COWLES COMMISSION FOR
RESEARCH IN ECONOMICS
Monograph No. 10

STATISTICAL INFERENCE IN DYNAMIC ECONOMIC MODELS

By

COWLES COMMISSION RESEARCH STAFF MEMBERS AND GUESTS

Edited by

TJALLING C. KOOPMANS

With Introduction by

JACOB MARSCHAK

JOHN WILEY & SONS, INC., NEW YORK
CHAPMAN & HALL, LIMITED, LONDON
1950

PRINTED IN THE UNITED STATES OF AMERICA

CONTRIBUTORS TO THIS VOLUME

RICHARD L. ANDERSON
North Carolina State College

THEODORE W. ANDERSON, JR.
Cowles Commission and Columbia University

TRYGVE HAAVELMO
Cowles Commission and University of Oslo

HAROLD HOTELLING
University of North Carolina

LEONID HURWICZ
Cowles Commission and University of Illinois

TJALLING C. KOOPMANS
Cowles Commission, The University of Chicago

ROY B. LEIPNIK
Cowles Commission (formerly), Institute for Advanced Study

JACOB MARSCHAK
Cowles Commission, The University of Chicago

HENRY B. MANN
Ohio State University

HERMAN RUBIN
Cowles Commission (formerly), Stanford University

ABRAHAM WALD
Columbia University

ACKNOWLEDGMENTS

The publication of this study, and the research of Cowles Commission staff members incorporated in it, has been made possible by generous grants received from the Rockefeller Foundation over an extended period.

Thanks are due to the guest authors who have made their contributions available for inclusion in this volume.

Mr. Dickson H. Leavens, Mr. B. A. de Vries, and the staff of the Cowles Commission, especially Mrs. Jane Novick, Editorial Secretary, Mr. Gerhard Stoltz, and Mrs. Lorraine Kirk, have put much care into editing and preparing the manuscript. Mr. Marvin Schuler has displayed skill, ingenuity, and perseverance in varityping the manuscript for reproduction by photo-offset.

TABLE OF CONTENTS

PART TWO: PROBLEMS SPECIFIC TO TIME SERIES

Trend and Seasonality

Estimation Problems

Continuous Stochastic Processes

PART THREE: SPECIFICATION OF HYPOTHESES

PRINCIPLES OF NOTATION

MATHEMATICAL SYMBOLS

α, β, η, χ, ... Greek characters denote known or unknown constants (parameters).

a, b, y, z, ... Latin characters denote quantities subject to a probability distribution, or exogenous variables assumed given.

α, a; π, p; ..... Unknown parameters and their estimates are denoted as much as possible by corresponding (Greek and Latin) characters.

A, A; B, B; ..... Greek and Latin capitals can be distinguished by the fact that all Greek characters are vertical, all Latin characters italicized.

$g = 1, \ldots, G$; .... Latin characters (lower case) are also used as subscripts for numbering of variables, equations or observation periods. In such cases the range of the subscript is often from 1 to a maximum denoted by the corresponding Latin capital letter.

f, F, φ, Φ, ... Greek and Latin characters are used without distinction to denote distribution functions.

a, b, A, B, ... Vectors are denoted by lower case type, matrices by capitals. Subscripts preceding a vector or matrix denote the deletion of elements, rows, columns, or sets thereof.

a_{12}, β_{gh}, x_k, ... Scalar elements of matrices or vectors are denoted in the normal manner by subscripts following the lower case character corresponding to the character denoting the matrix or vector.

α_g, y_{II}, M_{yz}, ... In some cases, explained in the text, subscripts denote sets of elements, rows or columns, to be included in a submatrix or subvector of the matrix or vector denoted by the symbol to which these subscripts are attached.

${}_{\text{I}}A$, ${}_{\text{III}}a$, ${}_{[1]}\Phi$, ... Preceding subscripts similarly denote elements, rows or columns excluded in forming subvectors

or submatrices.

$\mathcal{E}, \mathcal{P}, \mathcal{Q}, \mathcal{R}, \ldots\ldots$ Script letters denote operators.

$\mathfrak{G}, \mathfrak{H}, \ldots\ldots\ldots$ German letters denote sets.

REFERENCES

Items in the list of references at the end of this volume are referred to by author's name and where necessary by year of presentation or publication in square brackets: [Haavelmo, 1943].

Articles contained in this volume are referred to by Roman numerals in square brackets: [III].

Sections of articles are referred to by italicized Arabic numerals: [I – 2.5].

Formulae are referred to by vertical Arabic numerals in parentheses: (26).

STATISTICAL INFERENCE IN ECONOMICS:

AN INTRODUCTION

I. STATISTICAL INFERENCE IN ECONOMICS: AN INTRODUCTION

BY J. MARSCHAK

0. PURPOSE OF THE VOLUME

0.1. *The Problem*

Quantitative economic study has a threefold basis: it is necessary to formulate economic hypotheses, to collect appropriate data, and to confront hypotheses with data. The latter task, statistical inference in economics, was discussed at a Cowles Commission conference held at the University of Chicago from January 27 to February 1, 1945. Staff members of the Cowles Commission prepared,

and circulated in advance, some of the papers; others were delivered by the Commission's guests.[1]

In its *Annual Report* for 1944 and for subsequent years, the Commission stressed the importance of adapting statistical methods to the peculiarities of the data and the objectives of economic research. The economist's objectives are similar to those of an engineer but his data are like those of a meteorologist. The economist is often required to estimate the effects of a given (intended or expected) change in the "economic structure," i.e., in the very mechanism that produces his data. None of these changes can he produce beforehand, as in a laboratory experiment; and since some of the changes envisaged have never happened before, the economist often has to estimate the results of changes that he has never observed.

The economist can do this if his past observations suffice to estimate the relevant structural constants prevailing before the change. Having estimated the past structure the economist can estimate the effects of varying it. He can thus help to choose those variations of structure that would produce – from a given point of view – the most desirable results. That is, he can advise on policies (of a government or a firm).

Thus, practical considerations bring about the economist's concern with economic structure. Hypotheses about economic structure are also known as economic theories. They try to state relations that describe the behavior and environment of men and determine the values taken at any time by economic variables such as prices, output, and consumption of various goods and services, and the prices and amounts of various assets. As there are several variables the economic structure must involve several simultaneous relations to determine them. In this, economic theory is analogous to theories used in experimental science.

Also, economic variables as well as those of experimental science are, in principle, random (stochastic) variables: that is, their properties are described by probability distributions. In particular, the stochastic character of the observed data can often be ascribed to their dependence on stochastic nonobservable variables: such nonobservable variables are random "errors" in the observation of single variables or random "shocks" suffered by the relations connecting

---

[1]The guests included R. L. Anderson, T. Haavelmo, H. Hotelling, W. G. Madow, H. B. Mann, G. Tintner, and A. Wald. Staff members who participated in the conference were L. Hurwicz, L. R. Klein, T. C. Koopmans, R. Leipnik, J. Marschak, and H. Rubin. T. W. Anderson and T. Haavelmo joined the staff at a later date.

them. However, an experimenter could replace the natural conditions by laboratory conditions. To study one of the several relations, the experimenter observes the random values taken by one variable when the other observables that determine it are made reasonably free of the influence of errors and shocks. The economist cannot thus control variables and isolate relations. His data are produced by the existing economic structure, as described by a system of simultaneous relations between these random variables: the observables themselves, the errors, and the shocks. To use such data for the estimation of the system – "structural estimation" – is a new statistical problem.

This new statistical problem is thus forced upon the economist by the occurrence or consideration of structural changes (including policy changes on which his advice is sought), and by his inability to make experiments of either of two kinds: experiments producing in advance the considered change in structure (e.g., wind-tunnel experiments on airplanes), and experiments in which some of the random variables of nature are given fixed values (e.g., experiments to test fundamental laws of physics). See [J.Marschak, 1947B].

The role of simultaneous equations is familiar to economic theorists. But it has often been forgotten by economic statisticians who tried to estimate a single stochastic relation as if no other such relations had taken part in determining the observed values of the variables. On the other hand, economic theorists are apt to forget that the observed economic variables are, in general, stochastic. To be susceptible of empirical tests an economic hypothesis must be formulated as a statistical one, i.e., be specified in terms of probability distributions.

The statistical problem of the economist is complicated by the fact that many an economic relationship connects current and past values of the same or other variables involved. The economic structure determines, accordingly, not a set of constant values, one for each variable, but a set of probable paths, one for each variable, provided certain initial values are given. This dynamic character of economic structure creates, in the absence of experiments, further statistical difficulties: many economic data have the form of time series in which successive items are not independent. Statistical inference from time series of this kind involves further new problems.

Thus, economic data are generated by systems of relations that are, in general, stochastic, dynamic, and simultaneous. Occurring jointly, these three properties give rise to unsolved problems of statistical inference from the observed data to the relations.

Yet these very relations constitute economic theory and knowledge of them is needed for economic practice.

0.2. *The Discussion*

All these difficulties, under names like "pitfalls in demand and supply analysis" (Ragnar Frisch), "lack of independence in economic time series," etc., have caused uneasiness for a long time. Of the many attempts to grapple with the problem, Ragnar Frisch's contributions [1929, 1931, 1933, 1934, 1938] were probably the most stimulating ones. However, he did not take full account of the random disturbances (shocks) in the economic relations, nor of the simultaneous character of these relations. Moreover, Frisch's hypotheses on random disturbances (errors) in variables were not specified in probability terms. The latter, but not the former, defect was corrected in the early work of Koopmans [1937] and Wald [1940]. A new milestone was reached in 1943 when two articles were published in *Econometrica* by Haavelmo [1943], and by Mann and Wald [1943]. Haavelmo formulated the economist's simultaneous-equations model as a statistical hypothesis by assuming a random disturbance (shock) in each equation, in addition to random errors in each observable variable, and by specifying the distribution of these (unobservable) random quantities [Haavelmo, 1944, esp. Chapter III]. Mann and Wald outlined a solution of the estimation problem arising from the new formulation, though only for the case of large samples, and omitting the observation errors. For a set of short time series of interrelated variables, the contemporary (and incomplete) work on time series in a single variable has to be utilized as a start; and important suggestions can also be expected from the study of continuous random processes that is being developed currently in the service of other sciences. As to combining shocks and errors in one equation system, recent investigations of T. W. Anderson and L. Hurwicz [1947] were stimulated by discussions with G. Tintner.

For its quantitative studies of economic behavior, the Cowles Commission had to expect much from the criticism of statisticians who had contributed to the estimation of simultaneous equations and to the theory of time series. Such was the object of the conference. Earlier drafts of articles II, V, VI, XI, XII, XIV, XV, XVI, XVII, XVIII, XIX, were prepared for the conference. Articles III, VIII, X, XIII, are discussions that were contributed in the conference and written up shortly afterwards.

Articles or parts of articles added or substantially expanded later are I, II–2.3, II–3.3, II–4, IV, VI, VII, IX, XI–10.2, XIV, XV. The revision was helped by discussions with persons not present at the conference: especially on the problem of computations [II–4] with A. Adrian Albert and John von Neumann. An alternative method of structural estimation, that of "limited information," suggested by M. A. Girshick and worked out by T. W. Anderson and H. Rubin [1949], is briefly presented in this volume by T. W. Anderson [IX]. Other conference contributions, by L. R. Klein [1946 A] and by W. G. Madow [1945], have been published elsewhere. Madow's subject was explored further by R. B. Leipnik [1947].

The manuscript of the present volume was completed early in 1947, but publication has been delayed by typographical and other printing difficulties.

In the next two sections of the present introductory paper, the author has drawn freely on the results attained in the papers that follow and on suggestions made in the daily work and discussion within the Cowles Commission. His debt to Leonid Hurwicz and Tjalling C. Koopmans is particularly heavy.

We have tried to achieve conformity in terminology and, to some extent, in notation. However, since the several contributions differ in purpose as well as in emphasis, rigorous uniformity is neither possible nor desirable.

Most of the contributions to this volume presuppose on the part of the reader a general knowledge of mathematical principles of statistics; to explain these principles to a more general reader would take more space than is available. The present introduction, in summarizing the purpose and the main results of the studies collected in this volume, is addressed to the mathematically-minded economist rather than to the statistician. Hence – the use made of nonstochastic models (section 1) and the attention paid, in the stochastic case, to the properties of populations (section 2.5.1) as distinct from samples. This summary, too, has to be terse. For a less compressed treatment and further economic illustrations and applications the reader is referred to the following publications of Cowles Commission staff members: [Girshick and Haavelmo, 1947], [Haavelmo, 1943, 1944, 1947 A, 1947 B], [Hurwicz, 1947], [Klein, 1946 B, 1947, 1950], [Koopmans, 1945, 1949], [Marschak and Andrews, 1944], [Marschak, 1947 A, 1947 B]. The group continues to work on statistical inference in economics, both in general and with respect to specific economic models.[1] It is hoped that the present volume

[1] See the *Annual Report* of the Cowles Commission, in particular the *Five-Year Report* for 1942-46.

will stimulate further cooperation of mathematical statisticians and economists in solving the many problems that have been indicated but not solved in this volume. Plans are under way for a parallel monograph (No. 12) of a more expository character, in which emphasis is placed on a discussion of the main ideas and techniques developed in this volume with the help of simple illustrative models, rather than on formal mathematical proof.

1. NONSTOCHASTIC MODELS

Economic relations are, in general, stochastic. They involve variables whose properties are described with the aid of probability distributions. Moreover, the estimates of parameters of these relations, obtained by statistical methods from a limited number of observations, are also random variables.

However, important distinct properties of empirical economics can be brought out even if, for simplicity, we assume the data to be measured exactly and to satisfy exactly the relations of theory. The equations or inequalities serving to determine the parameters of such relations from observations are free from random variables. The problem of estimation degenerates into that of determination. This simplifies the study of certain "prestatistical" problems facing the economic statistician – in particular that of identification (section *1.3*), and also helps to see why, as indicated in section *0.1*, these problems originate in the need for policy decisions in the absence of experiments.

In the present section we shall deal with this special, or degenerate, case to meet in particular the habits of readers with economic rather than statistical background.

1.1. The Model

1.1.0. Denote the observable variables (or observables) by a vector $x \equiv (x_1, \ldots, x_N)$. The first, second, ..., Tth observations on x, succeeding each other in time, or arranged in space or in any other way, form a matrix:

$$X^o \equiv [\, x_n(t) \,], \qquad n = 1, \ldots, N, \quad t = 1, \ldots, T;$$

or

$$X^o \equiv \begin{bmatrix} x(1) \\ \cdots \\ x(T) \end{bmatrix}.$$

We call *a priori information* all statements (either true or false) arrived at independently of any knowledge of X^o. We call *model* $\mathfrak{S}$ the a priori information on a system of mutually consistent and independent equations

$$\varphi_g(x, \alpha_{(g)}) = 0, \qquad g = 1, \ldots, G, \tag{1.1}$$

where $\alpha_{(g)}$ is a vector of P_g parameters. We shall denote the vector of all parameters of the system by $\alpha \equiv (\alpha_{(1)}, \ldots, \alpha_{(G)})$, and write $\sum P_g = P$; we shall also write[1] $\bar{\varphi} \equiv (\varphi_1, \ldots, \varphi_G)$.

1.1.1. We shall assume throughout that $\mathfrak{S}$ defines a) the form of $\bar{\varphi}$, and b) the "a priori restrictions," i.e., equations or inequalities in parameters α.

1.1.2. To provide this information we must make full use of our independent knowledge of existing production conditions (technology, legal statutes, etc.) and of plausible, if not necessarily rational, individual behavior. The equations of the model must refer to individual agents in specified markets (as consumers or manufacturers of certain goods, or as workers, bankers, landlords, etc.).

1.1.3. However, to reduce the model to a manageable size, variables referring to single individuals in finely subdivided markets must be grouped into aggregates. Suppose the value of some variable relevant to a practical decision (see below, section *1.2.3*) is calculated on the basis of such an aggregative model. This value will contain an error inasmuch as it will deviate from the corresponding value calculated on the basis of a true, detailed model, with separate equations for each commodity and individual. Optimum aggregation should combine highest manageability (e.g., shortest computations) with smallest error. This aggregation problem (which includes that of index numbers) has not been solved or even formulated in detail, nor will it be studied in this volume.

1.1.4. The model $\mathfrak{S}$ is called *self-contained* if $G = N$; *sectional* if $G < N$.

We call a model *complete* if it is either self-contained or has the following property: a subset of x containing $K = N - G$ elements – call this subset $z \equiv (z_1, \ldots, z_K)$ – is determined by K

[1] Subscripts indicate the elements of a vector which are scalars or functions: x_n, φ_g. In the present article, subvectors formed from the elements of a vector are indicated by subscripts in parentheses: $\alpha_{(g)}$. In addition, it will here prove convenient to use a barred letter for a vector whose elements are functions: $\bar{\varphi}$.

possibly unknown "subsidiary" equations

$$\varphi_{G+k}(z) = 0, \qquad k = 1, \ldots, K, \tag{1.2}$$

which are independent of the equation (1.1); the equations (1.2) do not contain any elements of x that are not in z. Denote the latter elements by vector $y \equiv (y_1, \ldots, y_G)$. Then, assuming differentiability,

$$\frac{\partial \varphi_{G+k}}{\partial y_g} = 0, \qquad g = 1, \ldots, G, \qquad k = 1, \ldots, K. \tag{1.3}$$

A model that is not complete is called incomplete (or partial).

The observables z are called *exogenous* and the observables y *endogenous*, with respect to the model $\mathfrak{S}$.

Equation (1.1) can be rewritten as

$$\varphi_g(y, z; \alpha_{(g)}) = 0, \qquad g = 1, \ldots, G. \tag{1.4}$$

1.2. *Structural Changes and Policies*

1.2.1. Structure and reduced form.

1.2.1.1. We call *structure* S all properties of the equations (1.4), including the properties not known a priori. Any model $\mathfrak{S}$ is a class of structures. Each structure is defined by the functional forms of the equations and the values of the parameters occurring in them. We can write[1]

$$S \equiv (\bar{\varphi}, \alpha). \tag{1.5}$$

When the equations (1.4) are thus fully specified we call them *structural equations*. We call α the structural parameters. Concepts defined in section *1.1.4* with reference to the model ("complete," "sectional" model; variables exogenous to a model; etc.) will also be applied to a structure without causing ambiguity.

1.2.1.2. Given the structure S, equations (1.4) can be solved for y in terms of z, involving new parameters which we shall denote

[1] In what follows quantities depending on S will be introduced. They are, properly speaking, "functionals" with respect to the argument $\bar{\varphi}$ and "functions" with respect to the argument α, although we shall denote and refer to them as "functions" of S.

by vector π:

$$y = \bar{\eta}(z;\ \pi), \tag{1.6}$$

say, where $\bar{\eta}$ is a vector whose elements are functions. Equation system (1.6) is called *reduced form*.[1] $\bar{\eta}$ depends on $\bar{\varphi}$; π is obtained by applying a transformation to the parameters α in (1.4). This transformation itself depends on the functions $\bar{\varphi}$. If we call this transformation $\bar{\pi}_{\varphi}$, we can write, accordingly,

$$\pi = \bar{\pi}_{\varphi}(\alpha) = \bar{\pi}(S), \tag{1.7}$$

say; furthermore, we can also write $\bar{\eta}_{\varphi}$ instead of $\bar{\eta}$ to emphasize the dependence of the functions $\bar{\eta}$ on the functions $\bar{\varphi}$. Thus

$$y = \bar{\eta}_{\varphi}[z;\ \bar{\pi}_{\varphi}(\alpha)]. \tag{1.8}$$

1.2.1.3. If the structural functions φ are linear, the functions $\bar{\eta}$ and $\bar{\pi}_{\varphi}$ will also be linear, and the set of parameters $\pi = \bar{\pi}_{\varphi}(\alpha)$ of the reduced form corresponding to a given structure S will be unique. If the functions φ are nonlinear, several sets of parameters α may be compatible with the structure S. We shall neglect here this complication although it does occur in economic theory, as in the case of "multiple equilibrium" [Marshall, app. H].

1.2.1.4. Apart from very special cases, the reduced forms (each one characterized by a function set $\bar{\eta}$ and a parameter set π) compatible with a given structure S will be finite in number, or at least denumerable; in a linear model the reduced form is unique.

On the other hand the number of structures compatible with a given reduced form $(\bar{\eta},\ \pi)$ may or may not be (nondenumerably) infinite. Even if the model is linear, the structure may or may not be uniquely determined by the reduced form. (See also below, section *1.3.*)

1.2.2. Use of observations.

1.2.2.1. Suppose the period (or, say, the geographical area) of observations consists of elements 1, ..., T, and is so chosen that the structure maintains throughout it the value

$$S^{o} \equiv (\bar{\varphi}^{o},\ \alpha^{o}).$$

Then the following equations are satisfied by the observations X^{o}:

[1] See [IX] where $\bar{\varphi}$ and consequently $\bar{\eta}$ are linear functions.

$$\varphi_g^o[x(t);\ \alpha_{(g)}^o] = 0, \quad g = 1, \ldots, G, \quad t = 1, \ldots, T. \tag{1.9}$$

The structure S^o is called *observational structure*. We can similarly consider the equations

$$y(t) = \bar{\eta}^o[z(t),\ \pi^o], \qquad t = 1, \ldots, T, \tag{1.10}$$

where

$$\bar{\eta}^o \equiv \bar{\eta}^o_{\varphi^o} \tag{1.11}$$

and

$$\pi^o \equiv \bar{\pi}_{\varphi^o}(\alpha^o) \equiv \bar{\pi}(S^o); \tag{1.12}$$

the "observational reduced form" is given by $\bar{\eta}^o; \pi^o$. It follows from definitions that (1.10) is satisfied by observations X^o.

1.2.2.2. The form of the functions $\bar{\eta}_{\varphi^o}$ and $\bar{\pi}_{\varphi^o}$ is determined by the functions $\bar{\varphi}^o$ of the model. If the latter are linear so are the former. The model also provides restrictions on α^o, and these restrictions can be transformed, by (1.12), into equations or inequalities in π^o. If the model is linear, a unique set of parameters π^o will correspond to the set of structural parameters α^o. Furthermore, if T, the number of observations, is sufficiently large, the equations (1.9) together with the restrictions, determine the parameters π^o of the observational reduced form, given the observations X^o. We shall denote this operation by $\mathfrak{P}_1$, so that

$$\mathfrak{P}_1 X^o = \pi^o. \tag{1.13}$$

1.2.2.3. As stated in section *1.2.1.4*, a given reduced form may or may not determine the structure uniquely. Therefore, although, as just stated, the observational reduced form can be determined from the model, given the observations X^o (if their number T is sufficiently large), it may be impossible to determine a unique observational structure from the model and the observations *however large their number T*. It may sometimes be possible, in other words, to replace a given structure by any one of an infinite number of other structures without contradicting the observations. There exists thus a problem of *identifying a structure* (treated below in section *1.3*) but no problem of identifying a reduced form.

1.2.3. Structural changes and policies.

1.2.3.1. The observational structure S^o may be different from some structure S valid for a different period (or geographical area, etc.). Suppose we know the "structural changes" $\mathfrak{I}$, a transformation that carries S^o into S: $S = \mathfrak{I}S^o$. If, in addition, we know S^o from observations, we can obtain $S \equiv (\varphi, \alpha)$ and hence also the functions $\bar{\eta}_\varphi$ and $\bar{\pi}_\varphi$ and therefore the parameters π of the reduced form. It is then possible to evaluate[1] y for a given z.

1.2.3.2. We shall distinguish between two kinds of structural changes: the *controllable* ones, $\mathfrak{I}_c$; and the *uncontrollable* ones, $\mathfrak{I}_u$. The former ones are also called "structural policy;" the introduction, or abolition, of price control is an example. We can distinguish similarly between two sets of exogenous variables: the controllable ones, z_c; and the uncontrollable ones, z_u. The fixing of z_c is called "nonstructural" (or "routine") policy: for example the annual revision of tax rates [Marschak, 1947 A].

1.2.3.3. The policy-maker (on behalf of a government, of an individual firm, etc.) tries to maximize the "gain," or "welfare," ω, a certain function of the observables which, in principle, must be supposed to be known to him:

$$\omega = \omega(y, z) = \omega\{\bar{\eta}_\varphi[z; \bar{\pi}_\varphi(\alpha)], z\} = \omega_\varphi[z; \bar{\pi}_\varphi(\alpha)], \tag{1.14}$$

say. Thus the gain (welfare) function ω and the functions φ of the model combine to determine the function ω_φ of exogenous variables which is to be maximized, given the structure $(\bar{\varphi}, \alpha)$.

1.2.3.4. Suppose structural changes are neither intended nor expected for the period (or area) to which policy is to be applied, compared with the period (or area) of observation. (The policy consists, in this case, in fixing the value of z_c only: it is "nonstructural.") In this case, $\mathfrak{I}_c$ as well as $\mathfrak{I}_u$ is the identical transformation and we have $\varphi = \varphi^o$, $\alpha = \alpha^o$, and

$$\pi = \pi_{\varphi^o} \quad \text{(a set of constants)}. \tag{1.15}$$

Thus the gain

$$\omega = \omega_{\varphi^o}(z; \pi) = \omega_{\varphi^o}(z_c, z_u; \pi) \tag{1.16}$$

[1] But see section *1.2.1.3* for a qualification.

can be affected only by variations in z_c. Best policy is the value $\hat{z}_c$ of z_c that gives ω its maximum value $\hat{\omega}$:

$$\hat{\omega} = \max_{z_c} \omega_{\varphi^o}(z_c,\ z_u;\ \pi^o) = \omega_{\varphi^o}(\hat{z}_c,\ z_u;\ \pi^o). \tag{1.17}$$

By comparing the values of ω for varying z_c at fixed z_u, the best policy $\hat{z}_c$ can be determined for any given value of the uncontrollable variables z_u, provided the parameters π^o of the (observational) reduced form are known. But π^o can indeed be found from the observations X^o (section 1.2.2.2). The operation (1.13),

$$\mathfrak{P}_1\ X^o = \pi^o,$$

depends on the model only and is called "predictive determination when structure is unchanged."[1] Operation $\mathfrak{P}_1$ provides, then, the parameters π^o, to be used for the choice of policy under unchanged structure.

1.2.3.5. As a rule, however, some structural changes will be intended or expected, or both. That is, neither $\mathfrak{I}_c$ nor $\mathfrak{I}_u$ will, in general, be the identical transformation. We have $S = \mathfrak{I}S^o$, where $\mathfrak{I} = \mathfrak{I}_c\mathfrak{I}_u$ (neglecting the question of the order in which the transformations are applied). Further, by (1.14) and (1.15), we have

$$\omega = \omega_{\varphi}[z;\ \pi_{\varphi}(\alpha)] = \omega^*(z\ ;\ S), \tag{1.18}$$

say, where the form of the function ω^* depends on the form of the gain function ω only. Further,

$$\begin{aligned}\omega^*(z,\ S) &= \omega^*(z_c,\ z_u;\ \mathfrak{I}S^o)\\ &= \omega^{**}(z_c,\ z_u;\ \mathfrak{I}_c,\ \mathfrak{I}_u;\ \varphi^o,\ \alpha^o).\end{aligned} \tag{1.19}$$

Best policy is defined by values $\hat{z}_c$, $\hat{\mathfrak{I}}_c$ that jointly maximize the gain. Let the maximum gain be

$$\hat{\omega} = \omega^{**}(\hat{z}_c,\ z_u;\ \hat{\mathfrak{I}}_c,\ \mathfrak{I}_u\ ;\ \varphi^o,\ \alpha^o). \tag{1.20}$$

---

[1] See [VI].

By comparing the values of ω for varying z_c, $\mathfrak{I}_c$, given the values of the uncontrollable exogenous variables and given the uncontrollable changes in structure, one can determine the best policies $(\hat{z}_c, \hat{\mathfrak{I}}_c)$, provided the observational structure S^o is known. The practical procedure is, in principle, as follows: from S^o and any given $\mathfrak{I} = \mathfrak{I}_c \mathfrak{I}_u$ derive the new structure $S = \mathfrak{I} S^o$ for the period of policy application; from S derive the parameters $\pi = \pi(S)$; and compute the variables y and the gain ω as in (1.14).

1.2.3.6. An operation $\mathfrak{d} X^o = S^o$ determining S^o from the observations X^o will be called *structural determination;* the question of its existence will be discussed in section *1.3.* If structural determination is possible, it is possible to derive the parameters π by applying in succession the operations $\mathfrak{d}$ and $\mathfrak{I}$:

$$\pi = \bar{\pi}(S) = \bar{\pi}(\mathfrak{I}\, \mathfrak{d}\, X^o) = \mathfrak{P}_{\mathfrak{I}} X^o, \tag{1.21}$$

say. The operation $\mathfrak{P}_{\mathfrak{I}}$ thus defined is called *predictive determination when structure undergoes a given change* $\mathfrak{I}$.

1.2.3.7. Structural determination provides a master key for predictive determination, and for the calculation of alternative gains, for *any* of the various possible structural changes. The structural changes to be considered are seldom known long in advance. Therefore, although formally $\pi = \bar{\pi}(\mathfrak{I}\, \mathfrak{d}\, X^o)$ can be computed without a stop, it is preferable and often essential to pause at the step $\mathfrak{d} X^o = S^o$. The knowledge of observational structure means greater flexibility with regard to various alternative policies. This is one reason why people are interested in any kind of theory!

1.2.3.8. However, the considered transformations $\mathfrak{I}$ (structural policies and uncontrollable structural changes) and the subset of variables relevant to the evaluation of the gain, may happen to be such as to make the knowledge of all parameters α^o of the structure S^o unnecessary. A partial knowledge of S^o – some elements of α^o, or perhaps some functions of them – may be all that is needed.

1.2.3.9. It was required in section *1.1.2* that the equations of the model describe plausible behavior of specified economic agents, thus making full use of our a priori knowledge of behavior (rational or otherwise). We now see a further practical reason for shunning relations that do not refer to specified

economic agents. We call such relations "anonymous." Consider changes in human behavior, institutions, technology. The gain (personal or social) due to any such intended or expected change cannot be evaluated unless behavior, institutions, and technology are explicitly stated; such statements must be provided by the form, and by the values of parameters, of the equations of the model; that is, by the structure.

1.2.3.10. As an example, consider the following – admittedly oversimplified – model:

(1) demand for a commodity depends on its price and on national income;

(2) supply depends on price;

(3) demand equals supply;

and it is assumed that

(4) we can treat national income as exogenous.

Suppose we want to evaluate the effect of replacing free demand of the public by a fixed demand determined by the government; that is, relation (1) is replaced by

(1′) demand = a constant.

If we know the form and parameters of (2) we can evaluate, with the help of (1′) and (3), the price the suppliers will ask and the government will have to pay. Suppose, however, that instead of (1), (2), (3), we had at our disposal the following relation obtained by elimination of demand from (1), using (3), (2):

(5) price depends on income.

This "anonymous" relation (which, in this case, is a reduced form) can be computed from observations (section *1.2.2.2*) but cannot help to evaluate the effect of structural policy, i.e., the effect of replacing (1) by (1′). For another example, see [Marschak, 1947B].

1.3. Identification

1.3.1. It was remarked in section *1.2.2.3* that a given set of observations, however numerous, or a given reduced form, will not, in general, determine a unique structure. For a rigorous introduction to this *problem of identification* in the case of stochastic models we refer to [I–2] and [IV]. For the purposes of the present paper, it is convenient to approach the problem by using the concept of reduced form and studying it first in the nonstochastic case.

Consider equations (1.8) and the structure $S^o \equiv (\varphi^o, \alpha^o)$. Denote by $\mathfrak{S}^o_*$ the class of all structures $S^o_* \equiv (\varphi^o_*, \alpha^o_*)$ that are compatible with the model $\mathfrak{S}$ and yield the same reduced form as the structure S^o. The latter condition means that the following equations are satisfied identically in z:

$$y = \bar{\eta}_{\varphi^o}[z;\ \bar{\pi}_{\varphi^o}(\alpha^o)] = \bar{\eta}_{\varphi^o_*}[z;\ \bar{\pi}_{\varphi^o_*}(\alpha^o_*)]. \tag{1.22}$$

S^o is said to be *uniquely identifiable* by the model $\mathfrak{S}$ if S^o is the only element of $\mathfrak{S}^o_*$, i.e., if every $S^o_* = S^o$. Furthermore, S^o is said to be *incompletely identifiable* by the model $\mathfrak{S}$ if the class $\mathfrak{S}^o_*$ contains a nondenumerably infinite number of elements. Only in the former case is it possible to determine α^o uniquely from the parameters of the reduced form, $\pi^o = \bar{\pi}_{\varphi^o}(\alpha^o)$, provided the functions φ^o and the a priori restrictions on α^o are given.

1.3.2. The concept of identifiability can be easily extended to any subset of α^o, say $\alpha^o_{(i)}$ (*partial identification*): for example, all or some of the parameters of some of the equations (1.4) may be uniquely determinable from the parameters of the reduced form.

1.3.3. If a subset of α^o is nonidentifiable it is impossible to determine it from X^o *however large the number of observations* T. If α^o is completely and uniquely identifiable, it can be obtained from X^o using equations (1.9) jointly with the a priori restrictions, provided T is sufficiently large: this is structural determination, denoted in section *1.2.3.6* by $\mathfrak{d}$. A similar statement applies to any subset $\alpha^o_{(i)}$.

1.3.4. As an example, let $x = (y_1, y_2, z_1)$ be the coordinates of a point and let $\varphi_g(x) = 0$, $g = 1, 2$, be a pair of equations of two distinct planes. The observations on x will yield a set of collinear points which will determine a straight line – corresponding to (1.6) – permitting the prediction of y_1 or y_2 for a given z_1. But it will be impossible to reconstruct any particular pair of planes. A (partial) identification is, however, possible if, for example, it is known a priori that one of the planes is vertical; in which case this plane (but not the other one) is identifiable.

1.3.5. This corresponds to the economic example of section *1.2.3.10* with y_1 = demand (=supply), y_2 = price, z_1 = income. The parameters of (2) but not those of (1) are identifiable. If

(2) were modified into

(2′) supply depends on price and wage rate,

with wage rate assumed exogenous [Koopmans, 1945 p.451] all parameters of the structure would become identifiable.

1.3.6. The conditions of identifiability in a linear stochastic model are studied in [II–2]; their application to a nonstochastic model is easily derived. The most important criteria are supplied by the presence or absence of variables in each of the equations of the system (1.4). In particular, the occurrence of different exogenous variables in different equations contributes to identifiability.

1.3.7. Let the number of endogenous variables, $G = 1$; assume φ_1 is linear, and normalize the parameters α by choosing the coefficient of y_1 to be equal to 1:

$$y_1 - \sum \alpha_k z_k = 0. \tag{1.23}$$

The *linear uniequational structure* (1.23) *is always identifiable*, and $\alpha = \pi$, provided there exist no linear relations between the z's. The proposition can be easily extended to nonlinear uniequational structures, apart from trivial modifications.

1.3.8. Predictive determination when structure is unchanged, i.e., the operation $\mathfrak{P}_1 X^0$ in (1.13) is possible, regardless of whether or not the observational structure S^0 is identifiable.

1.3.9. Predictive determination when structure undergoes a known change $\mathfrak{I}$, i.e., the operation $\pi = \mathfrak{P}_{\mathfrak{I}} X^0$ in (1.21) is possible when all parameters α^0 are identifiable, and impossible when none of them are identifiable.

1.3.10. However, the structural change $\mathfrak{I}$ and the gain function ω may be such as to require the knowledge of some but not all elements of α^0 (section *1.2.3.8*); in this case partial identifiability (section *1.3.2*) is all that is needed. Also, $\mathfrak{I}$ and ω may be such as to require the knowledge, not of the parameters α^0 themselves, but of some functions of them; in which case unique identifiability of every single parameter separately is not necessary for the choice of best policy.

1.3.11. An incomplete model has more endogenous variables than it has equations (section *1.1.4*). For example a uniequational model involving two or more endogenous variables is incomplete. The parameters of such a model form a subset of the parameters of some complete model. The parameters of the incomplete model may or may

not be determinable from observations since the structure may or may not be partially identifiable (section 1.3.2), by the complete model, with respect to the particular subset of its parameters.

1.4. *Use of Experiments*

1.4.0. We shall now distinguish between: *original* structure S^{oo}, *future* structure $S = \mathfrak{I} S^{oo}$, and *observational* structure $S^o = \mathfrak{C} S^{oo}$. In the preceding section observations were supposed to be made on the original structure, so that $\mathfrak{C}$ was the identical transformation. This is indeed the situation in nonexperimental science. Experiments, on the other hand, consist in applying certain transformations that change the original structure and in getting observations from the new structure $S^o = \mathfrak{C} S^{oo}$ that is thus obtained. The transformations $\mathfrak{C}$ are chosen in such a way as to permit predictive determination under future structure S without determining either S or the (possibly nonidentifiable) original structure S^{oo}. According to the nature of the transformation $\mathfrak{C}$, there are various types of experiments. Two particular types of experiments deserve our attention [Marschak, 1947 B].

1.4.1. Experiments of type I: imitation of future structure (example: wind tunnels for testing airplanes). Here $\mathfrak{C} = \mathfrak{I}$, hence $S^o = S$ and $\pi^o = \pi$. The structural change occurs between S^{oo} and S. To evaluate y for a given z under the future structure S, and thus to evaluate the gain ω, it suffices to determine π^o by the operation $\mathfrak{P}_1 X^o = \pi$ (section 1.2.2.2). This does not require identifiability of either S^{oo} or S.

1.4.2. Experiments of type II: creation of uniequational complete structure (example: controlled experiments in the physical laboratory). Let the original structure S^{oo} correspond to a model

$$\varphi_g^{oo}(y, z; \alpha_{(g)}^{oo}) = 0, \qquad g = 1, \ldots, G. \tag{1.24}$$

Let $\mathfrak{C}_1$ be the operation of replacing all equations (1.24) but the first by the equations

$$y_g = z_{K+g}, \qquad g = 2, \ldots, G, \tag{1.25}$$

where the quantities $(z_{K+1}, \ldots, z_{K+G}) \equiv$ vector z_* are fixed by the experimenter at various values $z_*(1), \ldots, z_*(T)$, these

values being mutually independent. Denote the resulting observations by matrix X_1^o; they can be regarded as produced by an uni-equational complete structure involving a vector of parameters $\alpha_{(1)}^o$, as follows:

$$\mathbb{O}_1 S^{oo} = S_1^o = (\varphi_1^o, \alpha_{(1)}^o);$$

that is,

$$\varphi_1^o(y_1, z, z_*; \alpha_{(1)}^o) = 0, \tag{1.26}$$

where $\varphi_1^o = \varphi_1^{oo}$, and $\alpha_{(1)}^o = \alpha_{(1)}^{oo}$. The vector $\alpha_{(1)}^o$ of structural parameters is related by trivial transformations (e.g., dividing by a constant if φ_1^o is linear) to the vector of parameters $\pi_{(1)}^o$ $(= \alpha_{(1)}^{oo})$ of the reduced form

$$y_1 = \eta_{\varphi^o}(z, z_*; \pi_{(1)}^o). \tag{1.27}$$

Hence the subset $\alpha_{(1)}^{oo}$ $(= \alpha_{(1)}^o)$ can be determined (section *1.3.7*) even though the set α^{oo} may be nonidentifiable.

If it is desired to determine the whole set α^{oo}, experiments $\mathbb{O}_1, \mathbb{O}_2, \ldots, \mathbb{O}_G$ can be applied in succession.

2. STOCHASTIC MODELS

2.1. Random Disturbances

2.1.1. The stochastic character of economic relations will be recognized even by an out-and-out determinist. His world is, in principle, ruled by a set of very many equations in very many variables (both economic and other). But to make his theory verifiable by observation he will have to shorten the set considerably. The numerous causes that determine the error incurred in measuring a variable are not listed separately; instead, their joint effect is represented by the probability distribution of the error, a random variable. Also, the numerous causes that determine, say, the velocity of a gas particle are conveniently represented by the probability distribution of this velocity, a random variable. The economist acts similarly. He allows for

random errors of observation; and he represents the vagaries of, say, changing fashion by random "disturbances" or "shocks" that obey certain probability distributions; he thus cuts short the complicated causal explanation of why tastes fluctuate in the way they do.

The nondeterminist, on the other hand, will find it unnecessary to give any justification for the presence of random elements in economic models except by appealing directly to the "erratic," "unpredictable" character of certain types of events including human behavior; though he, too, will have to assume those events to be bound by certain probability distributions: even if actions are unpredictable, certain actions remain more probable than others.

2.1.2. Denote by $w \equiv (w_1, \ldots, w_J)$ the vector of nonobservable random disturbances affecting economic observations, and by $x \equiv (x_1, \ldots, x_N)$ the vector of observable variables. We shall call a stochastic model $\mathfrak{S}$ the a priori information on a system of equations

$$\varphi_g(x, w; \alpha_{(g)}) = 0, \qquad g = 1, \ldots, G, \tag{2.1}$$

and on the joint distribution density

$$f(w; \varepsilon), \tag{2.2}$$

where ε as well as $\alpha_{(g)}$ denote parameter vectors. As in section 1.1.0 we shall write $\alpha \equiv (\alpha_{(1)}, \ldots, \alpha_{(G)})$. We shall assume, in particular, that a priori information exists a) on the forms of the functions $\bar{\varphi} \equiv (\varphi_1, \ldots, \varphi_G)$ and f; and b) on some equations or inequalities in the parameters of these functions. We shall call *structure* all properties of the equations (2.1) and of the distribution (2.2) including the properties not known a priori.

If $w = 0$, we have the nonstochastic case treated in section 1 of this article.

2.1.3. If it is possible to substitute for w from (2.1), the distribution density function f of the disturbances can be transformed into the distribution density function of the observable vector x:

$$g_x(x), \tag{2.3}$$

say. Given the form of the functions φ and f and given their parameters α, ε, the distribution function g_x and its parameters are determined. Certain parts of the present volume [II-2 and IV] deal with the converse problem: given the distribution (2.3) of the observables, determine the properties of the equations (2.1) and the properties of the distribution (2.2) of the disturbances. Well-known methods permit estimation of the distribution of the observables (2.3) from the observations. But to use the knowledge of that distribution to determine the properties of (2.1) and (2.2) raises a new problem in inference, that of identification. The relation between the stochastic model and the distribution of the observables will show certain analogies with the relation, already studied, between the nonstochastic model and the reduced form.

2.1.4. The random disturbances w may include as a subset a vector of additive *disturbances in variables* (additive errors of observations, or briefly, *errors*) $v = (v_1, \ldots, v_H)$. Thus the model (2.1), (2.2) can be rewritten, slightly changing notation,

$$\begin{aligned} \varphi_g(x - v, w') &= 0 \\ f(v, w'), &\qquad g = 1, \ldots, G, \end{aligned} \tag{2.4}$$

where the vector w' is complementary to v in w, i.e., $w \equiv (v, w')$.

2.1.5. There are also random *disturbances in relations*, especially of the additive kind, which we call for brevity *shocks*, $(u_1, \ldots, u_G) \equiv u$, a subset of w'. If we denote by vector w'' the complement of u in w' so that $w \equiv (u, v, w'')$, we can rewrite the model (again slightly changing the meaning of φ):

$$\begin{aligned} \varphi_g(x - v, w'') &= u_g, \\ f(u, v, w''), &\qquad g = 1, \ldots, G. \end{aligned} \tag{2.5}$$

2.1.6. If w'' is empty the model may be called "simple shock-and-error model" [T. W. Anderson and Hurwicz, 1947]. In the present volume the contributions of the Cowles Commission staff are confined to the more special case where w'' and v are empty: the "simple shock model." Other studies, [Frisch and Mudgett, 1931], [Koopmans, 1937], [Wald, 1940], [Tintner, 1946], [Reiersøl, 1945], and [Geary, 1942], may be said to deal with another special case: the "simple error model," in which both w'' and u are empty.

2.1.7. Although even the simpler stochastic models present considerable difficulties, economists will probably be right in calling the statisticians' attention to more complicated models, i.e., those in which w'' is not empty, and this for three reasons:

(1) there may be nonadditive errors of observation;

(2) there may be nonadditive shocks, since an economic relation can be disturbed in a variety of ways (for example, a linear relation, say a demand curve, can fluctuate owing to random changes in its constant term as well as in any of the coefficients, see [XVIII], [XIX]);

(3) there occur "prospective"[1] variables, such as prospective profits, prices, etc., that affect people's behavior and must enter the equations of the model.

As a rule the economist cannot observe these (except by difficult questioning of a sample of people). But he may have hypotheses describing the determination of each of these variables in the minds of people: such "forecast equations" [Hurwicz, 1946, p.130 ff.] would relate the "prospective" variables to certain "actual" variables (such as the current or lagged national income, etc.) or to the observations on the "actual" variables. These additional structural equations will be themselves subject to psychological fluctuations, expressed as random disturbances, additive or nonadditive. By use of the forecast equations, the prospective variables can be eliminated. The model thus derived can be used in predicting effects of structural changes occurring in equations other than the forecast equations. The derived model will involve functions of the random disturbances that occurred in the forecast equations. Even if they occurred additively in the forecast equations, these disturbances will enter the model, generally, in a nonadditive fashion, i.e., as elements of w'' in (2.5).

However, the remaining parts of this article and the bulk of the monograph itself will deal with simple shock models only, though some of the statements might be generalized to apply to the general model (2.5).

2.2. *Shock Model and Structure*

2.2.1. The introduction of shocks, i.e., additive random

[1] The word "anticipated" variables is sometimes suggested. But this would seem to exclude expectations that do not come true. The term "expected" might be confused with the statistical term, expected (= mean) values.

disturbances in relations (excluding other random disturbances), can be considered as weakening the nonstochastic hypothesis (1.1): the observable variables $x \equiv (x_1, \ldots, x_N)$ are related by a system of equations

$$\varphi_g(x, \alpha_{(g)}) = u_g, \qquad g = 1, \ldots, G, \tag{2.6}$$

the distribution function (probability density) of the random vector $u \equiv (u_1, \ldots, u_G)$ being denoted by

$$f(u; \varepsilon) \tag{2.7}$$

where $\alpha_{(g)}$ and ε are parameter vectors. Thus each equation (other than definitional equations that can be properly eliminated from the system) is subject to a disturbance.

A priori information on the functions and parameters in (2.6), (2.7) constitutes a shock *model*, $\mathfrak{S}$. A shock *structure*, S, consists of all properties of those functions and parameters:

$$S \equiv (\varphi, f, \alpha, \varepsilon),$$

where vector $\alpha \equiv (\alpha_{(1)}, \ldots, \alpha_{(G)})$. If, as will be assumed, the model supplies the forms of the functions (in addition to some a priori restrictions on the parameters), it is permissible also to write, more briefly, $S \equiv (\alpha, \varepsilon)$, provided φ, f do not change. The structural properties to be estimated from the observations will then be the numerical values of α, ε; or, more generally, some relations, not known a priori, between the parameters.

2.2.2. We can apply the earlier definition (section *1.1.4*) of a self-contained model ($G=N$) and a sectional model ($G<N$). But the definition of a complete model and of exogenous vs. endogenous variables must be supplemented. The "subsidiary" system of equations will, in general, involve random shocks of its own, so that (1.2) becomes

$$\varphi_{G+k}(z) = u_k^{(k)}, \qquad k = 1, \ldots, K = N - G. \tag{2.8}$$

Rewriting the earlier condition (1.3) as

$$\frac{\partial \varphi_{G+k}}{\partial y_g} = 0, \qquad g = 1, \ldots, G, \qquad k = 1, \ldots, K, \tag{2.9}$$

we see that this condition is not sufficient to make the variables z determinable outside of the system (2.6). To make them thus determinable, and the considered model complete, we must add the condition that the shocks u of the considered model and the shocks $u^{(z)} \equiv (u_1^{(z)}, \ldots, u_K^{(z)})$ of the subsidiary model be distributed independently:

$$f_x(u,\ u^{(z)}) = f(u)\cdot f_z(u^{(z)}), \tag{2.10}$$

where f_x, f, f_z are density functions. If both conditions (2.9) and (2.10) are fulfilled, the model

$$\begin{aligned} &\varphi_g(y,\ z;\ \alpha_{(g)}) = u_g, \\ &f(u\,;\,\varepsilon), \qquad\qquad g = 1, \ldots, G, \end{aligned} \tag{2.11}$$

is said to be complete, with y as the endogenous and z the exogenous set of observables.

2.2.3. The distribution (2.3), $g_x(x)$, of the observable vector can always be represented as a product of a conditional and a marginal distribution

$$g_x(x',\ x'') \equiv g(x' \mid x'')\cdot g_{x''}(x''), \tag{2.12}$$

where x', x'' is any pair of mutually complementary subsets of x. Suppose now that the model is complete; put $x' = y$, $x'' = z$, and denote by vector λ the parameters of the conditional distribution $g(y\,|\,z)$:

$$g_x(y,\ z) \equiv g(y \mid z\,;\,\lambda)\cdot g_z(z), \tag{2.13}$$

say. We observe that, by section 2.2.2, z is determined by the subsidiary model and is independent of u, the random disturbances of the considered model. The parameters of the ("subsidiary") marginal distribution $g_z(z)$ are not related to the parameters λ of the other factor in (2.13), the conditional distribution of the endogenous variables

$$g(y \mid z\,;\,\lambda). \tag{2.14}$$

This distribution is determined by the structure S, as is seen by solving the structural equations in (2.11) for y. The *reduced form*,

$$y = \bar{\eta}_{\varphi}[z,\ u;\ \bar{\pi}_{\varphi}(\alpha)], \tag{2.15}$$

differs from that of the nonstochastic case (1.8) inasmuch as it involves the random vector u. Given z, the value taken by $u \equiv (u_1, \ldots, u_G)$ determines the value taken by $y \equiv (y_1, \ldots, y_G)$. The distribution of y depends on the elements of the structure $S \equiv (f, \varphi, \alpha, \varepsilon)$; this can be expressed thus:

$$g(y \mid z;\ \lambda) = g_{f,\varphi}(y \mid z;\ \lambda), \tag{2.16}$$

$$\lambda = \bar{\lambda}(\alpha,\ \varepsilon) = \bar{\lambda}_{f,\varphi}(\alpha,\ \varepsilon) = \bar{\lambda}(S). \tag{2.17}$$

Thus the structure S determines the form and the parameters of the conditional distribution $g(y \mid z;\ \lambda)$.

On the other hand, S does not determine the distribution $g_z(z)$, the other factor in the product in (2.13). For the purpose of estimating S from observations, we have to consider not the joint distribution $g_x(y,z)$ but merely the conditional distribution $g(y|z)$: that is, we can disregard the possibly random character of z, and treat z as if it were fixed in repeated samples of y.

2.2.4. The reduced form (2.15) is a system of stochastic equations. In the nonstochastic case, the prediction of y from z involved knowledge of π, the parameters of the reduced form. In the stochastic case, the prediction of y consists in estimating the parameters (mean, variance, etc.) of the conditional distribution of y, given z. That is, λ, the parameters of the conditional distribution $g(y|z;\ \lambda)$ are to be estimated. Now consider (2.17), and let

$$\lambda^0 = \bar{\lambda}(S^0) = \bar{\lambda}(\alpha^0,\ \varepsilon^0) \tag{2.18}$$

be the parameters of the conditional distribution during the observation period, that is, of the distribution

$$g^0[y(t) \mid z(t);\ \lambda^0], \qquad t = 1, \ldots, T. \tag{2.19}$$

It is possible to compute from observations an estimate l^0 of the vector λ^0 by an operation

$$\mathfrak{P}_1 X^o = l^o; \tag{2.20}$$

operation $\mathfrak{P}_1$ is called *predictive estimation when structure is unchanged.* The notation is analogous to but not identical with that of (1.13) of the nonstochastic case. There we dealt with predictive determination of the reduced form, here with predictive estimation of the distribution of observables.

2.2.5. If changes of structure are intended or expected, the operation $\mathfrak{P}_1$ will, in general, not suffice for the prediction of y from z.

Suppose it is possible to compute estimates (a^o, e^o) of the parameter vectors $(\alpha^o, \varepsilon^o)$ of the observational structure S^o. Call this operation $\mathfrak{E}$:

$$\text{est } S^o \equiv (a^o, e^o) = \mathfrak{E} X^o. \tag{2.21}$$

$\mathfrak{E}$ thus denotes *structural estimation* (analogous to the structural determination of section 1.2.3.6). Now let the observational structure undergo a change $\mathfrak{I}$: i.e., $\mathfrak{I} S^o = S$. The estimates l of the parameters λ of the new distribution of observables are obtained by the operation

$$\begin{aligned} l = \bar{\lambda}(a,e) &= \bar{l}[\mathfrak{I}(a^o, e^o)] \\ &= \bar{\lambda}(\mathfrak{I}\,\mathfrak{E} X^o) = \mathfrak{P}_{\mathfrak{I}} X^o, \end{aligned} \tag{2.22}$$

say; the operation $\mathfrak{P}_{\mathfrak{I}}$ is called *predictive estimation when structure undergoes a given change* $\mathfrak{I}$.

Whether optimal properties of the estimates a^o, e^o are preserved under subsequent transformations $\mathfrak{I}$, $\bar{\lambda}$, depends on the estimating methods; in particular, the function ψ of a maximum-likelihood estimate of a parameter θ is the maximum-likelihood estimate of the function $\psi(\theta)$.

2.2.6. One can interpret the operations (2.20) and (2.22) more generally. l^o and l may be used to denote, not the point estimates of the parameters λ^o and λ, but the estimates of the parameters of the joint distribution of those estimates or the estimates of their confidence regions. However, this interpretation need not be applied in the present article.

2.2.7. Structural estimation is needed, in particular, when

structural changes are intended, i.e., when *structural policies* are being undertaken. This was shown in section *1.2.3.5* with regard to structural determination when the model was nonstochastic. Section *2.5.1.3* illustrates the stochastic case.

We have to modify the earlier treatment since y and consequently the gain (welfare) ω as previously defined are random variables. For example, the real income (of a firm or of a nation, as the case may be) is the quotient of money income over price level, i.e., a function of endogenous variables y. Real income therefore depends on the value of exogenous variables z and on the distribution $g(y \mid z\,;\,\lambda)$, and is a random variable. We must, accordingly, redefine the gain (welfare) ω, i.e., the quantity which the policy-maker tries to maximize. The policy-maker will prefer a high mean value of real income to a low mean value; he may possibly at the same time prefer a small variance of income to a large variance. In short, he will prefer certain probability distributions of income to others. He will maximize a quantity ω that is a functional of the probability distribution function of income.[1] But the probability distribution of income depends on the value of the exogenous variables z and on the distribution $g(y \mid z\,;\,\lambda)$. Since the form of g is fixed a priori (depending on the function forms φ, f of the model), the gain ω is a known function of the distribution parameters λ and of z only. By (2.17) the parameters λ depend on the structure S in a way uniquely determined by the model. The structure S is, in turn, (analogously to section *1.2*) the result of controlled or uncontrolled changes applied to the observational structure. Thus

$$\lambda = \bar{\lambda}(S) = \bar{\lambda}(\mathfrak{I}_c \mathfrak{I}_u S^o). \tag{2.23}$$

Furthermore, the exogenous variables can be either controlled (z_c) or uncontrolled (z_u) by the policy-maker. His aim is thus to get

$$\hat{\omega} = \omega(\hat{z}_c,\ z_u,\ \hat{\mathfrak{I}}_c \mathfrak{I}_u S^o), \tag{2.24}$$

where ω is a known gain function, and $\hat{z}_c$, $\hat{\mathfrak{I}}_c$ are, respectively, the optimal values of controllable exogenous variables and of

[1] Compare [Hurwicz, 1946, p.132], [Tintner, 1942]. This statement can be shown to be implied in the statement that the mean value of utility of real income is maximized.

controllable structural changes. That is, $\hat{\omega}$ is a maximum of ω with respect to policies $(z_c, \mathfrak{I}_c)$.

To estimate the best policy $(\hat{z}_c, \hat{\mathfrak{I}}_c)$, for a given set $(z_u, \mathfrak{I}_u)$, of uncontrolled exogenous variables and uncontrolled structural changes, one has to have an estimate of the observational structure S^o, i.e., to have the estimates (a^o, e^o) of the observational structural parameters $(\alpha^o, \varepsilon^o)$.

In practice, the maximization of ω with respect to policies consists in estimating for each considered policy $(z_c, \mathfrak{I}_c)$, the parameters λ of the conditional distribution of y given z, and the consequent probability distribution of (say) real income. The policy which gives the probability distribution that is preferable to all others is then chosen.

2.2.8. *Relation between regression, reduced form, and structural equations.* Sometimes the series of alternative policies considered will be such as to make it unnecessary to know the joint conditional G-variate distribution (2.14) of all endogenous variables after the change in structure. It may suffice to know the distribution of some particular subset, say $y_{(d)}$ of y, that is, the distribution

$$g_{(d)}(y_{(d)}, z; \lambda_{(d)}). \tag{2.25}$$

Such conditional distribution is easily derived from the general conditional distribution (2.14).[1] The set $y_{(d)}$ is called the *predictand;* the *predictor* is, in this case, the whole exogenous set z. Practical importance may be attached in particular to the simplest case, viz., the univariate conditional distributions

$$g_1(y_1 \mid z; \lambda_1), \quad \ldots, \quad g_G(y_G \mid z; \lambda_G). \tag{2.26}$$

In particular, one may be interested in the first moments of such distributions, e.g., in the expectation

$$\mathcal{E}(y_1 \mid z) = R(z; \chi_{(1)}). \tag{2.27}$$

R is the regression function of y_1 on z. Its parameters, denoted here by the vector $\chi_{(1)}$, are called regression coefficients if R happens to be linear, so that, writing $\chi_{(1)} \equiv (\chi_{10}, \chi_{11}, \cdots, \chi_{1K})$,

[1] See [VI].

$$\mathcal{E}(y_1 \mid z) = \sum_1^K \chi_{1k} z_k + \chi_{10}. \tag{2.28}$$

A sufficient condition for the function R in (2.27) to be linear is that the functions φ, f of the model be linear and normal, respectively. We can then conveniently split the structural coefficients α into two subsets $\beta \equiv \{\beta_{gh}\}$ and $\gamma \equiv \{\gamma_{gk}\}$, $(g,h = 1, \ldots, G;$ $k = 0, \ldots, K)$, denoting respectively the coefficients of endogenous and of exogenous variables; further we can choose the constant terms γ_{g0} so as to make $\mathcal{E}u = 0$. The model (2.11) becomes:

$$\sum_{h=1}^G \beta_{gh} y_h + \sum_{k=1}^K \gamma_{gk} z_k + \gamma_{g0} = u_g, \quad g = 1, \ldots, G, \tag{2.29}$$

$$f(u),$$

where $f(u)$ is normal. In this case the reduced form (2.15) will also be linear; for example the first equation of the reduced form will be

$$y_1 = \sum_1^K \pi_{1k} z_k + \pi_{10} + \text{a linear function of } u. \tag{2.30}$$

Taking the expectation of y_1 given z, we find that the regression coefficients $\chi_{(1)}$ and the coefficients $\pi_{(1)}$ of the reduced form coincide: see (2.28).

Furthermore, suppose our complete model consists of one equation only $(G=1)$; if we choose the units of the z's so as to make $\beta_{11} = 1$, we can write the unique structural equation as

$$y_1 - \sum_{k=0}^K \gamma_{1k} z_k - \gamma_{10} = u_1. \tag{2.31}$$

In this particular case the coefficients π_1 of the reduced form will coincide not only with the regression coefficients $\chi_{(1)}$ but also with the structural coefficients $\gamma_{(1)}$. [This applies, of course, also to the case where a complete model involving $G > 1$ variables can be split into G uniequational complete models such as (2.31); in each of which, one endogenous variable depends on

exogenous variables only.] In general, however, there will be no such coincidence. An example will be given in section 2.5.1.2.

2.2.9. Predicting exogenous variables. If the conditional distribution (2.19) estimated from past observations is known to remain valid for $t > T$, this distribution, or any of its practically relevant conditional distributions (2.25) can be used, respectively, as the future distribution of y, or of a subset $y_{(d)}$, for given future values of z.

However, it may also be desirable to make predictions about exogenous variables, in particular the uncontrolled ones, z_u. It is seen from section 2.2.7 that the choice of best policy presupposes the knowledge of uncontrolled factors: z_u as well as $\mathfrak{I}_u$. It is possible to estimate future values of the variables z_u if they are related to observable variables that do not enter the model considered. For example, suppose prediction is done on behalf of a firm to help it in the choice of its policies. The firm's sales is an endogenous variable in the model describing the demand and production conditions for the firm's product. This model contains national income as an exogenous variable which the firm cannot control. The national income itself is an endogenous variable of another model; this may also include exogenous variables such as foreign crops (which affect the demand for this country's exports of manufactured goods, and hence affect this country's income). Thus the firm will be interested in predicting national income, using distributions such as (2.26), where y_1 would denote national income, and z will include foreign crops z_1. However, another variable endogenous to the national economy model – e.g., the imports y_2 – can be usefully included in the predictor set, if, in the future, information on national income becomes available later than information on imports. The prediction of y_1 for a given z and a specified value of y_2 is more accurate than the prediction of y_1 from z with the value of y_2 unspecified. Thus, it may be useful to derive from the observational distribution $g^o(y \mid z)$ a distribution

$$g^o_{(d,r)}(y_{(d)} \mid y_{(r)}, z), \tag{2.32}$$

say; this distribution is more general than (2.25) in that its predictor set (y_r, z) includes endogenous variables, $y_{(r)}$. In particular, we may derive a regression equation to estimate

(2.33) $$\mathcal{E}(y_1 \mid y_2, z).$$

We may also want to estimate a variable exogenous not only to the relevant model (e.g., the model of the firm) but also to the model (e.g., of the national economy) from the variables of which it has to be estimated. For example, foreign crops z_1 may be of direct interest to the firm and also may influence national imports y_2; but observations on factors determining foreign crops may be unavailable. If data on y_2 become known earlier than those on z_1, z_1 can be estimated from y_2. This case differs from the preceding one inasmuch as z_1 may be a nonstochastic variable. However, it may be possible to use the past distribution $g^o(y \mid z; \lambda^o)$ of the observables to derive a confidence region for z_1 (looked upon as a parameter): values of z_1 such as would give rise to the known values of y_2 only with a small probability (lower than a preassigned significance level) will lie outside the confidence limits for z_1.

2.2.9.1. These remarks are of some importance in view of many attempts to use endogenous variables as predictors; e.g., to predict national income from contemporaneous imports, or retail sales, etc. It may be possible to determine such a relation from past observations. But its usefulness is limited to the cases just mentioned. To choose between various national policies, it cannot be useful to be able to predict national income from imports, since the latter are themselves affected by any policy chosen. For purposes of private policies, on the other hand, such prediction can be useful. But such prediction is possible only inasmuch as information on imports is available earlier than that on national income; and only if, in addition, the relevant aspects of the national economic structure can be assumed unchanged between the period of observation and the time when the private policy is going to be applied. The same considerations are valid for the case when, because of a lag in the available data, variables endogenous to the national economy are used to estimate exogenous ones.

2.3. *Identification*

2.3.1. In section *1.3* it was shown for nonstochastic models

that structure can be determined to the extent that the a priori conditions (i.e., the model) suffice to make the structure uniquely identifiable. In this case, only one structure will correspond to a given reduced form. Or, in terms of parameters: if only one set α^o corresponds to a given set π^o, then, since π^o can always be determined from (sufficiently numerous) observations, the set α^o can also be so determined.

In the stochastic case, a structure is said to be uniquely identifiable by the model if only one structure compatible with the model corresponds to a given conditional distribution of the observations. Or, in terms of parameters: if only one set α^o, ε^o, corresponds to a given set λ^o, then, since λ^o can be estimated from the observations, the structural parameters α^o, ε^o can also be so estimated.

2.3.2. The structural transformations $\mathfrak{I}_c$, $\mathfrak{I}_u$ and the gain functional ω may be such as to make identifiability of all parameters α^o, ε^o unnecessary for the choice of best policy. Section *1.3.10* on partial identifiability applies accordingly.

2.3.3. If the structure, or a part of it, is not identifiable its estimation is not possible, however numerous the observations on the variables treated as observables in the model. However, observations on other variables may provide additional information (which is a priori with respect to the structure considered; see section *1.1.0*) such as to make the structure, or its relevant part, identifiable. The failure of a model to identify the structure is not a ground for rejecting the model; rather, it calls for additional information, to be provided by a new type of observations. Suppose, for example, that the relation describing the investment behavior of the aggregate of American firms is not identifiable within a model that involves, in addition to this relation, relations describing the behavior of consumers, lenders, etc. This would make it necessary to add a new type of information, based, e.g., on records of single firms, or on interviews with businessmen.

2.3.4. The a priori information provided by the model may involve the parameters α of the structural equations as well as the parameters ε of the shock distribution. If no a priori information on the parameters ε of the shock distribution exists, and if a nonstochastic structure

$$\varphi_g(y,\ z\ ;\ \alpha_{(g)}) = 0, \qquad g = 1,\ \ldots,\ G, \tag{2.34}$$

is not completely identifiable, then the parameters α of the stochastic structure

$$\varphi_g(y,\ z;\ \alpha_{(g)}) = u_g, \qquad f(u\ ;\ \varepsilon), \qquad g = 1,\ \ldots,\ G, \tag{2.35}$$

will also not be completely identifiable; but information on the distribution $f(u;\varepsilon)$, if available, may make (2.35) completely identifiable. Thus the example in section *1.3.5* may or may not hold for the stochastic case if information on the shock distribution is available: such as the knowledge that the shocks on the demand side are not correlated with those on the supply side; or knowledge of the ratio between the variances, etc.: [Frisch, 1933], [Mann and Wald, esp. p.219], [Marschak and Andrews, §§ 18 - 22], and sections 2.5.*1.1*, 2.5.*1.4* below.

2.3.5. In particular, a uniequational complete model is completely identifiable, apart from trivial transformations (section *1.3.7*). Hence the remarks in section *1.4.2*, on the role of experiments (of type II) in making the observational structure identifiable, will apply. If, in addition (section *2.2.8*), the uniequational complete model generated by an experiment is linear with shocks distributed normally, the structural coefficients will not only be identifiable but will coincide with the regression coefficients. The model can then also be regarded as one in which one (the "dependent") variable is subject to measurement errors while all others are free from such errors: a familiar case in the history of the application of statistics to experiments. It arises in nonexperimental science only if the mechanism producing the observations can be adequately represented by a model involving only one nonlagged endogenous variable.

2.4. *Dynamic Models*

2.4.1. A model is called dynamic if it has at least one of the following two properties: 1) at least one observable variable occurs in the structural equations with values taken at various points of time (this includes the case of time derivatives, differences, and integrals over time); 2) at least one equation contains functions of time (trend, seasonal fluctuations, etc.). If the first property is present the model is called *multitemporal*; if both properties are absent, it is called *unitemporal*.

2.4.2. We obtain a *discrete multitemporal* shock model if the equations in (2.11) are replaced by

$$\varphi_g[y'(t), \ldots, y'(t-\tau_{y'}), z'(t), \ldots, z'(t-\tau_{z'}); \alpha_{(g)}] = u_g(t), \tag{2.36}$$

and the distribution of shocks in (2.11) by

$$f[u(t), \ldots, u(t-\tau_{y'}); \varepsilon]. \tag{2.37}$$

The time interval between two successive observations is chosen as a time unit; time lags smaller than 1 are not admitted by the model; $\tau_{y'}$ and $\tau_{z'}$ denote the largest time lag with which the corresponding variables occur; and t takes all integral values through the time interval during which the model is supposed to be valid.

2.4.3. A case of great practical importance arises when successive shocks are mutually independent; that is,

$$\begin{aligned} &f[u(t), u(t-1), \ldots, u(t-\tau_{y'})] \\ &\quad = f[u(t)] \cdot f[u(t-1)] \cdot \cdots \cdot f[u(t-\tau_{y'})], \end{aligned} \tag{2.38}$$

say. If this condition is fulfilled then, as shown in [I-1] and [XVII], lagged variables can be treated as additional variables in judging the identifiability of a structure. Lagged endogenous variables can then be treated as if they were fixed in repeated samples (i.e., like exogenous variables). The following model can then be regarded as a complete one:

$$\varphi_g[y(t), z(t); \alpha_{(g)}] = u_g(t), \quad g = 1, \ldots, G, \tag{2.39}$$

$$f[u(t); \varepsilon], \tag{2.40}$$

where the notation in (2.36) has been changed as follows:

$$\begin{aligned} &y'(t) \equiv y(t) \equiv \text{"jointly dependent variables,"} \\ &[y'(t-1), \ldots, y'(t-\tau_{y'}), z'(t), \ldots, z'(t-\tau_{z'})] \equiv z(t) \\ &\qquad \equiv \text{"predetermined variables,"} \end{aligned} \tag{2.41}$$

following Koopmans' terminology [XVII].

2.4.4. It is plausible to assume, as in (2.38), the statistical independence of successive shocks if the time interval between two successive observations is not too short. It would be more realistic to study models in which at least some of the observables $y(t)$, $z(t)$, and also some of the shocks $u(t)$ are functions of a continuous time variable. Some properties of such continuous stochastic models are outlined in [XVI]. However, the remainder of this volume deals with discrete models and most of the time assumes condition (2.38) to be valid.

2.4.5. The reduced form defined in (2.15) applies to the multitemporal model with variables $y(t)$ and $z(t)$ defined by (2.41). The variables on the right-hand side of the reduced form (the predictor set) occur with lags no higher than those in the structural equations. However, other kinds of predictor sets can be considered. In the *separated form* each nonlagged variable is expressed as depending only on its own lagged values and on the exogenous variables with or without lags; this is possible because, by "shifting time back," one can obtain enough structural equations to eliminate the lagged values of other endogenous variables. One can go further and eliminate the lagged values of all endogenous variables, leaving only the exogenous variables (with and without lags) as the predictor set: the *resolved form*.

As shown by Tinbergen, the separated form (his "final equation") can be used to build up, year by year, the path of a single endogenous variable beginning with given initial values and giving effect every year to changes in exogenous variables; that is, the solution of each equation of the reduced form (looked upon as a difference equation) expresses the predictand variable as a function of time ("cyclical fluctuations") and of exogenous variables.

Corresponding to the reduced, the separated, and the resolved forms in which a multitemporal model can be written, there are three kinds of distributions with which predictive estimation under unchanged structure can be concerned, and which can be estimated (in principle) regardless of whether the structure is or is not identifiable. However, as will be shown below (section 2.5.3), only in the case of reduced form, but not in the case of separated and of resolved form, is the estimation amenable to known methods.

2.5. *Estimation*

2.5.0. Before we summarize the results obtained in this

volume with regard to the estimation of relevant parameters from finite samples, we shall restate and illustrate certain population properties. We shall then discuss the estimation of regression equations, and of complete and incomplete structures, and add a few remarks on the choice of models.

2.5.1. Population properties. Important differences between, on the one hand, the structural parameters α, and, on the other, the parameters π of the reduced form and the regression parameters χ, will be briefly restated. (It is understood that the parameters to be treated here are generated by the observational structure S^o; but the superscript ^{o} is omitted from the symbols where no ambiguity arises.) To fix the ideas we shall assume the model to be linear with normally distributed shocks and shall give examples when certain elements of χ or π are or are not equal to some elements of α.

While χ and π can always be estimated from observations, the same is not true of the structural parameters α unless they are identifiable. Furthermore (section 2.2.8) the coefficients $\pi_{(1)}$, of that equation of the reduced form which determines the endogenous variable y_1, are equal to the coefficients $\chi_{(1)}$ of the corresponding regression equation. If there exists a structural equation relating y_1 to the exogenous subset z and containing no endogenous variables, its coefficients $\alpha_{(1)}$ will also be equal to $\chi_{(1)}$; provided this structural equation, if stated together with the distribution – $f_1(u_1)$, say – of its shock-variable, constitutes a complete model. The latter condition implies (section 2.2.2) that u_1 is distributed independently of the shock variables $u_2, \ldots, u_G$ entering the remaining structural equations. If the structural equation considered contains lagged endogenous variables, and if successive shocks are independent, the statements just made can be extended so as to include in z not only the exogenous but also the lagged endogenous variables (section 2.4.3).

For example, structural and regression coefficients will coincide in each of the following complete models (each is assumed to be linear, with successive shocks independent and normally distributed; the choice of the "dependent" variable in the regression will be indicated in each case by the phrase describing the model). Of these four models, the first three are uniequational, while the last one can be partitioned into two uniequational complete models.[1]

[1] A more complicated model partitionable into four uniequational models is given in [Bentzel and Wold, p.104].

(1) Current yield per acre depends on current temperature and current rainfall.

(2) Current yield per acre depends on current temperature and rainfall, and also on the amount of fertilizers fixed by the experimenter.

(3) Current yield per acre depends on current and past temperature and rainfall, and on the decision farmers made in the previous year regarding the amount of fertilizers.

(4) Market price of a nonstorable good depends on supply (demand equation); supply depends on previous year's price (supply equation); the behavior of buyers and that of sellers undergo mutually independent shocks.

2.5.1.1. In the following example[1] on the other hand, each structural equation contains more than one endogenous variable, and cannot therefore (taken together with the distribution of its shock variable) be regarded as a complete model in itself; but the two equations together constitute a complete model. We shall show that the structural coefficients will not be equal to any of the regression coefficients. Let y_1 (national product, identical with national income, or the supply of all goods) and y_2 (demand for all goods) obey the following equations for any time t:

$$y_2(t) - \beta_1 y_1(t) - \beta_0 = u_1(t) \quad \text{(behavior of buyers)},$$

$$y_2(t) - y_1(t) = u_2(t) \quad \text{(behavior of producers)}.$$

The model does not contain predetermined variables. Its random shocks are $u_1(t)$ ("shift of demand") and $u_2(t)$ ("failure to adjust production to demand"). Each pair of values $u_1(t), u_2(t)$ determines

$$y_1(t) = [u_1(t) - u_2(t) + \beta_0] / (1 - \beta_1),$$

$$y_2(t) = [u_1(t) - \beta_1 u_2(t) + \beta_0] / (1 - \beta_1).$$

Suppose successive shocks are independent and the joint distribution $f[u_1(t), u_2(t)]$ of the shocks is normal and independent of time.

[1] For more fully developed examples from the same branch of economics see [Haavelmo, 1947 A].

Let its moments be

$$(2.42) \qquad \mathcal{E}\, u_1(t)\, u_2(t) = 0; \qquad \mathcal{E}\, u_1^2(t) = \sigma_{11}; \qquad \mathcal{E}\, u_2^2(t) = \sigma_{22}.$$

Then the distribution $g(y_1, y_2)$ of the observables is normal and has moments

$$(2.43) \qquad \begin{aligned} \mathcal{E}\, y_1 \equiv \mathcal{E}\, y_2 &= \beta_0 / (1-\beta_1); \quad \sigma_{y_1 y_2} = (\sigma_{11} + \beta_1 \sigma_{22}) / (1-\beta_1)^2; \\ \sigma_{y_1 y_1} &= (\sigma_{11} + \sigma_{22}) / (1-\beta_1)^2; \quad \sigma_{y_2 y_2} = (\sigma_{11} + \beta_1^2 \sigma_{22}) / (1-\beta_1)^2. \end{aligned}$$

These are four mutually independent equations (not counting the identity $\mathcal{E} y_1 \equiv \mathcal{E} y_2$) to determine the four unknown structural parameters $S = (\beta_1, \beta_0, \sigma_{11}, \sigma_{22})$ from the parameters of the distribution $g(y_1, y_2)$ of the observables. The structural parameters are completely identifiable.

2.5.1.2. Consider now the regression of y_2 on y_1:

$$\mathcal{E}(y_2 \mid y_1) = \chi_1 y_1 + \chi_0,$$

say, and

$$(2.44) \qquad \chi_1 = \sigma_{y_1 y_2} / \sigma_{y_1 y_1} = (\sigma_{11} + \beta_1 \sigma_{22}) / (\sigma_{11} + \sigma_{22}),$$

which approaches β_1 or 1 as $\sigma_{11} / \sigma_{22} \to 0$ or $\to \infty$, respectively. Thus β_1, the "marginal propensity to spend," is distinct from the regression coefficient χ_1 of spending (y_2) on income (y_1), except in a limiting case when the behavior of buyers is not subject to random shocks (while the behavior of producers is).

2.5.1.3. This example also illustrates the different practical purposes of estimating β_1, β_0 or χ_1, χ_0. Suppose a firm expects a given change in the economic structure; for example, a rise in the "marginal propensity to spend" whereby β_1 will be replaced by $k\beta_1$; suppose the other three structural parameters are known to stay unchanged. If the firm wants to use old observations on y_1, y_2 to predict the distribution of these variables under the new circumstances, it will have first to obtain β_1, β_0, σ_{11}, σ_{22} from observations, and then insert the new set of structural parameters

$(k\beta_1, \beta_0, \sigma_{11}, \sigma_{22})$ into (2.43). Furthermore, if the firm wants to estimate national spending from national income (because the latter is published earlier than the former; section 2.2.9.1), it will have to use, not the old regression coefficient (2.44), but a new regression coefficient, $(\sigma_{11} + k\beta_1\sigma_{22})/(\sigma_{11} + \sigma_{22})$. This cannot be obtained from χ_1, but can be obtained from the old structural parameters $(\beta_1, \beta_0, \sigma_{11}, \sigma_{22})$ since k is known. Only if the structure is known to remain unchanged can χ_1 be used.

2.5.1.4. If the structure were not identifiable, there would be no way to predict effects of known structural changes. This would be the case, for example, if the a priori assumption – in (2.42) – of noncorrelated shocks could not be admitted. We should then have five structural parameters to determine; the four equations (2.43) would therefore not suffice. On the other hand, the structure would be identifiable, even with the noncorrelation assumption dropped, if a different predetermined variable had been introduced into each of the two equations of the model.

