & u(1) \\ 0 & \cdots & 0 & u(p+1) & u(p) & \cdots & u(2) \\ \vdots & & \vdots & \vdots & & & \vdots \\ 0 & \cdots & 0 & u(T) & \cdots & & u(T-p+1) \end{bmatrix}$$

If ϕ and θ are temporarily regarded as being fixed, the maximizing value of ρ is given by

$$\rho(\phi,\theta) = (G'\Sigma_T^{-1}G)^{-1} G'\Sigma_T^{-1} Y \quad . \tag{3.3}$$

If the expression for $\hat{\rho}(\phi,\theta)$ is substituted back into (3.1) and the expression is simplified, there results

$$\begin{aligned} &\min_{w,\beta,\phi,\theta} (Y-\mu_Y)'\Sigma_T^{-1}(Y-\mu_Y) \\ &= \min_{\phi,\theta} (Y-G\hat{\rho})'\Sigma_T^{-1}(Y-G\hat{\rho}) \\ &= \min_{\phi,\theta} Y'\Sigma_T^{-1}(I-G(G'\Sigma_T^{-1}G)^{-1}G'\Sigma_T^{-1})Y \\ &= \min_{\phi,\theta} Y'\psi Y \quad , \end{aligned}$$

where

$$\psi = \Sigma_T^{-1} - \Sigma_T^{-1} G(G'\Sigma_T^{-1}G)^{-1} G' \Sigma_T^{-1} \quad .$$

The following theorem explores the structure of the matrix ψ. In particular it implies that ψ is independent of the AR parameters.

Theorem 3.1: Let Σ be a TxT covariance matrix and let

$$G = \left[\begin{array}{c|c} P & U_1 \\ \hline 0 & U_2 \end{array}\right]$$

be an Tx2p matrix of full rank. Let P be a pxp non-singular matrix. Let $S = \Sigma^{-1}$ and let S and Σ be partitioned in the same manner as is G, i.e.,

$$S = \begin{array}{c} \\ p \\ T-p \end{array} \overset{\begin{array}{cc} p & T-p \end{array}}{\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}} \qquad \Sigma = \begin{array}{c} \\ p \\ T-p \end{array} \overset{\begin{array}{cc} p & T-p \end{array}}{\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}} .$$

Then

$$S[I-G(G'SG)^{-1}G'S] = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{22}^{-1}[I-U_2(U_2'\Sigma_{22}^{-1}U_2)U_2' \Sigma_{22}^{-1}] \end{bmatrix} .$$

The proof of the theorem is long, and is broken down into several lemmas. The lemmas use the following notation.

$$(G'SG)^{-1} = \begin{bmatrix} P'S_{11}P & P'S_{11}U_1 + P'S_{12}U_2 \\ U_1'S_{11}P & U_1'S_{11}U_1 + U_2'S_{21}U_1 \\ +U_2'S_{21}P & +U_1'S_{12}U_2 + U_2'S_{22}U_2 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

Lemma 3.1. Using the same notation as in Theorem 3.1

$$A_{11}P'S_{11} + A_{12}U_1'S_{11} + A_{12}U_2'S_{21} = P^{-1} \quad .$$

Since $A_{12} = -A_{11}B_{12}B_{22}^{-1}$, it follows that

$$A_{11}P'S_{11} + A_{12}U_1'S_{11} + A_{12}U_2'S_{21}$$

$$= A_{11}[P'S_{11} - B_{12}B_{22}^{-1}(U_1'S_{11} + U_2'S_{21})] \tag{3.4}$$

but

$$P'S_{11} = B_{11}P^{-1} \tag{3.5}$$

and

$$U_1'S_{11} + U_2'S_{21} = B_{21}P^{-1} \quad . \tag{3.6}$$

Substituting 3.5 and 3.6 into (3.4), there results

$$A_{11}[B_{11}P^{-1} - B_{12}B_{22}^{-1}B_{21}P^{-1}]$$

$$= A_{11}[B_{11} - B_{12}B_{22}^{-1}B_{21}]P^{-1} = P^{-1} \quad .$$

Lemma 3.2. Using same notation as in Theorem 3.1

$$A_{21}P'S_{11} + A_{22}U_1'S_{11} + A_{22}U_2'S_{21} = 0.$$

Since

$$A_{21} = -A_{22}B_{21}B_{11}^{-1}$$

and

$$B_{11}^{-1} = P^{-1}S_{11}^{-1}(P')^{-1}$$

and

$$B_{21} = U_1'S_{11}P + U_2'S_{21}P \quad .$$

Then

$$A_{21} = -A_{22}[U_1'S_{11}P + U_2'S_{21}P]P^{-1}S_{11}^{-1}(P')^{-1}$$
$$= -A_{22}U_1'(P')^{-1} - A_{22}U_2'S_{21}S_{11}^{-1}(P')^{-1}$$

and

$$A_{21}P'S_{11} + A_{22}U_1'S_{11} + A_{22}U_2'S_{21}$$
$$= -A_{22}U_1'S_{11} - A_{22}U_2'S_{21} + A_{22}U_1'S_{11} + A_{22}U_2'S_{21} = 0 .$$

Lemma 3.3: Using the notation of Theorem 3.1

$$A_{21}P'S_{12} + A_{22}U_1'S_{12} + A_{22}U_2'S_{22} = (U_2'\Sigma_{22}^{-1}U_2)^{-1}U_2'\Sigma_{22}^{-1} .$$

Using

$$A_{21} = -A_{22}B_{21}B_{11}^{-1}$$
$$= -A_{22}B_{21}P^{-1}S_{11}^{-1}(P')^{-1} ,$$

it follows that

$$A_{21}P'S_{12} + A_{22}U_1'S_{12} + A_{22}U_2'S_{22}$$
$$= A_{22}[-B_{21}P^{-1}S_{11}^{-1}S_{12} + U_1'S_{12} + U_2'S_{22}]$$
$$= A_{22}[-(U_1'S_{11}P + U_2'S_{21}P)P^{-1}S_{11}^{-1} + U_1'S_{12} + U_2'S_{22}] \qquad (3.7)$$
$$= A_{22}[-U_1'S_{12} - U_2'S_{21}S_{11}^{-1}S_{12} + U_1'S_{12} + U_2S_{22}]$$
$$= A_{22}[U_2'(S_{22} - S_{21}S_{11}^{-1}S_{12})]$$
$$= A_{22}U_2'\Sigma_{22}^{-1} .$$

But

$$
\begin{aligned}
A_{22} &= [B_{22} - B_{21}B_{11}^{-1}B_{12}]^{-1} \\
&= [U_1'S_{11}U_1 + U_2'S_{21}U_1 + U_1'S_{12}U_2 + U_2'S_{22}U_2 \\
&\quad - U_1'S_{11}PP^{-1}S_{11}^{-1}P^{-1}PS_{11}U_1 \\
&\quad - U_1'S_{11}PP^{-1}S_{11}^{-1}P^{-1}PS_{12}U_2 \\
&\quad - U_2'S_{21}PP^{-1}S_{11}^{-1}P^{-1}PS_{11}U_1 \\
&\quad - U_2'S_{21}PP^{-1}S_{11}^{-1}PP^{-1}S_{21}U_2]^{-1} \\
&= [U_2'(S_{22} - S_{21}S_{11}^{-1}S_{12})U_2]^{-1} \\
&= [U_2'\Sigma_{22}^{-1}U_2]^{-1} .
\end{aligned}
$$

By substituting this result in (3.7), the lemma is proved.

Lemma 3.4: Using the notation of Theorem 3.1,

$$
\begin{aligned}
&P(A_{11}P'S_{12} + A_{12}U_1'S_{12} + A_{12}U_2'S_{22}) \\
&= S_{11}^{-1}S_{12}[I - U_2(U_2'\Sigma_{22}^{-1}U_2)^{-1}U_2'\Sigma_{22}^{-1}] - U_1(U_2'\Sigma_{22}^{-1}U_2)^{-1}U_2'\Sigma_{22}^{-1} .
\end{aligned}
$$

Proof: Let $C_2 = A_{21}P'S_{12} + A_{22}U_1'S_{12} + A_{22}U_2'S_{22}$ and let

$C_1 = A_{11}P'S_{12} + A_{12}U_1'S_{12} + A_{12}U_2'S_{22}$. Then

$$A_{12}A_{22}^{-1}C_2 = A_{12}A_{22}^{-1}A_{21}P'S_{12} + A_{12}U_1'S_{12} + A_{12}U_2'S_{22} + A_{11}PS_{12} - A_{11}PS_{12}$$

$$= C_1 - [A_{11} - A_{12}A_{22}^{-1}A_{21}]PS_{12}$$

$$= C_1 - B_{11}^{-1}PA_{12}$$

$$= C_1 - P^{-1}S_{11}^{-1}S_{12} \ .$$

By Lemma 3.3, $C_2 = (U_2' \Sigma_{22}^{-1}U_2)^{-1}U_2' \Sigma_{22}^{-1}$ and $A_{22} = (U_2' \Sigma_{22}^{-1}U_2)$, so that

$$C_1 = A_{12}U_2' \Sigma_{22}^{-1} + P^{-1}S_{11}^{-1}S_{12}$$

but $$A_{12} = -B_{11}^{-1}B_{21}'A_{22}$$

$$= -P^{-1}S_{11}^{-1}(P')^{-1}(P'S_{11}U_1 + P'S_{12}U_2)A_{22}$$

$$= -P^{-1}U_1 - P^{-1}S_{11}^{-1}S_{12}U_2 \ .$$

Hence,

$$PC_1 = S_{11}^{-1}S_{12}[I - U_2(U_2'\Sigma_{22}^{-1}U_2)^{-1}U_2'\Sigma_{22}^{-1}] - U_1(U_2'\Sigma_{22}^{-1}U_2)^{-1}\ U_2'\Sigma_{22}^{-1} \ .$$

Proof of Theorem 3.1: It follows from Lemmas 3.1 - 3.4 that

$$G(G'SG)^{-1}G'S = \begin{bmatrix} I_p & S_{11}^{-1}S_{12}[I - U_2(U_2'\Sigma_{22}^{-1}U_2)^{-1}\ U_2'\Sigma_{22}^{-1} \\ 0 & U_2(U_2'\Sigma_{22}^{-1}U_2)^{-1}U_2'\Sigma_{22}^{-1} \end{bmatrix}$$

Hence

$$S[I - G(G'SG)^{-1}G'S] = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{22}^{-1}[I - U_2(U_2'\Sigma_{22}^{-1}U_2)^{-1} U_2'\Sigma_{22}^{-1}] \end{bmatrix} .$$

Define the (T-p) x (T-p) matrix

$$\Omega = \Sigma_{22}^{-1} - \Sigma_{22}^{-1} U_2(U_2'\Sigma_{22}^{-1} U_2)^{-1} U_2'\Sigma_{22}^{-1} ,$$

where Σ_{22} and U_2 are obtained by partitioning Σ_T and U as

$$U = \begin{bmatrix} U_1 \\ \hline U_2 \end{bmatrix} \quad p \text{ rows} \qquad T = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right] \quad p \text{ rows} .$$

p cols.

The transformation that defines Y can be expressed in a manner that is more convenient at this point: let

$$Y = MZ = Z - H\Phi , \tag{3.8}$$

where

$$H = \left[\begin{array}{ccccc} 0 & 0 & \cdots & 0 & 0 \\ z(1) & 0 & \cdots & 0 & 0 \\ z(2) & z(1) & \cdots & 0 & 0 \\ \vdots & & & \vdots & \vdots \\ z(p-1) & z(p-2) & \cdots & z(1) & 0 \\ \hline z(p) & z(p-1) & \cdots & z(2) & z(1) \\ \vdots & \vdots & & \vdots & \vdots \\ z(T-1) & z(T-2) & \cdots & z(T-p+1) & z(T-p) \end{array}\right] = \left[\begin{array}{c} H_1 \\ \hline H_2 \end{array}\right] .$$

If Y and Z are partitioned as above, then it follows that

$$\min_{w,\beta,\phi,\theta} (Y-\mu_Y)'\Sigma_T^{-1}(Y-\mu_Y) = \min_{\phi,\theta} (Z_2-H_2\phi)'\Omega(Z_2-H_2\phi) \quad . \tag{3.9}$$

The minimizing value of ϕ is given by

$$\hat{\phi}(\theta) = (H_2'\Omega H_2)^{-1}H_2'\Omega Z_2 \quad . \tag{3.10}$$

C. Estimators of MA Parameters

The expression (3.10) could be substituted for ϕ in (3.9) to obtain an expression that is a function of only θ, and the MLE could be obtained by minimization. However, that direct approach results in a complicated function that makes minimization difficult.

A method that leads to an easier algorithm splits the parameters into two sets: one set comprising the θ-parameters and the other the ρ and ϕ parameters. A minimization is first performed with respect to

ρ and ϕ with θ held fixed. Then ρ and ϕ are held fixed at the minimizing values and the function is minimized with respect to θ. The process is then repeated until convergence is reached.

The MULE is obtained by substituting the MLE estimators of ϕ, ρ, and σ into the likelihood equation (2.11), and performing a series of numerical integrations. Gaussian quadrature (Acton, 1970) is used.

IV. PROPERTIES OF THE ALGORITHM

A. Discussion

The general approach of the algorithm to maximizing the likelihood function is to divide the parameters into two sets. The likelihood function is then maximized with respect to the first set while the second is held fixed, and then maximized with respect to the second while the first is held fixed. The above process is repeated until the likelihood function attains its maximum. The rationale for using this approach is that the maximum of the likelihood function with respect to ϕ and ρ, given θ, is relatively easy to obtain. The maximizing values of ϕ and ρ are obtained as the solution of a set of linear equations. Similarly, the solution of the likelihood equations for θ given ϕ and ρ is comparitively easy. The equations are not linear, but the first and second partial derivatives can be computed easily and without a great expense in computer time. The algorithm for computation of $\partial \ell / \partial \theta_k$ and $\partial^2 \ell / \partial \theta_k \partial \theta_j$ is given in the following paragraph.

Let $V_0 = (Y - E[Y])/\sigma$. Then the minus log likelihood function given ϕ, ρ, σ and Z is, except for additive terms constant in θ,

$$-\ell(\theta|\phi,\rho,\sigma,Z) = V_0' \Sigma_T^{-1} V_0/2 .$$

Let $QQ' = \Sigma_T^{-1}$ and define $V_{i+1} = Q'V_i$. The minus log likelihood function is

$$-\ell(\theta|\phi,\rho,\sigma,Z) = \frac{V_0' \Sigma_T^{-1} V_0}{2} = \frac{V_0' Q Q' V_0}{2} = V_1' V_1/2 .$$

It follows that

$$\frac{\partial \ell}{\partial \theta_k} = V_0' \frac{\partial Q}{\partial \theta_k} Q' V_0 .$$

But

$$\frac{\partial Q}{\partial \theta_k} = -R^{-1} J_k R^{-1} = -Q J_k Q$$

$$= -Q Q J_k$$

so that

$$\frac{\partial \ell}{\partial \theta_k} = - V_0' Q Q J_k Q' V_0 = - V_2' Q J_k V_1 . \qquad (4.1)$$

Similarly, it follows that

$$\frac{\partial^2 \ell}{\partial \theta_k \partial \theta_j} = V_2' J_k J_j' V_2 + 2V_3' J_{k+j} V_1 .$$

The V_i's can be computed recursively as follows:

$$V_{i+1} = Q' V_i = (R^{-1})' V_i$$

so that $R' V_{i+1} = V_i$. This implies the following relationship between the two vectors:

$$\begin{aligned} V_{i+1}(1) &= V_i(1) \\ V_{i+1}(2) &= V_i(2) - \theta_1 V_{i+1}(1) \\ V_{i+1}(3) &= V_i(3) - \theta_1 V_{i+1}(2) - \theta_2 V_{i+1}(1) \\ &\vdots \end{aligned}$$

and in general, for $k \geq p$,

$$V_{i+1}(k) = V_i(k) - \theta_1 V_{i+1}(k-1) - \ldots - \theta_p V_{i+1}(k-p) .$$

Given that derivatives are easy to obtain, there are several efficient numerical maximization algorithms. Newton-Raphson was selected simply because it was easy to program.

B. Convergence

Algorithms of this general type have been discussed by Oberhofer and Kmenta (1974). They prove the following convergence theorem:

Theorem 4.1: Let $f(\alpha)$ be a function which is to be maximized with respect to α, and $\alpha \varepsilon U$. Let α be partitioned as $\alpha = (\alpha_1 \ \alpha_2)$ with $\alpha_1 \varepsilon \ U_1 \subset R^n$ and $\alpha_2 \varepsilon \ U_2 \subset R^m$. Let $f(\alpha)$ have the following properties:

(i) There exists an s such that the set

$$S = \{\alpha | \alpha \varepsilon \ U_1 x U_2, \ f(\alpha) \geq s\}$$

is non-empty and bounded,

(ii) $f(\alpha)$ is continuous in S; and

(iii) the parameter space is closed, or U_2 is closed and $U_1 = R^n$.

Define the following iteration:

(i) Let α_1^0 be a vector of initial values of α_1 such that $\alpha_1^0 \varepsilon \ U_1$ and such that there exists an $\alpha_2 \varepsilon U_2$ such that $f(\alpha_1^0, \alpha_2) \geq s$.

(ii) Maximize $f(\alpha_1^0, \alpha_2)$ in U_2, say the maximum is reached at $\alpha_2 = \alpha_2^0 \varepsilon \ U_2$.

(iii) Suppose inductively that (α_1^j, α_2^j) have been obtained for all $0 \leq j \leq k$. Maximize $f(\alpha_1^k, \alpha_2)$ in U_2. The maximum will be reached at $\alpha_2 = \alpha_2^{k+1}$. Then maximize $f(\alpha_1, \alpha_2^{k+1})$ in U_1. The maximum will be attained at $\alpha_1 = \alpha_1^{k+1}$.

Then

(i) The sequence $\{\alpha^k\}$ has at least one accumulation point α^* in S.

(ii) If α^* and α^+ are two accumulation points of the sequence, then $f(\alpha^*) = f(\alpha^+)$.

(iii) For every accumulation point $\alpha^* = (\alpha_1^*, \alpha_2^*)$,

$$\max_{\alpha_1 \varepsilon U_1} f(\alpha_1, \alpha_2^*) = \max_{\alpha_2 \varepsilon U_2} f(\alpha_1^*, \alpha_2) = f(\alpha_1^*, \alpha_2^*) = f(\alpha^*).$$

This theorem is proved in Oberhofer and Kmenta (1974, p. 579-590).

In the case at hand, the properties (i), (ii) and (iii) hold (ℓ is bounded and continuous, θ is constrained to lie in a compact set), so that the theorem applies. Since ℓ is not in general convex, convergence to the global maximum is not guaranteed. However, in all the cases that have been examined, the algorithm encountered convergence problems only when the MA parameter was near or on the invertability boundary. Bounding the step size somewhat alleviated the problem; however, the final solution was to set the MA parameters to an arbitrary value near the boundary (e.g., $|\theta| = .99$ for case $p = 1$).

C. Asymptotic Variance of Parameter Estimates

An estimate of the covariance matrix of the parameter estimates can be obtained through maximum likelihood theory (Wilks, 1962). If β is taken to represent the vector of all the parameters, then asymptotically

$$\mathrm{Cov}(\hat{\beta}) = E\left[\frac{\partial \ell}{\partial \beta}\frac{\partial \ell'}{\partial \beta}\right]^{-1} .$$

A submatrix of the inverse covariance matrix will be termed an 'information matrix'. Selected terms of the inverse covariance matrix are evaluated below. Sufficient terms are evaluated to allow the inverse covariance matrix to be computed. The following lemma will be needed.

Lemma 4.1: Let D be a TxT matrix such that

$$D = \sum_{i=0}^{T-1} \tau_i J_i ,$$

where $J_0 = I$, J_1 is a matrix with ones on the first superdiagonal and zeros elsewhere, and $J_i = J_{i-1} J_1$, for $i = 2, \ldots, T-1$. Let $V = (v_1, v_2, \ldots, v_T)$ be a random T-vector distributed as MVN(0,I). Then for $1 \leq r \leq T-1$,

$$V'D J_r V = \sum_{j=1}^{T-r} \sum_{i=r+j}^{T} v_j \tau_{i-r-j} v_i ,$$

and for $k \geq 0$,

$$E[V'D J_r VV'DJ_{r+k}V] = \sum_{j=0}^{T-r-k-1} (T-r-k-j) \tau_j \tau_{j+k} .$$

Proof: It follows from the definitions of B and J_r that

$$V'D J_r V = V'\left(\sum_{i=0}^{T-1} \tau_i J_i\right) J_r V$$

$$= \sum_{i=0}^{T-r-1} \tau_i V'J_{i+r}V$$

$$= \sum_{i=0}^{T-r-1} \tau_i \sum_{j=1}^{T-i-r} v_j v_{i+r+j}$$

and the result is obtained by re-ordering the summation. Substituting this result into the expectation, there results

$$E[V'DJ_rVV'DJ_{r+k}V]$$

$$= \sum_{j=1}^{T-r} \sum_{i=r+j}^{T} \sum_{\ell=1}^{T-r-k} \sum_{m=r+k+\ell}^{T} \tau_{i-r-j} \tau_{m-r-k-\ell} E[v_i v_j v_\ell v_m] \quad .$$

But the expectation vanishes unless (a) $i = j$, $\ell = m$, or (b) $i = \ell$, $j = m$, or (c) $i = m$, $j = \ell$. However, cases (a) and (b) are impossible, since $r > 0$. Thus, the sum reduces to

$$\sum_{j=1}^{T-r-k} \sum_{i=r+j+k}^{T} \tau_{i-r-j} \tau_{i-r-j-k} \quad .$$

The result is obtained by a change of variable in the summation.

C.1 Information for $\hat{\rho}$.

The log likelihood can be taken as

$$\ell = -(Y - G\rho)' \Sigma_T^{-1} (Y - G\rho)/2\sigma^2 \quad .$$

Thus,

$$\frac{\partial \ell}{\partial \rho} = G' \Sigma_T^{-1} (Y - G\rho)/\sigma^2 \quad .$$

Recall that V_1 was defined as $Q'(Y - G\rho)/\sigma$, and V_1 is distributed as MVN(0,I). Thus,

$$\frac{\partial \ell}{\partial \rho} = \frac{G'Q}{\sigma} V_1 \tag{4.2}$$

and the information matrix of $\hat{\rho}$ is

$$E\left[\frac{\partial \ell}{\partial \rho} \frac{\partial \ell}{\partial \rho}\right] = E\left(\frac{G'QV_1V_1'Q'G}{\sigma^2}\right) = (G'\Sigma_T^{-1}G)/\sigma^2 \tag{4.3}$$

C.2 Information of $\hat{\theta}$ and $(\hat{\theta}, \hat{\rho})$.

From (4.1)

$$\frac{\partial \ell}{\partial \theta_k} = -V_1'Q\, J_k\, V_1 \ .$$

Applying Lemma 4.1, it follows that

$$\frac{\partial \ell}{\partial \theta_k} = -\sum_{j=1}^{T-K} \sum_{i=k+j}^{T} v_i\, \tau_{i-k-j}\, v_j \ , \tag{4.4}$$

where π_t, $0 \le t \le T-1$, is defined by (2.10). The lemma also implies that

$$E\left[\frac{\partial \ell}{\partial \theta_k} \frac{\partial \ell}{\partial \theta_{k+m}}\right] = \sum_{j=0}^{T-k-m-1} (T-k-m-j)\, \pi_j\, \pi_{j+m} \ .$$

From (4.2) and (4.4), the information for $\hat{\rho}$ and $\hat{\theta}$ can be obtained:

$$E[(\partial \ell/\partial \rho)(\partial \ell/\partial \theta_k)] = E\left[\frac{G'QV_1}{\sigma}\left(-\sum_{j=1}^{T-k} \sum_{i=k+j}^{T} v_i\, v_j\, \pi_{i-k-j}\right)\right] \ .$$

But since i > j, the expectation vanishes.

C.3 Information for $\hat{\phi}$.

From (3.2),

$$E[Y] = G\rho = G\begin{pmatrix} W \\ \beta \end{pmatrix} = [\Phi\ U]\begin{pmatrix} W \\ \beta \end{pmatrix} .$$

The product ΦW can also be written $F\phi$, where

$$F = \begin{bmatrix} W_1 & W_2 & \cdots & & W_P \\ 0 & W_1 & W_2 & \cdots & W_{P-1} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & & & W_1 \\ & & & & 0 \\ & & & & \vdots \\ 0 & & & & 0 \end{bmatrix}$$

and

$$\phi = (\phi_1, \phi_2, \ldots, \phi_p)' .$$

Except for constants, the log likelihood function is

$$\ell = \frac{1}{2\sigma^2} \left(MZ - M(F\ U)\begin{pmatrix} \phi \\ \beta \end{pmatrix}\right)' \Sigma_T^{-1} \left(MZ - M(F\ U)\begin{pmatrix} \phi \\ \beta \end{pmatrix}\right) .$$

Letting F(t) represent the t^{th} column of F, then

$$\frac{\partial \ell}{\partial \phi_t} = \frac{Z'J_tQV_1}{\sigma} - \frac{F'(t)QV_1}{\sigma}$$

But,

$$\frac{Z'}{\sigma} = V_1' R(M^{-1})' + \mu_z/\sigma$$

so that

$$\partial\ell/\partial\phi_t = V_1' R(M^{-1})' J_t Q V_1 + \frac{1}{\sigma}(\mu_z' J_t - F'(t))QV_1 \quad .$$

But R, $(M^{-1})'$, J_t and Q all commute, so

$$\partial\ell/\partial\phi_t = V_1'(M^{-1})' J_t V_1 + \frac{1}{\sigma}(\mu_z' J_t - F'(t))QV_1 \quad . \qquad (4.5)$$

It follows that

$$E[(\partial\ell/\partial\phi_t)(\partial\ell/\partial\phi_s)] =$$

$$E[V_1'(M^{-1})' J_t V_1 V_1'(M^{-1})' J_s V_1]$$

$$+ E[V_1'(M^{-1})' J_t V_1 V_1' Q'(J_s' \mu_z - F(s))]/\sigma$$

$$+ E[V_1'(M^{-1})' J_s V_1 V_1' Q'(J_t' \mu_z - F(t))]/\sigma$$

$$+ E[(\mu_z' J_t - F'(t))QV_1 V_1' Q'(J_s \mu_z - F(s))]/\sigma^2 \qquad (4.6)$$

The two middle terms of (4.5) have the form $E[(V_1' AV_1)V_1' b]$, where A is a TxT matrix and b is a Tx1 vector. But this expectation vanishes, since V_1 is distributed MVN(0,I). Since $E[V_1 V_1'] = I$, the fourth term of (4.6) is equal to

$$(\mu_z' J_t - F'(t))\,\Sigma_T^{-1}\,(J_s' \mu_z - F(s))/\sigma^2 \quad .$$

The first term of (4.5) can be evaluated by noting that

$$(M^{-1})' = \sum_{i=0}^{T-1} \lambda_i J_i$$

where the λ_1's are defined by

$$\begin{aligned} \lambda_0 &= 1 \\ \lambda_i &= \sum_{t=1}^{k} \phi_t \lambda_{k-t}, \quad k = \min(i,p), \quad i \geq 1 . \end{aligned} \tag{4.7}$$

Lemma 4.1 can be applied to complete the evaluation of (4.6). Thus, for $t \geq s$,

$$\begin{aligned} &E[(\partial \ell/\partial \phi_t)(\partial \ell/\partial \phi_s)] \\ &= \sum_{j=0}^{T-t-1} (T-t-j)\, \lambda_j \lambda_{j+t-s} \\ &+ (\mu_z' J_t - F'(t))\, \Sigma_T^{-1} (J_s' \mu_z - F(s))/\sigma^2 . \end{aligned} \tag{4.8}$$

From (4.2) and (4.5),

$$\begin{aligned} &E[(\partial \ell/\partial \rho)(\partial \ell/\partial \phi_t] \\ &= E\left[\frac{G'QV_1}{\sigma} V_1' (M^{-1})' J_t V_1\right] \\ &+ E\left[\frac{G'QV_1}{\sigma} V_1' Q(J_t' \mu_z - F(t))\right] \\ &= G' \Sigma_T^{-1} (J_t' \mu_z - F(t))/\sigma^2 \end{aligned} \tag{4.9}$$

From (4.1) and (4.5) ,

$$E[(\partial \ell/\partial \theta_k)(\partial \ell/\partial \phi_t)$$

$$= - E[V_1' Q J_K V_1 V_1' (M^{-1})' J_t V_1]$$

$$- E[V_1' Q J_K V_1 V_1' Q(J_r' \mu_z - F(t))] \quad .$$

The second term above vanishes, and the first term can be shown to be

$$- \sum_{j=0}^{T-k-1} (T - k-j)\pi_j \lambda_{j+k-t} \quad , \quad k > t$$

and

$$- \sum_{j=0}^{T-t-1} (T-t-j) \lambda_j \pi_{j+t-k} \quad , \quad t \geq k \qquad (4.10)$$

Using the equations above, all the elements of the inverse covariance matrix of the parameter estimates can be evaluated. Since the MULE should converge to the MLE, the same large sample covariance matrix can be used for both estimators.

D. Computational Aspects

The computations can be arranged to avoid well known numerical pitfalls. Only two matrix inversions are required: the inversion of a (2p) x (2p) matrix to obtain $\hat{\rho}$, and the inversion of a p x p matrix to obtain $\hat{\phi}$.

It is also possible to do the computations in an order that avoids large memory requirements. All of the sample cases have been run on a PDP-11/70 mini-computer in a 27K partition. A series of more than 200 observations can be processed in this size partition. For comparison on other computers, the PDP-11/70 uses double words for floating point numbers, so the effective partition size is somewhat less than 27K.

The computations exploit the factorization $\Sigma_T^{-1} = QQ'$ (given in (2.9)). In the following, a *-matrix will represent a matrix pre-multiplied by Q', e.g., $A^* = Q'A$, for any T x k matrix A.

From (3.1) the sum of squares in the likelihood function can be written

$$\begin{aligned}(Y - G\rho)' \Sigma_T^{-1} (Y - G\rho) &= (Q'Y - Q'G\rho)' (Q'Y - Q'G\rho) \\ &= (Y^* - G^*\rho)' (Y^* - G^*\rho)\end{aligned}$$

so that

$$\hat{\rho} = (G^{*\prime}G^*)^{-1} G^{*\prime} Y^* . \qquad (4.11)$$

It follows that

$$\begin{aligned}(Y^* - G^*\hat{\rho})' (Y^* - G^*\hat{\rho}) &= Y^*(I - G^*(G^{*\prime}G^*)^{-1}G^*)Y^* \\ &= Y^{*\prime} \Psi Y^*\end{aligned}$$

Theorem 3.1 implies that Ψ is independent of the AR parameters. From (3.8)

$$\begin{aligned}Y^* &= Q'Z - Q'H\phi \\ &= Z^* - H^*\phi\end{aligned}$$

so that

$$Y^{*\prime}\Psi Y^{*} = (Z^{*} - H^{*}\phi)'\Psi(Z^{*} - H^{*}\phi)$$

and

$$\hat{\phi} = (H^{*\prime}\Psi H^{*})^{-1}H^{*\prime}\Psi Z^{*} \quad . \tag{4.12}$$

Thus, the sum of squares as a function of the MA parameters is

$$S(\theta) = (Z^{*} - H^{*}\hat{\phi})'\Psi(Z^{*} - H^{*}\hat{\phi}) \quad . \tag{4.13}$$

The maximum likelihood estimate of σ^2 is given by

$$\hat{\sigma}^2 = S(\theta)/T \quad . \tag{4.14}$$

If (4.13) and (4.14) are substituted into the likelihood (2.11) we obtain the ULF(θ):

$$ULF(\theta) = \left(\frac{T}{S(\theta)}\right)^{T/2} \quad . \tag{4.15}$$

For moderate to large sample sizes, $T/S(\theta)$ should be scaled so the maximum is near one to avoid over- or under-flow.

Because many of the calculations required to find the MLE and the MULE are identical, it is efficient to use a single computer program to evaluate both estimates. Calculation of the MULE does not require iteration; hence, the MULE was computed first and used as a starting value in the iterative MLE calculation.

E. Backwards System Process

In the estimation procedures that have been developed the initial state of the system is estimated along with other parameters. There are instances in which an estimate of the system's state at some other time would conceptually be a more useful parameter. An attempt was made to

reparameterize the estimation procedure so that the state of the system at time T+1 was used as a parameter.

The attempt was not entirely successful. The model equations can be written so that the estimation procedures can be applied by reversing the data, so that $z_1 \rightarrow z_T^*$, $z_2 \rightarrow z_{T-1}^*, \ldots, z_T \rightarrow z_1^*$ etc. The MA parameters of the backwards system remain the same, because of the duality noted earlier. The AR and the regression parameters, however, are not invariant. This is not a difficulty for the regression parameters, but because the duality between the forward and backward system does not hold for the AR parameters of a forced system, the backwards representation is a non-stable system.

Several attempts were made to estimate the parameters using the backwards representation from sample data generated by a stable forward process. The resulting estimates were highly variable, and tended to be values giving a stable backward process. For example, if the data were generated using a forward process with an AR parameter of $\phi_f = 0.5$, then the backwards process AR parameter should be $\phi_b = 2.0$. But the estimates were generally such that $|\phi_b| < 1$.

After the fact the observed behavior is not surprising. For the sample input sequence the backwards system is almost surely explosively unstable. But the observed sample record is well-behaved, which is an unlikely behavior for an explosive system. Hence one should not expect estimates based on the likelihood function to lead to an unstable system. What remains is a perplexing problem. It would be useful to examine the system in the backward form. Direct inversion leads to an

intractable model. Is there another inversion which will meaningfully provide the analysis sought?

V. SIMULATION RESULTS AND EXAMPLE

A. Design of Simulation Studies

The simulation studies were intended to examine and compare the behavior of the MLE and the MULE. The first study was designed to investigate the small sample behavior of the two estimators, the second to examine the effect of sample size.

The studies were conducted on a DEC PDP 11-70, using the random number generator supplied by the manufacturer. The generator is based on the linear congruential algorithm

$$x_{i+1} = (ax_i + c) \bmod m \quad .$$

In particular the generator uses the constants

$$a = 2^{16} + 3$$

$$c = 0$$

$$m = 2^{32} \quad .$$

The generator was recently tested by Nicholson, et al. (1978). The period, length, and frequency of the generator were evaluated, and runs and lag products were tested. The tests indicated that similar sequences appeared to occur more frequently than expected, the distribution was reasonably uniform, more significant runs occurred than expected, and autocorrelation appeared in positive and negative "clumps". The general conclusion was that the generator would produce reasonable "random" numbers for most applications.

The random numbers were generated using a seed that was a function of the system clock time. Thus, repeated runs would in general produce different sequences of random numbers. The uniformly distributed numbers were transformed into normal deviates using the Box-Muller method (Hammersley and Handscomb, 1964).

The small sample tests were carried out using 40 samples of size 30. In order to investigate the behavior of the estimators over the entire parameter space, p was held equal to 1. Although the AR and regression parameters are the parameters of most intrinsic interest, it was anticipated that the value of the AR and MA parameters would have the most influence on the behavior of the estimator. Hence values of ϕ and θ were selected that spanned the parameter space. The values used were $\pm$ .95, $\pm$ .5 and $\pm$ .1 for both parameters, resulting in 36 combinations.

The regression parameter β and the variance parameter σ^2 were held constant at 2.0 and 2.25, respectively. The input function was a sine wave with period $\pi/4$. This constitutes a fairly pessimistic selection, since the magnitude of the noise term will frequently be greater than the magnitude of the accessible input term.

The sample size tests were conducted to investigate how seriously small sample sizes influenced the estimators. Four tests were planned at values of ϕ and θ of $\pm$ .5. An additional test was carried out at ϕ = -.95, θ = .95. For each test 10 samples of size 200 were generated. Parameter estimates were made using an increasing number of points in steps of 25.

B. Results of the Simulation Studies

The sample size simulation studies are summarized in Tables 1 through 6. In general the MULE appears to be better behaved than the MLE. The standard error of the MULE is smaller than that of the MLE except for nominal values near the boundary of the parameter space. However, as can be seen from Figures 7-12, the smaller SE (and MSE) of the MLE near the boundary is largely due to the fact that many of the MLE's are on the boundary.

Figures 1 through 6 are plots of some sample likelihood functions. The negative log of the ULF is plotted versus θ for nominal values of $\theta = .95$ and $\phi = \pm .95$, $\pm .5$ and $\pm .1$. The functions are generally skewed. Figures 3, 5, and 6 are particularly interesting because they represent concave likelihood functions. Figure 3 has nearly equal minima at both endpoints of the parameter space, implying that the likelihood is about the same for $\theta = +1$ as it is for $\theta = -1$. Similar behavior is exhibited in Figure 6, except that the likelihood function has three nearly equal maxima.

Confirmation of the fact that the selected likelihood plots were not unusually aberrant is presented in Figures 10, 11, and 12. These are scatter plots of the MULE and the MLE of ϕ versus the corresponding estimates of θ for nominal $\theta = .95$ and $\phi = -.1$, $-.5$, and $-.95$, respectively. For $\phi = -.1$, most of the MLE's of θ are clumped on the right boundary. For $\phi = -.5$, the spread increases and some MLE's appear on the left boundary. For $\phi = -.95$ the estimates appear in a band from the upper left to the lower right.

TABLE 1. SUMMARY OF SMALL SAMPLE SIMULATION TESTS FOR PHI = -0.95

MULE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.890	0.055	0.007	-0.914	0.071	0.006	1.975	0.204	0.042
-0.500	-0.511	0.189	0.036	-0.930	0.082	0.007	2.013	0.259	0.067
-0.100	-0.133	0.211	0.046	-0.916	0.091	0.009	1.995	0.374	0.140
0.100	0.028	0.290	0.089	-0.894	0.145	0.024	1.866	0.437	0.209
0.500	0.469	0.299	0.091	-0.861	0.163	0.034	1.867	0.701	0.509
0.950	0.140	0.431	0.842	-0.158	0.329	0.736	1.329	0.758	1.024

MLE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.978	0.044	0.003	-0.912	0.072	0.007	1.974	0.196	0.039
-0.500	-0.560	0.213	0.049	-0.927	0.087	0.008	2.010	0.258	0.067
-0.100	-0.151	0.242	0.061	-0.912	0.096	0.011	1.990	0.376	0.142
0.100	0.019	0.338	0.120	-0.887	0.167	0.032	1.858	0.444	0.217
0.500	0.555	0.377	0.145	-0.881	0.150	0.027	1.888	0.709	0.516
0.950	0.257	0.798	1.117	-0.269	0.548	0.764	1.497	0.930	1.119

TABLE 2. SUMMARY OF SMALL SAMPLE SIMULATION TESTS FOR PHI = -0.50

MULE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.890	0.047	0.006	-0.533	0.140	0.021	2.037	0.261	0.069
-0.500	-0.593	0.168	0.037	-0.448	0.186	0.037	1.961	0.300	0.092
-0.100	-0.069	0.338	0.115	-0.514	0.234	0.055	2.115	0.501	0.265
0.100	-0.031	0.376	0.159	-0.384	0.288	0.096	1.847	0.608	0.394
0.500	0.121	0.414	0.315	-0.212	0.335	0.195	1.838	0.768	0.615
0.950	0.592	0.259	0.195	-0.166	0.203	0.153	0.479	0.692	0.750

MLE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.984	0.038	0.003	-0.518	0.144	0.021	2.007	0.259	0.067
-0.500	-0.682	0.193	0.070	-0.410	0.209	0.052	1.910	0.317	0.109
-0.100	-0.086	0.439	0.193	-0.504	0.278	0.077	2.097	0.541	0.302
0.100	-0.095	0.600	0.398	-0.336	0.425	0.207	1.778	0.701	0.541
0.500	0.154	0.723	0.643	-0.250	0.525	0.338	1.903	0.913	0.843
0.950	0.886	0.274	0.079	-0.363	0.239	0.076	1.705	0.817	0.754

TABLE 3. SUMMARY OF SMALL SAMPLE SIMULATION TESTS FOR PHI = -0.10

MULE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.862	0.075	0.013	-0.170	0.151	0.028	2.096	0.336	0.122
-0.500	-0.359	0.281	0.099	-0.220	0.200	0.054	2.271	0.509	0.333
-0.100	0.137	0.372	0.195	-0.284	0.305	0.127	2.258	0.611	0.439
0.100	0.237	0.266	0.090	-0.242	0.269	0.093	2.268	0.663	0.512
0.500	0.335	0.391	0.180	0.005	0.308	0.106	1.894	0.766	0.597
0.950	0.742	0.317	0.144	0.023	0.211	0.060	1.880	0.793	0.644

MLE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.970	0.064	0.005	-0.147	0.158	0.027	2.061	0.342	0.121
-0.500	-0.512	0.392	0.153	-0.125	0.259	0.068	2.111	0.579	0.348
-0.100	0.205	0.644	0.507	-0.328	0.457	0.261	2.301	0.740	0.638
0.100	0.308	0.579	0.378	-0.300	0.459	0.251	2.363	0.903	0.947
0.500	0.419	0.610	0.378	-0.088	0.451	0.204	2.084	1.025	1.059
0.950	0.892	0.408	0.170	-0.058	0.263	0.071	2.031	0.886	0.786

TABLE 4. SUMMARY OF SMALL SAMPLE SIMULATION TESTS FOR PHI = 0.10

MULE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.840	0.174	0.042	-0.028	0.188	0.052	2.254	0.382	0.210
-0.500	-0.361	0.360	0.149	-0.068	0.253	0.093	2.416	0.610	0.545
-0.100	0.051	0.454	0.229	-0.141	0.356	0.185	2.508	0.837	0.958
0.100	0.128	0.433	0.188	-0.015	0.343	0.131	2.283	0.869	0.836
0.500	0.419	0.365	0.139	0.176	0.276	0.082	1.979	0.566	0.321
0.950	0.764	0.216	0.081	0.161	0.209	0.047	1.947	0.927	0.863

MLE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.939	0.184	0.034	-0.003	0.190	0.047	2.209	0.391	0.196
-0.500	-0.525	0.421	0.178	0.016	0.297	0.095	2.247	0.714	0.570
-0.100	0.044	0.708	0.522	-0.132	0.499	0.303	2.512	1.112	1.499
0.100	0.164	0.704	0.500	-0.082	0.515	0.298	2.415	1.123	1.433
0.500	0.498	0.479	0.229	0.115	0.328	0.108	2.086	0.623	0.396
0.950	0.888	0.328	0.112	0.113	0.230	0.053	2.037	0.967	0.936

TABLE 5. SUMMARY OF SMALL SAMPLE SIMULATION TESTS FOR PHI = 0.50

MULE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.861	0.111	0.020	0.464	0.073	0.007	2.110	0.239	0.067
-0.500	-0.464	0.307	0.095	0.444	0.114	0.016	2.179	0.330	0.141
-0.100	-0.077	0.322	0.104	0.435	0.194	0.042	2.217	0.624	0.436
0.100	0.099	0.304	0.092	0.403	0.217	0.057	2.353	0.860	0.865
0.500	0.513	0.220	0.049	0.457	0.159	0.027	2.108	0.734	0.550
0.950	0.862	0.086	0.015	0.449	0.186	0.037	2.037	1.062	1.128

MLE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.970	0.073	0.006	0.477	0.066	0.005	2.072	0.227	0.057
-0.500	-0.542	0.362	0.133	0.457	0.121	0.016	2.137	0.291	0.103
-0.100	-0.112	0.373	0.139	0.445	0.207	0.046	2.193	0.649	0.459
0.100	0.077	0.403	0.163	0.403	0.235	0.064	2.346	0.891	0.914
0.500	0.585	0.246	0.068	0.429	0.176	0.036	2.175	0.766	0.617
0.950	0.975	0.076	0.006	0.430	0.188	0.040	2.074	1.074	1.159

TABLE 6. SUMMARY OF SMALL SAMPLE SIMULATION TESTS FOR PHI = 0.95

MULE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.886	0.052	0.007	0.938	0.025	0.001	2.002	0.134	0.018
-0.500	-0.626	0.193	0.053	0.950	0.043	0.002	2.003	0.244	0.060
-0.100	-0.166	0.267	0.076	0.947	0.039	0.002	1.970	0.342	0.118
0.100	0.059	0.348	0.123	0.920	0.060	0.005	2.034	0.392	0.154
0.500	0.493	0.225	0.051	0.937	0.059	0.004	2.171	0.672	0.481
0.950	0.874	0.076	0.012	0.934	0.057	0.004	2.156	0.820	0.697