2.5.2. *Estimation of regression coefficients.* When structure is known to remain unchanged, estimates of regression coefficients χ, together with other parameters of the distribution of observables, help to estimate the future distribution of predictand variables for given values of others, and hence to estimate future positions of maximum gain. When, on the other hand, structure is known to undergo a given change, the estimation of future values of variables requires knowledge of the structural parameters (α, ε). We have seen (section 2.2.8) that in the special case of certain uniequational complete models (and of models that can be decomposed into such uniequational complete models), $\alpha = \chi$. In this case, if $q = q(X^o)$ is a function of observations X^o that is an unbiased estimate of the parameters χ, then $\mathcal{E}q = \chi = \alpha$; that is, q can serve as an unbiased estimate of structural parameters. In general, however, there is no equality between α and χ. Therefore a function of observations that is an unbiased estimate of χ cannot be an unbiased estimate of α, even for infinitely large samples. If such a function is used as an estimate of α for predictive estimation under a structure subjected to given changes, the future endogenous variables, and hence the future gains, will be estimated with a bias; consequently, other than optimal policies will be chosen.

This does not make the estimation of regression coefficients useless: uniequational complete models of the appropriate type may

exist; and structures do not always change. Hence our interest in reconsidering the properties of regression coefficients.

In a general multitemporal model, the variables are connected by a set of equations

$$\varphi_g[y_1(t), \ \ldots, \ y_1(t-\tau_1), \ y_2(t), \ \ldots, \ y_G(t-\tau_G), \ z] = u_g,$$

$$g = 1, \ \ldots, \ G,$$

where z denotes exogenous variables (lagged as well as nonlagged). The regression equation for y_1 – denoting by $(y_{(\tau)}, \ z)$ the whole predictor set – is

$$\mathcal{E}\,[y_1(t) \mid y_{(\tau)}, \ z] = R(y_{(\tau)}, \ z),$$

which has not been studied in general. The simplest case is that of the linear regression equation with $y_{(\tau)}$ empty:

$$\mathcal{E}\,[y_1(t) \mid z] = \sum_1^K \chi_{1k} z_k + \chi_{10}\,; \tag{2.45}$$

it is satisfied when $y_1(t)$ and z are connected by an uniequational unitemporal model, cf. (2.31):

$$y_1(t) - \sum_1^K \gamma_{1k} z_k - \gamma_{10} = u_1(t), \tag{2.46}$$

provided successive shocks $u_1(t)$, $u_1(t+1)$, ..., are normally and independently distributed. Certain optimal properties of the least-squares estimate of χ $(=\gamma)$ are well established for this case. The case has to be generalized in two important directions (possibly, but not necessarily, preserving the assumptions of linearity and normality of the structure): 1) the complete model may be made multiequational; 2) the complete model may be made multitemporal.

In the case of multiequational (but unitemporal) model, the properties of the least-squares estimate and the maximum-likelihood estimate of the regression coefficients, and in particular the conditions of equivalence of these two estimates, have been studied in [VI] and [VII] by Hurwicz and Koopmans for large as well as small

samples.

For the case of a uniequational multitemporal model with only one lag,

$$y_1(t) = \alpha y_1(t-1) + u_1(t), \tag{2.47}$$

where $u_1(t)$ is distributed normally with a constant variance and with zero covariances $\mathcal{E}u_1(t)u_1(t')$, $t \neq t'$. Hurwicz [XV] shows that the least-squares estimate of the regression (and structural) coefficient α has a bias. For $T = 20$ (a length of time series common to economic studies based on annual data of the interwar period), the bias approaches 9 per cent as $\alpha \to 0$. The bias seems to disappear as the sample increases or as $\alpha \to 1$.

2.5.3. This bias in the regression estimates from short time series will in general exist in every multitemporal model. If the model is multiequational, there arises, in addition, the question of the *choice of predictors*. Each predictand variable can be expressed as a function of the predictor variables and of the shock variables as in (2.15) and (2.30). Only contemporary shock variables will be involved provided the predictor variables occur with the same lags as those of the structural equations. Such is the case with the reduced form (see section 2.4.5). The predictand will then be related to the predictor by an equation such as (2.30), where the z's would stand for exogenous as well as lagged endogenous variables; and where the successive random terms will be independent. In this case, least-squares large-sample estimates of the coefficients of the variables will have the usual optimal properties. But it is different in the case of the "resolved" and the "separated" forms (section 2.4.5). In these cases, structural shocks relating to various points of time will be contained in the same equation, as the result of eliminating certain endogenous variables after replacing t by $t-1$, $t-2$, ..., or "shifting the time back." Therefore successive random terms of a "resolved" or a "separated" form are not independent; and the least-squares estimates of corresponding regression equations will, in general, be biased.[1]

2.5.4. As stated in section 2.3.2, expected or intended structural changes may be of such a nature that the estimation of all the parameters of a complete model is not necessary for purposes of

[1] Pointed out by Haavelmo. See [Klein, 1946 B, p.303 ff.].

prediction and policy: hence the practical importance of *partial structural estimation*, i.e., the estimation of a selected set of parameters. Such estimation is possible if the set of parameters in question is identifiable, even though some or all of the remaining parameters may not be identifiable. In particular, the estimation of the parameters of a few selected structural equations, or even of a single structural equation (e.g., demand equation) has an obvious practical interest.

On the other hand, the expected or intended changes of structure may not be known long beforehand. It is therefore often desirable to have estimated the structure completely.

2.5.5. The method of *complete structural estimation* most fully discussed in the volume is that of maximum likelihood. The joint probability density of all the observed values of the variables is regarded as a function (the likelihood function) of the structural parameters. Those values of the parameters for which this function attains its highest value are called maximum-likelihood estimates. In important classes of cases these estimates are "consistent" (they converge with probability 1 to the true values in the limit for infinitely large samples) and "efficient" (they have, in large samples, variances that never exceed those of any other normally distributed estimate). In [Mann and Wald] are discussed the maximum-likelihood estimates obtained when all available a priori information is used and when the model is as follows: a completely identifiable complete system of linear difference equations with no exogenous variables; shocks mutually independent; the model generates a stationary process, i.e., the observable variables would converge in time to constant values if shocks were absent. These authors proved the consistency (but not the efficiency) of the estimates under these conditions; they also showed that, in the absence of a priori restrictions, the structure is not identifiable. In [II-3] Koopmans and Rubin estimate the parameters of a complete linear model that is more general owing to the introduction of exogenous variables. Their discussion also covers the estimation of the parameters of some identifiable structural equations if other equations of the complete system are not identifiable. They give a proof of the consistency of the estimates which takes account of these two generalizing assumptions. In [XIV] Rubin extends the proof to a simple case of a nonstationary ("explosive") process.

There is an alternative maximum-likelihood method for obtaining consistent estimates of structural parameters: each equation of the structure is estimated separately via the reduced form as

described in section 2.5.6. For each equation this procedure has to use at least as much a priori information as is necessary to make that equation identifiable; but it leaves unused a (possibly large) part of the remaining a priori information. This method will therefore lead, in general, to less efficient estimates (i.e., larger sampling variances of the estimates) than the maximum-likelihood method using all information described above, although both kinds of estimates have the consistency property. For brevity, we call the two estimation methods "information-preserving maximum-likelihood estimation" and "limited-information maximum-likelihood estimation," respectively. The latter method is also known as the "method of reduced forms."

The practical usefulness of the consistency property of estimates diminishes if the sample becomes small. Small sample properties of structural estimates have not been studied, except [XV] for the uniequational multitemporal model, equation (2.47) of the present article. The bias found in the estimate in this case suggests that in general both proposed methods of estimating multitemporal structures are biased if applied to short time series.

2.5.6. Incomplete (partial) structural estimation. A complete model (and structure) was defined above, in sections *1.1.4, 1.2.1.1,* and *2.2.2.* Obviously the *limited-information maximum-likelihood* method just described can be used to estimate complete as well as incomplete structures. At the time when this introduction is being written this method appears to be best developed. But other suggestions for estimating incomplete structures will also be discussed below (section *2.5.6.1*) after adding a few more remarks on the reduced forms method.

The coefficients π of a linear reduced form (2.30) are regression coefficients of a nonlagged endogenous variables on exogenous and lagged endogenous variables. Their least-squares estimates have the consistency property. At the same time, they are functions of the structural parameters. This suggests that consistent estimates of structural coefficients can be obtained by applying appropriate transformations to the least-squares estimates of the coefficients of the reduced form. Similarly, parameters (e.g., the variance) of the estimated distribution of the random terms of the reduced form can be transformed into the parameters of the distribution of the shocks in the structural equations. The suggestion has been familiar for some time[1] but has been applied rigorously for the first

[1] See, for example, [Mann and Wald, p.219], [Haavelmo, 1944, pp. 103 - 104].

time by T. W. Anderson, M. A. Girshick, and H. Rubin; their joint work is summarized in [IX]. Since an equation of the reduced form has as many unknown coefficients as there are predetermined variables in the system, it will, in general, have more parameters than the structural equations (or equation) to be estimated. Hence equations connecting the unknown estimates of structural parameters with estimates of parameters of the reduced form may be more numerous than the number of these unknowns. To avoid this overdeterminacy, part of the available information has to be dropped. In particular, the method of reduced forms, which has been applied so far only to one equation at a time – in [Girshick and Haavelmo], [Haavelmo, 1947 A], [Klein, 1947, 1950] – does not use the observations on jointly dependent variables outside of this equation, though it does use the observations on predetermined variables of the system. The only a priori information this method uses are the linear restrictions on the parameters of the equation in question (including the prescription as to which variables enter this equation).

2.5.6.1. Other procedures of incomplete structural estimation were suggested by Koopmans [1945] and by Wald [VIII]. To estimate the parameters of F $(<G)$ structural equations, Koopmans proposed the following approximation method. The $G-F$ "complementary" equations of the model are "sketched in," e.g., by using admittedly biased single-equation least-squares estimates or some a priori guesses. Then proceed with the estimation of the parameters of the F equations to be estimated. The estimates of the remaining parameters of the complete model will then be improved compared with what they would be if the complementary part of the model were entirely neglected.

Wald's suggestion, [VIII], is different: even if we do not know anything about the complementary part of the model, our a priori knowledge about the F equations that interest the investigator may be sufficient to exclude hosts of originally admissible hypotheses about these equations, and thus to construct confidence regions for parameter estimates. For example, if we know that successive shocks are independent, then a set of values of the structural parameters must be rejected whenever the estimate of shocks computed from observations (the "residuals") fails to have approximately the characteristics of a random series. The elaboration of the method will consist of showing how to use a priori knowledge concerning an incomplete model to construct shortest confidence regions. Essentially, the difference between Koopmans' and Wald's suggestions on the estimation of incomplete models consists in

attaching different weight to our knowledge of the complementary part of the entire model.

The method of reduced forms makes use of the maximum-likelihood principle, since least-squares estimates of regression coefficients are maximum-likelihood estimates (under normally distributed shocks); however, it leaves unused certain types of available a priori knowledge. In Wald's approach to incomplete systems, still fewer known restrictions are used.

For a given number F $(\leq G)$ of equations that are to be estimated, the choice of method will depend on two considerations: the mathematical and computational simplicity on the one hand, and the degree of use of available information on the other. For $F = 1$, the reduced forms method is less laborious than the approach via the estimation of the complete system (Wald's method has not yet been studied in this respect). But the complete estimation has one advantage over the other methods: it utilizes in full the a priori information on the model as well as the observations on all variables of the system.

2.6. *The Choice of Model*

There are many competing sets of a priori restrictions that can be imposed upon the structural parameters without contradicting what we know of human behavior and environment; and there is a wide variety of functional forms to specify the relations and distributions involved. It is often asserted that the choice is much wider in economics than in other empirical sciences. The usual testing considers only one hypothesis (and its negation) at a time. This is an inadequate procedure when a number of hypotheses classifiable according to a large number of attributes are in competition. In this volume no attempt is made to approach this problem.[1] In fact, the present volume is little concerned with the testing of hypotheses. Yet, the following remarks implied in the basic ideas of this volume seem appropriate.

A completely identifiable structure is said to be "just identifiable" by the model if the omission of one of the a priori restrictions of the model makes the structure incompletely identifiable; a completely identifiable structure that is not just identifiable is called overidentifiable. If several alternative overidentifying models are acceptable on a priori grounds, each

[1] See [Wald, 1942, pp. 8 - 9], [Brookner]. This is the problem of "multiple (as distinguished from dual) decisions."

of them (and its negation) can be tested against data by the existing methods; this is, in fact, attempted in [IX-6]. However, new methods are needed to test the whole set of such models simultaneously.

Regarding the great variety of functions equally appropriate, on a priori grounds, to describe structural economic relations, one may expect some help from the statisticians' recent attempts at nonparametric estimation of distribution functions [Wald and Wolfowitz]. Certain weak a priori restrictions on the structural relations, such as the sign of certain partial derivatives, the independence of successive shocks, etc., the economist can assert with better conscience than the restrictions upon, say, the degree of polynomials chosen to describe the structural relations. If confidence limits for joint probability densities of the variables could be estimated on the basis of such weak restrictions, predictive estimation under properly defined structural changes might become possible without introducing stronger but less justifiable hypotheses.

3. THE PLAN OF THE VOLUME

3.1. This volume is concerned with empirical inference in economics. The bulk of its contents is determined by the *stochastic* character of economic models. Their character as *systems of simultaneous equations* requires certain modifications of the usual inference method. The usual method is appropriate when no changes in structure are expected or intended, so that no structural estimation is needed; or when structural estimation is based on experiments which isolate single equations that constitute the complete model by making all variables but one predetermined; or when experiment is used to reproduce the structural change considered (section *1.4* above). The modifications of the inference method that are necessary when none of these conditions is satisfied are treated in Part One: "Simultaneous Equation Systems." The *dynamic* character of economic models calls for modifications in the technique of inference, especially drastic in the treatment of small samples; these modifications are discussed in Part Two: "Problems Specific to Time Series." Part Three, on "Specification of Hypotheses," discusses the construction of models of the type analyzed in the earlier parts of this article, and other more general ones.

3.1.1. In particular, Part One deals with three subjects separately: identification, estimation, and computation. Koopmans, Rubin, and Leipnik treat these problems extensively in [II], and the discussion that follows is organized according to the three headings.

3.1.2. Koopmans and Rubin establish criteria for identification of linear dynamic systems. Their extensive treatment of this especially important case is followed by comments of three authors: Hurwicz, in [IV], provides the logical basis of the problem in its general formulation; Wald, in [III], contributes certain general criteria for identifiability; and Haavelmo, in [V], confronts Ragnar Frisch's technique of bunch maps (devised to establish the presence or absence of multicollinearity in the data of a model based on errors without specified distribution) with the theory of models of simultaneous equations involving shocks with specified distributions.

3.1.3. The logical foundations of structural estimation, compared with predictive estimation under changed or unchanged structure, is given by Hurwicz, in the introductory sections of [VI]. The more technical parts of this paper study the properties of least-squares estimates of regression coefficients when the complete model is multiequational. For one of Hurwicz's results – the equivalence of least-squares and maximum-likelihood estimates of regression coefficients – Koopmans gives an alternative proof in [VII].

The procedure of *complete* structural estimation by the maximum-likelihood method is developed in this volume more fully than the methods of *incomplete (partial)* structural estimation. To the former, the extensive article [II–3] by Koopmans, Rubin, and Leipnik is devoted. Methods of incomplete structural estimation, summarized above in section *2.5.6*, are discussed only briefly: by Wald, [VIII], who has sketched a nonparametric approach, and by T. W. Anderson, [IX], who presents the results obtained in a forthcoming larger article [T. W. Anderson and Rubin, 1949] on the method of reduced forms.

3.1.4. A difficult task, for the first time explored in this volume, is to develop appropriate techniques for the computation of maximum-likelihood estimates that utilize all a priori information contained in the multiequational model. When the model consists of a single equation, the maximization of the likelihood function with respect to structural parameters yields a system of

"normal equations" (familiar from multiple-regression theory) that is linear in the maximum-likelihood estimates: the usual single-equation least-squares method of estimating parameters is a special case of the maximum-likelihood method. When, however, the model has two or more equations, the "normal equations" have to be replaced by more general equations, from which to compute maximum-likelihood estimates of the structural parameters. (If the structure is not uniquely identifiable, these equations will be dependent and the estimates indeterminate in some degree.) The maximum-likelihood equations are nonlinear in the estimates. The direct solution of these equations is therefore troublesome. Koopmans, Rubin, and Leipnik [I-4] have provided iterative methods to solve these equations with any desired deree of approximation. On this problem, much help was derived from the advice of von Neumann and from Hotelling [X]. More recently, the methods involved were discussed and explored from a more general point of view by Chernoff [1949].

Since the likelihood function will have, in general, several maxima, it would be important to know their number and to ensure, by a proper choice of initial values, that successive iterations approach that set of values of parameters which corresponds to the highest of the likelihood maxima. These problems are as yet unsolved.

3.2. The fact that economic models are, in general, dynamic and have the nature of "stochastic processes" of a time variable gives rise to special problems treated in Part Two. Here the existence of simultaneous equations is provisionally relegated to the background and the study is occasionally confined to a uni-equational model.

3.2.1. The first division of Part Two ("Trend and Seasonals") may be said to specify the general (possibly nonlinear) model by using certain information on the possible role of the exogenous variable, time. A specified economic model of this kind is studied in Hurwicz's article [XI] on variable parameters in stochastic processes; structural estimation is the author's objective. In the two other articles of the section, given by guests of the Cowles Commission, Mann and R. L. Anderson, less reliance is placed on knowledge from economic theory; correspondingly, no attempt is made to estimate economic structure. Rather than estimating structural parameters, these two articles discuss how to test the hypothesis that no trend, or that no seasonal fluctuation, is present. Mann [XII] has devised a new nonparametric test

against trend (applicable in situations like quality control of successive products of a machine); R. L. Anderson [XII] gives account of the analysis of variance as used in testing for the presence of seasonality in a unitemporal (lagless) system.

3.2.2. The second division of Part Two ("Estimation Problems") considers the consequences of relaxing certain assumptions that in Part One were shown to be sufficient to make maximum-likelihood estimates unbiased: in particular, the assumption that the samples are large, and the assumption that the model describes a stationary process.

Problems of short time series were discussed at the conference by Madow and Hurwicz. The former's paper, [Madow, 1945], contained an exact formula through which the distribution of the autocorrelation coefficient of a discrete series when its population value differs from zero is connected with the corresponding distribution when the population value equals zero; the latter was derived in [Koopmans, 1942], [Dixon], and elsewhere. This formula applies to a circular series, i.e., a series in which the first and last elements are identical: a condition often satisfied in space series (as in the sampling of families around a block). The removal of this last condition is thus the only remaining step for the full solution of the problem. Hurwicz [XV] shows that, in short one-variable series, the least-squares estimate of the autoregression coefficient has a very considerable bias. This is a case when the regression coefficient equals the structural coefficient – hence the presence of bias in estimated structural coefficients is also proved. Moreover, in one of the cases treated (fixed initial value of the variable) the least-squares estimate equals the maximum-likelihood estimate. Hence, maximum-likelihood as well as least-squares estimates of both structural and regression coefficients are generally biased. This is an important warning since the time series that we have to use to estimate dynamic economic structures are short.

The assumption of stationary processes is discussed by Rubin [XIV]. He shows that, in the case of one variable, this assumption is not necessary for the consistency of maximum-likelihood estimates. This is important since economic "explosions" (speculative panics, etc., often described by economists as "cumulative processes") are known to occur.

3.2.3. The third division of Part Two, unlike the preceding two divisions, derives its importance not so much from specifying or relaxing the model of Part One as from correcting it. Koopmans'

article [XVI], which constitutes this division, suggests treating variables as functions of a continuous time variable instead of breaking up the continuous economic process into arbitrary periods of finite length. In addition to being more realistic the introduction of a continuous time variable helps to solve the following important paradox. In Part One of this volume, the answer to the question of whether a variable occurred with or without a time lag could, under certain circumstances, decide whether or not a system was identifiable and hence susceptible to estimation of all its parameters. The presence or absence of a time lag could also, under certain other circumstances, decide whether or not the estimation of some parameters by the single-equation least-squares method was subject to bias. Thus a lag of one day would appear to make an enormous difference for the statistician. Koopmans shows that this is not the case: as the time unit is being diminished, the assumption of independent successive disturbances becomes less and less permissible; and it is on the latter assumption that the above statements of Part One were based. A similar result is obtained for the discrete case by Hurwicz [XI]. If proper mathematical treatment of stochastic models can be developed, such models promise to be a more accurate and more flexible tool for inference in economics than the discrete models used heretofore.

3.3. Quite naturally, certain "maintained hypotheses" (as distinct from the "considered hypotheses" that are being tested on the basis of data) had to be specified throughout the volume. For example, linear relations were assumed, the effects of the presence of predetermined variables discussed, continuous and discrete models compared. Part Three is called "Specification of Hypotheses," with a view to more particular needs of economics.

It is introduced by Koopmans' article [XVII], in which he gives formal criteria for the successive delimitation of the exogenous and the predetermined variables. These criteria are not based on the distinction between "economic" and "noneconomic" variables; they refer, instead, to the algebraic role played by a given variable in the model. For example: unless significant lags exist between the economic facts and the political facts that are affected by them and that, in turn, affect economic facts, economic causation of politics must be treated as one of the relationships of the complete system. This shows the importance of the questions raised in Parts One and Two: how to deal with incomplete systems, and how to determine whether or not lags are significant.

The "nonadditive random disturbances" treated by Hurwicz in [XVIII] are studied to take into account the economists' view that, for example, the erratic variations in taste – from year to year or from person to person – do not necessarily affect the demand equation's constant term only; e.g., a change in slope of a linear demand curve may accompany its shift. This complicates the mathematics of statistical inference quite considerably. A beginning has been made in Rubin's brief article [XIX] that concludes the volume.

3.4. A number of major problems remain unsolved. It suffices to list some of those already mentioned in various places of this introduction: observation errors (and the problem of nonobservable variables in general); multiple decisions; problems of aggregation; continuous stochastic processes. Even where the statistical investigation has proceeded beyond the definition of the problem, the results obtained so far are subject to severe limitations. Structural estimation from small samples is in its beginnings; nonlinear relations, often called for in economic theory, have hardly been approached, etc. All these problems will require much further work on the part of statisticians.

PART ONE

SIMULTANEOUS EQUATION SYSTEMS

II. MEASURING THE EQUATION SYSTEMS OF DYNAMIC ECONOMICS

BY T. C. KOOPMANS, H. RUBIN, AND R. B. LEIPNIK[1]

[1]Rubin contributed to sections *1 - 3* and to an early draft of the methods developed in section *4*. Leipnik contributed to section *4* and directed most of the computations reported there. Further computations were directed by B. A. de Vries.

1. DESCRIPTION OF THE SYSTEMS CONSIDERED

1.1. The economic and statistical basis of a system of equations. The analysis and explanation of economic fluctuations has been greatly advanced by the study of systems of equations connecting economic variables. The construction of such a system is a task in which economic theory and statistical method combine. Broadly speaking, considerations both of economic theory and of statistical availability determine the choice of the variables. Economic theory predominates in the definition of the "behavior equations" describing a certain type of economic decisions taken by a certain category of economic agents, and in the specification of the variables that may possibly enter each behavior equation (i.e., of the conditions that may affect that decision by that group of agents). "Institutional equations" describe behavior patterns set by law or rule. Technical knowledge enters into the definition, and selection of variables, of the "technical equations" expressing the physical relation between input and output

in production. A fourth group of equations, usually referred to as "identities" (like the savings-investment identity), which occupy a place in economic literature out of proportion to their theoretical triviality, should be classified as deriving directly from the definitions of the variables through the principles of economic accounting. Theoretical preconceptions, statistical evidence, and sometimes mere assumption or approximation, are intermingled in the determination of the form of each equation, as regards linearity and as regards the occurrence and length of time lags. All these things being determined, it is almost entirely left to statistical methods to estimate the numerical values of the coefficients in the equations, and to assess the possible degree of error in those estimates, subject to the assumptions made.

Several equation systems of this kind have been constructed by Tinbergen [1939] and others for different countries and periods. We shall in this article assume a general knowledge of the nature of those systems, and of the uses to which they are put.

Tinbergen gives ample consideration to the economic assumptions on which these systems are based. Only recently has attention been directed systematically to the specific problems of statistical method involved in estimating the coefficients of any equation that forms part of such a system of equations. Haavelmo [1943, 1944] has pointed out that the methods developed for the measurement of a single relationship under conditions of experimental control over – or at least independent determination of – all variables except the one "dependent" variable, are inadequate if we are faced with a system of simultaneous equations between the variables. He has indicated the general principles of a statistical method appropriate to the latter situation. Mann and Wald [1943] have applied these principles to give a statistical treatment of large samples of a number of variables which satisfy an equal number of linear difference equations.

1.2. *Specifying the joint distribution of all variables.* The main principle advanced by Haavelmo is that the measurement of a system of equations should be based on a specification of the joint probability distribution of all values of all variables involved. This principle has been generally accepted in other applications of statistical method. Probably, economic statisticians have largely been unaware of the fact that their methods did not satisfy this requirement. Actually, the probability distributions that were employed always referred to one equation taken in isolation, and distributions specified with regard to different equations were usual-

ly incompatible.

1.3. Exogenous and endogenous variables. This article is concerned with linear systems of difference equations of the following general form:

$$(1.1)\qquad \sum_{i=1}^{G}\sum_{\tau=0}^{\tau^{\square}}\beta_{gi\tau}y_i(t-\tau)+\sum_{k=1}^{K}\sum_{\tau=0}^{\tau^{\square}}\gamma_{gk\tau}z_k(t-\tau)=u_g(t),$$

$$g=1,\ 2,\ \ldots,\ G;\quad t=1,\ 2,\ \ldots,\ T.$$

This form is slightly more general than that studied by Mann and Wald in that we consider G equations containing both G *endogenous* variables $y_i(t)$ and K *exogenous* variables $z_k(t)$. The latter are defined as variables that influence the endogenous variables but are not themselves influenced by the endogenous variables. It will be clear that at this stage the distinction between exogenous and endogenous variables is a theoretical, a priori distinction on which statistical evidence may or may not be obtained at a later stage. Because of the general interdependence of economic variables, exogenous variables are most likely to be found among noneconomic phenomena like temperature, rainfall, earthquake intensities, etc. Both endogenous and exogenous variables are assumed to be observable.

In the equations (1.1) the exogenous variables are treated as if they are given functions of time, the values of which remain the same in repeated samples. Another contribution to this volume, [XVII], is devoted to the justification of this procedure, which we shall here assume to be correct.

1.4. The disturbances in the equations. The distribution of the endogenous variables is then defined by means of the not directly observable *disturbances* $u_g(t)$. The latter are terms in the equations specified only to the extent that they are assumed to be subject to a joint probability distribution. Because of the presence of these terms, the system (1.1) is called a system of *stochastic equations.*

It will be assumed here that the $u_g(t)$ have a joint probability distribution of the form

$$(1.2)\qquad \prod_{t=1}^{T} f(u_1(t),\ \ldots,\ u_G(t))\,du_1(t)\ \ldots\ du_G(t).$$

The assumption (1.2) implies independence of disturbances in successive time intervals. It also implies that the disturbances are independent of the values of the exogenous variables. The conditions to be imposed on the distribution function f are not the same in the various sections of this article. In all cases, we shall assume that first-order moments exist and are equal to zero,

$$\mathcal{E}\, u_g(t) = 0, \tag{1.3}$$

and that second-order moments (variances and covariances),

$$\mathcal{E}\, u_g(t)\, u_h(t) = \sigma_{gh}, \tag{1.4}$$

also exist. In certain sections of this article we shall go further and assume that $f(u_1, \ldots, u_G)$ represents the joint normal distribution function of G variables.

The assumed distribution of the $u_g(t)$ defines the joint distribution of all values $y_i(t)$ of the endogenous variables for which $t = 1, \ldots, T$, provided we specify in addition that any values $y_i(t)$ for which $t \leq 0$ and which occur in (1.1) are regarded as given numbers that remain the same in repeated samples.

1.5. Economic interpretation of the disturbances in the equations. In each behavior equation, the disturbance is interpreted as representing the joint effect, on the behavior described by that equation, of all variables of minor individual importance that have not been explicitly introduced into the system of equations. For instance, random variation in consumers' tastes will lead to a certain amount of shifting in the curve of consumers' demand. Similarly, in the technical relations between input and output, a certain amount of random shifting in the relationships is due to a large number of minor causes of variation not explicitly studied. The term "random" is used here in the sense of the assumptions (1.2), (1.3), and (1.4), made regarding the disturbances in the equations.

1.6. Errors of measurement or disturbances in the variables. It is important to note that in the interpretation of disturbances just given, each disturbance is associated with an equation of the system, and not with a variable. This excludes the interpretation of the "disturbances in the equations" as errors of measurement. If errors of measurement occur to a marked degree, separate provision must be made for them in the probability dis-

tribution of observed variables by introducing additional "disturbances in the variables." In order to concentrate on the effect of disturbances in the equations, we shall assume in this study that all variables are measured without error. Systems in which "disturbances in the equations" (also called "shocks") and "disturbances in the variables" (also called "errors") occur side by side have been studied in [T. W. Anderson and Hurwicz].

1.7. Nonsingularity of Σ. For some purposes the mathematical treatment of systems like (1.1) is simplified if we can restrict ourselves to cases in which there is no *functional* relation (as distinct from *stochastic* dependence) between the G disturbances $u_g(t)$ for any t. This requires in particular that the matrix $\Sigma = [\sigma_{gh}]$ defined by (1.4) be nonsingular.

Now each "identity" that is present among the equations (1.1) makes all elements in the corresponding row and column of the matrix Σ vanish, because by their nature identities are not subject to disturbances. However, the variables entering into a given identity, and the coefficients with which they enter, are always known a priori (often the coefficients are $+1$ or -1). It is therefore possible, whenever the assumption of nonsingularity of Σ is desirable for mathematical reasons, to remove the identities from the system by elimination of as many variables as there are identities to be removed. For instance, the identity "volume of production equals real income" can be removed by replacing the variable "volume of production," wherever it occurs, by the variable "real income." A less trivial example: if the profit margin is conceived to be a determining factor of investment activity, the identity defining the profit margin can be removed by replacing the variable "profit margin" in the equation explaining investment activity by the linear combination "product price less the sum of factor prices per unit of product." The latter example shows that the elimination of variables defined by identities may introduce a priori proportionalities or other restrictions among the coefficients occurring in the remaining equations of the system. We shall revert to these a priori restrictions below.

In the case in which the identities have thus been disposed of, it is reasonable to assume that no functional relation exists between the disturbances in different behavior equations and technical equations. This can be seen if we ask ourselves what, for instance, would be implied in an exact proportionality of the disturbances in two given equations. This would mean, not only that

precisely the same minor causes would be operative in the random shifts of each of these relationships, but also that the relative strengths with which these causes operate in each equation are the same – an obvious impossibility. A similar, slightly more complicated argument applies to disqualify the assumption of a functional relation involving disturbances in more than two equations.

1.8. Jointly dependent variables and predetermined variables. Besides the distinction between endogenous and exogenous variables, it is desirable to introduce another classification of variables, which is based partly on the former distinction, and partly on the timing of each variable. That is, for the purposes of the classification now to be introduced, it will be necessary to regard for instance $y_i(t)$ and $y_i(t-1)$, and generally all variables measured with different time lags, as different variables.

The equations (1.1) for a given value of t are intended to describe the process of the formation of the endogenous variables $y_i(t)$ at time t, under the influence of earlier values $y_i(t-\tau)$, $\tau \geq 1$, of the endogenous variables, of the exogenous variables $z_k(t-\tau)$, $\tau \geq 0$, and of the disturbances $u_g(t)$. Generalizing a terminology of single-equation least-squares regression theory, the values $y_i(t)$, without time lag, of the endogenous variables may be called *jointly dependent variables at time t*. To bring out more clearly that the equations (1.1) describe the formation of the jointly dependent variables, these equations may be written in the form

$$\sum_{i=1}^{G} \beta_{gi_0}\, y_i(t) = - \sum_{i=1}^{G} \sum_{\tau=1}^{\tau^{\square}} \beta_{gi\tau}\, y_i(t-\tau) - \sum_{k=1}^{K} \sum_{\tau=0}^{\tau^{\square}} \gamma_{gk\tau}\, z_k(t-\tau) + u_g(t). \tag{1.5}$$

The right-hand member contains, besides the disturbances, a group of variables that we shall call *predetermined variables at time t*. The lagged values $y_i(t-\tau)$, $\tau \geq 1$, of the endogenous variables are predetermined in a temporal sense, in that their values $y_i(t-\tau)$ for a given value of t are determined by variables and disturbances relating to time intervals preceding t. In particular, they are unaffected by the disturbances $u_g(t)$ of the time interval t. The

exogenous variables $z_k(t)$ without time lag are predetermined in the logical sense that they are influenced only by causes outside the economic system studied, and are independent of all other variables and disturbances (measured at time t or earlier) included in the system of equations. The lagged exogenous variables $z_k(t-\tau)$, $\tau \geq 1$, are predetermined in both senses.

1.9. Nonsingularity of B_0. Although the need for the foregoing classification will become clear only when we come to estimation problems, it is introduced here because of a related basic assumption that is of general importance. If the right-hand member in (1.5) represents a set of causal factors in the determination of the dependent variables $y_i(t)$ in the left-hand members, without any causal action in the opposite direction, then it is necessary to specify that the matrix

$$B_0 = \begin{bmatrix} \beta_{110} & \cdots & \beta_{1G0} \\ \cdot & \cdots & \cdot \\ \beta_{G10} & \cdots & \beta_{GG0} \end{bmatrix} \tag{1.6}$$

be nonsingular. For if B_0 were singular, there would be a set of numbers λ_g, $g = 1, \ldots, G$, not all equal to zero such that

$$\sum_{g=1}^{G} \lambda_g \beta_{gi_0} = 0. \tag{1.7}$$

Writing $w_g(t)$ for the right-hand member in (1.5), the validity of (1.7) would then entail a linear restriction,

$$\sum_{g=1}^{G} \lambda_g w_g(t) = 0, \tag{1.8}$$

on the expressions represented by $w_g(t)$. Such a restriction, however, is contrary to the assumed direction of causation, and is in particular incompatible with the fact that there is no linear functional dependence between the stochastic variables $u_g(t)$.

The foregoing argument needs amendment in case (1.5) contains identities, because in that case the corresponding quantities $u_g(t)$ vanish. Suppose that the equations (1.5) for $g = 1, \ldots, G_0$ are identities, and that for $g = G_0 + 1, \ldots, G$ the equations involve disturbances of positive variance. Then a restriction (1.8) is compatible with the assumed direction of causation if and only if

$$\lambda_{G_0+1} = \cdots = \lambda_G = 0. \tag{1.9}$$

The only form of nonsingularity permissible in B_0 in the present case is therefore, according to (1.7), linear dependence among the first G_0 rows, each of which contains the coefficients of an identity. Such linear dependence is, of course, precluded by the simple reason that any linear dependence should be removed from the (fully known) identities before they are admitted to the system of equations.

There is, of course, no a priori reason why a number of behavior equations could not happen to be such that B_0 is singular. But there is good empirical evidence that this case can be ruled out, at least in dynamic equation systems. For if B_0 were singular (or even if its determinant value $\det B_0$ were very small compared with its term largest in absolute value), small disturbances in any direction in the space of the disturbance vector $u = (u_1, \ldots, u_G)$ incompatible with the linear restriction (1.7), would lead to infinite (or very large) *simultaneous* changes in the variables $y_i(t)$. Such phenomena have not been observed. Although small causes occasionally have great effects in economic developments, in such cases time is required for the effects to materialize.

One could not have equal confidence in a statement that the matrix $\bar{B}$ with elements $\bar{\beta}_{gi} = \sum_{\tau=1}^{\tau^0} \beta_{gi\tau}$, describing a corresponding static system obtained through the neglect of all time lags, is far from being singular. It is easily seen, however, that if $\bar{B}$ is singular for such a static system, then the deterministic dynamic system obtained from (1.5) by omitting the disturbances and giving *arbitrary* constant values to the exogenous variables does not in general have a solution asymptotically approaching a set of finite equilibrium values.

1.10. Timing of the variables. The nonsingularity of B_0 in particular excludes the possibility that one of the variables $y_i(t)$ would not occur at all in the equations (1.1) except with a positive time lag. A variable of that kind properly belongs among the exogenous variables because a past quantity cannot be influenced by present developments. An objection to this reasoning might be that the timing with respect to which the variables are defined may be varied at will by a transformation,

$$
\begin{aligned}
y_i^{\oplus}(t) &= y_i(t + \theta_i), \qquad i = 1, \ldots, G, \\
z_k^{\oplus}(t) &= z_k(t + \theta'_k), \qquad k = 1, \ldots, K.
\end{aligned}
\tag{1.10}
$$

However, such transformations cannot be regarded as permissible for our purposes – except in the trivial case where all quantities θ_i and θ_k are equal. The time variable is more than an index used to distinguish successive values of one and the same variable. It indicates historical time – the medium in which causation and interaction between economic and other variables takes place. Therefore the matrix B_0 must be such that the endogenous variables, which by their definition are in continuous and instantaneous interaction with each other, are all represented in the system (1.1) by simultaneous values with zero time lag ($\tau = 0$). Further light is thrown on this important point in another contribution already referred to [XVII].

1.11. The problem of identification. The statistical measurement of a system of equations like (1.1) involves two logically distinct and successive problems, which have here been called the problem of the *identification* of each equation and the problem of the *estimation of the parameters* of each equation. Section 2 is devoted to the former of these problems, which arises especially with regard to data governed by more than one equation at the same time. It originates from the fact that, if a system like (1.1) is viewed *only* as a mathematical specification of the joint probability distribution of the observable variables, it can be written in many different ways. Any linearly independent system of G linear combinations of the equations (1.1) with a correspondingly transformed distribution of the disturbance terms will be a mathematically equivalent way of defining the probability distribution of the variables.

Let a "way of writing" the system be called a *representation*

of the distribution of the variables. Two representations are called *observationally equivalent* if they define the same probability distribution of the variables. Haavelmo [1944, p. 91] uses the expression "indistinguishable on the basis of the observations" to describe two equivalent representations, because even if the probability distribution of the observations were fully known – the best that can be expected from statistical methods – there would still be no way to distinguish observationally equivalent representations. The distribution of the variables can be looked upon as determining the set of all observationally equivalent representations of it, and is completely defined by any of these representations. Mathematically speaking, it is immaterial which representation is employed, except that it will be desirable to choose a simple one. Economically, however, different representations of the same system are not at all equivalent.

The study of a system of equations like (1.1) derives its sense from the postulate – already implicit in earlier parts of this section – that there exists one and only one representation in which each equation corresponds to a specified law of behavior (attributed to a specified group of economic agents), to a specified technical law of production, or to a specified identity. Let us call these particular equations the *structural equations*, because they are the elements of which the dynamic economic structure of society is composed. The representation composed of the structural equations may be called the *representation according to economic structure*, or briefly the *structural representation*. Any discussion of the effects of changes in economic structure, whether brought about by gradual trends or by purposive policies, is best put in terms of changes in the structural equations. For those are the elements that can, at least in theory, be changed one by one, independently. For this reason, it is important to have the system (1.1) in a form in which the greatest possible number of its equations can be identified and recognized as structural equations.

Suppose for a moment that the structural representation be known to investigator *A*, who as a mathematical exercise derives another representation from it by taking linear combinations. In that process, the economic identity of the structural equations is lost, and when *A* hands the derived representation over to *B* without disclosing its source or method of computation, *B* is faced with the problem of identifying among all linear combinations of the equations of the representations given to him, the structural equations that alone reflect specified laws of economic behavior, of the

technique of production, or of economic accounting.

1.12. The a priori restrictions. The position of our investigator B corresponds exactly to the position of the econometrician who sets out to measure a system of economic relations. Statistical observation will in favorable circumstances permit him to estimate, with a precision again subject to estimation, the characteristics of the probability distribution of the variables. Under no circumstances whatever will passive statistical observation permit him to distinguish between different mathematically equivalent ways of writing down that distribution. Because he has no experimental control over economic variables, the simultaneous validity of all the structural equations prevents him from isolating and individually observing any one of them on a statistical basis alone. The only way in which he can hope to identify and measure individual structural equations implied in that system is with the help of a priori specifications of the form of each structural equation.

The most important instrument of identification is a specification as to *which variables may enter into which structural equations with which possible time lags.* Assuming now that the system (1.1) is the structural representation, this can be expressed mathematically by putting equal to zero all coefficients of terms that do not enter into the respective equations,

$$\begin{aligned} &\beta_{g_r i_r \tau_r} = 0, \qquad r = 1, 2, \ldots, R_\beta^{(1)}, \\ &\gamma_{g_r k_r \tau_r} = 0, \qquad r = R_\beta^{(1)}+1, \ldots, R_\beta^{(1)}+R_\gamma^{(1)}, \quad R_\beta^{(1)}+R_\gamma^{(1)} \equiv R_\alpha^{(1)}. \end{aligned} \tag{1.11}$$

Sometimes it is useful to state these restrictions on the coefficients in a slightly more general form which includes (1.11) as a special case,

$$\sum_{i=1}^{G} \sum_{\tau=0}^{\tau^\square} \chi_{ri\tau}^{(g_r)} \beta_{g_r i\tau} + \sum_{k=1}^{K} \sum_{\tau=0}^{\tau^\square} \psi_{rk\tau}^{(g_r)} \gamma_{g_r k\tau} = 0, \qquad r = 1, 2, \ldots, R_\alpha^{(1)}. \tag{1.12}$$

The special case (1.11) will be distinguished from (1.12) as the

case of single-parameter restrictions. The form (1.12) of restrictions, in which the quantities $\chi^{(g_\tau)}_{ri\tau}$ and $\psi^{(g_\tau)}_{rk\tau}$ are a priori known constants, permits inclusion of cases where the ratio of two coefficients in the same equation, or another linear relation between the coefficients of an equation, is given a priori. Examples of this type of restriction have already been given.

It will be noted that each condition (1.12) connects only coefficients that occur in the same structural equation. There is a further type of restrictions involving coefficients occurring in different equations. This can again be illustrated with the example of the profit margin. Assume that the profit margin enters as such into at least two behavior equations, whereas none of its constituents enters explicitly. Suppose further that the definition of the profit margin, with the help of which we wish to eliminate that variable, contains an unknown parameter. (This may happen, for instance, if the conversion factor by which the price of any given factor of production is related to the unit of product is not known.) The type of restriction arising from such a situation is one in which two coefficients, of the variables y_{i_τ} and $y_{i'_\tau}$, respectively, are required to have the same ratio in two different structural equations (numbered g_τ and g'_τ):

$$\begin{bmatrix} \beta_{g_\tau i_\tau \tau_\tau} & \beta_{g_\tau i'_\tau \tau_\tau} \\ \beta_{g'_\tau i_\tau \tau_\tau} & \beta_{g'_\tau i'_\tau \tau_\tau} \end{bmatrix} = 0. \tag{1.13}$$

Similar restrictions may arise from the approximation of a distributed time lag by a linear combination of terms with discrete lags. If a variable y_{i_τ} is supposed to occur with the same lag distribution in two equations numbered g_τ and g'_τ, this leads to a restriction of the type

$$\begin{bmatrix} \beta_{g_\tau i_\tau \tau_\tau} & \beta_{g_\tau i_\tau \tau'_\tau} \\ \beta_{g'_\tau i_\tau \tau_\tau} & \beta_{g'_\tau i_\tau \tau'_\tau} \end{bmatrix} = 0. \tag{1.14}$$

While the restrictions (1.12) are linear in the unknown coefficients

$\beta_{gi\tau}$, $\gamma_{gk\tau}$, restrictions like (1.13) and (1.14) are bilinear, and lead to greater mathematical complications in what follows. We shall assume that there are $R_\alpha^{(2)}$ bilinear restrictions of the types (1.13) and (1.14), possibly involving coefficients $\beta_{gi\tau}$, $\gamma_{gk\tau}$ of both endogenous and exogenous variables.

The restrictions (1.11) or (1.12), (1.13), and (1.14) – and such similar restrictions as we may wish to add later – will be called the *a priori restrictions*. In section 2 we investigate necessary and sufficient conditions under which the a priori restrictions suffice to identify a given equation (1.1) as a specified structural equation. On this basis we shall distinguish, and include in subsequent sections, the case, not covered by Mann and Wald [1943], in which one or more but not all of the structural equations can be identified within the system.

It will be seen that, even in the case where the a priori restrictions are insufficient in number *and variety* to permit identification of all structural equations, there may be among the a priori restrictions one or more that can be omitted without thereby removing further equations from the list of identifiable ones. "A priori" restrictions of this kind are in principle subject to statistical testing (on the basis of the remaining a priori restrictions). For this reason, statistical evidence was quoted, in the opening paragraph of this article, as one of the bases for a determination of the form of the structural equations. If restrictions supported to a degree by statistical evidence are nevertheless imposed a priori, this will in general reduce the sampling variances of the estimates of some or all parameters subject to estimation. The use of a priori restrictions should therefore be resorted to whenever the theoretical grounds are strong enough. To make possible a formal mathematical treatment according to established procedures of statistical inference, we shall in this article regard the "a priori restrictions" strictly as given a priori and imposed without reference even to the possibility of statistical test. It goes without saying that we should eliminate from consideration sets of a priori restrictions that are mutually incompatible or mutually dependent.

1.13. A priori restrictions on the distribution of disturbances. It may well happen that one or more specified structural equations cannot be identified on the basis of such a priori restrictions of the forms (1.12), (1.13), and (1.14) as are considered theoretically justified. We shall therefore study further a

priori restrictions on the matrix Σ of the variances and covariances of the disturbances, which require that the covariance between the disturbances in two specified equations shall vanish. The elements of Σ that are thereby required to vanish may or may not follow a regular pattern. One particular pattern is of interest both because of its special mathematical consequences and because the assumptions involved may present a fair approximation to reality. In this pattern (which is here formulated for systems from which all identities have been removed) it is supposed that the G equations can be classified into N groups of $G_1, G_2, \ldots, G_N$ equations respectively, with $G_1 + G_2 + \cdots + G_N = G$, such that the disturbances $u_g(t)$ of equations in different groups are independent. In that case, the matrix Σ can be partitioned into the form

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \ldots & 0 \\ 0 & \Sigma_2 & \ldots & 0 \\ . & . & \ldots & . \\ 0 & 0 & \ldots & \Sigma_N \end{bmatrix}, \tag{1.15}$$

where each of the matrices $\Sigma_1, \ldots, \Sigma_N$ is positive definite. Each element σ_{gh}, $g<h$, of Σ that is thereby prescribed to be zero gives rise to an a priori restriction, which turns out (see section 2) to be equivalent to a bilinear restriction in the elements of the corresponding rows of the coefficient matrix $[\,B \quad \Gamma\,]$. For this reason we shall denote by R_σ the total number of such restrictions, and write $R_\alpha^{(1)} + R_\alpha^{(2)} + R_\sigma = R$. The particular choice of vanishing elements indicated by the partitioning in (1.15) is simpler than other choices because it is invariant under inversion of Σ.

1.14. Inequalities as a priori restrictions. A further class of a priori restrictions that can often be based on economic considerations is inequalities. Frequently, the sign of coefficients $\beta_{gi\tau}$ or $\gamma_{gk\tau}$ is known beforehand. Sometimes it may be possible to prescribe the sign of, or set another limit to, the correlation of the disturbances in the structural equations. In the present article we do not study the question of how to give effect to restrictions of this kind.

1.15. Rules of normalization. The equations (1.1) as well as the restrictions (1.12), (1.13), and (1.14) are homogeneous in the coefficients $\beta_{gi\tau}$ and $\gamma_{gk\tau}$ and the parameters σ_{gh} for each value of g, i.e., they are unaffected by a change in scale

$$(1.16) \qquad \beta^{\oplus}_{gi\tau} = \upsilon_g \beta_{gi\tau}, \qquad \gamma^{\oplus}_{gk\tau} = \upsilon_g \gamma_{gk\tau}, \qquad \sigma^{\oplus}_{gh} = \upsilon_g \sigma_{gh} \upsilon_h,$$

of each equation (1.1). It will be useful sometimes to fix the scale factors υ_g by imposing a *normalization rule* on each equation. The precise form of the normalization rule is obviously a matter of choice, and different normalization rules are most convenient in different problems. We shall consider the following two of many possible alternative sets of G normalizing restrictions:

$$(1.17) \quad \begin{cases} (1.17a) & \beta_{g i_{g0}} = 1, \qquad g = 1, 2, \ldots, G, \\ (1.17b) & \sigma_{gg} = 1, \qquad g = 1, 2, \ldots, G. \end{cases}$$

In the case of the first rule (1.17a) there should of course be no conflict with (1.11) or (1.12). The second rule (1.17b) still leaves open the choice of the sign of one of the nonvanishing coefficients "β" or "γ" in each equation.

Because of the trivial nature of the question of normalization, we shall sometimes omit specification of a normalization rule. It is therefore useful to introduce the convention that the gth equation can be called completely identified by the a priori restrictions even if its scale has not been fixed, provided the ratios between all its coefficients and the quantities σ^{kk}_{gg}, σ_{gh}, $h \neq g$, are determinate. In case normalization rules are specified, they will be comprised in the term "a priori restrictions."

1.16. Summary of subsequent sections. In section 2, we discuss conditions for the identifiability of a given structural equation under a priori restrictions of the type (1.12), (1.13), (1.14), or (1.15). Necessary and sufficient conditions for identifiability under the restrictions (1.12) are derived (section 2.2). In section 2.4, the problem of extending these conditions to cases where restrictions of the types (1.13), (1.14), (1.15) are added is discussed but not solved. Means are indicated in section 2.3 to make possible the estimation of certain identifiable structural equations, even if certain other equations remain

unidentifiable. General observations indicating the incomplete state of our knowledge with regard to identification problems conclude this section.

Sections *3.1* and *3.2* deal with those properties of the likelihood function, before and after imposition of a priori restrictions, which are relevant to the maximum-likelihood method of estimation. It is found that whenever restrictions are imposed on structural equations that are not indispensable for the identification of those equations, the likelihood function is prevented from reaching its unrestricted absolute maximum, except in a set of samples of probability zero. Under such restrictions, maximum-likelihood estimates using all a priori information can only be obtained by computational procedures essentially more complicated than the least-squares method applied to the "reduced form" without regard to restrictions. Section *3.3* discusses and generalizes results regarding the limiting distribution of the maximum-likelihood estimates reached by earlier writers.

In section 4, iterative computation methods for the maximum-likelihood estimates are developed and discussed for the two cases in which the covariance matrix Σ of the disturbances is diagonal (section *4.3*), and unrestricted (section *4.4*). Unsolved problems connected with these methods are indicated.

Sections preceded by the symbol * can be passed over in a first reading without seriously affecting the understanding of the remaining parts of the article.

2. THE IDENTIFICATION OF ECONOMIC RELATIONS

2.1. The Concept of Identification

2.1.1. Earlier discussions of the identification problem. The first systematic discussion of the problem of identification was given by Frisch in an unpublished memorandum [1938]. Frisch's terminology is rather different from that employed here, and the concepts are slightly different in that the disturbances and their distribution are not explicitly introduced in his formulae. Nevertheless, the underlying ideas are to a large extent the same, and the present authors desire to acknowledge their indebtedness, and to emphasize the support found in Frisch's memorandum for the discussion of the problem of identification in this article.

Frisch indicates that what is here called the identification

problem arises from the passive nature of economic observations. There is no possibility of independently varying the several factors entering a given behavior equation. The only observations available are those which by assumption satisfy all structural equations simultaneously.

The same point is emphasized by Haavelmo, who has continued and extended Frisch's work in a very general discussion [Haavelmo, 1944, pp. 91-98] of one central problem in identification: the formulation of conditions under which *all* structural relations of the system can be identified. Haavelmo also does not use the term identification, but describes the above mentioned problem as the "problem of confluent relations" or, alternatively, as the "problem of arbitrary parameters," and classifies it under the heading "estimation." As regards this classification, it appears to the present authors that the identification problem is concerned with the unambiguous definition of the parameters that are to be estimated – a logical problem that precedes estimation. It is therefore not a problem in statistical inference, but a prior problem arising in the specification and interpretation of the probability distribution of the variables. As such it deserves separate classification.

Haavelmo's discussion of the "problem of confluent relations" is more general than the present discussion of identification problems in that he does not in any way restrict the functional form of the equations concerned. The conditions to be given below for the identifiability of *all* structural equations in a linear system could therefore be obtained as a specialization of Haavelmo's results, although we shall derive them directly. The present discussion, while restricted to linear systems, goes further in that we also discuss conditions under which any one particular structural equation can be identified.

2.1.2. *Notation.* It will be convenient to use a matrix notation for the equation system (1.5) in which the distinction between jointly dependent and predetermined variables introduced in section *1.8* is given explicit expression, whereas that between endogenous and exogenous variables is concealed. The variables and the disturbances will be represented by row vectors, and the coefficients will be regarded as the elements of matrices, as follows:

$$y(t) \equiv [y_1(t) \quad \cdots \quad y_G(t)], \tag{2.1}$$

...

$$z(t) \equiv \left[y_1(t-1) \quad \cdots \quad y_G(t-\tau^{\square}) \quad z_1(t) \quad \cdots \quad z_K(t-\tau^{\square}) \right],$$

$$u(t) \equiv \left[u_1(t) \quad \cdots \quad u_G(t) \right],$$

(2.1)
$$B \equiv \begin{bmatrix} \beta_{110} & \cdots & \beta_{1G0} \\ . & \cdots & . \\ \beta_{G10} & \cdots & \beta_{GG0} \end{bmatrix},$$

$$\Gamma \equiv \begin{bmatrix} \beta_{111} & \cdots & \beta_{1G\tau^{\square}} & \gamma_{110} & \cdots & \gamma_{1K\tau^{\square}} \\ . & \cdots & . & . & \cdots & . \\ \beta_{G11} & \cdots & \beta_{GG\tau^{\square}} & \gamma_{G10} & \cdots & \gamma_{GK\tau^{\square}} \end{bmatrix}.$$

Here the vector $y(t)$ of $G \equiv K_y$ elements comprises the variables jointly dependent at time t, and the vector $z(t)$ of K_z elements, say, comprises all variables predetermined at time t. The matrix B has previously been denoted by B_0. In this notation, the equations (1.1) can be written as follows:

(2.2)
$$By'(t) + \Gamma z'(t) = u'(t),$$

where y' denotes the column vector which is the transpose[1] of the row vector y. The probability density function $f\{u_1(t), \ldots, u_G(t)\}$ of the disturbances will be denoted by $f\{u(t)\}$. Occasionally, the argument t of y, z, u will be omitted.

For some purposes even the distinction between dependent and predetermined variables is irrelevant. Then we shall denote the equation system by

---

[1]For reasons to be stated in the footnote on p. 81, we have reversed the more usual notation in which the transposition sign denotes a row vector, its absence a column vector.

$$(2.3)\qquad \sum_{k=1}^{K_x} \alpha_{gk} x_k = u_g, \quad g = 1, \ldots, G, \quad K_x \equiv K_y + K_z \equiv G + K_z,$$

or $Ax' = u'$, where $A = [\, B \quad \Gamma \,]$, $x = [\, y \quad z \,]$.

2.1.3. Observationally equivalent structures. The concept of identifiability is to be defined with reference to the joint distribution function F_T of all observations, as determined by (1.2), (1.1), and the requirement that all values $z_k(t)$ and those values of $y_i(t)$ for which $t \leq 0$ are given constants. In order to write F_T explicitly as a distribution function in terms of the variables $y_i(t)$, $i = 1, \ldots, G$; $t = 1, \ldots, T$, it is necessary to regard (1.1) as a transformation expressing the $u_g(t)$ in terms of the $y_i(t)$. It is well known (see for instance [Wilks, p. 28]) that the absolute value of the volume element transforms according to

$$(2.4)\qquad \left| du_1(1)\, du_2(1) \cdots du_G(T) \right| = J_T \left| dy_1(1)\, dy_2(1) \cdots dy_G(T) \right|,$$

$$J_T \equiv \left| \frac{\partial\{u_1(1), \ldots, u_G(1); u_1(2), \ldots, u_G(2); \ldots; u_1(T), \ldots, u_G(T)\}}{\partial\{y_1(1), \ldots, y_G(1); y_1(2), \ldots, y_G(2); \ldots; y_1(T), \ldots, y_G(T)\}} \right|.$$

Here J_T represents the absolute value of the Jacobian determinant of the transformation (1.1). In evaluating the elements $\partial u_g(t)/\partial y_i(t')$ of the determinant in (2.4) we find that

$$(2.5)\qquad \frac{\partial u_g(t)}{\partial y_i(t)} = \beta_{gi0}, \qquad \frac{\partial u_g(t)}{\partial y_i(t')} = 0,$$

if $t' > t$. Viewed as a matrix, the Jacobian in (2.4) can therefore be partitioned as follows:

$$(2.6) \qquad \begin{bmatrix} B & & & \\ 0 & B & & \\ . & . & \dots & \\ 0 & 0 & \dots & B \end{bmatrix},$$

if, as before, $B = [\ \beta_{gi0}\ ]$. It follows that the determinant value of this matrix is independent of the elements $\partial u_g(t)/\partial y_i(t')$ with $t' < t$ [which have been left blank in (2.6)], and equals

$$(2.7) \qquad J_T(B) = |\det B|^T,$$

where the symbol "det" followed by a square matrix denotes the corresponding determinant. By assumption, the β_{gi0} are such that det B differs from zero (see section *1.9*). The distribution function therefore equals

$$(2.8) \qquad F_T = |\det B|^T \cdot \prod_{t=1}^{T} f\{ B\, y'(t) + \Gamma\, z'(t) \}.$$

In all integrations over the whole or part of the sample space, this function must, of course, be multiplied with the volume element $dy_1(1) \cdots dy_G(T)$.

The nature of the identification problem has already been explained in section *1.1*. We shall now formalize it by introducing the following concepts:

DEFINITION 2.1.3.1. *A structure S consists of a set of values of the coefficient matrices* B *and* Γ *(of which* B *is nonsingular), and a distribution function* $f(u)$ *of the vector u of disturbances.*

DEFINITION 2.1.3.2. *Two structures* $S = (B, \Gamma, f)$ *and* $S^{\oplus} = (B^{\oplus}, \Gamma^{\oplus}, f^{\oplus})$ *are called (observationally) equivalent if they imply the same probability distribution of the observations, i.e., if, for all values of T,* $y_i(t)$, $z_k(t)$,

$$\det{}^{T}(B)\prod_{t=1}^{T} f\{B\,y'(t) + \Gamma\,z'(t)\} \equiv$$

(2.9)

$$\det{}^{T}(B^{\oplus})\prod_{t=1}^{T} f^{\oplus}\{B^{\oplus}\,y'(t) + \Gamma^{\oplus}\,z'(t)\}.$$

This equivalence is denoted by $S \sim S^{\oplus}$. It follows from these definitions that equivalence of structures is transitive: if $S \sim S^{\oplus}$ and $S^{\oplus} \sim S^{\oplus\oplus\oplus}$ then $S \sim S^{\oplus\oplus}$.

We shall derive necessary and sufficient conditions for the equivalence of two structures. For that purpose, we restate certain restrictions which both structures are required to satisfy, and which are implied in the assumptions made in sections *1.4* and *1.9* respectively:

ASSUMPTION 2.1.3.3.

$$\mathcal{E}\{u(t) \mid z(t)\} = \mathcal{E}u(t) = 0.$$

This assumption is a consequence of the independence of $u(t)$ from the exogenous variables as well as from previous values $u(t')$, $t' < t$, of the disturbance vector, which together with exogenous variables determine the predetermined (lagged) endogenous variables now included in $z(t)$.

ASSUMPTION 2.1.3.4.

$$\det B \neq 0.$$

Solving (2.2) for y' through premultiplication with B^{-1} we obtain

$$y'(t) = -B^{-1}\,\Gamma\,z'(t) + B^{-1}u'(t). \tag{2.10}$$

For any equivalent structure $S^{\oplus}$ we likewise have

$$y'(t) = -B^{\oplus-1}\Gamma^{\oplus}z'(t) + B^{\oplus-1}u^{\oplus\prime}(t). \tag{2.11}$$

Since the identity (2.9) in Definition 2.1.3.2 should hold for all values of T, it implies the identity

$$(2.12)\qquad |\det B|\cdot f\{B\,y'(t) + \Gamma\,z'(t)\} \equiv |\det B^{\oplus}|\cdot f^{\oplus}\{B^{\oplus}\,y'(t) + \Gamma^{\oplus}\,z'(t)\}$$

of the conditional probability distributions of $y(t)$ for given $z(t)$ according to the two structures. It follows that, upon taking conditional expectations in (2.10) and (2.11) for given values of $z(t)$, and using Assumption 2.1.3.3, the same function of the elements of $z(t)$ must result from both structures:

$$(2.13)\qquad \mathcal{E}\{y'(t) \mid z(t)\} = -\,B^{-1}\,\Gamma\,z'(t) = -\,B^{\oplus-1}\,\Gamma^{\oplus}\,z'(t).$$

Consequently

$$(2.14)\qquad B^{-1}\,\Gamma = B^{\oplus-1}\,\Gamma^{\oplus}.$$

The square matrix of order G

$$(2.15)\qquad \Upsilon = B^{\oplus}\,B^{-1}$$

is nonsingular by Assumption 2.1.3.4, and satisfies, on account of (2.15) and (2.14),

$$(2.16)\qquad B^{\oplus} = \Upsilon\,B, \qquad \Gamma^{\oplus} = \Upsilon\,\Gamma, \qquad u'^{\oplus} = \Upsilon\,u',$$

the last equality in (2.16) being obtained from the first two by (2.2) and its counterpart for the equivalent structure $S^{\oplus}$.

Conversely, let S be a given structure, and let $S^{\oplus}$ now be a structure derived from S by the transformation (2.16) where Υ is any nonsingular matrix of order G. It is easily seen that $S^{\oplus}$ is then equivalent to S. For (2.16) now implies successively the nonsingularity of $B^{\oplus}$, (2.14), and

$$(2.17)\qquad B^{\oplus-1}\,u'^{\oplus}(t) = B^{\oplus-1}\,\Upsilon\,u'(t) = B^{-1}\,u'(t).$$

Hence (2.10) and (2.11) define the same conditional distribution (2.12) of the dependent variables, for any set of predetermined variables and for any distribution function $f(u)$ of u. The two structures S and $S^{\oplus}$ thus define the same distribution function (2.9) of the observed variables and are accordingly equivalent.

It will be noticed that in (2.16) the coefficient matrices B

and Γ occur in the same manner. It is therefore appropriate to express the result just obtained in the notation introduced by (2.1):

THEOREM 2.1.3.5. *A necessary and sufficient condition for the equivalence of two structures* $S = (\mathrm{A}, f(u))$ *and* $S^{\oplus} = (\mathrm{A}^{\oplus}, f^{\oplus}(u^{\oplus}))$ *satisfying Assumptions* 2.1.3.3 *and* 2.1.3.4 *is that they are connected by a linear transformation*

$$\mathrm{A}^{\oplus} = \Upsilon\, \mathrm{A}\,, \tag{2.18α}$$

$$u'^{\oplus} = \Upsilon\, u'\,, \tag{2.18u}$$

with nonsingular matrix Υ.

2.1.4. *Two interpretations of the implied transformation of the parameters* Σ. We note that the transformation (2.18α) implies a transformation for the covariance matrix Σ of the disturbances whenever that matrix exists. This transformation, together with (2.18α), can be written in matrix form as follows:

$$(2.18)\quad \begin{cases} \mathrm{A}^{\oplus} = \Upsilon\, \mathrm{A}, & (2.18\alpha) \\ \Sigma^{\oplus} = \Upsilon\, \Sigma\, \Upsilon'. & (2.18\sigma) \end{cases}$$

It should be stressed that the present discussion of the identification problem is based on the assumption that all available knowledge regarding the distribution function $f(u)$ of the disturbances is expressed by Assumption 2.1.3.3. If additional information on the functional form of f were available, the possibility exists that such information could be used for identification purposes. However, it is easily seen that there is an alternative case, in which the conclusions as regards identifiability of structural equations are precisely the same as under the present assumptions. This is the case in which the very general Assumption 2.1.3.3 is replaced by the special assumption that $f(u)$ represents a nonsingular joint normal distribution of the disturbances $u_1, \ldots, u_G$. In this case, the space of structures S becomes the space of the parameters α_{gk}, σ_{gh}, and (2.18σ) supplants (2.18u) in the definition of a linear transformation in the "parameter space." Whenever the transformation (2.18) is quoted in what follows, either of the two interpretations just given is applicable.

It so happens that any additional restrictions on $f(u)$ which we shall consider in what follows are restrictions on the parameters σ_{gh}. These restrictions have the same identifying effects whether $f(u)$ is previously restricted only by Assumption 2.1.3.3 and the assumption that Σ exists, or whether $f(u)$ is previously restricted to the form of the normal distribution. For this reason, it will be convenient from now on to discuss the identification problem in terms of points (A, Σ) in the parameter space rather than in terms of structures $S = (A, f(u))$. We therefore supplement Definition 2.1.3.3 by

DEFINITION 2.1.3.6. *Two points* (A, Σ) *and* $(A^{\oplus}, \Sigma^{\oplus})$ *in the parameter space are called (observationally) equivalent if they are connected with equivalent structures.*

Theorem 2.1.3.5 can now also be interpreted as stating conditions for the equivalence of two points in the space of the parameters A and Σ.

2.1.5. Equivalent points in the restricted parameter space. The identification problem in this article consists in the study of the extent to which there exist nontrivial transformations (2.18) which preserve the a priori restrictions. The following definitions are helpful in developing the concept of identification.

DEFINITION 2.1.5.1. *By the restricted parameter space we understand the set of those points* (A, Σ) *in the parameter space that satisfy the a priori restrictions.*

It will further be useful to rule out as irrelevant certain trivial transformations that do not affect the economic identity of the equations whose identification is studied.

DEFINITION 2.1.5.2. *A transformation* (2.18) *is called trivial with respect to the* g_0th *structural equation if it involves only a change of scale*

$$
\begin{aligned}
&\alpha^{\oplus}_{g_0 k} = v_{g_0 g_0}\,\alpha_{g_0 k}, \qquad \sigma^{\oplus}_{g_0 g_0} = v^2_{g_0 g_0}\,\sigma_{g_0 g_0}, \\
&\sigma^{\oplus}_{g_0 g} = v_{g_0 g_0} \sum_h \sigma_{gh} v_{gh}, \qquad g \neq g_0,
\end{aligned} \tag{2.19}
$$

in the parameters of that equation.

The concept of identifiability of a structural equation is now defined as follows:

DEFINITION 2.1.5.3. *The g_0th structural equation in* (2.3) *is said to be identifiable by a set of a priori restrictions, in the point* (A, Σ) *of the restricted parameter space, if each point* $(A^{\circledast}, \Sigma^{\circledast})$ *in the restricted parameter space, equivalent to* (A, Σ), *is obtainable from* (A, Σ) *by a transformation* (2.18) *which is trivial with respect to the g_0th equation.*

DEFINITION 2.1.5.4. *The system* (2.3) *of structural equations is said to be identifiable by a set of a priori restrictions, in the point* (A, Σ) *of the restricted parameter space, if each of its equations is thereby identifiable.*

If the latter definition is applied with reference to a set of a priori restrictions that includes an unambiguous normalization rule for each structural equation (2.3), the definition of identifiability of the system (2.3) is equivalent to requiring that the set of points in the restricted parameter space, equivalent to (A, Σ), consist only of the point (A, Σ).

2.2. *Identification of One Structural Equation under Linear Restrictions*

2.2.1. Necessary and sufficient conditions for identifiability of a given structural equation under linear single-parameter restrictions. Let us first consider the case of the identification of the g_0th equation by linear a priori restrictions of the single-parameter form (1.11), which require certain specified α_{gk} to vanish. It is useful to rearrange the conditions (1.11) in the order of the structural equations to which they apply:

$$\alpha_{gk_r} = 0, \tag{2.20}$$

$$r = \bar{R}_{g-1} + 1, \ \ldots, \ \bar{R}_g, \qquad \bar{R}_g - \bar{R}_{g-1} = R_g, \qquad g = 1, \ \ldots, \ G,$$

with $\bar{R}_0 \equiv 0$, $\bar{R}_G \equiv R_1 + R_2 + \cdots + R_G \equiv R_{\alpha}^{(1)}$.

THEOREM 2.2.1. *A necessary and sufficient condition for the*

identifiability, under Assumptions 2.1.3.3 *and* 2.1.3.4 *of the* g_0 th *structural equation in* (2.3) *by the a priori restrictions* (2.20), *is that the matrix*

$$\mathrm{A}^{(g_0)} \equiv [\, \alpha_{g k_r} \,], \tag{2.21}$$

$$g = 1, \ldots, G, \qquad r = \bar{R}_{g_0-1} + 1, \ldots, \bar{R}_{g_0},$$

obtained from the complete matrix A *of the coefficients* α_{gk} *by selecting those columns* $k = k_r$ *for which* $\alpha_{g_0 k}$ *is required to vanish, is of rank*[1] $G - 1$.

Obviously, $\mathrm{A}^{(g_0)}$ cannot have a rank higher than $G-1$ because its g_0 th row consists of zeros only. Stated in other words, the condition in Theorem 2.2.1 requires that from the g_0 th row of the matrix $\mathrm{A} = [\, \alpha_{gk} \,]$ we can select in at least one way $G-1$ elements that the a priori restrictions require to be zero, such that the determinant obtained by combining the columns of those elements with all other rows differs from zero. If this theorem is true, the following corollary ensues.

COROLLARY TO THEOREM 2.2.1. *A necessary condition for the identifiability of the* g_0 th *structural equation by the a priori restrictions* (2.20) *is that the number* R_{g_0} *of these restrictions involving coefficients of the* g_0 th *equation be at least equal to the number* G *of structural equations less one.*

In order to prove the theorem let us first assume that $\Upsilon = [\, \upsilon_{gh} \,]$ defines a transformation of the type (2.18), which preserves those restrictions (2.20) for which $r = \bar{R}_{g_0-1} + 1, \ldots, \bar{R}_{g_0}$. Then

$$\alpha^{\oplus}_{g_0 k_r} = \sum_{h=1}^{G} \upsilon_{g_0 h}\, \alpha_{h k_r} = 0, \qquad r = \bar{R}_{g_0-1} + 1, \ldots, \bar{R}_{g_0}. \tag{2.22}$$

Because of (2.20) the term with $h = g_0$ can be omitted from the

[1] A matrix X is said to be of rank ρ, if at least one of the determinants of order ρ, and none of the determinants of order $\rho+1$, that can be formed from the elements of X, is different from zero. Obviously, ρ cannot exceed the number of rows or columns in X.

summation. As a consequence of the condition specified in Theorem 2.2.1, the system of homogeneous equations (2.22), in which the $v_{g_0 h}$ with $h \neq g_0$ are now regarded as $G-1$ unknowns, has the rank $G-1$ (for the omission of a row of zeros from a matrix does not affect its rank). We can therefore in at least one way select from (2.22) $G-1$ equations, with $r = r_1^{(g_0)}, \ldots, r_{G-1}^{(g_0)}$, say, in which the determinant $\Delta(g_0; r_1^{(g_0)}, \ldots, r_{G-1}^{(g_0)})$ of the coefficients of the unknowns differs from zero. It follows that

$$v_{g_0 h} = 0 \tag{2.23}$$

for $h \neq g_0$, and the transformation Υ can only be of the type (2.19) admitted in Definition 2.1.5.2. This proves that the condition stated in Theorem 2.2.1 is sufficient for identifiability of the g_0 th structural equation.

That this condition is also necessary is seen if we now assume that the g_0 th equation is identifiable and that therefore the only nonsingular transformations Υ satisfying (2.22) if (2.20) holds are those that satisfy (2.23). Suppose that at the same time $A^{(g_0)}$ has a rank lower than $G-1$. Then (2.22) would possess at least two linearly independent solutions $v_{g_0 h}^{(1)}$ and $v_{g_0 h}^{(2)}$, say, of which the first can be taken to satisfy (2.23). The more general solution $v_{g_0 h} = \lambda_1 v_{g_0 h}^{(1)} + \lambda_2 v_{g_0 h}^{(2)}$ then satisfies (2.23) only if $\lambda_2 = 0$. Since λ_1 and the other rows $(g \neq g_0)$ of Υ can always be selected so that Υ coincides with the identical transformation (with unit matrix) when $\lambda_2 = 0$, there are values $\lambda_2 \neq 0$ of λ_2 (e.g., in a neighborhood of $\lambda_2 = 0$) for which Υ is nonsingular, but as stated does not satisfy (2.23). This contradicts the assumption made at the beginning of this paragraph. Therefore a rank $G-1$ of $A^{(g_0)}$ is also necessary.

2.2.2. *Identifiability conditions under more general linear restrictions.* Theorem 2.2.1 can easily be generalized to the case where the linear a priori restrictions take the form (1.12). It will be useful now to write these restrictions in matrix form:

$$\alpha(g)\Phi_g' = 0, \qquad g = 1, \ldots, G. \tag{2.24}$$

Here $\alpha(g)$ is a row vector[1] containing the elements of the gth row of A. The matrix Φ_g' is the transpose of a matrix Φ_g of rank R_g, containing in R_g rows and K_x columns the coefficients "X" and "ψ" of those restrictions (1.12) that refer to the gth structural equation in (2.3). (It is permissible to take the rank of Φ_g equal to the number of its rows, since otherwise the restrictions expressed by (2.24) would not be independent.) Again

$$R_1 + \cdots + R_G = R_\alpha^{(1)}. \tag{2.25}$$

We consider only such transformations (2.18) that preserve the restrictions (2.24). Hence, if υ_g is the gth row of $\Upsilon \equiv \Upsilon_{yy}$,

$$\alpha^{\oplus}(g)\,\Phi_g' = \upsilon_g\,\mathrm{A}\,\Phi_g', \qquad g = 1, \ldots, G. \tag{2.26}$$

By a repetition of the previous reasoning, it is seen that the g_0 th condition (2.26) will then and only then require the elements of υ_{g_0} (except $\upsilon_{g_0 g_0}$) to vanish, if $\mathrm{A}\Phi_{g_0}'$ (of which the g_0 th row consists of zeros only) has the rank $G - 1$. We therefore have

THEOREM 2.2.2. *A necessary and sufficient condition for the identifiability, under the Assumptions* 2.1.3.3 *and* 2.1.3.4, *of the* g_0 th *structural equation by the a priori restrictions* (2.24) *is that* $\mathrm{A}\Phi_{g_0}'$ *has the rank* $G - 1$.

Since the rank of Φ_g is assumed equal to the number R_g of its rows (the number of independent restrictions imposed on the gth structural equation), and since the rank of a matrix cannot increase through premultiplication with another matrix, we have

COROLLARY TO THEOREM 2.2.2. *A necessary condition for the*

[1] Vectors are here considered as one-row matrices rather than the more commonly used one-column matrices in order to be able to treat rows of A as vectors, with A in the form corresponding to the natural way (1.1) of writing the structural equations.

identifiability, under the assumptions of Theorem 2.2.2, *of the* g_0th *structural equation by the a priori restrictions* (2.24) *is that the number of independent restrictions expressed by* Φ_{g_0} *(i.e., the number of rows of* Φ_{g_0}*) be at least* $G - 1$.

2.2.3. *Identifiability almost everywhere in the parameter space.* It is of interest to note that the identifiability of the g_0 th structural equation depends only on the matrix Φ_{g_0} expressing the restrictions on that particular equation, and on the coefficient matrix A. In practice, however, the elements of A are unknown before estimation, and are not known exactly after estimation. Uncertainty with regard to the rank of $A\Phi'_{g_0}$ may therefore remain even after estimation. A question that can be answered before estimation, however, is whether, in cases where $\rho(\Phi_{g_0}) \geq G - 1$, the a priori restrictions (2.24) with $g \neq g_0$, i.e., those restricting structural equations other than the one whose identification is studied, do or do not reduce the rank of $A\Phi'_{g_0}$ below $G - 1$ identically, i.e., for all values of A permitted by (2.24). If the a priori restrictions are in the form (2.20), this question can be decided in a simple way by exhaustive study of a diagram of the elements of the matrix A, in which a zero is entered for every element required to be zero by (2.20), and a cross for every other element. A determinant Δ extracted from A is then not identically zero if at least one term of the determinant can be found that is the product of elements represented by crosses only.

The generalization of this technique to the case of restrictions in the form (2.24), although mathematically interesting, is somewhat complicated in operation. It appears preferable, for this particular purpose, to reduce to a minimum the number of restrictions not in the form (2.20). This can be achieved by retaining the identities as part of the equation system. This is a possible procedure because the assumption of nonsingularity of Σ has not been used in the present discussion of identification problems, and identities are therefore admissible to the equation system of which identification properties are studied. In this case, of course, no identification problem arises with respect to the identities themselves, since these are completely known already. If a few restrictions (2.24) remain that cannot be reduced to the form (2.20), it is advisable first to carry out the analysis indicated while ignoring

those particular restrictions, investigating thereafter whether or not the conclusions are changed by the presence of these restrictions.

Assuming that most situations arising in practice can be dealt with in such manner, we shall not attempt to formulate a general theorem covering all cases under which the rank of $A\Phi'_{g_0}$ may be *identically* below that of Φ_{g_0}. In section 2.2.4, however, we shall discuss some interesting special cases in which this occurs. Meanwhile, it is already possible to make the following generalization: All criteria for identifiability formulated above are in terms of ranks of matrices. Since a determinant is a linear function of any element, and of the elements in any row, a matrix that under linear restrictions of the type (2.20) or (2.24) attains a required rank in one point of the parameter space, attains the required rank in all points except for a set of measure zero. Thus a structural equation that is not identically unidentifiable under such restrictions, is identifiable almost everywhere in the parameter space.