MLE ESTIMATOR

NOMINAL THETA	MEAN THETA	SE THETA	MSE THETA	MEAN PHI	SE PHI	MSE PHI	MEAN BETA	SE BETA	MSE BETA
-0.950	-0.976	0.053	0.003	0.940	0.025	0.001	2.006	0.128	0.016
-0.500	-0.704	0.225	0.092	0.954	0.042	0.002	2.005	0.247	0.061
-0.100	-0.171	0.290	0.089	0.947	0.039	0.002	1.969	0.344	0.119
0.100	0.064	0.406	0.166	0.918	0.061	0.005	2.034	0.391	0.154
0.500	0.546	0.293	0.088	0.933	0.061	0.004	2.172	0.677	0.488
0.950	0.963	0.080	0.007	0.931	0.058	0.004	2.156	0.826	0.706

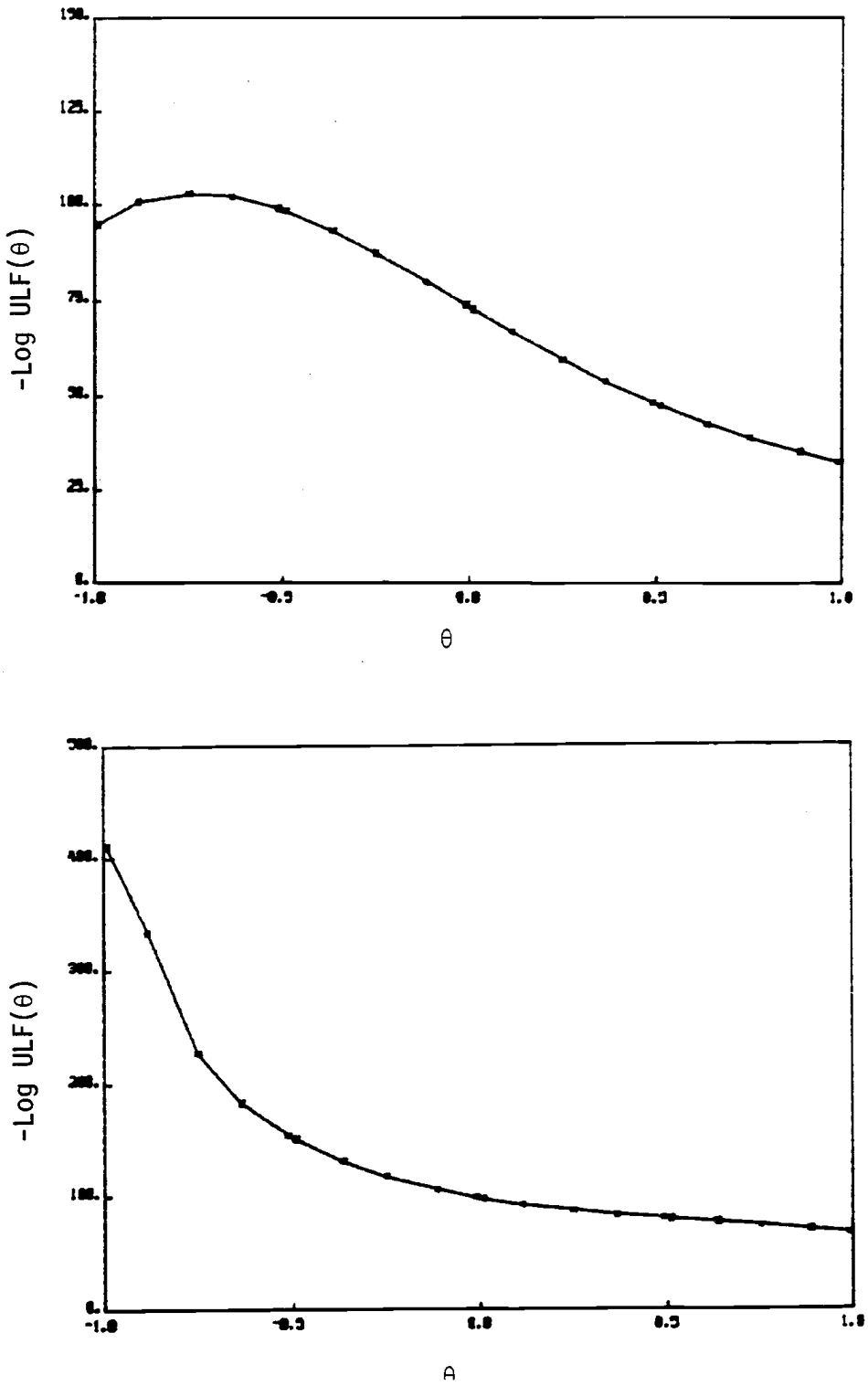


Figure 1. Realizations of ULF(θ) for nominal values θ = 0.95 ϕ = 0.95

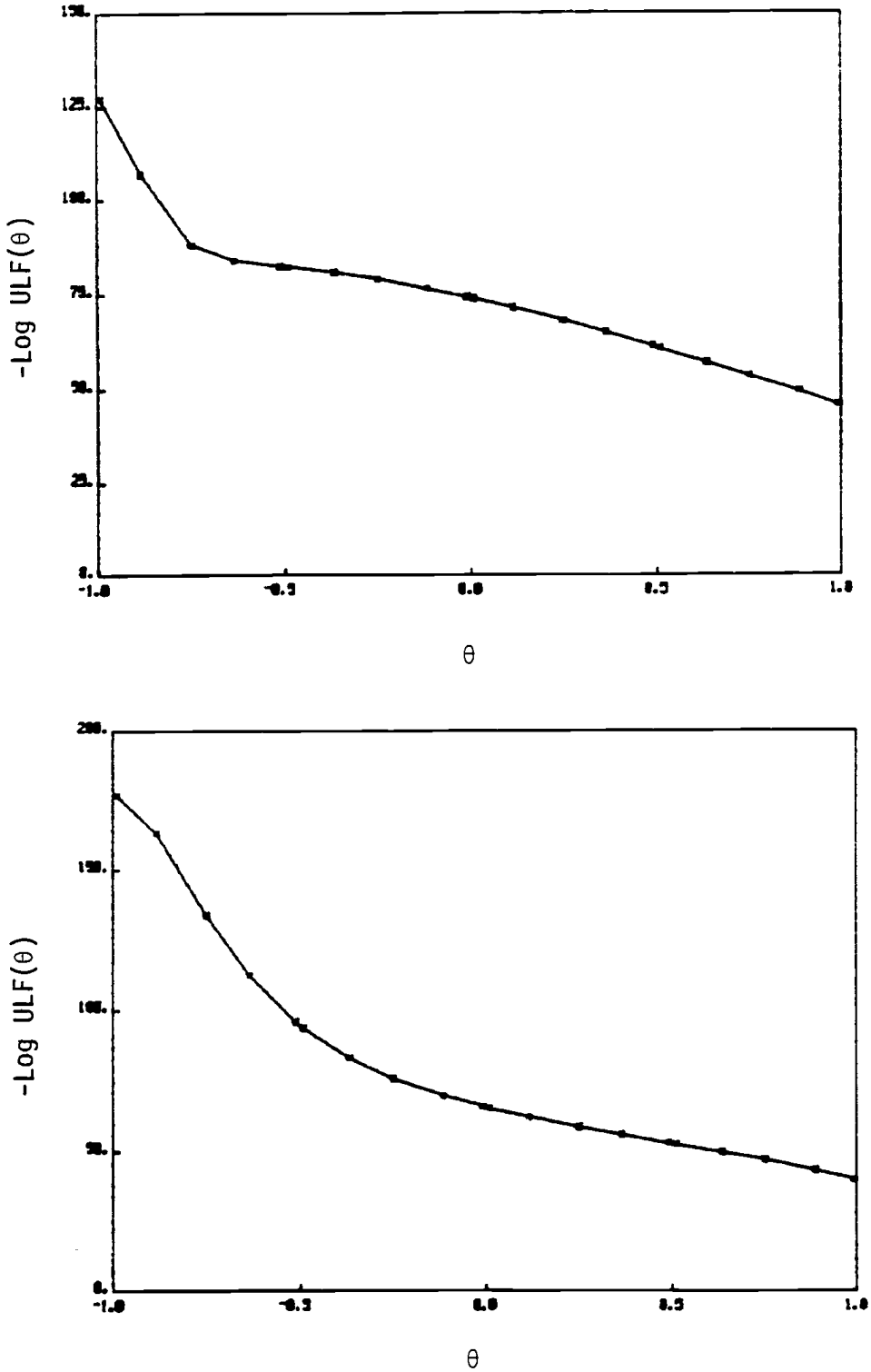


Figure 2. Realizations of ULF(θ) for nominal values θ = 0.95 ϕ = 0.50

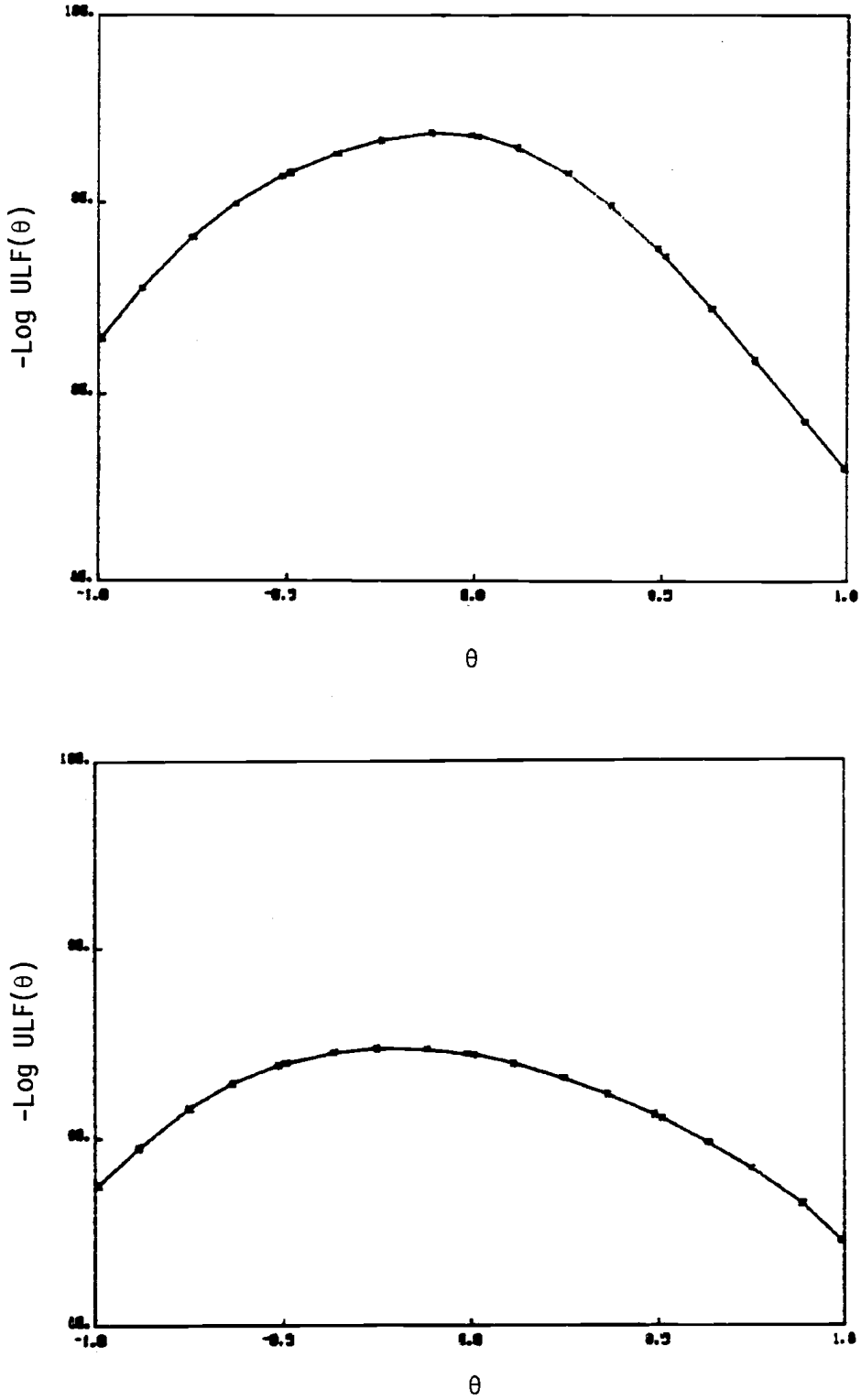


Figure 3. Realizations of ULF(θ) for nominal values θ = 0.95 ϕ = 0.10

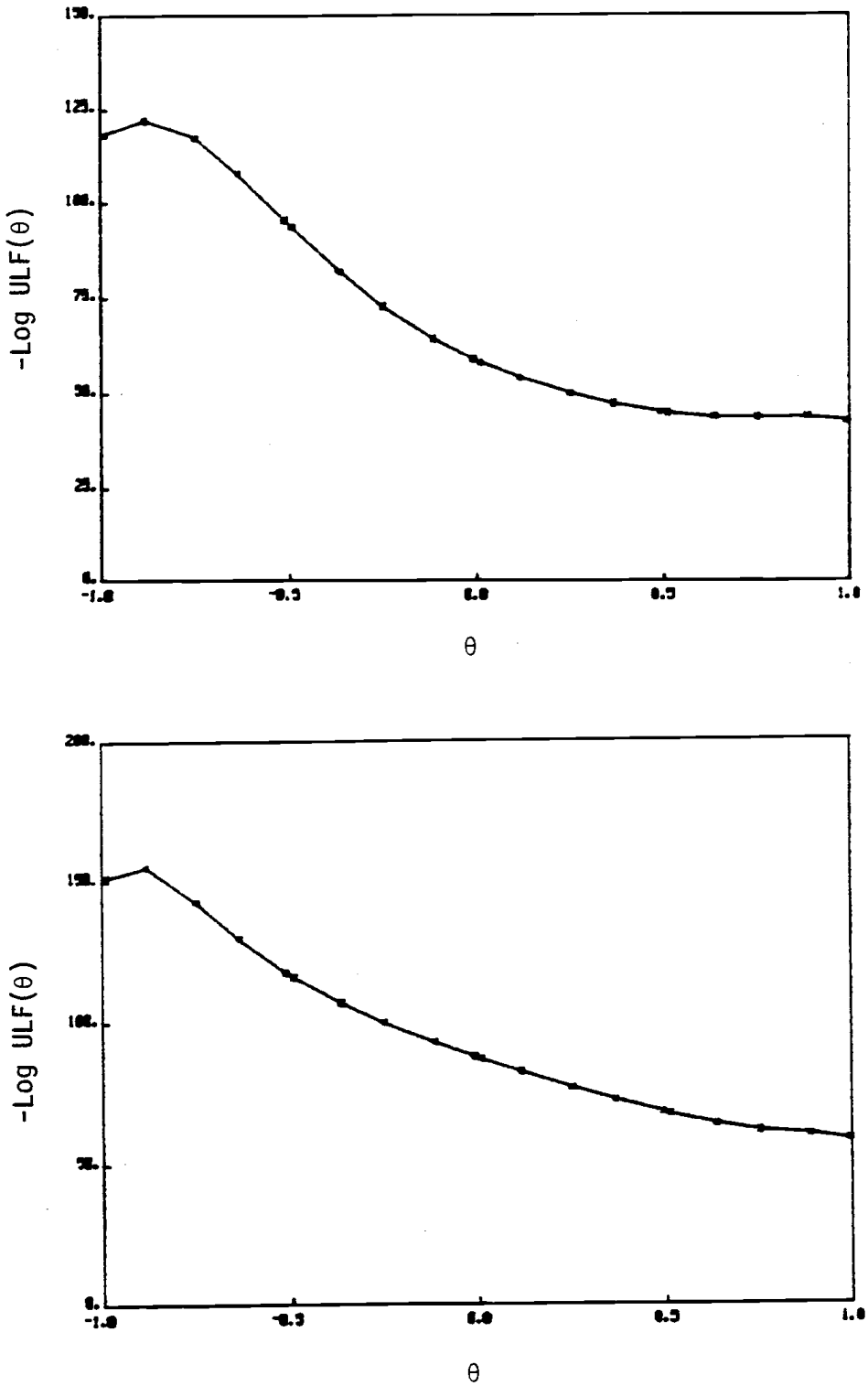
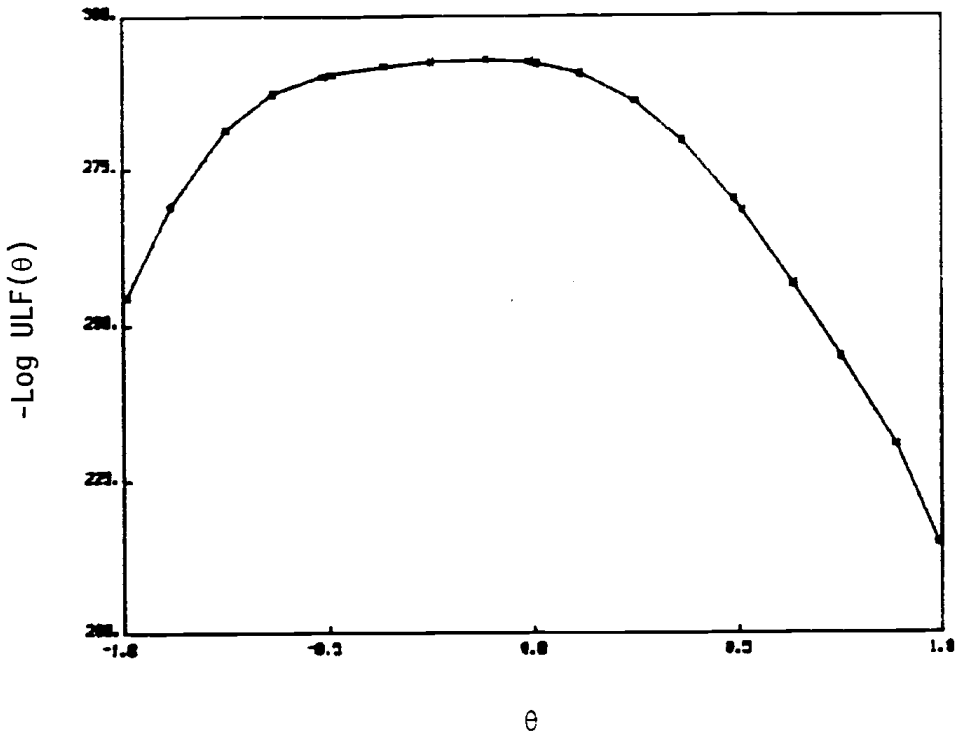


Figure 4. Realizations of ULF(θ) for nominal values θ = 0.95 ϕ = -0.10



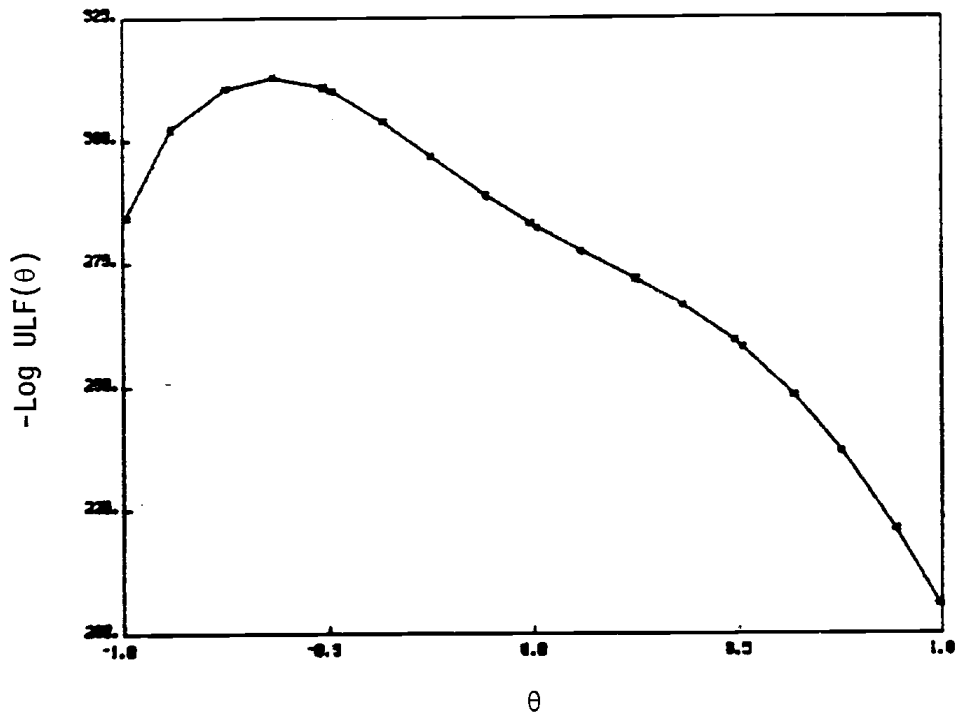
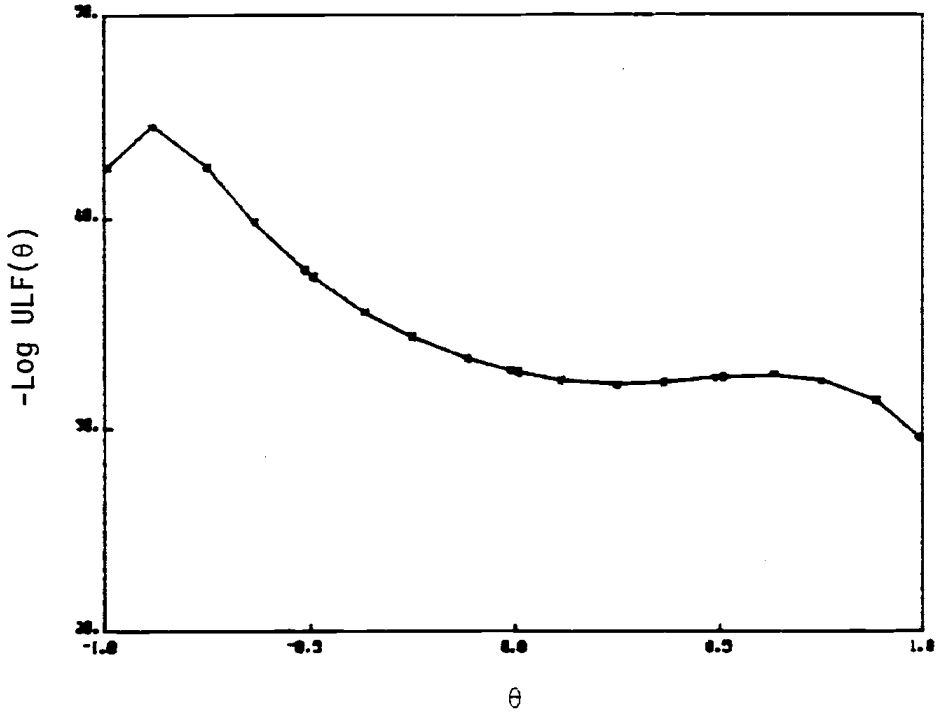


Figure 5. Realizations of ULF(θ) for nominal values θ = 0.95 ϕ = -0.50



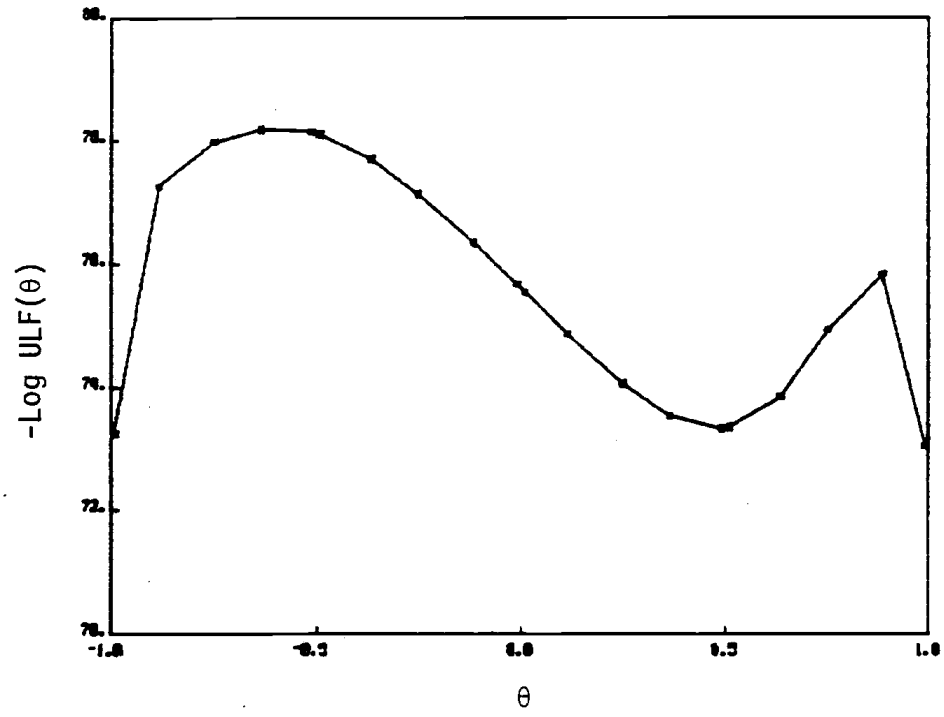
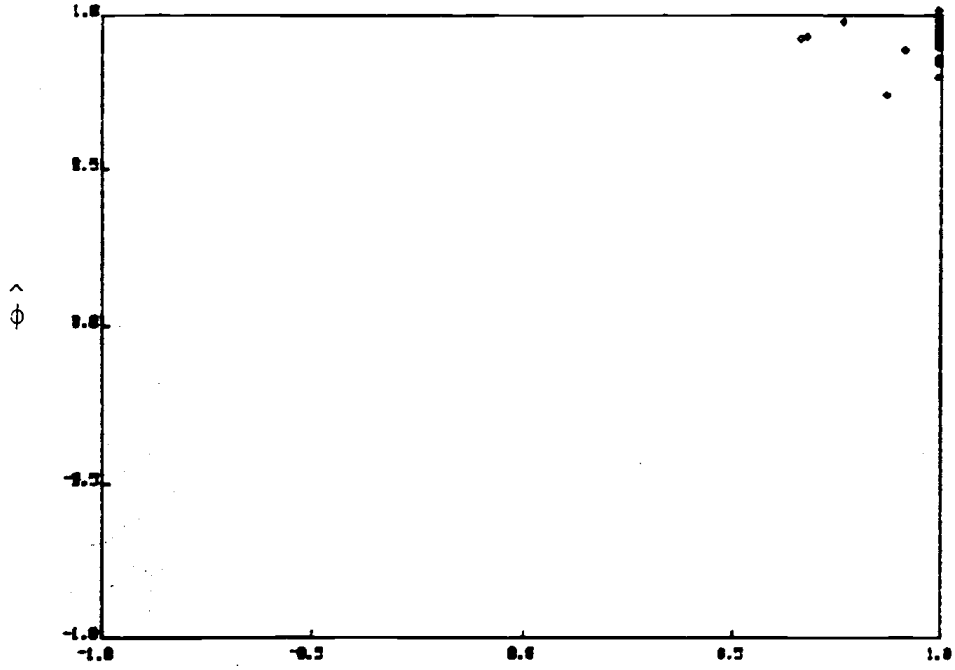


Figure 6. Realizations of ULF(θ) for nominal values $\theta = 0.95$ $\phi = -0.95$



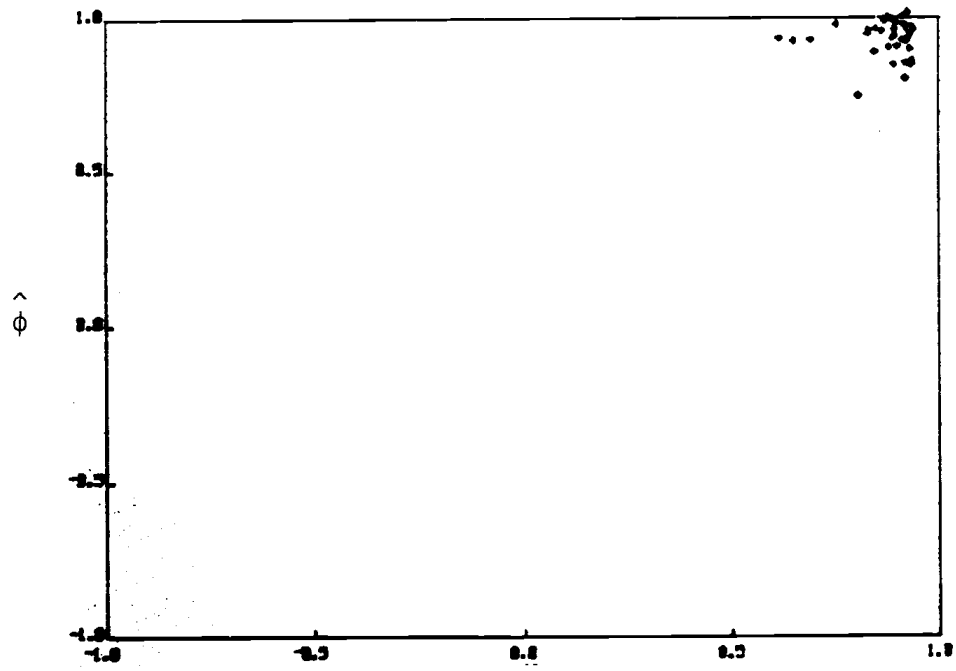
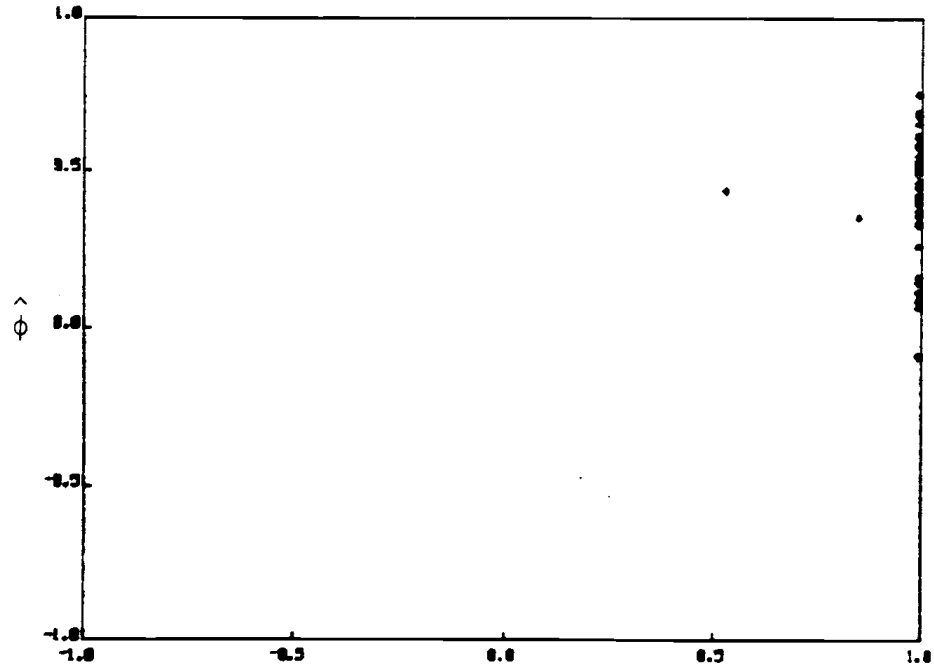


Figure 7. Scatter plots of MLE and MULE for nominal values $\theta = 0.95$, $\phi = 0.95$.



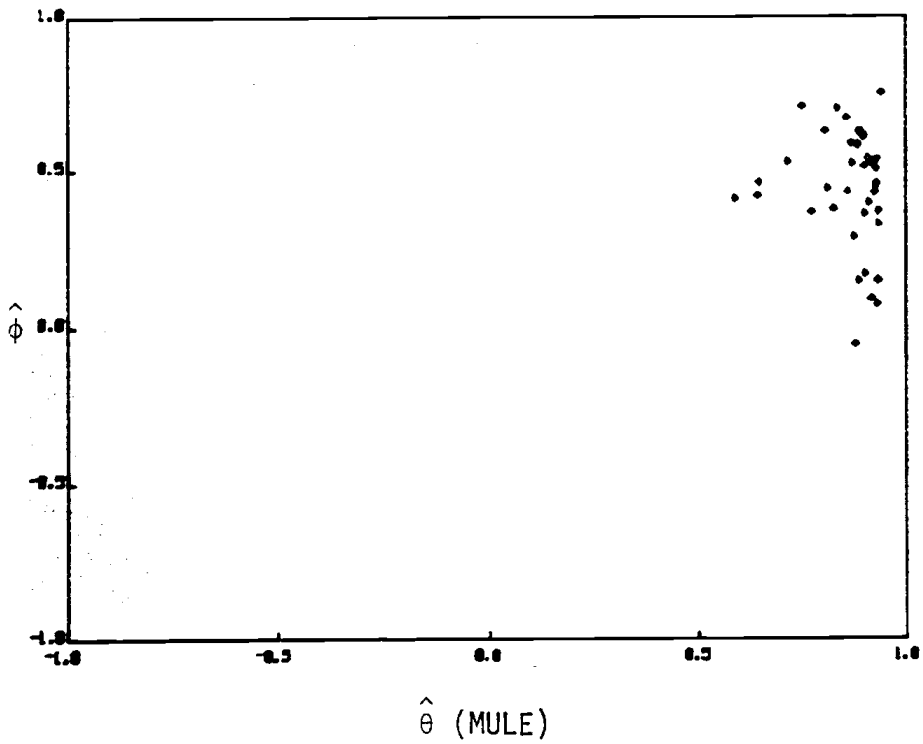
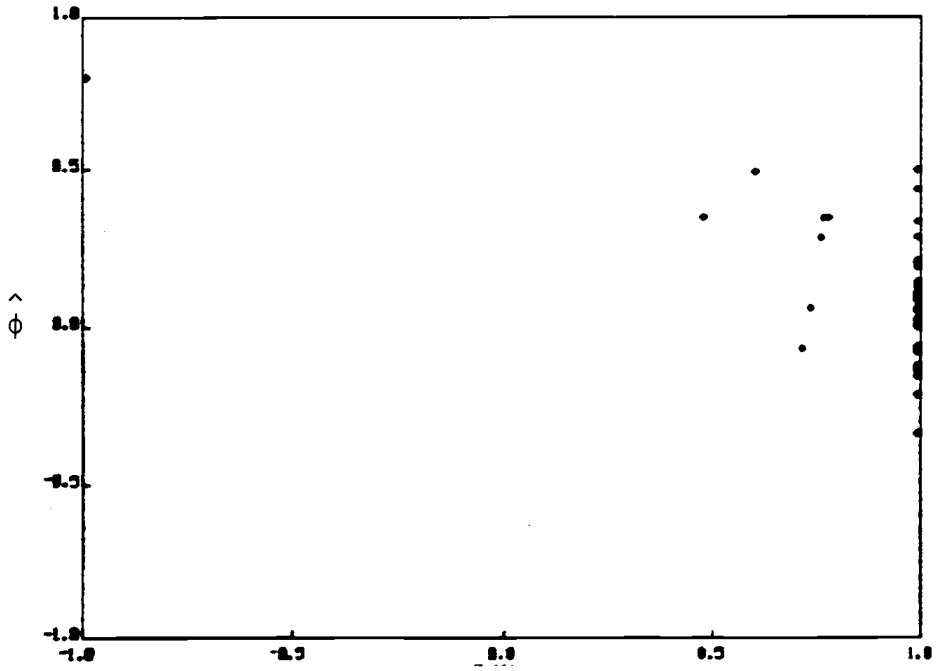


Figure 8. Scatter plots of MLE and MULE for nominal values $\theta = 0.95$, $\phi = 0.50$.



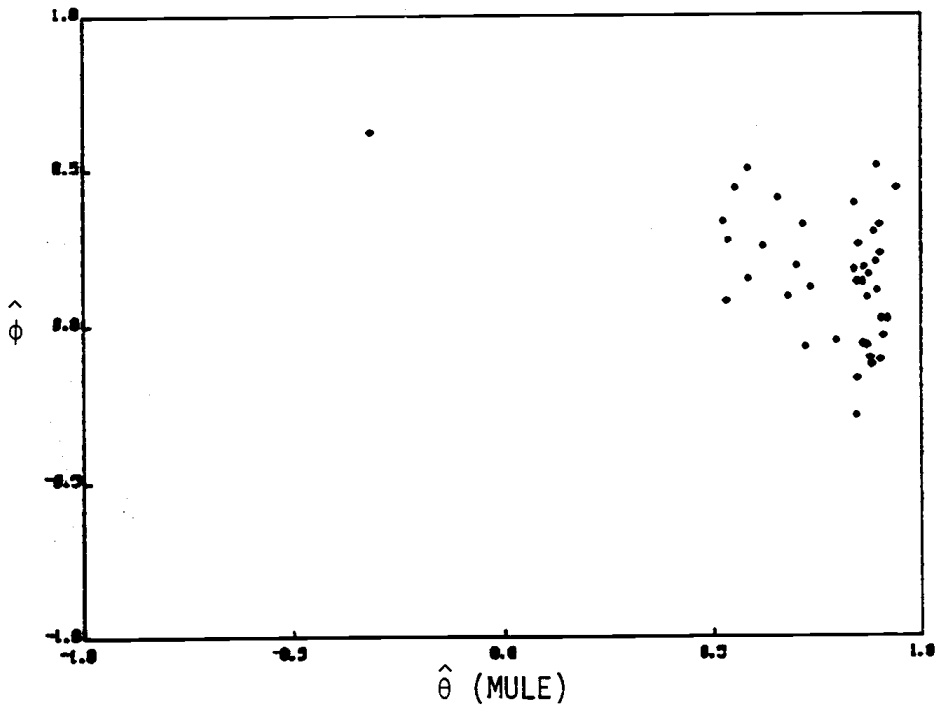


Figure 9. Scatter plots of MLE and MULE for nominal values $\theta = 0.95$, $\phi = 0.10$.

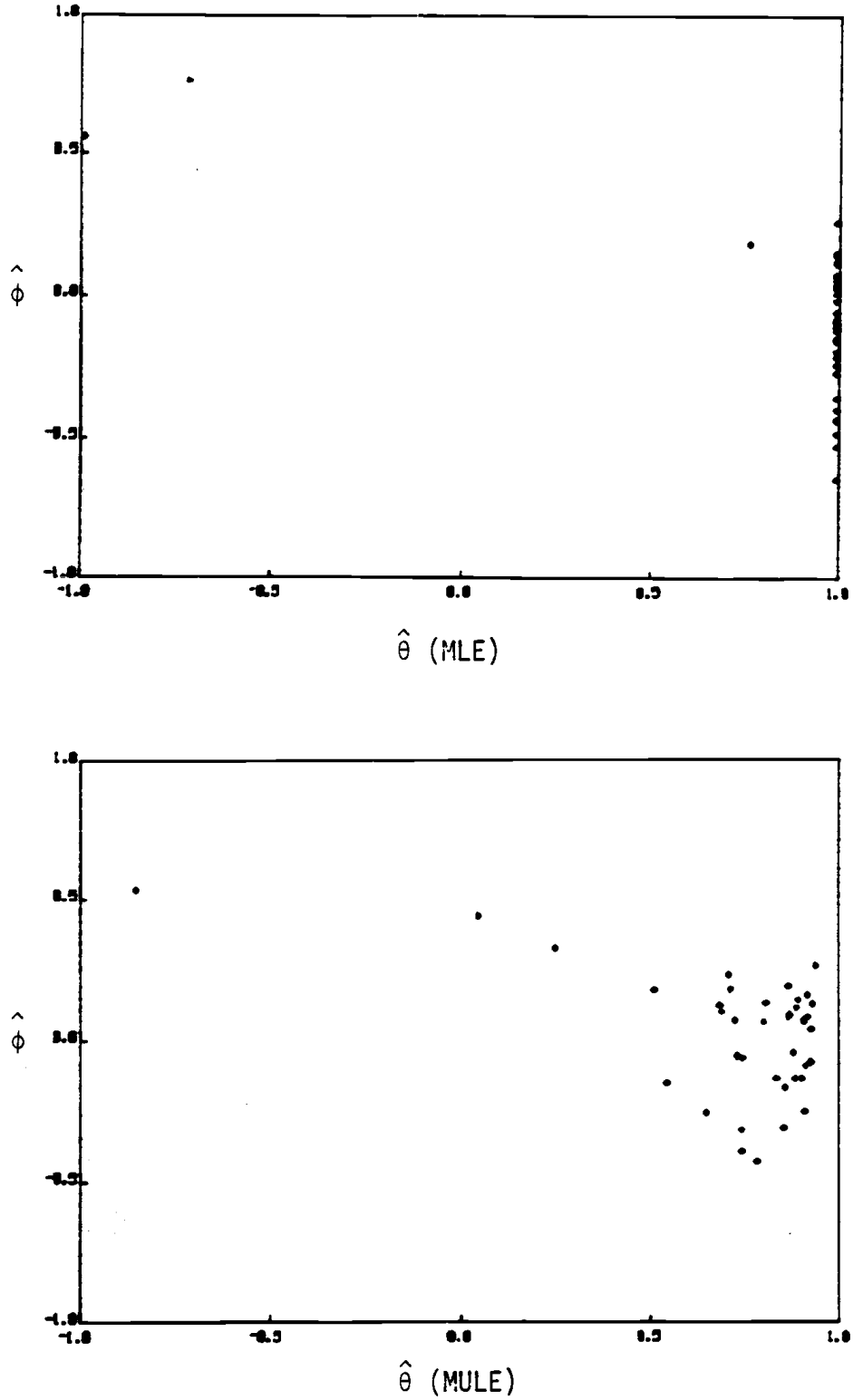
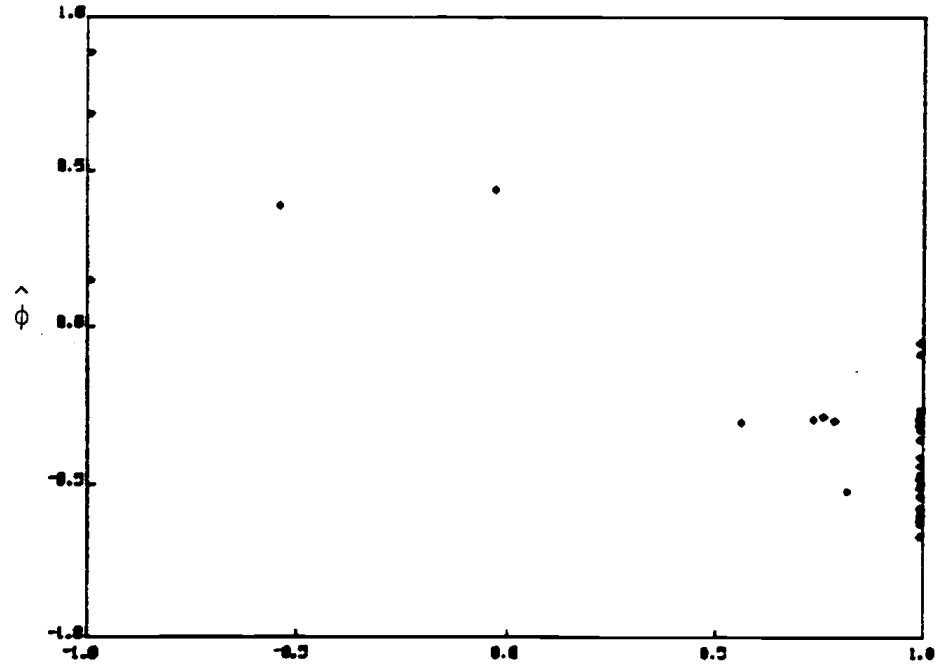


Figure 10. Scatter plots of MLE and MULE for nominal values $\theta = 0.95$, $\phi = -0.10$