*2.2.4. *Cases where the rank of* $A\Phi'_{g_0}$ *is identically below that of* Φ'_{g_0}. Let us now consider some special cases in which the rank of $A\Phi'_{g_0}$ is identically less than that of Φ'_{g_0} on account of the restrictions imposed on the other structural equations. Since the rank of Φ_{g_0} is at the same time the number R_{g_0} of rows in Φ_{g_0}, every nonvanishing vector λ_{g_0} containing R_{g_0} elements satisfies

$$\Phi'_{g_0}\,\lambda'_{g_0} \neq 0. \tag{2.27}$$

Let $\rho(X)$ represent the rank of any matrix X. Then, if

$$\rho(A\Phi'_{g_0}) < \rho(\Phi'_{g_0}) = R_{g_0} \tag{2.28}$$

for all values of A satisfying (2.24), there exists for every such value of A a nonvanishing vector λ_{g_0} such that

$$A\Phi'_{g_0}\,\lambda'_{g_0} = 0, \quad \text{or} \quad A\xi' = 0, \quad \text{where} \quad \xi' \equiv \Phi'_{g_0}\,\lambda'_{g_0} \neq 0. \tag{2.29}$$

Now there are two possible cases. It may occur that a *constant* vector λ exists for which (2.29) is satisfied by all values of A permitted by (2.24). In this case the restriction

$$\alpha(g)\,\xi' = 0, \tag{2.30}$$

ξ constant, must hold for each row $\alpha(g)$, $g = 1, \ldots, G$, of A, as a consequence of (2.24), i.e., there exist vectors λ_g such that

$$\Phi_g'\lambda_g' = \xi', \qquad g = 1, \ldots, G. \tag{2.31}$$

This means that the restrictions on the individual structural equations imply at least one restriction (2.30), which is common to all of them. The coefficients ξ of this common restriction do not depend on A, but can be determined or selected on the basis of the Φ_g, $g = 1, \ldots, G$, alone. It follows that the number K_x of variables x_k can be reduced by one without changing the nature of the problem studied. This is obvious in the special case that ξ has only one nonvanishing element, say ξ_1, because then the a priori restrictions imply that the variable x_1 does not actually occur in any one of the structural equations. The same conclusion can be drawn as follows if ξ is any other constant vector. Let Ξ be a nonsingular square matrix of order K_x containing ξ as its first row. Then the linear transformation

$$A\,\Xi \equiv A^{\oplus}, \qquad \Xi^{-1}x' \equiv x'^{\oplus}, \tag{2.32}$$

of variables x and coefficients A leads to a situation where the a priori restrictions imply $\alpha_{1g}^{\oplus} = 0$, $g = 1, \ldots, G$, that is, where the variable $x_1^{\oplus}$ does not actually occur.

Alternatively, it may occur that (2.29) can be satisfied only by a vector ξ which itself depends on A. A simple example is that of $G = 3$ equations where the a priori restrictions require a submatrix consisting of the kth and lth columns of A to be as follows:

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \alpha_{3k} & \alpha_{3l} \end{bmatrix}, \tag{2.33}$$

with α_{3k}, α_{3l}, and the remaining elements of $\alpha(1)$ and $\alpha(2)$ unrestricted. Here we must take

$$(2.34) \qquad \xi_k = -\alpha_{3l}, \qquad \xi_l = \alpha_{3k}, \qquad \xi_m = 0 \quad \text{if} \quad m \neq k, \; l,$$

to obtain a restriction (2.30) common to all structural equations. However, the first $G - 1 = 2$ structural equations have two independent constant restrictions in common[1] and are not identifiable.

*2.3. *Treatment of Unidentifiable Structural Equations by Linear Dummy Restrictions*

*2.3.1. *Dummy restrictions to produce formal identifiability.* In case one or more structural equations are not identifiable, this fact should not be allowed to interfere with the estimation of such other equations as are identifiable. We may even wish to go further and estimate such linear functions of the parameters of the unidentifiable equations as are not affected by the lack of identification. From the point of view both of estimation and computation, therefore, it is an important problem to write the parameters (A, Σ) as functions of two sets of parameters, θ_1 and θ_2 say, of which the first uniquely defines a set of points in the restricted (A, Σ) space, while variation of the second set of parameters θ_2 only causes the point (A, Σ) to vary within a set of equivalent points in that space. Using Wald's extension of the identifiability concept to individual parameters [III], we may say that the parameters (A, Σ) are written as functions of a set of identifiable parameters θ_1 and a set of unidentifiable parameters θ_2.

A device whereby the separation of θ_1 and θ_2 can be carried out is the addition to the original a priori restrictions of dummy restrictions such that the so augmented set of restrictions ensures identifiability. The dummy restrictions are then made to contain as many parameters θ_2 as are required to define a point within a set of equivalent points. It will be clear that this device is applicable only within a region of the restricted parameter space, in which the matrices $A\Phi_g'$, $g = 1, \ldots, G$, have constant ranks. For the sake of simplicity, we shall only discuss the case of a region in which the rank of each $A\Phi_g'$ equals that of the corresponding

[1] It is again true here that the variables x_k and x_l occur only in the linear combination $\alpha_{3k}x_k + \alpha_{3l}x_l$, but this does not permit us to reduce the number of variables because α_{3k} and α_{3l} are unknown.

restriction matrix Φ_g'. That is, we shall disregard the case in which the rank of an $A\Phi_g'$ is identically depressed on account of restrictions on the structural equations other than the gth.

*2.3.2. *A lemma.* In preparation for a theorem stating what can be achieved by dummy restrictions, we shall first prove

LEMMA 2.3.2. *If Φ and Ψ are two matrices with an equal number of rows, such that*[1]

$$\rho(\Phi \quad \Psi) < \rho(\Phi) + \rho(\Psi), \tag{2.35}$$

then there exist two nonvanishing vectors λ and μ such that

$$\Phi\lambda' + \Psi\mu' = 0, \qquad \Phi\lambda' \neq 0. \tag{2.36}$$

Let $c(\Phi)$ denote the number of columns of Φ, and assume first as a special case that

$$\rho(\Phi) = c(\Phi), \qquad \rho(\Psi) = c(\Psi). \tag{2.37}$$

(It is only with respect to this special case that the lemma is used in the present section; the further case (2.39) is added for later use, see sections *3.2.5* and *4.3.4.6.*) Then, from (2.35) and (2.37),

$$\rho(\Phi \quad \Psi) < c(\Phi \quad \Psi). \tag{2.38}$$

Hence there exists a nonvanishing vector $[\lambda \quad \mu]$ satisfying the equality in (2.36). However, this cannot be a vector such that λ vanishes, because then the equality in (2.36) would imply $\Psi\mu' = 0$ with $\mu' \neq 0$, which is precluded by the second condition in (2.37). Similarly, the first condition in (2.37) precludes the vanishing of $\Phi\lambda'$ now that $\lambda \neq 0$.

Now assume, more generally, that

$$\rho(\Phi) \leq c(\Phi), \qquad \rho(\Psi) \leq c(\Psi). \tag{2.39}$$

Then it is possible, wherever an inequality sign in (2.39) applies,

[1] Square brackets [ ] denoting matrices are omitted when a matrix appears as argument of the functions $\rho(\ )$, $r(\ )$, $c(\ )$.

to delete one or more columns from Φ or Ψ or both, so as to obtain matrices $\overline{\Phi}$ and $\overline{\Psi}$ respectively such that

$$\rho(\Phi) = \rho(\overline{\Phi}) = c(\overline{\Phi}), \qquad \rho(\Psi) = \rho(\overline{\Psi}) = c(\overline{\Psi}), \tag{2.40}$$

and, from (2.35) and (2.40),

$$\rho(\overline{\Phi} \quad \overline{\Psi}) \leq \rho(\Phi \quad \Psi) < \rho(\overline{\Phi}) + \rho(\overline{\Psi}). \tag{2.41}$$

The conditions (2.40) and (2.41) are equivalent to (2.35) and (2.37) with Φ and Ψ replaced by $\overline{\Phi}$ and $\overline{\Psi}$. Hence there exist nonvanishing vectors $\overline{\lambda}$ and $\overline{\mu}$ such that

$$\overline{\Phi}\,\overline{\lambda}' + \overline{\Psi}\,\overline{\mu}' = 0, \qquad \overline{\Phi}\,\overline{\lambda}' \neq 0. \tag{2.42}$$

By adding zero elements in the proper places, these can be completed to vectors λ and μ satisfying (2.36).

*2.3.3. *The number and type of dummy restrictions required.* We shall now study the impostition of dummy restrictions on the gth structural equation in the neighborhood of such a point A_0 in the space of the parameters A in which

$$\rho(A_0 \cdot \Phi_g') = \rho(\Phi_g'). \tag{2.43}$$

As before, the restriction matrix Φ_g is so chosen that

$$\rho(\Phi_g') = c(\Phi_g') = R_g \tag{2.44}$$

equals the number of independent restrictions imposed on the gth structural equation. Since the gth row of $A_0\Phi_g'$ consists of zeros only, we may as well operate with a matrix ${}_gA_0$ obtained from A_0 by deleting the gth row, and write instead of (2.43)

$$\rho({}_gA_0 \cdot \Phi_g') = \rho(\Phi_g'). \tag{2.45}$$

For reasons of notational symmetry, we shall in the remainder of this section 2.3 occasionally use the symbol K_y, introduced in

section 2.1.2 as synonymous with G (the number of structural equations and of dependent variables). Suppose now that

$$R_g < G - 1 \equiv K_y - 1, \tag{2.46}$$

that is, that the gth structural equation is not identified by the original restrictions (2.24). We shall continue to refer to the restrictions (2.24) as the a priori restrictions, and to regard them as the basis for the concept of equivalence according to Definition 2.1.3.6. We now wish to achieve identification of the gth equation by the addition to (2.24) of dummy restrictions which we denote

$$\alpha(g)\cdot\bar{\Phi}_g^{(0)} = 0, \qquad c(\bar{\Phi}_g^{(0)}) = \bar{R}_g. \tag{2.47}$$

We shall of course require that the dummy restrictions are independent of each other and of the a priori restrictions, i.e.,

$$\rho(\Phi_g' \quad \bar{\Phi}_g^{(0)}) = R_g + \bar{R}_g. \tag{2.48}$$

THEOREM 2.3.3. *If in a point* (A_0, Σ_0) *of the parameter space the following conditions are satisfied*

(a) *the a priori restrictions* (2.24),

(b) *the rank condition* (2.45),

(c) *the insufficiency* (2.46) *of the number* R_g *of independent* (2.44) *a priori restrictions on the* gth *structural equation to identify that equation,*

then there exists a neighborhood N *of* (A_0, Σ_0) *in the a priori restricted parameter space, and a matrix* $\bar{\Phi}_g^{(0)}$ *defining*

$$\bar{R}_g = G - 1 - R_g \tag{2.49}$$

dummy restrictions (2.47) *on the* gth *structural equation, which are independent, mutually and from the a priori restrictions, such that*

(i) *to each point* (A, Σ) *in* N *can be found at*

least one equivalent point satisfying the dummy restrictions (2.47),

(ii) *the a priori and dummy restrictions taken in combination identify the* gth *structural equation.*

The statement (i) in this theorem assures that no probability distribution of the observations permitted by the a priori restrictions is deprived of representation by imposing dummy restrictions.

Introducing the notation

$$\overline{\Phi}'(g)^{(0)} \equiv [\ \Phi_g' \quad \overline{\Phi}_g'^{(0)}\ ], \tag{2.50}$$

we shall first show that there exists a matrix $\overline{\Phi}_g'^{(0)}$ with the number of columns $\overline{R}_g$ as given by (2.49), such that

$$\rho({}_gA_0 \cdot \overline{\Phi}'(g)^{(0)}) = G - 1 = K_y - 1 = \rho(\overline{\Phi}'(g)^{(0)}). \tag{2.51}$$

The first step is to choose an arbitrary orthogonal complement of ${}_gA_0$, that is, a matrix $\overline{\overline{\Phi}}_g^{(0)}$ such that both

$$\rho(\overline{\overline{\Phi}}_g'^{(0)}) = c(\overline{\overline{\Phi}}_g'^{(0)}) = K_x - K_y + 1 = K_z + 1, \tag{2.52}$$

say, and

$${}_gA_0 \cdot \overline{\overline{\Phi}}'(g)^{(0)} = 0. \tag{2.53}$$

However this choice is made, we must have

$$\rho(\ \Phi_g' \quad \overline{\overline{\Phi}}_g'^{(0)}\ ) = c(\ \Phi_g' \quad \overline{\overline{\Phi}}_g'^{(0)}\ ) = R_g + K_z + 1. \tag{2.54}$$

For otherwise, according to Lemma 2.3.2, nonvanishing vectors λ_g and λ_g exist such that

$$\Phi_g' \cdot \lambda_g' + \overline{\overline{\Phi}}_g'^{(0)} \cdot \overline{\overline{\lambda}}_g' = 0, \tag{2.55}$$

and, on account of (2.53),

$${}_gA \cdot \Phi_g' \cdot \lambda_g' = 0, \tag{2.56}$$

which is incompatible with (2.44) and (2.45).

Because of (2.54), we can, as the next step, choose a matrix $\bar{\Phi}_g'^{(0)}$ with a number of rows $\bar{R}_g$ as given by (2.49), such that

$$(2.57) \qquad \bar{\bar{\Phi}}'(g) \equiv [\ \Phi_g' \quad \bar{\Phi}_g'^{(0)} \quad \bar{\bar{\Phi}}_g'^{(0)}\ ] \equiv [\ \bar{\Phi}'(g)^{(0)} \quad \bar{\bar{\Phi}}_g'^{(0)}\ ]$$

is a nonsingular square matrix of order $K_x \equiv K_y + K_z$. The nonsingularity of $\bar{\bar{\Phi}}'(g)^{(0)}$ ensures that (2.48) is satisfied. It follows further that

$$(2.58) \qquad \rho({}_gA_0 \cdot \bar{\bar{\Phi}}'(g)^{(0)}) = \rho({}_gA_0),$$

because postmultiplication with a nonsingular square matrix does not affect rank. On the other hand, from (2.53) and (2.57),

$$(2.59) \qquad \rho({}_gA_0 \cdot \bar{\bar{\Phi}}'(g)^{(0)}) = \rho({}_gA_0 \cdot \bar{\Phi}'(g)^{(0)}).$$

Finally, since the structural equations are independent [see also (1.6)],

$$(2.60) \qquad \rho(A_0) = K_y = 1 + \rho({}_gA_0).$$

Combining (2.58), (2.59), and (2.60), we have completed the proof of the first equality in (2.51). The second equality in (2.51) follows directly from (2.48), (2.49), and the nonsingularity of $\bar{\bar{\Phi}}(g)^{(0)}$

Since a determinant is a continuous function of its elements, it follows from (2.51) that

$$(2.61) \qquad \rho({}_gA \cdot \bar{\bar{\Phi}}'(g)^{(0)}) = K_y - 1 = \rho(\bar{\Phi}'(g)^{(0)})$$

in a neighborhood ${}_gN_\alpha$ of the point ${}_gA_0$ in the space of the parameters ${}_gA$ subject to the a priori restrictions (2.24).

Let N be a neighborhood of the point (A_0, Σ_0) in the restricted parameter space such that for all points (A_0, Σ_0) in N, the coor-

dinates ${}_gA$ define a point ${}_gA$ in ${}_gN_\alpha$. We shall now show that N and $\Phi_g^{(0)}$ have the properties required in the theorem.

To satisfy condition (i) we associate with any point (A, Σ) of N, a point

$$\bar{A} = \Upsilon A, \qquad \bar{\Sigma} = \Upsilon \Sigma \Upsilon', \tag{2.62}$$

by the following choice of the transformation Υ,

$$\Upsilon = \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ v_{g1} & \dots & v_{g,g-1} & 1 & v_{g,g+1} & \dots & v_{gG} \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix}, \tag{2.63}$$

in which the vector

$${}_gv(g) = [\, v_{g1} \quad \dots \quad v_{g,g-1} \quad v_{g,g+1} \quad \dots \quad v_{gG} \,] \tag{2.64}$$

of as yet unspecified elements in the gth row $v(g)$ of Υ is determined by

$${}_gv(g) \cdot {}_gA \cdot \bar{\Phi}'(g)^{(0)} = -\alpha(g) \cdot \bar{\Phi}'(g)^{(0)}. \tag{2.65}$$

As a consequence of (2.61), there is always one and only one solution ${}_gv(g)$ to this condition. Rewriting (2.65) as

$$\alpha(g) \cdot \bar{\Phi}'(g)^{(0)} = v(g) \cdot A \cdot \bar{\Phi}'(g)^{(0)} = 0, \tag{2.66}$$

we see, in connection with (2.50), that the point (2.62) satisfies both the a priori conditions (2.24) and the dummy restrictions (2.47).

Finally the uniqueness of the solution ${}_g\upsilon(g)$ of (2.65) implies that any other point $(\bar{A}^{\oplus}, \bar{\Sigma}^{\oplus})$ that is obtainable from (A, Σ) by a linear transformation (2.62) and satisfies the combined (a priori and dummy) restrictions is obtainable from $(\bar{A}, \bar{\Sigma})$ by a transformation that is trivial with respect to the gth structural equation (see Definition 2.1.5.2). For, the combined restrictions on the gth equation are expressed by (2.66), which permit free choice only of the diagonal element υ_{gg} in $\upsilon(g)$. This shows that condition (ii) of the theorem is also satisfied.

*2.3.4. *The degree of indeterminacy of an a priori unidentifiable structural equation.* It is of interest now to consider the reverse problem that arises after estimation has been carried out subject to dummy restrictions added to obtain formal identifiability: From a parameter point $(\bar{A}, \bar{\Sigma})$ that satisfies both the a priori and the dummy restrictions, reconstruct the set of all equivalent points (A, Σ) in the restricted parameter space; i.e., all points obtained through linear transformations inverse to (2.62) and satisfying the a priori restrictions but not necessarily the dummy restrictions. Since the identification problem under the restrictions (2.24) can be studied for each equation separately, it is sufficient to study those transformations for which A differs from $\bar{A}$ only as regards the gth row $\alpha(g)$ or

$$ {}_gA = {}_g\bar{A}. \tag{2.67} $$

This means that we can confine ourselves to transformations

$$ A = \Upsilon^{-1}\,\bar{A}, \qquad \Sigma = \Upsilon^{-1}\,\bar{\Sigma}\,\Upsilon'^{-1}, \tag{2.68} $$

with a matrix Υ^{-1} inverse to a matrix Υ of type (2.63). It is easily seen that Υ^{-1} itself then is of the same type, with nondiagonal elements in the gth row equal to

$$ \upsilon^{gh} = -\upsilon_{gh}, \qquad h \neq g. \tag{2.69} $$

For the gth row of A this gives us in particular

$$ \alpha(g) = \bar{\alpha}(g) - {}_g\upsilon(g)\cdot {}_g\bar{A}, \tag{2.70} $$

and, through postmultiplication with $\Phi_g'^{(0)}$ and $\overline{\Phi}_g'^{(0)}$ respectively, using (2.24), (2.50), and (2.66),

$$(2.71)\quad \begin{cases} (2.71r) & 0 = -\ {}_g\upsilon(g)\cdot {}_g\overline{\mathrm{A}}\cdot\Phi_g'^{(0)}, \\ (2.71\bar{r}) & \alpha(g)\cdot\overline{\Phi}_g'^{(0)} = -\ {}_g\upsilon(g)\cdot {}_g\mathrm{A}\cdot\overline{\Phi}_g'^{(0)}. \end{cases}$$

Of these conditions, (2.71r) expresses the restrictions on Υ^{-1}, arising through (2.69) from the fact that $\alpha(g)$ must satisfy the a priori restrictions. The left-hand member in (2.71$\bar{r}$) arises from the fact that $\alpha(g)$ is not bound by the dummy restrictions. It is easily seen that precisely those linear combinations of the elements of $\alpha(g)$ which are the elements of the vector

$$(2.72)\qquad \overline{\theta}_g \equiv \alpha(g)\ \overline{\Phi}_g'^{(0)}$$

can be chosen arbitrarily before, according to (2.61), (2.67), and (2.69), Υ^{-1} and therewith $\alpha(g)$ is fully determined. Therewith we have proved:

THEOREM 2.3.4. *If under the conditions of Theorem* 2.3.3 *the parameter point* $(\overline{\mathrm{A}}, \overline{\Sigma})$ *satisfies both the a priori restrictions* (2.24) *and the dummy restrictions* (2.47) *identifying the* gth *structural equation, the set of points* (A, Σ), *equivalent to* $(\overline{\mathrm{A}}, \overline{\Sigma})$ *but satisfying only the a priori restrictions, is obtained from the latter point by a transformation* (2.68) *with a matrix* Υ^{-1} *of which the* gth *row is through* (2.69) *determined from* (2.71), *with arbitrary choice of the vector* $\overline{\theta}_g$ *of dummy parameters* (2.72).

The theorem does not stipulate that the remaining rows of Υ^{-1} correspond to the form (2.63), since no assumptions were made as to the identifiability of the remaining structural equations. It will be clear, however, that the transformed point (2.68) must satisfy the a priori restrictions (2.24) on all structural equations.

2.4. *Identification of a Set of Structural Equations under Linear and Bilinear Restrictions*

2.4.1. Problems arising from additional types of restrictions. We shall now discuss identification problems that arise if two fur-

ther types of a priori restrictions are added, each of which binds the parameters of one structural equation to those of another. The first type is given by (1.13) or (1.14), which is rewritten in slightly different notation in (2.73α). The second type (2.73σ) expresses absence of correlation (or, under the normality assumption, independence) between the disturbances in two equations.

$$(2.73)\quad \begin{cases} (2.73\alpha) & \begin{bmatrix} \alpha_{g_r k_r} & \alpha_{g_r l_r} \\ \alpha_{h_r k_r} & \alpha_{h_r l_r} \end{bmatrix} = 0, \qquad r = 1, \ldots, R_\alpha^{(2)}, \\ & \qquad\qquad\qquad\qquad\qquad\qquad g_r < h_1, \\ (2.73\sigma) & \sigma_{g_r h_r} = 0, \qquad r = R_\alpha^{(2)} + 1, \ldots, R_\alpha^{(2)} + R_\sigma. \end{cases}$$

For the time being, we shall assume no particular pattern for the occurence of zeros among the elements of Σ. Later we shall say a few words concerning the special pattern (1.15).

If a given structural equation can already be identified on the basis of the restrictions (2.24) alone, the imposition of additional restrictions (2.73) of course does not detract from the identifiability of that equation. The only identification problem of interest in connection with the restrictions (2.73) is therefore under what circumstances an equation not identifiable on the basis of (2.24) alone can be identified if the restrictions (2.73) are added.

Each of the restrictions (2.73) connects two structural equations, numbered g_r and h_r, which we shall call the two equations *referred to* in that restriction. [Similarly, we shall speak of the one equation referred to by any one of the restrictions (2.24).] The restrictions (2.73) link up the identification problem of individual equations, and statements regarding identifiability will therefore in general relate either to the whole set of structural equations (2.3) or to subsets thereof which only in special cases may consist of one single equation.

2.4.2. *The additional restrictions are bilinear.* If the parameters A, Σ satisfy the a priori restrictions (2.24) and (2.73), to which we shall add the normalization rules (1.17a), the requirement that the transformed parameters (2.18) shall satisfy the same restrictions leads to the following expression of the

$R_\alpha^{(1)} + R_\alpha^{(2)} + R_\sigma + G$ a priori restrictions in terms of the rows υ_g, $g = 1, \ldots, G$, of the transformation matrix Υ:

$$(R_\alpha^{(1)} + G \text{ restrictions}) \begin{cases} (2.74lh) & \upsilon(g)\cdot A\cdot\Phi_g' = 0, \quad R \text{ restrictions}, \\ (2.74ln) & \upsilon(g)\cdot A\cdot\iota(i_g) = 1, \quad \text{one restriction}, \end{cases} \qquad g = 1, \ldots, G. \tag{2.74l}$$

$$(R_\alpha^{(2)} + R_\sigma \text{ restrictions}) \begin{cases} (2.74b\alpha) & \begin{bmatrix} \upsilon(g_r)\cdot A\cdot\iota'(k_r) & \upsilon(g_r)\cdot A\cdot\iota'(l_r) \\ \upsilon(h_r)\cdot A\cdot\iota'(k_r) & \upsilon(h_r)\cdot A\cdot\iota'(l_r) \end{bmatrix} = 0, \\ & r = 1, \ldots, R_\alpha^{(2)}, \\ (2.74b\sigma) & \upsilon(g_r)\cdot\Sigma\cdot\upsilon'(h_r) = 0, \\ & r = R_\alpha^{(2)} + 1, \ldots, R_\alpha^{(2)} + R_\sigma. \end{cases} \tag{2.74b}$$

(2.74)

Here $\iota(k)$ is the kth row of the unit matrix of order K_x; i.e., a vector of which the kth element is 1 and all other elements are 0. The normalization rules (2.74l) are nonhomogeneous in the elements of Υ.

All other restrictions in (2.74) will be referred to as the homogeneous restrictions. The two types of restrictions under (2.74b), while being quadratic in the elements of Υ, are linear in the elements of any row of Υ. For this reason we shall refer to (2.74b) as the *bilinear* restrictions.

The occurrence of bilinear restrictions greatly complicates the identification problem. The following discussion should be regarded as a first exploration of the field, and does not lead to firm criteria such as were derived for linear restrictions only.

2.4.3. *The solution* $\Upsilon = I$ *is always present.* It is important to note that, because the parameters A and Σ are assumed to satisfy the a priori restrictions (1.17a), (2.24), and (2.73), the system of restrictions (2.74) and any of its subsystems always

permit of one particular solution Υ, viz., the identical transformation

$$\Upsilon = I, \tag{2.75}$$

where I represents the unit matrix of order K. For this reason there can never be too many compatible restrictions for identification. There can only be either too few or a sufficient number.

2.4.4. *Unique, multiple, and complete identification.* We shall discuss the identifiability of a given subset S of the structural equations (2.3), containing the H equations for which $g = g_1, \ldots, g_H$, on the basis of a given subset R of the restrictions (2.74). This discussion is concerned with matrices of the type

$$\Upsilon^S \equiv \begin{bmatrix} v_{g_1 1} & \cdots & v_{g_1 G} \\ \cdot & \cdots & \cdot \\ v_{g_H 1} & \cdots & v_{g_H G} \end{bmatrix} \tag{2.76}$$

combining the rows, corresponding to the equations of S, of a solution Υ of the subset R of the restrictions (2.74).

DEFINITION 2.4.4.1. *A subset S of the structural equations will be said to be uniquely identifiable by a subset R' of the restrictions (2.74) that includes all normalization rules (2.74ln) relevant to S, if for all solutions Υ of R' we have*

$$\Upsilon = I_{[G_S, G]}, \tag{2.77}$$

where $I_{[G_S, G]}$ represents a unit matrix of order G_S adjoined to a zero matrix of G columns. We shall speak of multiple identification if to all solutions Υ of R' there corresponds a finite number of different matrices Υ^S that exceeds one, and of incomplete identification if the number of different matrices Υ^S is infinite. Complete identification means either unique or multiple identification.

Incomplete identifiability is, of course, synonymous with un-

identifiability. In order to obtain conformity with Definition 2.1.5.3 we add:

DEFINITION 2.4.4.2. *A subset S of the structural equations will be said to be uniquely, multiply, or incompletely identifiable with respect to a subset R of the homogeneous restrictions in* (2.74) *if, after addition to R of the normalization rules relevant to S, it is so identifiable in the sense of Definition* 2.4.4.1.

The possibility of complete but multiple identification arises, of course, from the presence of quadratic restrictions in (2.74). A simple example is that of three equations in three variables, in which the a priori restrictions assume the form

$$(2.78)\quad \begin{cases} (2.78a) & \alpha_{11} = \alpha_{22} = \alpha_{33} = 0, \\ (2.78b) & \sigma_{12} = \sigma_{23} = \sigma_{31} = 0, \quad \sigma_{11} = \sigma_{22} = \sigma_{33} = 1. \end{cases}$$

If for convenience we impose the normalization condition in (2.78b) only on the original matrix Σ but not on $\Sigma^{\oplus}$, we find that the remaining conditions (2.78) for the transformed matrices $A^{\oplus}$ and $\Sigma^{\oplus}$ permit, not only of any transformation with a diagonal matrix Υ corresponding to a change of scales, but also of the transformation of which the elements are obtained from

$$(2.79)\quad \begin{aligned} \upsilon_{11} &= \alpha_{12}\,\alpha_{13}\,(\alpha_{21}^2 + \alpha_{31}^2) + \alpha_{21}\,\alpha_{31}\,\alpha_{23}\,\alpha_{32}\,, \\ \upsilon_{12} &= \alpha_{31}\,(\alpha_{12}\,\alpha_{23}\,\alpha_{31} - \alpha_{13}\,\alpha_{32}\,\alpha_{21}), \\ \upsilon_{13} &= -\alpha_{21}\,(\alpha_{12}\,\alpha_{23}\,\alpha_{31} - \alpha_{13}\,\alpha_{32}\,\alpha_{21}), \end{aligned}$$

by cyclical permutation (followed again by any change of scales).

In the case of complete but multiple identification, the number of solutions can sometimes be reduced, or even unique identification can be achieved, through additional a priori restrictions in the form of inequalities (see section *1*). The use of such restrictions with respect to the elements of Σ in a case of incomplete identification has been demonstrated by Marschak and Andrews [1944].

In the remainder of this section we shall concentrate on the question of completeness or incompleteness of identifiability, irrespective of the number of solutions in the case of complete identification. We shall first make some remarks on the counting of

restrictions as an indication of identifiability. Thereafter, we shall discuss in which respects procedures based on counting alone may be insufficient to establish either identifiability or lack of identifiability.

*2.4.5. *Criteria of identifiability based on the counting of restrictions.* Apart from H normalization factors (one for each row), the matrix (2.76) contains $H(G-1)$ unknowns. If R contains at least $H(G-1)$ homogeneous restrictions, however, these may still be unevenly divided between the different rows of Υ^S, leaving some rows undetermined for lack of an adequate number of restrictions. In formulating principles for counting restrictions on individual rows, it is necessary to remember that each of the bilinear restrictions (2.74*b*) refers to two rows of Υ^S, and obviously should not be counted as a new restriction with regard to the identification of each of those two rows. These considerations lead to the following definition.

DEFINITION 2.4.5. *The subset R of the restrictions* (2.74) *will be said to be adequate in number and variety with respect to (the identification of) the subset S of the structural equations* (2.3) *if it is possible to assign each bilinear restriction* (2.74) *occurring in R to one of the two equations* (2.3) *to which it refers in such a way, that the number of homogeneous linear equations* (2.74*lh*) *in R referring to, plus the number of bilinear conditions* (2.74*b*) *in R assigned to, each equation of S is at least $G-1$.*

*2.4.6. *The completed subset of structural equations.* The concept introduced by Definition 2.4.5 can be applied in particular to the identification of the set of all equations (2.3) by the set of all restrictions (2.74). If some of the equations (2.3) cannot be identified for lack of an adequate number and variety of a priori restrictions, it becomes necessary to develop criteria that are of assistance in finding the largest subset S that can be identified.

DEFINITION 2.4.6.1. *The subset R_S of a priori restrictions* (2.74) *associated with a given subset S of the structural equations consists of all homogeneous linear conditions* (2.74*lh*) *referring to an equation of S and all bilinear conditions* (2.74*b*) *referring to two equations of S.*

DEFINITION 2.4.6.2. *A subset S_0 of the structural equations will be called a completed subset if a) the subset R_{S_0} of restric-*

tions (2.74) associated with S_0 is adequate in number and variety with respect to S_0 and b) there exists no larger subset S' that contains all the equations of S_0 and one or more others besides, and with respect to which the associated subset $R_{S'}$ of the restrictions (2.74) is adequate in number and variety.

THEOREM 2.4.6. *There is at most one completed subset S_0 of the structural equations (2.3).*

Suppose there are two different subsets S_1 and S_2 satisfying Definition 2.4.6.2. Obviously S_1 cannot be a subset of S_2 or vice versa, because then either S_1 or S_2 would not be a completed subset. On the other hand, if S_1 and S_2 have no equations in common, the combination $S_1 + S_2$ of both sets would possess an associated subset $R_{S_1+S_2} = R_{S_1} + R_{S_2}$ of restrictions (2.74) which is adequate in number and variety with respect to $S_1 + S_2$, contrary to the assumption made about S_1. In the third possible case, in which S_1 and S_2 have a set S_1S_2 in common, which differs from both S_1 and S_2, it can likewise be inferred that, contrary to the assumption regarding S_1, the set $R_{S_1+S_2}$ of restrictions (2.74) associated with the combination $S_1 + S_2$ is of the requisite number and variety with respect to $S_1 + S_2$. This is seen by assigning the bilinear restrictions (2.74b) in $R_{S_1+S_2}$ (which contains the combination $R_{S_1} + R_{S_2}$ of R_{S_1} and R_{S_2}) in the following way. The bilinear restrictions (A) (see Fig. 2.4.6) in R_S are assigned in the same way as they were assigned to meet the requirements of Definition 2.4.5 with respect to S_1. The equations of S_1 are thereby provided with an adequate number and variety of a priori restrictions. Then the bilinear restrictions (B) in R_{S_2} but not in R_{S_1} are assigned as they were to meet the requirements of Definition 2.4.5 with respect to S_2. This adequately provides the equations in $S_3 = S_2 - S_1S_2$, that is, the equations in S_2 but not in S_1, because the restrictions in R_{S_1} now excluded from consideration do not refer to these

equations. It follows that, irrespective of how any member of the third group of bilinear restrictions in $R_{S_1 + S_2}$, namely those neither in R_{S_1} nor in R_{S_2}, are assigned, the requirements of Definition 2.4.5 are always met with respect to $S_1 + S_2$. The assumption of two different completed subsets S_1 and S_2 has therewith been disproved.

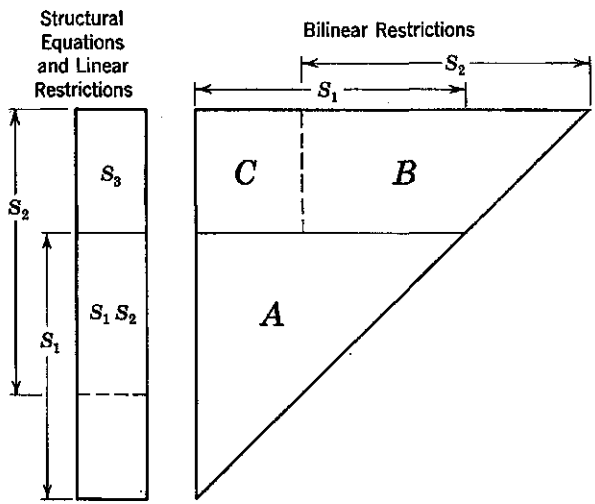


Figure 2.4.6

*2.4.7. *Construction of the completed subset.* It is of interest to give an example in which the completed subset can easily be constructed. This is the case in which the bilinear restrictions require Σ to be diagonal, while the structural equations can be ordered in such a way that there are $g-1$ homogeneous linear restrictions on the gth equation. Then the Gth structural equation is subject to an adequate number of linear restrictions alone, and the $G-1$ bilinear restrictions

$$\sigma_{Gg} = 0, \qquad g = 1, \ldots, G-1, \tag{2.80}$$

referring to it can be assigned one by one to each of the first $G-1$ equations. This provides the $(G-1)$th structural equation

with an adequate number of restrictions, and the $G-2$ remaining bilinear restrictions

$$\sigma_{G-1,g} = 0, \qquad g = 1, \ldots, G-2, \tag{2.81}$$

referring to it can now be assigned to the first $G-2$ equations. A repetition of this process shows that the completed subset S_0 contains all structural equations.[1]

This example suggests a practical procedure for constructing the completed subset S_0. First the set S_1 of those structural equations, for which the linear restrictions (2.74*b*) alone are adequate in number and variety, is included in S_0. Then bilinear restrictions connecting the equations of S_1 with the remaining ones are assigned to equations outside S_1, and the equations thereby provided with an adequate number of restrictions are included in S_0. This process is repeated until it has become impossible to include in S_0 further equations *one by one*. Thereafter, it may still be possible to include small sets of three or more, counting also bilinear restrictions connecting the equations being included as a set.

*2.4.8. *Lack of sufficiency of criteria based on counting.* It has already been indicated that counting of restrictions alone is inconclusive in establishing identifiability. The condition that a structural equation belongs to the completed subset S_0 is neither necessary nor sufficient for its identifiability. Unfortunately, the formulation of necessary and sufficient conditions generalizing those established for linear restrictions is a task beset with considerable difficulties owing to the presence of nonlinear restrictions. Nevertheless, cases in which equations inside S are unidentifiable, or equations outside S_0 are identifiable, are "exceptional" in one sense or another, and we shall presently discuss the nature of the "exceptions."

[1] An additional point of interest in this example is that the rows of Υ can be obtained successively (from the bottom row up), each as the solution of a linear equation system. Therefore, if, under the a priori restrictions stated, the set of structural equations is completely identifiable at all, it is uniquely identifiable, in spite of the fact that among the restrictions imposed, $G(G-1)/2$ are bilinear.

A structural equation belonging to S_0 may fail to be identifiable (as was also found in the case of linear restrictions discussed earlier in this section) through a functional dependence of the relevant restrictions (2.74) on Υ. Such functional dependence may come about because the parameters A, Σ happen to fall within a set of measure zero in the parameter space. Or it may even come about everywhere in the parameter space, as a result of the special nature of the a priori restrictions. Examples of the latter possibility were discussed in section 2.2.4. Another simple example is the case where the a priori restrictions are invariant for the interchange of two of the structural equations. These two equations are then inevitably unidentifiable, because the only case in which an interchange of two equations is innocuous, i.e., the case of complete equality of corresponding coefficients "α" and of corresponding covariances "σ" connecting the two equations with other equations of the system, is precluded by the assumed nonsingularity of B.

*2.4.9. *Lack of necessity of criteria based on counting.* A new element in the situation, which did not arise under linear restrictions only, is the fact that to belong to the completed subset is not even necessary for identifiability of a given structural equation. This is due to two restrictions on the parameter space which follow from the nature of the problem studied. The first of these is the restriction to real values of the parameters A, Σ. This restriction had no effect under linear a priori restrictions, since linear systems of equations with real coefficients only permit of real solutions. Quadratic or even bilinear systems possess no such property. Therefore, the possibility exists in the present case, that the a priori restrictions (including rules of normalization) confine Υ to a point set in the complex space of all its elements, of which all *real* points are such that the g_0th row of Υ equals the corresponding row of the unit matrix – even though the g_0th structural equation be outside of the completed subset S_0. The condition that this shall occur is in the nature of a tangency condition.[1] We have not attempted either to prove (by the construction of an example) the possibility of such an occurrence under bilinear restrictions, or to prove its impossibility. One would expect such a tangency to occur only on a point set of measure zero in the parameter space, but the possibility cannot be excluded

[1]An ellipsoid and a plane in three-dimensional space may have only one real point in common, although they represent only two inhomogeneous equations in three unknowns.

without proof that it could occur everywhere in the parameter space as a result of a special choice of a priori restrictions.

*2.4.10. *Restrictions requiring a certain partitioning of* A *and* Σ. The second relevant restriction on the parameter space is the nonsingularity of B, and therefore of Υ. That this restriction may affect the identification problem appears from a constructed example which was kindly brought to our attention by A. Wald. The following formulation contains Wald's example as a special case.

Let the a priori restrictions be such that, if the structural equations are in a certain way exhaustively subdivided into two sets, S_{I} and S_{II}, and if at the same time the dependent variables are arranged in a certain order, the matrices B and Σ partition as follows:

$$(2.82) \qquad B = \begin{bmatrix} B_{\mathrm{I\,I}} & B_{\mathrm{I\,II}} \\ 0 & B_{\mathrm{II\,II}} \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{\mathrm{I\,I}} & 0 \\ 0 & \Sigma_{\mathrm{II\,II}} \end{bmatrix}.$$

In addition to the restrictions implied in (2.82), there may be further linear and bilinear restrictions involving the elements of $\Sigma_{\mathrm{I\,I}}$, $\Sigma_{\mathrm{II\,II}}$, $B_{\mathrm{I\,I}}$, $B_{\mathrm{I\,II}}$, $B_{\mathrm{II\,II}}$ and the remaining elements of A.

In order that the transformation (2.16) shall preserve the partitioning (2.82), we must have

$$(2.83) \begin{cases} (2.83\beta) & B^{\oplus}_{\mathrm{II\,I}} = \Upsilon_{\mathrm{II\,I}} B_{\mathrm{I\,I}} = 0, \\ (2.83\sigma) & \Sigma^{\oplus}_{\mathrm{II\,I}} = \Upsilon_{\mathrm{II\,I}} \Sigma_{\mathrm{I\,I}} \Upsilon_{\mathrm{I\,I}} + \Upsilon_{\mathrm{II\,II}} \Sigma_{\mathrm{II\,II}} \Upsilon_{\mathrm{I\,II}} = 0. \end{cases}$$

Now the nonsingularity of B requires that $B_{\mathrm{I\,I}}$ be nonsingular. It follows from (2.83β) that

$$(2.84) \qquad \Upsilon_{\mathrm{II\,I}} = 0.$$

Consequently, the nonsingularity of Υ requires that $\Upsilon_{\mathrm{II\,II}}$ be nonsingular. From this, (2.83σ), (2.74), and the nonsingularity of $\Sigma_{\mathrm{II\,II}}$, it also follows that

$$(2.85) \qquad \Upsilon_{\mathrm{I\,II}} = 0,$$

so that Υ partitions as follows

$$\Upsilon = \begin{bmatrix} \Upsilon_{\mathrm{I\,I}} & 0 \\ 0 & \Upsilon_{\mathrm{II\,II}} \end{bmatrix}. \tag{2.86}$$

This result means that, if no *further* bilinear restrictions connecting a structural equation of S_{I} with one of S_{II} are introduced, the identification problems of the two sets of structural equations have been separated. For the only transformations permitted by (2.86) are those within each set of equations.

Let there be G_{I} structural equations in S_{I}, G_{II} in S_{II}, with $G_{\mathrm{I}} + G_{\mathrm{II}} = G$. Then a counting criterion for the separate identifiability of the equations of S_{I}, or a subset thereof, on the basis of restrictions *additional* to (2.82) which refer exclusively to equations of S_{I}, can be formulated as follows: A completed subset $S_{\mathrm{I}}^{(0)}$ of S_{I} is again defined by Definition 2.4.6.2, with this modification, that only restrictions additional to (2.82) are counted; and that their number and variety is deemed adequate if the requirements of Definition 2.4.5 are met with G_{I} substituted for G.

Now it is possible for $S_{\mathrm{I}}^{(0)}$ to be nonempty, or even to contain all equations of S_{I}, even though the unmodified application of Definition 2.4.6.2 to the total set $S_{\mathrm{I+II}}$ of G structural equations leads to an empty completed subset S_0 of $S_{\mathrm{I+II}}$. This is seen most clearly in Wald's example, which takes $G_{\mathrm{I}} = 1$, $G_{\mathrm{II}} \geq 2$, and assumes no restrictions besides (2.82). Then $G_{\mathrm{I}} - 1 = 0$, and $S_{\mathrm{I}}^{(0)}$ consists of the one equation in S_{I}, whereas S_0 is empty. Similar examples with higher values of G_{I} can easily be constructed.

The modified counting criterion just indicated for the identifiability of structural equations in S_{I}, and a similar criterion with reference to S_{II}, can be subsumed under a more general criterion applicable to the full set S of structural equations. In this general criterion both the restrictions (2.82) and any additional restrictions are counted under the unmodified Definitions 2.4.5 and 2.4.6.2, but the $G_{\mathrm{I}}\,G_{\mathrm{II}}$ restrictions on Σ in (2.82) are to be counted as equivalent to the *linear* restrictions (2.85) on $\Upsilon_{\mathrm{I\,II}}$, i.e., on the rows of Υ corresponding to the equations of

S_I, with G_{II} restrictions falling on each of the G_I equations of S_I. This procedure is justified because any attempt to construe the bilinear conditions (2.83σ) arising from (2.82) as restrictions on the equations of S_{II} would require $\Upsilon_{I\ II}$ to have a positive rank, which was shown to be incompatible with the nonsingularity of Υ. This general criterion can also be used if the additional restrictions contain further bilinear restrictions connecting an equation of S_I with one of S_{II}.

The foregoing discussion also applies in the more general case in which, instead of the partitioning (2.82) of B, we have (after some rearrangement of the variables x_k into two groups 1 and 2) a similar partitioning

$$\mathrm{A} = \begin{bmatrix} \mathrm{A}_{I1} & \mathrm{A}_{I2} \\ 0 & \mathrm{A}_{II2} \end{bmatrix} \tag{2.87}$$

of the rectangular matrix A, provided A_{I1} is square and nonsingular. The case (2.82) in which A_{I1} is a submatrix of B, however, is especially important, and will be studied further in section 3.2.7 in connection with estimation problems.

2.4.11. Other special cases. In special cases where the number of structural equations is moderate or the number of bilinear restrictions small, or both conditions hold, a more conclusive discussion of the identification problem than was given here for the general case, may be more easily possible. An example is the study of the measurement of production functions [Marschak and Andrews, 1944] already referred to.

Another example may be briefly indicated without attempting rigorous statement. This is the case in which no structural equation is subject to fewer than $G-2$ homogeneous linear restrictions, while for the set S of those equations that do not possess at least the adequate number $G-1$ of such restrictions, the deficiency is just made up by the bilinear restrictions between them. Further analysis then needs to be concerned only with the equations of S, and with the matrix Υ^S containing the corresponding rows of Υ. The elements of any row of Υ^S can now be expressed as a linear function of two of them. Writing θ_g for the ratio of those elements in the row υ_g, we easily see that the bilinear re-

lations take the form

$$\varkappa_{gh}\,\theta_g\,\theta_h + \lambda_{gh}\,\theta_g + \mu_{gh}\,\theta_h + \nu_{gh} = 0. \tag{2.88}$$

Since by assumption the number and variety of linear and bilinear restrictions on the equations of S is just adequate, each variable θ_h enters into two restrictions (2.88) together with θ_g and θ_i respectively, say. The elimination of θ_h from these two restrictions leads to the same type of restriction between θ_g and θ_i. Continuation of this process of elimination must finally lead to a restriction

$$\varkappa_g\,\theta_g^2 + \lambda_g\,\theta_g + \nu_g = 0 \tag{2.89}$$

connecting θ_g with itself. If all bilinear restrictions have been eliminated in this process, there is unique, multiple, or incomplete identification, according as (2.89) has one, two, or infinitely many solutions. If there are one or more other sets of bilinear restrictions connecting other subsets of the rows of Υ^S, these must be investigated in the same manner, until all bilinear restrictions have been accounted for.

2.5. *Incompleteness of the Present Discussion of Identification Problems*

2.5.1. Dummy restrictions if some a priori restrictions are bilinear. It has already been pointed out that the present approach has not led to necessary and sufficient conditions for identifiability in the general case of linear and bilinear restrictions, and is not likely to do so without considerable further study. For that reason, no study has been made of the degree of indeterminacy of a priori unidentifiable structural equations with the help of dummy restrictions.

2.5.2. Wald's criterion for identifiability. In [III], by a different approach, Wald obtains a criterion in terms of ranks of matrices, which is both necessary and sufficient, applies to each parameter separately rather than to all parameters of a structural equation as a group, and permits a much more general class of a priori restrictions. Against these advantages must be set the fact

that the matrices whose rank must be examined generally have a much higher order, equal to or exceeding the square of the number K_x of variables in the structural equations[1] (2.3).

2.5.3. Other indications of possibly incomplete identification. Both Wald's criterion, and the criteria developed in this section, are inevitably stated in terms of the unknown values of the parameters. It has already been indicated that there are exceptional point sets (of measure zero) in the parameter space, in which the matrices involved in the criteria suffer a decrease in rank, and in which therefore the parameters are subject to a greater degree of indeterminacy than was already recognized by the general analysis of the identification problem. The practical question then arises, whether a parameter point within the exceptional set could have produced the actual sample of observations with any degree of likelihood. Fortunately, there are further indications of such an occurrence: the maximum of the likelihood function must then be very "flat" with respect to one or more particular permissible directions in the parameter space – permissible as regards the a priori restrictions. In the most extreme case the maximum is completely "flat," i.e., it is reached along a curve, surface, etc., rather than in a point. Such situations reveal themselves 1) through slow convergence of the iterative computation procedure (in the extreme case through more rapid convergence to a solution which depends on the initial values used, section *4.3.3.11*), and 2) through very high (in the extreme case, infinite) values of the estimated sampling variances of parameters erroneously believed determinate. In this way deficiencies in the analysis of identification problems will come to light in later stages of the investigation. It will be clear, however, that the computational stage can be handled much more efficiently, if all indeterminacies in the parameter space have already been recognized through the study of identification problems.

2.5.4. Identification should be based on a minimum of firmly established assumptions. It is therefore important to make the prior analysis of identification problems as complete and general as possible. In particular, one should avoid as much as possible employing assumptions that might not be satisfied by the data, and which are at the same time essential to the conclusions reached

[1] In the equations (2.3), it will be remembered, values of the same economic variable measured with different time lags are to be considered as different variables.

regarding the identifiability of structural equations. For this reason, the present discussion of identification problems has been made independent of the assumption of normality of the distribution of disturbances. This is in contrast with those parts in subsequent sections, especially the evaluation of sampling variances of maximum-likelihood estimates, in which the normality assumption was already known to be relatively harmless, even if the data do not strictly correspond to it.

For the same reason, the present authors are inclined to rely more firmly and more extensively on restrictions involving the coefficients A that have a good basis in economic considerations, than on restrictions on the covariance matrix Σ of disturbances, at least until the nature of the disturbances has been more fully analyzed by theory and observation.

2.5.5. *Linearity of the structural equations.* The question should be raised whether the assumption that the structural equations are linear does not conceal from view possible further cases of indeterminacy in the measurement of economic relations that need not be strictly linear. To answer this question it is necessary to formulate what is the alternative to linearity. If the alternative is the addition of higher-degree terms in the variables to obtain polynomial expansions, the answer is that the present analysis can be extended to cover such cases as follows. First the present analysis is applied to the equation system obtained by omitting all nonlinear terms to find for any point in the restricted parameter space a set T_L of transformations preserving the a priori restrictions on the linear terms. As long as the vanishing of all nonlinear terms is not excluded a priori, the set T preserving, everywhere in the relevant part of the parameter space, any a priori restrictions on the nonlinear terms in addition to those on the linear terms, can only be a subset of T_L. Because of the algebraic independence of terms of different degrees under linear transformations, T_L can thus be narrowed down to T by successively applying the restrictions, if any, on the terms of each of the higher degrees. This reasoning indicates that the admission of nonlinear terms does not lead to new indeterminacies unsuspected in the linear case.

Another possible situation is that in which the a priori restrictions prescribe linearity for some structural equations, and for some other equations, a type of relationship that excludes linearity (e.g., hyperbolic or exponential). While the general mathe-

matical treatment of identification problems in such cases might be more difficult, we conjecture that again the set T would in some sense be narrower than in the corresponding case where all equations are linear. For this reason, we believe, "mixed" prescriptions of this type, as regards the form of the equations, are more likely to conceal than to reveal cases of indeterminacy of economic parameters, except where indubitable a priori evidence exists as regards the validity of such prescriptions.

2.5.6. *Transformations in the parameter space involving shifts in time.* It should be pointed out that among the assumptions on which the present discussion of identification problems is based, there is still at least one of the undesirable type against which we have just put in a word of warning. That is, there is one assumption that may not be too well fulfilled by the data, whereas its removal may open up new possibilities of indeterminacy. This is the assumption of independence between disturbances in successive time units.

The proof of Theorem 2.1.2 is based on that assumption. One example is sufficient to show that this basic theorem does not hold without the independence assumption. Suppose that we admit serial correlation between disturbances relating to successive time points, but do not think it justified to impose any particular mathematical form on the autocorrelation function[1]. Consider the system of two equations

$$\begin{aligned} \alpha_{110}\, x_1(t) \qquad\quad &+ \alpha_{111}\, x_1(t-1) + \alpha_{121}\, x_2(t-1) = u_1(t), \\ \alpha_{210}\, x_1(t) + \alpha_{220}\, x_2(t) \qquad\quad &= u_2(t), \end{aligned} \tag{2.90}$$

in which the open spaces indicate the coefficients prescribed to be zero. Each of these equations is identifiable under the assumptions of Theorem 2.1.2. But under the present assumptions, the transformation

$$\begin{aligned} u_1^{\oplus}(t) &= u_1(t) + \lambda u_2(t-1), \\ u_2^{\oplus}(t) &= u_2(t), \end{aligned} \tag{2.91}$$

[1] Hurwicz demonstrates in [XI-10.2] that certain specific assumptions regarding the form of the autocorrelation functions restore identifiability.

preserves the form of the equations (2.90), and the assumed form of the distribution of the disturbances, which now permits correlation between $u_1^{\oplus}(t)$ and $u_2^{\oplus}(t-1)$. The transformation (2.91) affects the coefficients of the first equation according to

$$(2.92) \qquad \alpha^{\oplus}_{110} = \alpha_{110}, \qquad \alpha^{\oplus}_{1g1} = \alpha_{1g1} + \lambda\alpha_{2g0}, \qquad g = 1, 2.$$

The first equation (2.90) has thus ceased to be identifiable.

The transformation (2.91) permits one of the structural equations to be shifted in its timing before it is linearly combined with another equation. If such transformations are permissible, the study of identification problems is greatly complicated even if the a priori restrictions are linear. It is argued in [XVI] that these problems can perhaps be studied more fruitfully if at the same time the time variable is made continuous rather than discrete.

3. ESTIMATION OF THE PARAMETERS

3.1. *Properties of the Unrestricted Likelihood Function*

3.1.1. Maximum-likelihood estimation using all a priori restrictions. We now turn to the problem of estimating the parameters $\beta_{gi\tau}$, $\gamma_{gk\tau}$, σ_{gh} of the distribution function (2.8). It is assumed that the study of identification problems has shown whether or not the various structural equations on which this distribution is built are uniquely identified by the a priori restrictions. It is further assumed that this analysis has revealed the extent and nature of the indeterminacy in the parameters of those equations that are not uniquely identified.

We shall now make a more restrictive assumption on the nature of the distribution function $f(u)$ of the disturbances, at least for the purpose of constructing estimates of the parameters:

ASSUMPTION 3.1.1. *The disturbances u_g have a joint normal distribution function with nonsingular covariance matrix* $\Sigma \equiv [\,\sigma_{gh}\,] \equiv [\,\sigma^{gh}\,]^{-1}$,

$$(3.1) \qquad f(u) = (2\pi)^{-\frac{1}{2}G} \det{}^{-\frac{1}{2}}\Sigma \exp -\tfrac{1}{2}\sum_{g,h=1}^{G} u_g\,\sigma^{gh}u_h .$$

We shall use as estimates those functions of the observations which for this choice of $f(u)$ constitute *maximum-likelihood estimates* of the parameters. If (3.1) is inserted in (2.8), the probability density (2.8) in any particular sample point, i.e., for any particular set of observations $y(t)$, $z(t)$, $t = 1, \ldots, T$, is a function of the parameters B, Γ, Σ, known as the likelihood function. The maximum-likelihood estimates here considered are those values

$$B, \quad C, \quad S \tag{3.2}$$

of the parameters for which, subject to all the a priori restrictions, the likelihood function reaches its highest maximum. Following Mann and Wald, the properties of these estimates can then be studied both under the same normality assumption for the distribution of the disturbances, and under some less restrictive assumption.

3.1.2. Maximum-likelihood estimates under partial disregard of a priori information. T. W. Anderson and Rubin have indicated[1] other estimates based on a suggestion of M. A. Girshick. These estimates are obtained by mathematically simpler, and in most cases less laborious, computational methods. These estimates are maximum-likelihood estimates under disregard of a suitably chosen part of the a priori information available. The simplification of computational problems is obtained at a cost of increased sampling variances of the estimates (reduced efficiency of the method of estimation). Further comparison with this elegant method, called the "reduced-form method" will be made in section *3.2.1*. In the remainder of this article, the term "maximum-likelihood estimates" will be used for such estimates obtained with the aid of all a priori information available. Where a distinctive expression is needed, the term "information-preserving maximum-likelihood method" will be used for the method of estimation here applied.

3.1.3. Classification of the variables. For most of the present section, the relevant distinction is that between "jointly dependent" and "predetermined" variables, made in the introduction. The importance of this distinction is based on the fact that the coefficients of the jointly dependent variables enter the Jacobian (2.2) of the transformation (2.1), whereas those of the predetermined variables do not. In the equations defining the maximum-likelihood estimates, and in the formulae for their estimated as-

[1] [1949] and unpublished manuscript. See also [IX].

ymptotic sampling variances and covariances, the position of the jointly dependent variables has similarities with that of the one dependent variable in the single-equation least-squares method.

On the contrary, in these equations and formulae the predetermined variables occur without any distinction as to whether they are exogenous variables, or lagged values of endogenous variables. The latter distinction is relevant in the present context only in one instance: in the proof of consistency of the maximum-likelihood estimates. That the distinction is irrelevant elsewhere, is, of course, connected with the fact that the present study is confined to large-sample approximations.

3.1.4. Notation. We shall therefore continue to use the notation introduced in section 2.1.2. We restate the partitioning of coefficients and variables

$$\text{(3.3)} \qquad \mathrm{A} \equiv [\ \mathrm{B} \quad \Gamma\ ], \qquad x \equiv [\ y \quad z\ ],$$

and the equation system (1.1) in this notation,

$$\text{(3.4)} \qquad \mathrm{A}x'(t) = \mathrm{B}y'(t) + \Gamma z'(t) = u'(t),$$

where $x'(t)$ is the transpose of x, and

$$\text{(3.5)} \qquad u(t) = [\ u_1(t) \quad \cdots \quad u_G(t)\ ], \qquad G \equiv K_y.$$

We shall more fully use the symmetric notation $K_y \equiv G$, K_z, and $K_x \equiv K_y + K_z$ for the number of jointly dependent, of predetermined, and of all variables respectively.

If (3.1) is substituted in the likelihood function (2.8) and logarithms are taken, we obtain in the new notation[1]

$$\text{(3.6)} \qquad \begin{aligned} \frac{1}{T}\log F \equiv L \equiv L(\mathrm{A}, \Sigma) = &-\frac{1}{2}K_y \log 2\pi + \log\det \mathrm{B} \\ &- \frac{1}{2}\log\det\Sigma - \frac{1}{2}\operatorname{tr}\Sigma^{-1}\mathrm{A}M_{xx}\mathrm{A}'. \end{aligned}$$

Here M_{xx} is the observed symmetric "moment" matrix

[1] tr X, the trace of a square matrix X, denotes the sum of all diagonal elements of X.

$$(3.7)\qquad M_{xx} \equiv \frac{1}{T}\left[\sum_{t=1}^{T} x_k(t)\, x_l(t)\right] = \frac{1}{T}\sum_{t=1}^{T} x'(t)\, x(t), \qquad k,\ l = 1,\ \ldots,\ K_x,$$

of all variables $x_k(t)$, and partitions according to

$$(3.8)\qquad M_{xx} = \begin{bmatrix} M_{yy} & M_{yz} \\ M_{zy} & M_{zz} \end{bmatrix}.$$

3.1.5. Positive definiteness of M_{xx}. In what follows we shall assume that the sample obtained is one of those, occurring with probability one, for which M_{xx} is nonsingular and therefore positive definite. (A symmetric matrix M_{xx} is called positive definite if $a\, M_{xx}\, a' > 0$ for every nonvanishing vector (one-row matrix) a. A nonsingular moment matrix is positive definite because $a\, M_{xx}\, a' = \frac{1}{T}\sum_t a\, x'(t)\, x(t)\, a'$ is a sum of squares of the vector products $a\, x'(t)$, $t = 1, \ldots, T$, while $a\, M_{xx}\, a' = 0$ for some nonvanishing a would entail the singularity of M_{xx}.) It follows that M_{yy} and M_{zz} have ranks equal to their respective orders K_y and K_z.

3.1.6. The reduced form of the structural equations and its parameters. We shall first study the maximum properties of L without imposing any a priori conditions on the parameters A, Σ. From Theorem 2.1.3.5, we know that any maximum properties of the likelihood function should be preserved by a transformation of the type (2.18), which we rewrite in the new notation

$$(3.9)\qquad \mathrm{A}^{\oplus} = \Upsilon\, \mathrm{A}, \qquad \Sigma^{\oplus} = \Upsilon\, \Sigma\, \Upsilon'.$$

A unique representation of each set of mutually equivalent points (see Definition 2.1.3.6) in the unrestricted parameter space is obtained by writing $\Upsilon = \mathrm{B}^{-1}$ in (3.9), which makes

$$B^{\oplus} = I, \tag{3.10}$$

if I denotes the unit matrix of order K_y. The fact that each equivalent point set contains one and only one point satisfying (3.10) corresponds to the fact that each system (3.4) can be written in one and only one way in what has been called the *reduced form* [Mann and Wald, p. 201, equation (85)] :

$$y_i(t) = \sum_{k=K_y+1}^{K_z} \pi_{ik}\, z_{k-K_y}(t) + v_i(t), \qquad i = 1, \ldots, K_y,$$

or (3.11)

$$y'(t) - \Pi z'(t) = v'(t),$$

in which each equation contains only one of the independent variables, $y_k(t)$, $k = 1, \ldots, K_y$, with unity as coefficient. Here we have written $-\Pi$ for the value of $A^{\oplus}$ in (3.9) corresponding to (3.10), and we shall write Ω for the corresponding value of $\Sigma^{\oplus}$,

$$\omega_{kl} = \mathcal{E} v_k(t) \cdot v_l(t), \qquad k, l = 1, \ldots, K_y,$$

or (3.12)

$$\Omega = \mathcal{E} v_y'(t) \cdot v_y(t).$$

It follows that Ω, and therefore also Ω^{-1}, are symmetric and positive definite, since B is nonsingular. In this notation, (3.9) becomes

$$[\, I \quad -\Pi \,] = B^{-1} A, \qquad \Omega^{-1} = B' \Sigma^{-1} B. \tag{3.13}$$

Since

$$\log \det B - \tfrac{1}{2} \log \det \Sigma = \tfrac{1}{2} \log \det(B' \Sigma^{-1} B), \tag{3.14}$$

the function L in (3.6) can now be written as

$$L = L(-\Pi, \Omega) = -\tfrac{1}{2} G \log 2\pi + \tfrac{1}{2} \log \det \Omega^{-1}$$
$$- \tfrac{1}{2} \operatorname{tr} \{\Omega^{-1}(M_{yy} - M_{yz}\, \Pi' - \Pi M_{zy} + \Pi M_{zz}\, \Pi')\}. \tag{3.15}$$

Because of the uniqueness of the reduced form, there is no transformation in the (Π, Ω) space, other than the identical transformation, which preserves the form of (3.15).

3.1.7. Rules for the differentiation of functions of matrices. Before proceding to maximize (3.15) it will be useful to state a few rules regarding the differentiation of a matrix $X(\xi) = [x_{mn}(\xi)]$ with respect to a scalar parameter ξ. If X is square or rectangular, and if Y is a constant matrix with the same number of rows and columns respectively, we have, because of the linearity of the trace operation,

$$\frac{d}{d\xi} \operatorname{tr}(XY') = \operatorname{tr}\left(\frac{dX}{d\xi} Y'\right). \tag{3.16}$$

If X is square, let X^{mn} denote the cofactor of x_{mn}, such that the typical element of the inverse X^{-1} of X is $x^{mn} = X^{nm} / \det X$. Then we have

$$\frac{d}{d\xi} \log \det X = \frac{\frac{d}{d\xi} \det X}{\det X} = \frac{\sum_{m,n} X^{mn} \frac{dx_{mn}}{d\xi}}{\det X} = \operatorname{tr} X^{-1} \frac{dX}{d\xi} . \tag{3.17}$$

An expression for $\frac{dX^{-1}}{d\xi}$ is derived as follows:

$$XX^{-1} = I, \qquad \frac{dX}{d\xi} X^{-1} + X \frac{dX^{-1}}{d\xi} = 0, \qquad \frac{dX^{-1}}{d\xi} = - X^{-1} \frac{dX}{d\xi} X^{-1}. \tag{3.18}$$

Therefore, if Y is also square, of the same order, and constant,

$$\frac{d}{d\xi} \operatorname{tr}(X^{-1} Y') = - \operatorname{tr}\left(X^{-1} \frac{dX}{d\xi} X^{-1} Y'\right). \tag{3.19}$$

3.1.8. Maximizing the likelihood function with respect to Π. The study of the maximum properties of $L(-\Pi, \Omega)$ is facilitated by

maximizing L in two successive steps as follows: First we consider Ω as a given matrix, and determine the value P of Π at which the quadratic form $L(-\Pi, \Omega)$ in the elements of Π has a maximum (for variations of Π only).

Writing

$$\Pi = P + \varepsilon \underset{\rightarrow}{P} \tag{3.20}$$

in (3.15), we have

$$\begin{aligned} \frac{dL(-\Pi, \Omega)}{d\varepsilon} &= -\frac{1}{2}\mathrm{tr}\{\Omega^{-1}(M_{yz}\underset{\rightarrow}{P}' + \underset{\rightarrow}{P}M_{zy} - \Pi M_{zz}\underset{\rightarrow}{P}' - \underset{\rightarrow}{P}M_{zz}\Pi')\} \\ &= -\mathrm{tr}\{\Omega^{-1}(M_{yz} - \Pi M_{zz})\underset{\rightarrow}{P}'\} \end{aligned} \tag{3.21}$$

in connection with the symmetry[1] of M_{xx} and Ω^{-1}. Furthermore

$$\frac{d^2L}{d\varepsilon^2} = -\mathrm{tr}(\Omega^{-1}\underset{\rightarrow}{P}M_{zz}\underset{\rightarrow}{P}') \tag{3.22}$$

identically in ε, and in particular for $\varepsilon = 0$.

A necessary condition for a maximum of $L(-\Pi, \Omega)$ in $\Pi = P$ is that $\left(\frac{dL}{d\varepsilon}\right)_{\varepsilon=0}$ shall vanish for all possible values of $\underset{\rightarrow}{P}$. It is seen from (3.21) that this requires

$$M_{yz} - PM_{zz} = 0, \quad \text{or} \quad P = M_{yz}M_{zz}^{-1}, \tag{3.23}$$

using the nonsingularity of M_{zz}. There is therefore only one extremum of L, which is reached in a point $\Pi_{yz} = P_{yz}$ which proves to be independent of Ω. Moreover, this extremum is actually a maximum (and since we are dealing with a quadratic function, an absolute maximum) because (3.22) is a negative definite quadratic form in the elements of $\underset{\rightarrow}{P}$. For the matrices Ω^{-1} and M_{zz}, being positive definite, can be decomposed (see [9], p. 246) according to

$$\Omega^{-1} = \Psi'\Psi, \qquad M_{zz} = RR', \tag{3.24}$$

[1] Use has been made of the properties $\mathrm{tr}\, XY' = \mathrm{tr}\, Y'X = \mathrm{tr}\, YX'$ which follows directly from the definition of the trace.

where Ψ and R are real. Thereby (3.22) turns into the negative sum of squares[1]

$$\frac{d^2L}{d\varepsilon^2} = -\operatorname{tr}\{(\Psi \underset{\rightarrow}{P} R)(\Psi \underset{\rightarrow}{P} R)'\}. \tag{3.25}$$

The reader will have noticed that the elements

$$\hat{\pi}_{ik} = \sum_{l=K_y+1}^{K_x} m_{il}\, m^{lk}_{(zz)}, \qquad \text{with} \qquad [m^{lk}_{(zz)}] = M^{-1}_{zz}, \tag{3.26}$$

$$l,\ k = K_y+1,\ \ldots,\ K_x,$$

of the ith row of the matrix P represent the coefficients of the *elementary regression* of the dependent variable $y_i(t)$, $1 \le i \le K_y$, on the predetermined variables $z_k(t)$, $k = K_y+1, \ldots, K_x$, i.e., the coefficients estimated by the single-equation least-squares method.

3.1.9. Maximizing the likelihood function with respect to Ω. The second step is to insert (3.23) in the expression (3.15) for L, which on account of the symmetry of M_{zz} becomes

$$L(\Omega) \equiv -\frac{1}{2}K_y \log 2\pi + \frac{1}{2}\log\det\Omega^{-1} - \frac{1}{2}\operatorname{tr}(\Omega^{-1}\cdot {}^{z}M_{yy}), \tag{3.27}$$

where

$${}^{z}M_{yy} \equiv M_{yy} - M_{yz}\cdot M^{-1}_{zz}\cdot M_{zy} \tag{3.28}$$

is again positive definite, because it is the moment matrix of the "residuals" $v^{\oplus}_i(t) \equiv y_i(t) - \sum_{l=1}^{K_z} \hat{\pi}_{i,K_y+l}\, z_l(t)$, $i = 1, \ldots, K_y$, from the elementary regressions of each of the dependent variables separately, on all predetermined variables. We shall now maximize (3.27) with respect to the variations of Ω or of Ω^{-1}. If we write

$$\Omega^{-1} = V + \eta \underset{\rightarrow}{V}, \qquad V \equiv W^{-1}, \qquad \underset{\rightarrow}{V} = \underset{\rightarrow}{V}', \tag{3.29}$$

[1] Use is made of the properties $(XY')' = YX'$ and $\operatorname{tr}(XX') = \sum_{m,n}(x_{mn})^2$.

the assumed symmetry of $\underset{\rightarrow}{V}$ and V and the positive definiteness of V will ensure positive definiteness of Ω^{-1} in an η-neighborhood of $\eta = 0$ for any particular value of $\underset{\rightarrow}{V}$. For, if ω_1^{-1} and ω_1^{-1} represent the absolute minima of $v\,\Omega^{-1}\,v'$ and $v\,V\,v' \equiv v\,W^{-1}\,v'$ respectively, under the restriction $v\,v' = 1$, and $|\,\eta_1\,|$ the similarly restricted absolute maximum of $|v\,\underset{\rightarrow}{V}\,v'|$, we have $\omega_1^{-1} > 0$, $\omega_1^{-1} \geq \omega_1^{-1} - |\eta|\cdot|\,\eta_1\,| > 0$ for sufficiently small values of $|\eta|$. Using (3.16) and (3.17), we have from (3.27)

$$\frac{dL(\Omega)}{d\eta} = \frac{1}{2}\,\mathrm{tr}(\Omega\,\underset{\rightarrow}{V}) - \frac{1}{2}\,\mathrm{tr}(\underset{\rightarrow}{V}\cdot{}^{z}M_{yy}) = \frac{1}{2}\,\mathrm{tr}\{(\Omega - {}^{z}M_{yy})\,\underset{\rightarrow}{V}\}\,, \tag{3.30}$$

and, using (3.19),

$$\left(\frac{d^2L}{d\eta^2}\right)_{\eta=0} = -\frac{1}{2}\,\mathrm{tr}(W\,\underset{\rightarrow}{V}\,W\,\underset{\rightarrow}{V}). \tag{3.31}$$

From (3.30) we see that $L(\Omega)$ has one stationary value, which is reached if Ω equals

$$W \equiv {}^{z}M_{yy}\,. \tag{3.32}$$

This stationary value is a maximum because the quadratic form (3.31) can be shown to be negative definite in the elements of $\underset{\rightarrow}{V}$, by an argument similar to that used in the case of (3.22), and using the symmetry of $\underset{\rightarrow}{V}$.

There are various ways of proving that (3.32) indicates the absolute maximum of $L(\Omega)$ in the space of symmetric and positive definite matrices Ω^{-1}. Perhaps the most elementary proof is as follows: If $\Omega^{-1}(1)$ is a matrix in that space different from W^{-1}, the matrix

$$\Omega^{-1}(\theta) \equiv \theta\,\Omega^{-1}(1) + (1-\theta)W^{-1}, \qquad 0 \leq \theta \leq 1, \tag{3.33}$$

is easily shown to belong to the same space. The function

$$\bar{L}(\theta) \equiv L\{\,\Omega^{-1}(\theta)\} \tag{3.34}$$

possesses continuous first and second derivatives with respect to θ for $0 \leq \theta \leq 1$. These derivatives satisfy the two conditions

$$(3.35) \qquad \left(\frac{d\bar{L}(\theta)}{d\theta}\right)_{\theta=0} = 0, \qquad \frac{d^2\bar{L}(\theta)}{d\theta^2} < 0 \quad \text{for} \quad 0 \leq \theta \leq 1.$$

The first condition is satisfied because a stationary value $L(\Omega)$ is reached for $\Omega^1 = \Omega^1(0)$. The second condition is satisfied because the negative definiteness of (3.31) is not dependent on W satisfying the maximum conditions (3.32). It follows from (3.35) by use of Taylor's theorem, that

$$(3.36) \qquad L\{\Omega(1)\} = \bar{L}(1) = \bar{L}(0) + \frac{1}{2}\left(\frac{d^2\bar{L}(\theta)}{d\theta^2}\right)_{\theta=\theta'} < \bar{L}(0) = L(W^{-1}),$$

where θ' represents some number between 0 and 1.

3.1.10. The absolute maximum of the likelihood function. By inverting the transformation (3.13) we can summarize the maximum properties of the likelihood function in the following

THEOREM 3.1.10. *In the absence of any a priori restrictions the logarithmic likelihood function* (3.6) *has one and only one maximum value*

$$(3.37) \qquad L_{\max} = -\frac{1}{2}K_y(1 + \log 2\pi) - \frac{1}{2}\log\det(M_{yy} - M_{yz}M_{zz}^{-1}M_{zy}),$$

which is an absolute maximum. This maximum is reached in each point

$$(3.38) \qquad \begin{aligned} &B = \text{any nonsingular square matrix of order } K_y, \\ &\Gamma = -BP, \\ &\Sigma = BWB, \end{aligned}$$

of the set of points equivalent to the point

$$(3.39) \qquad B = I, \quad \Gamma = -P = -M_{yz}M_{zz}^{-1}, \quad \Sigma = W = M_{yy} - M_{yz}M_{zz}^{-1}M_{zy}.$$

This theorem establishes the uniqueness of the maximum of the likelihood function in the unrestricted parameter space, in the sense that there is one and only one set of mutually equivalent points on which the maximum is reached.

We add an expression for the likelihood function that is derived from (3.15) with the help of (3.39)

(3.40)
$$L = -\frac{1}{2}K \log 2\pi + \frac{1}{2} \log \det \Omega^{-1} - \frac{1}{2} \operatorname{tr}\left(\Omega^{-1}\{W + (P-\Pi) M_{zz} (P-\Pi)'\}\right),$$

and that brings out clearly the significance of the statistics P and W established by Theorem 3.1.10.

3.2. *Properties of the Restricted Likelihood Function*

3.2.1. The case in which the restricted likelihood function can attain its absolute maximum. In the case where a priori restrictions are introduced, a somewhat weaker theorem can be formulated as long as the a priori restrictions do not prevent the likelihood function from attaining its absolute maximum (3.37).

THEOREM 3.2.1. *Under a priori restrictions of any kind that permit the likelihood function to attain its absolute maximum in some point* (A, Σ), *this maximum is attained only in all points* $(A^{\oplus}, \Sigma^{\oplus})$ *of the restricted parameter space that are equivalent to the point* (A, Σ),

This theorem follows immediately from Theorem 3.1.10 and Definition 2.1.5.1. It should be noted that Theorem 3.2.1 does not preclude the existence of one or more relative maxima where the likelihood function attains a value lower than (3.37).

The question of whether or not the a priori restrictions permit the likelihood function to attain its absolute maximum is important for two reasons. In the first place this question is connected with the relations between the reduced-form method[1] based essentially on single-equation least-squares procedures, and the maximum-likelihood method preserving all a priori information, as applied in this article. According to Theorem 3.1.10, as long as the absolute maximum of the likelihood function can be reached, the information-preserving maximum-likelihood method of estimation is mathematically equivalent to the single-equation least-squares

[1] See section *3.1.2* and also [IX].

method applied to each equation of the reduced form (3.11). For the respective rows of P in (3.39) are identical with the estimates obtained for the coefficients of the corresponding equations (3.11) by the latter method. After P and W have been determined from (3.39), it is then possible to determine the transformation (3.38) so as to satisfy the a priori restrictions.

The second reason is connected with the computation of maximum-likelihood estimates, and is a consequence of the first reason. In case L attains the value $L_{\max}$, the procedure just described always leads to the absolute maximum of the likelihood function. In case L cannot attain $L_{\max}$, the maximum-likelihood equations are essentially nonlinear, and the only practicable methods of computation available are iterative methods. So far we do not know with certainty under what conditions each of these methods converges to the absolute maximum. One may possibly be led to a relative maximum, depending on the initial values chosen at the start of the iterative procedure, and the particular method of iteration used. As far as our present results reach, therefore, the case where L cannot attain the value $L_{\max}$ is subject to an uncertainty which is absent when $L_{\max}$ can be attained.