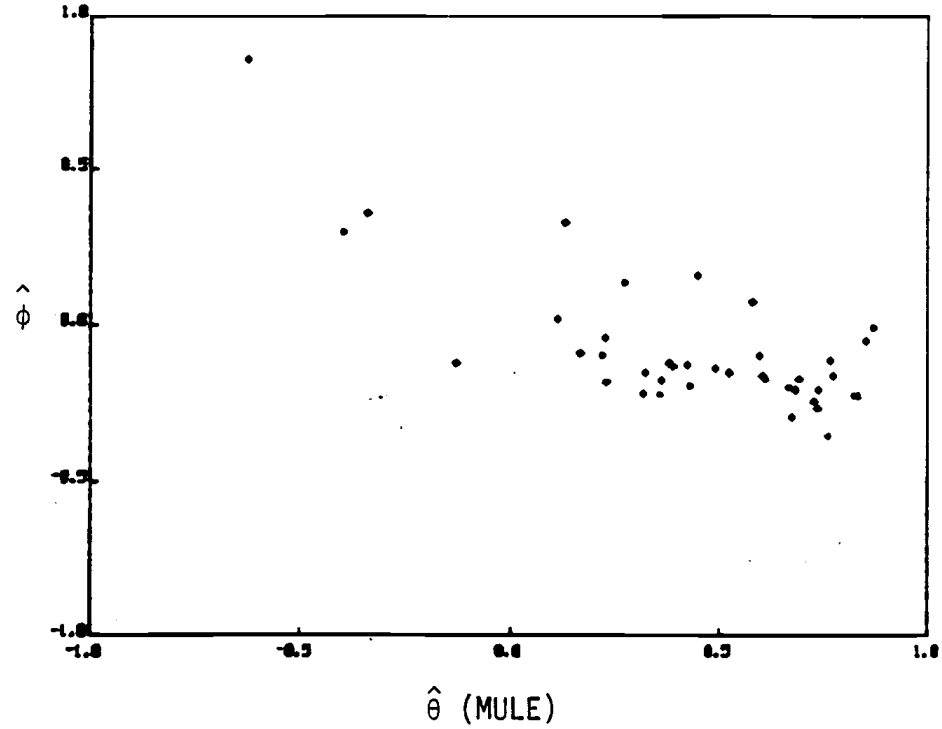
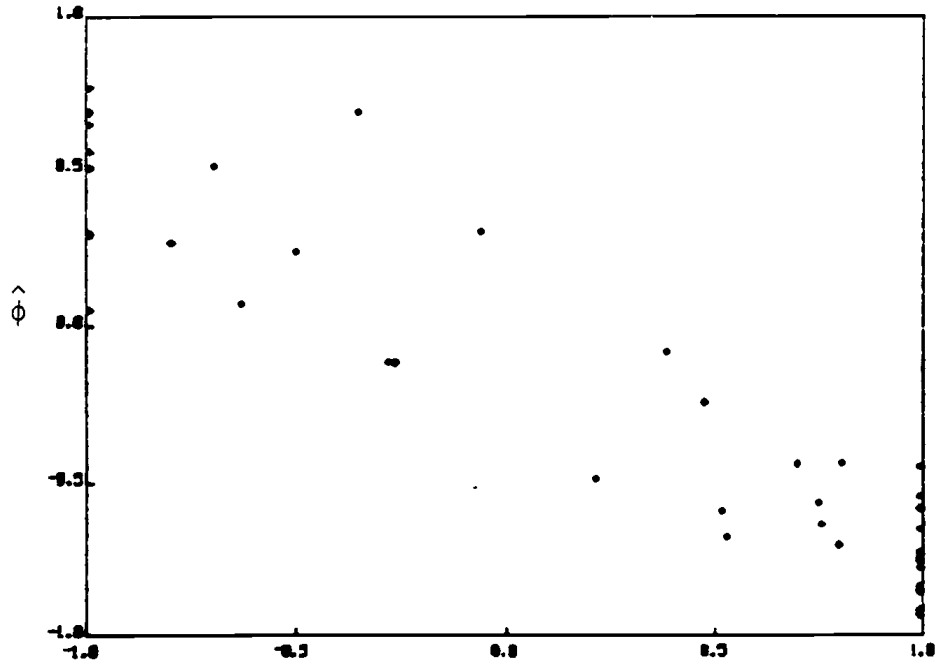


Figure 11. Scatter plots of MLE and MULE for nominal values $\theta = 0.95$, $\phi = -0.50$



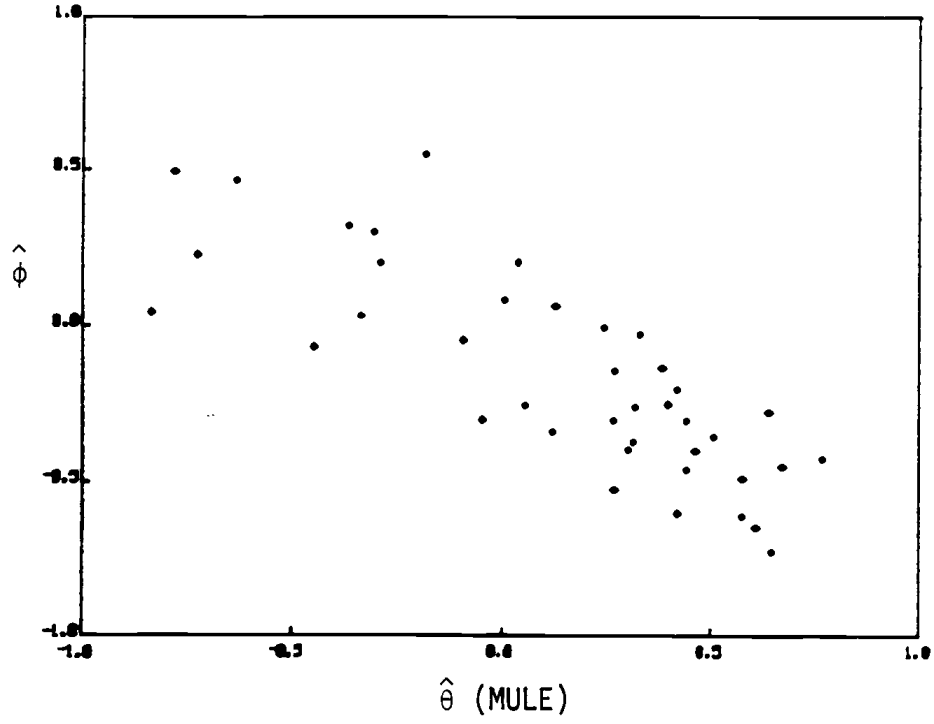


Figure 12. Scatter plots of MLE and MULE for nominal values $\theta = 0.95$, $\phi = -0.95$.

TABLE 7. SUMMARY OF SAMPLE SIZE TESTS FOR THETA = 0.50 AND PHI = 0.50

MEAN VALUES

N	MULE THETA	MLE THETA	MULE PHI	MLE PHI	MULE SIGMA	MLE SIGMA	MULE BETA	MLE BETA
25	0.511	0.598	0.513	0.490	1.411	1.400	2.635	2.708
50	0.496	0.536	0.507	0.489	1.488	1.484	2.065	2.114
75	0.524	0.545	0.510	0.502	1.485	1.485	2.063	2.082
100	0.490	0.510	0.521	0.513	1.481	1.481	1.992	2.011
125	0.490	0.506	0.528	0.522	1.473	1.473	1.997	2.011
150	0.474	0.487	0.539	0.535	1.470	1.470	1.983	1.995
175	0.489	0.498	0.522	0.519	1.477	1.477	2.081	2.090
200	0.503	0.509	0.523	0.521	1.484	1.484	2.051	2.056

STANDARD DEVIATIONS

N	MULE THETA	MLE THETA	MULE PHI	MLE PHI	MULE SIGMA	MLE SIGMA	MULE BETA	MLE BETA
25	0.221	0.271	0.147	0.161	0.261	0.261	0.505	0.539
50	0.156	0.191	0.121	0.139	0.203	0.303	0.415	0.455
75	0.088	0.089	0.064	0.065	0.144	0.144	0.271	0.278
100	0.093	0.094	0.070	0.072	0.123	0.123	0.290	0.291
125	0.076	0.074	0.063	0.064	0.113	0.113	0.309	0.309
150	0.076	0.073	0.054	0.055	0.083	0.083	0.284	0.282
175	0.060	0.056	0.055	0.054	0.080	0.080	0.280	0.277
200	0.062	0.058	0.057	0.057	0.099	0.099	0.294	0.290

TABLE 8. SUMMARY OF SAMPLE SIZE TESTS FOR THETA = 0.50 AND PHI = -0.50

MEAN VALUES

N	MULE THETA	MLE THETA	MULE PHI	MLE PHI	MULE SIGMA	MLE SIGMA	MULE BETA	MLE BETA
25	0.127	0.147	-0.168	-0.204	1.419	1.353	1.486	1.549
50	0.343	0.469	-0.342	-0.441	1.404	1.377	1.854	2.013
75	0.463	0.550	-0.415	-0.467	1.438	1.415	1.919	2.002
100	0.432	0.529	-0.398	-0.466	1.478	1.464	1.851	1.977
125	0.447	0.496	-0.413	-0.444	1.469	1.460	1.888	1.952
150	0.491	0.493	-0.470	-0.454	1.471	1.463	1.889	1.885
175	0.450	0.455	-0.427	-0.414	1.473	1.470	1.833	1.807
200	0.405	0.415	-0.378	-0.372	1.484	1.480	1.771	1.754

STANDARD DEVIATIONS

N	MULE THETA	MLE THETA	MULE PHI	MLE PHI	MULE SIGMA	MLE SIGMA	MULE BETA	MLE BETA
25	0.542	0.929	0.451	0.651	0.209	0.198	0.693	0.945
50	0.516	0.757	0.446	0.578	0.123	0.122	0.770	0.981
75	0.352	0.572	0.361	0.520	0.127	0.131	0.703	0.879
100	0.319	0.602	0.329	0.564	0.119	0.119	0.611	0.890
125	0.279	0.477	0.293	0.467	0.124	0.124	0.497	0.752
150	0.234	0.437	0.244	0.429	0.100	0.101	0.414	0.661
175	0.168	0.426	0.176	0.417	0.094	0.094	0.370	0.541
200	0.210	0.414	0.229	0.408	0.077	0.077	0.317	0.465

TABLE 9. SUMMARY OF SAMPLE SIZE TESTS FOR THETA = -0.50 AND PHI = 0.50

MEAN VALUES

N	MULE THETA	MLE THETA	MULE PHI	MLE PHI	MULE SIGMA	MLE SIGMA	MULE BETA	MLE BETA
25	-0.512	-0.752	0.404	0.470	1.442	1.403	2.189	2.011
50	-0.504	-0.559	0.452	0.471	1.471	1.468	2.118	2.069
75	-0.459	-0.500	0.454	0.466	1.490	1.489	2.066	2.034
100	-0.484	-0.513	0.464	0.473	1.507	1.506	2.073	2.051
125	-0.497	-0.520	0.466	0.472	1.494	1.494	2.070	2.053
150	-0.470	-0.488	0.457	0.463	1.495	1.494	2.107	2.092
200	-0.462	-0.475	0.469	0.472	1.485	1.485	2.093	2.083

STANDARD DEVIATIONS

N	MULE THETA	MLE THETA	MULE PHI	MLE PHI	MULE SIGMA	MLE SIGMA	MULE BETA	MLE BETA
25	0.300	0.272	0.173	0.122	0.265	0.244	0.592	0.433
50	0.209	0.179	0.075	0.069	0.175	0.171	0.287	0.282
75	0.169	0.153	0.041	0.040	0.141	0.140	0.210	0.207
100	0.153	0.145	0.040	0.038	0.106	0.106	0.210	0.207
125	0.148	0.143	0.042	0.040	0.089	0.089	0.217	0.214
150	0.104	0.097	0.030	0.029	0.079	0.079	0.194	0.191
175	0.097	0.090	0.035	0.034	0.074	0.074	0.200	0.196
200	0.081	0.074	0.031	0.030	0.052	0.052	0.188	0.186

TABLE 10. SUMMARY OF SAMPLE SIZE TEST FOR THETA = -0.50 AND PHI = -0.50

MEAN VALUES

N	MULE THETA	MLE THETA	MULE PHI	MLE PHI	MULE SIGMA	MLE SIGMA	MULE BETA	MLE BETA
25	-0.648	-0.792	-0.461	-0.411	1.462	1.423	2.082	2.010
50	-0.504	-0.548	-0.528	-0.506	1.494	1.493	2.074	2.045
75	-0.470	-0.510	-0.538	-0.517	1.486	1.485	2.067	2.040
100	-0.512	-0.538	-0.522	-0.509	1.484	1.484	2.048	2.030
125	-0.513	-0.532	-0.513	-0.503	1.486	1.485	2.049	2.037
150	-0.527	-0.537	-0.498	-0.493	1.494	1.494	2.029	2.022
175	-0.521	-0.531	-0.506	-0.502	1.487	1.487	2.035	2.029
200	-0.515	-0.524	-0.497	-0.492	1.495	1.495	2.010	2.003

STANDARD DEVIATIONS

N	MULE THETA	MLE THETA	MULE PHI	MLE PHI	MULE SIGMA	MLE SIGMA	MULE BETA	MLE BETA
25	0.186	0.217	0.166	0.203	0.241	0.236	0.256	0.302
50	0.099	0.108	0.105	0.117	0.130	0.129	0.244	0.255
75	0.114	0.114	0.125	0.131	0.132	0.132	0.170	0.179
100	0.080	0.078	0.115	0.117	0.114	0.114	0.175	0.176
125	0.069	0.061	0.102	0.103	0.106	0.105	0.172	0.173
150	0.064	0.059	0.096	0.095	0.086	0.086	0.160	0.160
175	0.054	0.047	0.096	0.095	0.079	0.079	0.147	0.144
200	0.052	0.046	0.101	0.100	0.058	0.058	0.148	0.147

TABLE 11. SUMMARY OF SAMPLE SIZE TESTS FOR THETA = 0.95 AND PHI = -0.95

MEAN VALUES

N	MULE THETA	MLE THETA	MULE PHI	MLE PHI	MULE SIGMA	MLE SIGMA	MULE BETA	MLE BETA
25	0.147	0.111	-0.126	-0.116	1.431	1.336	1.049	1.041
50	0.173	0.254	-0.213	-0.289	1.512	1.383	1.285	1.431
75	0.351	0.450	-0.358	-0.446	1.505	1.495	1.455	1.568
100	0.392	0.544	-0.381	-0.525	1.531	1.519	1.389	1.554
125	0.454	0.550	-0.461	-0.554	1.489	1.481	1.428	1.522
150	0.400	0.496	-0.392	-0.492	1.474	1.469	1.370	1.436
175	0.440	0.473	-0.445	-0.471	1.485	1.479	1.429	1.436
200	0.490	0.585	-0.483	-0.566	1.497	1.492	1.433	1.487

STANDARD DEVIATIONS

N	MULE THETA	MLE THETA	MULE PHI	MLE PHI	MULE SIGMA	MLE SIGMA	MULE BETA	MLE BETA
25	0.465	0.926	0.339	0.582	0.187	0.143	0.533	0.762
50	0.386	0.808	0.372	0.680	0.184	0.188	0.664	0.938
75	0.223	0.491	0.287	0.497	0.116	0.115	0.501	0.665
100	0.252	0.553	0.244	0.535	0.115	0.116	0.462	0.696
150	0.233	0.476	0.224	0.441	0.112	0.112	0.422	0.417
175	0.225	0.472	0.216	0.432	0.089	0.086	0.328	0.390
200	0.198	0.405	0.190	0.379	0.078	0.078	0.290	0.298

The apparently poor behavior of the estimator in the cases when ϕ is near $-\theta$ can be explained by noting that a cancellation of factors occurs in the noise transfer function. In the ARMA model used by Box and Jenkins, the asymptotic inverse covariance matrix of the parameter estimates is singular for this case (Box and Jenkins, 1974, p. 246). If a forcing function is present, the inverse covariance matrix is no longer singular. However, if the signal to noise ratio is low, the covariance matrix may be ill-conditioned.

These observations can be illustrated by an examination of the inverse covariance matrix for the case p = 1. From Section 4.C, the inverse covariance matrix of the estimated parameter vector $(\hat{\rho}, \hat{\phi}, \hat{\theta})'$ is

$$\begin{bmatrix} (G' \Sigma_T^{-1} G)/\sigma^2 & & \\ (G' \Sigma_T^{-1} A)/\sigma^2 & \sum_{j=0}^{T-2} (T-1-j)\phi^{2j} & \\ & + A' \Sigma_T^{-1} A/\sigma^2 & \\ 0 & - \sum_{j=0}^{T-2} (T-1-j)\phi^j (-\theta)^j & \sum_{j=0}^{T-2} (T-1-j)\theta^{2j} \end{bmatrix}$$

with symmetric terms in the upper triangular portion, and where $A = J_1' \mu_Z - F(1)$. The vector A is related to the information about the AR parameter contributed by the accessible input. If there were no accessible input, then $A \approx 0$. If, in addition, $\phi = -\theta$, then the above matrix is obviously nearly singular.

This suggests that a statistic based on $A' \Sigma_T^{-1} A/\sigma^2$ might be a reasonable measure of the signal to noise ratio. A statistic that could be useful is

$$\xi = Tr(A' \Sigma_T^{-1} A)/\sigma \quad ,$$

where $Tr(\cdot)$ is the trace operator. ξ might be useful in testing whether the accessible input had accounted for most of driving input to the system.

A series of simulations were made to examine the utility of ξ. These simulations were run with $\theta = 0.95$, $\phi = -0.95$, and $\beta = 2.0$. Five runs were made for each of the values 1.0, 0.75, 0.50, 0.25, 0.10, and 0.05 for σ. Some summary results are given in Table 12.

Table 12. Summary of Signal to Noise Ratio Tests

σ	Mean $\hat{\xi}$	Mean $\hat{\theta}$	Mean $\hat{\phi}$
1.00	5.4	0.197	-0.274
0.75	7.0	0.288	-0.277
0.50	6.4	0.678	-0.610
0.25	11.4	0.578	-0.543
0.10	35.2	0.757	-0.812
0.05	79.4	0.823	-0.876

As expected the estimates do improve as the noise variance decreases, and large values of ξ indicate better estimates. Also the correlation between $\hat{\theta}$ and $\hat{\phi}$ tended to be around -0.9. A low value for $\hat{\xi}$ and a high negative correlation between $\hat{\theta}$ and $\hat{\phi}$ should be sufficient grounds to

use the estimates carefully.

Secondly, the behavior of the estimator is only 'apparently poor' in another sense. The measure of goodness that was computed was a measure of distance in the parameter space. However, the real interest is in the behavior of the system in the trajectory space. The widely scattered estimates when $\phi = -\theta$ simply reflect the fact that the system is insensitive to the values of the parameters along the line $\phi = -\theta$. It was noted above that this behavior of the parameters is reflected in the correlation matrix. In an effort to build a realistic model, this knowledge is as useful as more precise parameter estimates would be, for it indicates that the model may be over parameterized. Model order selection procedures are generally based on the behavior of the residual mean square (Jenkins and Watts, 1968, Box and Jenkins, 1971). The correlation matrix and the ξ statistic could be valuable adjuncts to such procedures.

Summary results of the sample size tests are presented in Tables 7 through 11. As expected the standard errors of both the MULE and the MLE decrease with increased sample size, and generally both estimators get closer to the nominal values. The mean values for the samples of size 200 were quite close to the nominal values except for the case $\theta = .95$, $\phi = -.95$. Generally the MULE appeared to have smaller standard errors than the MLE and to have a mean value that was as good or better than the MLE.

C. Example Using Oregon Sheep Supply Data

The data for this example are taken from Brown and Fawcett (1974). Their paper examines the relationship between the number of sheep in Oregon, meat and wool prices, and predator control policy. Two basic models were proposed: a geometrically distributed lag model and a polynomial lag model, both of which are commonly used in econometrics. The geometric lag is structurally similar to the forced ARMA model, but the parameters are usually estimated by ordinary least squares, leading to well documented problems.

The geometric lag model presented by Brown and Fawcett has the form

$$q(t) = \phi q(t-1) + \beta_0 + \beta_1 P(t-1) + \beta_2 X(t-1) \qquad (5.1)$$

where $q_{(t)}$ is the number of stock sheep and lambs in Oregon in year t, P(t) is a price index consisting of a weighted combination of meat and wool prices, and X(t) is a dummy variable used to express the influence of predator control policy. The policy was changed in 1965, so the variable X(t) is set equal to one for the years 1947 through 1964 and zero for the years 1965 through 1972, the last year of the data set.

The basic problem in the analysis of this data set is ancillary to the choice of the correct model and parameter estimation. This problem involves the logical difficulties of drawing an inference from observational data. The inferential techniques that are used for experimental data cannot be applied indiscriminately to observational data. In an experimental situation great care is exercised to either eliminate or

control extraneous factors or to insure that both treated and control groups are homogeneous in those factors. The assurance that treated and control groups differ systematically only in treatments permits the assertion of casuality when a difference is observed.

Comparable control is not available in the collection of observational data. In some instances these difficulties can be countered by careful selection and/or stratification of the sample. Mantel and Haenszel (1959) discuss some of the techniques that can be used to analyze retrospective clinical data.

In the case of the data set that Brown and Fawcett analyze, such techniques cannot be used. The problem is one of trying to infer the effect of a policy change in a non-experimental situation. Only limited data are available, and there is no possibility of obtaining another data set to be used for a control. It is noted that this is the prevailing circumstance in the context of this dissertation. As noted earlier, most applications are properly regarded as unique realizations. However, one does wish to investigate causal inferences in this context.

In proceeding with the analysis of this data set, several general principles should be followed. The first is that the model used to describe the data must provide a good fit to the data. If care is not taken to insure an adequate model, then the supposed significant effect attributed to a policy change may in fact be the partial correction of a bad model. A second principle is that the data should be analyzed independently of prior concepts. Of course model formulation should

be guided by theory, and the nature of the question being investigated has a bearing on the analysis. But the data should point to the conclusion. A third principle that holds for all data analysis is that both null and alternate hypotheses should be clearly stated before any hypothesis testing is done. It is possible to be somewhat lax with this principle when analyzing experimental data, since the experimental design usually restricts both null and alternate hypotheses. This is not the case in the analysis of observational data. Incorrect specification of hypotheses can lead to the wrong analysis and to improper conclusions. As an illustration suppose that one wishes to determine if a policy change made in, say, 1960 had an effect on the behavior of some system. If H_0: 'The behavior of the system did not change' is tested against H_1: 'The behavior of the system changed in 1960' then any test is almost surely biased in favor of H_1 because the hypotheses H_0 and H_1 are not exhaustive. For suppose an event occurred in 1958 that did have an effect on the system behavior. Unless the event is corrected for by the model, a test of H_0 versus H_1 will show significance which might be wrongly attributed to the policy change in 1960.