3.2.2. Attainability of the absolute maximum under linear and bilinear restrictions. For these reasons, to which another will be added in section *4.3.3.4*, it is important to know under which conditions $L_{\max}$ can be attained, that is, under which conditions (3.38) is compatible with the a priori restrictions. If the latter consist of the linear restrictions (2.24) combined with the bilinear restrictions (2.73), this question must be answered from an equation system obtained by inserting (3.38) in these restrictions:

$$\beta(g)[-I \quad P]\,\Phi_g^i = 0, \qquad g = 1, \ldots, K_y, \tag{3.41lh}$$

$$\beta(g)[-I \quad P]\,\iota'(i_g) = 1, \tag{3.41ln}$$

$$\begin{bmatrix} \beta(g_r)[-I \quad P]\,\iota'(k_r) & \beta(g_r)[-I \quad P]\,\iota'(l_r) \\ \beta(h_r)[-I \quad P]\,\iota'(k_r) & \beta(h_r)[-I \quad P]\,\iota'(l_r) \end{bmatrix} = 0, \qquad r = 1, \ldots, R_\alpha^{(2)}, \tag{3.41bα}$$

$$(3.41b\sigma) \qquad \beta(g_r)\, W\, \beta'(h_r) = 0, \qquad r = R_\alpha^{(2)} + 1, \ \ldots, \ R_\alpha^{(2)} + R_\sigma ,$$

where $\iota(k)$ is again the kth row of the unit matrix of order K_x. We shall, for convenience, refer to these equations as follows:

$$(3.41) \begin{cases} (3.41l) \begin{cases} (3.41lh) \\ (3.41ln) \end{cases} \\ (3.41b) \begin{cases} (3.41b\alpha) \\ (3.41b\sigma) \end{cases} \end{cases}$$

The matrix $[-I \quad P]$ in (3.41) is put together in analogy to

$$(3.42) \qquad [-I \quad \Pi] = -B^{-1}[B \quad \Gamma] = -B^{-1}A ,$$

as defined in (3.13).

The equations (3.41) are similar in form to the equations (2.74), and the present problem is therefore closely related to the identification problem. Nevertheless, there are two important differences in the two problems, one in the assumptions, and one in the question to be answered. In the identification problem, it is known by assumption that the equations (2.74) have at least one real solution $\Upsilon = I$, and the question to be answered is under what conditions there is only one, or a finite number of real solutions. In the present problem it is not known whether there is at all a real solution B to the equations (3.41), and the question is under what conditions there is at least one solution. To obtain an answer to the present question, the counting of equations and variables is even less conclusive than in the identification problem. For even in a case in which, on the basis of counting, the number of real or complex solutions B is believed to be finite almost everywhere in the space of P and W, we cannot without further analysis say that the part of the sample space in which all solutions are complex is of measure zero.

We have so far not succeeded in finding general conditions for the existence of at least one real solution. However, the iterative computation procedures for solution of the maximum-likelihood

equations to be described in section 4 lead to such solutions if they exist, provided the computation is started with suitable initial values – although again we do not know precisely which initial values are suitable.

3.2.3. Attainability of the absolute maximum under linear restrictions only. It is not difficult to state exact conditions for the attainability of the absolute maximum of L in the case where only linear a priori restrictions of the type (2.24) are introduced. This leads to the conditions (3.41*lh*) which we wish to be satisfied by at least one real solution B.

THEOREM 3.2.3.1. *A necessary condition for the attainability of the absolute maximum of the likelihood function under the homogeneous linear a priori restrictions* (2.24) *is that a) none of the matrices* $P\Phi_g'$, $g = 1, \ldots, K_y$, *has a rank exceeding* $K_y - 1$. *A necessary and sufficient condition is that, in addition to a), b) the consequently nonempty set of solutions* B *of* (3.41*lh*) *contains at least one nonsingular solution* (det $B \neq 0$).

THEOREM 3.2.3.2. *A necessary condition for the attainability, almost everywhere in the sample space, of the absolute maximum of the likelihood function under the homogeneous linear a priori restrictions* (2.24) *is that none of the matrices* Φ_g, $g = 1, \ldots, K_y$, *has a rank exceeding* $K_y - 1$.

Theorem 3.2.3.1 follows directly from the conditions for the existence of a solution of a homogeneous system of linear equations. Theorem 3.2.3.2 follows because the condition that the rank of $P\Phi_g'$ shall fall below K_y if the rank of Φ_g' is at least K_y entails a restriction on P satisfied only on a point set of measure zero in the sample space.

3.2.4. Connections between attainability of the absolute maximum of the likelihood function and identifiability of structural equations. It is of interest to compare Theorem 3.2.3.2 with Theorem 2.2.2 and its corollary. The latter states that identifiability of the gth structural equation requires the rank of Φ_g to be at least $K_y - 1$. Theorem 3.2.3.2 states that attainability of the absolute maximum requires that rank to be at most $K_y - 1$ (for all values of G). Thus, if independent linear restrictions on the coefficients of the gth equation are added one by one (beginning in a situation where L_{max} is attainable), the point at which

in general complete identification of the gth structural equation is attained almost everywhere in the parameter space, is at the same time the point beyond which no further restrictions referring to the equation can be added without preventing $L_{\max}$ from being attainable almost everywhere in the sample space.

A similar situation is found under quite general a priori restrictions, which we shall denote

$$\varphi_r(A, \Sigma) = 0, \qquad r = 1, \ldots, R. \tag{3.43}$$

We shall assume these restrictions to be independent[1], compatible, and to imply normalization of all structural equations. We shall further assume that the functions φ_r possess continuous derivatives with respect to the parameters A, Σ.

DEFINITION 3.2.4.1. *By the restricted reduced parameter space we understand the space of the parameters*

$$\Pi = -B^{-1}\Gamma, \qquad \Omega = B^{-1}\Sigma B'^{-1}, \qquad \det \Omega \neq 0, \tag{3.44}$$

subject to such restrictions,

$$\psi_r(\Pi, \Omega) = 0, \tag{3.45}$$

if any, as are a consequence of (3.43).

The usefulness of this definition is based on the fact, recognized above, that the parameters Π, Ω of the reduced form uniquely specify the distribution of the observed variables. In other words, there is a one-to-one correspondence between the points of the reduced parameter space and the sets of mutually equivalent points in the restricted parameter space (see Definitions 2.1.4 and 2.1.5.1).

DEFINITION 3.2.4.2. *The space of the statistics M_{xx} will be called the moment space. The space of the statistics P and W defined by* (3.39) *and* (3.42) *will be called the reduced moment space. We exclude from these spaces any singular values of M_{xx} or W, which will be referred to as arising from "singular" samples.*

[1]The concept of independence for the purposes of this discussion is sharply defined by Definition 3.2.5 below.

On the basis of the foregoing definitions, the condition (3.38) for attainment of the absolute maximum of the likelihood function can be rewritten as

$$\Pi = P, \qquad \Omega = W, \tag{3.46}$$

for this is obtained if the expressions (3.38) for A and Σ are substituted in (3.44). We thus have:

THEOREM 3.2.4.1. *A necessary and sufficient condition for the attainability of the absolute maximum of the likelihood function for a given nonsingular sample of observations, is that, to the point P, W in the reduced moment space, there corresponds, by* (3.46), *a point Π, Ω that belongs to the restricted reduced parameter space, i.e., satisfies the restrictions* (3.45).

This theorem shows that the attainability of the absolute maximum of L under given restrictions depends only on the statistics P, W, not on the statistics M_{zz} on which, as shown in (3.40), the likelihood function also depends. Moreover, the attainability of $L_{\max}$ does not even depend on P and W if the set of restrictions (3.45) is empty.

We can now state an important theorem, which indicates the connection between identifiability of structural equations and attainability of the absolute maximum of the likelihood function, alluded to at the beginning of the present section *3.2.4*.

THEOREM 3.2.4.2. *Let N_m be a region of positive measure in the reduced moment space, in each point of which the likelihood function can attain its absolute maximum under the restrictions* (3.43). *Let N_η be the corresponding region, according to* (3.46), *in the reduced parameter space. Assume that in every point of N_η the structural equations belonging to a nonempty set S are completely identifiable. Then, the addition to* (3.43) *of one further restriction, which is independent in N_η (in the sense of Definition 3.2.5 below) of the original restrictions* (3.43), *and which refers to equations of S only, will prevent the likelihood function from attaining its absolute maximum almost everywhere in N_η.*

*3.2.5. *Proof of Theorem 3.2.4.2.* Denote by $\theta \equiv [\ \theta_S \quad \theta_{-S}\ ]$ a vector (one-row matrix) containing, under the notation θ_p, $p = 1, \ldots, P$, all elements of A and Σ. Let θ_S contain, under the notation

θ_p, $p = 1, \ldots, P_S$, all elements of those rows of A corresponding to structural equations belonging to the set S and those elements of Σ referring to two equations of S. Let θ_{-S} with elements θ_p, $p = P_S + 1, \ldots, P$, contain all remaining parameters. Let the vector $\eta = \eta(\theta)$ contain all $P^{\oplus}$ elements of Π and Ω, and h those of P and W.

The assumptions with regard to N_m and N_η respectively, stated in the theorem, imply, owing to Theorem 3.2.4.1, that in every point η of N_η, the equation system (3.43), (3.44) admits at least one solution θ, and further that among those solutions, there is only a finite number of different values of θ_S. It follows from theorems regarding implicit functions, that (3.43) and (3.44) define θ_S as an implicit (multivalued) function of η, of which the derivatives are found from

$$\begin{aligned} \Phi_S \cdot \delta\theta_S' + \Phi_{-S} \cdot \delta\theta_{-S}' &= 0, \\ H_S \cdot \delta\theta_S' + H_{-S} \cdot \delta\theta_{-S}' &= \delta\eta', \end{aligned} \tag{3.47}$$

by elimination of $\delta\theta_{-S}$. Here

$$\Phi_S \equiv \begin{bmatrix} \frac{\partial\varphi_1}{\partial\theta_1} & \cdots & \frac{\partial\varphi_1}{\partial\theta_{P_S}} \\ \cdot & \cdots & \cdot \\ \frac{\partial\varphi_R}{\partial\theta_1} & \cdots & \frac{\partial\varphi_R}{\partial\theta_{P_S}} \end{bmatrix}, \qquad \Phi_{-S} \equiv \begin{bmatrix} \frac{\partial\varphi_1}{\partial\theta_{P_S+1}} & \cdots & \frac{\partial\varphi_1}{\partial\theta_P} \\ \cdot & \cdots & \cdot \\ \frac{\partial\varphi_R}{\partial\theta_{P_S+1}} & \cdots & \frac{\partial\varphi_R}{\partial\theta_P} \end{bmatrix}, \tag{3.48}$$

and H_S, H_{-S} are defined similarly with respect to the elements of $\eta = \eta(\theta)$.

The fact that (3.47) possesses at least one solution $[\ \delta\theta_S \quad \delta\theta_{-S}\ ]$ for any $\delta\eta$ requires that

$$\rho\begin{pmatrix} \Phi_S & \Phi_{-S} \\ H_S & H_{-S} \end{pmatrix} = \rho(\ \Phi_S \quad \Phi_{-S}\ ) + \rho(\ H_S \quad H_{-S}\ ). \tag{3.49}$$

For the only alternative to (3.49) is that the right-hand member exceeds the left-hand member, in which case there would, according to Lemma 3.2.2, exist two nonvanishing vectors φ, η such that

$$(3.50) \qquad [\ \bar{\varphi} \quad \bar{\eta}\ ] \begin{bmatrix} \Phi_S & \Phi_{-S} \\ H_S & H_{-S} \end{bmatrix} = 0 .$$

But then we could conclude from the existence of a solution of (3.47) that

$$(3.51) \qquad \bar{\varphi}\, 0' + \bar{\eta}\, \delta\eta' = \bar{\eta}\, \delta\eta' = 0,$$

and values of $\delta\eta$ not satisfying (3.51) would not permit a solution of (3.47), contrary to the assumption made.

On the other hand, it is known that, after elimination of $\delta\theta_{-S}$, (3.47) has only a finite number of solutions $\delta\theta_S$, which in view of the linearity of the system (3.47) can only be one. The uniqueness of this solution is essentially a property of the homogeneous system of equations obtained from (3.47) by writing $\delta\eta = 0$. From the uniqueness of $\delta\theta_S$ it follows that

$$(3.52) \qquad \rho\begin{pmatrix} \Phi_S & \Phi_{-S} \\ H_S & H_{-S} \end{pmatrix} = \rho\begin{pmatrix} \Phi_S \\ H_S \end{pmatrix} + \rho\begin{pmatrix} \Phi_{-S} \\ H_{-S} \end{pmatrix} .$$

For otherwise, according to Lemma 2.3.2, two nonvanishing vectors $\bar{\theta}_S$, $\bar{\theta}_{-S}$ would exist such that

$$(3.53) \qquad \begin{bmatrix} \Phi_S & \Phi_{-S} \\ H_S & H_{-S} \end{bmatrix} \begin{bmatrix} \bar{\theta}'_S \\ \bar{\theta}'_{-S} \end{bmatrix} = 0 ,$$

and a scalar multiple of $\bar{\theta} \equiv [\ \bar{\theta}_S \quad \bar{\theta}_{-S}\ ]$ could be added to the solution $\delta\theta = [\ \delta\theta_S \quad \delta\theta_{-S}\ ]$ of (3.47) to produce other solutions

for the same value of $\delta\eta$, which differ in regard to $\delta\theta_S$. In addition, we have

$$\rho\begin{pmatrix}\Phi_S\\ H_S\end{pmatrix} = P_S, \tag{3.54}$$

which is the highest rank a matrix of P_S columns can attain. For, if the left-hand member in (3.54) were less than P_S, then a nonvanishing vector $\bar{\theta} = [\ \bar{\theta}_S \quad 0_{-S}\ ]$ could be found, a scalar multiple of which could be added to the solution $\delta\theta = [\ \delta\theta_S \quad \delta\theta_{-S}\ ]$ of (3.47) to produce other solutions for the same value of $\delta\eta$, which differ in regard to $\delta\theta_S$.

Now suppose that an additional a priori restriction

$$\varphi(\theta_S) = 0 \tag{3.55}$$

is imposed such that N_η contains at least one point η_0 in which (3.43), (3.44), and (3.55) have a solution θ. (If no such point exists the theorem is already true.) We shall investigate what are the conditions to be satisfied by the derivatives of φ and φ_r in that point in order that the absolute maximum of the likelihood function is attainable everywhere in a neighborhood of the corresponding point $h_0 = \eta_0$ of the reduced moment space. For that to be so, it is necessary that if the row

$$[\ \varphi_S \quad 0_{-S}\ ] \equiv \left[\ \frac{\partial\varphi}{\partial\theta_1} \quad \cdots \quad \frac{\partial\varphi}{\partial\theta_p} \quad 0_{P_S+1} \quad \cdots \quad 0_{P_S}\ \right] \tag{3.56}$$

is added to the matrix $[\ \Phi_S \quad \Phi_{-S}\ ]$, the system (3.47) so enlarged or

$$\begin{aligned} \varphi_S \cdot \delta\theta'_S &= 0, \\ \Phi_S \cdot \delta\theta'_S + \Phi_{-S} \cdot \delta\theta'_{-S} &= 0, \\ H_S \cdot \delta\theta'_S + H_{-S} \cdot \delta\theta'_{-S} &= \delta\eta', \end{aligned} \tag{3.57}$$

satisfies the properties established for the original system (3.47). This would have the following consequences: From (3.54), applied both to the original and to the enlarged system, it follows that

$$\rho\begin{pmatrix}\Phi_S\\ \\ H_S\end{pmatrix} = P = \rho\begin{pmatrix}\varphi_S\\ \Phi_S\\ H_S\end{pmatrix}. \tag{3.58}$$

From (3.52) applied to both members of (3.58) follows

$$\rho\begin{pmatrix}\Phi_S & \Phi_{-S}\\ \\ H_S & H_{-S}\end{pmatrix} = \rho\begin{pmatrix}\varphi_S & 0\\ \Phi_S & \Phi_{-S}\\ H_S & H_{-S}\end{pmatrix}, \tag{3.59}$$

since the addition of a row of zeros to the last matrix in (3.52) does not change its rank. From (3.49) applied to both members of (3.59)

$$\rho\begin{pmatrix}\Phi_S & \Phi_{-S}\end{pmatrix} = \rho\begin{pmatrix}\varphi_S & 0\\ \\ \Phi_S & \Phi_{-S}\end{pmatrix}. \tag{3.60}$$

This, then, is a necessary condition for the existence of a solution of the enlarged system (3.57) for every value of $\delta\eta$. It will now be shown that if (3.60) is not satisfied, the only values of $\delta\eta$ permitting such a solution are those subject to a linear restriction of the type (3.51). If (3.60) is not true, we must have

$$\rho\begin{pmatrix}\Phi_S & \Phi_{-S}\end{pmatrix} = \rho\begin{pmatrix}\varphi_S & 0\\ \\ \Phi_S & \Phi_{-S}\end{pmatrix} - 1\,. \tag{3.61}$$

Suppose then that $\delta\eta$ is such that a solution of the enlarged

system (3.57) exists. The validity of (3.58) and (3.59) is not affected by the fact that $\delta\eta$ now represents one particular value, instead of all possible values. For the proof of the equalities for the enlarged system, equivalent to (3.52) and (3.54) respectively, depends only on the uniqueness of the solution with regard to $\delta\theta_S$, and this was already recognized as being a property of the homogeneous system obtained from (3.57) by taking $\delta\eta = 0$.

From (3.59), (3.61), and (3.49), we conclude

$$(3.62)\qquad \rho\begin{pmatrix} \varphi_S & 0 \\ \Phi_S & \Phi_{-S} \\ H_S & H_{-S} \end{pmatrix} = \rho\begin{pmatrix} \varphi_S & 0 \\ \Phi_S & \Phi_{-S} \end{pmatrix} + \rho\begin{pmatrix} H_S & H_{-S} \end{pmatrix} - 1,$$

from which, as before, we can derive the existence of a linear restriction on $\delta\eta$ of the type (3.51).

It is well known that, if (3.60) is satisfied everywhere in N_η, then $\varphi(\theta_S) \equiv \varphi([\ \theta_S \quad 0_{-S}\ ])$ is a function of the remaining functions $\varphi_r(\theta)$, and (3.55) is either dependent on or incompatible with (3.43). The present theorem could probably be proved on the assumption that (3.60) holds in N_η only on a set of measure zero. We shall make a somewhat different assumption, which is better adapted to this particular proof, and is sufficient for our purposes:

DEFINITION 3.2.5. *The restriction* (3.55) *is called independent in* N_η *of the restrictions* (3.43), *if* (3.60) *is not satisfied in any point in* N_η *in which* (3.43), (3.44), *and* (3.45) *permit a solution* θ (*or* A, Σ).

If this is the case, the set of points η_0 in N_η in which (3.43), (3.44), and (3.55) permit a solution θ can only be of measure zero, because from any one such point, any neighboring points can now only be reached by variations $\delta\eta$ subject to a linear restriction.

A special case in which the additional restriction satisfies Definition 3.2.5 is, of course, that in which (3.60) is not satisfied in any point in N_η.

*3.2.6. *Tabular summary of possible cases.* We shall now apply Theorem 3.2.4.2 to the case of linear and bilinear a priori restrictions (2.24) and (2.27). It may be useful to set out the various

TABLE 3.2.6

CONNECTION BETWEEN IDENTIFIABILITY OF STRUCTURAL EQUATIONS AND ATTAINABILITY OF THE ABSOLUTE MAXIMUM OF THE LIKELIHOOD FUNCTION

Note: This classification excludes point sets of measure zero in the parameter space (col. 3) and in the sample space (col. 4) and is subject to other exceptions discussed in sections *2.4.8 – 10*. It is assumed that the a priori restrictions are compatible and that they are mutually independent in the sense of Definition 3.2.5.

Possible Cases		Statements relating to these cases	
(1)	(2)	(3)	(4)
The completed subset[1] S_0 of the structural equations	The a priori restrictions in the associated subset[2] R_{S_0} of (3.41) are with respect to S_0	The following structural equations are completely identifiable:	Can the likelihood function attain its absolute maximum?
(A) is empty.		none.	Only if among the solutions B of (3.41) there is a real solution.[4]
(B) is not empty but does not contain all structural equations.	(1) just adequate in number and variety.[3]	only those of S_0.	Only if among the solutions B of (3.41) there is a real solution.[4]
	(2) more than just adequate in number and variety.[3]	only those of S_0.	No.
(C) contains all structural equations.	(1) just adequate in number and variety.[3]	all structural equations.	Only if among the solutions B of (3.41) there is a real solution.[4]
	(2) more than just adequate in number and variety.[3]	all structural equations.	No.

[1]See Definition 2.4.6.2. [2]See Definition 2.4.6.1. [3]See Definition 3.2.6.

[4]If the a priori restrictions consist of linear restrictions only, this clause can be replaced by an unqualified "yes." It may be stated without

cases as to identifiability and attainability of the absolute maximum of L in a tabular form based on the counting of restrictions, even though the validity of this criterion is subject to exceptions already noted. In connection with Thm. 3.2.4.2, it is desirable to supplement Definition 2.1.5.5 by

DEFINITION 3.2.6. *A subset R of the a priori restrictions* (2.28) *will be said to be just adequate in number and variety with respect to (the identification of) a subset S of the structural equations if it is adequate in the sense of Definition* 2.1.5.5 *but loses that property if any of the restrictions in R are omitted.*

3.2.7. A factorization of the likelihood function. A further remark may be made about the case where the a priori restrictions imply a simultaneous partitioning (2.82) or

$$B = \begin{bmatrix} B_{\mathrm{I\,I}} & B_{\mathrm{I\,II}} \\ 0 & B_{\mathrm{II\,II}} \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{\mathrm{I\,I}} & 0 \\ 0 & \Sigma_{\mathrm{II\,II}} \end{bmatrix}, \tag{3.63}$$

of the matrices B and Σ. It is easily seen that this entails a factorization of the likelihood function, expressed by the following splitting-up of its logarithm (3.6) into two terms

$$L(A, \Sigma) = L_{\mathrm{I}}(A_{\mathrm{I}}, \Sigma_{\mathrm{II}}) + L_{\mathrm{II}}(A_{\mathrm{II}}, \Sigma_{\mathrm{II\,II}}), \tag{3.64}$$

where

$$\begin{aligned} L_{\mathrm{I}} &= -\frac{1}{2}K_{\mathrm{I}} \log 2\pi + \log\det B_{\mathrm{I\,I}} - \frac{1}{2}\log\det \Sigma_{\mathrm{I\,I}} \\ &\quad - \frac{1}{2}\operatorname{tr}(\Sigma_{\mathrm{I\,I}}^{-1} \cdot A_{\mathrm{I}} \cdot M_{xx} \cdot A_{\mathrm{I}}'), \end{aligned} \tag{3.65}$$

$$\begin{aligned} L_{\mathrm{II}} &= -\frac{1}{2}K_{\mathrm{II}} \log 2\pi + \log\det B_{\mathrm{II\,II}} - \frac{1}{2}\log\det \Sigma_{\mathrm{II\,II}} \\ &\quad - \frac{1}{2}\operatorname{tr}(\Sigma_{\mathrm{II\,II}}^{-1} \cdot A_{\mathrm{II}} \cdot M_{xx} \cdot A_{\mathrm{II}}'). \end{aligned}$$

In [XVII] this factorization property of the likelihood function

---

proof that the "yes" applies even if the number of *linear* restrictions referring to *each* structural equation is $K_y - 1$.

is used to justify the concept of exogenous variables. Here it is sufficient to remark that, if no further a priori restrictions connect the structural equations with coefficients A_I with those having coefficients A_{II}, the estimation problems of these two subsystems of the system of structural equations have been effectively separated. For the two terms in (3.64) then depend on entirely independent sets of parameters, and the function (3.64) can only reach its maximum if each of the two terms reaches its own maximum.

It should be noted that the expressions (3.65) for L_I and L_{II} are of precisely the same form as the original logarithmic likelihood function (3.6). The variables indicated by the subscript II occur as dependent variables in L_{II} while the variables corresponding to the subscript I do not occur in L_{II} (the moments $M_{y_I x}$ are multiplied into vanishing coefficients). The variables corresponding to the subscript II occur as predetermined variables in L_I, in which the variables corresponding to the subscript I represent the dependent variables.

3.3. *Large-Sample Properties of the Maximum-Likelihood Estimates*

3.3.1. Assumptions. In this section *3.3* we shall discuss the large-sample properties of the maximum-likelihood estimates of the parameters of the system (1.1) of structural equations. Following Mann and Wald, [1943, p. 192], we shall assume that the equation system is stable. Reverting to the notation of section *1*, we express this by the following two assumptions:

ASSUMPTION 3.3.1.1. *All roots* ρ *of the equation*

$$\det\left[\sum_{\tau=0}^{\tau^{\square}} B(\tau)\,\rho^{\tau^{\square}-\tau}\right] = \det\left[\sum_{\tau=0}^{\tau^{\square}} \beta_{g i \tau}\,\rho^{\tau^{\square}-\tau}\right] = 0, \qquad g,\ i = 1, \ldots, K_y, \tag{3.66}$$

satisfy

$$|\rho| < 1. \tag{3.67}$$

ASSUMPTION 3.3.1.2. *If*

$$(3.68)\qquad m_{z_k z_l}(\tau, \tau', T) \equiv \frac{1}{T}\sum_{t=1}^{T} z_k(t-\tau)\, z_l(t-\tau'), \qquad k, l = 1, \ldots, K,$$

is a moment of two exogenous variables $z_k(t)$ *and* $z_l(t)$, *there exists a finite limit*

$$(3.69)\qquad \lim_{T\to\infty} m_{z_k z_l}(\tau, \tau', T) = {}^{\infty}\mu_{z_k z_l}(\tau, \tau')$$

for every k, l, τ, *and* τ'.

Regarding the distribution of the disturbances we shall make two alternative assumptions:

ASSUMPTION 3.3.1.3. *The distribution function* $f(u_1, \ldots, u_{K_y})$ *of the disturbances possesses finite* $(4+\varepsilon)$*th-order moments for some* $\varepsilon > 0$. *Its first-order moments vanish and its second-order moments form a nonsingular matrix* Σ.

Alternatively, we shall specify a particular distribution admitted under Assumption 3.3.1.3.

ASSUMPTION 3.3.1.4. *The disturbances* $u_1, \ldots, u_{K_y}$ *have a joint normal distribution* (3.1) *with mean zero and nonsingular second-order moment matrix* Σ.

Values $y(t)$, $t \le 0$, of endogenous variables, with a timing preceding the period $1 \le t \le T$ during which the dependent variables are observed, are treated (together with the values of the exogenous variables) as given constants which remain the same in repeated samples.

3.3.2. Quasi-maximum-likelihood estimates. Under Assumption 3.3.1.4, the distribution function of the observations $x_1(1), \ldots, x_{K_y}(T)$ is

$$(3.70)\qquad F(M_{xx}, A, \Sigma) = (2\pi)^{-\frac{1}{2}\cdot K_y T}\cdot \det{}^{T}B \cdot \det{}^{-\frac{1}{2}\cdot T}\Sigma \cdot \exp\{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1} A M_{xx} A')\}.$$

As a function of the parameters A, Σ, we have called (3.70) the likelihood function, and defined maximum-likelihood estimates as the values of the parameters that, subject to the a priori restrictions, maximize this function. Under the wider Assumption 3.3.1.3, (3.70) has no necessary connection with the distribution of the observations. Nevertheless, we can use the function (3.70) to define estimates of the parameters by the same maximizing procedure. In these circumstances, we shall call (3.70) the quasi-likelihood function, and call the maximizing values of its parameters quasi-maximum-likelihood estimates. We shall also discuss some large-sample properties of these estimates.

3.3.3. Results of Mann and Wald. A very thorough analysis of large-sample properties of the quasi-maximum-likelihood estimates has been given by Mann and Wald [1943]. The system considered by these authors satisfies Assumption 3.3.1.1 and a slightly more restrictive version of Assumption 3.3.1.3. Their system does not contain exogenous variables $z_k(t)$ (except a constant term in each equation). Finally, they assume that each equation is completely identified. Our main concern in the present section *3.3* is to indicate that Mann and Wald's results can be extended to the case where exogenous variables satisfying Assumption 3.3.1.2 are present, and to the case where some but not all of the structural equations are identifiable. We shall first discuss the large-sample properties of the moment matrix M_{xx}, and thereafter those of the quasi-maximum-likelihood estimates.

3.3.4. Asymptotic distribution of the moments. Extended to include systems with exogenous variables, Mann and Wald's results regarding the moments can be stated as follows:

THEOREM 3.3.4. *Under Assumptions* 3.3.1.1, 3.3.1.2, *and* 3.3.1.3, *the expected value*

$$\mathcal{E} M_{xx} \equiv \mathrm{M}_{xx} \tag{3.71}$$

of the moment matrix M_{xx} *possesses the properties* a) *that*

$$\lim_{T \to \infty} \mathrm{M}_{xx} \equiv \mathrm{M}_{\infty} \tag{3.72}$$

exists and is finite and b) *that those elements of*

(3.73) $$M_{xx} - \mathrm{M}_{xx}$$

which are subject to sampling variation have a joint asymptotically normal distribution with a variance-covariance matrix of order T^{-1}.

The matrix M_{xx} comprises the square and cross moments of the endogenous variables y_i (with and without time lags), the cross moments between the endogenous variables y_i and the exogenous variables z_k (with and without time lags), and the square and cross moments (3.68) of the exogenous variables. Since the latter variables are treated as given functions of time (see [XVII]) not subject to a probability distribution, the elements (3.68) of M_{xx} are likewise given functions of T, equal to the corresponding elements of M_{xx}.

We shall not indicate in detail the incorporation of exogenous variables in Mann and Wald's proof, since a somewhat different proof including exogenous variables will be published by one of the present authors [Rubin, 1948].

3.3.5. A property of the logarithmic quasi-likelihood function. In the remainder of this section *3.3*, we shall notationally combine in one vector θ all elements of A and Σ, and we shall write M, M instead of M_{xx}, M_{xx}. Occasionally, in particular in the present section *3.3.5*, we shall distinguish notationally between the true values θ of these parameters, and the argument $\bar{\theta}$ of the quasi-likelihood function (3.70). The logarithmic quasi-likelihood function (divided by T),

(3.74) $$L(M, \bar{\theta}) = \frac{1}{T} \log F(M, \bar{\theta})$$

as written out in (3.6), is linear in the moment matrix M. Its expected value therefore equals

(3.75) $$\mathcal{E}L(M, \bar{\theta}) = L(\mathrm{M}, \bar{\theta}),$$

a function we shall refer to as the expected logarithmic quasi-likelihood function. The expected moment matrix M occurring in (3.75) depends, for any given T, only on the fixed values of the exogenous variables $z_k(t)$, on the requisite number of initial values $x_k(t)$, $t \leq 0$, of all variables, and on the true values θ of

the parameters. We express the last-mentioned dependence by

$$M = M(\theta). \tag{3.76}$$

The function (3.75) possesses the following important property:

THEOREM 3.3.5. *Under Assumption* 3.3.1.3, *the expected logarithmic likelihood function*

$$L\{M(\theta),\ \bar{\theta}\} \tag{3.77}$$

reaches its (unrestricted) absolute maximum with respect to the parameters $\bar{\theta}$ *in the point*

$$\bar{\theta} = \theta. \tag{3.78}$$

We shall first prove this theorem under the normality Assumption 3.3.1.4. In that case the function $F(M,\ \theta)$ serves both as distribution function of the observations, and as a function defining the maximum-likelihood estimates. Therefore, if dx stands for $dx_1(1) \cdots dx_{K_y}(T)$, and $\int dx$ for integration over the whole sample space, we have $\int F(M,\ \theta)dx = 1$, and

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int F(M,\theta)dx = \int \frac{\partial F}{\partial\theta}dx = \int F\frac{\partial \log F}{\partial\theta}dx \\ &= \int F(M,\theta)\left(\frac{\partial \log F(M,\bar{\theta})}{\partial\bar{\theta}}\right)_{\bar{\theta}=\theta}dx \\ &= \left[\frac{\partial}{\partial\bar{\theta}}\int F(M,\theta)\ \log F(M,\bar{\theta})\,dx\right]_{\bar{\theta}=\theta} = \left[\frac{\partial}{\partial\bar{\theta}}\ \mathcal{E}\log F(M,\bar{\theta})\right]_{\bar{\theta}=\theta}, \\ &= T\left[\frac{\partial}{\partial\bar{\theta}}\mathcal{E}L(M,\bar{\theta})\right]_{\bar{\theta}=\theta} = T\left[\frac{\partial}{\partial\bar{\theta}}L\{M(\theta),\ \bar{\theta}\}\right]_{\bar{\theta}=\theta}, \end{aligned} \tag{3.79}$$

using (3.75) in the last equality. The differentiations with respect to θ and $\bar{\theta}$ are performed without regard to the a priori restrictions. On the other hand, we know from Theorem 3.3.10, that the function (3.77) of an unrestricted $\bar{\theta}$ is stationary only in points where its absolute maximum is reached. It follows from

(3.79) that that maximum is reached for $\bar{\theta} = \theta$.

The moments M entering in the definition (3.71) of the function (3.76) can be expressed, for any value of T, as quadratic or linear functions of the disturbances $u_g(t)$. This is seen most readily by repeated substitution of the right-hand member of the reduced form (3.11) for the $y_i(t)$ in the definition (3.7) of the moments, taking $t = T, T-1, \ldots, 1$, successively. It follows that the function $M(\theta)$ remains the same under the more general Assumption 3.3.1.3 regarding the distribution of the disturbances. Consequently (3.79), and therewith Theorem 3.3.5, are also valid under Assumption 3.3.1.3.

3.3.6. Consistency of quasi-maximum-likelihood estimates of identifiable parameters. It will be clear that any statement regarding consistency[1] of quasi-maximum-likelihood estimates can relate only to the estimation of parameters that are uniquely identifiable in a neighborhood of the true parameter point θ. Since maximum-likelihood estimation is invariant for functional transformation in the parameter space, we can achieve greater generality and flexibility by formulating our statements in terms of identifiable functions of the parameters θ, defined as follows:

Let the a priori restrictions be denoted, as in (3.43), by

$$(3.80) \qquad \varphi(\theta) \equiv \left[\varphi_1(\theta) \quad \cdots \quad \varphi_R(\theta) \right] = 0 .$$

The restrictions (3.80) define the restricted parameter space, within which as before (section 2.1.5) we distinguish sets of mutually equivalent points θ, or briefly "equivalent point sets."

DEFINITION 3.3.6.1. *A parameter $\zeta = \zeta(\theta)$ is called uniquely identifiable in a region ${}^{\tau}N$ of the restricted parameter space if it is constant, within ${}^{\tau}N$, on any set of mutually equivalent points.*

Let $\eta(\theta) \equiv \left[\eta_1(\theta) \quad \cdots \quad \eta_{P^{\oplus}}(\theta) \right]$ represent, as before, the $P^{\oplus}$ parameters Π, Ω of the reduced form (3.11) of the structural equations.

---

[1] An estimate t_q of a parameter θ_q, derived from a sample of size T, is called consistent if, for any $\varepsilon > 0$, $\lim_{T \to \infty} P(|t_q - \theta_q| > \varepsilon) = 0$, where $P(E)$ denotes the probability of an event E. This relationship of t_q and θ_q is also denoted by $\operatorname{plim}_{T \to \infty} t_q = \theta_q$.

DEFINITION 3.3.6.2. *The a priori restrictions* (3.80) *will be called regular in the parameter point* θ *if in a neighborhood* N_θ *of that point in the unrestricted parameter space the following three conditions are satisfied:*

(i) *the functions* $\varphi_r(\theta)$ *possess continuous third derivatives,*

(ii)

$$\rho\left(\frac{d\varphi}{d\theta'}\right) = R, \tag{3.81}$$

(iii)

$$\rho\left(\frac{d\eta}{d\theta'} \qquad \frac{d\varphi}{d\theta'}\right) = P^{\oplus} + R^{\oplus\oplus}, \tag{3.82}$$

say, is constant.

On the basis of these definitions, we shall prove:

THEOREM 3.3.6. *Let* $\zeta(\theta) \equiv \left[\zeta_1(\theta) \quad \cdots \quad \zeta_Q(\theta)\right]$ *be a set of* Q *parameters that*

(i) *are uniquely identifiable in a neighborhood* ${}^{r}N_\theta$ *of the true parameter point* θ *in the parameter space as restricted by means of a priori restrictions* (3.80) *regular in that point,*

(ii) *possess continuous third derivatives in a neighborhood* N_θ *of* θ *in the unrestricted parameter space containing* ${}^{r}N_\theta$, *and*

(iii) *in* N_θ *satisfy*

$$\rho\left(\frac{d\zeta}{d\theta'} \qquad \frac{d\varphi}{d\theta'}\right) = Q + R. \tag{3.83}$$

Then the quasi-maximum-likelihood estimates $\hat{\zeta}$ *of* ζ *are consistent, and have an asymptotically normal distribution with variance-covariance matrix of order* T^{-1}.

*3.3.7. *Three lemmas.* In order to prove this theorem, we shall first establish the following three lemmas.

LEMMA 3.3.7. *If the restrictions* (3.80) *are regular in the point* θ, *they imply exactly*

(3.84) $$R^{\oplus} \equiv R - R^{\oplus\oplus}$$

independent restrictions

(3.85) $$\psi(\eta) \equiv \left[\psi_1(\eta) \quad \cdots \quad \psi_{R^{\oplus}}(\eta) \right] = 0, \qquad \rho\left(\frac{d\psi}{d\eta'} \right) = R^{\oplus},$$

on the parameters $\eta(\theta)$ *of the reduced form, in a neighborhood* N_θ *of the point* θ.

Since the parameters of the reduced form are independent,

(3.86) $$\rho\left(\frac{d\eta}{d\theta'} \right) = P^{\oplus},$$

in any region of the parameter space (excluding, of course, the points with $\det B = 0$). It follows from (3.81), (3.82), (3.84), and (3.86) that

(3.87) $$0 \leq R^{\oplus\oplus} \leq R, \qquad \text{so} \qquad 0 \leq R^{\oplus} \leq R.$$

If $R^{\oplus} > 0$ and hence $R^{\oplus\oplus} < R$, it follows from (3.81), (3.82), and (3.86) that there exists a vector function of $R^{\oplus}$ elements $\psi(\eta, \varphi) \equiv \left[\psi_1(\eta, \varphi) \quad \cdots \quad \psi_{R^{\oplus}}(\eta, \varphi) \right]$ such that in N_θ

(3.88) $$\psi\{\eta(\theta), \varphi(\theta)\} = 0, \qquad \rho\begin{pmatrix} \dfrac{\partial\psi}{\partial\eta'} \\ \dfrac{\partial\psi}{\partial\theta'} \end{pmatrix} = R^{\oplus}.$$

Moreover, these functions must be such that, in the point set $N_{\eta,\varphi}$ on which N_θ is mapped through the functions $\eta(\theta)$ and $\varphi(\theta)$,

(3.89) $$\rho\left(\frac{\partial\psi}{\partial\eta'} \right) = \rho\left(\frac{\partial\psi}{\partial\varphi'} \right) = R^{\oplus},$$

For, if for instance $\rho(\partial\psi / \partial\eta')$ were less than $R^{\oplus}$, there would exist a vector function $\varkappa(\eta, \varphi)$ containing $R^{\oplus}$ elements such that on $N_{\eta,\varphi}$,

$$\frac{\partial\psi}{\partial\eta'}\,\varkappa' = 0, \qquad \frac{\partial\psi}{\partial\varphi'}\,\varkappa' \equiv \bar{\varkappa}' \neq 0. \tag{3.90}$$

In this case the equations

$$\begin{bmatrix} \dfrac{d\eta}{d\theta'} & \dfrac{d\varphi}{d\theta'} \end{bmatrix} \begin{bmatrix} \dfrac{\partial\psi}{\partial\eta'} \\ \dfrac{\partial\psi}{\partial\varphi'} \end{bmatrix} = 0 \tag{3.91}$$

obtained from (3.88) by differentiation with respect to θ' would possess a linear combination

$$\begin{bmatrix} \dfrac{d\eta}{d\theta'} & \dfrac{d\varphi}{d\theta'} \end{bmatrix} \begin{bmatrix} \dfrac{\partial\psi}{\partial\eta'} \\ \dfrac{\partial\psi}{\partial\varphi'} \end{bmatrix} \varkappa' = \frac{d\varphi}{d\theta'}\,\bar{\varkappa}' = 0, \qquad \bar{\varkappa}' \neq 0, \tag{3.92}$$

in contradiction with the regularity condition (3.81) on the a priori restrictions. Writing now

$$\psi(\eta) \equiv \psi(\eta, 0) \tag{3.93}$$

(3.85) follows from (3.88) and (3.89).

We note for later use that in ${}^{r}N_{\theta}$, as a consequence of (3.91) and (3.93)

$$\rho\left(\frac{d\eta}{d\theta'}\frac{d\psi(\eta)}{d\eta'} \quad \frac{d\varphi}{d\theta'}\right) = \rho\left(-\frac{d\varphi}{d\theta'}\frac{\partial\psi(\eta,\varphi)}{\partial\varphi'} \quad \frac{d\varphi}{d\theta'}\right) = \rho\left(\frac{d\varphi}{d\theta'}\right). \tag{3.94}$$

LEMMA 3.3.7.2. *If* Z *and* Ξ *are two matrices with equal numbers of rows and columns respectively, and* Φ *is a third matrix with an equal number of rows, such that*

$$\rho(\ (Z - \Xi)\quad \Phi\ ) = \rho(\Phi), \tag{3.95}$$

then

$$\rho(Z\quad \Phi) = \rho(\Xi\quad \Phi). \tag{3.96}$$

Proof: It follows from (3.95) that there exists a matrix Π such that

$$Z - \Xi = \Phi\,\Pi. \tag{3.97}$$

Hence

$$\rho(Z\quad \Phi) = \rho(\ (\Xi + \Phi\,\Pi)\quad \Phi) = \rho(\Xi\quad \Phi). \tag{3.98}$$

LEMMA 3.3.7.3. *If Ξ, Ψ, and Φ are three matrices with an equal number of rows, such that*

$$\rho(\Xi\quad \Phi) = \rho(\Xi) + \rho(\Phi), \quad \rho(\Xi) = c(\Xi), \quad \rho(\Phi) = c(\Phi), \tag{3.99}$$

and

$$\rho(\Psi\quad \Phi) = \rho(\Phi), \tag{3.100}$$

then

$$\rho(\Xi\quad \Psi) = \rho(\Xi) + \rho(\Psi). \tag{3.101}$$

Proof: It follows from (3.100) that there exists a matrix P such that

$$\Psi = \Phi\,P. \tag{3.102}$$

Now, if the left-hand member in (3.101) were smaller than the right-hand member, there would according to Lemma 2.3.2 exist two vectors λ and μ such that

$$\Xi\lambda' + \Psi\mu' = \Xi\lambda' + \Phi P\mu' = 0, \qquad \Xi\lambda' \neq 0. \tag{3.103}$$

Regarding $P\mu'$ as a new vector $\bar{\mu}'$, this is in contradiction with (3.99), since the second condition in (3.103) precludes the vanishing of $[\lambda\quad \bar{\mu}]$. It is easily seen that the last two conditions in (3.99) only facilitate the proof, and can be dispensed with if necessary.

*3.3.8. *First part of the proof of Theorem* 3.3.6. It was noted in section *3.1.6* that the parameter vector $\eta(\theta)$ of the reduced form is constant on each equivalent point set in the unrestricted parameter space, and that $\eta(\theta)$ assumes different values on any two different equivalent sets. Consequently, the same is true in the restricted parameter space. It follows from Definition 3.3.6.1 that $\zeta(\theta)$ is in ${}^{r}N_{\theta}$ a one-valued function $\xi(\eta)$ of $\eta(\theta)$:

$$\zeta(\theta) = \xi\{\eta(\theta)\} \quad \text{whenever} \quad \varphi(\theta) = 0. \tag{3.104}$$

Since $\zeta(\theta)$, $\eta(\theta)$, and $\varphi(\theta)$ have continuous third derivatives with respect to the elements of θ in an unrestricted neighborhood N_{θ} of θ, $\xi(\eta)$ must have continuous third derivatives with respect to the elements of η. Therefore, (3.104) implies that in a restricted neighborhood ${}^{r}N_{\theta}$ of θ, viz., in the set ${}^{r}N_{\theta}$ of those points in N_{θ} for which $\varphi(\theta) = 0$,

$$\rho\left(\left(\frac{d\zeta}{d\theta'} - \frac{d\eta}{d\theta'}\frac{d\xi}{d\eta'}\right) \quad \frac{d\varphi}{d\theta'}\right) = \rho\left(\frac{d\varphi}{d\theta'}\right) = R. \tag{3.105}$$

Thus the matrices $Z \equiv d\zeta/d\theta'$, $\Xi \equiv (d\eta/d\theta')(d\xi/d\eta')$ and $\Phi \equiv d\varphi/d\theta'$ satisfy the condition of Lemma 3.3.7.2, and, from (3.83) and (3.96), we have

$$\rho\left(\frac{d\eta}{d\theta'}\frac{d\xi}{d\eta'} \quad \frac{d\varphi}{d\theta'}\right) = Q + R. \tag{3.106}$$

Since this is the maximum possible rank for a matrix of $Q + R$ columns, we also have in ${}^{r}N_{\theta}$

$$\rho\left(\frac{d\eta}{d\theta'}\frac{d\xi}{d\eta'}\right) = Q. \tag{3.107}$$

According to (3.81), (3.101), (3.106), and (3.107), the matrices $\Xi \equiv (d\eta/d\theta')(d\xi/d\eta')$, $\Psi \equiv (d\eta/d\theta')(d\psi/d\eta')$, and $\Phi \equiv d\varphi/d\theta'$, satisfy the conditions of Lemma 3.3.7.3. It follows, using (3.89), (3.101), and (3.107), that in ${}^{r}N_{\theta}$

$$(3.108)\qquad \rho\left(\frac{d\eta}{d\theta'}\left[\frac{d\xi}{d\eta'} \quad \frac{d\psi}{d\eta'}\right]\right) = \rho\left(\frac{d\eta}{d\theta'}\frac{d\xi}{d\eta'} \quad \frac{d\eta}{d\theta'}\frac{d\psi}{d\eta'}\right) = Q + R^{\oplus}.$$

Since the rank of a matrix product does not exceed the rank of either of the two factors, we must have

$$(3.109)\qquad \rho\left(\frac{d\xi}{d\eta'} \quad \frac{d\psi}{d\eta'}\right) \geq Q + R^{\oplus}$$

in the point set ${}^{\tau}N_{\eta}$ on which ${}^{\tau}N_{\theta}$ is mapped by the function $\eta(\theta)$. However, $Q + R^{\oplus}$ is also the number of columns of the matrix in (3.109). Hence, in ${}^{\tau}N_{\eta}$

$$(3.110)\qquad \rho\left(\frac{d\xi}{d\eta'} \quad \frac{d\psi}{d\eta'}\right) = c\left(\frac{d\xi}{d\eta'} \quad \frac{d\psi}{d\eta'}\right) = Q + R^{\oplus},$$

and, because of the continuity of the functions involved, (3.110) holds also in a neighborhood N_{η}, of the point set ${}^{\tau}N_{\eta}$, in the space of the unrestricted parameters η - provided ψ is regarded as that function $\psi(\eta)$ of η only, defined by (3.93).

It follows from (3.110) that

$$(3.111)\qquad S^{\oplus} \equiv P^{\oplus} - Q - R^{\oplus} \geq 0.$$

The equality sign holds only if ζ represents a complete set of identifiable parameters. However, whenever $S^{\oplus} > 0$ we can because of (3.110) choose a vector function $\chi(\eta) \equiv \left[\chi_1(\eta) \quad \cdots \quad \chi_{S^{\oplus}}(\eta)\right]$ in N_{η} having $S^{\oplus}$ elements, with continuous third derivatives, such that in N_{η}

$$(3.112)\qquad \rho\left(\frac{d\xi}{d\eta'} \quad \frac{d\psi}{d\eta'} \quad \frac{d\chi}{d\eta'}\right) = Q + R^{\oplus} + S^{\oplus} = P^{\oplus}.$$

The matrix in (3.112) has thus been made square and nonsingular, and there exists in N_{η} an inverse function $\eta(\xi, \psi, \chi)$, i.e., a one-valued function with continuous third derivatives such that identically in N_{η},

$$(3.113)\qquad \eta\{\xi(\eta), \psi(\eta), \chi(\eta)\} = \eta, \qquad \xi(\eta) = \zeta.$$

3.3.9. Second part of the proof of Theorem 3.3.6. We shall summarize the results reached in the previous section *3.3.8*. At the same time, we shall revert to the notation used in section *3.3.5*, whereby the argument $\bar{\theta}$ of the likelihood function, and functions $\bar{\eta} = \bar{\eta}(\bar{\theta})$, $\bar{\zeta} = \bar{\zeta}(\bar{\theta})$, etc., of $\bar{\theta}$, are distinguished from the true parameter point θ and the corresponding functional values $\eta = \bar{\eta}(\theta)$, $\zeta = \bar{\zeta}(\theta)$, etc., by placing bars on the former quantities.

It has been found that the $P^{\oplus}$ parameters $\bar{\eta}$ of the reduced form of the structural equation can be expressed in a neighborhood N_{η} of η in the unrestricted space of the parameters $\bar{\eta}$, as one-valued and uniquely invertible functions $\bar{\eta} = \bar{\eta}(\bar{\zeta}, \bar{\psi}, \bar{\chi})$ possessing continuous third derivatives, of

(i) the Q identifiable parameters $\bar{\zeta} = \bar{\zeta}(\bar{\theta})$,

(ii) the $R^{\oplus}$ functions $\bar{\psi}(\bar{\eta})$ expressing the restrictions (3.85) on $\bar{\eta}$ arising from the a priori restrictions (3.80) on $\bar{\theta}$.

(iii) $S^{\oplus}$ auxiliary parameters $\bar{\chi} = \bar{\chi}(\bar{\theta})$, with $S^{\oplus} \equiv P^{\oplus} - Q - R^{\oplus} \geq 0$.

We go on to describe maximum-likelihood estimation of the parameters ζ under the restrictions (3.80) in terms of the functions that have been introduced. We start from the likelihood function (3.74) in the reduced form (3.15), now to be denoted

$$L = L^{\oplus}(M, \bar{\eta}), \tag{3.114}$$

a function possessing continuous derivatives of all orders. In this function we substitute

$$\bar{\eta} = \bar{\eta}(\bar{\zeta}, 0, \bar{\chi}) \tag{3.115}$$

– thus automatically satisfying the restrictions (3.80) – and maximize with respect to $\bar{\chi}$ for any constant $\bar{\zeta}$. Let the maximizing value of $\bar{\chi}$ be denoted

$$\hat{\chi} = \hat{\chi}(M, \bar{\zeta}). \tag{3.116}$$

This function is one-valued in a neighborhood of $M = \mathrm{M}$, $\bar{\zeta} = \zeta$, because of (3.112), and of the negative definiteness of $\partial^2 L(\mathrm{M}, \bar{\eta}) / \partial\bar{\eta}' \partial\bar{\eta}$ in the point $\bar{\eta} = \eta$, to be shown below in (3.121). Furthermore, it possesses continuous second-order derivatives because

of the continuity of the third-order derivatives of $\bar{\eta}(\bar{\zeta}, 0, \bar{\chi})$. We insert the value (3.116) in (3.115),

$$\bar{\eta} = \bar{\eta}\{\bar{\zeta}, 0, \hat{\chi}(M, \zeta)\}, \tag{3.117}$$

and write

$$L = L^{\oplus}[M, \bar{\eta}\{\bar{\zeta}, 0, \hat{\chi}(M, \bar{\zeta})\}] \equiv L^{\oplus\oplus}(M, \bar{\zeta}) \tag{3.118}$$

for the function so obtained. It follows from the invariance of any maximizing process for continuous functional transformation of the parameters that the value $\hat{\zeta}$ of $\bar{\zeta}$ maximizing $L^{\oplus\oplus}(M, \bar{\zeta})$ represents the maximum-likelihood estimate of the parameter vector ζ. Explicitly,

$$\hat{\zeta} = \bar{\zeta}(\hat{\theta}), \tag{3.119}$$

if $\hat{\theta}$ is that value of $\bar{\theta}$ maximizing the likelihood function in its original form (3.6), subject to the restrictions (3.80).

It was shown in the proof of Theorem 3.1.10 that the matrix

$$\frac{\partial^2 L^{\oplus}(M, \bar{\eta})}{\partial \bar{\eta}' \, \partial \bar{\eta}} \tag{3.120}$$

is negative definite in any point $\bar{\eta} = \bar{\eta}(\bar{\theta})$ such that the original likelihood function $L(M, \bar{\theta})$ reaches its unrestricted absolute maximum in $\bar{\theta}$. Theorem 3.3.5 states that $L(\mathrm{M}, \bar{\theta})$ reaches its unrestricted absolute maximum in the true parameter point θ. Since $\eta = \bar{\eta}(\theta)$, it follows that

$$\Lambda^{\oplus} \equiv \left(\frac{\partial^2 L^{\oplus}(\mathrm{M}, \bar{\eta})}{\partial \bar{\eta}' \, \partial \bar{\eta}} \right)_{\bar{\eta} = \eta} \tag{3.121}$$

is negative definite. We shall now prove that in consequence

$$\Lambda^{\oplus\oplus} \equiv \left(\frac{\partial^2 L^{\oplus\oplus}(\mathrm{M}, \bar{\zeta})}{\partial \bar{\zeta}' \, \partial \bar{\zeta}} \right)_{\bar{\zeta} = \zeta} \tag{3.122}$$

is also negative definite, where $\zeta = \bar{\zeta}(\theta)$ is the true value of the parameter vector ζ.

If we insert M for M in (3.117) and regard M as constant, $\bar{\eta}$ is expressed as a function

$$\bar{\eta} \equiv \bar{\eta}\{\bar{\zeta},\ 0,\ \hat{\chi}(M,\ \bar{\zeta})\} \equiv \eta(\bar{\zeta})\,, \tag{3.123}$$

say, of $\bar{\zeta}$ alone, which possesses continuous first and second derivatives, the first being

$$\left(\frac{\partial\bar{\eta}}{\partial\bar{\zeta}'} + \frac{\partial\hat{\chi}}{\partial\bar{\zeta}'}\frac{\partial\bar{\eta}}{\partial\bar{\chi}'}\right)_{\bar{\psi}=0,\ \bar{\chi}=\hat{\chi}(M,\bar{\zeta})} \equiv \frac{d\bar{\eta}}{d\bar{\zeta}'}\,, \tag{3.124}$$

say. In particular, owing to Theorem 3.5.5,

$$\bar{\eta}(\zeta) = \eta\,. \tag{3.125}$$

Differentiating (3.118) with respect to $\bar{\zeta}$, after substituting M for M, we have,

$$\frac{\partial L^{\oplus\oplus}(M,\ \bar{\zeta})}{\partial\bar{\zeta}'} = \frac{d\bar{\eta}}{d\bar{\zeta}'}\left(\frac{\partial L^{\oplus}(M,\ \bar{\eta})}{\partial\bar{\eta}'}\right)_{\bar{\eta}=\bar{\eta}(\bar{\zeta})}\,. \tag{3.126}$$

Because of Theorem 3.3.5, the quantities $\partial L^{\oplus}/\partial\bar{\eta}'$ in (3.126) vanish for $\bar{\zeta} = \zeta$. Therefore, and because continuous second derivatives of $\bar{\eta}(\bar{\zeta})$ exist, we have, using (3.125),

$$\Lambda^{\oplus\oplus} = H\,\Lambda^{\oplus}\,H'\,, \tag{3.127}$$

where

$$H \equiv \left(\frac{d\bar{\eta}}{d\bar{\zeta}'}\right)_{\bar{\zeta}=\zeta}\,. \tag{3.128}$$

From (3.127) we conclude that the quadratic form

$$z\,\Lambda^{\oplus\oplus}\,z' = z\,H\,\Lambda^{\oplus}\,H'\,z' = y\,\Lambda^{\oplus}\,y'\,, \tag{3.129}$$

say, is equal to a negative definite quadratic form, in which, owing to (3.112), $y = z\,H$ vanishes only if z vanishes. Hence $\Lambda^{\oplus\oplus}$ is negative definite, and therefore nonsingular.

It follows from Theorem 3.3.5 and from the definition of the

maximum-likelihood estimate $\bar{\zeta}$ of ζ, that the vector function

$$l^{\oplus\oplus}(\bar{M}, \bar{\zeta}) \equiv \frac{\partial}{\partial \bar{\zeta}} L^{\oplus\oplus}(\bar{M}, \bar{\zeta}) \tag{3.130}$$

vanishes in the point $\bar{M} = \mathrm{M}$, $\bar{\zeta} = \zeta$. Let us consider the Taylor expansion

$$0 = \overset{*}{l}{}^{\oplus\oplus}(M, \hat{\zeta}) - l^{\oplus\oplus}(\mathrm{M}, \zeta) = \left[\frac{\partial}{\partial \bar{\zeta}} \operatorname{tr}\left\{(M - \mathrm{M}) \frac{\partial L^{\oplus\oplus}(\mathrm{M}, \bar{\zeta})}{\partial \mathrm{M}'}\right\}\right]_{\bar{\zeta}=\zeta} + (\hat{\zeta} - \zeta)\Lambda^{\oplus\oplus} + \cdots \tag{3.131}$$

from which we can solve for $\hat{\zeta} - \zeta$ by postmultiplication with $(\Lambda^{\oplus\oplus})^{-1}$,

$$\hat{\zeta} - \zeta = -\left[\frac{\partial}{\partial \bar{\zeta}} \operatorname{tr}\left\{(M - \mathrm{M}) \frac{\partial L^{\oplus\oplus}(\mathrm{M}, \bar{\zeta})}{\partial \mathrm{M}'}\right\}\right](\Lambda^{\oplus\oplus})^{-1} + \cdots. \tag{3.132}$$

Reference to the form (3.6) of the likelihood function $L(M, \theta)$ shows that the coefficients of the elements $M - \mathrm{M}$ in (3.132) do not all vanish. It follows that $\hat{\zeta} - \zeta$ is of the same order of magnitude $T^{-\frac{1}{2}}$ as $M - \mathrm{M}$. The usual analysis of the quadratic term, omitted from (3.131), which is to be taken in a point intermediate between the two points $(M, \hat{\zeta})$ and (M, ζ), will show that this term is of order T^{-1}, because of the continuity of the second derivative of the likelihood function (3.114) and of the functions $\bar{\eta}(\bar{\zeta}, 0, \bar{\chi})$ and $\bar{\chi}(\bar{M}, \bar{\zeta})$. Therefore, $\hat{\zeta} - \zeta$ is linear in $M - \mathrm{M}$ up to terms of order T^{-1}. Theorem 3.3.6 follows from this observation combined with Theorem 3.3.4.

3.3.10. Asymptotic sampling variances and covariances of the maximum-likelihood estimates $\hat{\zeta}$. Mann and Wald's analysis shows that the expressions for the asymptotic sampling variances and covariances of the maximum-likelihood estimates $\hat{\theta}$ are greatly simplified by the normality Assumption 3.3.1.4 regarding the distribution of the disturbances. We shall here deal only with that case. The following derivation differs from that given by Mann and Wald [1943, pp. 213-214] only in that it applies to any set of parameters $\bar{\zeta}$ uniquely identifiable in a neighborhood of ζ, rather than to a complete set of identifiable parameters $\bar{\theta}$.

Under Assumption 3.3.1.4, the function F in (3.70), now to be denoted $F(M, \bar{\theta})$, also represents the probability density in the sample space. Writing

$$[\bar{\xi} \quad \bar{\chi}] \equiv \bar{\omega}, \tag{3.133}$$

we define, analogously to (3.114),

$$F(M, \bar{\theta}) = F^{\oplus}(M, \bar{\eta}) \tag{3.134}$$

and, somewhat differently from (3.118),

$$F^{\oplus}\{M, \bar{\eta}(\bar{\xi}, 0, \bar{\chi})\} \equiv \tilde{F}(M, \bar{\omega}), \tag{3.135}$$

with similar formulae in terms of $L = (1/T)\log F$. Finally, we define

$$\tilde{l}(M, \bar{\omega}) \equiv \frac{\partial}{\partial \bar{\omega}} \tilde{L}(M, \bar{\omega}). \tag{3.136}$$

Then, in the point $\bar{\omega} = \omega \equiv [\xi \quad \chi]$,

$$\begin{aligned}
\mathcal{E}\,\tilde{l}'(M,\omega)\,\tilde{l}(M,\omega) &= T^{-2}\int \tilde{F}(M,\omega)\,\frac{\partial \log \tilde{F}(M,\omega)}{\partial \omega'}\,\frac{\partial \log \tilde{F}(M,\omega)}{\partial \omega}\,dx \\
&= T^{-2}\int \frac{\partial \tilde{F}(M,\omega)}{\partial \omega'}\,\frac{\partial \log \tilde{F}(M,\omega)}{\partial \omega}\,dx \\
&= T^{-2}\int\left[\frac{\partial}{\partial \omega'}\left\{F(M,\omega)\frac{\partial \log \tilde{F}(M,\omega)}{\partial \omega}\right\}\right. \\
&\qquad \left. - \tilde{F}(M,\omega)\,\frac{\partial^2 \log \tilde{F}(M,\omega)}{\partial \omega'\,\partial \omega}\right]dx \qquad (3.137) \\
&= T^{-2}\,\frac{\partial}{\partial \omega'}\int \tilde{F}(M,\omega)\,\frac{\partial \log \tilde{F}(M,\omega)}{\partial \omega}\,dx \\
&\qquad - T^{-2}\int \tilde{F}(M,\omega)\,\frac{\partial^2 \log \tilde{F}(M,\omega)}{\partial \omega'\,\partial \omega}dx \\
&= 0 - T^{-1}\mathcal{E}\,\frac{\partial^2 \tilde{L}(M,\omega)}{\partial \omega'\,\partial \omega}
\end{aligned}$$

$$= -T^{-1}\frac{\partial^2 L(\mathrm{M},\omega)}{\partial\omega'\,\partial\omega} \equiv -T^{-1}\,\tilde{\Lambda},$$

say.

On the other hand, we have a Taylor expansion

$$-\tilde{l}(M,\omega) = \tilde{l}(M,\hat{\omega}) - \tilde{l}(M,\omega) \tag{3.138}$$
$$= (\hat{\omega}-\omega)\frac{\partial^2 L(M,\omega)}{\partial\omega'\,\partial\omega} + \cdots,$$

where the omitted term is of order $T^{-1/2}$ relative to the term shown, because of the continuity of the third derivatives of the likelihood function (3.114) and of the function $\eta(\zeta, 0, \chi)$. The nonsingularity of the matrix

$$\frac{\partial^2 \tilde{L}(M,\omega)}{\partial\omega'\,\partial\omega} \equiv \tilde{L}, \tag{3.139}$$

in a neighborhood of the point $M = \mathrm{M}$ follows directly from that of the matrix $\Lambda^{\oplus}$ defined by (3.121) and from the continuity of the derivatives involved. Therefore,

$$\hat{\omega}-\omega = -\tilde{l}(M,\omega)\tilde{L}^{-1} + \cdots, \tag{3.140}$$

and from (3.137)

$$\mathcal{E}(\hat{\omega}'-\omega')(\hat{\omega}-\omega) = -T^{-1}\tilde{L}^{-1}\tilde{\Lambda}\tilde{L}^{-1} + \cdots, \tag{3.141}$$

in which the omitted term is of order $T^{-1/2}$ relative to the term shown.

Owing to Theorem 3.3.4, property a), and the continuity of the relevant derivatives of the likelihood function,

$$\lim_{T\to\infty}\tilde{\Lambda} \equiv \tilde{\Lambda}_{\infty} \tag{3.142}$$

exists and is finite. Furthermore, because of property b) of the same theorem,

$$\operatorname*{plim}_{T\to\infty} \tilde{L} = \tilde{\Lambda}_\infty . \tag{3.143}$$

It follows from (3.141), (3.142), and (3.143), that

$$\operatorname*{plim}_{T\to\infty} T\mathcal{E}(\hat{\omega}' - \omega')(\hat{\omega} - \omega) = -\tilde{\Lambda}_\infty^{-1} .$$

This is the desired result in case χ is empty, i.e., in case $\bar{\zeta} = \bar{\omega}$ represents a complete set of unrestricted parameters. If $\bar{\zeta}$ is not complete, we shall use subscripts ζ and χ to indicate the partitioning of matrices illustrated by

$$\tilde{\Lambda} \equiv \begin{bmatrix} \tilde{\Lambda}_{\zeta\zeta} & \tilde{\Lambda}_{\zeta\chi} \\ \tilde{\Lambda}_{\chi\zeta} & \tilde{\Lambda}_{\chi\chi} \end{bmatrix}. \tag{3.144}$$

Our problem then is to evaluate

$$\operatorname*{plim}_{T\to\infty} T\mathcal{E}(\hat{\xi}' - \zeta')(\hat{\xi} - \zeta) = -(\tilde{\Lambda}_\infty^{-1})_{\zeta\zeta} \tag{3.145}$$

in terms that permit estimation on the basis of the quantities $\bar{\zeta}$ only. For this purpose we shall use the identity [Hotelling, 1943-1, p. 4]

$$(\tilde{\Lambda}^{-1})_{\zeta\zeta} = (\tilde{\Lambda}_{\zeta\zeta} - \tilde{\Lambda}_{\zeta\chi}\tilde{\Lambda}_{\chi\chi}^{-1}\tilde{\Lambda}_{\chi\zeta})^{-1}. \tag{3.146}$$

We recall the function $\hat{\chi}(M, \bar{\zeta})$ defined in (3.116), which we now need only for the argument $M = \mathrm{M}$. Besides the possession of a sufficient number of derivatives, the only property of this function used in the proof of Theorem 3.3.6 is that

$$\hat{\chi}(\mathrm{M}, \zeta) = \chi . \tag{3.147}$$

At present, we must also use the property that, owing to the definition of $\hat{\chi}(M, \bar{\zeta})$,

$$\left(\frac{\partial \tilde{L}(\mathrm{M}, \bar{\omega})}{\partial \bar{\chi}'}\right)_{\bar{\chi} = \hat{\chi}(\mathrm{M}, \bar{\zeta})} = 0 . \tag{3.148}$$

We differentiate (3.148) with respect to $\bar{\zeta}$,

$$\left(\frac{\partial^2 \tilde{L}(M,\omega)}{\partial \bar{\chi}'\,\partial \bar{\zeta}} + \frac{\partial^2 \tilde{L}(M,\bar{\omega})}{\partial \bar{\chi}'\,\partial \bar{\chi}}\,\frac{\partial \hat{\chi}(M,\bar{\zeta})}{\partial \bar{\zeta}}\right)_{\bar{\chi}=\hat{\chi}(M,\bar{\zeta})} = 0\,, \tag{3.149}$$

substitute $\bar{\zeta} = \zeta$ using (3.147), and solve as follows:

$$\frac{\partial \hat{\chi}'(M,\zeta)}{\partial \zeta} = -\tilde{\Lambda}_{\chi\chi}^{-1}\tilde{\Lambda}_{\chi\zeta}\,, \tag{3.150}$$

using the nonsingularity of $\tilde{\Lambda}_{\chi\chi}$ which follows from the negative definiteness of $\Lambda^{\oplus}$.

On the other hand, if we write

$$\tilde{L}(M,\bar{\omega}) \equiv \tilde{L}(M,\bar{\zeta},\bar{\chi})\,, \tag{3.151}$$

we have, from a comparison with (3.118),

$$L^{\oplus\oplus}(M,\zeta) = \tilde{L}\{M,\bar{\zeta},\hat{\chi}(M,\bar{\zeta})\}\,, \tag{3.152}$$

and hence, using (3.148),

$$\frac{\partial L^{\oplus\oplus}(M,\bar{\zeta})}{\partial \bar{\zeta}} = \left(\frac{\partial \tilde{L}(M,\bar{\zeta},\bar{\chi})}{\partial \bar{\zeta}}\right)_{\bar{\chi}=\hat{\chi}(M,\bar{\zeta})}\,. \tag{3.153}$$

Differentiating once more with respect to $\bar{\zeta}$ and substituting $\bar{\zeta}=\zeta$, we obtain, using (3.150),

$$\begin{aligned}\Lambda^{\oplus\oplus} = \frac{\partial^2 L^{\oplus\oplus}(M,\zeta)}{\partial \zeta'\,\partial \zeta} &= \frac{\partial^2 \tilde{L}(M,\zeta,\chi)}{\partial \zeta'\,\partial \zeta} + \frac{\partial^2 \tilde{L}(M,\zeta,\chi)}{\partial \zeta'\,\partial \bar{\chi}}\,\frac{\partial \hat{\chi}'(M,\zeta)}{\partial \zeta} \\ &= \tilde{\Lambda}_{\zeta\zeta} - \tilde{\Lambda}_{\zeta\chi}\tilde{\Lambda}_{\chi\chi}^{-1}\tilde{\Lambda}_{\chi\zeta}\,.\end{aligned} \tag{3.154}$$

Comparison with (3.145) and (3.146) now yields

$$\text{(3.155)} \qquad \operatorname*{plim}_{T\to\infty} T\mathcal{E}(\hat{\zeta}' - \zeta')(\hat{\zeta} - \zeta) = -(\Lambda_{\infty}^{\oplus\oplus})^{-1}.$$

In practice, the matrix $\Lambda_{\infty}^{\oplus\oplus}$ must be estimated from a large sample. According to (3.143), a consistent estimate of $\Lambda^{\oplus\oplus}$ is also a consistent estimate of $\Lambda_{\infty}^{\oplus\oplus}$. According to Theorems 3.3.4 and 3.3.6 and the continuity of the relevant derivatives, the former quantity can again be estimated consistently by substituting M for M and maximum-likelihood estimates $\hat{\zeta}$ for ζ. This completes the proof of

THEOREM 3.3.10. *Under Assumption* 3.3.1.3 *(normally distributed disturbances), the product of the number of observations T and the matrix of sampling variances and covariances of the maximum-likelihood estimates* $\hat{\zeta}$ *of a set of parameters satisfying the conditions of Theorem* 3.3.6 *is consistently estimated by*

$$\text{(3.156)} \qquad \operatorname{est} T\mathcal{E}(\hat{\zeta}' - \zeta')(\hat{\zeta} - \zeta) = -\left(\frac{\partial^2 L^{\oplus\oplus}(M, \bar{\zeta})}{\partial\bar{\zeta}'\,\partial\zeta}\right)_{\bar{\zeta}=\hat{\zeta}}$$

as defined further by (3.118).

In section *4.4.13* this theorem will be used to determine sampling variances and covariances of the estimates of the parameters A in cases where the sampling variances and covariances of the estimates of Σ are not required.

4. COMPUTATION OF THE MAXIMUM-LIKELIHOOD ESTIMATES

4.1. Introductory Remarks

4.1.1. Nature of the computation problem. Apart from special cases, the equations to be satisfied by the maximum-likelihood estimates of the parameters A, Σ are essentially nonlinear and of a type that does not lend itself easily to direct solution. We shall therefore study iterative methods in which a sequence of successive approximations to the solution is obtained in such a way that the essential step in the determination of each approximation constitutes a linear problem.

The present discussion is exploratory. In section 4.5 we men-

tion several important problems that are left unsolved.

The authors wish to acknowledge very valuable help received from J. von Neumann with respect to the present problem, in the form of suggestions and advice only partially acknowledged by specific reference in what follows. Much support was also found in analogies with Hotelling's iterative method [Hotelling, 1943] for inverting a matrix.

4.1.2. Notation. We shall follow the rule of denoting functions of the observations by italic characters, using that notation also for the maximum-likelihood estimates, A, S, and for successive approximations, A_n, S_n, to these estimates. This notation will also be used for the initial values A_0, S_0, even though the latter need not (but frequently will) be functions of the observations. We shall continue to use A, Σ for the arguments of the likelihood function in general. Occasionally we shall use $\mathfrak{M}$ to denote an arbitrary matrix of the same number of rows and columns as A, but which is not necessarily subject to the restrictions imposed on A.

4.1.3. Positive definiteness of M_{xx}. As before, we shall assume throughout that the moment matrix M_{xx} of the observed variables is positive definite. This assumption fails to be fulfilled only in cases occurring with probability zero, provided all linear identities are eliminated from the structural equation beforehand.

4.1.4. A special case of a priori restrictions. Before discussing the computation problem under the most general types of a priori restrictions that we have studied, it may be useful in a special and simple case to indicate a heuristic principle which has led to the computation methods discussed in what follows. In this case we assume that there is no correlation between the disturbances in different structural equations, and that normalization is imposed by taking

$$\Sigma = I. \tag{4.1}$$

We shall further assume that the only restrictions on the coefficients of the structural equations are single-parameter restrictions prescribing that certain coefficients are zero, the number of such restrictions on each equation being sufficient for its unique identification everywhere in a region N of the parameter space that contains the highest restricted maximum of the likelihood function as an internal point. The logarithm of the latter function, from

(3.6) and (4.1), is found to be, after division by T,

$$(4.2)\qquad \frac{1}{T}\log F = L(\mathrm{A}) = \text{const} + \log\det \mathrm{B} - \frac{1}{2}\operatorname{tr}(\mathrm{A}\, M_{xx}\, \mathrm{A}').$$

4.1.5. The iterative procedure now involves only the coefficients of the dependent variables. It will be noted that the coefficients Γ of the predetermined variables occur only in the last (quadratic) term in (4.2). It is therefore useful first to maximize the likelihood function with respect to the nonprescribed elements of Γ only: the maximizing values $\hat{\Gamma}$ so obtained being functions of the elements of B. The last terms in (4.2) can be written as a sum of G terms of the type

$$(4.3)\qquad -\frac{1}{2}\alpha(g)\cdot M_{xx}\cdot\alpha'(g) = -\frac{1}{2}\{\beta(g)\cdot M_{yy}\cdot\beta'(g) + 2\,\beta(g)\cdot M_{yz}\cdot\gamma'(g) + \gamma(g)\cdot M_{zz}\cdot\gamma'(g)\},$$

each term containing only coefficients of the corresponding structural equation indicated by g. Let the vector $\alpha^g \equiv [\beta^g \quad \gamma^g]$ be obtained from the gth row $\alpha(g) = [\beta(g) \quad \gamma(g)]$ of A by deleting all elements that are prescribed to be zero. Let $M^g \equiv M^g_{xx} \equiv M'^g_{xx}$ be obtained from M_{xx} by deleting the corresponding rows and columns. Then we wish to maximize

$$(4.4)\qquad -\beta^g\cdot M^g_{yz}\cdot\gamma'^g - \frac{1}{2}\gamma^g\cdot M^g_{zz}\cdot\gamma'^g$$

by variation of γ^g only. It is easily seen that the maximizing values $\hat{\gamma}^g$ of γ^g are

$$(4.5)\qquad \hat{\gamma}^g = -\beta^g\cdot M^g_{yz}\cdot(M^g_{zz})^{-1}.$$

When these values are inserted in (4.3), (4.2) becomes

$$(4.6)\qquad L^*(\mathrm{B}) = \text{const} + \log\det \mathrm{B} - \frac{1}{2}\sum_{g=1}^{G}\beta^g\cdot {}_zM^g_{yy}\cdot\beta'^g,$$

where

$$(4.7) \qquad {}^{z}M^{g}_{yy} = M^{g}_{yy} - M^{g}_{yz}\cdot(M^{g}_{zz})^{-1}\cdot M^{g}_{zy} .$$

The problem has now been reduced to finding the maximizing value B of B in (4.6). After this has been determined, the corresponding maximizing value C of Γ can be evaluated from (4.5). The computational advantage of this procedure is that the elements of C do not need to be recomputed with each iteration in the determination of B, but can be found directly from the result of the last iteration determining B.

4.1.6. Revision of a single row. Let B_0 represent a suitable initial value from which, through successive improvements, we attempt to reach the value B maximizing (4.6). The heuristic principle referred to above consists in revising only one row of B_0 at a time, as follows:

We write

$$(4.8) \qquad B_0 = B_{0,0} ,$$

and determine another matrix $B_{0,1}$, which equals $B_{0,0}$ in all elements except those of the first row, the latter being determined so as to maximize (4.6). This leads to the first-order condition[1]

$$(4.9) \qquad d_{0,1}\cdot \operatorname{cof} b^{1}_{0,1} - b^{1}_{0,1}\cdot {}^{z}M^{1}_{yy} = 0 ,$$

where $\operatorname{cof} b^{g}_{0,1}$ stands for a row vector containing as elements the cofactors in $B_{0,1}$ of the corresponding elements of $b^{g}_{0,1}$, and where the scalar quantity $d_{0,1}$ equals

$$(4.10) \qquad d_{0,1} \equiv \det{}^{-1} B_{0,1} = (b^{g}_{0,1}\cdot \operatorname{cof}' b^{g}_{0,1})^{-1} ,$$

according to the Laplace expansion of $\det B_{0,1}$. Since the elements of $\operatorname{cof} b^{g}_{0,1}$ are independent of the quantities $b^{g}_{0,1}$ now regarded as unknowns, we have a system (4.9) of linear equations in the unknowns $b^{g}_{0,1}$, $d_{0,1}$ and one quadratic equation (4.10).

Because of the positive definiteness of M_{xx}, and therefore that of ${}^{z}M^{1}_{yy}$, the unknowns $b^{1}_{0,1}$ can be solved uniquely from (4.9) in

[1]Second-order conditions will be discussed in the general case below, see section *4.3.3.3.*

terms of $d_{0,1}$ as

$$b^1_{0,1} = d_{0,1} \cdot \operatorname{cof} b^1_{0,1} \cdot ({}^{z}M^1_{yy})^{-1}. \tag{4.11}$$

The one remaining unknown $d_{0,1}$ is found, from (4.10) and (4.11), to satisfy

$$(d_{0,1})^2 = \{\operatorname{cof} b^1_{0,1} \cdot ({}^{z}M^1_{yy})^{-1} \cdot \operatorname{cof}' b^1_{0,1}\}^{-1}, \tag{4.12}$$

and can if desired be computed as the positive or negative square root of the right-hand member of (4.12). This indeterminacy of the sign of $d_{0,1}$ was to be expected, since the normalization rule (4.1) admits simultaneous changes in sign of all elements in any row of A.

4.1.7. Successive versus simultaneous revision of rows of B_0. There are two important alternative ways in which the principle of revision of a row of B_0, just described in terms of the first row, can be applied to all rows. In the first alternative, to be called successive revision of the rows of B_0, the next step is to find a matrix $B_{0,2}$ which equals $B_{0,1}$ in all elements except those of the second row, the latter being determined again so as to maximize (4.6). In this way all rows are modified successively, the result of revision of the last row

$$B_{0,G} = B_1 \tag{4.13}$$

being considered the result of the first complete iterative revision of the initial matrix B_0.

The row-by-row revision just described, by which B_1 is obtained from B_0, is economical for relatively small orders G of B, so that the computation of new cofactors after the revision of each row is not too laborious. Where economical, the row-by-row revision has a special flexibility in that one may depart from strict successive revision to give a higher frequency of revision to slowly converging subsets of the structural equations.

For larger systems, however, economy in terms of both quantity and standardization of work favors an alternative definition of B_1, which requires simultaneous revision of all rows of B_0. In this procedure, $B_{0,g}$ is defined, for every g, as being obtained from $B_{0,0}$ by the same process described above for $B_{0,1}$. Finally B_1 is such

that its gth row,

$$b_1(g) = b_{0,g}(g), \tag{4.14}$$

is equal to the gth row of $B_{0,g}$. Although one would expect slower convergence per iteration of this procedure, the saving through simultaneous computation of the cofactors of all elements cuts down the work per iteration to an extent increasing with G. All procedures studied in what follows require simultaneous revisions of all rows of B_0. An additional reason for this choice is that generalization of the procedure to the case where no restrictions are imposed on the covariance matrix Σ of the disturbances is easier and more natural if A_1 is defined by simultaneous revision of all rows of A_0.

4.1.8. Use of arbitrary scale factors in the approximations. A slight further saving arises if we realize that the only nonlinear operation in the procedure, viz., the computation of $d_{0,1}$ from (4.12), which serves only to determine common scale factors for the elements of each row of B_1, need not be carried out for any iteration except the last one. [The term "scale factor" or "scale" is used here as distinct from "normalization" because it applies only to a row $\alpha(g)$ of A, not to the corresponding row $\sigma(g)$ or column $\sigma'(g)$ of Σ. While normalization is a matter of choice, the scale factor adjusting the absolute value of $\alpha(g)$ to that of $\sigma(g)$ or σ_{gg} is of course determined by maximizing the likelihood function. See also equation (4.112) below.]

The elements of any row of B_n enter linearly and homogeneously in all relevant operations, their ratios being the relevant unknowns. For the determination of these ratios it is therefore permissible to choose any suitable value for $d_{0,g}$, for instance, unity. In what follows, it is found most suitable to substitute the known quantity $\det^{-1} B_{0,0}$ for the unknown quantity $\det^{-1} B_{0,g}$, even though the latter would give better scale factors in successive approximations. The effect of this substitution on the scale factors of the approximations B_n will be studied below. Its advantage is that each iteration is thereby completely reduced to a linear process, which can now be written

$$b_1^g \cdot {}^z M^g_{yy} = \iota(g) \cdot B_0'^{-1} \cdot \Phi^{g\prime}. \tag{4.15}$$

Here Φ^g is a matrix of which each element is either 0 or 1, such that $\alpha(g)\Phi^{g\prime}$ is a vector containing only those elements of α which are not prescribed to be zero, and $\iota(g)$ is the gth row of the identity matrix of appropriate order.

4.1.9. Saving through factorization of the likelihood function. In what follows we consider quite general homogeneous linear restrictions on the elements of A, each restriction involving elements of one row $\alpha(g)$ only. Before specializing the restrictions on Σ, we again draw attention to the possibility that the restrictions on A and Σ taken together imply a factorization of the likelihood functions as a result of the partitioning (3.63) of these matrices. In this case it is permissible to treat the two corresponding subsystems of the structural equations separately, and a considerable saving in computation work results.

The two factors of the likelihood function indicated by (3.65) and connected with the two subsystems are of the same general form as the likelihood function for the total system of structural equations. For this reason no loss of generality is involved if we assume that the equation systems considered in the remainder of this section cannot be further reduced to subsystems in this manner.

4.1.10. Two cases regarding the a priori restrictions on Σ. In two subsections we shall consider successively the case (4.3) where the disturbances are uncorrelated, and hence Σ is diagonal, and the case (4.4) where no restrictions at all are imposed on Σ. The latter case has simpler mathematical properties, although some of the formulae contain more terms and for that and other reasons the computations are more laborious. The former case, which was treated in the second place in the discussion of identification problems, is now taken up first. This is done mainly because the application of the heuristic principle indicated above is more straightforward in the case where Σ is restricted to be diagonal, while the experience so gained will be helpful in extending the methods to the case of an unrestricted Σ.

4.1.11. Dummy restrictions to insure identifiability. We shall assume throughout that each equation of the system is uniquely identifiable within a region N of the parameter space of which the highest restricted maximum of the likelihood function is an internal point. If this condition is not met initially, it must be met by adding dummy restrictions on the parameters as described in section 2.3. It was proved there that this can always be

done without further restricting the distribution function of the variables, in the case where Σ is unrestricted. No such proof was given for the case where Σ is required to be diagonal, because conditions of identifiability in that case were not fully analyzed. It may nevertheless be possible to apply the present computation methods in individual cases, either because identifiability can already be established without using the diagonality of Σ, or because a moderate size of the system permits the analysis of identification by *ad hoc* methods.

Before proceeding to the two specializations of the restrictions on Σ here considered, we shall in section 4.2 introduce a new formalism for the treatment of the restrictions on A which will facilitate the discussion of computation problems.

4.2. *A Complete Set of Unrestricted Parameters*

4.2.1. The basic matrices Φ^g. In previous sections, linear homogeneous a priori restrictions on the rows of A were used in the explicit form (2.24), which we rewrite[1]

$$\alpha(g)\,\Phi_g' = 0\,, \qquad \rho(\Phi_g) = r(\Phi_g) = R_g\,, \qquad g = 1,\ \ldots,\ G\,. \tag{4.16}$$

It will now be preferable to give implicit effect to the a priori restrictions on A by expressing all elements of A as linear functions of a basic set of unrestricted parameters. The most convenient choice of basic parameters is made separately for each row $\alpha(g)$ of A, by the use of an orthogonal complement Φ^g of Φ_g. The matrix Φ^g is defined, except for premultiplication by a nonsingular square matrix, through

$$\Phi^g\,\Phi_g' = 0, \qquad \rho(\Phi^g) = r(\Phi^g) = Q_g = K_2 - R_g\,. \tag{4.17}$$

We shall call the Φ_g, $g = 1, \ldots, G$, the *restriction matrices*, and the Φ^g the *basic matrices*. It follows from (4.16) and (4.17) that

$$\Phi(g) \equiv \begin{bmatrix} \Phi^g \\ \Phi_g \end{bmatrix} \equiv [X(g) \quad \Psi(g)]\,, \tag{4.18}$$

[1] $\rho(X)$ indicates the rank, $r(X)$ and $c(X)$ the number of rows and columns, respectively, of a matrix X.

where $X(g)$ has K_y and $\Psi(g)$ has K_z columns, is square and nonsingular and that the transformation

$$\theta(g) \equiv \theta(*g)\ \Phi(g) \equiv \theta^g\ \Phi^g + \theta_g\ \Phi_g \tag{4.19}$$

establishes a one-to-one correspondence between the space of the vector $\theta(g)$ and that of the vector

$$\theta(*g) \equiv [\ \theta^g \quad \theta_g\ ] \tag{4.20}$$

of an equal number K_x of elements. If a vector $\alpha(g)$ satisfying the restriction (4.16) is substituted for $\theta(g)$ in (4.19), we find through postmultiplication with Φ_g', using (4.17), that

$$\alpha_g\ \Phi_g\ \Phi_g' = 0 \tag{4.21}$$

and hence, from the rank condition in (4.16), that $\alpha_g = 0$. Thus $\alpha(g)$ is expressible as

$$\alpha(g) = \alpha^g\ \Phi^g. \tag{4.22}$$

Conversely, any vector so expressible satisfies the restrictions (4.16). The components of the G vectors

$$\alpha^g, \qquad g = 1, \ldots, G, \tag{4.23}$$

represent an unrestricted set of parameters[1] except for such rules of normalization as it may be convenient to impose in certain cases.

The freedom of premultiplication by a nonsingular matrix in choosing each Φ^g should be used to make its form as simple as possible for computational purposes. Often the restrictions on A arise from the elimination of variables connected by identities to the variables retained in the system. Such elimination leads directly to the matrices Φ^g, without need for prior evaluation of the matrices Φ_g.

---

[1]The notation α^g employed in section 4.1 represents a special case of the present notation. If matrices Φ^g were constructed for that special case, they would contain elements 1 and 0 only, with at most one 1 to each row or column. An example is contained in formula (4.128) below.

For certain purposes, especially in presenting theory, it is convenient to choose the rows of Φ^g orthogonal to each other, so that, after suitable normalization,

$$\Phi^g \Phi'^g = I. \tag{4.24}$$

From a computational point of view, such orthogonalization is often not necessary, and if carried out may increase the computational labor.

4.2.2. Normalization. If, as in subsection 4.3, we assume that Σ is diagonal, it is convenient for most purposes to normalize by equating the diagonal elements σ_{gg} of Σ to unity,

$$\Sigma = I. \tag{4.25}$$

For some purposes, however, it may be convenient not to restrict the σ_{gg}, but to normalize on one element of each α^g, by

$$\alpha^g \iota'(1) = 1, \qquad g = 1, \ldots, G, \tag{4.26}$$

say, if as before $\iota(1)$ indicates a vector of the appropriate order, of which the first element is 1 and all other elements are 0. These purposes include the calculation of sampling variances and covariances of the estimates a^g of the α^g. The normalization (4.26) will at any rate be applied in subsection 4.4 where Σ is unrestricted.

4.2.3. The matrix A *treated as a vector.* A set of parameters θ_{gk}, $g = 1, \ldots, K_y$ $(\equiv G)$, $k = 1, \ldots, K_x$, can be considered either as a matrix

$$\Theta = \begin{bmatrix} \theta(1) \\ \vdots \\ \theta(K_y) \end{bmatrix} \tag{4.27}$$

k of K_y rows and K_x columns, or as a vector θ defined by

$$\theta \equiv \operatorname{vec} \Theta \equiv [\, \theta(1) \quad \theta(2) \quad \cdots \quad \theta(K_y) \,]. \tag{4.28}$$

For certain operations, notably forming the determinant and the inverse of B, the matrix representation is convenient. Other operations, in particular those connected with the a priori restrictions, are simpler in the vector representation (4.28), because the restrictions are different for different rows of A. Finally, if we define M by[1]

$$(4.29) \qquad M \equiv \begin{bmatrix} M_{xx} & 0 & \cdots & 0 \\ 0 & M_{xx} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & M_{xx} \end{bmatrix} = I_{[G]} \otimes M_{xx},$$

and, if H, η are connected in a manner analogous to (4.28), the relations

$$(4.30) \qquad \operatorname{tr} \circledast H' = \Theta \eta', \qquad \operatorname{vec}(\circledast M_{xx}) = \Theta M, \qquad \operatorname{tr} \circledast M_{xx} \circledast' = \Theta M \Theta',$$

connect expressions of a simple type in both representations. The formal framework of the following analysis of computation problems will be an alternating use of the matrix and vector representations of the parameter space, taking advantage of the special properties of each.

4.2.4. *Projection of a matrix on the restricted parameter space.* We define the matrix

$$(4.31) \qquad \Phi(*) \equiv \begin{bmatrix} \Phi^* \\ \Phi_* \end{bmatrix} \equiv \begin{bmatrix} \Phi^1 & 0 & \cdots & 0 \\ 0 & \Phi^2 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \Phi^G \\ \Phi_1 & 0 & \cdots & 0 \\ 0 & \Phi_2 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \Phi_G \end{bmatrix},$$

[1] The symbol $\otimes$ denotes the "direct product" or "Kronecker product" of two matrices; see [MacDuffee, p. 81].

which, owing to (4.16) and (4.17), is nonsingular and satisfies

$$\Phi^* \Phi_*' = 0. \tag{4.32}$$

Φ^* is called the basic matrix, Φ_* the restriction matrix. The transformation (4.19) for the individual vectors $\theta(g)$, $g = 1, \ldots, G$, can now be summarized in

$$\theta = \theta(*)\,\Phi(*) = \theta^* \Phi^* + \theta_* \Phi_*, \tag{4.33}$$

where

$$\theta(*) = [\theta^* \ \theta_*], \qquad \begin{cases} \theta^* = [\theta^1 \ \theta^2 \ \cdots \ \theta^G], \\ \theta_* = [\theta_1 \ \theta_2 \ \cdots \ \theta_G]. \end{cases} \tag{4.34}$$

In particular, if $\alpha = \operatorname{vec} A$ arises from a matrix A satisfying the a priori restrictions (4.16), which in the new notation take the form

$$\alpha \Phi_*' = 0, \tag{4.35}$$

we must have

$$\alpha^* = [\alpha^1 \ \alpha^2 \ \cdots \ \alpha^G], \qquad \alpha_* = 0. \tag{4.36}$$

Thus, under the normalization rule (4.25), the elements of the vector α^* constitute a complete set of unrestricted parameters through which the original restricted parameters are expressed by

$$\alpha = \alpha^* \Phi^*. \tag{4.37}$$

Under the normalization (4.26), the complete set of unrestricted parameters consists of those elements of α^* not prescribed by (4.26), plus the other diagonal elements σ_{gg}, or all elements σ_{gh}, $h \geq g$, of the symmetric matrix Σ, according to the case considered.

Through (4.33) the arbitrary vector θ is expressed uniquely as the sum of two vectors, the first $\theta^* \Phi^*$ lying within the restricted parameter space, the second orthogonal to that space. This decomposition plays an essential role in what follows. Since it will also be applied to cases where Θ is given as a matrix

product, it is convenient to introduce, in addition to the notations used in (4.33), the operator notations

$$\theta(*) \equiv \mathrm{vec} * \Theta, \qquad \theta^* \equiv \mathrm{vec}^* \Theta, \qquad \theta_* \equiv \mathrm{vec}_* \Theta. \tag{4.38}$$

In these terms (4.33) runs

$$\theta = \mathrm{vec}\ \Theta = (\mathrm{vec}^*\ \Theta)\,\Phi^* + (\mathrm{vec}_*\ \Theta)\,\Phi_*, \tag{4.39}$$

from which we can solve for $\mathrm{vec}^*\ \Theta$ through postmultiplication by $\Phi'^*(\Phi^*\ \Phi'^*)^{-1}$ using (4.32), thus obtaining

$$\mathrm{vec}^*\Theta = (\mathrm{vec}\ \Theta)\ \Phi'^*(\Phi^*\ \Phi'^*)^{-1}. \tag{4.40}$$

Conversely, we define

$$\Theta \equiv \mathrm{mat}\ \theta \equiv \mathrm{mat} * \theta(*) \equiv \mathrm{mat} * \mathrm{vec} * \Theta, \tag{4.41}$$

and

$$\mathrm{mat} * \theta^* \equiv \mathrm{mat} * [\theta^* \quad 0_*], \qquad \mathrm{mat} * \theta_* \equiv \mathrm{mat} * [0^* \quad \theta_*]. \tag{4.42}$$

The decomposition (4.39) can thus be written in matrix coordinates as

$$\Theta = \mathcal{Q}\Theta + \mathcal{R}\Theta, \tag{4.43}$$

where

$$\mathcal{Q}\Theta \equiv \mathrm{mat} * \mathrm{vec}^*\Theta, \qquad \mathcal{R}\Theta \equiv \mathrm{mat} * \mathrm{vec}_*\ \Theta. \tag{4.44}$$

The operation $\mathcal{Q}\Theta$ can be regarded as a projection of the matrix Θ on the restricted parameter space. In particular, the a priori restrictions (4.35) on the matrix A are equivalent to

$$A = \mathcal{Q}A, \quad \text{or} \quad \mathcal{R}A = 0. \tag{4.45}$$

We shall operate mostly in the vector space of θ^*, but occasionally return to the restricted matrix space of $\mathcal{Q}\Theta$ through the transformation mat *.

If the matrix H satisfies the restrictions $\eta_* = 0$, we have the

important property that

$$\text{(4.46)} \qquad \operatorname{tr} \Theta H' = \theta \eta' = (\theta^* \Phi^* + \theta_* \Phi_*) \Phi'^* \eta'^* = \theta^* \Phi^* \Phi'^* \eta'^*$$

because of (4.32) does not depend on θ_*. In particular we have:

LEMMA 4.2.4. *A necessary and sufficient condition that* $\operatorname{tr} \Theta H' = 0$ *for all values of* H *satisfying the a priori restrictions* $\eta_* = \operatorname{vec}_* H = 0$ *is that* $\theta^* \equiv \operatorname{vec}^* \Theta = 0$.

The proof follows from (4.46) and the nonsingularity of $\Phi^* \Phi'^*$ due to the rank condition in (4.17).

4.3. *The Case of Uncorrelated Disturbances*

4.3.1. *The nature of the problem.*

4.3.1.1. The maximum-likelihood equations. In the present subsection, the matrix Σ of variances and covariances of the disturbances is assumed to be diagonal. Unless otherwise stated, normalization will be based throughout on (4.25) where Σ is equated to the unit matrix.

We shall now write down the first-order conditions for a maximum of the logarithmic likelihood function (4.2) which we rewrite:

$$\text{(4.47)} \qquad L(A) = \text{const} + \log \det B - \tfrac{1}{2} \operatorname{tr}(A M_{xx} A').$$

Let A_0 be a trial value of A satisfying the restrictions $\alpha_* = 0$, and write (δA_0 here denoting a *finite* change in A_0)

$$\text{(4.48)} \qquad A = A_0 + \delta A_0, \qquad \operatorname{vec}_* \delta A_0 = 0,$$

which insures that A again satisfies the restrictions. The Taylor expansion of $L(A)$ in the neighborhood of $\delta A_0 = 0$ then contains the following constant and linear terms, derived with the use of (3.16) and (3.17),

$$\begin{aligned} L(A) &= L(A_0) + \operatorname{tr}\{B_0'^{-1} (\delta B_0)'\} - \operatorname{tr}\{A_0 M_{xx} (\delta A_0)'\} + \cdots \\ &= L(A_0) + \operatorname{tr}\{(B_0'^{-1} I_{[K_y K_x]} - A_0 M_{xx})(\delta A_0)'\} + \cdots. \end{aligned} \qquad \text{(4.49)}$$

(Here $I_{[K_y K_x]}$ is a submatrix, of K_y rows and K_x columns, of the unit matrix of order K_x, as follows: $I_{[K_y K_x]} = [I_{[K_y]} \quad 0]$, the first matrix in the right-hand member being the unit matrix of order K_y). According to Lemma 4.2.4, the necessary first-order condition for A_0 to coincide with a restricted maximum A of the likelihood function is therefore

$$(4.50)\quad \begin{cases} (4.50q) & \operatorname{vec}^* (B'^{-1} I_{[K_y K_x]} - A M_{xx}) = 0, \\ (4.50r) & a_* \equiv \operatorname{vec}_* A = 0. \end{cases}$$

These are the maximum-likelihood equations that are to be solved by an iterative process. If desired, the operators vec^* and vec_* can be replaced by $\mathfrak{D}$ and $\mathfrak{R}$, thus reverting to a matrix form.

It will be noted that the number of conditions equals the number of unknowns. If identification is incomplete, the equations become interdependent.

4.3.1.2. Solutions without restrictions on A. Further light is thrown on the mathematical nature of this problem if we first consider the case where no restrictions at all are imposed on the matrix A. Then a_* has no elements and the symbol vec^* can be omitted in (4.50q), so that

$$(4.51)\quad \begin{cases} (4.51y) & B'^{-1} - A M_{xy} = 0, \\ (4.51z) & A M_{xz} = 0. \end{cases}$$

Of these equations (4.51z) is solved by expressing A linearly in terms of an orthogonal complement of $M'_{xz} = M_{zx}$ as follows:

$$(4.52)\qquad A = B[I_{[K_y]} \quad -M_{yz} M_{zz}^{-1}],$$

with B an arbitrary nonsingular matrix of order K_y. Substituting this result in (4.51) we obtain as the condition on B

$$(4.53)\qquad B'^{-1} - B(M_{yy} - M_{yz} M_{zz}^{-1} M_{zy}) = 0,$$

or

$$B'B = (M_{yy} - M_{yz} M_{zz}^{-1} M_{zy})^{-1}, \tag{4.54}$$

which is solved, but for an arbitrary orthogonal matrix O, by

$$B = O(M_{yy} - M_{yz} M_{zz}^{-1} M_{zy})^{-\frac{1}{2}}, \tag{4.55}$$

provided the inverse square root is taken to be symmetric. The right-hand member in (4.54) can also be denoted by $(M_{xx}^{-1})_{yy}$. The solutions (4.52) and (4.55) can also be obtained from the conditions (3.58) for the absolute maximum of the likelihood function by choosing a value B of B which will make Σ equal to the unit matrix as at present required.

It thus appears that in the absence of all restrictions on A, our problem is of the nature indicated by (4.54), leaving an arbitrary orthogonal matrix O in the solution. If identifying restrictions are now added gradually, more and more restrictions are imposed on O. The existence of a solution A of (4.51) within the restrictions is the necessary and sufficient condition for the likelihood function to be able to attain its absolute maximum. Since we assume here that the total set of restrictions (a priori and dummy combined) identifies each structural equation, there is just one special case in which a solution of (4.51) is still possible. This is the case in which the total set of restrictions is just adequate in number and variety, in accordance with Definition 3.6, to identify all structural equations. In that case a possible computation procedure would be to find one particular solution of (4.54) and then to determine O in such a way that A satisfies the restrictions. However, even in this case, computational economy may still favor the iterative methods developed below.

As soon as the restrictions are more than adequate in number and variety with respect to at least one structural equation, equation (4.54) cannot in general be satisfied any longer. We then have a more general problem where (4.54) must, in some sense, be satisfied as nearly as possible within the linear restrictions on B arising from those on A.

4.3.2. *The methods* $\mathfrak{P}_1$, $\mathfrak{P}_h$, *and* $\mathfrak{P}_{h_n}$.

4.3.2.1. Choice of the linear path toward the next approximation. Suppose now that in A_0 a restricted stationary value of the

likelihood function is not reached, but in point A in the neighborhood of A_0 a restricted maximum is reached. If we write as the next approximation

$$A_1 = A_0 + h\Delta A_0, \tag{4.56}$$

the matrix ΔA_0 indicates the direction on the linear path on which the next approximation is sought and the scalar h determines the distance traveled along that path. In sections 4.3.2 to 4.3.4 inclusive, we shall be concerned with methods based on the following choice of ΔA_0:

$$(4.57)\quad \begin{cases} (4.57^*) & \operatorname{vec}^*(\Delta A_0 \cdot M_{xx}) = \operatorname{vec}^*(B_0'^{-1} I_{[K_y K_x]} - A_0 M_{xx}), \\ (4.57_*) & \Delta a_{0*} = \operatorname{vec}_* \Delta A_0 = 0. \end{cases}$$

We shall first show that a sufficiently small value of h will always lead to an increase in the likelihood function. If $h\Delta A_0$ is substituted for δA in (4.49), we have, on account of (4.49), (4.57), and (4.46),

$$\begin{aligned} L(A_1) - L(A_0) &= h \operatorname{tr}\{(B_0'^{-1} I_{[K_y K_x]} - A_0 M_{xx})(\Delta A_0)'\} + \cdots \\ &= h \{\operatorname{vec}^*(B_0'^{-1} I_{[K_y K_x]} - A_0 M_{xx})\} \Phi^* \cdot \Phi'^* \cdot \Delta a_0^* + \cdots \\ &= h \{\operatorname{vec}^*(\Delta A_0 \cdot M_{xx})\} \Phi^* \cdot \Phi'^* \cdot \Delta a_0^* + \cdots \\ &= h \operatorname{tr}(\Delta A_0 \cdot M_{xx} \cdot \Delta A_0') + \cdots. \end{aligned} \tag{4.58}$$

Because of the positive definiteness of M_{xx}, the coefficient of h in this expansion is positive unless ΔA_0 vanishes identically, in which case a restricted stationary value of the likelihood function would already have been reached in A_0. If h is sufficiently small but positive, therefore, the linear term in (4.58) will exceed in absolute value the sum of all subsequent terms, and $L(A_1) > L(A_0)$.

4.3.2.2. Choices of the next approximation on the linear path selected. The process $\mathfrak{P}_1$. We shall postpone until section 4.3.2.5 the proof that (4.57) always admits of one and only one solution ΔA_0, and now discuss possible choices of h. We shall first show that the process obtained through the choice

$$h = 1, \tag{4.59}$$

now to be called the process $\mathcal{P}_1$, is equivalent to the process demonstrated in a special case in section 4.1. For that choice of h, A_1 is defined by

$$(4.60)\quad \begin{cases} (4.60^*) & \text{vec}^*\,(A_1\,M_{xx}) = \text{vec}^*\,[B_0'^{-1} \quad 0]\,, \\ (4.60_*) & a_{1*} = \text{vec}_*\,A_1 = 0\,. \end{cases}$$

Furthermore, in the case referred to, all restrictions are of the single-parameter type which require certain elements of A to vanish. In this case, a suitable choice for each Φ^g in (4.22) is obtained by deleting from the unit matrix $I_{[K_x]}$ all rows corresponding to elements in $\alpha(g)$ that are required to vanish. We shall return (4.60) to matrix form by applying mat $*$ to all members:

$$(4.61)\quad \begin{cases} (4.61\mathcal{Q}) & \mathcal{Q}(A_1\,M_{xx}) = \mathcal{Q}[B_0'^{-1} \quad 0], \\ (4.61\mathcal{R}) & \mathcal{R}A_1 = 0. \end{cases}$$

because in the present case the operation $\mathcal{Q}$ consists simply in replacing by zero all elements of the matrix on which $\mathcal{Q}$ operates corresponding to elements of A that are required to vanish by (4.61$\mathcal{R}$). In this case, therefore, if the partitioning of Θ corresponding to $A = [B \quad \Gamma]$ is denoted momentarily by $\Theta = [H \quad Z]$, the definition of $\mathcal{Q}$ can be extended to submatrices of Θ through

$$(4.62)\qquad \mathcal{Q}\Theta \equiv [\mathcal{Q}H \quad \mathcal{Q}Z],$$

and (4.61$\mathcal{Q}$) partitions into

$$(4.63\mathcal{Q})\quad \begin{cases} (4.63\mathcal{Q}_y) & \mathcal{Q}(A_1\,M_{xy}) = \mathcal{Q}B_0'^{-1}, \\ (4.63\mathcal{Q}_z) & \mathcal{Q}(A_1\,M_{xz}) = 0. \end{cases}$$

Of these conditions, (4.63$\mathcal{Q}_z$) is equivalent to (4.5), and (4.63$\mathcal{Q}_y$) to (4.15). This result establishes a presumption that $\mathcal{P}_1$ will have satisfactory convergence properties, except perhaps with respect to a common scale factor for each row of A_n.

4.3.2.3. The process $\mathcal{P}_{1/2}$. In a special borderline case the

choice $h = \frac{1}{2}$ leading to the process $\mathfrak{P}_{1/2}$ has superior convergence properties. This is the case where there are no predetermined variables and no restrictions on the matrix B. (This, of course, implies dropping, for the present example, the previous assumption that each equation is completely identified.) In this case all matrices involved are square matrices of order G, and $A \equiv B$, $M_{xx} \equiv M_{yy}$. The process $\mathfrak{P}_{1/2}$ now runs

$$A_1 = \frac{1}{2}(A_0'^{-1} \cdot M_{xx}^{-1} + A_0). \tag{4.64}$$

Simple calculations will show that this process possesses the property

$$A_1 M_{xx} A_1' - I = \frac{1}{4}(A_0 M_{xx} A_0' - I)(A_0 M_{xx} A_0')^{-1}(A_0 M_{xx} A_0' - I). \tag{4.65}$$

Since under the present assumptions the maximum-likelihood equations (4.51) to be solved iteratively are equivalent to

$$A M_{xx} A' = I, \tag{4.66}$$

(4.65) implies a high rate of convergence of $\mathfrak{P}_{1/2}$, once convergence is obtained initially. The extent to which A_1 fails to satisfy (4.66), as measured by $A_1 M_{xx} A_1' - I$, is of second order compared with the corresponding quantity $A_0 M_{xx} A_0' - I$ in terms of A_0. It follows that the number of decimal places which is correct in the nth iteration increases geometrically with n.

It will be clear that in cases where $\mathfrak{P}_{1/2}$ possesses this very desirable property, a process $\mathfrak{P}_h$ based on any other constant value of h will produce a lower-order speed of convergence.

Of course, the solution of (4.66) is determined but for an orthogonal transformation, and it depends on the initial value A_0 which particular solution of (4.66) is approached by successive iterations. If the order G of the matrix A is reduced to one, the indeterminacy disappears, and (4.64) is specialized to a well-known iterative procedure,

$$a_1 = \frac{1}{2}(\frac{m^{-1}}{a_0} + a_0), \tag{4.67}$$

to obtain the square root of a scalar m^{-1}.

4.3.2.4... The process $\mathfrak{P}_{h_n}$. Von Neumann has suggested determining h afresh with each iteration from the requirement that the sum of the linear and quadratic terms in the Taylor expansion of the likelihood function with respect to h shall have a vanishing first derivative. Extending the expansion (4.58), with the aid of (3.18), to

$$L(A_1) - L(A_0) = h \operatorname{tr}\{(B_0'^{-1} I_{[K_y K_x]} - A_0 M_{xx})(\Delta A_0)'\}$$

$$+ \frac{1}{2}h^2 \operatorname{tr}\{-B_0'^{-1}(\Delta B_0)' B_0'^{-1}(\Delta B_0)' \tag{4.68}$$

$$- (\Delta A_0) M_{xx} (\Delta A_0)'\} + \cdots,$$

and using, as in (4.58), the definition of ΔA_0 given by (4.57), we find that the value,

$$h_0 = \frac{\operatorname{tr} \Delta A_0 \cdot M_{xx} \cdot (\Delta A_0)'}{\operatorname{tr}\left(\{B_0'^{-1}(\Delta B_0)'\}^2 + \Delta A_0 \cdot M_{xx} \cdot (\Delta A_0)'\right)}, \tag{4.69}$$

of h satsifies the criterion mentioned. This procedure, which we denote by $\mathfrak{P}_{h_n}$, may be expected to have an asymptotic speed of convergence superior to that obtained by any constant value of h. In particular, in cases like the preceding example where one particular constant value of h leads to a higher-order speed of convergence than all other constant values, the value h_n according to (4.69) may be expected to converge to that constant for $n \to \infty$.

4.3.3. *Asymptotic convergence properties of* $\mathfrak{P}_1$, $\mathfrak{P}_h$, $\mathfrak{P}_{h_n}$

4.3.3.1. Orthogonalizing the basic matrix Φ^*. The foregoing tentative discussion of various processes can be made more definite by studying their properties in the neighborhood of a restricted maximum A of the likelihood function. For this purpose, it is useful to assume that the basic matrix Φ^* defined by (4.31) is orthogonal according to (4.24). If the restriction matrix Φ_* is similarly orthogonalized, we have

$$\Phi(*)\,\Phi'(*) = I, \qquad \text{so} \quad \Phi^{-1}(*) = \Phi'(*). \tag{4.70}$$

The relationships (4.30) can then be extended in a simple way to the θ-space. We register for future use:

$$\begin{gathered} \operatorname{tr}(\Theta\, H') = \theta(*)\,\eta'(*), \qquad \operatorname{vec}_*(\Theta\, M_{xx}) = \theta(*)\, M(*), \\ \operatorname{tr}(\Theta\, M_{xx}\, \Theta') = \theta(*)\, M(*)\, \theta'(*), \end{gathered} \tag{4.71}$$

where

$$M(*) = \Phi(*)\, M\, \Phi'(*) = M'(*) = \begin{bmatrix} M^{**} & M^{*}_{*} \\ M_{*}^{*} & M_{**} \end{bmatrix}. \tag{4.72}$$

In particular we have, if H satisfies the restrictions $\eta_* = 0$, the important identities

$$\begin{aligned} \operatorname{tr}\, \Theta\, H' &= \theta^*\, \eta'^* \\ \operatorname{vec}^*(H\, M_{xx}) &= \eta^*\, M^{**}, \\ \operatorname{tr}(H\, M_{xx}\, H') &= \eta^*\, M^{**}\, \eta'^*. \end{aligned} \tag{4.73}$$

Incidentally, it follows from the second equality in (4.73) and the nonsingularity of M^{**} that (4.57) always admits of one and only one solution ΔA_0, a statement previously made without proof. This conclusion remains true even when identification is incomplete, since the nonsingularity of M^{**} is not affected thereby.

4.3.3.2. Analysis of asymptotic convergence properties. Let successive iterations be

$$A_n = A + \bar{A}_n, \qquad \bar{a}_{n*} = \operatorname{vec}_* \bar{A}_n = 0, \qquad n = 0, 1, \ldots . \tag{4.74}$$

We shall only consider linear and sometimes quadratic terms in Taylor expansions with respect to $\bar{A}_0$, $\bar{A}_1$. In terms of $\bar{A}_n$ the iterative process (4.56) runs

$$\bar{A}_1 = \bar{A}_0 + h(\Delta \bar{A}_0), \tag{4.75}$$

where $\Delta\bar{A}_0$ is identical with ΔA_0 as defined by (4.57). The Taylor expansion of (4.57*) now becomes, because of (4.50q),

$$\text{(4.76)} \qquad \text{vec}^*\,(\Delta\bar{A}_0 M_{xx}) = \text{vec}^*\,(-B'^{-1}\bar{B}_0'\,B'^{-1}I_{[K_y K_x]} - \bar{A}_0\,M_{xx}) + \cdots .$$

We shall study this relation in the space of the unrestricted vectors $\bar{a}_0^* \equiv \text{vec}^*\,\bar{A}_0$ and $\Delta\bar{a}_0^* \equiv \text{vec}^*\,\Delta\bar{A}_0$. The term shown in the right-hand member of (4.76) is most easily understood in relation to the quadratic term in $\bar{A}_0$ in the Taylor expansion of the likelihood function $L(A_0)$ based on the point A. We now write for the latter term

$$\text{(4.77)} \qquad \begin{aligned} \frac{1}{2}\,L_{(2)}(A_0) &\equiv \frac{1}{2}\,\text{tr}(-B'^{-1}\bar{B}_0'\,B'^{-1}\,\bar{B}_0' - \bar{A}_0\,M_{xx}\,\bar{A}_0') \\ &\equiv \frac{1}{2}\,\bar{a}_0^*\,L^*\,\bar{a}_0'^*, \end{aligned}$$

thereby uniquely defining a symmetric matrix L^*. The explicit evaluation of L^* is immaterial at this stage, and will be demonstrated below in formulae (4.183) and (4.188) dealing with an analogous matrix L_0^*.

Differentiating the middle member of (4.77) with respect to $\bar{a}_0^*$, we have on the one hand, using the first relation (4.73), a row vector

$$\text{(4.78)} \qquad \begin{aligned} \frac{1}{2}\frac{dL_{(2)}}{d\bar{a}_0^*} &= \frac{\frac{1}{2}\,\text{tr}\left(\dfrac{dL_{(2)}}{d\bar{A}_0}\,d\bar{A}_0'\right)}{d\bar{a}_0^*} \\ &= \frac{\text{tr}\{(-B'^{-1}\,\bar{B}_0'\,B'^{-1}\,I_{[K_y K_x]} - \bar{A}_0\,M_{xx})\,d\bar{A}_0'\}}{d\bar{a}_0^*} \\ &= \frac{\text{vec}^*\,(-B'^{-1}\,\bar{B}_0'\,B'^{-1}I_{[K_y K_x]} - \bar{A}_0\,M_{xx})\,d\bar{a}_0^*}{d\bar{a}_0^*} \\ &= \text{vec}^*\,(-B'^{-1}\,\bar{B}_0'\,B'^{-1}\,I_{[K_y K_x]} - \bar{A}_0\,M_{xx}). \end{aligned}$$

Comparing this with the result of differentiating the last member of (4.77), we find that (4.76) is equivalent to

$$\Delta\bar{a}_0^* M^{**} = \bar{a}_0^* L^* + \cdots, \tag{4.79}$$

in which the first member is obtained through the second relation (4.73). Through one further transformation

$$\begin{aligned} M^{**} &\equiv R^* R'^*, & L^\dagger &\equiv (R^*)^{-1} L^* (R^*)'^{-1}, \\ \Delta a_0^\dagger &\equiv \Delta\bar{a}_0^* R^*, & \bar{a}_0^\dagger &\equiv \bar{a}_0^* R^*, \end{aligned} \tag{4.80}$$

this goes over into

$$\Delta\bar{a}_0^\dagger = \bar{a}_0^\dagger L^\dagger + \cdots, \tag{4.81}$$

so that

$$\bar{a}_1^\dagger = \bar{a}_0^\dagger (I + hL^\dagger) + \cdots. \tag{4.82}$$

In the absence of higher-order terms, this iterative process has been studied for constant h by Hotelling [1933, 1936]. Its properties depend on the characteristic values k_q, $q = 1, \ldots, Q$, of the matrix

$$K^\dagger \equiv I + hL^\dagger. \tag{4.83}$$

The choice of R^* in (4.80) is determined except for postmultiplication by an orthogonal matrix, and this freedom can be used to make $L^\dagger$, and hence $K^\dagger$, diagonal. The diagonal elements of these matrices then equal the characteristic values l_q and k_q, respectively, which are connected by

$$k_q = 1 + hl_q. \tag{4.84}$$

We are free to arrange the values (4.84) in descending order of algebraic magnitude[1] through suitable choice of R^*. The elementary vectors

$$\iota(q) = [0_1 \quad \cdots \quad 0_{q-1}\ 1_q \quad 0_{q+1} \cdots \quad 0_Q], \quad q = 1, \ldots, Q, \tag{4.85}$$

[1] Since we restrict ourselves to positive values of h, the descending order applies simultaneously to l_q and k_q.

form an orthogonal set of corresponding characteristic vectors, i.e.,

$$\iota(q)K^{\dagger} = k_q\,\iota(q), \qquad \iota(q)L^{\dagger} = l_q\,\iota(q),$$

$$\iota(q_1)\,\iota'(q_2) = \begin{cases} 0 & \text{if } q_1 \neq q_2, \\ 1 & \text{if } q_1 = q_2. \end{cases} \tag{4.86}$$

If

$$\bar{a}_n^{\dagger} = \sum_{q=1}^{Q} \{\bar{a}_n^{\dagger}\,\iota'(q)\}\,\iota(q)\ , \tag{4.87}$$

the iterative process (4.82) consists - to the first order of magnitude - of a multiplication

$$\bar{a}_1^{\dagger}\,\iota'(q) = k_q\{\bar{a}_0^{\dagger}\,\iota'(q)\} \tag{4.88}$$

of the qth component $\bar{a}_0^{\dagger}\,\iota'(q)$ underlying $\bar{A}_0$ by the factor k_q. Apart from the effect of higher-order terms (which is smaller, the nearer the initial value A_0 is to the solution A), convergence is assured if all characteristic values k_q are smaller than unity in absolute value.

4.3.3.3. Identifiability of structural equations and nonsingularity of $L^{\dagger}$. We have assumed that in A a maximum of the likelihood function is reached under completely identifying restrictions. In connection with the definition (4.77) of L^{*}, the fact that A is a maximum precludes the possibility that any characteristic value l_q of $L^{\dagger}$ be positive. It does not necessarily preclude singularity of $L^{\dagger}$, since a maximum might be reached in a point where second derivatives vanish in a certain direction (e.g., the function $-x^4$ in the point $x = 0$). However, unless all third derivatives of the likelihood function vanish in A, complete identifiability of structural equations implies that L^{*} will be negative definite. If that is the case, all characteristic values l_q are negative, and a sufficiently small value of h will insure that $|k_q| < 1$ for all $q = 1, \ldots, Q$, and will thus insure convergence from initial values A_0 sufficiently near to A. It is also clear from (4.84) that too small a value of h will make convergence quite

slow.

4.3.3.4. Importance of the case in which the maximum of the likelihood function is not depressed by the restrictions. In one important special case it is not difficult to make further statements about the characteristic values of $L^{\dagger}$, and therefore about those of the matrix (4.83). This is the case, discussed in section 3.2.2, in which the a priori restrictions do not depress the likelihood function (4.47) below its absolute maximum. It was shown above that this will be the case if and only if (4.51) permits a solution within the restrictions.

It was also noted that, since we are assuming complete identification of each structural equation, this case can occur identically in the sample space only if the restrictions are just adequate in number and variety for complete identification. This will rarely be so in systems of appreciable size, unless the investigator chooses to ignore the excess of a priori information over the minimum essential for identification. If excess information is available and is used, the case of a nondepressed likelihood function can still occur with probability zero in the sample space if a sample is drawn with "exceptional" values of M_{xx} that permit a restricted solution A of (4.51). Again this remark would be of little practical value, were it not that, in a large majority of sufficiently large samples, values of M_{xx} are obtained that are not far removed from a value M_{xx} that permits the likelihood function to reach its absolute maximum, provided the a priori information embodied in the restrictions is actually valid in the population. For in that case, Theorems 3.3.5 and 3.3.4 apply. In the first place, if the expectation,

$$\mathrm{M}_{xx} = \mathcal{E} M_{xx}, \tag{4.89}$$

of the moment matrix is inserted in the likelihood function, the absolute maximum can be reached by inserting the true values of the parameters A, Σ, values which obviously tally with any valid a priori information about these parameters. Secondly, we have

$$\operatorname{plim}(M_{xx} - \mathrm{M}_{xx}) = 0. \tag{4.90}$$

Thus any statement about the characteristic values of $L^{\dagger}$ that is based on the assumption that (4.51) possesses a solution under valid restrictions is approximately true with high probability in sufficiently large samples.

4.3.3.5. Characteristic values of L^* if the maximum of the likelihood function is not depressed. It is well known that the characteristic values l_q of $L^\dagger$ are the stationary values of the quadratic forms in $\bar{a}^\dagger$,

$$L_{(2)} = \bar{a}^\dagger L^\dagger \bar{a}'^\dagger = \operatorname{tr}\{-(B'^{-1}\bar{B}')^2 - \bar{A} M_{xx} \bar{A}'\}, \tag{4.91}$$

under the restrictions on $\bar{a}^\dagger$,

$$1 = \bar{a}^\dagger \bar{a}'^\dagger = \operatorname{tr} \bar{A} M_{xx} \bar{A}'. \tag{4.92}$$

Since a one-to-one correspondence has been established between the vectors $\bar{a}^\dagger$ and the matrices $\bar{A}$ satisfying the a priori restrictions, the values l_q are also the stationary values of the last member of (4.91) subject to the a priori restrictions on $\bar{A}$ plus the restrictions (4.92).

Let us now make use of the assumption that A satisfies (4.51) to supplement it to a nonsingular matrix,

$$H \equiv \begin{bmatrix} A \\ D \end{bmatrix} \equiv \begin{bmatrix} B & C \\ 0 & F \end{bmatrix}, \quad \text{where} \quad F' F = M_{zz}^{1}, \tag{4.93}$$

which is such that

$$H M_{xx} H' = I, \quad \text{or} \quad M_{xx} = H^{-1} H'^{-1}. \tag{4.94}$$

If we now transform $\bar{A}$ uniquely by

$$\bar{A} \equiv \tilde{A} H \equiv \tilde{B} A + \tilde{C} D, \quad \text{or} \quad \tilde{A} \equiv \bar{A} H^{-1} \equiv [\tilde{B} \quad \tilde{C}], \tag{4.95}$$

the linear a priori restrictions on $\bar{A}$ entail similar restrictions on $\tilde{A}$. We shall refer to the latter restrictions as the $\tilde{A}$-restrictions, it being implied in the use of this expression that the a priori restrictions on A permit the likelihood function to reach its absolute maximum. Upon inserting (4.95) in (4.91) and (4.92) and using (4.93) and (4.94), we find that the l_q are the stationary values of the quadratic form

$$(4.96)\qquad L_2 = \operatorname{tr}(-\tilde{B}'^2 - \tilde{A}\,\tilde{A}') = -\sum_{g,h=1}^{K_y} \tilde{a}_{gh}\,\tilde{a}_{hg} - \sum_{g=1}^{K_y}\sum_{k=1}^{K_x} \tilde{a}_{gk}^2,$$

subject to the $\tilde{A}$-restrictions and the additional restriction

$$(4.97)\qquad \operatorname{tr}\tilde{A}\,\tilde{A}' = \sum_{g=1}^{K_y}\sum_{k=1}^{K_x} \tilde{a}_{gk}^2 = 1,$$

where $\tilde{a}_{gk}$ denotes the element of $\tilde{A}$ in the gth row and kth column.

Let us first revert to the case where no a priori restrictions are imposed on $\bar{A}$ and hence none on $\tilde{A}$. Then the values l_q, $q = 1, \ldots, P$, are those values of l that permit a solution $\tilde{A}$ of

$$(4.98)\qquad \frac{1}{2}\frac{d}{d\tilde{A}}(L_{(2)} - l\operatorname{tr}\tilde{A}\,\tilde{A}') = -\tilde{B}'\,I_{[K_y K_x]} - (1+l)\tilde{A} = 0.$$

The following table, whose entries are easily verified by substitution, contains all possible solutions, since the sum of the multiplicities of the characteristic values (determined as the corresponding number of linearly independent characteristic "vectors" $\tilde{A}$) equals $P = K_y K_x$.

$$(4.99)$$

(a) Value of l	(b) Value of $k = 1 + hl$	(c) Multiplicity	(d) Characteristic "vectors" $\tilde{A}$ satisfy
0	1	$\frac{1}{2}K_y(K_y - 1)$	$\tilde{B} = -\tilde{B}'$, $\tilde{C} = 0$
-1	$1 - h$	$K_y K_z$	$\tilde{B} = 0$, $\tilde{C}$ arbitrary
-2	$1 - 2h$	$\frac{1}{2}K_y(K_y + 1)$	$\tilde{B} = \tilde{B}'$, $\tilde{C} = 0$

Since the values $l = 0$ and $l = -2$ are the extrema of the form (4.96) under the sole restriction (4.97), we have

$$(4.100)\qquad -2 \le L_{(2)} \le 0 \quad \text{if} \quad \operatorname{tr}\tilde{A}\,\tilde{A}' = 1.$$

Consequently, if a priori restrictions are now introduced that do not restrict the maximum of the likelihood function, the new values

l_q and k_q must satisfy

$$(4.101) \quad -2 \leq l_q \leq 0, \qquad 1-2h \leq k_q \leq 1, \qquad q = 1, \ldots, Q.$$

Any particular characteristic value in (4.98) will remain a characteristic value under the a priori restrictions if the ensuing $\tilde{A}$-restrictions permit the corresponding condition in (4.99), column (d), to be satisfied. Its multiplicity then equals the number of independent "vectors" $\tilde{A}$ satisfying that condition and the restrictions.[1]

4.3.3.6. Exclusion of the characteristic value $l = 0$ through complete identification. Under the present restrictions (4.25) on Σ, but in the absence of any restrictions on A, the transformations

$$(4.102) \qquad A^{\oplus} = \Upsilon A$$

preserving the form of the likelihood function are orthogonal:

$$(4.103) \qquad \Upsilon \Upsilon' = I.$$

Let $A + \tilde{A}$ be obtained from A by such a transformation. Then

$$(4.104) \qquad \tilde{C} = 0, \qquad \tilde{B} = \Upsilon - I,$$

and, up to the first-order terms in $\tilde{B}$,

$$(4.105) \qquad \Upsilon \Upsilon' = (I + \tilde{B})(I + \tilde{B}') = I + \tilde{B} + \tilde{B}' + \cdots = I,$$

so that in first approximation $\tilde{B} = -\tilde{B}'$. Thus the characteristic vectors (4.99) associated with $l = 0$ represent only directions of change from A corresponding to orthogonal transformations of A. In the case of complete (unique or multiple) identification of all structural equations, all such transformations leading from a point A satisfying the restrictions to a point $A + \bar{A}$ in a sufficiently small neighborhood of A are excluded by the restrictions on $A + \bar{A}$. In that case, therefore, the characterictic value $l = 0$ is no longer present, and

[1] Except that, with probability zero in the sample space, a "new" characteristic value $l = -1$ may be added, with a "vector" that is a linear combination of "vectors" corresponding to $l = -2$ and $l = 0$, respectively, in (4.99).

(4.106) $-2 \leq l_q < 0, \quad 1 - 2h \leq k_q < 1, \quad q = 1, \ldots, Q.$

4.3.3.7. Complete identification through restrictions on A *permits only trivial solutions with* $l = -2$. Let us now assume that identification is now complete on the basis of the restrictions on A alone, i.e., without recourse to (4.25). Then all points

(4.107) $$A + \bar{A} = (I + \tilde{B})A, \qquad (\tilde{C} = 0),$$

permitted by the restrictions on $A + \bar{A}$ are such that

(4.108) $\tilde{B}$ is diagonal.

Since one can select only K_y linearly independent diagonal matrices of order K_y, the value $l = -2$ now has its multiplicity reduced to K_y,

(4.109) $$l_q = -2, \qquad q = Q - K_y + 1, \ldots, Q.$$

The accompanying characteristic "vectors" (4.108) correspond to the application of arbitrary scale factors $1 + c_{gg}$ to the rows $a(g)$ of A.

It follows that the remaining characteristic values now satisfy

(4.110) $-2 < l_q < 0, \quad 1 - 2h < k_q < 1, \quad q = 1, \ldots, Q - K_y,$

and that the corresponding characteristic "vectors" satisfy $\tilde{C} \neq 0$.

4.3.3.8. Asymptotic properties of $\mathfrak{P}_1$. The foregoing analysis suggests that, among processes $\mathfrak{P}_h$ employing a constant value of h, $\mathfrak{P}_1$ is suitable if we have a large sample from a distribution satisfying known restrictions on the matrix A that are sufficient for its identification. For all relevant characteristic values are smaller in absolute value than unity, and convergence can be expected if the initial value A_0 is sufficiently close to the solution A. The only characteristic of A that does not participate in the convergence is a set of scale factors for the respective rows $a_n(g)$, $g = 1, \ldots, G$. It is useful to define the scales of any approximation A_n by

$$m\{a_n(g)\} \equiv \{a_n(g)\, M_{xx}\, a_n'(g)\}^{1/2} \tag{4.111}$$

because the corresponding expressions for A satisfy

$$m\{a(g)\} \equiv \{a(g)\, M_{xx} a'(g)\}^{1/2} = 1, \tag{4.112}$$

as is easily proved by proportional variations of all elements of $\alpha(g)$ in the likelihood function (4.47). Now if $m\{a_0(g)\}$ is sufficiently different from unity, successive approximations of $\mathfrak{P}_1$ will, under the present assumptions, exhibit scales $m\{a_n(g)\}$ differing from unity by an amount asymptotically constant in absolute value but alternating in signs since the corresponding characteristic value satisfies $k = 1 + l = -1$. If desired for aesthetic or practical reasons, this oscillation of scales can be reduced by occasional modification of the scales so as to make $m\{a_n(g)\} = 1$, or by occasional application of $\mathfrak{P}_{1/2}$ instead of $\mathfrak{P}_1$, whereby the characteristic values concerned become $1 + \frac{1}{2}l = 0$.

It has already been shown in section *4.1.8* that the alternating behavior of scales does not affect the speed of convergence of the ratios between the elements of any row $a_n(g)$. The asymptotic value of that speed is determined by the largest of the quantities $|k_q|$, $q = 1, \ldots, Q - K_y$. If knowledge of these values were available beforehand, it would be possible to choose another constant value of h that minimizes the largest of the $|k_q|$, $q = 1, \ldots, Q - K_y$. In the absence of such knowledge, the following considerations favor the use of $\mathfrak{P}_1$ in the circumstances at present assumed:

(a) The cost of computation per iteration for $\mathfrak{P}_1$ is below that for any other constant value of h, and considerably below that for the process $\mathfrak{P}_{h_n}$ with the variable value (4.69) of h_n.

(b) The interval to which l_q is confined according to (4.110) is such that values of h larger than unity may lead to divergent processes. Of the admissible values $0 < h \leq 1$, only the value $h = 1$ leads to an interval (4.110) for k_q of which 0 is the midpoint.

The second consideration is based on complete ignorance of the range of relevant values l_q inside the interval $(-2, 0)$, and may lose its weight if more experience about these values is accumu-

lated from economic data, or if the interplay of the restrictions with the conditions in the last column of (4.99) is analyzed theoretically more completely than is done here. In any case, the risk that some values l_q might be close to -2 presents an additional reason for interspersing the iterations of $\mathfrak{P}_1$ with an occasional application of $\mathfrak{P}_{1/2}$, which cuts down those components $\bar{a}^{\dagger} \cdot \iota'(q)$ of $\bar{A}$ corresponding to characteristic values l_q nearest to -2.

4.3.3.9. Case where identification depends on the diagonality restriction on Σ. If the restrictions on A alone are insufficient for identification of all structural equations, there exist nondiagonal matrices $\tilde{B}$ such that $A + \bar{A}$ in (4.107) again satisfies the restrictions. The possibility exists that among those matrices $\tilde{B}$ at least one symmetric matrix can be found. In that case, at least one of the characteristic values l_q, $q = 1, \ldots, Q - K_y$, not associated with trivial scale factors, equals -2, and if $\mathfrak{P}_1$ is applied iteratively, the corresponding component $\bar{a}_0^{\dagger}\, \iota'(q)$ in the initial value is not reduced to zero in successive iterations. It may be possible, by criteria similar to those developed in section 2, to determine whether or not the restrictions on A permit at least one of the nondiagonal matrices $\tilde{B}$ in (4.107) to be symmetric.[1] Alternatively, one may apply $\mathfrak{P}_h$ with a constant value of $h < 1$, say $h = 3/4$, or one may insert $\mathfrak{P}_{1/2}$ at regular intervals. One initial application of $\mathfrak{P}_{1/2}$ would indeed cut out, up to the first order of magnitude, those components $\bar{a}_0^{\dagger}\, \iota'(q)$ of $\bar{A}$ that cannot be reduced by $\mathfrak{P}_1$, but the presence of higher-order terms in (4.88) may reintroduce those components to some extent, thus requiring that $\mathfrak{P}_{1/2}$ now be inserted with regularity.

4.3.3.10. Asymptotic properties of $\mathfrak{P}_{h_n}$. Alternatively, whether nontrivial characteristic values $l = -2$ are present or not, one may pay the higher cost per iteration of the process $\mathfrak{P}_{h_n}$ already described in order to obtain at each stage a value h_n of h that is already optimal (in a first-order sense) in relation to the relative sizes of the various components $\bar{a}_n^{\dagger}\, \iota'(q)$ present in $\bar{A}_n$. By transformation of ΔA_0 in (4.69) to the space of $\Delta \bar{a}_0^{\dagger}$ we obtain

[1] Such an inquiry might also throw further light on questions left unanswered in section 2.4.

$$h_0 = -\frac{\Delta\bar{a}_0^\dagger \cdot \Delta\bar{a}_0'^\dagger}{\Delta\bar{a}_0^\dagger \cdot L^\dagger \cdot \Delta\bar{a}_0'^\dagger} = -\frac{\bar{a}_0^\dagger\,(L^\dagger)^2\,\bar{a}_0'^\dagger}{\bar{a}_0^\dagger\,(L^\dagger)^3\,\bar{a}_0'^\dagger} + \cdots$$

(4.113)

$$= -\frac{\sum_{q=1}^{Q} l_q^2\,\{\bar{a}_0^\dagger\,\iota'(q)\}^2}{\sum_{q=1}^{Q} l_q^3\,\{\bar{a}_0^\dagger\,\iota'(q)\}^2} + \cdots .$$

It is easily seen that the lowest-order term shown in the last member of (4.113) represents that value of h_0 which minimizes the lowest-order term shown in the last member of

$$\bar{a}_1^\dagger\,\bar{a}_1'^\dagger = \sum_{q=1}^{Q} \{\bar{a}_1^\dagger\,\iota'(q)\}^2$$

(4.114)

$$= \sum_{q=1}^{Q} \{1 + h_0\,l_q\}^2 \{\bar{a}_0^\dagger\,\iota'(q)\}^2 + \cdots ,$$

in keeping with the principle from which the process $\mathfrak{P}_{h_n}$ is derived.

Another interesting property of $\mathfrak{P}_{h_n}$ may be mentioned. If we postmultiply (4.81) by $(L^\dagger)^{-1}$ and insert the resulting expression for $\bar{a}_n^\dagger$ in (4.82),[1] we obtain as the equivalent of the iterative procedure (4.82) in terms of $\Delta\bar{a}_n^\dagger$, using the value (4.113) of h_n,

$$\Delta\bar{a}_1^\dagger = \Delta\bar{a}_0^\dagger - \frac{\Delta\bar{a}_0^\dagger \cdot \Delta\bar{a}_0'^\dagger}{\Delta\bar{a}_0^\dagger \cdot L^\dagger \cdot \Delta\bar{a}_0'^\dagger} \cdot \Delta\bar{a}_0^\dagger \cdot L^\dagger + \cdots . \tag{4.115}$$

Postmultiplication with $\Delta\bar{a}_0'^\dagger$ shows that successive adjustments $\Delta\bar{a}_n^\dagger$ of $\bar{a}_n^\dagger$ are, to a first approximation, orthogonal to each other. Figure 4.3.3.10 demonstrates the first-order properties of (4.115). It is seen from (4.115) that first-order terms in $\mathfrak{P}_{h_n}$ are homogeneous of degree zero in the characteristic values l_q of $L^\dagger$. It fol-

[1] Taking $n = 0$ and 1, respectively.

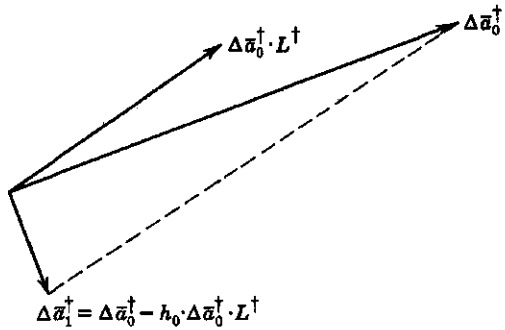


Figure 4.3.3.10

lows that only the ratios of the values l_q affect the asymptotic speed of convergence of $\mathfrak{P}_{h_n}$. This circumstance appears in the figure, in that only the direction, not the length, of the vector $\Delta\bar{a}_0^{\dagger} L^{\dagger}$ determines the vector $\Delta\bar{a}_1^{\dagger}$. This shows that characteristic values l_q near to zero, indicating proximity to a situation of incomplete identification using all restrictions, are a more fundamental difficulty in computing maximum-likelihood estimates than characteristic values l_q near to or equal -2, indicating "almost" or "altogether" lack of identification using restrictions on A alone. Components corresponding to the latter characteristic values can always be reduced by a suitable value of h.

4.3.3.11. Convergence properties of $\mathfrak{P}_h$ and $\mathfrak{P}_{h_n}$ under incomplete identification. It should be added that if any of the processes so far discussed is applied in cases where identification is incomplete (using all restrictions), and where therefore at least one characteristic value l_q vanishes, this circumstance is not revealed by the convergence properties of the process. Corresponding components $\Delta\bar{a}_0^{\dagger}$ in the initial displacement $\bar{A}_0$ are preserved through iterative multiplication by a factor unity. In other words, while reasonably fast convergence may be obtained if no other characteristic values near to zero exist, the limiting value $A_{\infty} = A + \bar{A}_{\infty}$ now depends on the initial components $\Delta\bar{a}_0^{\dagger}$ concerned.

4.3.3.12. Comparisons between $\mathfrak{P}_h$ and $\mathfrak{P}_{h_n}$. Experience with

actual data is needed to determine whether the greater speed of convergence that might be expected of $\mathcal{P}_{h_n}$ as compared with $\mathcal{P}_h$ for any constant h justifies the greater cost of computation per iteration. One would expect $\mathcal{P}_{h_n}$ to be particularly economical if the ratios of the characteristic values l_q are close to unity. In any case, $\mathcal{P}_{h_n}$ would seem to be less subject to as yet unknown risks connected with the distribution of the characteristic values l_q, with the selection of the initial approximation A_0, and with the sampling variations of M_{xx} around its expected value (4.89), which leads in general to depression of the restricted maximum of the likelihood function.

*4.3.3.13. *Effect of the $\tilde{A}$-restrictions on the characteristic values l_q.* On the basis of (4.99) we decompose $\tilde{A}$ uniquely in terms of characteristic "vectors" of four types:

$$\begin{aligned}\tilde{A} &= [\tilde{B} \quad \tilde{C}] \\ &= [\tilde{B}_{\text{dia}} \quad 0] + [\tilde{B}_{\text{sym}} \quad 0] + [\tilde{B}_{\text{ant}} \quad 0] + [0 \quad \tilde{C}] \\ &= \tilde{B}_{\text{dia}} I_{[K_y K_x]} + \tilde{B}_{\text{sym}} I_{[K_y K_x]} + \tilde{B}_{\text{ant}} I_{[K_y K_z]} + \tilde{C} I'_{[K_x K_z]} ,\end{aligned} \tag{4.116}$$

$\tilde{B}_{\text{dia}}$ is diagonal,

$\tilde{B}_{\text{sym}} = \tilde{B}'_{\text{sym}}$ is symmetric with vanishing diagonal elements,

$\tilde{B}_{\text{ant}} = -\tilde{B}'_{\text{ant}}$ is antisymmetric.

In the absence of any restrictions, (4.116) gives the components of $\tilde{A}$ according to the subspaces corresponding to the three different characteristic values of $L^{\dagger}$. We have further decomposed the symmetric component of $\tilde{B}$ into a diagonal component $\tilde{B}_{\text{dia}}$ and a component $\tilde{B}_{\text{sym}}$ with vanishing diagonal elements, because of the trivial nature of the components $\tilde{B}_{\text{dia}}$.

With a priori restrictions of the type here considered, whenever $\tilde{A}$ satisfies the $\tilde{A}$-restrictions, the component $\tilde{B}_{\text{dia}} I_{[K_y K_x]}$ satisfies the same restrictions. In any study of the effect of these restric-

tions on the characteristic values l_q, $q = 1, \ldots, Q$, therefore, the component $\tilde{B}_{\text{dia}} I_{[K_y K_x]}$ and the corresponding characteristic values l_q, $q = Q - K_y + 1, \ldots, Q$, play no role.

If $\tilde{A}$-restrictions are imposed on $\tilde{A}$, the decomposition (4.116) is still unique but it does not in general represent $\tilde{A}$ as a linear combination of characteristic "vectors." It will do so if and only if each of the last three components in (4.116) also satisfies the $\tilde{A}$-restrictions. Whenever an $\tilde{A}$-restricted $\tilde{A}$ exists for which at least one of those three components fails to satisfy the $\tilde{A}$-restrictions, at least one new "intermediate" characteristic value l_q has been introduced of which the (each) characteristic "vector" $\tilde{A}^{(i)}$ is a linear combination $\tilde{B}^{(i)}_{\text{sym}} I_{[K_y K_x]} + \tilde{B}^{(i)}_{\text{ant}} I_{[K_y K_x]} + \tilde{C}^{(i)} I'_{[K_x K_z]}$ with at least two nonvanishing components. This is seen as follows: If, after imposing the $\tilde{A}$-restrictions, a complete set of new characteristic "vectors" could be chosen in such a manner that each of them consisted of one component (4.116) only, the unique expression of $\tilde{A}$ as a linear combination of those new characteristic "vectors" [derived from (4.87) by transformation to the $\tilde{A}$-space] would coincide with the unique decomposition (4.116). This would contradict the assumption that at least one component (4.116) of $\tilde{A}$ fails to satisfy the $\tilde{A}$-restrictions.

*4.3.3.14. *Case where all relevant characteristic values coincide at* $l = -2$. We have met already with one example where the a priori restrictions imply $\tilde{A}$-restrictions that are preserved in the decomposition (4.116). This is the case, discussed in section 4.3.2.3 in connection with $\mathbb{P}_{1/2}$, where the only a priori restrictions are

(4.117) $$C = 0, \quad \text{and hence} \quad \bar{C} = 0.$$

From (4.93) and (4.95) the corresponding $\tilde{A}$-restrictions are seen to be

(4.118) $$\tilde{C} = 0.$$

This wipes out all characteristic values $l = -1$, while leaving unaffected the values $l = 0$ and $l = -2$ and the corresponding vectors. It follows that one application of $\mathbb{P}_{1/2}$ will reduce all components corresponding to $l = -2$ to second-order magnitude.

This accounts for the high speed of convergence of $\mathcal{P}_{1/2}$ found in this special case. It is seen as follows, however, that a single additional restriction on B would introduce a characteristic value $-2 < l < 0$ with probability one in the sample space. Let the additional restriction be denoted $\operatorname{tr} \bar{A}\, \Phi' = 0$. The $\tilde{A}$-restriction $\operatorname{tr} \tilde{B}\, A\, \Phi' = 0$ will only then be identically satisfied by $\tilde{B}_{\text{sym}}\, I_{[K_y K_x]}$ if it also implies $\operatorname{tr} \tilde{B}' A\, \Phi' = \operatorname{tr} \Phi\, A' \tilde{B} = \operatorname{tr} \tilde{B}\, \Phi\, A' = 0$, which requires $A\, \Phi' = \Phi\, A'$, an occurrence of probability zero.

*4.3.3.15. *Case where all characteristic values coincide at* $l = -1$. It is of interest to inquire whether there exists a counterpart to the foregoing case, in which the nontrivial component $\tilde{B}_{\text{sym}}\, I_{[K_y K_x]}$ with the characteristic value $l = -2$ is wiped out by restrictions without introducing any intermediate characteristic values. Such a case can be found easily if we also require complete identification, i.e., wiping out of the component $\tilde{B}_{\text{ant}}\, I_{[K_y K_x]}$ with characteristic value $l = 0$. For in that case the restrictions must imply that

$$\tilde{a}_{gh} = \bar{a}(g)\, H^{-1}\, \iota'(h) = \bar{a}(g) \begin{bmatrix} B^{-1} \\ 0 \end{bmatrix} \iota'(h)$$

$$= \bar{a}^g\, \mathrm{X}^g\, B^{-1}\, \iota'(h) = 0 \quad \text{for} \quad g \neq h, \quad g, h \leq G, \tag{4.119}$$

whatever $\bar{a}^g$, or that

$$\mathrm{X}^g\, B^{-1}\, \iota'(h) = 0 \quad \text{for} \quad g \neq h. \tag{4.120}$$

Here X^g is a submatrix of Φ^g as defined in (4.18). This again requires the existence of column vectors λ'^g such that

$$\mathrm{X}^g = \lambda'^g\, b(g), \tag{4.121}$$

so that the matrix B is restricted by

$$\beta(g) = \alpha^g\, \mathrm{X}^g = \alpha^g\, \lambda'^g\, b(g) \equiv \lambda\, b(g), \tag{4.122}$$

where λ is the scalar quantity $\alpha^{g}\ \lambda'^{g}$. Thus, the ratios of all elements of any row of B must be prescribed by the restrictions, and B is the product $\Lambda_{yy}\ B^{\oplus}$ of an unknown nonsingular diagonal matrix Λ_{yy} with a known nonsingular matrix $B^{\oplus}$. Hence a known nonsingular transformation,

$$y' = B^{\oplus}\ y'^{\oplus}, \tag{4.123}$$

of the dependent variables alone will then bring the system of structural equations into the form

$$\Lambda_{yy}\ y'^{\oplus} + \Gamma\ z' = u', \tag{4.124}$$

which differs from the reduced form only by the principle of normalization. In this form, therefore, maximum-likelihood estimation is equivalent to the least-squares principle applied to each equation separately.

Thus, the exclusion of both components $\tilde{B}_{\text{sym}}\ I_{[K_y\ K_x]}$ and $\tilde{B}_{\text{ant}}\ I_{[K_y\ K_x]}$ leads to a trivial case which can be treated by more elementary methods. In fact, one single application of $\mathfrak{P}_1$, which is the optimal procedure in the present case, is the equivalent to the least-squares procedure and leads to the exact solution of the maximum-likelihood equation in one iteration. However, if only a single restriction on B is relaxed, nondiagonal values of $\tilde{B}$ are made possible that are in general neither symmetric nor antisymmetric, so that new intermediate characteristic values are introduced. A simple example is obtained if in (4.124) we assume only $K_y = 2$ dependent variables $y^{\oplus}$ and leave λ_{12} unrestricted while λ_{21} is still required to vanish.

4.3.3.16. Asymptotic speed of convergence of $\mathfrak{P}_h = \mathfrak{P}_{h_n}$ in the foregoing case. This discussion shows that cases where all relevant characteristic values coincide are rare and, from the point of view of the problem of measuring economic relationships, either trivial or accidental. In general, different characteristic values are present, and the discrepancy $\bar{A}_{n+1} = A_{n+1} - A$ between the result of the $(n + 1)$th iteration and the solution of the maximum-likelihood equations is even asymptotically a nonvanishing fraction of the corresponding difference $\bar{A}_n$ computed from the result of the nth iteration with any of the methods so far discussed.

As to speed of convergence per iteration, therefore, the present methods fall short of the well-known Newton method, in which by definition $\bar{A}_{n+1}$ is of second-order magnitude compared with $\bar{A}_n$.

4.3.4. *Arrangement of computations for* $\mathfrak{P}_1$, $\mathfrak{P}_h$, $\mathfrak{P}_{h_n}$

4.3.4.1. A constructed example for numerical illustration of $\mathfrak{P}_h$, $\mathfrak{P}_{h_n}$ *involving only single-parameter restrictions.* Before exploring the application of the Newton method to our present problem, we shall give a few numerical illustrations of the procedures so far discussed, and give further comments of a practical character with regard to computational procedure.

We have constructed an example of a three-equation system characterized by the matrices

$$A = \begin{bmatrix} 0 & 1 & 4 & 1 & 0 & 0 \\ 1 & 0 & -3 & 0 & 1 & 0 \\ -2 & 1 & 0 & 0 & 0 & 1 \end{bmatrix},$$

(4.125)

$$\Sigma = \begin{bmatrix} 0.2 & 0.1 & 0.0 \\ 0.1 & 0.2 & 0.1 \\ 0.0 & 0.1 & 0.3 \end{bmatrix},$$

$$M_{zz} = I,$$

in which the variables z are regarded as fixed in repeated samples. The matrix M_{yx} then fluctuates from sample to sample, but, in order to abstract from the effect of sampling fluctuations, we have assumed that the observed moment matrix in the sample imagined to be drawn is equal to its expected value, which is easily computed from (4.125):

$$M_{yx} = \mathrm{M}_{yx} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{E} y'(t)\, x(t) \tag{4.126}$$

$$= B^{-1}\,\Sigma\,B'^{-1}\,I_{[K_y K_x]} - B^{-1}\,\Gamma\,M_{zz}\,[-\Gamma'\,B'^{-1} \quad I_{[K_z]}]$$

$$= \begin{bmatrix} 0.417 & 0.484 & -0.021 & -0.300 & -0.400 & 0.300 \\ 0.484 & 1.568 & -0.192 & -0.600 & -0.800 & -0.400 \\ -0.021 & -0.192 & 0.073 & -0.100 & 0.200 & 0.100 \end{bmatrix}.$$

In the present section 4.3 the diagonality restriction is imposed on Σ, although the example has been constructed with a non-diagonal Σ. Comparison of the "maximum-likelihood estimates" of A so obtained with the true values will illustrate the effect of the diagonality assumption regarding Σ in a case where it is incorrect.

The restrictions on A are that those elements that are zero in (4.125) are known and required to be zero. In that case it is profitable to state the definition (4.57) of ΔA_0 in matrix form

$$\mathcal{Q}(\Delta A_0\, M_{xx}) = \mathcal{Q}(B_0'^{-1}\, I_{[K_y K_x]} - A_0\, M_{xx}),$$
$$\mathcal{Q}\,\Delta A_0 = \Delta A_0, \tag{4.127}$$

because the operator $\mathcal{Q}$ now consists in replacing by zero the elements numbered (1,1), (1,5), (1,6), (2,2), (2,4), (2,6), (3,3), (3,4), (3,5), and can according to (4.62) also be applied to submatrices. Alternatively, the restrictions can be expressed through the set of basic matrices

$$\Phi^1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

$$\Phi^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \tag{4.128}$$

$$\Phi^3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

4.3.4.2. Elimination of the unknowns C_n under single-parameter restrictions. If $\mathcal{P}_1$ is applied,

$$\mathcal{Z}(A_n M_{xz}) = 0, \qquad \mathcal{Z}(\Delta A_n M_{xz}) = 0 \tag{4.129}$$

will automatically be satisfied for $n \geq 1$. The condition (4.129) will be recognized as the condition (4.5) for maximization of the likelihood function with respect to the parameters Γ only. For other choices of h_n it is possible and desirable to insure the validity of (4.129) for all n by imposing it as a condition on the initial value A_0. Any initial value A_0 derived from any set of single-equation least-squares estimates in which one of the variables y_i, $i = 1, \ldots, K_y$, is selected as "dependent" variable for each value of g, satisfies that condition. Since (4.129) also holds for A, a similar condition is satisfied by $\bar{A}_n = \tilde{A}_n H$. Let the submatrices of H'^{-1} partitioned similarly to (4.93) be denoted by subscript combinations yy, yz, etc. Then, since $(H'^{-1})_{yz} = 0$ and $(H'^{-1})_{zz}$ is nonsingular, substitution of (4.94) in (4.129) shows that the latter condition secures the identical vanishing of $\mathcal{Z}\tilde{C}_n$ rather than its general reduction in successive iterations. It was already recognized in section 4.1.5 in connection with $\mathcal{P}_1$ that this cuts down the number of unknowns that need to be determined in each iteration. The unknowns participating in the iterations are the unrestricted elements of $\mathcal{Z}B_n$ corresponding, through $\bar{B}_n = \tilde{B}_n B$, to the first three components of $\tilde{A}_n$ in the decomposition (4.116).

4.3.4.3. Arrangement of computations under single-parameter restrictions. Because the restrictions imposed are of the single-parameter type, the notation α^g simply indicates that the elements in $\alpha(g)$ prescribed to be zero are deleted. Solving for $\mathcal{Z}\Delta C_n$ and $\mathcal{Z}C_n$ from (4.129) and substituting in (4.127), we have as the defi-

nition of $\mathfrak{D}\Delta B_0$ in that notation,

$$\Delta b_0^g \cdot {}^z M_{yy}^g = \iota(g) \cdot B_0'^{-1} \cdot \Phi'^g - b_0^g \cdot {}^z M_{yy}^g , \tag{4.130}$$

from which we solve for Δb_0^g for computational purposes,

$$\Delta b_0^g = \iota(g)\, B_0'^{-1}\, \Phi'^g \,({}^z M_{yy}^g)^{-1} - b_0^g . \tag{4.131}$$

If $\mathfrak{P}_{h_n}$ is applied, the value (4.69) of h is obtained most conveniently from

$$h_0 = \frac{\sum_g \Delta b_0^g \cdot {}^z M_{yy}^g \cdot \Delta b_0'^g}{\mathrm{tr}\{(B_0'^{-1} \cdot \Delta B_0'^{-1})^2 + \sum_g \Delta b_0^g \cdot {}^z M_{yy}^g \cdot \Delta b_0'^g\}} \tag{4.132}$$

in which the first term in the denominator is not transformed to vector coordinates.

First the K_y matrices ${}^z M_{yy}^g$ and their inverses are prepared from M_{xx} and the restrictions. Then an initial value A_0 (a set of vectors b_0^g) is selected. Each row of $({}^z M_{yy}^g)^{-1}$ represents a least-squares regression as a possible choice of initial vector b_0^g. The initial vectors are then normalized by

$$b_0^g \cdot {}^z M_{yy}^g \cdot b_0'^g = 1, \tag{4.133}$$

and $B_0'^{-1}$ is computed by any suitable method of inversion. Then either Δb_0^g or $b_0^g + \Delta b_0^g$ is determined from (4.131), from which b_1^g is obtained, in the case of $\mathfrak{P}_{h_n}$ with the help of (4.132), and the next iteration can proceed.

If the process is terminated after the nth iteration, $\mathfrak{D}C_n$ is obtained from $\mathfrak{D}B_n$ by (4.5) or

$$c_n^g = -b_n^g \cdot M_{yz}^g \cdot (M_{zz}^g)^{-1} . \tag{4.134}$$

4.3.4.4. Application of $\mathfrak{P}_1$, $\mathfrak{P}_{3/4}$, $\mathfrak{P}_{h_n}$ to the constructed example. Table 4.3.4.4 gives the results of applying $\mathfrak{P}_1$, $\mathfrak{P}_{3/4}$,

TABLE 4.3.4.4. Numerical illustration of $\mathcal{P}_1$,

Method	Iteration n =	Ratios of elements of $b_n(g)$			Scales: $m\{a(g)\} \equiv \{a_n(g)\ M_{xx}\ a_n'(g)\}^{1/2}$		
		$-\frac{b_{12}^n}{b_{13}^n}$	$-\frac{b_{21}^n}{b_{23}^n}$	$-\frac{b_{32}^n}{b_{31}^n}$	$m\{a_1(g)\}$	$m\{a_2(g)\}$	$m\{a_3(g)\}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\mathcal{P}_1$	0	0.25000	0.22960	0.42887	1.00000	0.99805	1.00000
	1	0.23673	0.32274	0.46075	1.00787	1.05568	1.01041
	2	0.23277	0.33141	0.47046	0.99301	0.94762	0:99060
	3	0.23230	0.33408	0.47160	1.00709	1.05530	1.00858
	4	0.23219	0.33440	0.47187	0.99292	0.94757	0.99087
	5	0.23217	0.33448	0.47191	1.00716	1.05368	1.00947
	6	0.23217	0.33448	0.47192	0.99286	0.94758	0.99091
$\mathcal{P}_{3/4}$	0	0.25000	0.22960	0.42887	1.00000	0.99805	1.00000
	1	0.23991	0.29946	0.45275	1.00443	1.03172	1.00592
	2	0.23524	0.32130	0.46427	0.99884	0.98748	0.99837
	3	0.23336	0.32965	0.46895	1.00073	1.00677	1.00105
	4	0.23263	0.33265	0.47078	0.99639	0.99669	0.99950
	5	0.23235	0.33379	0.47148	1.00021	1.00025	1.00025

$\mathcal{P}_{3/4}$, and $\mathcal{P}_{h_n}$ when Σ is required to be diagonal.