Additionally the inference of causality based on observational data is at best of uncertain validity. Even if there is compelling evidence of concurrent policy and system changes, the conclusion that the system change resulted from the policy change does not necessarily follow. The randomization element, which allows inference of causality in formal experimentation, is lacking. The association could be a matter of chance coincidence, which the procedures of sound experimentation are designed

to rule out. But if it is not established that the system response change was concurrent with (or possibly subsequent to) a policy change, then the inference of causality is not warranted under any circumstances.

Thus, the problem of detecting a system change resulting from a policy change has two aspects: establishing that a change occurred and establishing that the change occurred in a time frame consistent with the policy change. The second aspect can be examined using a procedure much like stepwise regression. Suppose observations $y(1), y(2),\ldots,$ $y(T)$ are available. Let $f(t)$ represent the system model, so

$$y(t) = f(t) + \eta(t)$$

where $\eta(t)$ is a random disturbance term. Generalize the model to account for a change in system behavior at time j by letting the variable $X_j(t)$ represent the structure

$$X_j(t) = \begin{cases} 1 , & t \leq j \\ 0 , & t > j \end{cases}$$

for $j = 1,2,\ldots,T$. The system model becomes $f(t, X_j(t))$.

The variables $X_1, X_2,\ldots,X_T$ could be entered into the equation sequentially to determine which single variable gives the most improvement in model fit. Say the best single variable is X_k. Relabel X_k as $X_{(1)}$. Repeat the process using the model $f(t, X_{(1)}(t), X_j(t))$ and the variables $X_1, X_2,\ldots,X_{k-1}, X_{k+1},\ldots,X_{T-1}$, and select $X_{(2)}$ as the variable which gives the most improvement in model fit given that $X_{(1)}$ is in the model. Continue until the addition of further variables does not

produce significant model improvement. If one of the variables $X_{(1)}$, $X_{(2)},\ldots,X_{(s)}$ that has entered the model is consistent with a policy change, then the data do present some evidence for an effect of the policy change.

For some models f(t) the analogy to stepwise regression is exact, and commonly available stepwise regression programs could be used. However the F-tests used in stepwise regression are not always appropriate. If parameters are being estimated using maximum likelihood, then significance tests could be constructed using likelihood ratio tests.

The above procedure was not carried out <u>in toto</u> for this example. Instead, several models with an increasing number of X_i's were fit to the data. First order lags and first order MA noise processes were used for all models.

At first a model with no policy variable present was tried. The model equation is

$$q(t) = \phi q(t-1) + \beta_0 + \beta_1 P(t-1) + a_{(t)} + \theta a_{(t-1)} \cdot \qquad (5.2)$$

The parameter estimates and their standard deviations are given in Table 13.

Table 13. Model 1 Parameter Estimates

Parameter	Estimate	Standard Deviation
ϕ	1.112	0.014
β_0	-216.680	37.012
β_1	3.713	1.038
θ	.482	0.180
σ	26.054	

The data and the fitted model are plotted in Figure 13. Figure 14 is a plot of the standardized residuals versus time. As can be seen in Figure 13, the model does not track the data. The residuals exhibit a definite pattern. It can safely be concluded that the model is not adequate.

The primary interest of the modelling effort is in the system structure. This interest is reflected in the choice of objective function and in the way fitted values used to compare with observations are computed. From (3.1) the objective function is

$$(Y-G\rho)' \Sigma_T^{-1} (Y-G\rho).$$

Following the standard procedure of using estimated parameter values in model equations to obtain fitted values, there results

$$\hat{Y} = \hat{G}\,\hat{\rho}$$

and from (2.9)

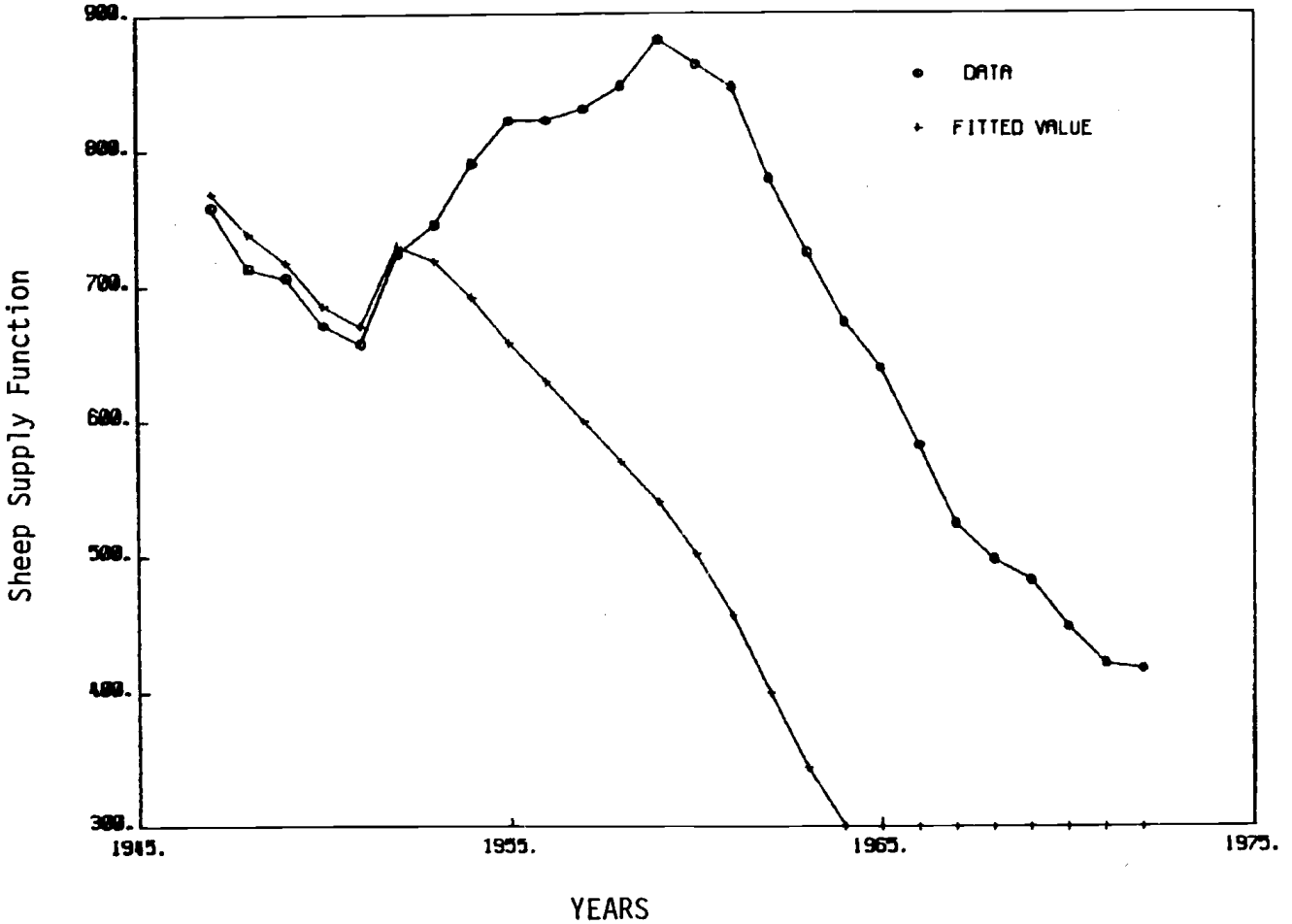


Figure 13. Model 1 Sheep Supply Function

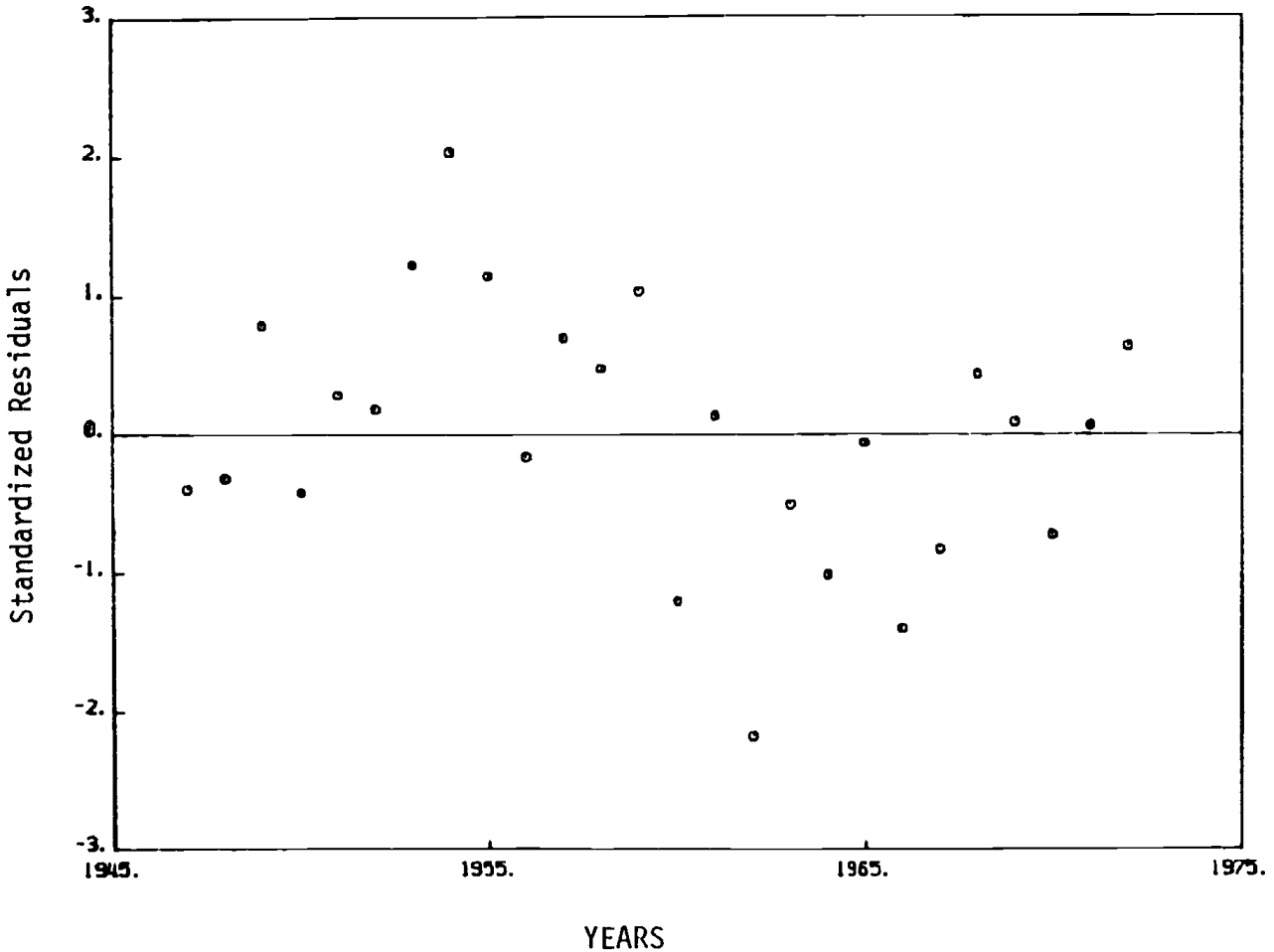


Figure 14. Model 1 Standardized Residuals

$$\hat{Z} = \hat{M}^{-1} \hat{G} \hat{\rho} \quad . \tag{5.3}$$

None of the factors on the right hand side of (5.3) depend on the observed values of Z except through the parameter estimates. This is in contrast to the method commonly used to compute $\hat{Z}$ for lagged models:

$$\hat{Z}_k = \hat{\phi} Z_{k-1} + \ldots \quad . \tag{5.4}$$

The differences in the two means of calculating $\hat{Z}$ can be striking. The second method, represented by (5.4), can present an overly optimistic view of how well the model fits the data. The same parameter values used in Figure 13 were also used in Figure 15. The fitted values were computed using (5.4) instead of (5.3). The fit appears to be much better.

Adding the dummy variable associated with the policy change being tested does improve the model fit. The model equation becomes

$$q(t) = \phi q(t-1) + \beta_0 + \beta_1 P(t-1) + \beta_2 X_{19}(t) + a(t) + \theta a(t-1) \tag{5.5}$$

The parameter estimates and their standard deviations are in Table 14.

Table 14. Model 2 Parameter Estimates

Parameter	Estimate	Standard Deviation
ϕ	0.986	0.059
β_0	-148.612	55.519
β_1	41.985	18.453
β_2	3.407	1.030
θ	0.521	0.178
σ	25.113	

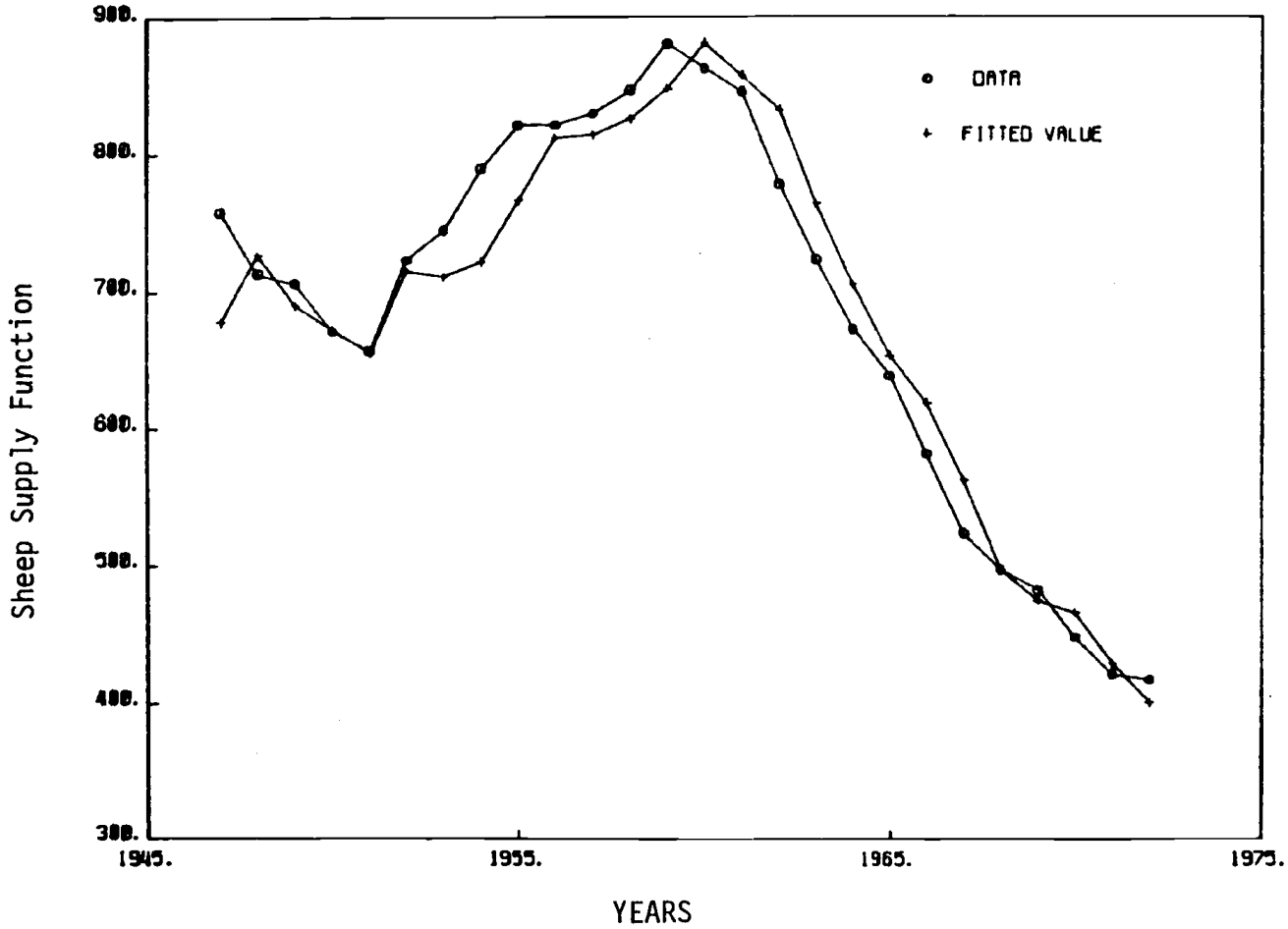


Figure 15. Model 1 Sheep Supply Function Using $\hat{Z}(k) = \hat{\phi}Z(k-1)$

The model and standardized residuals are plotted in Figures 16 and 17, respectively. There is still substantial evidence of lack of fit and the residuals still have a definite pattern. However it is clear that the fit improved.

Although no exact test is available, a good indication of the significance of β_2, and, but for inferential problems, of the effects of the policy variable, can be obtained through the use of the log likelihood ratio. The above estimates are MULE; however, the MULE should have all of the asymptotic properties of MLE. Thus

$$\lambda = -2 \log \left[\frac{\underset{\beta_2 = 0}{\text{Max}} \quad \text{ULF}}{\text{Max} \quad \text{ULF}} \right] \tag{5.6}$$

should have an approximate $\chi^2(1)$ distribution. Applying the test gives $\lambda = 3.24$, which is almost significant at the 5% level.

If the analysis were concluded at this point, then one might be tempted to infer that the policy change had an effect on the system. However, such a conclusion is premature because concurrency has not been established. Specifically, one has not rejected the