$\mathcal{P}_{h_n}$							
h_n = 1.06540	0	0.25000	0.22961	0.42887	1.00000	0.99805	1.00000
0.57158	1	0.23591	0.32884	0.46283	1.00902	1.06286	1.01181
0.87201	2	0.23399	0.33053	0.46747	0.99893	0.99330	0.99860
0.59751	3	0.23259	0.33291	0.47089	1.00088	1.00521	1.00118
0.85300	4	0.23238	0.33368	0.47140	0.99983	0.99898	0.99971
0.60588	5	0.23223	0.33434	0.47176	1.00011	1.00075	1.00016
	6	0.23220	0.33436	0.47184	0.99995	0.99986	0.99997
Newton	0	0.25000	0.22960	0.42887	1.00000	0.99805	1.00000
	1	0.23197	0.33566	0.47251	1.00283	1.00458	1.00542
	2	0.23224	0.33463	0.47192	0.99924	0.99989	1.00044
	3	0.23218	0.33432	0.47185	0.99988	1.00022	1.00032
Solutions		$\frac{b_{12}}{b_{13}}$	$-\frac{b_{21}}{b_{23}}$	$-\frac{b_{32}}{b_{31}}$	$\frac{b_{14}}{b_{13}}$	$-\frac{b_{25}}{b_{23}}$	$-\frac{b_{36}}{b_{31}}$
Σ diagonal (from $\mathcal{P}_1$ above)		0.23217	0.33448	0.47192	0.23931	0.33796	0.48876
Σ nondiagonal (true values)		0.25000	0.33333	0.50000	0.25000	0.33333	0.50000

and $\mathcal{P}_{h_n}$ respectively in the examples given. It also gives the application, to the same data, of the Newton method, to be discussed below. Initial values were the same in all cases and were based on the rows of $({}^{z}M^{g}_{yy})^{-1}$ numbered 3, 1, and 2 for $g = 1, 2, 3$, respectively.

Comparison of the three methods shows that in the present case $\mathcal{P}_1$ is superior to $\mathcal{P}_{3/4}$, and even slightly better than $\mathcal{P}_{h_n}$, with respect to the most essential property: the speed of convergence of the ratios of elements of each $a_n(g)$. $\mathcal{P}_1$ shows the characteristic alternation in successive values $m\{a_n(g)\}$, $n = 0, 1, \ldots$. These values converge gradually in $\mathcal{P}_{3/4}$ and slightly faster in $\mathcal{P}_{h_n}$. The wide variation in successive values of h_n is remarkable. The apparent alternation of these values is peculiar to the present example. Results substantially similar to those shown here were obtained in another constructed example of only two equations, but the sequence of values h_n was found to be more irregular in that case.[1]

*4.3.4.5. *Arrangement of computations under more general restrictions.* Under the simple type of restrictions imposed in the foregoing example the vectors $a(g)$ differ from the corresponding vectors a^g_n only in that they contain in addition certain vanishing elements. With more general basic matrices Φ^g it is necessary to decide whether the quantities a^g_n are actually to be computed as a separate step in each iteration. We shall show that this can be avoided, with a resulting saving of computational labor.

Using successively (4.40), the second relation of (4.30), (4.37), and (4.72), we can rewrite the definition (4.57) of ΔA_0 in the form

$$\Delta a_0^* M^{**} = \operatorname{vec}[B_0'^{-1} \quad 0]\,\Phi'^* - a_0^* M^{**}. \tag{4.135}$$

Even with orthogonalization of Φ^*, which we do not now require, the nonsingular matrix $(\Phi^* \Phi'^*)^{-1}$ appearing in (4.40) can be and has been omitted from (4.135) because it appears originally as postmultiplicand to all terms.

[1] This example was discussed as "case I" in another article by one of the present authors [Koopmans, 1945]. Initial values were the least-squares regression with x_1 as dependent variable in both structural equations. Successive values of h_n were 0.394, 0.839, 1.133, 0.582, 0.795, ...

From (4.135) we derive

$$\Delta a_0^* = \mathrm{vec}[B_0'^{-1} \quad 0]\, \Phi'^{*}\, (M^{**})^{-1} - a_0^*, \tag{4.136}$$

to which we again apply (4.37) to obtain

$$\Delta a_0 = \mathrm{vec}[B_0'^{-1} \quad 0]\, \Phi'^{*}\, (M^{**})^{-1}\, \Phi^{*} - a_0. \tag{4.137}$$

Defining

$$P^{(g)} \equiv \Phi'^{g}\, (\Phi^{g}\, M\, \Phi'^{g})^{-1}\, \Phi^{g} \equiv P_{xx}^{(g)} \equiv \begin{bmatrix} P_{yy}^{(g)} & P_{yz}^{(g)} \\ P_{zy}^{(g)} & P_{zz}^{(g)} \end{bmatrix}, \tag{4.138}$$

we have

$$P \equiv \Phi'^{*}\, (M^{**})^{-1}\, \Phi^{*} = \begin{bmatrix} P^{(1)} & 0 & \cdots & 0 \\ 0 & P^{(2)} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & P^{(G)} \end{bmatrix}. \tag{4.139}$$

Therefore, (4.137) is equivalent to

$$\Delta a_0(g) = \iota(g)\, B_0'^{-1}\, P_{yx}^{(g)} - a_0(g), \qquad g = 1, \ldots, G. \tag{4.140}$$

The preparations for any of the iterative procedures based on (4.140) now consist in the evaluation of the G matrices $M^g = \Phi^g\, M\, \Phi'^g$ from (4.72), their inversion, and the transformation of the inverses $(M^g)^{-1}$ back to the x-coordinates by (4.138), to obtain the $P_{yx}^{(g)}$.

*4.3.4.6. *Canonical form of the basic matrices* Φ^g. In many cases a further saving, similar to that obtained under simpler restrictions, can be secured by choosing an initial value satisfying (4.129). It will be noted that $P_{zz}^{(g)}$ does not occur in (4.140), and that $P_{yz}^{(g)}$ is not needed to obtain B by an iterative process based on (4.140). This suggests that the inversion of M^{**} can profitably be replaced by successive inversions of lower-order matrices.

In order to obtain the full benefit of this consideration, we must analyze the basic matrices Φ^g with the help of the following lemma:

LEMMA 4.3.4.6. *If $\bar{\Phi}^g$ is such that its number of rows equals its rank,*

$$r(\bar{\Phi}^g) = \rho(\bar{\Phi}^g) = Q_g, \tag{4.141}$$

and if Q_g^{I}, Q_g^{II}, Q_g^{III} are defined by

$$Q_g \equiv \rho(\bar{X}^g) + Q_g^{\mathrm{III}} \equiv \rho(\bar{\Psi}^g) + Q_g^{\mathrm{I}} \equiv Q_g^{\mathrm{I}} + Q_g^{\mathrm{II}} + Q_g^{\mathrm{III}}, \tag{4.142}$$

then there exists a nonsingular transformation matrix[1] Ω *such that*

$$\Omega\,\bar{\Phi}^g = \Phi^g = [X^g \quad \Psi^g] = \begin{bmatrix} X_{\mathrm{I}}^g & 0 \\ X_{\mathrm{II}}^g & \Psi_{\mathrm{II}}^g \\ 0 & \Psi_{\mathrm{III}}^g \end{bmatrix}, \tag{4.143}$$

with

$$\begin{aligned} \rho(X_{\mathrm{I}}^g) &= r(X_{\mathrm{I}}^g) = Q_g^{\mathrm{I}}, \\ \rho(\Psi_{\mathrm{III}}^g) &= r(\Psi_{\mathrm{III}}^g) = Q_g^{\mathrm{III}}, \\ \rho(\Phi_{\mathrm{II}}^g) &= r(\Phi_{\mathrm{II}}^g) = Q_g^{\mathrm{II}}, \end{aligned} \tag{4.144}$$

and there exists no nonsingular transformation matrix Ω such that in (4.143) X_{I}^g or Ψ_{III}^g or both have higher ranks than given by (4.144).

Proof: Owing to (4.142) there exist matrices Ω_{I}, Ω_{III} such that

$$\rho(\Omega_{\mathrm{I}}) = r(\Omega_{\mathrm{I}}) = Q_g^{\mathrm{I}}, \qquad \rho(\Omega_{\mathrm{III}}) = r(\Omega_{\mathrm{III}}) = Q_g^{\mathrm{III}}, \tag{4.145}$$

[1] For brevity no subscript g is appended to Ω.

and

$$\Omega_{III}\,\bar{X}^{g} = 0, \qquad \Omega_{I}\,\bar{\Psi}^{g} = 0. \tag{4.146}$$

We must have

$$\rho\begin{pmatrix}\Omega_{I}\\ \Omega_{III}\end{pmatrix} = \rho(\Omega_{I}) + \rho(\Omega_{III}) \tag{4.147}$$

because, if the right-hand member in (4.147) exceeded the left-hand member, there would, according to Lemma 2.3.2, exist nonvanishing vectors $\lambda_{I}, -\lambda_{III}$ such that

$$\mu = \lambda_{I}\,\Omega_{I} = \lambda_{III}\,\Omega_{III} \neq 0. \tag{4.148}$$

In that case (4.146) would imply

$$\mu[\bar{X}^{g} \quad \bar{\Psi}^{g}] = \mu\,\bar{\Phi}^{g} = 0, \qquad \mu \neq 0, \tag{4.149}$$

contrary to the assumption (4.141) about the rank of $\bar{\Phi}^{g}$.

Owing to (4.147) we can find a matrix Ω_{II} with

$$\rho(\Omega_{II}) = r(\Omega_{II}) = Q_{g}^{II} \geq 0 \tag{4.150}$$

such that the matrix Ω defined by

$$\Omega \equiv \begin{bmatrix}\Omega_{I}\\ \Omega_{II}\\ \Omega_{III}\end{bmatrix} \tag{4.151}$$

is nonsingular (rank Q_{g}). This matrix Ω produces the partitioning required in (4.143) with submatrices having the number of rows required in (4.144). Since Φ^{g} has the same rank (4.141) as $\bar{\Phi}^{g}$, the ranks in (4.144) cannot fall below the corresponding number of rows.

Finally, if a nonsingular matrix Ω existed such that the rank and hence the number of rows of X_{I}^{g}, say, exceeded Q_{g}^{I}, we should have, using successively (4.141), (4.143), and (4.142),

$$(4.152) \qquad Q_g - Q_g^{\mathrm{I}} > \rho(\Phi^g) - \rho(X_{\mathrm{I}}^g) = \rho(\Psi^g) = Q_g - Q_g^{\mathrm{I}},$$

an obvious contradiction. This completes the proof of Lemma 4.3.4.6. The form (4.143) of Φ^g will be called its canonical form. It is worth stressing that this form neither requires nor precludes orthogonalization of Φ^g according to (4.24).

*4.3.4.7. *Elimination of the unknowns* $a_{n,\,\mathrm{III}}^g$. Let us now assume that the basic matrices Φ^g are already in the canonical form. Then the expression (4.22) for $\alpha(g)$ in terms of a set of unrestricted parameters partitions as follows:

$$(4.153) \qquad \begin{aligned} \beta(g) &= \alpha_{\mathrm{I}}^g X_{\mathrm{I}}^g + \alpha_{\mathrm{II}}^g X_{\mathrm{II}}^g, \\ \gamma(g) &= \alpha_{\mathrm{II}}^g \Psi_{\mathrm{II}}^g + \alpha_{\mathrm{III}}^g \Psi_{\mathrm{III}}^g. \end{aligned}$$

Thus the parameters α_{III}^g do not enter into the Jacobian B. Iterative processes involving B_n only can therefore be constructed on the basis of the parameters

$$(4.154) \qquad {}_{\mathrm{III}}\alpha^g \equiv [\alpha_{\mathrm{I}}^g \quad \alpha_{\mathrm{II}}^g]$$

alone. With each approximation ${}_{\mathrm{III}}a_n^g$ to ${}_{\mathrm{III}}\alpha^g$, there are associated "silent" values $a_{n,\,\mathrm{III}}^g$ which are those linear functions of ${}_{\mathrm{III}}a_n^g$ that maximize the likelihood function with ${}_{\mathrm{III}}a_n^g$ inserted for ${}_{\mathrm{III}}\alpha^g$. Only at the termination of iterations do these values need to be determined explicitly or implicitly. In the computational arrangement of the Newton method discussed below, an explicit determination from equation (4.185) involves no extra cost. Since we have for the present decided against explicit evaluation of any part of a_n^g, we shall operate equivalently from (4.140) and (4.138) on the basis of properties of inverses of partitioned matrices.

We define

$$(4.155) \qquad (M^g)^{-1} \equiv N^g \equiv \begin{bmatrix} N^g_{\mathrm{I\,I}} & N^g_{\mathrm{I\,II}} & N^g_{\mathrm{I\,III}} \\ N^g_{\mathrm{II\,I}} & N^g_{\mathrm{II\,II}} & N^g_{\mathrm{II\,III}} \\ N^g_{\mathrm{III\,I}} & N^g_{\mathrm{III\,II}} & N^g_{\mathrm{III\,III}} \end{bmatrix}.$$

As before, the postsubscripts I, II can be subsumed in the presubscript III, the postsubscripts II, III in the presubscript I. For the iterations in (4.140) we need only

$$(4.156) \qquad P^g_{yy} = X'^g N^g X^g = {}_{\mathrm{III}}X'^g_{\cdot\,\mathrm{III\,III}}N^g_{\cdot\,\mathrm{III}}X^g,$$

where ${}_{\mathrm{III\,III}}N^g$ is obtained from N^g by deleting the rows and columns intersecting in $N^g_{\mathrm{III\,III}}$. If the ultimate evaluation of C_n is based on (4.140), we need in addition

$$(4.157) \qquad P^g_{yz} = X'^g N^g \Psi^g = {}_{\mathrm{III}}X'^g_{\cdot\,\mathrm{III\,I}}N^g_{\cdot\,\mathrm{I}}\Psi^g.$$

Thus $N^g_{\mathrm{III\,III}}$ is not needed. ${}_{\mathrm{III\,III}}N^g$ is computed from[1]

$$(4.158) \qquad {}_{\mathrm{III\,III}}N^g = \{{}_{\mathrm{III\,III}}M^g - {}_{\mathrm{III}}M^g_{\mathrm{III}}(M^g_{\mathrm{III\,III}})^{-1}\,{}_{\mathrm{III}}M'^g_{\mathrm{III}}\}^{-1}$$

and the submatrix ${}_{\mathrm{III}}N^g_{\mathrm{III}}$ needed in addition for (4.157) is obtainable from[1]

$$(4.159) \qquad {}_{\mathrm{III}}N^g_{\mathrm{III}} = -{}_{\mathrm{III\,III}}N^g_{\cdot\,\mathrm{III}}M^g_{\mathrm{III}}\cdot(M^g_{\mathrm{III\,III}})^{-1},$$

using once more a matrix ${}_{\mathrm{III}}M^g_{\mathrm{III}}(M^g_{\mathrm{III\,III}})^{-1}$ already computed for (4.158).

These formulae show that the most important saving from the use of the canonical form (4.143) – avoiding the calculation of $N^g_{\mathrm{III\,III}}$ – is due to the separation of Φ^g_{III} from ${}_{\mathrm{III}}\Phi^g$. The further separation of Φ^g_{I} from Φ^g_{II} leads to a minor additional saving by reducing the

[1] [Hotelling, 1943-A, p.4].

number of elements involved in the second matrix multiplication in (4.157).

In general the ranks of P^g_{yy} and P^g_{yz} are lower than the maximum compatible with the number of rows and columns. This expresses the fact that the elements of the vectors $a_n(g)$ depend linearly on a smaller number of parameters a^g_n.

4.3.4.8. The final evaluation of the a^g_{III}. The return from B_n to C_n on the basis of (4.140) requires in addition the inversion of B_n, which then serves for all values of g,

$$c_n(g) = \iota(g)\, B_0'^{-1}\, P^g_{yz}. \tag{4.160}$$

Depending on the circumstances, an alternative formula for $c_n(g)$ may be more economical. This is based on the expression for the "silent" values,

$$a^g_{III} = -\,{}_{III}a^g_{\cdot\,III}M^g_{III}\cdot(M^g_{III\;III})^{-1}, \tag{4.161}$$

which is the equivalent of (4.134) under the present form of the Φ^g. From (4.22) and (4.143) we have

$$b_n(g) = {}_{III}a^g_{n\cdot\,III}X^g, \qquad c_n(g) = a^g_{n,II}\cdot\Psi^g_{II} + a^g_{n,III}\cdot\Psi^g_{III}. \tag{4.162}$$

The first of these relations is solved for ${}_{III}a^g_n$ by

$$ {}_{III}a^g_n = b_n(g)\,\Xi, \tag{4.163}$$

where the relation

$$ {}_{III}X^g\cdot\Xi = {}_{III\;III}I \tag{4.164}$$

defines all that needs to be defined concerning Ξ. The condition (4.164) may in simple cases offer more ready ways of finding a suitable value of Ξ than the explicit calculation of the particular solution

$$\Xi = {}_{III}X'^g\,({}_{III}X^g\cdot{}_{III}X'^g)^{-1}. \tag{4.165}$$

Combining the second relation (4.162) with (4.161) and (4.163), we have

$$
\begin{aligned}
c_n(g) &= b_n(g) \equiv \{ {}_{\mathrm{III}}\Psi^g - {}_{\mathrm{III}}M^g_{\mathrm{III}}\,(M^g_{\mathrm{III}\,\mathrm{III}})^{-1}\,\Psi^g_{\mathrm{III}} \} \\
&\equiv b_n(g)\, J(g)\,,
\end{aligned}
\tag{4.166}
$$

say. The application of this formula requires the evaluation of G matrices $J(g)$, involving in principle G inversions (4.165) of orders equal to the respective values of $(Q^{\mathrm{I}}_g + Q^{\mathrm{II}}_g)$.

4.3.4.9. A formula for the computation of h_n. In the application of $\mathfrak{P}_{h_n}$, the formula (4.166) is preferable to (4.160) because it also holds if $a_n(g)$ is replaced by the scalar multiple $\Delta a_n(g)$ of the difference between two successive approximations. It can therefore be used to derive from (4.69) the formula

$$
h_0 = \frac{\sum_{g=1}^{G} \Delta b_0(g) \cdot M_{yy}(g) \cdot \Delta b_0'(g)}{\mathrm{tr}(B_0'^{-1} \cdot \Delta B_0') + \sum_{g=1}^{G} \Delta b_0(g) \cdot M_{yy}(g) \cdot \Delta b_0'(g)}
\tag{4.167}
$$

for the evaluation of h_0, in which

$$
M_{yy}(g) = M_{yy} + J(g)\, M_{zy} + M_{yz} J'(g) + J(g)\, M_{zz} J'(g)\,.
\tag{4.168}
$$

4.3.5. *The Newton method*

4.3.5.1. The principle of the Newton method. Unlike the methods discussed so far, the principle of the Newton method has no connection with the particular form of the likelihood function. Its application to our problem proceeds as follows. If we write

$$
A_1 \equiv A_0 + \Delta A_0\,,
\tag{4.169}
$$

the Taylor expansion (4.68) of the likelihood function $L(A_1)$ in terms of ΔA_0 can be written

$$L(A_1) - L(A_0) = \operatorname{tr}\{(B_0'^{-1} I_{[K_y K_x]} - A_0 M_{xx}) \Delta A_0'\}$$

(4.170)

$$+ \frac{1}{2} \operatorname{tr}\{-(B_0'^{-1} \Delta B_0')^2 - \Delta A_0 M_{xx} \Delta A_0'\} + \cdots .$$

This expansion, transformed to the unrestricted parameters Δa_0^* defined as in (4.37), may be denoted

$$L(a_1^*) - L(a_0^*) = l_0^* \Delta a_0'^* + \frac{1}{2} \Delta a_0^* L_0^* \Delta a_0'^* + \cdots . \tag{4.171}$$

The vector l_0^* and symmetric matrix L_0^* so defined depend, of course, on the initial value A_0, as distinct from the vector $l^* = 0$ and the matrix L^* defined by (4.77) on the basis of the solution A.

The Newton method determines Δa_0^* from the requirement that the first two terms in the expansion

$$\frac{d L(a_1^*)}{d \Delta a_0^*} = l_0^* + \Delta a_0^* L_0^* + \cdots \tag{4.172}$$

shall cancel out. This leads to

$$a_1^* L_0^* = (a_0^* + \Delta a_0^*) L_0^* = a_0^* L_0^* - l_0^* \tag{4.173}$$

as the formula defining a_1^*.

4.3.5.2. Comparisons between the Newton method and $\mathfrak{P}_h$, $\mathfrak{P}_{h_n}$.

The following differences between this method and the procedures based on the earlier choice (4.57) of ΔA_0 deserve discussion.

The Newton method seeks any stationary point of the likelihood function to which the initial value A_0 is sufficiently near. The earlier methods converge only to maxima. This establishes a presumption that the Newton method requires for convergence a closer proximity of the initial value to the maximum sought. It also means that after a stationary point has been obtained, second-order conditions must now be investigated to determine whether the point found is actually a maximum. Saddle points, maxima, and minima can

be distinguished through indefiniteness, negative definiteness, and positive definiteness, respectively, of L^*.

The Newton method breaks down in the case of incomplete identification because then L^* is singular, and hence cannot be inverted.

If A_0 is sufficiently close to the desired maximum A, the speed of convergence per iteration in the Newton method is superior. Writing again $a_n^* = a^* + \bar{a}_n^*$, it is seen as follows that $\bar{a}_1^*$ is quadratic in $\bar{a}_0^*$ – a property which was found to be present in the earlier methods only in the rare case that all relevant characteristic values of $L^\dagger$ coincide.

From the expansion of the likelihood function $L(\alpha^*)$ with respect to $\bar{\alpha}^* = \alpha^* - a^*$,

$$L(\alpha^*) - L(a^*) = \frac{1}{2}\,\bar{\alpha}^*\,L^*\,\bar{\alpha}'^* + \cdots, \tag{4.174}$$

we have

$$l_0^* \equiv \left(\frac{d\,L(\alpha^*)}{d\bar{\alpha}^*}\right)_{\bar{\alpha}^* = \bar{a}_0^*} = \bar{a}_0^*\,L^* + \cdots. \tag{4.175}$$

From (4.173) and (4.175) we have the relation

$$\bar{a}_1^* = \bar{a}_0^* - l_0^*\,(L_0^*)^{-1} = \bar{a}_0^*\,\{\,I - L^*\,(L_0^*)^{-1} + \cdots\,\}, \tag{4.176}$$

from which it is easily seen, by expanding L_0^* in terms of $\bar{a}_0^*$, that the first-order term in $\bar{a}_0^*$ in this expression vanishes, and that the quadratic term in $\bar{a}_0^*$ depends on the third derivatives of the likelihood function in the point a^*.

Against the superior speed of convergence per iteration in the Newton method must be set the greatly increased computational labor per iteration. Disregarding for a moment the saving arising in both methods from the canonical form of the basic matrices, the Newton method requires for each iteration afresh the calculation of the matrix L_n^* and the solution of the linear equations (4.173) in Q unknowns a_1^*. In contrast, the matrix M^{**} occurring in the left-hand member $\Delta a_0^* M^{**}$ of (4.57) remains the same for all iterations, so that its inversion paves the way for evaluation of successive values Δa_n^* by matrix multiplication only. Finally, under

the type of restrictions here considered, M^{**} partitions into diagonal blocks M^g, and its inversion therefore requires only the inversion of G matrices of orders $Q_1, \ldots, Q_G$. The matrix L^* is not similarly partitionable, although it has other regularities of which a smaller advantage could probably be taken.

*4.3.5.3. *Computational procedure in the Newton method.* Since the matrix L_n^* depends on n, there is no incentive in the Newton method to avoid the explicit appearance of the vectors a_n^* in the computations. We shall therefore develop the formulae largely in terms of *-coordinates.

As in the other methods, there is a possibility of saving computational work whenever certain elements of α^* do not enter the Jacobian B. Assume, therefore, that Φ^* is in canonical form. There will be a further computational advantage in assuming that at least the submatrices Φ_{I}^g, Φ_{II}^g, Φ_{III}^g are made mutually orthogonal by suitable choice of the Ω in (4.143). To simplify the formulae, we shall assume row-by-row orthogonality

$$\Phi^* \Phi'^* = I \tag{4.177}$$

of Φ^*, although in actual computations it need not be economical to go that far.

We shall now relate l_0^* and L_0^* in (4.173) to the initial vector a_0^*. Using (4.177), we evaluate l_0^* from (4.46) and (4.49) as

$$l_0^* \equiv \left(\frac{d\,L(\alpha^*)}{d\,\alpha^*}\right)_{\alpha^* = a_0^*} = \mathrm{vec}^*\,(B_0'^{-1}\, I_{[K_y K_x]} - A_0\, M_{xx}). \tag{4.178}$$

On the other hand we have, for any $\mathrm{A} = \mathrm{mat}*\,\alpha^*$, from the definition (4.171) of L_0^*,

$$\begin{aligned} \alpha^*\, L_0 &= \frac{1}{2}\frac{d}{d\alpha^*}(\alpha^*\, L_0^*\, \alpha'^*) \\ &= \mathrm{vec}^*(-B_0'^{-1}\, \mathrm{B}'\, B_0'^{-1}\, I_{[K_y K_x]} - \mathrm{A}\, M_{xx}), \end{aligned} \tag{4.179}$$

and, hence, in particular,

$$a_0^* L_0^* = \mathrm{vec}^*(-B_0'^{-1} I_{[K_y K_x]} - A_0 M_{xx}). \tag{4.180}$$

When (4.178) and (4.179) are inserted in (4.173), the terms containing $A_0 M_{xx}$ cancel. Using (4.179) again with a_1^* substituted for α^*, we see that (4.173) is equivalent to

$$\begin{aligned} -a_1^* L_0^* &\equiv \mathrm{vec}^*(B_0'^{-1} B_1' B_0'^{-1} I_{[K_y K_x]} + A_1 M_{xx}) \\ &= 2\,\mathrm{vec}^*(B_0'^{-1} I_{[K_y K_x]}). \end{aligned} \tag{4.181}$$

Computation can conveniently be based on this formulation of the Newton method or on an equivalent formulation in terms of ΔA_0 instead of A_1.

We shall use postsubscripts and presubscripts I, II, III to denote submatrices of Φ^*, L_n^*, and subvectors of α^*, etc., corresponding to the canonical form of Φ^*. For instance

$$_{\mathrm{III}}X_{\mathrm{I}}^* \equiv \begin{bmatrix} {}_{\mathrm{III}}X^1 & 0 & \cdots & 0 \\ 0 & {}_{\mathrm{III}}X^2 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & {}_{\mathrm{III}}X^G \end{bmatrix}. \tag{4.182}$$

It is seen from (4.181) [or from the definition (4.171) of L_0^*] that

$$L_0^* = \begin{bmatrix} {}_{\mathrm{III}\,\mathrm{III}}L_0^* & -{}_{\mathrm{III}}M_{\mathrm{III}}^{**} \\ -{}_{\mathrm{III}}M_{\mathrm{III}}'^{**} & -M_{\mathrm{III}\,\mathrm{III}}^{**} \end{bmatrix}, \tag{4.183}$$

because the first term in the second member of (4.181) does not give rise to any III-components of L_0^*. Similarly, the last member in (4.181) has vanishing III-components. It follows that

$$(4.184)\qquad a_1^* \begin{bmatrix} -{}_{\mathrm{III}}M^{**}_{\mathrm{III}} \\ \\ -M^{**}_{\mathrm{III}\,\mathrm{III}} \end{bmatrix} = -{}_{\mathrm{III}}a_1^* \cdot {}_{\mathrm{III}}M^{**}_{\mathrm{III}} - a^*_{1,\mathrm{III}} \cdot M^{**}_{\mathrm{III}\,\mathrm{III}} = 0,$$

from which $a^*_{1,\mathrm{III}}$ can be solved according to (4.161) and inserted in (4.181) to give

$$(4.185)\qquad {}_{\mathrm{III}}a_1^* \cdot {}^{\mathrm{III}}L_0^* = {}_{\mathrm{III}}\mathrm{vec}^* \begin{bmatrix} B_0'^{-1} & 0 \end{bmatrix},$$

with

$$(4.186)\qquad {}^{\mathrm{III}}L_0^* \equiv {}_{\mathrm{III}\,\mathrm{III}}L_0^* + {}_{\mathrm{III}}M^{**}_{\mathrm{III}} \cdot (M^{**}_{\mathrm{III}\,\mathrm{III}})^{-1} \cdot {}_{\mathrm{III}}M'^{**}_{\mathrm{III}}.$$

The following steps arise in the application of (4.185). Initial "overhead" work consists in orthogonalization of Φ^*, determination of M^{**} from (4.72), and calculation of the last term in (4.186), which remains the same in all iterations. The inversion of $M^{**}_{\mathrm{III}\,\mathrm{III}}$ can be carried out for each of its diagonal blocks $M^g_{\mathrm{III}\,\mathrm{III}}$ separately. Then one chooses B_0, calculates its inverse and uses it in calculating both

$$(4.187)\qquad {}_{\mathrm{III}}\mathrm{vec}^* \begin{bmatrix} B_0'^{-1} & 0 \end{bmatrix} = \mathrm{vec}\begin{bmatrix} B_0'^{-1} & 0 \end{bmatrix} \cdot {}_{\mathrm{III}}\Phi'^*$$

and

$$(4.188)\qquad {}_{\mathrm{III}\,\mathrm{III}}L_0^* = -{}_{\mathrm{III}}\mathrm{X}^* \cdot K_0 \cdot {}_{\mathrm{III}}\mathrm{X}'^* - {}_{\mathrm{III}\,\mathrm{III}}M^{**}.$$

The matrix K_0 is defined by

$$(4.189)\qquad \mathrm{tr}(B_0'^{-1}\, B')^2 \equiv {}_{\mathrm{III}}a^* \cdot {}_{\mathrm{III}}\mathrm{X}^* \cdot K_0 \cdot {}_{\mathrm{III}}\mathrm{X}'^* \cdot {}_{\mathrm{III}}a'^*,$$

and has as elements

$$(4.190)\qquad k_{gi,hj} = \iota(i) \cdot B_0^{-1} \cdot \iota'(h) \cdot \iota(j) \cdot B_0^{-1} \cdot \iota'(g),$$

arranged (with $G = K_y$) according to

$$
(4.191)\qquad
\begin{bmatrix}
k_{11,11} & \cdots & k_{11,1G} & \cdots & k_{11,G1} & \cdots & k_{11,GG} \\
\cdot & \cdots & \cdot & \cdots & \cdot & \cdots & \cdot \\
k_{1G,11} & \cdots & k_{1G,1G} & \cdots & k_{1G,G1} & \cdots & k_{1G,GG} \\
\cdot & \cdots & \cdot & \cdots & \cdot & \cdots & \cdot \\
k_{G1,11} & \cdots & k_{G1,1G} & \cdots & k_{G1,G1} & \cdots & k_{G1,GG} \\
\cdot & \cdots & \cdot & \cdots & \cdot & \cdots & \cdot \\
k_{GG,11} & \cdots & k_{GG,1G} & \cdots & k_{GG,G1} & \cdots & k_{GG,GG}
\end{bmatrix} .
$$

Finally, ${}^{III}L_0^*$ is put together from (4.186), and ${}_{III}a_1^*$ solved from (4.185), leading to

$$
(4.192)\qquad B_1 = \mathrm{mat}\,({}_{III}a_1^* \cdot {}_{III}X^*) .
$$

At the termination of iterations, C_n is obtained from

$$
\begin{aligned}
C_n &= \mathrm{mat}\,({}_{III}a_n^* \cdot {}_{III}\Psi^* + a_{n\,III}^* \cdot \Psi_{III}^*) \\
&= \mathrm{mat}\,[{}_{III}a_n^* \{ {}_{III}\Psi^* - {}_{III}M_{III}^{**} \cdot (M_{III\,III}^{**})^{-1} \cdot \Psi_{III}^* \}] .
\end{aligned}
\qquad (4.193)
$$

It appears from (4.188) that the matrix ${}_{III\,III}L_0^*$ has considerable regularity in its make-up. The problem of how to best utilize those regularities for the inversion of ${}_{III\,III}L_0^*$ or for the solution of (4.185) has not been investigated by us.

4.3.5.4. Numerical illustration of the Newton method. This method has likewise been applied to the constructed example already discussed in which the basic matrices have the simple form (4.128). The superior speed of convergence of the Newton method comes out clearly in the results shown in Table 4.3.4.4. More experience with actual data is required to determine whether and in what circumstances the greater speed of convergence is adequate compensation for the greatly increased labor per iteration.

4.3.5.5. Estimated sampling variances and covariances of the

estimated coefficients a_{gk}. Even if another method is used to obtain a satisfactory approximation A_n to A, it is still advisable to make one final iteration with the Newton method in order to obtain the matrix of estimated sampling variances and covariances

$$\text{(4.194)} \qquad \begin{aligned} \operatorname{est} \mathcal{E}\{(a'^* - \alpha'^*)(a^* - \alpha^*)\} &= -\frac{1}{T}\left[\frac{\partial^2 L(\alpha^*)}{\partial \alpha'^* \, \partial \alpha^*}\right]^{-1}_{\alpha^* = a^*} \\ &= -\frac{1}{T} L^{*-1} \end{aligned}$$

of the estimated parameters a^* as a by-product. It was shown in Theorem 3.3.10 that the estimates (4.194) are consistent. A suitable method of obtaining L^{*-1} in the present circumstances is the partitioning method whereby ${}_{\text{III III}}(L^{*-1})$ is obtained as the inverse

$$\text{(4.195)} \qquad {}_{\text{III III}}(L^{*-1}) = ({}^{\text{III}}L^*)^{-1}$$

of ${}^{\text{III}}L^*$ as a step in solving ${}_{\text{III}}a_1^*$ from (4.185), and $(L^{*-1})_{\text{III III}}$ and ${}_{\text{III}}(L^{*-1})_{\text{III}}$ are found from similar formulae [Hotelling, 1943 A, p. 4], quoted and used before.

Because of the normalization rule (4.25) here employed, the sampling variances (4.194) are not in the form in which they are normally expressed. One will usually regard as final parameters the ratios

$$\text{(4.196)} \qquad \frac{\alpha^{g}{}_{\cdot \iota'}(q)}{\alpha^{g}{}_{\cdot \iota'}(1)}, \qquad q = 2, \ldots, Q_g, \qquad g = 1, \ldots, G,$$

of the elements of each α^g. Since the estimates (4.194) themselves are first-order approximations that become only asymptotically exact as the sample size T tends to infinity, sampling variances and covariances of the estimates $a^{g}{}_{\cdot \iota'}(q) \,/\, a^{g}{}_{\cdot \iota'}(1)$ of the parameters (4.196), of an equal order of approximation, can be found by Taylor expansions in which only terms linear in the estimates (4.194) are retained.

Alternatively, one may normalize on the $\alpha^{g}{}_{\cdot \iota'}(1)$ by (4.26) and treat the diagonal elements σ_{gg} of Σ as unknown parameters. This procedure may lead to a further saving in computational labor because the parameters σ_{gg} so introduced fall in the same category as

the parameters α^{g}_{III} : for any ${}_{\text{III}}a^{g}_{n}$ the corresponding maximizing values of α^{g}_{III} and σ_{gg} are easily found. The order of the inversion (4.195) can therefore be further reduced by the number G of parameters σ_{gg}. It is not necessary to go into the details of this procedure since the application of the normalization (4.26) will be demonstrated in section 4.4 in the case where Σ is entirely unrestricted.

From the estimated sampling variances and covariances of the $a^{g}\cdot\iota'(q)$ / $a^{g}\cdot\iota'(1)$ we may revert to the singular matrix of estimated sampling variances and covariances of the estimates a_{gk} of the structural coefficients α_{gk} through the transformations (4.22).

4.4. *The Case of Unrestricted Correlations between the Disturbances*

4.4.1. No restrictions on Σ. We shall now study the case in which no a priori restrictions are imposed on the matrix Σ of variances and covariances of the disturbances in the structural equations (1.1) except the symmetry and positive-definiteness conditions arising from its definition. The discussion can be brief in those aspects of the problem that are also found in the case of uncorrelated disturbances just discussed. The main emphasis will be on points of difference between the two cases.

4.4.2. Normalization. With the nondiagonal elements of Σ unrestricted in any case, it is not convenient to impose normalization through the diagonal elements σ_{gg} of Σ. We shall either impose no normalization at all or normalize through one element of each vector α^{g}, for which we may conveniently take the first element $\alpha^{g}\cdot\iota'(1) = 1$. In the latter case we shall employ the notation

$$\begin{aligned}\alpha^{*}_{[1]} &\equiv [\ \alpha^{1}\cdot\iota'(1) \quad \alpha^{2}\cdot\iota'(1) \quad \dots \quad \alpha^{G}\cdot\iota'(1)\ ] \\ &= [\ 1 \quad 1 \quad \dots \quad 1\ ]\end{aligned} \tag{4.197}$$

to express the normalization rule, and introduce similar notations

$$\Phi^{g} = \begin{bmatrix} {}^{g}\Phi_{[1]} \\ {}_{[1]}\Phi^{g} \end{bmatrix},$$

$$\Phi^{*}_{[1]} \equiv \begin{bmatrix} \Phi^{1}_{[1]} & 0 & \cdots & 0 \\ 0 & \Phi^{2}_{[1]} & \cdots & 0 \\ . & . & \cdots & . \\ 0 & 0 & \cdots & \Phi^{G}_{[1]} \end{bmatrix},$$

$$[1]\Phi^{*} \equiv \begin{bmatrix} {}_{[1]}\Phi^{1} & 0 & \cdots & 0 \\ 0 & {}_{[1]}\Phi^{2} & \cdots & 0 \\ . & . & \cdots & . \\ 0 & 0 & \cdots & {}_{[1]}\Phi^{G} \end{bmatrix},$$

(4.198)

for the corresponding partitioning of the basic matrices.

4.4.3. Elimination of the parameters Σ. We shall first maximize the likelihood function

$$L(\mathrm{A}, \Sigma) = \text{const} + \log \det \mathrm{B} - \frac{1}{2} \log \det \Sigma - \frac{1}{2} \operatorname{tr}(\Sigma^{-1} \mathrm{A} M_{xx} \mathrm{A}') \tag{4.199}$$

with respect to the unrestricted parameters Σ while the parameters A are kept constant. From (3.17) and (3.18) it is easily seen that the first derivatives $\partial L / \partial \sigma_{gh}$, $g, h = 1, \ldots, G$, vanish if

$$\Sigma = \hat{\Sigma} \equiv \mathrm{A} M_{xx} \mathrm{A}' = \hat{\Sigma}'. \tag{4.200}$$

The derivation of (4.200) must take account of the required symmetry of Σ but the result is not affected thereby. It follows from (3.35) that (4.200) indicates the unique and absolute maximum of the function (4.199) with respect to Σ. Upon inserting (4.200) in (4.199), we obtain

$$(4.201) \qquad L(\mathrm{A}) = \text{const} + \log\det \mathrm{B} - \frac{1}{2}\log\det(\mathrm{A}\, M_{xx}\, \mathrm{A}')$$

as the likelihood function after maximization with respect to Σ. This function is homogeneous of degree 0 in each row of A, i.e., it is invariant for changes in normalization of A through premultiplication with a nonsingular diagonal matrix Υ. This is easily verified directly or can be seen as a consequence of the invariance of (4.201) under the wider group of nonsingular transformations implied in Theorem 2.1.3.5.

The maximum of (4.201) in the absence of restrictions on A has already been studied in the analysis leading to Theorem 3.1.10.

4.4.4. *The maximum-likelihood equations.* We shall continue to use the symbol $\hat{\Sigma}$ as an abbreviation for the expression (4.200) in terms of A. Similarly, we shall use the abbreviations

$$(4.202) \qquad S_n = A_n M_{xx} A_n', \qquad S = A M_{xx} A'.$$

Again writing $\mathrm{A} = A_0 + \Delta A_0$, where A_0 is a trial value, we have, using (3.16) and (3.17),

$$\begin{aligned} L(\mathrm{A}) - L(A_0) &= \operatorname{tr}(B_0'^{-1}\Delta B_0') \\ &\quad - \frac{1}{2}\operatorname{tr}\{S_0^{-1}(A_0 M_{xx}\Delta A_0' + \Delta A_0 M_{xx} A_0')\} + \cdots \\ &= \operatorname{tr}\{B_0'^{-1} I_{[K_y K_x]} - S_0^{-1} A_0 M_{xx}\}\Delta A_0' + \cdots . \end{aligned} \tag{4.203}$$

The restrictions are

$$(4.204) \qquad \alpha = \alpha^* \Phi^*, \qquad a_0 = a_0^* \Phi^*, \qquad \operatorname{vec}\Delta A_0 = (\operatorname{vec}^*\Delta A_0)\Phi^* .$$

In the absence of normalizing restrictions on A, A_0 can coincide with a restricted maximum A of the likelihood function only if the linear term in (4.203) vanishes for all values of $\operatorname{vec}^*\Delta A_0$. The first-order maximum-likelihood conditions in this case are therefore, owing to Lemma 4.2.4,

$$\text{vec}^*\,(B'^{-1}\, I_{[K_y\, K_x]} - S^{-1} A\, M_{xx}) = 0\,, \tag{4.205}$$

with S again depending on A according to (4.202). As before, these conditions permit premultiplication of A by a diagonal matrix, through which we may satisfy the normalization (4.197) if desired.

4.4.5. The processes $\mathcal{P}_h$ and $\mathcal{P}_{h_n}$. For the generalization of these processes to the present case, we shall for the time being not impose a normalization rule on A. Given an initial value A_0, the following extension of the definition (4.57) of the direction matrix ΔA_0 seems natural as well as the simplest possible:

$$\text{vec}^*(S_0^{-1}\, \Delta A_0\, M_{xx}) = \text{vec}^*(B_0'^{-1}\, I_{[K_y\, K_x]} - S_0^{-1} A_0\, M_{xx})\,. \tag{4.206}$$

Comparison with (4.203) shows that a property similar to (4.58) in the previous case again holds: If $A_1 = A_0 + h\,\Delta A_0$, a sufficiently small value of h will always lead to $L(A_1) > L(A_0)$ if a stationary value of $L(A)$ is not already reached in A_0.

One can again choose a suitable constant value of h, or a value h_0 determined from the principle underlying $\mathcal{P}_{h_n}$. To obtain the latter we write as an extension of (4.203), using also (3.19),

$$\begin{aligned} L(A_1) - L(A_0) &= \\ &= h\,\text{tr}\{\,(B_0'^{-1}\, I_{[K_y\, K_x]} - S_0^{-1} A_0\, M_{xx})\,\Delta A_0'\,\} \\ &+ \frac{1}{2}\,h^2\,\text{tr}\,\{-(B_0'^{-1}\,\Delta B_0')^2 + S_0^{-1}(A_0\, M_{xx}\,\Delta A_0' \\ &\qquad + \Delta A_0'\, M_{xx}\, A_0')\, S_0^{-1} A_0\, M_{xx}\,\Delta A_0' - S_0^{-1}\,\Delta A_0\, M_{xx}\,\Delta A_0'\,\} \\ &+ \cdots\,. \end{aligned} \tag{4.207}$$

Using (4.206) we find that the sum of the two terms shown in (4.207) is maximized if h is given the value

(4.208)
$$h_0 = \frac{\operatorname{tr}(S_0^{-1}\,\Delta A_0\,M_{xx}\,\Delta A_0')}{\operatorname{tr}\{(B_0'^{-1}\,\Delta B_0') - S_0^{-1}(A_0\,M_{xx}\,\Delta A_0' + \Delta A_0\,M_{xx}\,A_0)S_0^{-1}A_0\,M_{xx}\,\Delta A_0' + S_0^{-1}\,\Delta A_0\,M_{xx}\,\Delta A_0'\}}.$$

We define matrices M_0^* and L_0^* such that the numerator and denominator in (4.208) are identical with those in

(4.209)
$$h_0 = \frac{\Delta a_0^*\,M_0^*\,\Delta a_0'^*}{-\Delta a_0^*\,L_0^*\,\Delta a_0'^*},$$

postponing their explicit evaluation until the discussion of computational arrangements below.

4.4.6. Asymptotic properties of $\mathfrak{P}_h$ *and* $\mathfrak{P}_{h_n}$. If we write as before $A_n = A + \bar{A}_n$, the expansion of (4.206) in terms of $\bar{A}_0$ is, by use of (4.205),

(4.210)
$$\operatorname{vec}^*(S^{-1}\,\Delta\bar{A}_0\,M_{xx}) + \cdots$$
$$= \operatorname{vec}^*\{-B'^{-1}\,\bar{B}_0'\,B'^{-1}\,I_{[K_y\,K_x]} + S^{-1}(A\,M_{xx}\,\bar{A}_0' + \bar{A}_0\,M_{xx}\,A')S^{-1}\,A\,M_{xx} - S^{-1}\,\bar{A}_0\,M_{xx}\}.$$

Omitting bars from Δa_n and ΔA_n (see p.174), this can be written as

(4.211)
$$\Delta a_0^*\,M^* + \cdots = \bar{a}_0^*\,L^* + \cdots,$$

with suitable definitions of the matrices M^* and L^*, which are now the same functions of A and M_{xx} as the matrices M_0^* and L_0^*, respectively, are of A_0 and M_{xx}.

The study of (4.211) is exactly similar to that of (4.79) in the previous case of uncorrelated disturbances. Formulae (4.80) through (4.88) and the discussion connected therewith remain valid

with the new definitions of M^* and L^*. Limits on the characteristic values l_q, $q = 1, \ldots, Q$, of L^* can again be determined in the case that the maximum of the likelihood function is not depressed by the restrictions, as follows: Retaining for that case the definition (4.93) of H, we have instead of (4.94)

$$H M_{xx} H' = \begin{bmatrix} S & 0 \\ 0 & I_{[K_z]} \end{bmatrix} = T , \tag{4.212}$$

say. Writing for the moment, instead of (4.95),

$$\bar{A} = \tilde{A}^{\oplus} H , \tag{4.213}$$

we have from (4.212) and (4.202)

(4.214)

$$\begin{aligned} L_{(2)} &\equiv \operatorname{tr}\{-(B'^{-1}\bar{B}')^2 + S^{-1}(AM\bar{A}' + \bar{A}M_{xx}A')S^{-1}AM_{xx}\bar{A}' - S^{-1}\bar{A}M_{xx}\bar{A}'\} \\ &= \operatorname{tr}\{-(\tilde{B}'^{\oplus})^2 + (\tilde{B}'^{\oplus} + S^{-1}\tilde{B}^{\oplus}S)\tilde{B}'^{\oplus} - S^{-1}(\tilde{B}^{\oplus}S\tilde{B}'^{\oplus} + \tilde{C}^{\oplus}\tilde{C}'^{\oplus})\} \\ &= -\operatorname{tr}(S^{-1}\tilde{C}^{\oplus}\tilde{C}'^{\oplus}) \end{aligned}$$

and

$$M_{(2)} \equiv \operatorname{tr}(S^{-1}\bar{A}M_{xx}\bar{A}') = \operatorname{tr}\{S^{-1}(\tilde{B}^{\oplus}S\tilde{B}'^{\oplus} + \tilde{C}^{\oplus}\tilde{C}'^{\oplus})\} . \tag{4.215}$$

Through a further transformation

$$S = UU' , \qquad U^{-1}\tilde{A}^{\oplus}\begin{bmatrix} U & 0 \\ 0 & I_{[K_z]} \end{bmatrix} = \tilde{A} , \tag{4.216}$$

it is seen that, in the absence of restrictions on A, the characteristic values l_q are the stationary values of the quadratic form

$$L_{(2)} = -\operatorname{tr}(\tilde{C}\tilde{C}') = -\sum_{g=1}^{K_y}\sum_{h=K_y+1}^{K_x}\tilde{a}_{gh}^2 \tag{4.217}$$

under the restrictions

$$M_{(2)} = \operatorname{tr}(\tilde{A}\,\tilde{A}') = \sum_{g=1}^{K_y}\sum_{h=1}^{K_x} \tilde{a}_{gh}^2 = 1. \tag{4.218}$$

We record in one formula combining (4.213) and (4.216) the transformation

$$\bar{A} = U\,\tilde{A}\begin{bmatrix} U^{-1} & 0 \\ 0 & I_{[K_z]} \end{bmatrix} H = U\,\tilde{B}\,U^{-1}A + U\,\tilde{C}\,F\,I_{[K_z\,K_x]} \tag{4.219}$$

through which the forms (4.217) and (4.218) have been derived. These forms lead to the following complete table of characteristic values and vectors.

(4.220)

(a) Value of l	(b) Value of $k = 1 + h\,l$	(c) Multiplicity	(d) Characteristic "vectors" $\tilde{A}$ satisfy
0	1	$K_y\,K_z$	$\tilde{C} = 0$
-1	$1 - h$	$(K_y)^2$	$\tilde{B} = 0$

Therefore, under any a priori restrictions that permit the likelihood function to attain its absolute maximum,

$$-1 \le l_q \le 0, \quad 1 - h \le k_q \le 1, \quad q = 1, \ldots, Q. \tag{4.221}$$

Furthermore, under any such restrictions that in addition ensure (as here supposed) complete identification of each structural equation,

$$-1 \le l_q < 0, \quad 1 - h \le k_q < 1, \quad q = 1, \ldots, Q - K_y, \tag{4.222}$$

from which we exclude the K_y characteristic values $l = 0$, connected with the freedom of normalization of A (choice of diagonal elements of $\tilde{B}^{\oplus} = U\,\tilde{B}\,U^{-1}$) and unaffected by homogeneous restrictions (4.204).

4.4.7. Considerations in choosing a constant value of h. It follows that, among processes $\mathcal{P}_h$ with a constant value of h, $\mathcal{P}_1$ does

not have the excellence it possesses in one important case with uncorrelated disturbances. In large samples under valid restrictions, $\mathcal{P}_1$ confines the characteristic values $k_q = 1 + h l_q$ approximately to the interval $0 \leq k_q < 1$, so that, unless all k_q vanish, $\max |k_q|$ can be decreased by taking $h > 1$.

As a guide in determining how far above 1 to choose h, it is of interest to ask what type of restrictions will exclude the characteristic value $l = -1$. This value will remain present as long as the restrictions permit an addition to A of the type

$$\bar{A} = U \tilde{G} F I_{[K_z K_x]} = \bar{C} I_{[K_z K_x]} \tag{4.223}$$

containing only the second term of (4.219). It follows that, for the exclusion of the characteristic value $l = -1$, it is necessary and sufficient that in the canonical form of the basic matrix Φ^* the submatrix Φ^*_{III} be absent. If this is the case, the same reasoning from ignorance that previously favored $\mathcal{P}_1$, now leads to the recommendation of $\mathcal{P}_2$: the relevant values k_q are then confined to the interval $-1 < k_q < 1$. However, if Φ^*_{III} is present, any constant value of h should be chosen below 2, and the nearer to 2, the nearer the highest of the values l_q, $q = 1, \ldots, Q - K_y$, is suspected of being to zero.

*4.4.8. *Problems in the arrangement of computations for* $\mathcal{P}_h$, $\mathcal{P}_{h_n}$. We shall now write (4.206), using (4.40) and the orthogonalization (4.177) of Φ^*, in the form

$$\Delta a_n^* M_n^* = \operatorname{vec}(B_n'^{-1} I_{[K_y K_z]}) \cdot \Phi'^* - a_n^* M_n^*, \qquad n = 0, 1, \ldots, \tag{4.224}$$

where M_n^* is defined by

$$M_n^* \equiv \Phi^* M_n \Phi'^*, \tag{4.225}$$

$$M_n \equiv S_n^{-1} \otimes M_{xx} \equiv \begin{bmatrix} s_n^{11} M_{xx} & s_n^{12} M_{xx} & \cdots & s_n^{1G} M_{xx} \\ s_n^{21} M_{xx} & s_n^{22} M_{xx} & \cdots & s_n^{2G} M_{xx} \\ \cdot & \cdot & \cdots & \cdot \\ s_n^{G1} M_{xx} & s_n^{G2} M_{xx} & \cdots & s_n^{GG} M_{xx} \end{bmatrix}, \tag{4.226}$$

with s_n^{gh} denoting the elements of

$$S_n^{-1} \equiv [\, s_n^{gh} \,], \tag{4.227}$$

which are again functions of A_n. Combining (4.225) and (4.226), we can alternatively write for M_n^*, using (4.31),

$$M_n^* = \begin{bmatrix} M_n^{11} & \cdots & M_n^{1G} \\ \cdot & \cdots & \cdot \\ M_n^{G1} & \cdots & M_n^{GG} \end{bmatrix}, \tag{4.228}$$

with the further definition

$$M_n^{gh} \equiv s_n^{gh}\, \Phi^g\, M_{xx}\, \Phi'^h \equiv s_n^{gh}\, V^{gh}, \tag{4.229}$$

say. The symbol V^{gh} is merely an abbreviation for the matrix product it represents. There is no meaning, in the present context, in putting the matrices V^{gh} together to form a larger matrix V^*. In the special case that $S_n = I_{[K_y]}$, M_n^* goes over into M^{**} as defined in (4.72), in which $M_n^{gh} = 0$ for $g \neq h$.

Since M_n^* changes from one iteration to the next, there is no advantage in avoiding explicit use of q-coordinates. Likewise, in solving for Δa_n^* from (4.224), there is no greater advantage from the use of the canonical form of the basic matrix Φ^* than there is in general from the use of any partitioning method for the inversion of a matrix or the solution of linear equations. The new element in the present situation, as compared with the application of P_h in the case of uncorrelated disturbances, is that now M_n^* does not partition into diagonal blocks. In principle, therefore, we now have one high-order inversion job instead of K_y lower-order inversions – a situation such as was already encountered in the Newton method in the case of uncorrelated disturbances (because L_n^* likewise does not partition). The main problem of computational economy now is to find an efficient method of solving for Δa_n^* from (4.224) which takes advantage of the special form of M_n^*. This problem again has not been systematically investigated by us. The following considerations seem relevant.

Of the matrices entering into the definition of M_n^*, those re-

maining the same through all iterations are M_{xx} and Φ^*. This suggests that it will be advantageous to go as far as possible toward the solution of Δa_n^* on the basis of these matrices alone before the matrix S_n specific to the nth iteration is brought into play. One possible procedure would be to start from the matrices V^{gh} as basic material, developing functions of these matrices that facilitate the solution of Δa_n^* from (4.224) for all n. This method would be similar to the partitioning method of matrix inversion, although a complete inversion of M_n^* may not be needed.

A perhaps more powerful method would be to utilize the common origin of all V^{gh} in M_{xx} on the basis of one initial inversion of M_{xx} used in

$$M_n^{-1} = S_n \otimes M_{xx}^{-1}, \tag{4.230}$$

followed by

$$M_n^{*-1}(*) = \Phi'^{-1}(*)\, M_n^{-1}\, \Phi^{-1}(*). \tag{4.231}$$

[The evaluation of (4.231) may be facilitated by orthogonalization of $\Phi(*)$.] This approach requires an economical method of finding M_n^{*-1} if both M_n^* and $M_n^{-1}(*)$ are available.

*4.4.9. *Processes $\mathfrak{P}_h$ and $\mathfrak{P}_{h_n}$ modified by normalization.* In the derivation of the first-order maximum-likelihood conditions (4.205) from (4.203) we have not imposed any normalization on A_0 and A. If, alternatively, we had required that both A_0 and A satisfy the normalization rule (4.197), ΔA_0 would have been restricted by

$$\operatorname{vec}^*_{[1]}\, \Delta A_0 = 0, \tag{4.232}$$

and the first-order condition for a maximum would be expressed by

$$\begin{cases} [1]\operatorname{vec}^*(B'^{-1} I_{[K_y K_x]} - S^{-1} A M_{xx}) = 0, & (4.233,-1) \\ a^*_{[1]} = [\,1\ 1\ \cdots\ 1\,]. & (4.233,+1) \end{cases} \tag{4.233}$$

It follows from the homogeneity properties of the likelihood function that (4.233) is equivalent to (4.205). However, the anal-

ogous iterative procedures using ΔA_0 defined by

$$(4.234)\begin{cases} (4.234,-1) & {}_{[1]}\mathrm{vec}^*(S_0^{-1}\,\Delta A_0\,M_{xx}) \\ & \quad = {}_{[1]}\mathrm{vec}(B_0'^{-1}\,I_{[K_y K_x]} - S_0^{-1}\,A_0\,M_{xx}), \\ (4.234,+1) & \Delta a^*_{0,[1]} = 0, \quad \text{or} \\ & \Delta a_0 = (\Delta_{[1]}a_0^*)_{[1]}\Phi^*, \end{cases}$$

are essentially different from those based on (4.206). For, even if A_0 satisfies the normalization rule (4.197), the solution ΔA_0 of (4.206) cannot satisfy (4.234, +1) for all possible choices of that row of each Φ^β on which normalization is based. For if that were so, Δa_0^* and therewith ΔA_0 would vanish identically. In general, therefore, the substitution of (4.234, +1) for an equal number K_y of the equations (4.206) leads to nonproportional changes in the elements of the solution ΔA_0.

It follows that the convergence properties of the modified processes $\mathfrak{P}_h$, $\mathfrak{P}_{h_n}$ based on (4.234) differ from those derived from (4.206) and depend on what rows of Φ^* have been selected for normalization purposes. We have not investigated the effect of this application of the normalization rule (4.197) on the asymptotic convergence properties. There is reason to believe that the effect is not a radical modification. For, whatever value of h is chosen, the characteristic value $l = 0$ corresponding to the diagonal elements of $\tilde{B}^\oplus = U\,\tilde{B}\,U^{-1}$ in (4.220) connected with the scales of the rows of A leads to $k = 1$. There is therefore no alternation or other unsteadiness in scales in the application of (4.206). Thus, in the first approximation, the elements of Δa_0^* determined from (4.206) differ from zero only to an extent required for improving the ratios of the elements of a_0^* in the next approximation a_1^*. Hence the fixing of certain elements of Δa_0^* at the value zero while relaxing an equal number of the conditions (4.206) might somewhat retard, but need not destroy, convergence.

This point is important because the modification of $\mathfrak{P}_h$ through normalization reduces by K_y the number of unknowns in Δa_n^* to be

determined in each iteration. The resulting saving of labor per iteration may be considerable except possibly in methods based on the initial inversion of M_{xx} for use in (4.230).

The computational arrangement for the modified processes $\mathcal{P}_h$, $\mathcal{P}_{h_n}$ differs from that described earlier only in easily perceived details.

4.4.10. Numerical illustrations of $\mathcal{P}_1$, $\mathcal{P}_{5/4}$. We have applied $\mathcal{P}_1$ and $\mathcal{P}_{5/4}$ without normalization, and $\mathcal{P}_1$ modified by normalization, to the constructed example discussed before. As initial values we have taken the result of the 6th iteration with $\mathcal{P}_1$ in the case where Σ was assumed to be diagonal. The lower cost per iteration under that assumption is a good reason to apply it initially, albeit only to get closer to the maximizing value A without restrictions on Σ. The results are shown in Table 4.4.10.

**4.4.11. The Newton method.* All properties previously derived from the general formulation (4.173) of the Newton method carry over, of course, to the present case. All that is here required is to derive new expressions for the first and second derivatives of the likelihood function in the initial point A_0, generalizing formulae (4.178) and (4.179) of the previous case.

It will be remembered that the Newton method requires the matrix of second derivatives of the likelihood function to be nonsingular. As before, complete identification, and, as a new requirement, use of the normalization rule (4.197), are therefore now indispensable. Instead of the previous formulation (4.172) we thus obtain as the definition of the Newton method

$$a_1^* = a_0^* + \Delta a_0^*,$$

$$(\Delta_{[1]} a_0^*)\ {}_{[1]}L_0^* = -\ {}_{[1]}l_0\,, \qquad \Delta a^*_{0,[1]} = 0\,. \tag{4.235}$$

Assuming again complete orthogonality (4.177) of Φ^* (which is compatible with any choice of one row of each Φ^g for normalization purposes), we have from (4.203)

$$[1]l_0 \equiv \left(\frac{d\,L(\alpha^*)}{d\ {}_{[1]}\alpha^*}\right)_{\alpha^* = a_0^*} = {}_{[1]}\mathrm{vec}^*\{B_0'^{-1}\, I_{[K_y\, K_x]} - S_0^{-1}\, A_0\, M_{xx}\}\,. \tag{4.236}$$

TABLE 4.4.10

Numerical illustrations of $\mathfrak{P}_1$ *and* $\mathfrak{P}_{5/4}$ *without restrictions on* Σ.

Method	Iteration $n=$	Row $g=$	Matrices B_n $k=1$	2	3	Matrices S_n $h=1$	2	3	Scale Factors[2] $a_{n;g,g+3}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	0	1	0.00000	0.97020	4.17885	0.19385	0.08535	−0.02277	
		2	1.00207	0.00000	−2.99581	0.08535	0.20000	0.08700	
		3	−2.04599	0.96553	0.00000	−0.02277	0.08700	0.29509	
$\mathfrak{P}_1$	1	1	0.00000	0.99521	4.04308	0.19834	0.09630	−0.00515	1.00140
		2	0.99144	0.00000	−2.97534	0.09630	0.19674	0.09717	1.05436
without normalization[1]		3	−2.01349	0.99600	0.00000	−0.00515	0.09717	0.29990	0.99498
	2	1	0.00000	0.99892	4.00812	0.19958	0.09919	−0.00103	0.99753
		2	0.99789	0.00000	−2.99534	0.09919	0.19929	0.09948	1.00128
		3	−2.00275	0.99930	0.00000	−0.00103	0.09948	0.30000	0.99882
	3	1	0.00000	0.99976	4.00157	0.19990	0.09982	−0.00021	0.99937
		2	0.99951	0.00000	−2.99918	0.09982	0.19986	0.09991	1.00012
		3	−2.00058	0.99989	0.00000	−0.00021	0.09991	0.30001	0.99970
True Values		1	0.00000	1.00000	4.00000	0.20000	0.10000	0.00000	
		2	1.00000	0.00000	−3.00000	0.10000	0.20000	0.10000	
		3	−2.00000	1.00000	0.00000	−0.00000	0.10000	0.30000	

[1] As described in section *4.4.8*, no normalization has been imposed as part of the computation of each iteration result from the preceding iteration result. See, however, note 2.

[2] For the purpose of comparison of successive iteration results, each iteration result has been re-normalized by $a_{14} = a_{25} = a_{36} = 1$ before being entered in this table. Column (10) gives values of $a_{n,14}$, $a_{n,25}$, $a_{n,36}$ obtained, before such re-normalization, by one application of $\mathfrak{P}_1$ (without normalization) to the matrices B_{n-1}, S_{n-1} as stated in the table.

TABLE 4.4.10
(*Continued*)

Method	Iteration $n =$	Row $g =$	Matrices B_n $k = 1$	2	3	Matrices S_n $h = 1$	2	3	Scale Factors[2] $a_{n;g,g+3}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	0	1	0.00000	0.97020	4.17885	0.19385	0.08535	−0.02277	1.00000
		2	1.00207	0.00000	−2.99581	0.08535	0.20000	0.08700	1.00000
		3	−2.04599	0.96553	0.00000	−0.02277	0.08700	0.29509	1.00000
$\mathcal{P}_{5/4}$	1	1	0.00000	1.00146	4.00914	0.20059	0.09890	0.00002	1.00140
		2	0.98878	0.00000	−2.97022	0.09890	0.19595	0.09957	1.05436
without		3	−2.00537	1.00362	0.00000	0.00002	0.09957	0.30200	0.99498
normalization[1]	2	1	0.00000	0.99944	3.99793	0.19978	0.10020	0.00000	1.00215
		2	1.00258	0.00000	−3.00778	0.10020	0.20104	0.10020	0.99025
		3	−1.99864	0.99936	0.00000	0.00000	0.10020	0.29961	1.00302
	3	1	0.00000	1.00013	4.00054	0.20005	0.09995	0.00000	0.99947
		2	0.99935	0.00000	−2.99806	0.09995	0.19974	0.09995	1.00263
		3	−2.00034	1.00017	0.00000	0.00000	0.09995	0.30010	0.99934
True Values		1	0.00000	1.00000	4.00000	0.20000	0.10000	0.00000	
		2	1.00000	0.00000	−3.00000	0.10000	0.20000	0.10000	
		3	−2.00000	1.00000	0.00000	0.00000	0.10000	0.30000	

[1]As described in section *4.4.8*, no normalization has been imposed as part of the computation of each iteration result from the preceding iteration result. See, however, note 2.

[2]For the purpose of comparison of successive iteration results, each iteration result has been re-normalized by $a_{14} = a_{25} = a_{36} = 1$ before being entered in this table. Column (10) gives values of $a_{n,14}$, $a_{n,25}$, $a_{n,36}$ obtained, before such re-normalization, by one application of $\mathcal{P}_{5/4}$ (without normalization) to the matrices B_{n-1}, S_{n-1} as stated in the table.

TABLE 4.4.10
(Continued)

Method	Iteration	Row	Matrices B_n			Matrices S_n		
	n =	g =	k = 1	2	3	h = 1	2	3
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	0	1	0.00000	0.97020	4.17885	0.19385	0.08535	−0.02277
		2	1.00207	0.00000	−2.99581	0.08535	0.20000	0.08700
		3	−2.04599	0.96553	0.00000	−0.02277	0.08700	0.29509
$\mathcal{P}_1$	1	1	0.00000	0.99331	4.03780	0.19760	0.09666	−0.00542
		2	0.99910	0.00000	−3.00006	0.09666	0.19986	0.09740
modified by normalization[1]		3	−2.00830	0.99236	0.00000	−0.00542	0.09740	0.29796
	2	1	0.00000	0.99772	4.00567	0.19911	0.09911	−0.00126
		2	0.99923	0.00000	−2.99997	0.09911	0.19987	0.09938
		3	−2.00071	0.99768	0.00000	−0.00126	0.09938	0.29915
	3	1	0.00000	0.99931	4.00083	0.19973	0.09976	−0.00031
		2	0.99973	0.00000	−3.00002	0.09976	0.19996	−0.09984
		3	−1.99991	0.99931	0.00000	−0.00031	0.09984	0.29972
		1	0.00000	1.00000	4.00000	0.20000	0.10000	0.00000
True Values		2	1.00000	0.00000	−3.00000	0.10000	0.20000	0.10000
		3	−2.00000	1.00000	0.00000	0.00000	0.10000	0.30000

[1] As described in section *4.4.9*.

TABLE 4.4.10
(Continued)

Method	Iteration	Row	Matrices B_n			Matrices S_n		
	$n =$	$g =$	$k = 1$	2	3	$h = 1$	2	3
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\wp_{5/4}$ modified by normalization[1]	0	1	0.00000	0.97020	4.17885	0.19385	0.08535	−0.02277
		2	1.00207	0.00000	−2.99581	0.08535	0.20000	0.08700
		3	−2.04599	0.96553	0.00000	−0.02277	0.08700	0.29509
	1	1	0.00000	0.99909	4.00254	0.19964	0.09950	−0.00038
		2	0.99835	0.00000	−3.00112	0.09950	0.19983	0.09999
		3	−1.99888	0.99907	0.00000	−0.00038	0.09999	0.29952
	2	1	0.00000	0.99972	4.00037	0.19995	0.09997	−0.00003
		2	0.99963	0.00000	−3.00031	0.09997	0.20000	0.09999
		3	−1.99973	0.99975	0.00000	−0.00003	0.09999	0.29996
	3	1	0.00000	0.99997	3.99998	0.20001	0.10002	0.00002
		2	0.99999	0.00000	−3.00002	0.10002	0.20001	0.10000
		3	−1.99998	0.99998	0.00000	0.00002	0.10000	0.30002
True Values		1	0.00000	1.00000	4.00000	0.20000	0.10000	0.00000
		2	1.00000	0.00000	−3.00000	0.10000	0.20000	0.10000
		3	−2.00000	1.00000	0.00000	0.00000	0.10000	0.30000

[1]As described in section *4.4.9*.

Furthermore, because $\Delta a^*_{0,[1]} = 0$, we obtain

$$(\Delta_{[1]}a_0^*)\,{}_{[1]}L_0^* = \frac{1}{2}\frac{d}{d(\Delta_{[1]}a_0^*)}\{(\Delta_{[1]}a_0^*)\,{}_{[1]}L_0^*\,(\Delta_{[1]}a_0'^*)\}$$

$$= \frac{1}{2}\frac{d}{d(\Delta_{[1]}a_0^*)}\,(\Delta a_0^*\; L_0^*\;\Delta a_0'^*)\,. \tag{4.237}$$

Combining (4.235), (4.236), (4.237), and the definition (4.208) – (4.209) of L_0^*, we obtain

$$\begin{aligned}(\Delta_{[1]}a_0^*)\,{}_{[1]}L_0^* = {}&{}_{[1]}\mathrm{vec}^*\{-[B_0'^{-1}\,\Delta B_0'\,B_0'^{-1} \qquad 0]\\ &+ S_0^{-1}(A_0\,M_{xx}\,\Delta A_0 + \Delta A_0\,M_{xx}\,A_0')\,S_0^{-1}\,A_0\,M_{xx}\\ &- S_0^{-1}\,\Delta A_0\,M_{xx}\}\\ = {}&{}_{[1]}\mathrm{vec}^*(-B_0'^{-1}\,I_{[K_y\,K_x]} + S_0^{-1}\,A_0\,M_{xx})\end{aligned} \tag{4.238}$$

as a formulation of the Newton method adapted to computational use. The middle member of (4.238) serves to define ${}_{[1]}L_0^*$ through

$$\Delta A_0 \equiv \mathrm{mat}^*\{(\Delta_{[1]}a_0^*)\,{}_{[1]}\Phi^*\}\,. \tag{4.239}$$

The repeated evaluation of ${}_{[1]}L_n^*$ from (4.238) and (4.239) and its inversion (or other method of solving for $\Delta_{[1]}a_n^*$) are laborious. The problem of how to take advantage of the regularities in ${}_{[1]}L_n^*$ for its inversion also appears more formidable than in the case of $\mathfrak{P}_h$ where only M_n^* as defined in (4.225) needs to be inverted.

*4.4.12. *Numerical experiment with the Newton method.* A numerical experiment in which (4.238) was applied to the constructed example discussed earlier with the (least-squares) initial value A_0

used in Table 4.3.4.4 did not produce convergent results, in contrast with the modified $\mathfrak{P}_1$, as defined by (4.234,- 1), which led to convergent iterations from the same initial value. We have not repeated the experiment with the (closer) initial value used in Table 4.4.10.

*4.4.13. *Estimated sampling variances and covariances of the estimated coefficients* a_{gk}. It follows from Theorem 3.3.10 that

$$\text{(4.240)}\quad \operatorname{est}\mathcal{E}\{({}_{[1]}a'^{*} - {}_{[1]}\alpha'^{*})({}_{[1]}a^{*} - {}_{[1]}\alpha^{*})\} = -\frac{1}{T}({}_{[1]}L^{*})^{-1}$$

defines consistent estimates of the sampling variances and covariances of the estimates ${}_{[1]}a^{*}$ of the parameters ${}_{[1]}\alpha^{*}$. Their evaluation requires inversion of the matrix ${}_{[1]}L_n^{*}$ computed from the final value A_n with which iterations are terminated. If this is done from a value A_n obtained by a method other than the Newton method, a check on A_n is obtained at small extra cost by employing $({}_{[1]}L_n^{*})^{-1}$ for one more iteration by the Newton method.

If estimated sampling variances and covariances of the estimates ${}_{[1]}a^{*}, S$ of all parameters ${}_{[1]}\alpha^{*}, \Sigma$ are desired, it is necessary to operate with the second-derivative matrix of the original likelihood function $L(\mathrm{A}, \Sigma)$ defined by (4.199). Denoting by

$$\text{(4.241)}\quad \begin{aligned}\sigma &= [\ \sigma_{11}\ \ \sigma_{12}\ \ \cdots\ \ \sigma_{1G}\ \ \sigma_{21}\ \ \cdots\ \ \sigma_{2G}\ \ \cdots\ \ \sigma_{GG}\ ] \\ &= \operatorname{vec}\Sigma\end{aligned}$$

a row vector containing all independent elements of Σ, we have for any direction $\delta\Sigma = \delta\Sigma'$ of variation of Σ

$$\text{(4.242)}\quad \frac{\partial L(\mathrm{A}, \Sigma)}{\partial\sigma}\delta\sigma' = -\frac{1}{2}\operatorname{tr}\{\Sigma^{-1}\delta\Sigma(I - \Sigma^{-1}\mathrm{A}M_{xx}\mathrm{A}')\}.$$

We introduce the notational definition

$$\left(\begin{bmatrix} \dfrac{\partial^2}{\partial_{[1]}\alpha'^* \, \partial_{[1]}\alpha^*} & \dfrac{\partial^2}{\partial_{[1]}\alpha'^* \, \partial\sigma} \\ \dfrac{\partial^2}{\partial\sigma' \, \partial\alpha^*} & \dfrac{\partial^2}{\partial\sigma' \, \partial\sigma} \end{bmatrix} L(\mathrm{A}, \Sigma) \right)_{\substack{\mathrm{A} = A \\ \Sigma = S}}$$

(4.243)

$$= \begin{bmatrix} {}_{[11]}\hat{L}^* & {}_{[1]}\hat{L}^*_{\sigma} \\ {}_{[1]}\hat{L}'^*_{\sigma} & \hat{L}^*_{\sigma\,\sigma} \end{bmatrix} = \hat{L}.$$

Then, if $\delta a^*_{[1]} = 0$, we have from (4.242) and (4.202), after using (3.16) and (3.19),

$$\delta_{[1]}a^* \cdot {}_{[11]}\hat{L}^* \cdot \delta_{[1]}a'^* = \mathrm{tr}\{-(B'^{-1}\,\delta A')^2 - S^{-1}\,\delta A\, M_{xx}\,\delta A'\},$$

$$\delta_{[1]}a^* \; {}_{[1]}\hat{L}^*_{\sigma}\,\delta s' = \mathrm{tr}\{S^{-1}\,\delta\Sigma\, S^{-1}\, A\, M_{xx}\,\delta A'\},,$$

(4.244)
$$\delta s\, \hat{L}^*_{\sigma\sigma}\,\delta s' = -\frac{1}{2}\,\mathrm{tr}(S^{-1}\,\delta\Sigma)^2,$$

where s is defined analogously to σ in (4.241). These formulae serve to evaluate $\hat{L}$. The desired sampling covariances are now obtained from

$$\mathrm{est}\,\mathcal{E}\begin{bmatrix} ({}_{[1]}a'^* - {}_{[1]}\alpha'^*) \\ (s' - \sigma') \end{bmatrix}\begin{bmatrix} ({}_{[1]}a^* - {}_{[1]}\alpha^*) & (s - \sigma) \end{bmatrix}$$

(4.245)
$$= -\frac{1}{T}\,\hat{L}^{-1},$$

possibly by the partitioning method of inversion already quoted

[Hotelling, 1943-A, p. 4]. Of course, the matrix ${}_{[1]}L^{*-1}$ inverse to the matrix ${}_{[1]}L^{*}$ defined by (4.238) is a principal submatrix of $\hat{L}^{-1}$ located in the upper left corner.

4.5. *Concluding Remarks*

4.5.1. Nature of the concluding remarks. In this concluding section 4.5 we shall first make rough comparisons between the costs of computation in the various methods discussed. These comparisons will give occasion to recall certain problems of matrix computation which have not been investigated by us and to make some remarks on suitable methods for the various matrix inversions required. Secondly, we shall indicate a possible generalization of the restrictions on A. Finally, we shall draw attention to important problems connected with the number and nature of different maxima of the likelihood function which require solution before full reliance can be placed in the methods developed.

4.5.2. Uncertainty in computation costs. A good measurement of computation cost requires counts of the number of operations involved (initially and per iteration), distinguishing additions, multiplications, and divisions, and indicating the number of decimals required in intermediate steps for a given decimal accuracy in the result. If such measurements were available, cost comparisons would still depend on insufficiently known relative speeds of convergence per iteration. However, even initial cost and cost per iteration cannot be measured by the counts indicated because in applications to economic equation systems so much depends on the precise form of the basic matrix Φ^*. In addition, there is still considerable uncertainty about the most economical method of inversion or solution of linear equations in cases where Φ^* is already specified.

4.5.3. Cost comparisons between various methods. For these reasons we shall confine ourselves to setting out in Table 4.5.3 in a comparative fashion the main features of each method affecting cost of computation. In reading this table, which requires reference to earlier formulae for detailed comparisons, it must be remembered that the inversion of a matrix of order N involves a number of operations proportional to N^3 and that the multiplication of an $N_1 \times N_2$-matrix into an $N_2 \times N_3$-matrix requires $N_1 N_2 N_3$ multiplications and an almost equal number of additions.

TABLE 4.5.3
Comparison of Various Methods

Method	I. Case of uncorrelated disturbances (Σ diagonal)	II. Case of unrestricted variances and covariances Σ of disturbances
A. Methods $\mathcal{P}_h$ and $\mathcal{P}_{h_n}$ based on M_n^*	1. $M_n^* = M^{**} = \Phi^*(I_{[K_y]} \otimes M_{xx})\,\Phi'^*$ is constant in iterations. Hence (a) α^*_{III} can be eliminated, and (b) the required extent of the inversion of M^{**}, and the transformations $B = \text{mat}({}_{\text{III}}\alpha^* \cdot {}_{\text{III}}\Phi^*)$, can be carried out for all iterations at once. 2. M^{**} partitions into diagonal blocks, reducing the number of operations in its inversion in the ratio[1] $(\Sigma\, {}^{\text{III}}Q_g)^3$ to $\Sigma({}^{\text{III}}Q_g)^3$. 3. Each iteration requires the inversion of B_n.	1. $M_n^* = \Phi^*(S_n^{-1} \otimes M_{xx})\,\Phi'^*$ is not constant in iterations. Its inversion (or other processing in solving for α^*_{n+1}) must be repeated for each iteration. 2. There are probable advantages from the regularities in M^* for its inversion or other processing, perhaps dispensing with inversion of S_n for each iteration. 3. Each iteration requires the inversion of B_n.

[1] ${}^{\text{III}}Q_g$ is defined as $Q_g^{\text{I}} + Q_g^{\text{II}}$.

TABLE 4.5.3
(Continued)

B. Newton method based on L_n^*	1. L_n^* contains only two terms.	1. $_{[1]}L_n^*$ contains four terms.
	2. The first term requires for its evaluation the inversion of B_n for each iteration. This term vanishes outside $_{\text{III III}}L_n^*$.	2. Its evaluation requires the inversion of B_n and S_n for each iteration.
	3. The remaining term $-M_n^*$ is constant in iterations, so that use of $^{\text{III}}L_n^*$ permits the elimination of α_{III}^*. The required inversion of $M_{\text{III III}}^{**}$ is facilitated by partitioning into diagonal blocks.	3. No submatrix of $_{[1]}L_n^*$ remains constant in iterations. Its inversion or other processing must be repeated for each iteration.
	4. The inversion of $^{\text{III}}L_n^*$ (or other processing in solving for $_{\text{III}}a_{n+1}^*$) must be repeated for each iteration.	4. The advantage from regularities in $_{[1]}L_n^*$ for its inversion or other processing are highly uncertain.
	5. There may be advantages in the regularities of $^{\text{III}}L_n^*$ for its inversion or other processing.	

The relative cheapness of $\mathfrak{P}_h$ and $\mathfrak{P}_{h_n}$ in the case of uncorrelated disturbances stands out clearly from this table. Not only can M^{**} be inverted (to the extent required) once for all iterations, but its partitioning into diagonal blocks greatly reduces the amount of work involved in that inversion. Precise comparisons between the remaining three methods (entries BI, AII, BII of the table) are made difficult by the uncertainties already mentioned. The general inference can be made that each transition, either from method A to method B within the same case, or from case I to case II within the same method, leads to a considerable increase in cost of computation.

4.5.4. *Methods of inversion.* If we include the computation of sampling variances and covariances of estimated parameters a^*, the inverse of each of the matrices B_n, S_n, ${}_{\text{III III}}M^g$, ${}^{\text{III}}L_n^*$, M_n^*, ${}_{[1]}L_n^*$, is required in at least one of the methods or cases. The problems encountered in taking advantage of the regularities in the definitions (4.186), (4.225) and (4.208) – (4.209) of the last three of these matrices have already been mentioned. Here we only point to the importance of these problems for the methods discussed, and to their inherent mathematical interest. Our further remarks are directed to those inversions where such "advantages" are not present or are not real advantages in the sense that their exploitation is not worth the cost. This will often be the case in computing the relevant submatrices of $(M^g)^{-1}$, which could all be derived from one larger-order inverse M_{xx}^{-1} with the help of orthogonalized basic matrices Φ^g.

Five of the seven inversion jobs listed have to be repeated in successive iterations. This places a premium on iterative methods of inversion since, for instance, S_{n-1}^{-1} can serve as initial value for the iterative inversion of S_n. Iterative methods for inverting matrices have been discussed by Hotelling [1943–A, especially paragraphs 5, 7, 9, 10]. When such a method is applied, the approximation to S_n^{-1} must not be pushed beyond a certain level corresponding to the degree of approximation to A expected to be reached by A_{n+1}. A certain balance between the frequency of iterations in the various parts of the whole calculation should thus be preserved.

The inversion of B_n may offer special opportunities for economies because usually many of its elements are prescribed to vanish. In such cases it is advisable to permute the rows and the first K_y

columns of A (i.e., structural equations and dependent variables, respectively) so as to bring B as nearly as possible into triangular form. Partitioning of B according to (2.82) will be noticed as a by-product of such analysis. If the corresponding partitioning (2.82) of Σ is not assumed, the partitioning (2.82) of B still facilitates the inversion of B_n. In cases where permutation of rows and columns can only produce a compact block of zeros in the southwest corner of B that does not extend to the main diagonal, considerable savings will still be encountered in any of the variants of the Doolittle method applied to the inversion of B_n.

4.5.5. *Generalization of the restrictions on* A. The formulae for all methods discussed admit, without serious complications, of a generalization of the restrictions on A which has already been mentioned in earlier sections. This is the restriction (2.73α) requiring two pairs of coefficients occurring in different structural equations to have the same unknown ratio. In combination with suitably chosen normalization rules, such a restriction can be given the linear form

$$(4.246)\quad \begin{cases} (4.246\text{k}) & \alpha_{g_1 k_1} = \alpha_{g_2 k_2} = 1, \\ (4.246\text{l}) & \alpha_{g_1 l_1} - \alpha_{g_2 l_2} = 0, \end{cases}$$

which differs from restrictions considered earlier only in that elements of different rows of A enter into the same restriction (4.246l). The treatment of restrictions of the type (4.246k) has already been demonstrated above. A restriction of the type (4.246l) can be introduced into the various iterative procedures by incorporating in Φ^*, as defined by (4.31), the row

$$(4.247)\quad \begin{bmatrix} 0(1) & \dots & 0(g_1-1) & \varphi(g_1) & 0(g_1+1) \\ & & \dots & 0(g_2-1) & \varphi(g_2) & 0(g_2+1) & \dots & 0(K_y) \end{bmatrix}$$

with

$$(4.248)\quad \begin{aligned} \varphi(g_1) &\equiv \begin{bmatrix} 0_1 & \dots & 0_{l_1-1} & 1 & 0_{l_1+1} & \dots & 0_{K_x} \end{bmatrix}, \\ \varphi(g_2) &\equiv \begin{bmatrix} 0_1 & \dots & 0_{l_2-1} & 1 & 0_{l_2+1} & \dots & 0_{K_x} \end{bmatrix}, \end{aligned}$$

where the zeros in (4.247) represent vectors of order K_x, whereas the zeros in (4.248) represent scalar components. This new row makes Φ^* different from all the Φ^* previously considered. Previously Φ^* was a matrix with vanishing elements except in the diagonal blocks occupied by Φ^1, Φ^2, ... Φ^G, where Φ^i expressed restrictions on the parameters of the ith equation only. It will be clear that the number of restrictions (4.246) that can be expressed in this manner is limited by the fact that only one normalization rule can be imposed on each structural equation. The only computational complication arising from the presence of rows like (4.247) in Φ^* is that M^* partitions into fewer and larger diagonal blocks.

4.5.6. Unsolved problems in distinguishing the highest maximum of the likelihood function. An important class of unsolved problems, presumably requiring methods quite different from those here employed, is connected with the question of how to make sure that any maximum of the likelihood function found is the highest maximum or, if possible, of how to ensure by choice of initial values that iterations will converge to the highest maximum. Of course, the theory of the asymptotic properties of the maximum-likelihood estimates a^*, S has approximative value only if the highest maximum is well above the next highest. But how will proximity of the two highest maxima be recognized? Will it necessarily be revealed by high sampling variances of the estimates?

The condition

$$\det \mathrm{B} = 0 \tag{4.249}$$

divides the space of the elements α_{gh}, $g, h = 1, \ldots, K_y$, into two connected regions. The logarithmic likelihood function

$$L(\mathrm{A}) = \text{const} + \log \det \mathrm{B} - \frac{1}{2} \operatorname{tr} \mathrm{A} M_{xx} \mathrm{A}' \tag{4.250}$$

in the case of uncorrelated disturbances approaches $-\infty$ whenever B approaches the boundary (4.249) of the two regions. It follows that, whatever the linear restrictions on A, there are at least two maxima of the likelihood function (4.250). Under linear restrictions that are more than adequate in number and variety for complete identification of the structural equations, many more maxima can be expected to arise: there will be at least one maximum for each connected part of the restricted-parameter space cut out by the condition (4.249).

In the case where Σ is unrestricted, no difficulties arise in the subspace of the parameters Σ because the positive definiteness of Σ precludes the passing of a border analogous to (4.249), and for a given A only one maximum (4.200) with respect to Σ exists. Further discussion can therefore be based on the function (4.201) which we rewrite as

$$
\begin{aligned}
L(\mathrm{A}) &= \text{const} + \log\det \mathrm{B} - \frac{1}{2}\log\det \mathrm{A}\, M_{xx}\, \mathrm{A}' \\
&= \text{const} - \frac{1}{2}\log\det \mathrm{B}^{-1}\, \mathrm{A}\, M_{xx}\, \mathrm{A}'\, \mathrm{B}'^{-1} \qquad (4.251) \\
&= \text{const} - \frac{1}{2}\log\det \begin{bmatrix} -I_{[K_y]} & \Pi_{[K_y K_z]} \end{bmatrix} M_{xx} \begin{bmatrix} -I_{[K_y]} \\ \Pi_{[K_y K_z]} \end{bmatrix}.
\end{aligned}
$$

This function will still approach $-\infty$ if B approaches a point B_0 on the boundary (4.249), provided Γ does not simultaneously approach a point Γ_0 such that $\Pi_{[K_y K_z]}$ has a finite limit. It is easily seen that, if the point A_0 approached by A is finite, $\Pi_{[K_y K_z]}$ can remain finite only if A_0 is of rank $K_y - 1$. Points A_0 of this character form a "bridge" across the boundary (4.249) which may complicate the analysis of the number of maxima under linear restrictions on A.

These remarks may suffice to indicate a class of difficult problems, the solution of which is vital to the computation methods here developed. Pending a systematic attack on these problems, the best one can do is to accumulate "practical" experience by trying out various alternative initial values in order to learn from what range of plausible initial values the same maximum is approached.

4.5.7. Choice of initial values A_0. The single-equation least-squares estimates for various choices of "dependent variables" in each equation, obtained anyhow as a by-product of the preparations for the simpler ones of the iterative processes discussed, would seem to be suitable material for such experimenting. If divergence of iterations or convergence to a different maximum for different least-squares initial values occurs frequently, or even occasionally, it will be an important problem to find initial values as near as possible to the highest maximum of the likelihood

function. In the case of unrestricted Σ, probably the best possible initial values for that purpose are obtained by the reduced-form method developed by Anderson and Rubin.[1] While more costly than the least-squares estimates, the reduced-form estimates are superior in that they are consistent estimates. They are, moreover, maximum-likelihood estimates under sacrifice of an amount of a priori information that is perhaps in some sense the minimum sacrifice consistent in general with direct (i.e., noniterative) methods of computation. If so, these estimates are in a sense the nearest one can get to the highest maximum of the likelihood function by direct methods. They may, however, be less economical than least-squares estimates in cases where no doubt exists as to the identity of the highest maximum of the likelihood function. An intermediate choice is given by the maximum-likelihood estimates with diagonally restricted Σ, using all a priori information relating to A, and determined iteratively. These estimates are, of course, not consistent if Σ is actually nondiagonal.

[1] See [IX].

III. NOTE ON THE IDENTIFICATION OF ECONOMIC RELATIONS

BY A. WALD

T. C. Koopmans and H. Rubin have discussed the problem of identification of economic relations in [II – 2], and have obtained a number of very interesting results. In this note the problem treated by Koopmans and Rubin is somewhat generalized and a different approach to its solution is briefly discussed.

1. Definitions and Formulation of the Problem

Let $x_1, \ldots, x_K$ be a set of K variables[1] and let $A = [\alpha_{gk}]$ $(g = 1, \ldots, G;\ k = 1, \ldots, K)$ be a given matrix of rank G. Denote the linear form $\sum_{k=1}^{K} \alpha_{gk} x_k$ by l_g $(g = 1, \ldots, G)$ and let $\Sigma = [\sigma_{gh}]$ $(g, h = 1, \ldots, G)$ be a given symmetric and positive definite matrix. Furthermore, let

$$\varphi_r(\alpha_{11}, \alpha_{12}, \ldots, \alpha_{GK}, \sigma_{11}, \sigma_{12}, \ldots, \sigma_{GG}) = 0 \qquad (r = 1, \ldots, R) \tag{1.1}$$

be a given system of equations, called a priori restrictions, that are satisfied by the quantities α_{gk} and σ_{gh}. For any nonsingular matrix $\Upsilon = [\upsilon_{gh}]$ $(g, h = 1, \ldots, G)$ we shall denote the matrix ΥA by $A(\Upsilon)$ and the elements of $A(\Upsilon)$ by $\alpha_{gk}(\Upsilon)$ $(g = 1, \ldots, G;\ k = 1, \ldots, K)$. Thus, $\alpha_{gk}(\Upsilon) = \sum_{h=1}^{G} \upsilon_{gh} \alpha_{hk}$. Furthermore, we shall

[1] The integer K corresponds to what was denoted K_x in [II].

denote the matrix $\Upsilon \Sigma \Upsilon'$ (Υ' is the transpose of Υ) by $\Sigma(\Upsilon)$ and the elements of $\Sigma(\Upsilon)$ by $\sigma_{gh}(\Upsilon)$ $(g, h = 1, \ldots, G)$. Finally, the linear form $\sum_{h=1}^{G} v_{gh} l_h$ will be denoted by $l_g(\Upsilon)$.

DEFINITION 1.1. *A nonsingular matrix* $\Upsilon = [\, v_{gh} \,]$ $(g, h = 1, \ldots, G)$ *is said to be an admissible transformation if and only if the equations*

$$\varphi_r\{\alpha_{11}(\Upsilon), \alpha_{12}(\Upsilon), \ldots, \alpha_{GK}(\Upsilon), \sigma_{11}(\Upsilon), \sigma_{12}(\Upsilon), \ldots, \sigma_{GG}(\Upsilon)\} = 0 \qquad (r = 1, \ldots, R) \tag{1.2}$$

are fulfilled.

DEFINITION 1.2. *An element* α_{gk} *of the matrix* A *is said to be identifiable*[1] *if* $\alpha_{gk}(\Upsilon)$ *takes only a finite number of different values over the domain of all admissible transformations* Υ. *Similarly, an element* σ_{gh} *of* Σ *is said to be identifiable if* $\sigma_{gh}(\Upsilon)$ *takes only a finite number of different values over the domain of all admissible transformations* Υ.

DEFINITION 1.3. *The linear form* l_g *is said to be identifiable if the coefficients* $\alpha_{g1}, \ldots, \alpha_{gK}$ *are identifiable.*

The matrix A has GK elements and the matrix Σ has $(G^2 + G)/2$ elements. Thus, the total number of elements in the two matrices A and Σ is equal to $GK + (G^2 + G)/2 = P$ (say). Consider the elements of A and Σ arranged in an ordered sequence and denote them by $\theta_1, \ldots, \theta_P$, respectively. The set $\theta = (\theta_1, \ldots, \theta_P)$ can be represented by a point in the P-dimensional space, called parameter space. For any nonsingular transformation Υ we shall denote the point $(\theta_1(\Upsilon), \ldots, \theta_P(\Upsilon))$ by $\theta(\Upsilon)$.

DEFINITION 1.4. *A coordinate* θ_p *of a point* θ *will be said to be locally identifiable if there exists an open set* ω *containing* θ *such that for any admissible transformation* Υ *either* $\theta_p(\Upsilon) = \theta_p$ *or* $\theta(\Upsilon)$ *lies outside* ω.

---

[1] This concept corresponds to what was called multiple identifiability in [II-2.4.4].

The problem considered in this note is to formulate conditions under which a coordinate θ_p of a point θ of the parameter space is identifiable or is locally identifiable.

2. *Two Lemmas*

In this section we shall prove two lemmas which will then be used for deriving necessary and sufficient conditions for the identification of θ_p.

Consider the quadratic form

$$X = \sum_{h=1}^{G} \sum_{g=1}^{G} \sigma^{gh} l_g l_h , \tag{2.1}$$

where $[\sigma^{gh}]$ is the inverse of $[\sigma_{gh}]$. Let ξ_{kl} $(k, l = 1, \ldots, K)$ denote the coefficient of $x_k x_l$ in X. For any nonsingular transformation Υ we shall put

$$X(\Upsilon) = \sum_{h=1}^{G} \sum_{g=1}^{G} \sigma^{gh}(\Upsilon) l_g(\Upsilon) l_h(\Upsilon) , \tag{2.2}$$

where $[\sigma^{gh}(\Upsilon)]$ denotes the inverse of $[\sigma_{gh}(\Upsilon)]$. We shall denote by $\xi_{kl}(\Upsilon)$ the coefficient of $x_k x_l$ in $X(\Upsilon)$.

LEMMA 2.1. *For any nonsingular transformation* Υ *we have* $\xi_{kl}(\Upsilon) = \xi_{kl}$ $(k, l = 1, \ldots, K)$.

Proof: Denote by l the row vector $[l_1 \ldots l_G]$. Using matrix notation we can write

$$X = l \Sigma^{-1} l' \tag{2.3}$$

and

$$X(\Upsilon) = l(\Upsilon) \Sigma^{-1}(\Upsilon) l'(\Upsilon) , \tag{2.4}$$

where l' is the transpose of l and Σ^{-1} is the inverse of Σ. We have

$$l'(\Upsilon) = \Upsilon l' , \tag{2.5}$$

$$l(\Upsilon) = l\,\Upsilon', \tag{2.6}$$

and

$$\Sigma^{-1}(\Upsilon) = \Upsilon'^{-1}\,\Sigma^{-1}\,\Upsilon^{-1}. \tag{2.7}$$

Hence, from (2.4) – (2.7) we obtain

$$X(\Upsilon) = l\,\Upsilon'\,\Upsilon'^{-1}\,\Sigma^{-1}\,\Upsilon^{-1}\,\Upsilon\,l = l\,\Sigma^{-1}\,l' = X, \tag{2.8}$$

and Lemma 2.1 is proved.

Let $\theta^* = (\theta_1^*, \ldots, \theta_p^*)$ be a parameter point different from θ and denote by l_g^*, X^*, and ξ_{kl}^* the expressions we obtain from l_g, X, and ξ_{kl}, respectively, by substituting θ^* for θ. Now we shall prove the following lemma.

LEMMA 2.2. *If θ^* is a point such that $\xi_{kl}^* = \xi_{kl}$ $(k, l = 1, \ldots, K)$, then there exists a nonsingular transformation Υ such that*

$$\xi_{kl}^* = \xi_{kl}(\Upsilon) \qquad (k, l = 1, \ldots, K). \tag{2.9}$$

Proof: From $\xi_{kl}^* = \xi_{kl}$ it follows that $X^* = X$ identically in $x_1, \ldots, x_K$. Thus we have

$$X^* = \sum\sum \sigma^{*gh}\, l_g^*\, l_h^* = \sum\sum \sigma^{gh}\, l_g\, l_h = X. \tag{2.10}$$

Since $l_1, \ldots, l_G$ are independent linear forms and since $[\sigma_{gh}]$ is nonsingular, the rank of the quadratic form X is equal to G. Hence, the rank of X^* is also equal to G and, therefore, $l_1^*, \ldots, l_G^*$ are independent linear forms. From this and (2.10) it follows that each linear form l_g^* is a linear combination of the forms $l_1, \ldots, l_G$. Hence there exists exactly one nonsingular transformation Υ such that

$$l_g(\Upsilon) = l_g^* \qquad (g = 1, \ldots, G). \tag{2.11}$$

From Lemma 2.1 it follows that

$$\sum\sum \sigma^{gh}(\Upsilon)\, l_g(\Upsilon)\, l_h(\Upsilon) = \sum\sum \sigma^{gh}\, l_g\, l_h. \tag{2.12}$$

From (2.10), (2.11), and (2.12) we obtain

$$\sum\sum \sigma^{gh}(\Upsilon)\, l_g^*\, l_h^* = \sum\sum \sigma^{*gh}\, l_g^*\, l_h^*. \tag{2.13}$$

Hence

$$\sigma^{gh}(\Upsilon) = \sigma^{*gh} \qquad (g, h = 1, \ldots, G), \tag{2.14}$$

and, therefore,

$$\sigma_{gh}(\Upsilon) = \sigma^*_{gh} \qquad (g, h = 1, \ldots, G). \tag{2.15}$$

1

Since (2.11) implies that $\alpha_{gk}(\Upsilon) = \alpha^*_{gh}$, Lemma 2.2 is proved.

The coefficients ξ_{kl} $(k, l = 1, \ldots, K)$ depend, of course, on the parameter point θ. To make this evident, we shall occasionally replace ξ_{kl} by $\xi_{kl}(\theta)$, and $\xi_{kl}(\Upsilon)$ by $\xi_{kl}\{\theta(\Upsilon)\}$. Since $\xi_{kl}\{\theta(\Upsilon)\} = \xi_{kl}(\theta)$, we shall say that the functions $\xi_{kl}(\theta)$ are invariant under nonsingular transformations Υ.

Let $F(\theta)$ be a function of θ. We shall say that $F(\theta)$ is invariant under nonsingular transformations if for any nonsingular transformation Υ we have $F\{\theta(\Upsilon)\} = F(\theta)$. Clearly, if $F(\theta)$ is a function of $\xi_{11}(\theta)$, $\xi_{12}(\theta)$, ..., $\xi_{KK}(\theta)$, then $F(\theta)$ is invariant under nonsingular transformations. We shall show that the converse is also true. Let $F(\theta)$ be a function such that $F\{\theta(\Upsilon)\} = F(\theta)$ for all nonsingular transformations Υ. Suppose that $F(\theta)$ is not a function of $\xi_{11}(\theta)$, $\xi_{12}(\theta)$, ..., $\xi_{KK}(\theta)$. Then there exist two points θ'' and θ''' such that

$$\xi_{kl}(\theta'') = \xi_{kl}(\theta''') \qquad (k, l = 1, \ldots, K) \tag{2.16}$$

and

$$F(\theta'') \neq F(\theta'''). \tag{2.17}$$

From Lemma 2.2 and (2.16) it follows that there exists a nonsingular transformation Υ such that

$$\theta''(\Upsilon) = \theta'''. \tag{2.18}$$

But then

$$F(\theta'') \neq F\{\theta''(\Upsilon)\}, \tag{2.19}$$

which contradicts our assumption about $F(\theta)$. Hence $F(\theta)$ must be a

function of $\xi_{11}(\theta)$, $\xi_{12}(\theta)$, ..., $\xi_{KK}(\theta)$. Thus, the functions $\xi_{11}(\theta)$, $\xi_{12}(\theta)$, ..., $\xi_{KK}(\theta)$ form a fundamental set of invariants.

3. *Necessary and Sufficient Conditions for the Identification of a Coordinate θ_p of a Parameter Point θ*

Let θ be a parameter point satisfying the a priori conditions (1.1). And further, let $\theta^* = (\theta_1^*, \ldots, \theta_p^*)$ be an unknown parameter point and consider the equations in θ^*:

$$\xi_{kl}(\theta^*) = \xi_{kl}(\theta) \qquad (k, l = 1, \ldots, K) \tag{3.1}$$

and

$$\varphi_r(\theta^*) = 0 \qquad (r = 1, \ldots, R) \text{ (a priori conditions).} \tag{3.2}$$

The following two theorems are immediate consequences of Lemmas 2.1 and 2.2.

THEOREM 3.1. *A necessary and sufficient condition that θ_p be identifiable is that the equations* (3.1) *and* (3.2) *in the unknowns θ_1^*, ..., θ_P^* should admit of only a finite number of solutions for θ_p^*.*

THEOREM 3.2. *A necessary and sufficient condition that θ_p be locally identifiable is that there exists a finite neighborhood ω of θ such that for any solution θ^* in ω of the equations* (3.1) *and* (3.2) *we have $\theta_p^* = \theta_p$.*

In what follows in this section we shall assume that the R equations (3.2) have unique solutions in R unknowns, i.e., in R coordinates of θ^*. We may assume without loss of generality that these R coordinates are the last ones, i.e., θ_{P-R+1}^*, ..., θ_P^*. Thus, equations (3.2) can be written as

$$\theta_p^* = \psi_p(\theta_1^*, \ldots, \theta_{P-R}^*) \qquad (p = P - R + 1, \ldots, P). \tag{3.3}$$

We shall assume that the functions ψ_p admit of continuous first-order partial derivatives. For any parameter point $\theta = (\theta_1, \ldots, \theta_P)$ in the P-dimensional parameter space we shall denote by $\bar{\theta}$ the parameter point in the $(P - R)$-dimensional space obtained from θ by omitting the last R coordinates, i.e., $\bar{\theta} = (\theta_1, \ldots, \theta_{P-R})$.

Denote by $\bar{\xi}_{kl}(\bar{\theta})$ the function we obtain from $\xi_{kl}(\theta)$ by substituting $\psi_p(\bar{\theta})$ for θ_p $(p = P - R + 1, \ldots, P)$. Then the system of equations (3.1) and (3.2) is equivalent to the system

$$\bar{\xi}_{kl}(\bar{\theta}^*) = \bar{\xi}_{kl}(\bar{\theta}) \qquad (k, l = 1, \ldots, K) \tag{3.4}$$

and

$$\theta_p^* = \psi_p(\bar{\theta}^*) \qquad (p = P - R + 1, \ldots, P). \tag{3.5}$$

Denote the $(P - R)$-dimensional parameter space by $\bar{\Omega}$. For any point $\bar{\theta}$ of $\bar{\Omega}$ we shall denote by $\Delta(\bar{\theta})$ the Jacobian of the functions $\bar{\xi}_{11}(\bar{\theta})$, $\bar{\xi}_{12}(\bar{\theta})$, ..., $\bar{\xi}_{KK}(\bar{\theta})$ taken at the point $\bar{\theta}$. A point $\bar{\theta}$ of $\bar{\Omega}$ will be called regular if the following condition is satisfied: Any minor of the Jacobian of the $K^2 + P - R$ functions $\bar{\xi}_{11}(\bar{\theta})$, $\bar{\xi}_{12}(\bar{\theta})$, ..., $\bar{\xi}_{KK}(\bar{\theta})$, $d_p(\bar{\theta}) = \theta_p$ $(p = 1, \ldots, P - R)$ is either unequal to zero at $\bar{\theta}$ or is *identically* zero in some finite neighborhood of $\bar{\theta}$.

THEOREM 3.3. *Let $\bar{\theta}^0$ be a regular point and denote by $\delta_p(\bar{\theta})$ the Jacobian of the $K^2 + 1$ functions $\bar{\xi}_{11}(\bar{\theta})$, $\bar{\xi}_{12}(\bar{\theta})$, ..., $\bar{\xi}_{KK}(\bar{\theta})$, $d_p(\bar{\theta}) = \theta_p$ for any value of p satisfying $p \leq P - R$. A necessary and sufficient condition that θ_p be locally identifiable for any point $\bar{\theta}$ in a finite neighborhood $\bar{\theta}^0$ is that the rank of $\Delta(\bar{\theta}^0)$ be equal to the rank of $\delta_p(\bar{\theta}^0)$.*

Proof: Since $\bar{\theta}^0$ is a regular point, a necessary and sufficient condition that $\bar{\theta}_p$ be a single valued function of $\bar{\xi}_{11}(\bar{\theta})$, $\bar{\xi}_{12}(\bar{\theta})$, ..., $\bar{\xi}_{KK}(\bar{\theta})$ in a finite neighborhood of $\bar{\theta}^0$ is that the rank of $\Delta(\bar{\theta}^0)$ be equal to that of $\delta_p(\bar{\theta}^0)$. Theorem 3.3 follows from this and the fact that the functions $\bar{\xi}_{11}(\bar{\theta})$, $\bar{\xi}_{12}(\bar{\theta})$, ..., $\bar{\xi}_{KK}(\bar{\theta})$ form a fundamental set of invariants.

IV. GENERALIZATION OF THE CONCEPT OF IDENTIFICATION[1]

BY LEONID HURWICZ

1. *Introduction*

1.0. A model[2] may or may not possess the property of being *structure-identifying* (or, for briefness, *identifying*). When the model is *not* structure-identifying, no amount of empirical information will make it possible to determine the *structure*[3] of the system under investigation. Thus it is desirable to deal with identifying models when the knowledge of the structure is needed; this need frequently arises when forecasts for policy purposes are to be made.

There are several kinds and "degrees" of identification power.[4] *Totally unique identification power* ("totally" is usually omitted) implies that if an infinite sample of observations were available at most one structure consistent with the model can be made to

[1]Part of the work on this paper was done in 1945–46 during the author's tenure of the Guggenheim Memorial Fellowship.
Thanks are due to Tjalling Koopmans and Herman Rubin for valuable suggestions.

[2]Say a particular theory of the "business cycle."

[3]I.e., the set of numerical values of the parameters characterizing the model; e.g., the numerical values of the marginal propensity to consume or of the effects that profits may have on investment, etc.

[4]An identifying model is said to possess identification power.

"fit" a given set of observational data; in a finite sample only one "best (in some given sense) fit" (estimate) consistent with the model is possible. When more than one structure is compatible with the observed data, it will, in general, so happen that all structures compatible with a given set of data will have a certain property in common.[1] Then the model is said to be *uniquely identifying with regard to that property*. In such a case we speak of *partially* unique identification power of the model, unless the model is uniquely identified with regard to *all* its properties so that *totally* unique identification power, as already defined, is present.

A model lacking unique identifying power with regard to a given property may nevertheless possess *complete* identifying power with regard to that property if the set of structures compatible with any given data is finite or, at worst, denumerably infinite.

In what follows these concepts are defined in a more rigorous manner; a partial summary is given in section 7. The definitions given differ from those found elsewhere[2] in that they are independent of the nature of the models and the structures involved. In particular, linearity of the "structural relations" and the normality of the disturbance distribution are not assumed in the present paper.[3] It is not necessary that either the structural relations or the disturbance distributions should have a parametric form. In fact, the definitions given would apply to situations where structures are defined in terms other than those of "structural relations" and disturbance distributions.

One may hope that these generalizations will be of some help in clarifying the logic of the identification problems; they are not a substitute for the study of conditions under which specified models of practical importance possess one of the various types of identification power.

2. *Structures and Models*

2.1.1. The cumulative probability distribution G of the ob-

[1] E.g., just the consumption functions may be uniquely determined, but not the investment function; it might even conceivably happen that the effect of income on consumption should be identified, but not the effect on consumption of the assets held.

[2] E.g., [II-2].

[3] Illustrative examples have, however, been provided which are based on normal linear, hence parametric, systems. This may facilitate the task of relating theory and results already in existence to the generalizations of this note.

served variate (vector) x may be regarded as produced by an operation (usually a transformation) $\mathfrak{I}$ performed on the distribution F of the (nonobservable) disturbance (vector) u, i.e,.

$$G = \mathfrak{I} \to F. \tag{1}$$

2.1.2. As an illustration, consider the usually treated parametric case where $\mathfrak{I}$ is a linear transformation and F (hence also G) is normal. Adopting two simplifying assumptions[1], viz., that the mean of u is zero and that the predetermined variables are absent, one may in this case represent F and G by their respective covariance matrices Σ^u and Σ^x while $\mathfrak{I}$ is represented by the inverse A^{-1} of the structural coefficient matrix[2] A. We then have, as a counterpart of (1), the equation

$$\Sigma^x = A^{-1}\,\Sigma^u\,(A^{-1})', \tag{1'}$$

where Σ^x is produced by performing the operation $A^{-1}\{\ldots\}(A^{-1})'$ on Σ^u. [The operation consists in premultiplying by A^{-1} and postmultiplying by $(A^{-1})'$.]

2.2. It will be noted that (1) does not require a parametric representation for F, G, or $\mathfrak{I}$. It is sufficient that F and G should be functions of u and x, respectively, and that these functions should belong to the class of distribution functions.

$\mathfrak{I}$ may be any well-defined operation carrying one distribution function into another.

u and x need not be of the same dimensionality.[3]

2.3. We shall refer to

$$S = (F, \mathfrak{I}) \tag{2}$$

as a structure. An a priori postulated class $\mathfrak{S}_0$ of structures S that is

---

[1]The theory developed in this note is not subject to these restrictions. In particular, it is valid for systems containing fixed variates as well lagged endogenous variables.

[2]In terms of operations on variables (rather than on distributions) we have $Ax = u$, or, if the transformation need not be linear, $\Phi(x) = u$. In the latter case $\mathfrak{I}$ is represented by Φ^{-1}. This is the symbol used in [1].

[3]When the dimensionality of u exceeds that of x we are, except for trivial cases, dealing with *nonadditive disturbances*. Cf. [XVIII].

a proper subset of the class $\mathfrak{S}$ of all structures is called a model.[1]

When (1) holds, the structure S, as defined in (2), is said to *generate* the distribution G (written $S :. G$); G is then said to be *generated by* S (written $G .: S$).

3. *Structural Estimation and Identification*

3.1. The process of *structural estimation* consists in estimating G and then obtaining S generating that particular G. (The two stages may be combined into one computational process.)

In general, more than one structure generating a given G can be found. Hence structural estimation would lead to a class of structures, say $\mathfrak{S}_0$, containing all structures S that could have generated the given G. However, when certain additional assumptions are made with regard to S – these are the "identifying restrictions" – the class $\mathfrak{S}_0$ may be narrowed down to a proper subset $\mathfrak{S}_{00}$ of $\mathfrak{S}_0$. (If the identifying restrictions exclude a priori all structures not in $\mathfrak{S}_1$, say, we have $\mathfrak{S}_{00} = \mathfrak{S}_0\mathfrak{S}_1$.) When $\mathfrak{S}_{00}$ is denumerably infinite or contains a finite number N of elements, we speak of *complete identification power of the model: unique* identification power if $N = 1$, *multiple* if $N > 1$; when $\mathfrak{S}_{00}$ is nondenumerable, we speak of *incomplete identification power.*

3.2.1. The nature of the additional (identifying) assumptions mentioned in section *3.1* will now be stated and a more explicit definition of identification power provided.

3.2.2. The identifying assumptions postulate that the structure S which has generated a given G belongs to a certain class $\mathfrak{S}_1$ of structures.[2]

The elements of $\mathfrak{S}_1$ may be distinguished from those of $\operatorname{comp}\mathfrak{S}_1 \equiv \mathfrak{S} - \mathfrak{S}_1$ (where $\mathfrak{S}$ is, as before, the class of all structures) through restrictions on F, $\mathfrak{I}$, both F and $\mathfrak{I}$, or, more generally, through a relation to be satisfied by F and $\mathfrak{I}$. In the parametric

---

[1] Thus a structure is defined by one completely specified distribution F and one completely specified operation $\mathfrak{I}$. In the parametric case treated above, section *2.1.2*, the structure is given if both matrices Σ^{uu} and A are numerically given. On the other hand, postulating, say, zeros in certain parts of either matrix without restricting other elements of these matrices defines a whole class of structures, i.e., a model.

[2] Thus, by a previous definition, $\mathfrak{S}_1$ is a model.

case used for illustrative purposes (cf. section 2.1.2) restrictions are imposed on $\mathfrak{A}$, Σ^u, or both.

4. *Fundamental Definition of Identification Power of a Model*

We shall now define the class $\mathfrak{G}_1$ of all G's generated by the elements of $\mathfrak{S}_1$ (the class of all structures not excluded by the identifying restrictions as defined in the preceding section), so that symbolically

$$(3.1) \qquad G_1 \in \mathfrak{G}_1 : \supset : \exists\, S_1 \in \mathfrak{S}_1 \ \ni \cdot\ S_1 :. \ G_1$$

(for every element G_1 of $\mathfrak{G}_1$ there exists a structure S_1 in $\mathfrak{S}_1$ such that S_1 generates G_1) and

$$(3.2) \qquad S_1 \in \mathfrak{S}_1 : \supset : \exists\,|\, G_1 \in \mathfrak{G}_1 \ \ni \cdot S_1 :. \ G_1$$

(for every element S_1 of $\mathfrak{S}_1$ there exists *one and only one* G_1 in $\mathfrak{G}_1$ such that G_1 is generated by S_1). It may so happen that $\mathfrak{S}_1$ has the property that for any element G_1 in $\mathfrak{G}_1$ there exists *one and only one* S_1 in $\mathfrak{S}_1$ such that S_1 generates G_1, i.e.,

$$(4) \qquad G_1 \in \mathfrak{G}_1 : \supset : \exists\,|\, S_1 \in \mathfrak{S}_1 \ \ni \cdot S_1 :. \ G_1 .$$

When (4) holds we say that the model $\mathfrak{S}_1$ is *uniquely identifying*.[1] If for any G_1 in $\mathfrak{G}_1$ the set $\mathfrak{S}_{1,G_1}$ of all S_1 in $\mathfrak{S}_1$ which generate G_1 is finite or denumerably infinite, the model $\mathfrak{S}_1$ is said to be *completely identifying*. When $\mathfrak{S}_{1,G_1}$ has a finite number of elements for all G_1 in $\mathfrak{G}_1$, $\mathfrak{S}_1$ is said to be *multiply* or *uniquely identifying* depending on whether or not $N > 1$ for some G_1. (This definition of unique identification power is, of course, equivalent to that given earlier.)

5. *Models Identifying over Certain Submodels*

5.1. The above definition will now be generalized so as to take care of a situation frequently arising in applications. Sup-

[1] It will be noted that this definition of identification could have been stated in identical terms for definitions of S and of the relation $S :. G$ entirely different from those given in section 2.3.

pose one is willing to postulate that the unknown structure is an element of the model $\mathfrak{S}_1$. We write

$$\mathfrak{S}_1 = \mathfrak{S}_{11} + \mathfrak{S}_{12} , \tag{5}$$

where $\mathfrak{S}_{11}$ is defined as the set of *all*[1] elements of $\mathfrak{S}_1$ that possess the following property: if S_{11} is an element of $\mathfrak{S}_{11}$ and if S_{11} generates the distribution G_{11} in $\mathfrak{G}_1$ (where $\mathfrak{G}_1$ is defined as before), then there exists no other element S_1 of $\mathfrak{S}_1$ that also generates G_{11}. It follows that if S_{12} is an element of $\mathfrak{S}_{12}$ and if S_{12} generates G_{12}, there exists at least one other element S_{12} of $\mathfrak{S}_{12}$ that also generates G_{12}; however, G_{12} cannot be generated by any elements of $\mathfrak{S}_{11}$. In such a case we shall say that the (sub-)model $\mathfrak{S}_{11}$ is *uniquely identifying in the model* $\mathfrak{S}_1$. Alternatively, it will be said that $\mathfrak{S}_1$ is *uniquely identifying over* $\mathfrak{S}_{11}$. The definition of identification power given in section 4 is consistent with the one just given. In section 4, $\mathfrak{S}_1$ was identifying over (or in) itself, so that in the present notation $\mathfrak{S}_1 = \mathfrak{S}_{11}$ and $\mathfrak{S}_{12}$ is empty.

5.2. One may find it helpful to use a graph depicting the situation described in the previous paragraph. In Figure 1 the lower axis corresponds to the set $\mathfrak{G}$ of all G's, while the upper one corresponds to the set $\mathfrak{S}$ of all S's. A point on the lower axis corresponds to a given G, a point on the upper axis to a given S. A

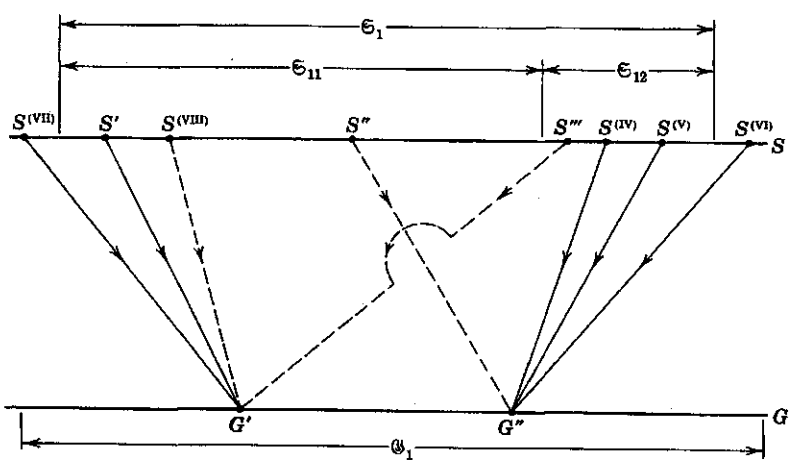


Figure 1.

[1] It is important to note the word "all" in this definition.

solid line (with an arrow pointing downward) connects a structure S with the distribution G which that particular structure generates. Broken lines (with arrows) indicate "prohibited" generating relationships in the case where $\mathfrak{S}_1$ is uniquely identifying over $\mathfrak{S}_{11}$.

Thus it can be seen that, when $\mathfrak{S}_1$ is uniquely identifying over $\mathfrak{S}_{11}$, G' which is assumed to be generated by S' cannot be generated by S''' or $S^{(\text{viii})}$, although it can be generated by $S^{(\text{vii})}$. On the other hand, G'' can be generated by both $S^{(\text{iv})}$ and $S^{(\text{v})}$ as well as $S^{(\text{vi})}$, although it cannot be generated by S''.

Similar diagrams could be constructed for other types of identification.

5.3. As an illustration of the preceding concepts, consider the parametric case $Ax = u$, where $x = (x_1\ x_2\ x_3\ x_4)$, $u = (u_1\ u_2\ u_3\ u_4)$, and $A = [\alpha_{ij}]$, $i, j = 1, 2, 3, 4$. Now let $\mathfrak{S}_1$ be defined by the following zeros in Σ^u:

$$\begin{bmatrix} \times & \times & 0 & 0 \\ \times & \times & 0 & 0 \\ 0 & 0 & \times & \times \\ 0 & 0 & \times & \times \end{bmatrix}$$

and by the following zeros in A:

$$\begin{bmatrix} 1 & \times & \times & 0 \\ \times & 1 & 0 & \times \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where $\times$ indicates an unrestricted element and 1 indicates a nonvanishing element which has been equated to unity by the normalization rule chosen. Then $\mathfrak{S}_{12}$ is defined by zeros for α_{13} or α_{24} or both, so that, for S in $\mathfrak{S}_{12}$, A is one of the following:

$$\begin{bmatrix} 1 & \times & 0 & 0 \\ \times & 1 & 0 & \times \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & \times & \times & 0 \\ \times & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & \times & 0 & 0 \\ \times & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The requirement that the structure should belong to $\mathfrak{S}_{11}$ is in this case the "rank condition" treated in [II-2.2.2 and 2.2.3].

5.4.1. $\mathfrak{S}_{11}$ in section *5.1* is said to be the *maximal set of* $\mathfrak{S}_1$ *of multiplicity* 1.[1]

$\mathfrak{S}_1$ will be said to be *multiply identifying* (of multiplicity N) *over* $\mathfrak{S}_{11}$ if 1) for every G_{11} in $\mathfrak{G}_1$, the subset $\mathfrak{S}_{1,G_{11}}$ of all structures S_1 in $\mathfrak{S}_{11}$ which generate G_{11} contains a finite number N of elements (possibly less than N for some, but not all, G_{11}), and 2) $\mathfrak{S}_{11}$ contains all structures generating G's with property 1). $\mathfrak{S}_{11}$ is then called the maximal set of $\mathfrak{S}_1$ of multiplicity N.

Similarly, $\mathfrak{S}_1$ is said to be *denumerably identifying* when the above definition holds with denumerable infinity substituted for N. Thus the maximal set whose multiplicity is denumerable infinity is defined as in the case of finite multiplicity.[2]

Now let $\{\mathfrak{S}_1^{(M_N)}\}$, $N = 2, 3, \ldots$, be the (finite or denumerably infinite) sequence of the maximal sets of $\mathfrak{S}_1$ of multiplicities N, and write $\mathfrak{S}_1^{(1)}$, $\mathfrak{S}_1^{(D)}$ for the maximal sets of $\mathfrak{S}_1$ of multiplicities 1 and denumerable infinity, respectively. Then $\mathfrak{S}_1$ is said to be *multiply* identifying over $\mathfrak{S}_1^{(M)} = \sum_{1<N<\infty} \mathfrak{S}_1^{(M_N)}$, *finitely* identifying over $\mathfrak{S}_1^{(F)} = \mathfrak{S}_1^{(1)} + \mathfrak{S}_1^{(M)}$, *completely* identifying over $\mathfrak{S}_1^{(C)} = \mathfrak{S}_1^{(F)} + \mathfrak{S}_1^{(D)}$, and *incompletely* identifying over $\mathfrak{S}_1^{(I)} = \mathfrak{S}_1 - \mathfrak{S}_1^{(C)}$. (Nondenumerable infinity is defined as the multiplicity in the case of incomplete identification power.)

5.4.2. It is convenient to have a term for the property of being a subset of a given maximal set.

Let $\mathfrak{S}_1'$ be any subset of $\mathfrak{S}_1^{(1)}$. We then say that $\mathfrak{S}_1'$ is uniquely identifying *over and possibly beyond* $\mathfrak{S}_1'$. (If it is known that $\mathfrak{S}_1'$ is a proper subset of $\mathfrak{S}_1^{(1)}$, we omit "possibly." If it is known that $\mathfrak{S}_1' = \mathfrak{S}_1^{(1)}$, we omit "and possibly beyond.") Clearly, if $\mathfrak{S}_1$ is

[1] There exist smaller sets of structures with no two elements of the set generating the same G_1, but $\mathfrak{S}_{11}$ contains all such sets; hence the term *maximal*.

[2] It is important to note that the choice of $\mathfrak{S}_1$ determines all the maximal sets. Some of the maximal sets may, of course, happen to be empty.

uniquely identifying over and possibly beyond $\mathfrak{S}_1'$, it is so identifying over any subset of $\mathfrak{S}_1'$. Similar language may be used for multiple identification power and identification power of other multiplicities.

This terminology helps formulate the correspondence between the language of this note and that of [II]. When [II] says that a given set of equations is uniquely identified over some subset Ω_{11} of the parameter space Ω, this can be expressed in the language of the present note as saying that the model is uniquely identifying *over and possibly beyond* the set $\mathfrak{S}_{11}$ corresponding to Ω_{11}.

The language of [II] does not specify whether Ω_{11} is a maximal set.

6. *Identification Power with Regard to a Criterion*

6.1. Since often only some of the properties of the structure are of interest, it is desirable to broaden the concept of identification so as to cover situations where some properties of S can be determined uniquely from the knowledge of G while other properties of S perhaps cannot.

6.2. Consider a criterion C which establishes a partition of the class $\mathfrak{S}$ of all structures into a system of nonoverlapping subclasses $\mathfrak{S}^{(i)}$ $(\sum_i \mathfrak{S}^{(i)} = \mathfrak{S})$; this partition, in general, need not be finite or denumerable. By definition, if both S_1 and S_2 belong to $\mathfrak{S}^{(i)}$ they are indistinguishable with regard to C.

Let there be given a model $\mathfrak{S}_1$ and its not necessarily proper subset $\mathfrak{S}_{11} \subseteq \mathfrak{S}_1$. Let $\mathfrak{G}_1$ be the class of all G's generated by the elements of $\mathfrak{S}_1$ and let G_1 in $\mathfrak{G}_1$ be generated by S_{11} in $\mathfrak{S}_{11}$. Consider the set $\mathfrak{S}_{1,G_1}$ of all structures that generate G_1. Then if, for every G_1 in $\mathfrak{G}_1$, all the elements of $\mathfrak{S}_{1,G_1}$ belong to the same subclass, say $\mathfrak{S}^{(i_0)}$, it is said that $\mathfrak{S}_{11}$ is *uniquely identifying* in $\mathfrak{S}_1$ *with regard to* C.

6.3. It may again be helpful to present the matter diagrammatically. In Figure 2 solid generating lines are the permissible ones while the broken ones are prohibited in the case of unique (partial) identification with regard to the criterion on which the partition is based.

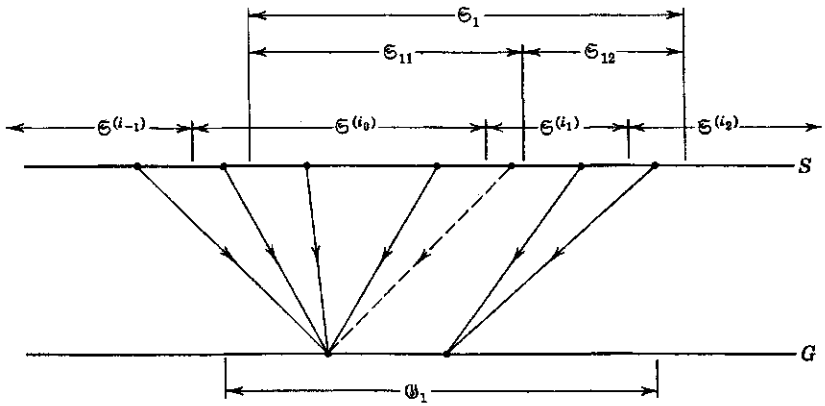


Figure 2.

6.4. As an illustrative example we use the one given in section 5.3 except that for S in $\mathfrak{S}_1$ we now have A with the following zeros:

$$\begin{bmatrix} 1 & \times & 0 & 0 \\ \times & 1 & 0 & \times \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

$\mathfrak{S}_{12}$ is here given by a matrix A with $\alpha_{24} = 0$, i.e.,

$$\begin{bmatrix} 1 & \times & 0 & 0 \\ \times & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix};$$

here as criterion C we have the values of the elements of the first row of A. This is the familiar case of only one of the equations of the system being identified, to use the language of [II].

With other identifying restrictions, the value of a single element of A could have been chosen as C, cf. [III].

6.5. Multiplicities other than 1 of identification power with regard to a given criterion are defined by analogy with earlier sections of this note.

Now let $\mathfrak{S}_1$ possess identification power of a given multiplicity over $\mathfrak{S}_{11}$ with regard to *all* criteria C. Then $\mathfrak{S}_1$ is said to possess *total* identification power over $\mathfrak{S}_{11}$ of the appropriate multiplicity. ("Total" may be omitted where no danger of confusion exists.)

On the other hand, let $\mathfrak{S}_1$ possess identification power over $\mathfrak{S}_{11}$ of varying multiplicities with regard to different criteria. Then we say that $\mathfrak{S}_1$ has *total* identification power over $\mathfrak{S}_{11}$ of the *highest* multiplicity and *partial* of all other multiplicities.

Thus, for instance, if we say that $\mathfrak{S}_1$ is *partially* denumerably identifying over $\mathfrak{S}_{11}$, the following is implied: over $\mathfrak{S}_{11}$, $\mathfrak{S}_1$ is denumerably identifying with regard to some criteria and *incompletely* with regard to some others.[1] (There may or may not be some criterion with regard to which $\mathfrak{S}_1$ is uniquely or multiply identifying over $\mathfrak{S}_{11}$.)

As in section *5.4.1*, $\mathfrak{S}_1$ can be divided into the maximal sets $\mathfrak{S}_1^{(1)}$, $\mathfrak{S}_1^{(M)}$, ..., with regard to a given criterion C. This process is too obvious to require detailed description.

7. Summary

At this stage it may be desirable to provide a verbal summary of the more important concepts introduced.

A model is said to be *uniquely identifying with respect to a given criterion* if all structures compatible with the probability distribution of the observations are indistinguishable on the basis of that criterion.

A model is said to be *multiply identifying with respect to a given criterion* if all structures compatible with the probability distribution of the observations can be grouped into a *finite* number of classes such that within each class structures are indistinguishable on the basis of that criterion. The number N of such groups is the *multiplicity* of identification power of the model.

A model is said to be *denumerably identifying with respect to a given criterion* if all structures can be grouped into a *denumer-*

[1] If it were completely identifying with regard to all criteria, and denumerably with regard to some, it would be (totally) denumerably identifying.

ably infinite number of such classes.

A model uniquely or multiply identifying with respect to a given criterion is said to be *finitely identifying with respect to that criterion.*

A model *finitely* or *denumerably* identifying with respect to a given criterion is said to be *completely identifying with respect to that criterion.*

A model *not* completely identifying with regard to a given criterion is said to be *incompletely identifying with regard to that criterion.*

The multiplicity of identification power of a model with regard to a given criterion is, in general, defined as 1, N, denumerable infinity, or nondenumerable infinity when the identification power is unique, multiple (of multiplicity N as defined above), denumerable, or incomplete, respectively. Multiplicity N is *higher* than 1, etc.[1]

If a model possesses identification power of a given multiplicity with respect to *all* criteria, it is said to possess *total* identification power of that multiplicity. Thus, if a model is uniquely identifying with respect to all criteria, it is said to be totally-uniquely identifying, or – more simply – uniquely identifying.

When a model is uniquely identifying, there exists only one structure compatible with the probability distribution of the observations.

If a model possesses identification power of different multiplicities with respect to different criteria, it is said to possess *total* identification power of the *highest* multiplicity and partial identification power of all other multiplicities.

Thus, if a model is uniquely identifying with respect to one set of properties, multiply with respect to another, and denumerably with respect to a third set of properties, then the model is said to be totally denumerably identifying as well as partially multiply and partially uniquely identifying.[2]

---

[1] Thus the higher the multiplicity, the more difficult it is to determine the structure from the probability distribution of the observations.

[2] When the model is said to be *partially* uniquely identifying with regard to a given criterion, "partially" serves as a warning that there exist other criteria with regard to which the model possesses identification power of a higher multiplicity, i.e., multiple, denumerable, or even incomplete, identification power.

8. *General Remarks*

8.1. In practice we are not often certain of the proper choice of identifying assumptions. In fact, if the choice of $\mathfrak{S}_1$ is made before G is known, it may happen that the structures S that generate G are not among the elements of $\mathfrak{S}_1$. In this case we obtain no S whatever.

Also, a given G will, in general, yield different structures, say S_1', S_1'', ..., corresponding to different models $\mathfrak{S}_1'$, $\mathfrak{S}_1''$, ..., even if each of them is uniquely identifying. If no a priori information permits us to choose one of these models, or, equivalently, when the model is not completely identifying, general principles of making decisions (e.g., minimization of maximum risk) may be applied. However, a solution will not always exist.[1]

8.2. Identification power, as here defined, is a property of a model (with reference to specified criteria). This in itself makes it clear why the problem of identification does not arise in prediction under unchanged structure.[2] Such prediction involves only certain properties of G, and the structure does not even enter the picture. On the other hand, for prediction under changed structure it is, in general, desirable to know S; to do this requires, in general, an identifying model.

8.3. Since the concept of identification refers to structure and to the distributions thus generated, it should be clear that the problem of identification is entirely independent of any sampling aspects of structural estimation. The latter can only refer to the relationship of the estimate of G obtained from the sample to the true value of G. The former refers to the possibility of ascribing a unique S to a given G.

---

[1] Cf. [Hurwicz, 1946].

[2] Cf. [VI].

V. REMARKS ON FRISCH'S CONFLUENCE ANALYSIS AND ITS USE IN ECONOMETRICS

BY TRYGVE HAAVELMO

Confluence Analysis [Frisch, 1934] was written in part as a protest against a mechanical and uncritical use of the classical least-squares method to estimate demand functions and other economic relations. Frisch pointed out that economic data, in general, do not satisfy the conditions required to justify the use of the classical method of least-squares. It is of course true that the strict conditions of a theoretical model perhaps never are exactly fulfilled in any observational material. And good theoretical models should be able to absorb moderate discrepancies between model and facts without the inference drawn becoming valueless or nonsensical. In many respects the classical model of the least-squares method fulfills this requirement. One important exception, however, is the case where the "independent variables" are themselves highly intercorrelated while, at the same time, they are subject to "errors" which tend to hide such intercorrelation. Then the application of the classical method of least-squares might give unreliable or even nonsensical results. I shall try to explain this a little more explicitly.

1. *Confluence Analysis and the Markoff Theorem on Least-Squares*

Let x'_{it}, $i = 1, 2, \ldots, n$, $t = 1, 2, \ldots, T$, denote a system of values of nT variables, fixed in repeated samples, and interconnected by the linear relation

$$x'_{1t} = \alpha_2 x'_{2t} + \alpha_3 x'_{3t} + \ldots + \alpha_n x'_{nt}, \qquad t = 1, 2, \ldots, T. \tag{1.1}$$

(The case of a separate constant term is covered by setting, e.g., x'_{nt} equal to 1, identically, for all values of t.) We make the

following assumptions [David and Neyman, 1938]:

ASSUMPTION I. *The* $(n-1)T$ *quantities* $x'_{2t}, x'_{3t}, \ldots, x'_{nt}$, $t = 1, 2, \ldots, T$, *are known observations while, instead of the values of* x'_{1t}, *we can only observe the values of* x_{1t} *defined by*

$$x_{1t} = x'_{1t} + x''_{1t}, \qquad t = 1, 2, \ldots, T, \tag{1.2}$$

where x''_{1t} *for each value of* t *is a random variable that cannot be observed separately..*

ASSUMPTION II. *The* T *random variables* x''_{1t}, $t = 1, 2, \ldots, T$, *are independent random variables with* $\mathcal{E}(x''_{1t}) = 0$ *and* $\mathcal{E}(x''^2_{1t}) = \sigma^2$ *for all values of* t.

ASSUMPTION III. *The* $(n-1)$*-rowed and* T*-columned matrix* $[x'_{it}]$, $i = 2, 3, \ldots, n$, $t = 1, 2, \ldots, T$, *is of rank* $(n-1)$.

If these assumptions are fulfilled, and if $x_{11}, x_{12}, \ldots, x_{1T}$, is a sample of the T variables x_{1t}, $t = 1, 2, \ldots, T$, Markoff's theorem on least-squares states that of all the unbiased estimates of α_i that are linear in the variables x_{1t} the one that has the smallest variance is given by the value of a_i which minimizes the sum $\sum_{t=1}^{t=T} (x_{1t} - a_2 x'_{2t} - a_3 x'_{3t} - \cdots - a_n x'_{nt})^2$. Further, the variance of this estimate is given by

$$\text{variance of } a_i = \sigma^2 \frac{M'_{ii}}{M'}, \qquad i = 2, 3, \ldots, n, \tag{1.3}$$

where M' is the $(n-1)$th order determinant of the matrix $\left[\sum_{t=1}^{t=T} x'_{kt} x'_{jt}\right]$, $k, j = 2, 3, \ldots, n$, while M'_{ii} is the $(n-2)$th order principal minor obtained from this determinant by omitting the ith row and column.

If the conditions of the Markoff theorem are met by the data to which the theorem is applied, the accuracy of the estimation of the α's is shown by their variances (1.3). For example, even if Assumption III is very near to being violated, the least-squares method will still give unbiased estimates of the α's, but their

variances might then become very large. The estimate, s^2, say, of σ^2, and hence that of the variance of a_i, is not impaired by the fact that the independent variables are highly intercorrelated. Therefore, if the fact that Assumption III is near to being violated leads to very uncertain estimates of the α's, this will be shown by large values of the estimated variances of the a's. It is only when M' is exactly equal to zero that the method breaks down.

Frisch pointed out, however, that in practice, and in particular when dealing with economic statistics, it is hardly ever correct to assume that the expected values of one variable are linearly related to the *observable* values of the other variables. It is more realistic to assume that every one of the variables contains some random elements having no functional or stochastic connection with the other variables. Instead of (1.2) above, Frisch therefore assumed that all we can observe are variables x_{it} defined by

$$x_{it} = x'_{it} + x''_{it}, \tag{1.4}$$

$$i = 1, 2, \ldots, n, \qquad t = 1, 2, \ldots, T,$$

where all the variables x''_{it} are assumed to be uncorrelated and to have stochastic properties similar to those of the variables x''_{1t} in (1.2). As Frisch did not use a probabilistic approach, his "language" in describing the model differs considerably from that employed here. The stochastic interpretation of Frisch's model as outlined above is due to Koopmans [1937].

Suppose now that, by analogy, we attempt to estimate α_i in (1.1) by the value of a_i^* that satisfies

$$\sum_{t=1}^{t=T} (x_{1t} - a_2^* x_{2t} - a_3^* x_{3t} - \cdots - a_n^* x_{nt})^2 = \text{minimum}. \tag{1.5}$$

Clearly, the Markoff theorem cannot be called upon as a justification for this procedure as the conditions of the theorem are no longer met. And the resulting "estimates" of the α's obtained from (1.5) will have the following properties:

1. No matter whether Assumption III is fulfilled or not, and no matter how many observations we have, the statistics a_i^* will no longer be unbiased estimates of the α_i's. The bias will depend on

the variances of the disturbances x''_{it}, $i = 2, 3, \ldots, n$. Even if Assumption III is not fulfilled and, therefore, the α's could not possibly be determined uniquely, the procedure (1.5) will yield *some* apparently unique values of the quantities a_i^*, depending mainly on the variances of the disturbances x''_{it}.

2. If, by analogy, the variance of a_i^* is taken as

$$\text{(1.6)} \qquad \text{variance of } a_i^* = \sigma^2 \frac{M_{ii}}{M}, \qquad i = 2, 3, \ldots, n,$$

where the determinants M are obtained from the determinants M' in (1.3) by replacing x' by x, this "variance" will no longer control the accuracy with which the α's are being estimated by (1.5), because (1.6) is simply not the variance of an unbiased estimate of α_i.

Frisch was mainly interested in deriving a method of control which would reveal the particularly nonsensical results that this procedure might lead to if Assumption III of the Markoff theorem is not fulfilled. The result was his now well-known "bunch-map" method, which is a graphical method of exposing the manner in which the T sample points $(x_{1t}, x_{2t}, \ldots, x_{nt})$, $t = 1, 2, \ldots, T$, are clustered in the n-dimensional scatter diagram of the n observable variables. The most important objective of the "bunch analysis," therefore, is to reveal a possible lack of fulfillment of Assumption III of the Markoff theorem when this fact – contrary to the assumptions of the theorem – is hidden by errors in all the variables.

The method of "bunch analysis" has been much discussed, and, because of its partly subjective nature, it is no doubt open to criticism. We shall not add to this discussion here as we shall be interested in another and perhaps more fundamental question, namely the question of whether or not the *statistical model* discussed above actually represents a workable model for the type of relationships we may expect between economic variables.

2. *Confluence Analysis and Econometrics*

If the set of "systematic parts" x'_{it} defined above satisfies one or more independent linear relations *in addition* to (1.1), the set $(x'_{1t}, x'_{2t}, \ldots, x'_{nt})$, in Frisch's terminology, is said to

be multicollinear. The occurrence of multicollinearity might sometimes be due to pure accident. An alternative is that the data under observation are ruled, simultaneously, by two or more linear relations that are permanent and inherent in the economic structure that produces the data observed.

If, however, Frisch's statistical model (or, as a special case, that of Markoff) is accepted, I think it will be very difficult to construct any realistic economic theory that would imply such multicollinearity in the strict sense. In fact, I believe that the model and the method based upon it are tenable only in connection with a single-equation approach where other relations, if any, between the variables are accidental, or "given from outside." I shall try to make this clear.

Let $x_{1t}, x_{2t}, \ldots, x_{nt}$ be n observable series of economic variables. Usually, no exact functional relationship holds between such observable variables. It is much more likely that they would be interconnected through some sort of stochastic relationship. And the stochastic elements involved would usually be not only "errors of measurements" but random variables of a more fundamental nature that are characteristic elements of economic actions and decisions. If, therefore, the disturbances x''_{it} in (1.4) are thought of only as errors of measurement, the "true" variables x'_{it} would hardly ever satisfy an exact relation such as (1.1); nor could there be any case of multicollinearity in the strict sense, except by accident. In other words, if the "true" variables are thought of as economic variables which we might be able to observe if we had better statistics, then any assumption of exact linear relations between these variables is unrealistic[1]

If, nevertheless, we would require such relations to be exact, one or more of the variables x'_{it} would have to be considered as theoretical constructions for the purpose of building a model. The remainder, x''_{it}, would then no longer consist only of errors of measurements. Moreover, the behavior and, indeed, the economic meaning of these theoretical variables would then *depend radically on the whole network of stochastic economic relations between the variables studied* rather than on the particular relation (1.1) under investigation. Now, the real reason for splitting the variables into "systematic parts" and "disturbances" would seem to be, first, the idea that the systematic parts of the variables observed are the would-be "true economic variables" which according to eco-

[1] For a more extensive discussion of this subject reference is made to [Haavelmo, 1943, 1944].

nomic theory would fit exactly the theoretical relation considered, second, that this relation would be autonomous in the sense that it would hold regardless of whether or not other economic relations were fulfilled, and, third, that there might be some hope of eventually being able to experience and observe the "true" variables so that the exact relation between them would have value for prediction purposes. But a set of "systematic parts" derived from a simultaneous system of stochastic equations will in general not meet these conditions.

It might be useful to illustrate these points by a simple example. Let x_{1t}, x_{2t}, x_{3t}, $t = 1, 2, \ldots, T$, be three observable economic series, each observation x_{it} containing some error of measurement x''_{it}, the true parts $x'_{it} = x_{it} - x''_{it}$ being the observations we should actually make if there were no errors of measurement. Then any relationship assumed to exist between the systematic parts x'_{1t}, x'_{2t}, and x'_{3t} (except in the case of a bookkeeping identity or a similar, "uninteresting," relation) would have to be of a stochastic type rather than an exact functional relation. Suppose now that economic theory has led us to a relation

$$x'_{1t} = \alpha_2 x'_{2t} + \alpha_3 x'_{3t} + u_{1t}, \qquad t = 1, 2, \ldots, T, \tag{2.1}$$

and that we want to estimate the parameters α_2 and α_3, the variables u_{1t} being nonobservable random elements. Suppose further that, perhaps without the knowledge of the investigator, another, similar relation

$$x'_{1t} = \beta_2 x'_{2t} + \beta_3 x'_{3t} + u_{2t}, \qquad t = 1, 2, \ldots, T, \tag{2.2}$$

is also fulfilled. Let us further assume that the series x'_{3t} are "exogenous variables" having a fixed value for each value of t. Both x'_{1t} and x'_{2t} will then have to be stochastic variables, their stochastic properties being determined by those of the u's and the transformations (2.1) and (2.2). Let us for simplicity assume that all the $2T$ u's are normally and independently distributed and that $\mathcal{E}(u_{1t}) = \mathcal{E}(u_{2t}) = 0$ and $\mathcal{E}(u_{1t}^2) = \sigma_1^2$, $\mathcal{E}(u_{2t}^2) = \sigma_2^2$, for all values of t. We assume that the σ's are unknown parameters.

On the basis of the definitions above it is possible to split the variables x_{1t}, x_{2t}, x_{3t}, artificially, into systematic parts and disturbances as follows:

$$(2.3) \qquad x_{1t} = \left(\frac{\alpha_2 \beta_3 - \alpha_3 \beta_2}{\alpha_2 - \beta_2} x'_{3t}\right) + \left(\frac{\alpha_2 u_2 - \beta_2 u_1}{\alpha_2 - \beta_2} + x''_{1t}\right),$$

$$(2.4) \qquad x_{2t} = \left(\frac{\beta_3 - \alpha_3}{\alpha_2 - \beta_2} x'_{3t}\right) + \left(\frac{u_2 - u_1}{\alpha_2 - \beta_2} + x''_{2t}\right),$$

$$(2.5) \qquad x_{3t} = (x'_{3t}) + (x''_{3t}).$$

If the first term of the right-hand side in each of these equations is taken as the systematic part of x_{1t}, x_{2t}, x_{3t}, respectively, any two among these three systematic parts are linearly related. But the artificially constructed "disturbances" in the second term of (2.3) and (2.4) would here, in general, be stochastically dependent, which is another expression for the fact that the "true" systematic parts x'_{1t}, x'_{2t}, x'_{3t} are not multicollinear according to Frisch's definition.

It is well known that in the example discussed above *no unique estimate* exists for any of the parameters α_2, α_3, β_2, β_3, unless some additional a priori knowledge is available. In Koopmans' and Rubin's terminology [II-*1.11*] neither (2.1) nor (2.2) can be identified. But there exist two well-defined and identifiable relationships among the three variables x'_{1t}, x'_{2t}, x'_{3t}, namely the expected value of x'_{1t}, given x'_{2t} and x'_{3t}, and the expected value of x'_{2t}, given x'_{1t} and x'_{3t}. These relations are

$$(2.6) \qquad \mathcal{E}(x'_{1t} \mid x'_{2t}, x'_{3t}) = \frac{\alpha_2 \sigma_2^2 + \beta_2 \sigma_1^2}{\sigma_1^2 + \sigma_2^2} x'_{2t} + \frac{\alpha_3 \sigma_2^2 + \beta_3 \sigma_1^2}{\sigma_1^2 + \sigma_2^2} x'_{3t},$$

$$(2.7) \qquad \mathcal{E}(x'_{2t} \mid x'_{1t}, x'_{3t}) = \frac{\alpha_2 \sigma_2^2 + \beta_2 \sigma_1^2}{\alpha_2^2 \sigma_2^2 + \beta_2^2 \sigma_1^2} x'_{1t} - \frac{\alpha_2 \alpha_3 \sigma_2^2 + \beta_2 \beta_3 \sigma_1^2}{\alpha_2^2 \sigma_2^2 + \beta_2^2 \sigma_1^2} x'_{3t}.$$

Apart from bias caused by the presence of the disturbances x''_{1t}, x''_{2t}, and x''_{3t}, these two prediction equations could be estimated from the observations x_{1t}, x_{2t}, x_{3t}. Now it is easy to see that for certain values of the unknown parameters the two planes (2.6) and (2.7) might not differ very much, in which case the method of "bunch analysis," by its very nature, would not show any sign of multicollinearity. This, however, does not mean that the α's in the original equation (2.1) could be estimated. Thus, the apparent absence of multicollinearity would be no guarantee that the equation which we are seeking could actually be identified.

The purely geometric properties of a set of points in the sample space are insufficient as a basis for statistical inference. In fact, a sample of observations is just a set of cold, uninteresting numbers unless we have a theory concerning the stochastic mechanism that has produced them. To know the meaning of the results that a certain statistical method will yield we shall have to adopt a well-defined stochastic model. Then we can usually also find more efficient statistical tools than those that are not based on a probabilistic approach.

VI. PREDICTION AND LEAST SQUARES[1]

BY LEONID HURWICZ

0. INTRODUCTION

0.1. Prediction and Structural Changes

0.1.1. Predictive estimation. Among the problems confronting the statistician is that of making predictions. As an example, one may be asked to forecast next year's national income on the basis of this year's income and government spending planned for the next year. Next year's income is the *predictand,* this year's

[1]Part of the work on this paper was done in 1945 – 46 during the author's tenure of the Guggenheim Memorial Fellowship.
Some of the problems considered arose in connection with the author's research at the Institute of Meteorology at the University of Chicago in 1944.

income and next year's planned government spending are the *predictors*. The *prediction* is a statement concerning, say, the most likely value of the predictand given the predictors; or the prediction may be formulated as an interval within which the predictand is likely to fall given the predictor values. Thus, in general, *prediction is a statement about the (conditional) probability distribution of the predictand given the predictors.*[1]

In order to make the prediction, the statistician must obtain this conditional distribution on the basis of past observations. Thus, one may *estimate,* on the basis of past observations, the conditional distribution of a given year's income, with income in the preceding year and government spending planned for the given year specified.

However, the past observations can only tell what relationship between the values of the predictors and probability distribution of the predictand existed during the (past) observation period. Will the same relationship hold in the *future* period for which the prediction is to be made?

An affirmative answer implies that *no structural changes* have occurred, or are expected to occur, between the observation period and the period for which prediction is made. A *structural change*[2] is a change in any of the not directly observable ("structural") properties of the system. Thus structural changes occur if people's tastes shift or industrial productivity increases, provided neither tastes nor productivity are observed directly; should they have been observed, they could simply have been included among the variables of the system.

Hence, if *no structural changes* are expected to occur between

[1]Actually, there are three steps in the prediction process: first, a *rule* is derived which enables us to construct the probability distribution of the predictand for any given values of the predictors (the regression equation is an example of such a rule); second, the relevant predictor values are found; third, with the help of the above-mentioned rule, the given predictor values are used to construct the relevant predictand distribution.

[2]A structural change transforms one *structure* into another. An example may be given as follows: Let an equation system consist of a supply equation and a demand equation for some commodity. Then the structure is given by the parameters of these two equations ("the structural parameters") and the parameters of the distribution of the disturbances. (More rigorous definitions are given below in section *1*; they are also discussed in connection with the identification concept in [IV].) For briefness we often refer to the structure before and after the change as the "old" and "new" structures, respectively.

the observation period and the period for which prediction is to be made, all the statistician is required to do for prediction purposes is to provide a "best" (in some well-defined sense) estimate of the relationship that existed during the observation period between the probability distribution of the predictand and the values of the predictors. Once the relationship has been estimated, the forecast is made by specifying the predictor values to be substituted in that relationship and the latter will yield the appropriate (best estimate of the) probability distribution of the predictand.

The situation is fundamentally different when *structural changes are expected* to occur (or are already known to have taken place) between the observation period and the period for which prediction is to be made. Here it is not legitimate to apply to the future the past relationship between (the probability distribution of) the predictand and (the values of) the predictors: the structural changes, in general, will modify that relationship and it is this modified relationship that has to be determined. Once the modified relationship has been determined, a forecast can again be made by substituting into it the values of predictors; the relationship will then again yield (the best estimate of the probability distribution of) the predictand.

The problem is thus reduced to the *determination of the modified predictor-predictand relationship* after structural changes have taken place. How can this be accomplished? To begin with, it should be evident that if nothing is known about the nature of the expected structural changes, the new predictor-predictand relationship cannot be determined. Thus the nature of these changes must first be stated.

A structural change, by definition, transforms the "old" structure into the "new" (modified) structure. The nature of the *structural change* must be *completely specified* in the following sense: enough must be known about the change to make possible the derivation of the "new" structure *if* the "old" one were known. One example of such specification is that of the structural change consisting in, say, a productivity coefficient increasing by 10 per cent; this statement in itself does not tell us either the old or the new value of the coefficient, but it would enable us to find the new value if the old one were known.

Assuming that the structural change is completely specified, the "new" predictor-predictand relationship that is being sought may be determined in the following manner. From the available observations we estimate the ("old") joint probability of the ob-

served variates. Then, provided the model is *structure-identifying*,[1] we can determine the "old" structure, i.e., the structure during the observation period. Next, we derive the "new" structure from the "old" one. (This can be done since the structural change is completely specified.)

Then from this "new" structure we derive the "new" distribution that the observed variates will have after the structural change has occurred.[2] The predictor-predictand relationship is "embedded" in the distribution of observed variates, so that from the distribution the relationship can always be obtained. (The converse is not true.) This procedure – obtaining an estimate of the modified predictor-predictand relationship given past observations, a completely specified structural change, and, possibly, an identifying (i.e., structure-identifying) model[3] – is called *predictive estimation under changed structure*.

It may be found helpful to have a diagrammatic presentation of this procedure as given in Figure 1. As indicated on the diagram, the link B (from 2″ to 3) requires that the model be identifying. Hence, in general, *prediction under changed structure* requires that the identification power[3] of the model be established so as to make *structural estimation* possible. (Structural estimation consists of links A″ and B, but computationally one can go directly from 1 to 3; this, of course, does not eliminate the need for identification power.[4])

The point to be emphasized is that, in general, one cannot get to 6 directly from 2′: it is necessary to go through 2″, 3, 4, and 5. However, *if no structural change occurs*, the "new" predictor-predictand relationship 6 is identical with the "old" one 2′. Hence there is no need for going all around from 1 through the stages 2″, 3, 4, and 5, to 6. Instead one may go directly from 1 to 2′, using the link A′.[5] Here the link B is not used and hence there is *no need for identification power* in the case of *prediction under unchanged structure*.

---

[1] I.e., provided enough a priori knowledge is available to make the structure determinable from the distribution of observed variates; cf. [IV].