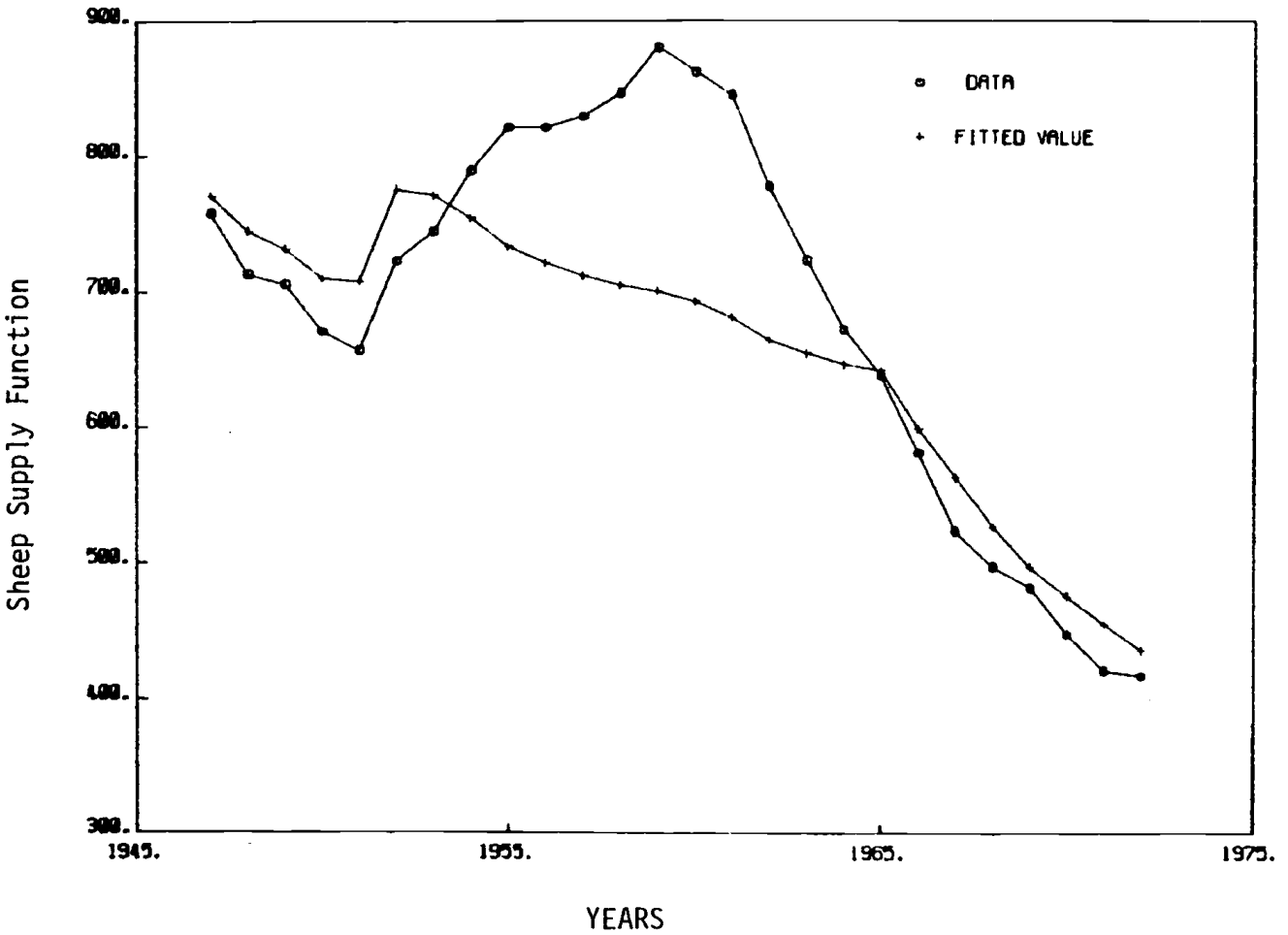


Figure 16. Model 2 Sheep Supply Function

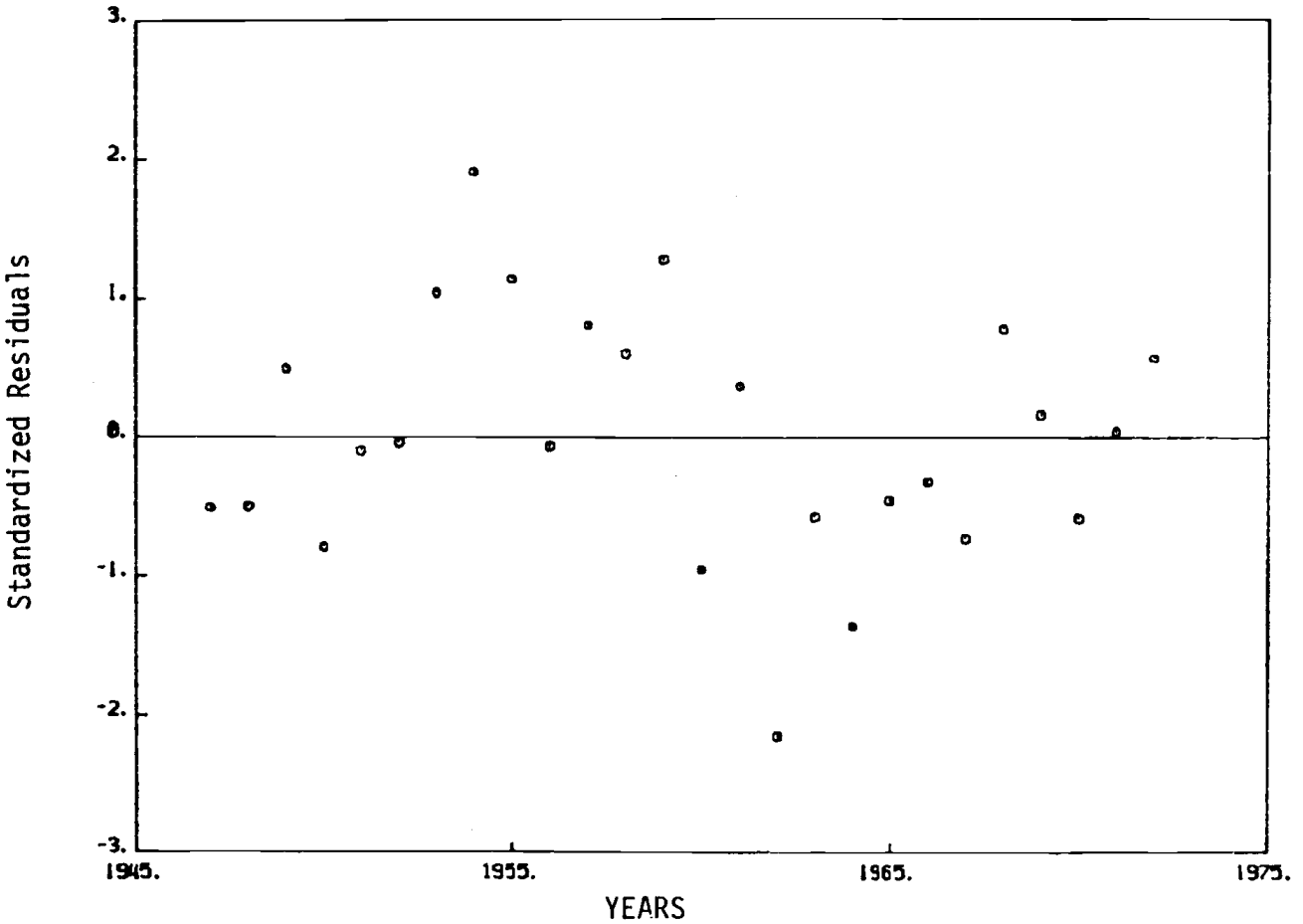


Figure 17. Model 2 Standardized Residuals

associated alternate hypotheses that the system behavior changed in 1964, or in 1966, or in any of a number of other possible years. It is clear that a test of the null hypothesis that a change occurred in any of these years would lead to a similar indication of change. Before one can accept the specific alternative of interest, these other alternatives must be examined, and either rejected or incorporated into the model.

The sheep supply function has three apparently distinct phases (Figure 13). In order to duplicate this behavior, a model with three levels instead of two was used. In essence this constitutes a subjective "guess" of the best three X_k's to enhance the model. This model (Model 3) had the form

$$q(t) = \phi q(t-1) + \beta_1 X_5(t) + \beta_2 X_{13}(t) + \beta_3 X_{26}(t) + \beta_4 P(t-1) + a(t) + \theta a(t-1) \tag{5.7}$$

where X_5 was equal to one from 1947 to 1951 and zero otherwise, X_{13} was equal to one from 1952 to 1959 and zero otherwise, and X_{26} was zero prior to 1960 and equal to one afterwards. The estimated parameters and their standard deviations are in Table 15.

Table 15. Model 3 Parameter Estimates

Parameter	Estimate	Standard Deviation
ϕ	0.989	0.038
β_1	-79.063	46.371
β_2	-33.971	51.552
β_3	-98.760	46.990
β_4	2.040	0.840
θ	0.215	0.202
σ	17.375	

The model and standardized residuals are plotted in Figure 17 and 18, respectively. The model seems to track the data quite well and the residuals do not exhibit any obvious patterns.

At this point the policy variable X_{19} was brought back into the equation to assess the effect of predator control policy. The model is

$$q(t) = \phi q(t-1) + \beta_1 X_1(t) + \beta_2 X_6(t) + \beta_3 X_{14}(t) + \beta_4 X_{19}(t) + \beta_5 P(t-1) + a(t) + \theta a(t-1) \quad (5.7)$$

The parameter values and their standard deviations are given in Table 16. The model and standardized residuals are plotted in Figure 20 and 21, respectively. From (5.6) $\lambda = 2.57$, which is not significant. Thus, this data set does not appear to present substantial evidence that the change in predator control policy effected the number of sheep in Oregon.

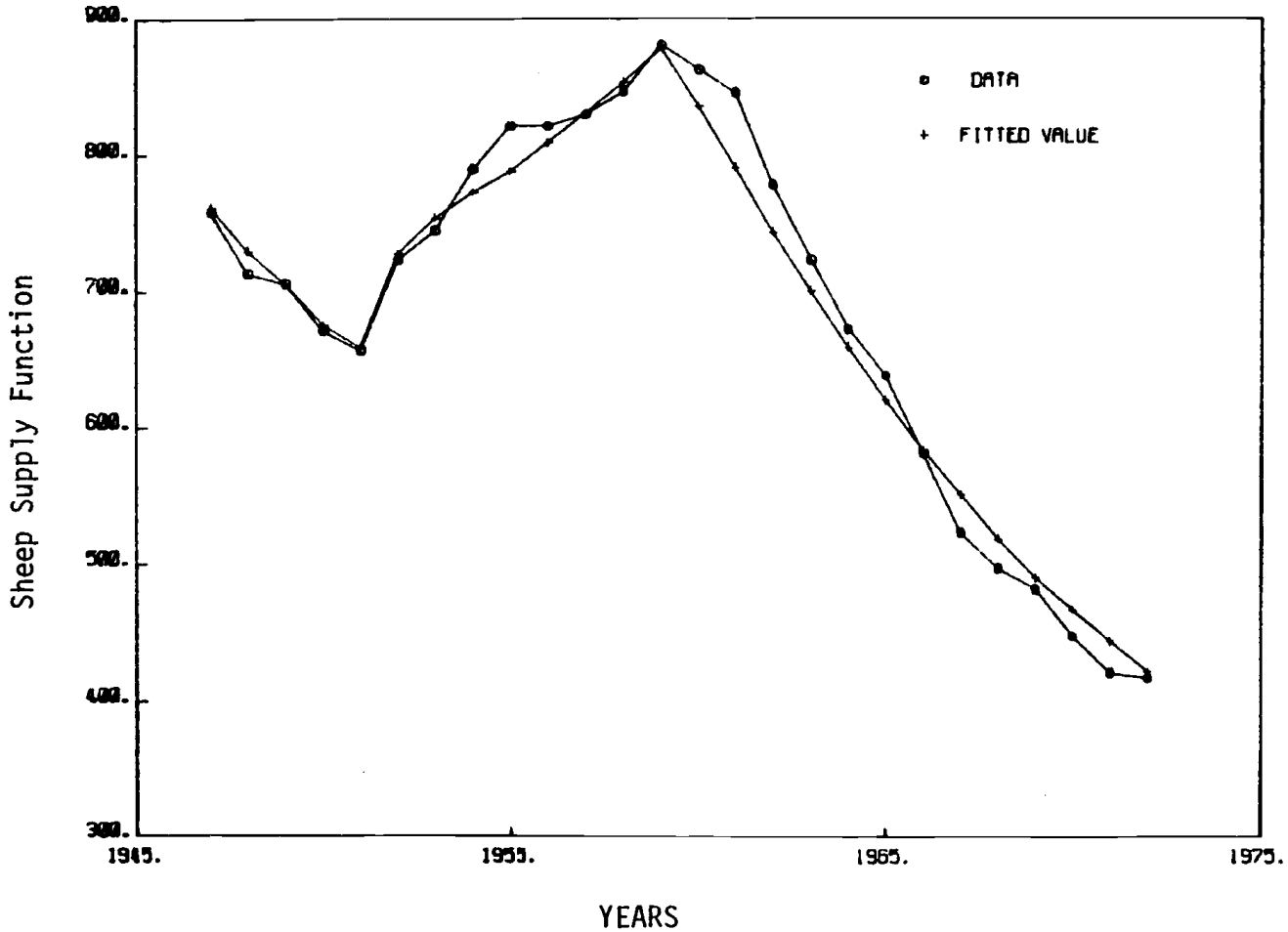


Figure 18. Model 3 Sheep Supply Function

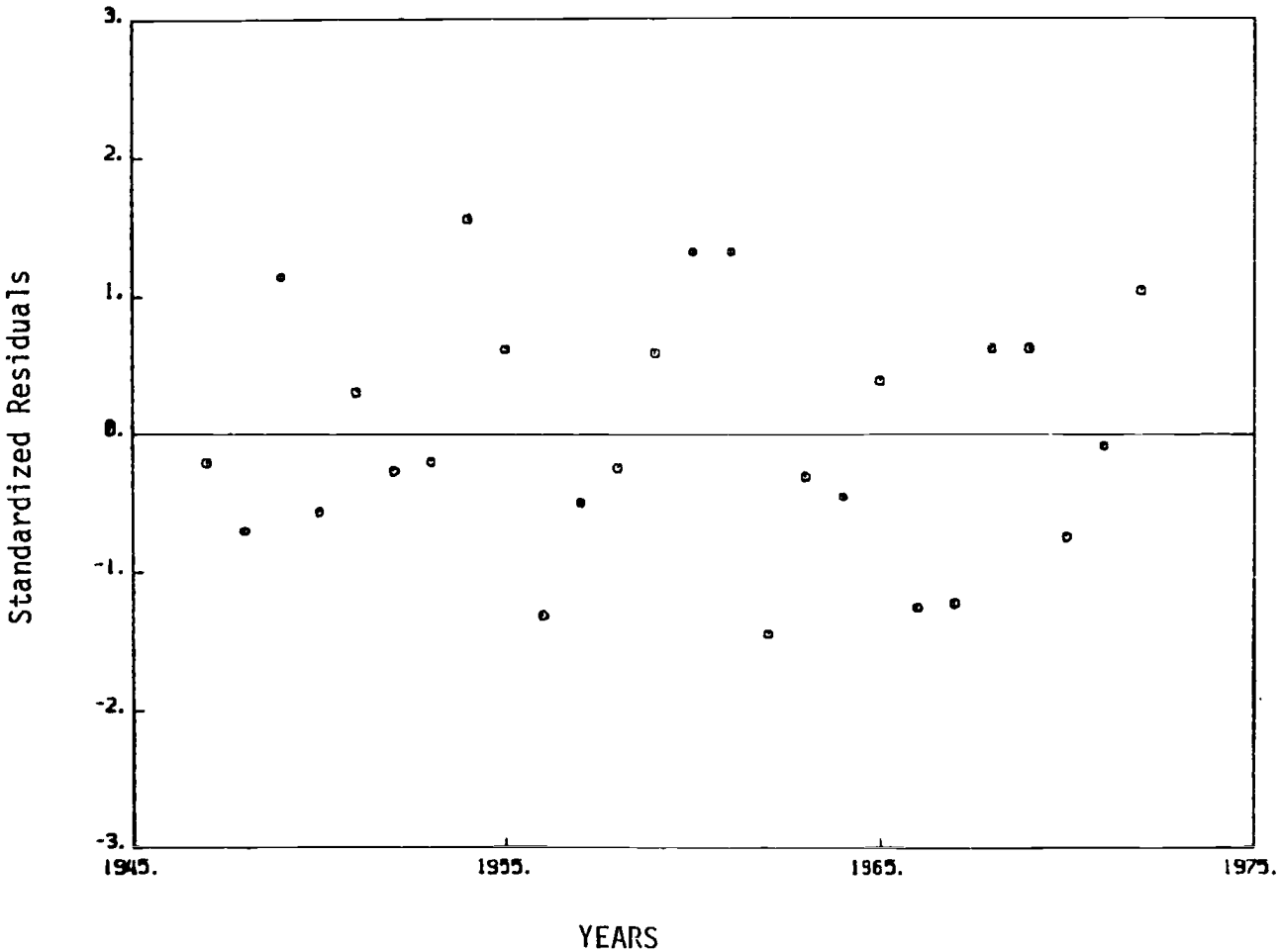


Figure 19. Model 3 Standardized Residuals

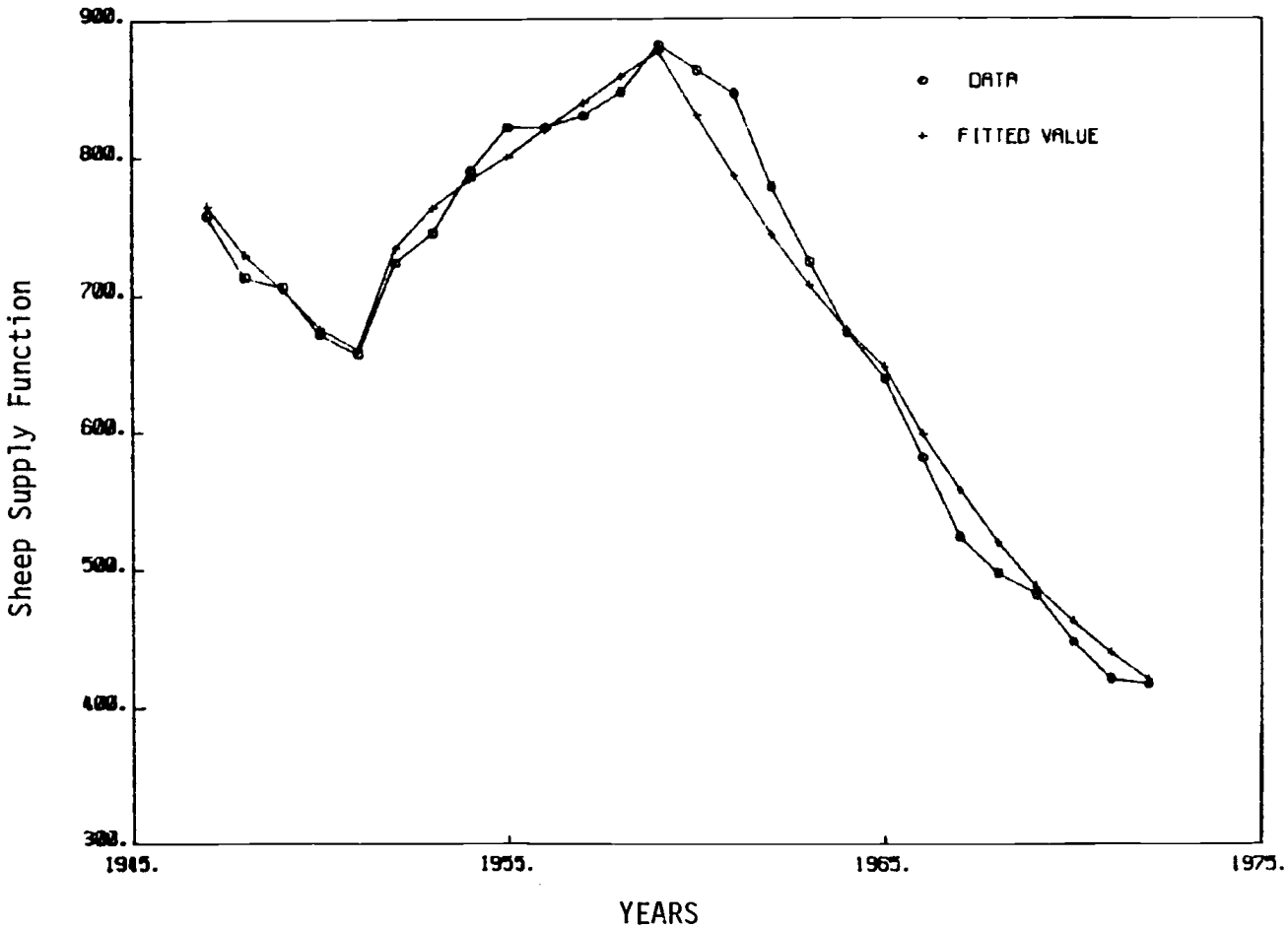


Figure 20. Model 4 Sheep Supply Function

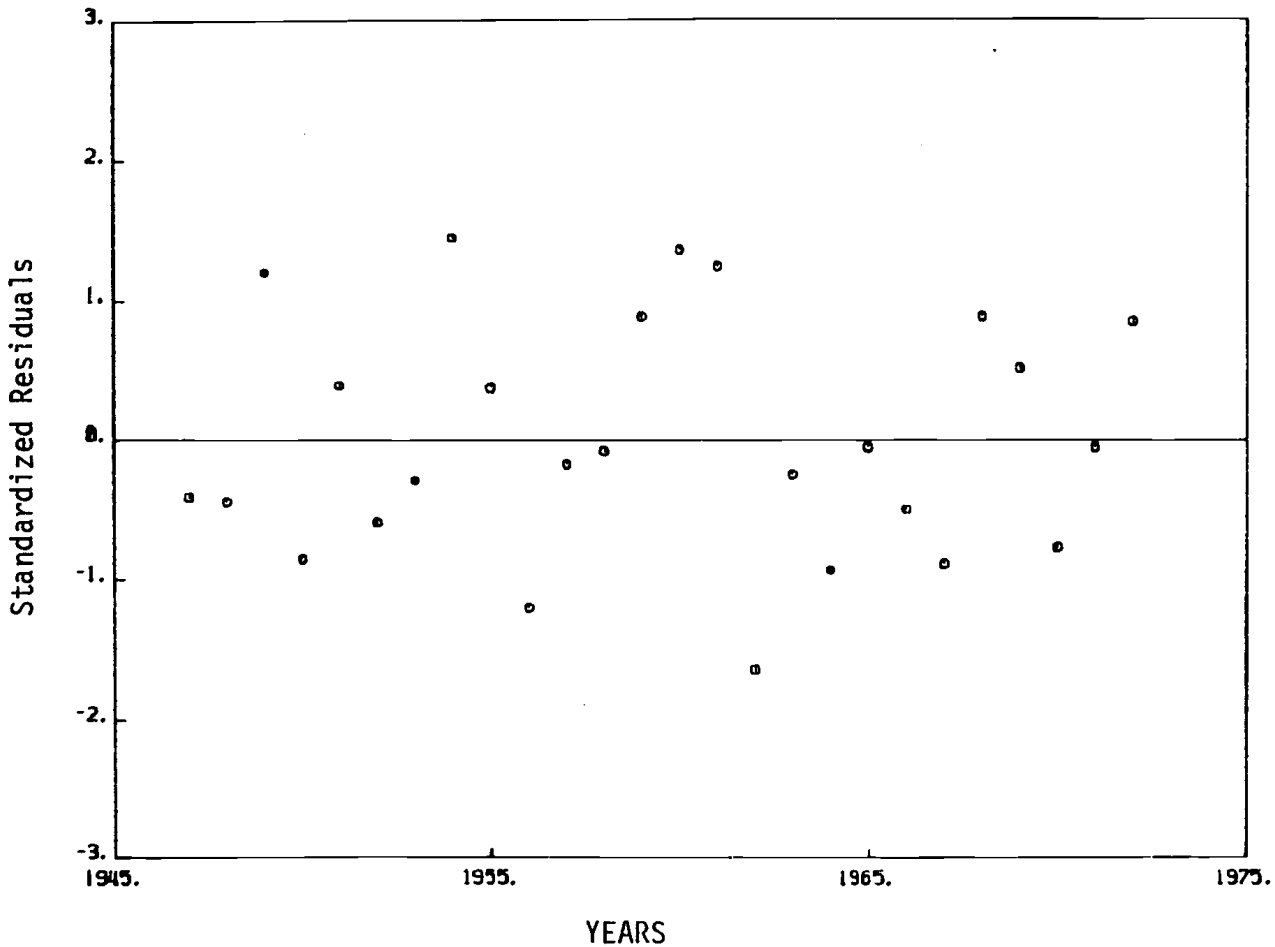


Figure 21. Model 4 Standardized Residuals

Table 16. Model 4 Parameter Estimates

Parameter	Estimate	Standard Deviation
ϕ	0.912	0.061
β_1	-41.778	54.231
β_2	9.475	61.329
β_3	-51.751	56.932
β_4	25.830	16.861
β_5	1.775	0.874
θ	0.388	0.193
σ	16.991	

The Oregon sheep supply data set was used as an example for several reasons. First, the sheep supply, price, and predator relationship is a subsystem that can reasonably be approximated by a linear system model even though the macro socio-economic system is almost certainly nonlinear. Also, the relationships within the sybsystem are not static: they depend on the interface to the macro system. Hence the constant coefficient linear system approximation is adequate for only a limited length of time. Analysis methods based on long-term stationary behavior are not conceptually suitable for this data set.

Secondly, our perception of the data set is as a unique realization. There is no realistic possibility of repeating the circumstances that lead to the generation of the data set in order, say, to determine if the results would be different if a policy change were not made. Furthermore, temporal extension of the data set would not be likely to

add information pertinent to the question of the effect of policy change. Any insight into the results of the change of predator control policy must be gleaned from the data presently in hand.

Thirdly, the data set represents a situation where there is probably an additional input that has not been identified. The residuals from Model 1 and Model 2 (Figures 14 and 17) clearly indicate a change in system behavior. This change was adjusted for and identified by the use of dummy variables, but further research is indicated to determine economic factors associated with these dummy variables.

Finally, the example illustrates the relevancy of the techniques developed in this dissertation. There is a need to learn from dynamic systems whose very existance may be ephemeral, and hence a need to learn from small sets of data collected on dynamic systems.

VI. MULTIPLE OUTPUT SYSTEMS

The estimation theory and the computational algorithm extend without difficulty to multiple output systems. The computational algorithm has additional steps because cross-covariance terms now have to be considered.

In Chapter II shift operators $\Phi_p(F)$ were defined as p^{th}-order polynomials in F with scalar coefficients. For multiple output systems the definition needs to be extended by allowing the coefficients to be matrices:

$$\Phi_p^{ts}(F) = \phi_1^{ts} F^{p-1} + \phi_2^{ts} F^{p-2} + \ldots + \phi_p^{ts}$$

where ϕ_i^{ts} is a txs matrix with elements $\phi_{k_j}(i)$. The multidimensional analog to (2.4) is then

$$F^p Z(k) = \Phi_p^{tt}(F)Z(k) + B_r^{ts}(F)U(k) + \Theta_q^{ts}(F)A(k),$$

where

$$Z(k) = (z_1(k), z_2(k), \ldots, z_t(k))'$$
$$U(k) = (u_1(k), u_2(k), \ldots, u_s(k))'$$
$$A(k) = (a_1(k), a_2(k), \ldots, a_m(k))' .$$

The number of parameters increases rapidly as the number of outputs increases. In order to minimize notational burden, the estimation theory and computational algorithm will be developed for a two output

system with two correlated inaccessible inputs and a single accessible input. This representation is sufficiently complex to serve as an illustration of the general algorithm.

The system model will be the pair of linear difference equations

$$z_1(k) = \phi_{11}(1)z_1(k-1) + \phi_{11}(2)z_1(k-2) + \phi_{12}(1)z_2(k-1) + \phi_{12}(2)z_2(k-2) + \beta_1 u(k-1) + a_1(k-1) + \theta_1 a_1(k-2)$$

$$z_2(k) = \phi_{21}(1)z_1(k-1) + \phi_{21}(2)z_1(k-2) + \phi_{22}(1)z_2(k-1) + \phi_{22}(2)z_2(k-2) + \beta_2 u(k-1) + a_2(k-1) + \theta_2 a_2(k-2).$$

Additional accessible inputs and lagged terms in both the accessible and inaccessible inputs can be added without effect on the algorithm.

The inaccessible input is assumed to be a zero mean Gaussian sequence with the following properties:

$$E[a_i(k)a_j(\ell)] = \begin{cases} 0 & k \neq \ell \\ \gamma_i^2, & i = j,\ k = \ell \\ \gamma_{12}, & i \neq j,\ k = \ell \end{cases} . \qquad (6.1)$$

The parameters γ_1^2, γ_2^2, and γ_{12} will be termed 'variance parameters' and γ will denote the matrix

$$\begin{bmatrix} \gamma_1^2 & \gamma_{12} \\ \gamma_{12} & \gamma_2^2 \end{bmatrix} .$$

The initial state of the system is again parameterized so that its appearance is gradually shifted out of the model. Thus, let

$$W = (w_{11}\ w_{12}\ w_{21}\ w_{22})'$$

be the initial state vector, i.e., state at time 0. Then

$$\begin{aligned} z_j(1) &= \phi_{j1}(1)w_{11} + \phi_{j1}(2)w_{12} + \phi_{j2}(1)w_{21} + \phi_{j2}(2)w_{22} \\ &\quad + \beta_j u(0) + a_j(0), \\ z_j(2) &= \phi_{j1}(1)z_1(1) + \phi_{j1}(2)w_{11} + \phi_{j2}(1)z_2(1) + \phi_{j2}(2)w_{21} \\ &\quad + \beta_j u(1) + a_j(1) + \theta_j a_j(0), \\ &\vdots \\ z_j(k) &= \phi_{j1}(1)z_1(k-1) + \phi_{j1}(2)z_1(k-2) + \phi_{j2}(1)z_2(k) + \phi_{j2}(2)z_2(k-1) \\ &\quad + \beta_j u_1(k-1) + a_j(k-1) + \theta_j a_j(k-2), \end{aligned}$$

for $k \geq 3$, and for $j = 1,2$.

As in the single output case, the model will be transformed as follows:

$$
\begin{aligned}
y_j(1) &= z_j(1) \\
y_j(2) &= z_j(2) - \phi_{j1}(1)z_1(1) - \phi_{j2}(1)z_2(1) \\
&\vdots \\
y_j(k) &= z_j(k) - \phi_{j1}(1)z_1(k-1) - \phi_{j1}(2)z_1(k-2) - \phi_{j2}(1)z_2(k-1) - \\
&\qquad \phi_{j2}(2)z_2(k-2)
\end{aligned} \tag{6.2}
$$

for $k \geq 3$, and for $j = 1,2$.

Letting $Y = (y_1(1)y_1(2) \ldots y_1(T)y_2(1)y_2(2) \ldots y_2(T))'$

$Z = (z_1(1)z_1(2) \ldots z_1(T)z_2(1)z_2(2) \ldots z_2(T))'$

and

$$
M = \left[\begin{array}{ccccc:ccccc}
1 & & & & 0 & 0 & & & & 0 \\
-\phi_{11}(1) & 1 & & & & -\phi_{12}(1) & 0 & & & \\
-\phi_{11}(2) & -\phi_{11}(1) & 1 & & & -\phi_{12}(2) & -\phi_{12}(1) & 0 & & \\
 & \ddots & \ddots & \ddots & & & \ddots & \ddots & \ddots & \\
0 & & -\phi_{11}(2) & -\phi_{11}(1) & 1 & 0 & & -\phi_{12}(2) & -\phi_{12}(1) & 0 \\
\hdashline
0 & & & & 0 & 1 & & & & 0 \\
-\phi_{21}(1) & 0 & & & & -\phi_{22}(1) & 1 & & & \\
-\phi_{21}(2) & -\phi_{21}(1) & 0 & & & -\phi_{22}(2) & -\phi_{22}(1) & 1 & & \\
 & \ddots & \ddots & \ddots & & & \ddots & \ddots & \ddots & \\
0 & & -\phi_{21}(2) & -\phi_{21}(1) & 0 & 0 & & -\phi_{22}(2) & -\phi_{22}(1) & 1
\end{array}\right] \tag{6.3}
$$

then $Y = MZ$.

Theorem 6.1: The 2Tx2T matrix M as defined by (6.3) has unit determinant.

Proof: Denote by $|M_T|$ the determinant of the 2Tx2T matrix. Expand $|M_T|$ by minors of the first row. This row has a 1 in position (1,1) and zero elsewhere, so $|M_T| = |M_T^*(1,1)|$, where $M_T^*(1,1)$ is the matrix that results from deleting the first row and first column of M_T. The T^{th} row of $M_T^*(1,1)$ is all zeros except for a 1 in the (T,T) position. Expand $|M_T^*(1,1)|$ by minors of the T^{th} row to get

$$|M_T| = |M_T^*(1,1)| = |M_{T-1}| ,$$

since the matrix that results from deleting the T^{th} row and T^{th} column of $M_T^*(1,1)$ is just M_{T-1}. The above sequence can be repeated to obtain

$$|M_T| = |M_{T-1}| = \ldots |M_1| = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1 .$$

It is convenient to re-order the Y vector at this point. Let

$$Y^* = (y_1(1)y_2(1) \quad y_1(2)y_2(2) \quad \ldots \quad y_1(T)y_2(T))' .$$

The expected value of Y* is

$$E[Y^*] = \left[\begin{array}{cccc|cc} \phi_{11}(1) & \phi_{12}(1) & \phi_{11}(2) & \phi_{12}(2) & u(0) & 0 \\ \phi_{21}(1) & \phi_{22}(1) & \phi_{21}(1) & \phi_{22}(2) & 0 & u(0) \\ \phi_{11}(2) & \phi_{12}(2) & 0 & 0 & u(1) & 0 \\ \phi_{21}(2) & \phi_{22}(2) & 0 & 0 & 0 & u(1) \\ \hline & & & & \vdots & \vdots \\ & & 0 & & u(T-1) & 0 \\ & & & & 0 & u(T-1) \end{array}\right] \left[\begin{array}{c} w_{11} \\ w_{21} \\ w_{12} \\ w_{22} \\ \beta_1 \\ \beta_2 \end{array}\right]$$

$$= G^*\rho .$$

Let Σ^*_{2T} be the covariance matrix of Y^*. Since Y^* has a multivariate normal distribution, the log likelihood is, except for additive constants,

$$\ell(\phi, \theta, \rho, \gamma|Z) = - \log|\Sigma^*_{2T}| - (Y^*-G^*\rho)'\Sigma^{*-1}_{2T}(Y^*-G^*\rho) \quad (6.4)$$

and the maximum likelihood estimate of ρ is given by

$$\hat{\rho}(\phi, \theta, \gamma) = (G^{*\prime}\Sigma^{*-1}_{2T}G^*)^{-1} G^{*\prime}\Sigma^{*-1}_{2T}Y^* \quad . \quad (6.5)$$

If this result is substituted for ρ in equation (6.4), the resulting equation is

$$\ell = \log|\Sigma^*_{2T}| - Y^{*\prime}(\Sigma^{*-1}_{2T} - \Sigma^{*-1}_{2T}G^*(G^{*\prime}\Sigma^{*-1}_{2T}G^*)^{-1} G^{*\prime}\Sigma^{*-1}_{2T})Y^*$$

Theorem 3.1 applies, so the matrix

$$\psi^* = \Sigma^{*-1}_{2T} - \Sigma^{*-1}_{2T}G^*(G^{*\prime}\Sigma^{*-1}_{2T}G^*)^{-1}G^{*\prime}\Sigma^{*-1}_{2T}$$

is independent of $\Phi = (\phi_{11}(1), \phi_{11}(2), \ldots, \phi_{22}(2))'$. If the rows of G* are reordered to correspond to Y, and if Σ_{2T} is the covariance matrix of Y, then the same result, with the *'s omitted, holds. That is, the matrix

$$\psi = \Sigma_{2T}^{-1} - \Sigma_{2T}^{-1} G(G'\Sigma_{2T}^{-1}G)^{-1}G'\Sigma_{2T}^{-1}$$

is independent of Φ.

Then analogous to (3.10) the estimate of ϕ is

$$\hat{\phi}(\theta,\gamma) = (H'\psi H)^{-1} H'\psi Z \tag{6.6}$$

where

$$H = \begin{bmatrix} H_T & 0 \\ 0 & H_T \end{bmatrix}$$

and

$$H_T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ Z_1(1) & 0 & Z_2(1) & 0 \\ Z_1(2) & Z_1(1) & Z_2(2) & Z_2(1) \\ \vdots & \vdots & \vdots & \vdots \\ Z_1(T-1) & Z_1(T-2) & Z_2(T-1) & Z_2(T-2) \end{bmatrix}$$

Let $A = (a_1(1)\ a_1(2)\ \ldots\ a_1(T)\ a_2(1)\ \ldots\ a_2(T))'$. It follows from equation (6.1) that

$$\mathrm{Cov}(A) = \Gamma_A = \begin{bmatrix} \gamma_1^2 I_T & \gamma_{12} I_T \\ \gamma_{21} I_T & \gamma_2^2 I_T \end{bmatrix} .$$

Equations (6.2) imply that

$$\Sigma_{2T} = R' \Gamma_A R$$

where

$$R = \left[\begin{array}{cccc:cccc} 1 & \theta_1 & & & & & & \\ & 1 & \theta_1 & & & 0 & & \\ & & & \theta_1 & & & & \\ & & & 1 & & & & \\ \hdashline & & & & 1 & \theta_2 & & \\ & & & & & 1 & \theta_2 & \\ & & 0 & & & & & \theta_2 \\ & & & & & & & 1 \end{array}\right] = \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} .$$

Analogous to the results for the single output case, we have

$$R^{-1} = \begin{bmatrix} R_1^{-1} & 0 \\ 0 & R_2^{-1} \end{bmatrix} = Q$$

with $$R_i^{-1} = Q_i = \sum_{j=0}^{T-1} \pi_j J_j, \quad i = 1,2 .$$

Moreover,

$$\Gamma_A^{-1} = \frac{1}{\gamma_1^2 \gamma_2^2 - \gamma_{12}^2} \begin{bmatrix} \gamma_2^2 I & -\gamma_{12} I \\ -\gamma_{12} I & \gamma_1^2 I \end{bmatrix}$$

so that the inverse of Σ_{2T} has an explicit form:

$$\Sigma_{2T}^{-1} = Q \Gamma_A^{-1} Q' .$$

Also, since $|Q| = 1$, it follows that

$$|\Sigma_{2T}| = |\Gamma_A| = |\gamma|^T = (\gamma_1^2 \gamma_2^2 - \gamma_{12})^T .$$

The matrix Σ_{2T}^* can be written as

$$\Sigma_{2T}^* = R^{*\prime} \Gamma^* R^* ,$$

where

$$\Gamma^* = \begin{pmatrix} \gamma & & & \\ & \gamma & & \\ & & \ddots & \\ & & & \gamma \end{pmatrix}$$

and

$$R^* = \begin{pmatrix} 1 & 0 & \theta_1 & & & & & \\ & 1 & 0 & \theta_2 & & & & \\ & & 1 & 0 & \theta_1 & & & \\ & & & 1 & 0 & \theta_2 & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & 0 & \theta_1 & \\ & & & & & 1 & 0 & \theta_2 \\ & & & & & & 1 & 0 \\ & & & & & & 0 & 1 \end{pmatrix}$$

Let

$$(Y^* - G\rho)R^* = (d_1\ d_2\ \ldots\ d_{2T})'$$

and let

$$D_j = (d_{2j-1},\ d_{2j})' \quad , \text{ for } j = 1,2,\ldots,T \quad ,$$

so that

$$(Y^* - G^*\rho)R^* = (D_1'\ D_2'\ \ldots\ D_T')' \quad .$$

Then the log likelihood function can be written

$$\ell(\phi,\ \theta,\ \rho,\ \gamma|z) = -\log\ |\gamma)^T - \sum_j D_j'\ \gamma^{-1}\ D_j \quad .$$

Let

$$\Delta_\gamma = \begin{pmatrix} \gamma_1^2 & 0 \\ 0 & \gamma_2^2 \end{pmatrix}$$

and

$$\Delta_{D_{ij}} = \begin{pmatrix} d_{2j-1}^2 & 0 \\ 0 & d_{2j}^2 \end{pmatrix}$$

Then

$$\frac{\partial \ell}{\partial \gamma^{-1}} = T(2\gamma - \Delta\gamma) - \sum_{j=1}^{T} (2D_j D_j' - \Delta_{D_{jj}}) \ .$$

It follows that

$$\hat{\gamma}(\phi, \theta, \rho) = \sum_{j=1}^{T} D_j D_j'/T \qquad (6.7)$$

is the MLE of γ.

Equations (6.5), (6.6), and (6.7) can be used to form an iteration to compute ULF(θ). Given starting values for γ, (6.6) can be solved for ϕ, then (6.5) for ρ, and then an updated value of γ can be obtained from (6.7).

It can be shown using Theorem 3.1 that if $\theta_1 = \theta_2$, then the estimates of β in (6.5) and ϕ in (6.6) are independent of γ. Also, D_j does not depend on W for $j \geq 3$. Hence the iteration can be started by tak-

ing $\theta_1 = \theta_2$, computing $\hat{\beta}$ and $\hat{\phi}$, and then using (6.7) to compute $\hat{\gamma}$, beginning the sum with $j = 3$. Theorem 4.1 applies, and since the likelihood is convex in ρ, ϕ, and γ, convergence to the global maximum is assured.

Computation of the MULE could very quickly become prohibitively expensive because of the required multidimensional numerical integration. The single dimensional ULF seemed to be quite smooth except near the boundary of the parameter space. A numerical quadrature using a fine grid near the edges and a coarse grid in the center may make the computation of the MULE feasible.

BIBLIOGRAPHY

Acton, F. S., 1970, Numerical Methods that Work, Harper & Row, New York.

Barnard, G. A., 1959, 'Control Charts and Stochastic Processes', Journal of the Royal Statistical Society B, 21:2, 239-257.

Barnard, G. A., G. M. Jenkins, and C. B. Winston, 1962, 'Likelihood Inference and Time Series', Journal of the Royal Statistical Society A, 125, 321-352.

Bellman, R., H. Kagiwanda, and R. Kalaba, 1965, 'Identification of Linear Systems Via Numerical Inversion of Laplace Transforms', IEEE Trans. Auto. Cont., AC-10: 111-112.

Berman, M., et al., 1962a, 'The Routine Fitting of Kinetic Data to Models', Biophysics J., 2: 275-287.

Berman, M., et al., 1962b, 'Some Formal Approaches to the Analysis of Kinetic Data in Terms of Linear Compartmental Systems', Biophysics J., 2: 289-316.

Birnbaum, A., 1962, 'On the Foundations of Statistical Inference', Journal of the American Statistical Association, 57, 269-306.

Box, G. E. P., and G. M. Jenkins, 1970, Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, California.

Brown, W. G., and D. Fawcett, 1974, 'Estimated Economic Losses by Oregon Sheep Growers Associated With Restricted Predator Control, 1965-1972: Some Preliminary Findings', Spec. Rept. 418, Agr. Expt. Sta., Oregon State University, Corvallis, Oregon.

Chan, S., S. Chan, and S. Chan, 1972, Analysis of Linear Networks and Systems, Addison-Wesley, Reading, Mass.

De Russo, P. M., R. J. Roy, and C. M. Close, 1965, State Variables for Engineers, John Wiley and Sons, New York.

Desai, R. C., and R. Oldenburger, 1969, 'Identification of Impulse Response From Normal Operating Data Using the Delay Line Synthesizer Principle', IEEE Trans. Auto. Cont., AC-14, 580-582.

Dhrymes, P. J., R. Berner, and D. Cummins, 1974, 'A Comparison of Some Limited Information Estimators for Dynamic Simultaneous Equations Models With Autocorrelated Errors', Econometrica, 42:2, 311-332.

Dudley, D. G., 1977, 'Fitting Noisy Data With a Complex Exponential Series', Lawrence Livermore Laboratories, UCRL-52242.

Durbin, J., 1959 , 'Efficient Estimation of Parameters in Moving-Average Models', Biometrika 46, 306-316.

Durbin, J., 1960a , 'The Fitting of Time Series Models', Rev. Int. Inst. Stat., 28:3, 233-244.

Durbin, J., 1960b , 'Estimation of Parameters in Time-Series Regression Models', JRSS(B), 22:1, 139-153.

Fair, Ray C., 1970, 'The Estimation of Simultaneous Equation Models With Lagged Endogenous Variables and First Order Serially Correlated Errors', Econometrica, 38:3, 507-516.

Freeman, H., 1965, Discrete Time Systems, John Wiley & Sons, Inc., New York.

Griliches, Z., 1967, 'Distributed Lags: A Survey', Econometrica 35, 16-49.

Hammersley, J. M., and D. C. Handscomb, 1964, Monte Carlo Methods, Methuen, London.

Hsia, T. C., 1976, 'On Least Squares Algorithms for System Parameter Identification', IEEE Trans. Auto. Cont., AC-21, 104-108.

Jacquez, John A., 1972, Compartmental Analysis in Biology and Medicine, Elsevier Publishing Co., New York.

Jenkins, G. M., and D. G. Watts, 1968. Spectral Analysis and its Applications, Holden-Day, San Francisco, California.

Johnston, J. 1972, Econometric Methods, 2nd Ed. McGraw Hill, New York.

Kashyap, R. L., 1970, 'Maximum Likelihood Identification of Stochastic Linear Systems', IEEE Transactions on Automatic Control, AC-15:1, 25-34.

Levenstein, H., 1960, 'Use Difference Equations to Calculate Frequency Response From Transient Response', Control Engineering, 4:4, 90-95.

McMichael, F. C., and Hunter, J. S., 1972, 'Stochastic Modeling of Temperature and Flow in Rivers', Water Resources Research 8:1, 87-98.

Mantel, N., and W. Haenszel, 1959, 'Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease', J. of Nat. Can. Ins., 22:4, 719-748.

Meditch, J. S., 1969, Stochastic Optimal Linear Estimation and Control, McGraw-Hill, New York.

Nicholsen, P. R., J. M. Thomas, and C. R. Watson, 1978, 'Characterization of PDP-11 Pseudo-Random Number Generators', in Proceedings of the Digital Equipment Computer Users Society, 5:2, Digital Equipment Corporation, Maynard, Mass.

Oberhofer, W., and J. Kmenta, 1974, 'A General Method for Obtaining Maximum Likelihood Estimates in Generalized Regression Models', Econometetrica, 42:3, 579-590.

Prasad, R. M., and A. K. Sinha, 1977, 'On Bootstrap Identification Using Stochastic Approximation', IEEE Trans. Auto. Cont., AC-22:4, 671-672.

Rescigno, A., and G. Segre, 1962, "Analysis of Multicompartmented Biological Systems", J. Theoretical Biology, 3:149-163.

Sage, A. P., and J. L. Melsa, 1971, System Identification, Academic Press, New York.

Saridis, G. N., and G. Stein, 1968a, 'Stochastic Approximation Algorithms for Linear Discrete-Time System Identification', IEEE Trans. Auto. Cont., AC-13:5, 515-523.

Saridis, G. N., and G. Stein, 1968b, 'A New Algorithm for Linear System Identification', IEEE Trans. Auto. Cont., AC-13:5, 592-594.

Shaman, P., 1973, 'On the Inverse of the Covariance Matrix for an Autoregressive-Moving Average Process', Biometrika 60:1, 193-196.

Shaman, P., 1975, 'An Approximate Inverse for the Covariance Matrix of Moving Average and Autoregressive Processes, Annals of Statistics, 3:2, 532-538.

Siddiqui, M. M., 1958, 'On the Inversion of the Sample Covariance Matrix in a Stationary Autoregressive Process', Annals of Math. Stat. 29, 585-588.

Steiglitz, K., and L. E. McBride, 1965, 'A Technique for the Identification of Linear Systems', IEEE Trans. Auto. Cont., AC-10, 461-464.

Stoica, P., and T. Soderstrom, 1977, 'A Method for the Identification of Linear Systems Using the Generalized Least Squares Principle', IEEE Trans. Auto. Cont., AC-22:4, 631-634.

Wilks, S. S., 1962, Mathematical Statistics, John Wiley & Sons, New York.