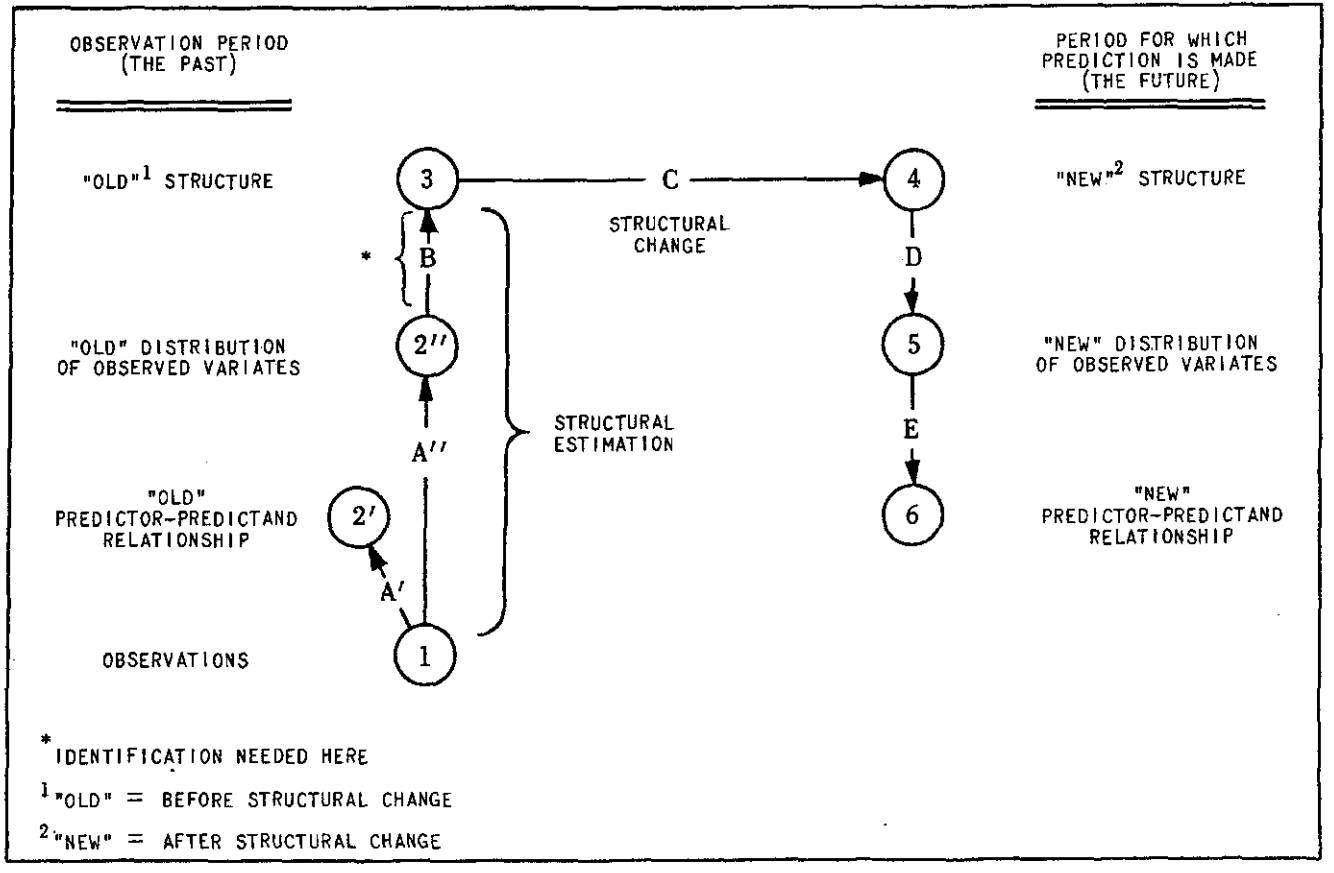
[2] The distribution of the observed variates can *always* be found when the structure is known; the converse is true only in structure-identifying models.

[3] A model that is structure-identifying is said to possess identification power.

[4] Cf. footnote on page 271.

[5] One could also go from 1 to 2″ and from there to 2′. This would be analogous to using the link E.

Figure 1. *Predictive Estimation*



0.1.2. Prediction for structural policies. Structural changes may be due either to causes consciously controlled by man ("policies") or to other factors. (However, not all policies are of this *structural* type. Some policies are carried out through certain observable variables, say the quantity of money, and need not imply any structural changes.) If several courses of action are available, the policy decision will be made on the basis of the prediction of the likely effects of the different policies. If the policy is of the type that implies a structural change, the decision requires that prediction be made of behavior of certain economic variables under the different possible structural changes. Thus one is forced to go through the procedure of predictive estimation under changed structure, a phase of which is structural estimation, and one must have an identifying model.[1]

0.1.3. Prediction concept generalized. The term *prediction* may be given a somewhat broader meaning than that of making probability statements about the future on the basis of the past. Sometimes it is desired to estimate the values of a variable given a contemporary of later value of another variable. The need for such a procedure may be due to the fact that the variable to be "predicted" from subsequent values of another is not directly available or, perhaps, is available only with a considerable delay.

Thus, the essence of the prediction concept lies not in the temporal sequence of the predictor and the predictand, but rather in the conditional nature of the (probability) statement about some set of variables given another such set.

Also, prediction need not apply to one variable only. One may, for instance, wish to know the future correlation of some two variables given the past values of a third one, etc.

0.2. Regression: A Special Case of Predictive Estimation

0.2.1. Prediction of a single variable under unchanged structure. When the structure is unchanged, there is, as we have seen, no need for estimation of the structure of the system; our objec-

---

[1] There are exceptional situations where some short-cuts can be made: it might even happen that *partial identification power* (see [IV]) is sufficient, but in general this will not be the case. In certain case where prediction of policy effects does require *total identification power (ibid.)*, it is nevertheless possible to decide on a preferable course of action when such identification power *(ibid., 8.1)* does not exist.

tive is 2′ in Figure 1 and this can be reached directly from 1 through the link A′. This link consists in obtaining the conditional distribution of the (univariate) predictand. The distribution may be regarded as given by its moments. If the distribution is normal, the first two moments (say the mean and the variance) suffice to determine the distribution; in general, the higher moments are also needed, but they will not be considered here.

The relationship between the mean of the predictand and (the values of) the predictors is the *regression* function; the relationship between the variance of the predictand and the predictors is the scedastic function.

0.2.2. Methods of estimating the regression functions. The problem of statistical inference is now reduced to finding *optimal methods of estimating* these functions. We shall see that, under rather general conditions, in models with normally distributed disturbances, the *maximum-likelihood* estimates of the (true) regression coefficients will be given by the application of the *least-squares* procedure. Under even more general conditions, the least-squares method will yield *consistent*, but probably somewhat inefficient, estimates of the regression coefficients.

Hence, if prediction under unchanged structure is intended, and the sample is sufficiently large, the method of least squares is a safe one (in the sense of consistency) to use, although not necessarily the best one (in the sense of efficiency).

But for small and medium samples, the asymptotic properties of consistency and efficiency are not adequate safeguards against serious sampling errors. It is therefore of interest to investigate the small-sample properties of the least-squares estimates both for the case where the least-squares estimates satisfy the maximum-likelihood criterion and the case where the maximum-likelihood estimates differ from the least-squares estimates.

It is found that in both cases the least-squares estimates are strongly *biased*[1] when the series is of the *autoregressive* type, i.e., where the variables satisfy a system of stochastic difference equations with *lagged endogenous* variables present.[2] On the other hand,

[1] An estimate is biased if its expectation, for any given size sample, is not identically equal to the true value of the parameter.

[2] The example given in section *0.1.1* was of the autoregressive type since national income, which is an endogenous variable, appeared both as a predictand and, with a lag of one year, as a predictor.

in *lagless* (nonautoregressive) systems, the least-squares estimates of the regression coefficients do possess an important optimal property: they are the best[1] unbiased linear estimates of those coefficients. It is important to note that this property is not asymptotic, but holds for any finite-sized sample.[2]

0.2.3. Regression coefficients and structural coefficients. The question of the extent to which the least-squares estimates are "good" estimates of the regression coefficients, treated in the preceding paragraph, should not be confused with the following one: under what conditions do the structural coefficients of an equation in a simultaneous stochastic system coincide with a set of regression coefficients in some predictor-predictand relationship?

When such a coincidence does occur, it is clear that the problem of structural estimation (for that particular equation) is exactly identical with the problem of predictive estimation under unchanged structure, when the predictors and the predictand are suitably chosen. Hence in such a case the least-squares estimates are "good" or "best" estimates, not only of the regression coefficients, but also of the structural coefficients. Just when this does happen is discussed elsewhere in the volume.[3]

0.3. Outline of the Remainder of This Paper

The remainder of this paper is divided into three sections. Section *1* provides definitions of the concepts used as well as the principles of notation. The very brief section *2* is devoted to

---

[1] I.e., with smallest sampling variance among all the unbiased linear estimates.

[2] It is conceivable that *for certain populations* there exist better unbiased linear estimates than the least-squares estimates, but there are no linear estimates unbiased for *all* populations and better than the least-squares estimates.

[3] [I], [II]. One case where structural and regression coefficients coincide is when in a given equation all variables but one are either lagged or exogenous. In general, when a least-squares regression has a high multiple-correlation coefficient and the standard errors of the regression coefficients are small, this only proves that we have a good (and probably accurately estimated) predictive relationship, but not necessarily that this relationship coincides with one of the structural equations (even if the variables entering the predictive formula are exactly the same as those entering a given structural equation). See [I] and [II – *3.3.7*] in the special case where the set I contains only one structural equation. See also *1.2.1* below.

policy aspects of the prediction problem. In section *3* the field of inquiry is narrowed down to that of regression estimation (see section *3.0*): the optimal properties of the maximum-likelihood, Markoff (see *3.1.1.2*), and least-squares regression estimates and the domain of their equivalence are investigated. A more detailed outline of section *3* is given in *3.0.3*.

1. NOTATION AND DEFINITIONS: DISTRIBUTIONS OF OBSERVED VARIATES; STRUCTURE AND STRUCTURAL CHANGES; PREDICTIVE ESTIMATION

1.1. Let the system involve a set (denoted by capitals)

$$Y = (y_0, y_1, \ldots, y_{K_y}) \tag{1}$$

of the $K_y + 1$ *observed stochastic* variates y_g and a set

$$Z = (z_1, z_2, \ldots, z_{K_z}) \tag{2}$$

of the K_z *observed fixed*[1] variates z_j. Let the cumulative distribution function of Y be

$$F \equiv F_y \equiv F_y(y; z)\,, \tag{3}$$

where y and z are vectors whose components are the elements of the sets Y and Z. For any nonempty[2] proper[3] subset Y^* of Y (i.e., $\Lambda \neq Y^* \subset Y$) we have the marginal cumulative distribution function

$$F_* \equiv F_{y^*} \equiv F_{y^*}(y^*; z)\,. \tag{4}$$

---

[1] A variable z is said to be a "fixed" variate if its tth observation has the cumulative distribution function given by

$$F(z_t) = \begin{cases} 0 \text{ for } z_t < z_{0t}\,, \\ 1 \text{ for } z_t \geq z_{0t}\,, \end{cases} \tag{2'}$$

where the z_{0t}'s form an arbitrarily chosen sequence of constants.

[2] The empty set will be denoted by Λ.

[3] $E' \supset E''$ means that E'' is a subset of E' but $E' \neq E''$. $E' \supseteq E''$ means that E'' is a subset of E', but their equality is not excluded.

Given two sets of y^* and y^{**} such that[1]

$$\Lambda \neq Y^* \subset Y \tag{5}$$

and

$$Y^{**} \equiv Y - Y^*, \tag{6}$$

we define the *conditional* marginal cumulative distribution function

$$F_{*\,|\,**} \equiv F_{y^*\,|\,y^{**}} \equiv F_{y^*\,|\,y^{**}}(y^* \mid y^{**};\ z). \tag{7}$$

We shall refer to F in (3) as the *complete* distribution, while the marginal and conditional distributions in (4) and (7) will be called *derivates* of F, or derived distributions.

1.2. Let the *structure*[2] of the system be denoted by S. Then S determines F, which may be written as

$$F = \varphi(S). \tag{8}$$

Now $F_{*\,|\,**}$ is a derivate of F, hence

$$F_{*\,|\,**} = \varphi_{*\,|\,**}(S). \tag{9}$$

Suppose a *structural change* takes place, so that the "old" structure $S^{(0)}$ is replaced by the "new" one $S^{(1)}$. We have

$$S^{(1)} = \mathfrak{I}(S^{(0)}), \tag{10}$$

where $\mathfrak{I}$ must be so specified that when $S^{(0)}$ is given, $S^{(1)}$ can be uniquely determined. Then

$$F^{(1)} = \varphi(S^{(1)}) = \varphi\{\mathfrak{I}(S^{(0)})\} = \varphi(S^{(0)};\ \mathfrak{I}) \tag{11}$$

and

---

[1] It will be noted that Y^{**} is not required to precede Y^* in time. The opposite may well be true.

[2] For general definitions of structure, model, and identification, see [IV]. In the usually treated parametric normal linear case described by the matrix equation $Ax = u$ (where A is the structural coefficient matrix, u the nonobservable disturbance, and x the observed variate), the structure is defined by A and the disturbance covariance matrix Σ^u so that $S = (A, \Sigma^u)$.

$$F^{(1)}_{*|**} = \varphi_{*|**}(S^{(1)}) = \varphi_{*|**}\{\mathfrak{I}(S^{(0)})\} \equiv \varphi_{*|**}(S^{(0)}, \mathfrak{I}) \tag{12}$$

while

$$F^{(0)}_{*|**} = \varphi_{*|**}(S^{(0)}) . \tag{13}$$

In general, there does *not* exist a relationship

$$F^{(1)}_{*|**} = \psi_{*|**}\{F^{(0)}_{*|**}; \mathfrak{I}\} . \tag{14}$$

That is, apart from special cases, the knowledge of $S^{(0)}$ is necessary in order that $F^{(1)}_{*|**}$ should be obtained.

1.2.1. The situation discussed in the preceding section may be illustrated by the following example. Let our model consist of equations

$$y_0 = \beta_1 y_1 + \gamma_1 z_1 + u_1 , \qquad y_1 = \beta_2 y_0 + \gamma_2 z_2 + u_2 ,$$

where the z's are exogenous with a moment matrix $[\zeta_{ij}]$ and the u's are the disturbances with zero means and a covariance matrix $[\sigma_{ij}]$. If we are to predict y_0 given y_1 and z_1, we must have the regression function

$$\mathcal{E}(y_0 \mid y_1, z_1) = \chi_1^y y_1 + \chi_1^z z_1 .$$

The χ's can be expressed in terms of the structural parameters of the system. Thus, for instance,

$$\chi_1^y - \beta_1 = (1 - \beta_1\beta_2)(\sigma_{12} + \beta_2\sigma_{11}) \times \left\{ \sigma_{22} + \beta_2^2\sigma_{11} + 2\beta_2\sigma_{12} + \gamma_2^2\left(\frac{\zeta_{22}}{\zeta_{11}} - \frac{\zeta_{12}^2}{\zeta_{11}}\right)\right\}^{-1} .$$

Unless $\sigma_{12} + \beta_2\sigma_{11} = 0$, in which case the single-equation approach is applicable, we find that the difference $\Delta\beta_1$ between a "new" value $\beta_1^{(1)}$ and an "old" value $\beta_1^{(0)}$ is not sufficient to determine the corresponding increment $\Delta\chi_1^y$. Hence, in general, we need additional knowledge of the structural parameters in the right-hand member of the above expression.

It may be added that the expression for $\chi_1^y - \beta_1$ gives the magnitude of the bias involved in using the single-equation least-squares value χ_1^y as a measure of β_1. This bias may be high even if the multiple correlation coefficient ρ^2 (y_0 on y_1 and z_1) is high and the (conventionally computed) standard errors σ_χ low. Thus, for instance, an example has been constructed by the author with $\beta_1 = 0$ while $\chi_1^y = .99$ despite the fact that $\rho^2 = .9901$ and $\sigma^2_{\chi_1^y} = .005$.

1.3. Now let the observations (prior to the structural change $\mathfrak{I}$) be given by the matrix $X \equiv [X_{it}]$; $(i = 0, 1, \ldots, K_y + K_z;$ $t = 1, 2, \ldots, T)$.[1] We obtain an estimate, denoted by $\operatorname{est} F^{(0)}(X)$, of $F^{(0)}$. Then, if we are dealing with an identifying model[2] we can obtain an estimate of $S^{(0)}$ from, say,[3] $\operatorname{est} S^{(0)} = \psi(\operatorname{est} F^{(0)}; \mathfrak{S}_0)$. By virtue of (10), we have

$$\operatorname{est} S^{(1)} = \mathfrak{I}(\operatorname{est} S^{(0)}) \tag{15}$$

and, consequently, we obtain

$$\operatorname{est} F^{(1)}_{*|**} = \varphi_{*|**}(\operatorname{est} S^{(1)}) \equiv \varphi_{*|**}\Big[\mathfrak{I}\{\psi[\operatorname{est} F^{(0)}(X); \mathfrak{S}_0]\}\Big]. \tag{16}$$

The process of obtaining $\operatorname{est} F^{(1)}_{*|**}$ from X is called *predictive* estimation. Y^* is the *predictand*, Y^{**} and Z the *predictors*. In general, prediction involves the use of X, $\mathfrak{I}$, and $\mathfrak{S}_0$. However, when $\mathfrak{I} = \mathfrak{A}$, where $\mathfrak{A}$ is the identity transformation, we clearly have

$$\operatorname{est} F^{(1)}_{*|**} = \operatorname{est} F^{(0)}_{*|**}\,. \tag{17}$$

This is a special case where a relationship of the type (14) does exist, hence it is not necessary here to obtain $S^{(0)}$, etc. In particular, therefore, it is not necessary that the model be identifying.

It is with regard to this special case of *predictive estimation*

---

[1] In lagless systems, t may, but need not necessarily, mean time.

[2] See [IV].

[3] $\mathfrak{S}_0$ is the set of all structures permitted by the identifying restrictions.

under unchanged structure ($\mathfrak{I} = \mathfrak{A}$) that we examine the optimal properties of the least-squares method in section *3*.

2. PREDICTION FOR POLICY PURPOSES

One of the most important cases of need for prediction under changed structure is that of certain types of economic policies. There are cases where structural changes are not due to policy decisions (e.g., spontaneous changes of tastes); nor do all policies imply structural change; some only involve manipulation of Z (nonstructural policies). But when a change is to be made between two policies implying the respective structural changes $\mathfrak{I}'$ and $\mathfrak{I}''$, one cannot decide between them (no matter what the policy objectiv tive) without knowing just how they will affect certain variables. Here, since $\mathfrak{I} \neq \mathfrak{A}$ for one or both of $\mathfrak{I}'$ and $\mathfrak{I}''$, we deal with predictive estimation under structural change; hence in general we must resort to structural estimation, and it is necessary that the model possess a certain degree of identification power.

On the other hand, suppose that the policy contemplated involves *no structural change*, so that it is either one of inaction or manipulation of Z. Then, regardless of whether or not Z is to be manipulated, we have $\mathfrak{I} = \mathfrak{A}$, and the effects of such a policy (nonstructural policy or complete inaction) can be predicted without resort to structural estimation; hence it is not necessary to have an identifying model.

3. OPTIMAL PROPERTIES AND MUTUALRRELATIONSHIPS OF ALTERNATIVE METHODS (MAXIMUM-LIKELIHOOD, MARKOFF, LEAST-SQUARES) OF ESTIMATING THE REGRESSION FUNCTIONS

3.0. Regression and its Estimation: A Special Case of Predictive Estimation

3.0.1. The remainder of this paper is devoted to prediction under unchanged structure (see section *1.3* above).

The case considered is not the most general one even under unchanged structure. Three important specializing assumptions are made:

(a) *The choice of the sets* Y^* *and* Y^{**}.

(18) $$Y^* = (y_0),$$

so that Y^* is a one-element set.

(19) $$Y^{**} = Y - Y^* = (y_1, y_2, \ldots, y_{K_y}),$$

i.e., Y^{**} is the complement of Y^* in Y. [1]

(b) *Admissible distributions.* It will be assumed that the regression of y_0 on the elements of Y^{**} is linear and homoscedastic,[2] so that[3]

(20) $$\mathcal{E}(y_0 \mid Y^{**}; Z) = \sum_{i=1}^{K_y} \chi_i^y y_i + \sum_{j=1}^{K_z} \chi_j^z z_j + \chi_0$$

and

(21) $$\sigma^2(y_0 \mid Y^{**}; Z) = \text{const.}$$

(c) *Properties to be estimated.* While the general definition of prediction calls for estimation of $F_{*|**}$, our study will be confined to the estimation of the first moment of $F_{*|**}$, i.e., to the estimation of the χ's (*regression coefficents*) in (21).

3.0.2. Estimation principles. Three important principles have been applied in the past in estimating the regression coefficients. They are: (a) maximum-likelihood, (b) best unbiased linear (Mar-

---

[1] (18) does narrow down the scope of the study, but (19) only seems to. For suppose that, contrary to (19), we have

(19′) $$Y - Y^* - Y^{**} \neq \Lambda;$$

then we form the marginal distribution of $Y^{***} \equiv Y^* + Y^{**}$, say F_{***}, and consider it as the parent distribution; obviously (19) now holds, provided Y is replaced by Y^{***}, and the nature of the problem is unchanged.

[2] The restriction of homoscedasticity is not difficult to relax, provided ratios of conditional variances are assumed known or even under more general conditions.

[3] χ_0 in (21) may, without loss of generality, be assumed to vanish since it is always permissible to have one of the Z's identically equal to unity.

koff), and (c) least-squares. It is well known that the three methods are related and that there are cases where all three, or at least two, are equivalent. But examples of divergence also exist. Therefore, it becomes important to investigate the domain of their equivalence in relation to the nature of the model.

3.0.3. Outline of the remainder of this section. Section *3.1* of this paper will be devoted to the study of the domain of equivalence of the best unbiased linear and least-squares estimates; section *3.2* to equivalence of maximum-likelihood and least-squares estimates. Each section is divided into two subsections; the first subsection of each section (i.e., *3.1.1* and *3.2.1*) treats of lagless models.

The problem of additional restrictions on the parameters is treated in section *3.3* for both best unbiased linear and maximum-likelihood estimates, with or without lags.

3.1. Equivalence of Best Unbiased Linear and Least-Squares Estimates of the Regression Coefficients

3.1.1. Lags absent.

3.1.1.1. Consider a sample of size T where the tth observation is a $(1 + K_y + K_z)$-dimensional vector (y_t, z_t) with

$$(22) \qquad \begin{aligned} y_t &= (y_{0t}, y_{1t}, \ldots, y_{K_y t}), \\ z_t &= (z_{1t}, z_{2t}, \ldots, z_{K_z t}). \end{aligned}$$

It will be assumed that the observations for successive values of t are independent in the probability sense, but they need not come from the same universe. (This will be qualified later.)

Given the cumulative distribution function $F_t(y_t, z_t)$ of the universe from which the tth observation is drawn, the joint cumulative distribution function of the sample is

$$(23) \qquad \prod_{t=1}^{T} F_t(y_t; z_t).$$

Now from each of the F_t it is possible to derive the conditional cumulative distribution function of y_0 given the remaining y's and z's. These conditional distributions, to be written as

$$F_t^C(y_{0t} \mid y_t^{(0)} ; z_t), \tag{24}$$

where $y_t^{(0)} = (y_{1t}, \ldots, y_{K_y t})$, will be assumed identical for all t so that the subscript t of F_t^C may be dropped. Moreover, the first two moments of this conditional distribution F^C are assumed to satisfy equations (20) and (21), so that the regression of y_0 on the other y's and z's is linear and homoscedastic.

We now seek to form (out of observations of the sample described above) an estimate q for any one of the regression coefficients χ in (20). (For the moment the affixes of χ and q may be omitted.) These estimates are to be *linear* in the y_0 and can thus be written as

$$q = \sum_{t=1}^{T} \varphi_t y_{0t}, \tag{25}$$

where

$$\varphi_t = \varphi_t(y_1^{(0)}, y_2^{(0)}, \ldots, y_T^{(0)}; z_1, z_2, \ldots, z_T). \tag{26}$$

The expectation of q is given by the Stieltjes integral

$$\mathcal{E}(q) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \sum_{t=1}^{T} \varphi_t y_{0t} \prod_{t=1}^{T} dF_t(y_t; z_t) \tag{27}$$

and it may or may not equal χ. Whether this is the case depends a) on the choice of the φ_t's, and b) on the nature of the function in (23).

For a given choice of φ_t's, it may happen that q would equal χ only provided the function in (23) belongs to a specified family $\mathfrak{F}_0$ of such functions. In such a case we say that the estimate q corresponding to that particular choice of the φ_t's is (conditionally) *unbiased with regard to* $\mathfrak{F}_0$. If the family $\mathfrak{F}_0$ is the family $\mathfrak{F}$ of all distribution functions, the estimate q will be said to be *absolutely unbiased.*

We shall now derive a *necessary and sufficient condition* to be satisfied by the φ_t's if q is to be an *absolutely unbiased estimate* of χ. The regression is assumed to be linear, that is, (20) holds.

Using (20) and the well-known properties of conditional distri-

butions, we obtain[1] from (27)

$$\mathcal{E}(q) = \int \cdots \int \sum_t \varphi_t y_{0t} \prod_t dF^C(y_{0t} \mid y_t^{(0)} ; z_t) \prod_t dF_t^M(y_t^{(0)} ; z_t)$$

$$= \int \cdots \int \sum_t \varphi_t \mathcal{E}(y_{0t} \mid y_t^{(0)} ; z_t) \prod_t dF_t^M(y_t^{(0)} ; z_t) \tag{28}$$

$$= \int \cdots \int \sum_t \varphi_t \left(\sum_i \chi_i^y y_{it} + \sum_j \chi_j^z z_{jt} \right) \prod_t dF_t^M(y_t^{(0)} ; z_t),$$

or, rearranging,

$$\mathcal{E}(q) = \sum_i \chi_i^y \int \cdots \int \left(\sum_t \varphi_t y_{it} \right) \prod_t dF_t^M + \sum_j \chi_j^z \int \cdots \int \left(\sum_t \varphi_t z_{jt} \right) \prod_t dF_t^M , \tag{29}$$

where F^C is a conditional, and F_t^M a marginal, cumulative distribution function.

So far neither q nor the φ_t's have had any affixes referring to the χ which is being estimated. Now, for definiteness, we shall assume that it is χ_1^y that is being estimated; its estimate will be written q_1^y; the corresponding φ_t's will be denoted by $\varphi_t^{y_1}$, but the superscripts may be omitted where no danger of confusion exists.

Now $\mathcal{E}(q_1^y)$ will be given by (29) with proper affixes attached to q and the φ_t's, thus yielding

$$\mathcal{E}(q_1^y) = \sum_i \chi_i^y \int \cdots \int \sum_t \varphi_t^{y_1} y_{it} \prod_t dF_t^M + \sum_j \chi_j^z \int \cdots \int \sum_t \varphi_t^{y_1} z_{jt} \prod_t dF_t^M . \tag{30}$$

If q_1^y is to be an absolutely unbiased estimate of χ_1^y, the right-hand member of (30) must equal χ_1^y for any choice of $\prod_t dF_t^M$. This

[1] The integrals are between $-\infty$ and $+\infty$ unless otherwise indicated.

will be the case if and only if

$$
\begin{aligned}
&\int \cdots \int \Big(\sum_t \varphi_t^{\gamma_1} \, y_{it} - 1 \Big) \prod_t dF_t^M \equiv 0 , \\
&\int \cdots \int \sum_t \varphi_t^{\gamma_1} \, y_{it} \prod_t dF_t^M \equiv 0 , \quad \text{for all } i \neq 1 , \\
&\int \cdots \int \sum_t \varphi_t^{\gamma_1} \, z_{jt} \prod_t dF_t^M \equiv 0 , \quad \text{for all } j .
\end{aligned} \tag{31}
$$

It is easily seen that a set of *sufficient* conditions for (31) to hold is:

$$
\begin{aligned}
&\sum_t \varphi_t^{\gamma_1} \, y_{it} - 1 \equiv 0 , \\
&\sum_t \varphi_t^{\gamma_1} \, y_{it} \equiv 0 , \quad \text{for all } i \neq 1 , \\
&\sum_t \varphi_t^{\gamma_1} \, z_{jt} \equiv 0 , \quad \text{for all } j .
\end{aligned} \tag{32}
$$

But it can be shown that the identities (32) are also *necessary* for q_1^{γ} to be absolutely unbiased. To carry out the proof it is enough if a family $\mathfrak{F}^M$ of $\{F_t^M; (t = 1, 2, \ldots, T)\}$ is found for which the vanishing of the integrals in (31) implies the vanishing of their integrands, as given in (32). One such family is given by making the y_{it}'s normally independently distributed with a common variance but different means.[1] Many more could, no doubt, be found. Therefore, the identities (32) are the *necessary and sufficient* conditions for q_1^{γ} to be absolutely unbiased. But if q_1^{γ} were to be only conditionally unbiased, depending on the family with regard to which lack of bias is postulated, either (31) or (32) would have to be used.

3.1.1.2. We shall now procede to define the *best* absolutely unbiased linear estimate of, say, χ_1^{γ}. This estimate will be denoted by $\bar{q}_1^{\gamma}$ and will have the following properties:

---

[1] [Wald, 1944].

(a) linearity [i.e., (25) with (26)],

(b) absolute unbiasedness [i.e., (32)],

(c) minimum sampling variance (to be defined below).

The property (c) implies that the sampling variance $\sigma^2_{\bar{q}_1^y}$ of $\bar{q}_1^y$ will not exceed the sampling variance of any other linear absolutely unbiased estimate of χ_1^y.

Mathematically, the problem of finding an estimate possessing the properties (a), (b), (c) (a *"Markoff estimate"*) requires minimization of the integral giving $\sigma^2_{q_1^y}$ [where q_1^y has properties (a) and (b)] subject to the side relation (32).

Now

$$\sigma^2_{q_1^y} = \int \cdots \int \Big(\sum_t \varphi_t y_{0t} - \chi_1^y \Big)^2 \prod_t dF_t(y_t ; z_t) \tag{33}$$

or

$$\sigma^2_{q_1^y} = \int \cdots \int \Big(\sum_t \varphi_t y_{0t} \Big)^2 \prod_t dF_t - (\chi_1^y)^2 . \tag{34}$$

$\mathcal{E}(q_1^y)^2$ is given by the integral in the right-hand member of (34).

Now, using the properties of homoscedasticity and applying transformations analogous to those in (28), we have

$$\mathcal{E}(q_1^y)^2 = \int \cdots \int \sum_{t', t''} \varphi_{t'} \varphi_{t''} y_{0t'} y_{0t''} \prod_t dF_t^C \prod_t dF_t^M \tag{35}$$

which, after simplification gives

$$\sigma^2_{q_1^y} = \sigma^2 \int \cdots \int \sum_t \varphi_t^2 \prod_t dF_t^M . \tag{36}$$

Hence the problem is reduced to that of minimizing the integral (36) subject to the restrictions (32). Formally, this is a Calculus of Variations problem, but actually, since no derivatives enter the integrand, we may simply minimize the integrand $\sum_t \varphi_t^2$

subject to (32).

At this stage the mathematical problem is exactly the same as that treated in the older version of the Markoff Theorem.[1] Therefore, the estimate $\tilde{q}_1^y$, obtained by minimizing the sampling variance $\sigma^2_{\tilde{q}_1^y}$ with regard to the φ's subject to conditions (a) and (b) listed above, is identically equal to the least-squares estimate $\tilde{q}_1$, obtained by minimizing the expression

$$(37) \qquad S \equiv \sum_t (y_{0t} - \sum_i q_i^y y_{it} - \sum_j q_j^z z_{jt})^2$$

with regard to the q's. Thus "best" (in the Markoff sense) *absolutely unbiased linear* estimates are *identical with the least-squares* estimates.

3.1.1.3. In the preceding section we considered Markoff estimates that were *absolutely* unbiased. They turned out to be identical with the least-squares estimates.

In general, this is not the case for Markoff estimates *conditionally* unbiased with regard to a family of distributions for which (31) does not imply (32).[2]

Since this is a proposition of a negative nature, it can be proved by demonstration of a special case. We shall assume that only χ_1^y (to be written simply as χ) is different from zero, while all the other χ's vanish.

Here we again proceed to minimize σ_q^2 subject to

$$(38) \qquad \int \cdots \int (\sum_t \varphi_t y_{1t} - 1) \prod_t dF_t^M \equiv 0$$

which is all that remains of (31).

Now in this case (35) gives

$$(39) \qquad \mathcal{E}(q^2) = \int \cdots \int \{\sigma^2 \sum_t \varphi_t^2 + \chi^2 (\sum_t \varphi_t y_{1t})^2\} \prod_t dF_t^M$$

or

[1] [David and Neyman].

[2] If the estimates are to be conditionally unbiased with regard to a family of distributions for which (31) does imply (32), then such conditionally unbiased Markoff estimates would, of course, still be identical with the least-squares estimates.

or

$$(40)\quad \mathcal{E}(q^2) = \sigma^2 \int \cdots \int \{ \sum_t \varphi_t^2 + \frac{\chi^2}{\sigma^2} (\sum_t \varphi_t y_{1t})^2 \} \prod_t dF_t^M .$$

It is evident that we may minimize the integral in (40) instead of σ_q^2. Writing

$$(41)\quad \nu^2 = \frac{\chi^2}{\sigma^2} ,$$

we may therefore reformulate the problem as that of minimizing the integral

$$(42)\quad M = \int \cdots \int \{ \sum_t \varphi_t^2 + \nu^2 (\sum_t \varphi_t y_{1t})^2 \} \prod_t dF_t^M$$

subject to (38).

Now

$$(43)\quad M = \int \cdots \int \sum_{t', t''} \varphi_{t'} \varphi_{t''} s_{t't''} \prod_t dF_t^M ,$$

where

$$(44)\quad s_{t't''} = \delta_{t't''} + \nu^2 y_{1t'} y_{1t''} ,$$

and where $\delta_{t't''}$ is the Kronecker symbol.

Now the integrand in (43) may be written in matrix form as

$$(45)\quad \varphi' S \varphi$$

where[1]

$$(46)\quad \varphi = \{\varphi_1, \ldots, \varphi_T\} , \qquad S = [s_{t't''}] ,$$

and φ' is the transpose of φ.

---

[1]The notation used here is the customary one (φ is a column vector, φ' a row vector) rather than that used in other parts of this volume.

Similarly, the integrand of (38) is

$$\varphi' y_1 - 1 . \tag{47}$$

We now minimize (45) with respect to φ' subject to the vanishing of (47) and denote the Lagrange multiplier by μ (a scalar).

Differentiating the expression

$$\varphi' S \varphi - \mu \varphi' y_1 \tag{48}$$

with respect to φ', we obtain

$$\frac{\partial}{\partial \varphi'} \{ \varphi' S \varphi - \mu \varphi' y_1 \} = S \varphi - \mu y_1 = 0 , \tag{49}$$

and, hence,

$$\varphi = \mu S^{-1} y_1 . \tag{50}$$

To evaluate

$$S^{-1} y_1 \tag{51}$$

we observe that

$$S = I + \nu^2 y_1 y_1' . \tag{52}$$

Hence

$$\begin{aligned} S y_1 &= (I + \nu^2 y_1 y_1') y_1 = I y_1 + \nu^2 (y_1' y_1) y_1 \\ &= I \{1 + \nu^2 (y_1' y_1)\} y_1 \end{aligned} \tag{53}$$

and

$$S^{-1} S y_1 = \{1 + \nu^2 (y_1' y_1)\} S^{-1} y_1 , \tag{54}$$

so that the desired expression is

$$S^{-1} y_1 = \frac{1}{1 + \nu^2 (y_1' y_1)} y_1 . \tag{55}$$

Thus, substituting (55) into (56), we have

$$\varphi = \mu \frac{1}{1 + \nu^2 (y_1' y_1)} y_1 \tag{56}$$

Only the multiplier μ remains to be evaluated.

Premultiplying (56) with y_1' and integrating, we find that

$$\int \cdots \int y_1' \varphi \prod_t dF_t^M = \mu \int \cdots \int \frac{y_1' y_1}{1 + \nu^2 (y_1' y_1)} \prod_t dF_t^M , \tag{57}$$

where the left-hand member equals unity by (42). Therefore,

$$\mu^{-1} = \int \cdots \int \frac{y_1' y_1}{1 + \nu^2 (y_1' y_1)} \prod_t dF_t^M , \tag{58}$$

and, finally,

$$\varphi = \frac{y_1}{\{1 + \nu^2 (y_1' y_1)\} \int \cdots \int \frac{y_1' y_1}{1 + \nu^2 (y_1' y_1)} \prod_t dF_t^M} . \tag{59}$$

Thus, if matrix notation is abandoned, $\overline{q}$, the Markoff "estimate"[1] of the regression coefficient χ is

$$\overline{q} = \frac{\sum_t y_{1t} y_{0t}}{(1 + \nu^2 \sum_t y_{1t}^2) \int \cdots \int \frac{\sum_t y_{1t}^2}{1 + \nu^2 \sum_t y_{1t}^2} \prod_t dF_t^M} , \tag{60}$$

---

[1] q depends not only on the observations, but also on ν^2 which is a population parameter and, in general, is not known; an analogous "estimate" was obtained in [Johnson], where the knowledge of the coefficient of variation is needed in order to obtain a least variance (about true value, not necessarily unbiased) estimate of the mean in a normal population.

In order to avoid this type of solution one should subject the above Calculus of Variations procedure to the additional restrictions $\partial \varphi_t / \partial \chi =$

while the least-squares estimate $\tilde{q}$ of χ would have been

$$\tilde{q} = \frac{\sum_t y_{1t} y_{0t}}{\sum_t y_{1t}^2} . \tag{61}$$

3.1.1.4. Three observations should be made concerning the foregoing result.

(a) $\bar{q}$ is actually a better "estimate" than $\tilde{q}$, i.e.,

$$\sigma_{\bar{q}}^2 < \sigma_{\tilde{q}}^2 . \tag{62}$$

To see this we shall calculate both variances. Writing

$$\begin{aligned} y_1' y_1 &= p , \\ 1 + \nu^2 p &= k , \end{aligned} \tag{63}$$

we may, using (56), write the integrand $Q = \varphi' S \varphi$ of M in (43) as

$$Q = \mu^2 (p/k) \tag{64}$$

where μ, given by (58), may be written as

$$\mu = \{ \int \cdots \int (p/k) \prod_t dF_t^M \}^{-1} . \tag{65}$$

Hence the integral M in (43) is

$$\begin{aligned} M &= \int \cdots \int Q \prod_t dF_t^M = \int \cdots \int \mu^2 \frac{p}{k} \prod_t dF_t^M \\ &= \mu^2 \int \cdots \int \frac{p}{k} \prod_t dF_t^M = \mu^2 \mu^{-1} = \mu , \end{aligned} \tag{66}$$

0, $\partial \varphi_t / \partial \sigma = 0$. This has not yet been done, but it would be of interest to see whether these additional restrictions would imply $\bar{q} = \tilde{q}$. (Asymptotically, at least, this would seem to be the case because of the efficiency property of $\tilde{q}$ as a maximum-likelihood estimate of χ; see below, section *3.2*. Exceptions might conceivably arise if linear unbiased estimates of χ exist whose asymptotic distributions are not normal).

and, using (65) and (66),

$$(67)\quad \sigma^2_{\bar{q}} = \sigma^2 M - \chi^2 = \sigma^2 \left\{ \int \cdots \int \frac{p}{k} \prod_t dF_t^M \right\}^{-1} - \chi^2 .$$

Having obtained the variance of the Markoff "estimate" $\bar{q}$ of χ, we shall now evaluate the variance of the least-squares estimate $\tilde{q}$ of χ.

Since

$$(68)\quad \tilde{q} = \frac{y_1' y_0}{y_1' y_1}$$

is also unbiased,[1] we have

$$(69)\quad \sigma^2_{\tilde{q}} = \mathcal{E}(\tilde{q}^2) - \chi^2 ,$$

with

$$\begin{aligned} \mathcal{E}(\tilde{q}^2) &= \int \cdots \int \frac{1}{p} \sum_{t',t''} y_{1t'} y_{1t''} \int \cdots \int y_{0t'} y_{0t''} \\ &\qquad \times \prod_t dF^C(y_{0t} \mid y_{1t}) \prod_t dF_t^M(y_{1t}) \\ &= \sigma^2 \int \cdots \int \frac{1}{p^2} y_1' S y_1 \prod_t dF_t^M \qquad (70) \\ &= \sigma^2 \int \cdots \int \frac{1}{p^2} (p + \nu^2 p^2) \prod_t dF_t^M \\ &= \sigma^2 \int \cdots \int \frac{k}{p} \prod_t dF_t^M , \end{aligned}$$

and, hence,

$$(71)\quad \sigma^2_{\tilde{q}} = \sigma^2 \int \cdots \int \frac{k}{p} \prod_t dF_t^M - \chi^2 .$$

---

[1] In fact it was shown (cf. section *3.1.1.2*) to be the best *absolutely* unbiased estimate.

Comparing this with (67) we see that in order to prove (62) we must show that

$$(72) \qquad \int \cdots \int \frac{k}{p} \prod_t dF_t^M > \{ \int \cdots \int \frac{p}{k} \prod_t dF_t^M \}^{-1} .$$

But (72) follows from a multivariate analogue of a well-known inequality.[1] Thus q is actually better than $\tilde{q}$.

(b) From a practical viewpoint, (62) offers little consolation since q cannot be computed unless the F_t^M's and ν^2 are known, and ν^2 is not likely to be known.

Unless an unbiased linear estimate not requiring the knowledge of ν^2 can be found with a variance between $\sigma_{\tilde{q}}^2$ and $\sigma_{\bar{q}}^2$, *the least-squares estimate* $\tilde{q}$ *is still the best practical solution,* although it does not utilize whatever knowledge we may possess about the F_t^M's.

(c) However, it is of considerable value to have the expression for $\sigma_{\bar{q}}^2$ since a lower bound is obtained for the variance of all the *linear unbiased* estimates of χ. If the use of a priori information concerning ν^2 is permitted, $\sigma_{\bar{q}}^2$ is obviously the greatest lower bound. If the use of ν^2 is *not* permitted, $\sigma_{\bar{q}}^2$ may or may not be the greatest lower bound, but we can say at least that the greatest lower bound does not exceed[2] $\sigma_{\tilde{q}}^2$ and is not less than $\sigma_{\bar{q}}^2$. It would, of course, be of interest to find such a greatest lower bound for the class of all (linear and nonlinear) unbiased estimates.

3.1.2. (Best unbiased linear and least-squares estimates.) Lags present. In (1) some of the y's may be lagged values of other y's. Then the right-hand member of (20) might contain lagged values of y_0.

We shall consider the simplest case of this type, where

$$(73) \qquad y_{1t} \equiv y_{0,\,t-1}$$

---

[1] [Hardy, Littlewood, Pólya; pp. 150-151, Theorem 204]. (72) would become an equality only if $(1 + \rho^2 p)/p = \text{const.}$

[2] It would seem that asymptotically $\sigma_{\tilde{q}}^2$ is the greatest lower bound because of the maximum-likelihood properties of $\tilde{q}$ (see section 3.2), and their thereby implied efficiency. Exceptions might be due to the existence of estimates whose asymptotic distribution is nonnormal.

and there are no other y's or z's present. In this case (20) and (21) are reduced to

$$\mathcal{E}(y_t \mid y_{t-1}) \equiv \mathcal{E}(Y_t \mid Y_{t-1} = y_{t-1}) = \chi\, y_{t-1}, \qquad \sigma^2(y_t \mid y_{t-1}) = \text{const.}, \tag{74}$$

where y_t is written for y_{0t} and the affixes of χ are dropped.

The sample available is that of T observations on y, say

$$(y_1, \ldots, y_T). \tag{75}$$

When χ is being estimated by the least-squares method under the assumption (74), it is considered that $T-1$ pairs of observations of the type (y_t, y_{t-1}) are given, with the first element in the pair as the "dependent" variable. The estimate $\tilde{q}$ of χ obtained from the least-squares principle is

$$\tilde{q} = \frac{\sum_{t=2}^{T} y_t\, y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2}. \tag{76}$$

The question to be answered is whether $\tilde{q}$ is a best (conditionally or absolutely) unbiased estimate. The complete solution of this problem is not known. So far there is no known class of cumulative distribution functions of $(y_1, \ldots, y_T)$ satisfying (74) with regard to which $\tilde{q}$ is (conditionally) unbiased. On the other hand, an important class of cumulative distribution functions for which (74) holds but where $\tilde{q}$ *is* biased has been found.[1] Thus it is certain the $\tilde{q}$ is not in any case an *absolutely unbiased estimate* of χ, and a fortiori not a *best* absolutely unbiased estimate.

Several questions of interest remain unanswered. Are there cumulative distribution functions satisfying (74) with regard to which $\tilde{q}$ is (conditionally) unbiased? What is the best (linear or nonlinear) unbiased[2] estimate of χ? (Asymptotically, $\tilde{q}$ probably is a best unbiased estimate due to its maximum-likelihood properties.)

---

[1] However, this bias tends to zero, though not very rapidly, as the sample size increases. Cf. [XV].

[2] It is shown in [XV] that an unbiased estimate of χ does exist.

3.2. *Equivalence of Maximum-Likelihood and Least-Squares Estimates of the Regression Coefficients*

3.2.1. *Lags absent.*

3.2.1.1. The results given in this section are not in any way new, but there does not seem to be an easily accessible source treating the general case and giving the necessary proofs.

As is well-known, the maximum-likelihood method leads to least-squares estimates only when a normal universe is assumed and this assumption will be implicit in all that follows.

Thus, consider a set Y of jointly normally distributed variates [Y given in (1)] with the following properties:

$$\mathcal{E}(y_{gt}) = \sum_{j=1}^{K_z} \pi_{gj} z_{jt} \qquad (g = 0, 1, \ldots, K_y),$$

$$\operatorname{cov}(y_{gt'}, y_{ht''}) = \omega_{gh}\, \delta_{t't''} \qquad (g, h = 0, 1, \ldots, K_y), \tag{77}$$

where the z's are (observed) fixed variates, the π's and ω's are unknown parameters[1] and $\delta_{t't''}$ the Kronecker symbol. Then, clearly,

$$\mathcal{E}(Y_{0t} \mid Y^{**}; Z) = \sum_{i=1}^{K_y} \chi_i^y y_{it} + \sum_{j=1}^{K_z} \chi_j^z z_{jt} \tag{78}$$

(which is identical with (20) where $\chi_0 = 0$) and the regression, as is well-known, is homoscedastic.

The regression coefficients χ are related to the parameters in (77) by

$$\chi_i^y = -\frac{\omega^{0i}}{\omega^{00}}, \tag{79}$$

$$\chi_j^z = \sum_g \pi_{gj} \chi_g^y . \tag{80}$$

---

[1] The problem of possible restrictions on the π's and ω's is discussed below in section *3.3*.

3.2.1.2. It is now to be shown that the least-squares estimates $\tilde{q}$ and the maximum-likelihood estimates $\hat{\chi}$ of the χ's are identical.[1] The proof will be divided into two parts: first it will be shown that $\tilde{q}^y = \hat{\chi}^y$, and then that $\tilde{q}^z = \hat{\chi}^z$. In each case we first derive the least-squares estimates and the the maximum-likelihood estimates.

I. To show that $\tilde{q}^y = \hat{\chi}^y$, we first obtain the least-squares estimates by minimizing with respect to the q's the expression S given in (37).

Setting

$$q_0^y = -1 , \tag{81}$$

we may rewrite S as

$$S = \sum_{t=1}^{T} \Big(\sum_{g=0}^{K_y} q_g^y \, y_{gt} + \sum_{j=0}^{K_z} q_j^z \, z_{jt} \Big)^2 \tag{82}$$

or, for short

$$S = \sum_t \Big(\sum\nolimits^y q^y y + \sum\nolimits^z q^z z \Big)^2 \tag{83}$$

where $\sum^y \equiv \sum_0^{K_y}$, and $\sum^z \equiv \sum_1^{K_z}$.

Differentiating S with respect to the q's we obtain

$$\frac{1}{2} \frac{\partial S}{\partial q_i^y} = \sum_t \Big[\Big(\sum_h^y q_h^y \, y_h + \sum_j^z q_j^z \, z_j \Big) y_i \Big] = 0 \tag{84}$$

$$(i = 1, 2, \ldots, K_y)$$

and

$$\frac{1}{2} \frac{\partial S}{\partial q_k^z} = \sum_t \Big[\Big(\sum_h^y q_h^y \, y_h + \sum_j^z q_j^z \Big) z_l \Big] = 0 \tag{85}$$

$$(l = 1, 2, \ldots, K_z).$$

---

[1] The proof given in the text is a reproduction of the one originally presented at the Cowles Commission Conference in February, 1945. Following this presentation Koopmans suggested an alternative proof [VII], which is much simpler, more elegant, and more easily generalized.

Using notation similar to that of [II] for sample moments, we write

$$
\begin{aligned}
m_{y_g\, y_h} &= \frac{1}{T} \sum_t y_{gt}\, y_{ht} \ , \\
m_{y_g\, z_j} &= \frac{1}{T} \sum_t y_{gt}\, z_{jt} \ , \\
m_{z_j\, z_l} &= \frac{1}{T} \sum_t z_{jt}\, z_{lt} \ ,
\end{aligned} \tag{86}
$$

where T is the size of the sample. We may then rewrite equations (84) and (85), thus obtaining

$$
\sum_h^y m_{y_i\, y_h}\, q_h^y + \sum_j^z m_{y_i\, z_j}\, q_j^z = 0 \qquad (i = 1, 2, \ldots, K_y), \tag{87}
$$

$$
\sum_h^y m_{z_l\, y_h}\, q_h^y + \sum_j^z m_{z_l\, z_j}\, q_j^z = 0 \qquad (l = 1, 2, \ldots, K_z). \tag{88}
$$

Now write

$$
\sum_h^y m_{y_0\, y_h}\, q_h^y + \sum_j^z m_{y_0\, z_j}\, q_j^z = \varphi, \tag{89}
$$

where the value of φ is immaterial, and consider (89) as the first equation of a system consisting of (87) and (89). Then (87) with (89) may be written in matrix form as

$$
M_{yy}\, q^y + M_{yz}\, q^z = \varphi\, e \ , \tag{90}
$$

where $e = [1 \ 0 \ \ldots \ 0]$. Similarly, (88) may be written

$$
M_{zy}\, q^y + M_{zz}\, q^z = 0 \ , \tag{91}
$$

where $M_{zy} = M_{yz}'$ is the transpose of M_{yz}, and $0 = [0 \ 0 \ \ldots \ 0]$.

It follows from (91) that

$$q^z = - M_{zz}^{-1} M_{zy} q^y , \tag{92}$$

and substituting this value into (90) we obtain

$$M_{yy} q^y - M_{yz} M_{zz}^{-1} M_{yz} q^z = \varphi e , \tag{93}$$

or, finally,

$$q^y = \tilde{q}^y = (M_{yy} - M_{yz} M_{zz}^{-1} M_{zy})^{-1} \varphi e . \tag{94}$$

Now, in obtaining the maximum-likelihood estimates, we see from the elementary properties of determinants that

$$\sum_h^y w_{gh} w^{0h} = \delta_{0g} | W | \qquad (g = 0, 1, \ldots, K_y) \tag{95}$$

where w_{gh} is the maximum-likelihood estimate of ω_{gh}. Dividing by w^{00}, we have

$$\sum_h^y w_{gh} = \delta_{0g} \frac{| W |}{w^{00}} . \tag{96}$$

Thus, by virtue of (80)[1] and setting $q_0^y = -1$, we have

$$\sum_h^y w_{gh} \hat{\chi}_h^y = - \delta_{0g} \frac{| W |}{w^{00}} , \tag{97}$$

which may be written in matrix form as

$$W \hat{\chi}^y = \psi e , \tag{98}$$

where $\psi = - \dfrac{| W |}{w^{00}}$ and $e = [1 \ 0 \ \ldots \ 0]$ as before.

It will be noted that the value of ψ does not affect $\hat{\chi}^y$. Now, as has been shown in [II],

---

[1] This operation is justified since (79) and (80) hold for the maximum-likelihood estimates as well as for true values of the parameters.

(99) $$W = M_{yy} - M_{yz} M_{zz}^{-1} M_{zy}$$

so that (98) becomes

(100) $$(M_{yy} - M_{yz} M_{zz}^{-1} M_{zy}) \hat{\chi}^z = \psi e$$

or

(101) $$\hat{\chi}^y = (M_{yy} - M_{yz} M_{zz}^{-1} M_{zy})^{-1} \psi e ,$$

which implies

(102) $$\tilde{q}^y = \hat{\chi}^y$$

since the expressions in (101) and (94) are identical (except possibly for φ and ψ which are without effect on $\tilde{q}^y$ and $\hat{\chi}^y$).

II. To show that $\tilde{q}^z = \hat{\chi}^z$, we first note, as stated in (92), that

(103) $$\tilde{q}^z = - M_{zz} M_{zy} \tilde{q}^y .$$

Since, as is shown in [II], the maximum-likelihood estimate P of π is given by

(104) $$P = - M_{yz} M_{zz}^{-1} ,$$

the maximum-likelihood estimate $\hat{\chi}^z$ of χ^z is obtained by substituting P from (104) and $\hat{\chi}^z$ from (101) into the sample counterpart of (80), thus obtaining

(105) $$\hat{\chi}^z = - M_{zz}^{-1} M_{zy} \hat{\chi}^y ,$$

with the symmetry of M_{zz} taken into account.

Hence, by virtue of (102),

(106) $$\hat{\chi}^z = - M_{zz}^{-1} M_{zy} \tilde{q}^y ,$$

where the right-hand member is identical with that in (103). Thus

(107) $$\tilde{q}^z = \hat{\chi}^z ,$$

and the second part of the proof is completed.

3.2.2. Lags present. No distinction is made in the derivation of the maximum-likelihood (or least-squares) estimates between "fixed variates" and lagged values of the endogenous variables Y. Hence we may regard the z's in section *3.2* as being "predetermined"[1] (i.e., including lagged y's). The stochastic difference equations are covered by the above proof.

The only word of warning concerns the nature of the *initial observation* or observations. An example will clarify this point. Let there be only one variable y and a sample

$$(y_1, \ldots, y_T). \tag{108}$$

We assume that y_t satisfies a stochastic difference equation of, say, first order:

$$y_t = \chi y_{t-1} + u_t , \tag{109}$$

where the u's are normally independently distributed with a zero mean and a constant variance.

To specify the problem completely, we must still provide information concerning the (marginal) distribution of y_1 (i.e., the initial observation). Now if y_1 is assumed to be a fixed variate, the maximum-likelihood estimate of χ is identical with the least-squares estimate. But if the (marginal) distribution of y_1 equals the (marginal) distribution of any other y_t (i.e., y_1 has a zero mean and a variance $\sigma_{y_1}^2 = \sigma_u^2/(1-\chi^2)$, the resulting maximum-likelihood estimate of χ is different from that obtained from least-squares.[2] This discrepancy will, in general, occur whenever the distribution of y_1 depends on χ.

In most economic problems it is the writer's opinion that it is not realistic to consider y_1 as a fixed variate; the only merit this assumption seems to have is that of expediency in that it leads to least-squares estimates.[3]

---

[1] See section *1.1*.

[2] [Koopmans, 1942].

[3] As the sample size increases, the maximum-likelihood estimate obtained for the case where y_1 is not a fixed variate will tend stochastically to the least-squares estimate.

3.3. *Regression Estimates under A Priori Restrictions*

The proofs of the optimal properties of the least-squares estimates $\tilde{q}$ of the regression coefficients χ were implicitly based on the following assumption: the a priori restrictions (if any) on the nature of the structure[1] S do not imply any restriction on the nature of the distribution F of the observed variates.[2]

Now consider the case where the a priori restrictions on the nature of the structure S are such that F itself is subject to restrictions.

$$F \equiv F^{C}_{*|**}(y^* \mid y^{**};\ z;\ \chi, \sigma^2)\ F^{M}_{**}(y^{**};\ z;\ \theta^M)\,. \tag{110}$$

(The regression in $F_{*|**}$ is assumed linear in y^{**} and z, with regression coefficients χ, and homoscedastic with variance σ^2.)

We may then distinguish three cases of restrictions[3] on F (implied by the a priori restrictions on S):

The restrictions on F do *not* involve χ:

$$\psi_{(\chi)}(\sigma^2;\ \theta^M) = 0\,. \tag{111}$$

The optimal properties of $\tilde{q}$ are not affected.

The restrictions on F involve χ *only:*

$$\psi_{\chi}(\chi) = 0\,. \tag{112}$$

Then it is easily shown that the least-squares method, *where minimization is carried out subject to* (112), gives results equivalent to both best unbiased linear and maximum-likelihood estimates when the latter also take (112) into account (whenever the equivalence

---

[1] In the linear normal case $S = (\mathrm{A}, \Sigma^{u})$ where A is the structural coefficient matrix and Σ^{u} the covariance matrix of the disturbances. As an example of restrictions on S one may mention the specification of zeros in A or Σ^{u}.

[2] In the linear normal case, normality of F is postulated, but the first two moments remain unrestricted.

[3] The discussion in the remainder of this section treats restrictions in the form of equalities only; the extension to inequalities is straightforward.

would exist in the absence of restrictions).

Restrictions involve both the regression coefficients, and some other parameters of F (σ^2 or θ^M or both); they may then be written as

$$\psi(\chi, \sigma^2; \theta^M) = 0. \tag{113}$$

In this case the equivalence of least-squares with best unbiased linear and maximum-likelihood estimates, in general, no longer holds.

It may be remarked that in the latter two cases, if the existing restrictions are ignored, the unbiasedness or consistency is not removed, but the variance of the estimates is no longer minimal.

VII. THE EQUIVALENCE OF MAXIMUM-LIKELIHOOD AND LEAST-SQUARES ESTIMATES OF REGRESSION COEFFICIENTS

BY T. C. KOOPMANS

The equivalence of maximum-likelihood and least-squares estimates, under the conditions assumed by Hurwicz [VI–3.2], can also be established on the basis of concepts and propositions developed in other contributions to this volume. The equation[1] {(78)} of which we desire to estimate the coefficients may be combined with the equations {(77)} with $g = 1, \ldots, K_y$ (i.e., omitting the first equation with $g = 0$) to give a system

(1)

$$(1.0)\quad y_0 - \overset{y}{\chi}_1 y_{1t} - \cdots - \overset{y}{\chi}_{K_y} y_{K_y t} - \overset{z}{\chi}_1 z_{1t} - \cdots - \overset{z}{\chi}_{K_z} z_{K_z t} = v^*_{0t},$$

$$(1.1)\quad y_{1t} - \pi_{11} z_{1t} - \cdots - \pi_{1K_z} z_{K_z t} = v_{1t},$$

$$\vdots$$

$$(1.k')\quad y_{K_y t} - \pi_{K_y 1} z_{1t} - \cdots - \pi_{K_y K_z} z_{K_z t} = v_{K_y t},$$

in which the $z_1, \ldots, z_{K_z}$ are fixed variables, and the $v^*_{0t}, v_{1t}, \ldots, v_{K_y t}$ have the same joint normal distribution for each value of t. In particular, since the first equation (1.0) in (1) represents the regression {(78)} of one variable (y_0) in a normal system on the others ($y_1, \ldots, y_{K_y}$), the quantity

$$(2)\quad v^*_0 = y_0 - \mathcal{E}(y_0 \mid y_1, \ldots, y_{K_y}; z_1, \ldots, z_{K_z})$$

is distributed independently of $y_1, \ldots, y_{K_y}$, and therefore also

[1] Equation numbers in braces { } refer to formulae in article [VI] by Hurwicz.

independently of $v_1, \ldots, v_{K_y}$,

$$\mathcal{E}\, v_0^*\, v_k = 0, \qquad k = 1, \ldots, K_y. \tag{3}$$

In fact, this condition alone uniquely identifies[1] the equation (1.0) within the system (1), given the fact that each of the remaining equations (1.1), ..., (1.k′) is uniquely identified by the ("reduced"[2]) form prescribed for it.

Because of the independence of v_0^* from $v_1, \ldots, v_{K_y}$, and because of the special form of the matrix of coefficients of $y_0, y_1, \ldots, y_{K_y}$ in (1), the system (1) meets the conditions, stated elsewhere,[3] for a factorization

$$\begin{aligned} F = {} & F_1(\chi_i^y, \chi_j^z, \sigma_{v_0^*}; y_{it}, z_{jt}) \\ & \times F_2\{\pi_{gk}\ (g \geq 1),\ \omega_{gh}\ (g, h \geq 1),\ y_{it}, z_{it}\} \end{aligned} \tag{4}$$

of the likelihood function.[4] Here each factor depends, besides on the observed variables, on a different set of parameters (as indicated), relating to the equation (1.0) and to the set of equations (1.1), ..., $(1.K_y)$, respectively. The two factors can therefore be maximized separately. In particular, the variables $y_1, \ldots, y_{K_y}$ can in equation (1.0) be regarded as "predetermined"[5] by the equations (1.1), ..., $(1.K_y)$. That is, for purposes of maximizing the first factor

$$F_1 = \text{constant} \times \sigma_{v_0^*}^{-1} \exp - \frac{1}{2\sigma_{v_0^*}^2} \sum_t (v_{0t}^*)^2 \tag{5}$$

of the likelihood function, where v_{0t}^* is given by (1.0), these variables can be treated in (1.0) as if they were fixed variables. In

---

[1] See [II, Definition 2.4.4.2].

[2] See [II-*3.1.6*].

[3] See [XVII-*3 & 6*].

[4] *F* here denotes the probability *density* in the sample space, not the cumulative distribution function.

[5] See [XVII-*6*].

view of the form of (5), this establishes the equivalence of maximum-likelihood and least-squares procedures with regard to the estimation of the parameters of (1.0).

To complete the link with other articles in this volume, the question should be raised as to what extent the equivalence just established remains valid if the system (1) is obtained by linear transformation from a system of structural equations as discussed in [II], the parameters of which are subject to a number of a priori restrictions, and if the likelihood function is maximized subject to those restrictions.

It is seen from formula (3.13) in [II] that the parameters of the second factor F_2 in the likelihood function (4) depend on the structural parameters through

$$\Pi = -B^{-1}\Gamma, \qquad \Omega = (B^{-1}\Sigma B'^{-1})^{-1}. \tag{6}$$

It should be noted that the elements in the zero row of Π and those in the zero row (and column) of Ω do not enter in F_2. On the other hand, the parameters in F_1, according to {(77)} and (1), satisfy

$$\chi_i^y = -\frac{\omega^{0i}}{\omega^{00}}, \qquad \chi_j^z = -\sum_{g=0}^{K_y} \chi_g^y \pi_{gj},$$
$$\sigma_{v_0^*}^2 = \sum_{g,h=0}^{K_y} \chi_g^y \omega_{gh} \chi_h^y, \tag{7}$$

where $\chi_0^y = -1$. It follows that the parameters in F_1 can be given any arbitrarily chosen values by suitable choice of the elements in the zero rows in Π and Ω. The parameters in F_1 and those in F_2 are, therefore, truly independent of each other.

This remains true under such restrictions on the structural parameters A, Σ as do not entail any restrictions on the possible values of Π and Ω. In the terminology of [II–3.2], the equivalence of maximum-likelihood and least-squares procedures remains valid under a priori restrictions that do not depress the likelihood function below its absolute maximum, because both the likelihood function[1] and the parameters of (1.0) depend on A, Σ only through the parameters Π and Ω which in that case are unrestricted.

[1] See either (4) and (7) above or formula (3.15) of [II].

On the other hand, maximization of the likelihood function under restrictions on A, Σ that do not permit the likelihood function to attain its absolute maximum, can only "by accident" be equivalent to the least-squares method, depending on how the vectors χ^y, χ^z of regression coefficients are affected by those restrictions (as explained in [VI–3.3]).

VIII. REMARKS ON THE ESTIMATION OF UNKNOWN PARAMETERS IN INCOMPLETE SYSTEMS OF EQUATIONS

BY A. WALD

1. *Formulation of the Problem*

Let $x_1, \ldots, x_N$ be N variables and denote by x_{nt} the value of x_n at the time point t where t can take any integral value. It is assumed that these variables satisfy a system of H $(H < N)$ stochastic equations

$$f_h(x_{1t}, x_{1,t-1}, \ldots, x_{1,t-\tau^\square}; \ldots; x_{Nt}, x_{N,t-1}, \ldots, x_{N,t-\tau^\square}; \alpha_1, \ldots, \alpha_P) = u_{ht}, \quad (h = 1, \ldots, H) \tag{1}$$

where $\alpha_1, \ldots, \alpha_P$ are unknown parameters, f_h is a given function of the variables $x_{n,t-\tau}$ $(n = 1, \ldots, N;\ \tau = 0, 1, \ldots, \tau^\square)$ and of the parameters α_p $(p = 1, \ldots, P)$, and $u_{1t}, \ldots, u_{Ht}$ are random variables. It is assumed that the distribution of the random vector $u_t = (u_{1t}, \ldots, u_{Ht})$ is independent of t and that the vectors $u_1, u_2, \ldots,$ are independently distributed. It is also assumed that the unknown distribution function of u_t is known to be an element of a given finite-parameter family Ω of cumulative distribution functions. Denote by $\theta_1, \ldots, \theta_Q$ the unknown parameters involved in the distribution function of u_t.

The problem considered here is to estimate all or some of the unknown parameters $\alpha_1, \ldots, \alpha_p; \theta_1, \ldots, \theta_Q$ on the basis of the observed values x_{nt} $(n = 1, \ldots, N; t = 1, \ldots, T)$.

Since $H < N$, the distribution of the random vectors $u_t = (u_{1t}, \ldots, u_{Ht})$, $t = 1, \ldots, T$, does not determine the distribution of the variables x_{nt}, $n = 1, \ldots, N$, but only imposes a restriction on the latter distribution. For this reason, the title of this note refers to incomplete systems of equations. It will appear from what follows that the estimation problems in incomplete systems are essentially different from those in complete systems discussed in other contributions to this volume.

2. *Existence of Consistent Estimates*

A basic problem is, of course, the question of the existence of consistent estimates of the unknown parameters. We shall say that a parameter point $\alpha = (\alpha_1, \ldots, \alpha_p)$ satisfies condition C if the following conditions are simultaneously fulfilled: (1) the distribution of the vector

$$u_t(\alpha) = \Big[f_1(x_{1t}, \ldots, x_{1,t-\tau^{\square}}; \ldots; x_{Nt}, \ldots, x_{N,t-\tau^{\square}}; \alpha_1, \ldots, \alpha_p),$$

$$\ldots, f_H(x_{1t}, \ldots, x_{1,t-\tau^{\square}}; \ldots; x_{Nt}, \ldots, x_{N,t-\tau^{\square}})\Big]$$

is independent of t; (2) the vectors $u_1(\alpha)$, $u_2(\alpha)$, $\ldots$, are independently distributed; (3) the distribution of $u_t(\alpha)$ is an element of Ω. It is clear that if there exist two parameter points α' and α'' such that both satisfy condition C, no consistent estimate of the parameter point α exists. On the other hand, if there exists one and only one parameter point α that satisfies condition C, then, under some further mild restrictions that we do not propose to discuss here, a consistent estimate of α will exist.

Whether or not there exist several parameter points α satisfying condition C depends on the joint probability distribution of the observable variables x_{nt} $(n = 1, \ldots, N; t = 1, \ldots, T)$. Thus, the existence of consistent estimates depends on the joint probability distribution of the observable variables. Since this distribution is usually unknown a priori, we cannot be sure that a consistent estimate exists. This difficulty, however, is not as serious as it would appear at first sight. In fact, instead of

point estimates, we are usually more interested in constructing a confidence region for the unknown parameters corresponding to a given confidence coefficient. We shall see in the next section that a confidence region can be obtained irrespective of the existence of consistent estimates. The only effect of the nonexistence of consistent estimates on the confidence region is that the diameter of the confidence region (maximum distance between two points of the confidence region) will not approach zero as the number of observations approaches infinity.

3. *Construction of Confidence Regions for the Parameter Point* α

To construct a confidence region for the parameter point α we may proceed as follows: Suppose that the prescribed confidence coefficient is δ. For any given parameter point $\alpha^0 = (\alpha_1^0, \ldots, \alpha_p^0)$ we construct a critical region $W(\alpha^0)$ of size $1 - \delta$ for testing the hypothesis that $\alpha = \alpha^0$. The critical region $W(\alpha^0)$ is a subset in the HT-dimensional space of the variables u_{ht} $(h = 1, \ldots, H;\ t = 1, \ldots, T)$ and we reject the hypothesis that $\alpha = \alpha^0$ if and only if the observed point $u_{11}(\alpha^0), u_{12}(\alpha^0), \ldots, u_{Ht}(\alpha^0)$ falls in $W(\alpha^0)$. Of course, $W(\alpha^0)$ is to be constructed in such a way that the probability measure of $W(\alpha^0)$ is equal to $1 - \delta$ under any distribution of u_t that is an element of Ω. Usually such a region $W(\alpha^0)$ can be constructed without any difficulty. Now consider all possible parameter points α^0 and the corresponding family of critical regions $W(\alpha^0)$. The set of all parameter points α^0 that are not rejected on the basis of this test procedure will form a confidence set with confidence coefficient δ.

Thus, we see that there is no serious difficulty in obtaining a confidence region for α. There will be, in general, infinitely many possible confidence regions, since $W(\alpha^0)$ can usually be chosen in infinitely many different ways. The problem of proper choice of $W(\alpha^0)$ is not yet solved. The theory of confidence intervals as developed by J. Neyman [1937] does not apply, since Neyman deals with the parametric case, while in our problem the class of all possible distribution functions of the observable variables x_{nt} $(n = 1, \ldots, N;\ t = 1, \ldots, T)$ compatible with the relations (1) cannot be described by a finite number of parameters owing to the fact that $H < N$.

4. *Some Examples*

In this section we shall give a few examples to illustrate the procedure for the construction of confidence regions. It should be emphasized, however, that the confidence regions described later in this section are by no means "best." As a matter of fact they may be very inefficient under certain conditions and it is not our intention to recommend them for practical use.

Consider the following problem: Let x_1, x_2, and x_3 be three observable variables and denote by x_{nt} $(n = 1, 2, 3;\ t = 1, 2, \ldots)$ the value of x_n at the time point t. Suppose that x_{1t}, x_{2t}, and x_{3t} satisfy the relation

$$x_{1t} + \alpha_2 x_{2t} + \alpha_3 x_{3t} + \alpha_0 = u_t , \tag{2}$$

where α_2, α_3, α_0 are unknown constants and u_1, u_2, ... are independently and normally distributed random variables with zero means and a common variance σ^2. We shall construct a confidence region for the parameters α_2, α_3, and α_0 on the basis of the observed values x_{nt} $(n = 1, 2, 3;\ t = 1, \ldots, T)$. For simplicity we shall assume that $T = 3V$, where V is a positive integer. Denote by $\bar{x}_n^m$ the arithmetic mean of the observations on x_n in the mth group of V observations, i.e.,

$$\bar{x}_n^m = \frac{1}{V} \sum_{t=(m-1)V+1}^{mV} x_{nt} , \qquad m = 1, 2, 3 . \tag{3}$$

Furthermore, let

$$s_m^2 = \frac{1}{V-1} \sum_{t=(m-1)V+1}^{mV} \left[(x_{1t} - \bar{x}_1^m) + \alpha_2 (x_{2t} - \bar{x}_2^m) + \alpha_3 (x_{3t} - \bar{x}_3^m) \right]^2 , \quad m = 1, 2, 3 . \tag{4}$$

It is clear that $(V-1)s_1^2/\sigma^2$, $(V-1)s_2^2/\sigma^2$, and $(V-1)s_3^2/\sigma^2$ are independently distributed, each having the χ^2-distribution with $V-1$ degrees of freedom. Thus, each of the expressions

$$(5) \qquad t_m = \frac{(\bar{x}_1^m + \alpha_2 \bar{x}_2^m + \alpha_3 \bar{x}_3^m + \alpha_0) V^{1/2}}{s_m}, \qquad m = 1, 2, 3,$$

has the t-distribution with $V - 1$ degrees of freedom, and t_1, t_2, and t_3 are independently distributed.

A confidence region for $(\alpha_2, \alpha_3, \alpha_0)$ with confidence coefficient δ can be constructed as follows: Let λ_δ be the value for which the probability that $-\lambda_\delta \leq t \leq \lambda_\delta$ is equal to $\delta^{1/3}$. Here t denotes a random variable that has the t-distribution with $V - 1$ degrees of freedom. The set of all parameter points $(\alpha_2, \alpha_3, \alpha_0)$ that satisfy simultaneously the following three inequalities,

$$(6) \qquad |t_m| \leq \lambda_\delta, \qquad m = 1, 2, 3,$$

forms a confidence region with confidence coefficient δ.

The confidence region given by (6) is certainly far from being the best possible one. There is a loss of efficiency in taking three different estimates s_1^2, s_2^2, s_3^2 for σ^2. A better procedure would be to combine these three estimates into a single one given by $s^2 = \frac{1}{3}(s_1^2 + s_2^2 + s_3^2)$. This, however, would make the variables t_1, t_2, and t_3 dependent and would complicate the derivation of a confidence region.

If the determinant

$$(7) \qquad \begin{vmatrix} \bar{x}_2^1 & \bar{x}_3^1 & 1 \\ \bar{x}_2^2 & \bar{x}_3^2 & 1 \\ \bar{x}_2^3 & \bar{x}_3^3 & 1 \end{vmatrix}$$

is bounded away from zero as $T \to \infty$, and if s_1^2, s_2^2, and s_3^2 are bounded functions of T, the diameter of the confidence region given by (6) will approach zero as $T \to \infty$. If the distribution of the vector $v_t = (x_{1t}, x_{2t}, x_{3t})$ is independent of t, if v_1, v_2, ... are independently distributed, and if the first two moments of x_{nt} exist, then the determinant (7) will converge stochastically to zero and the probability is one that the diameter of the confidence

set given by (6) will not approach zero as $T \to \infty$.

Another possible procedure for obtaining a confidence region is the following: Consider the expression

$$F = \frac{V \sum_{m=1}^{3} (\bar{x}_1^m + \alpha_2 \bar{x}_2^m + \alpha_3 \bar{x}_3^m + \alpha_0)^2}{(V-1)(s_1^2 + s_2^2 + s_3^2)} \frac{T-3}{3}. \tag{8}$$

This has the F-distribution with 3 and $T-3$ degrees of freedom. Let F_δ be a positive value for which the probability that $F \leq F_\delta$ is δ. Then the set of all parameter points α_2, α_3, α_0 that satisfy the inequality

$$F \leq F_\delta \tag{9}$$

will form a confidence set with confidence coefficent δ.

A confidence region for α_2, α_3, α_0 can also be obtained as follows: Consider the serial correlation

$$r = \frac{\sum_{t=1}^{T-1} \left[(x_{1t} + \alpha_2 x_{2t} + \alpha_3 x_{3t} + \alpha_0) \times (x_{1,t+1} + \alpha_2 x_{2,t+1} + \alpha_3 x_{3,t+1} + \alpha_0) \right]}{\sum_{t=1}^{T} (x_{1t} + \alpha_2 x_{2t} + \alpha_3 x_{3t} + \alpha_0)^2} \tag{10}$$

and let r_δ be a value such that the probability that $|r| \leq r_\delta$ is equal to δ. Then the set of all parameter points $(\alpha_2, \alpha_3, \alpha_0)$ for which $|r| \leq r_\delta$ will form a confidence set with confidence coefficient δ.

IX. ESTIMATION OF THE PARAMETERS OF A SINGLE EQUATION BY THE LIMITED-INFORMATION MAXIMUM-LIKELIHOOD METHOD

BY T. W. ANDERSON

1. *The Purpose of the Method*

In many economic studies the investigator wishes to study only a single equation out of a complete system. For example, an economist may be interested in the demand function for a given product. This interest would involve only one structural relation out of a set of such relations, among which might be a supply equation, an income equation, etc. If a particular equation is to be studied, a complete model must exist that explains the formation of all the variables in this equation that are considered as endogenous. It is possible to set up the complete model explicitly and, if all the equations are linear, estimate all of the coefficients by the methods described elsewhere in this volume.

To obviate this often complicated process a method has been developed which gives consistent estimates of the coefficients of a single equation without requiring estimates of the complete system. This method is known under two names, the "limited-information maximum-likelihood method" or the "reduced form method," each of which refers to an important aspect of the method.

The only information about the system that is required for this method is the knowledge of all the predetermined variables. The restrictions on the system that are imposed in the estimation procedure are that the coefficients of certain variables be zero in the equation estimated. The estimates are then derived on the basis of maximum likelihood. Since there is in general a larger number of effective a priori restrictions (i.e., depressing the likelihood function[1]) on the system than the restrictions on a single equation, this method of estimation is not as efficient as the maximum-likelihood method using all a priori restrictions on the complete system.

This article describes the estimates and gives some of their statistical properties. A full mathematical treatment is given in [T. W. Anderson and Rubin].

2. *Derivation of the Coefficients from the Population Reduced Form*

Let the equation in which we are interested be

$$\beta y_t' + \gamma z_t' = u_t , \tag{1}$$

where y_t is a row vector of H ($\leq G$) jointly dependent variables;[2] z_t is a row vector of all the predetermined variables in the complete system (K in number), and u_t is the random disturbance associated with this equation. Suppose that there are K^{**} coordinates of γ that are known and prescribed to be zero. Then (1) can be written as

$$\beta y_t' + \gamma^* z_t^* = u_t , \tag{2}$$

where z_t^* has K^* ($= K - K^{**}$) of the coordinates of z_t. Here γ^* is the vector of the K^* coefficients of the coordinates of z_t^*. We shall assume that (2) is identified by the zero coefficients prescribed through the omission of z_t^{**} and possibly certain jointly dependent variables which are not coordinates of y_t (if $H < G$).

---

[1] See [II-3.2.1].

[2] For the definition of *jointly dependent* and *predetermined* variables see [II-1.8].

This condition implies that[1] $K^{**} \geq H - 1$.

The reduced form of the complete system is the system written in the form of a regression on the predetermined variables. The part of the reduced form[2] of the complete system that involves the H coordinates of y_t in (2) is

$$y_t' = \Pi^* z_t^{*\prime} + \Pi^{**} z_t^{**\prime} + v_t' , \tag{3}$$

where v_t is a row vector of disturbances. If we premultiply (3) by β we obtain

$$\beta y_t' = \beta \Pi^* z_t^{*\prime} + \beta \Pi^{**} z_t^{**\prime} + \beta v_t' . \tag{4}$$

Since (4) must be identical with (2), we have

$$\gamma^* = - \beta \Pi^* , \tag{5}$$

$$0 = \beta \Pi^{**} . \tag{6}$$

If β is unknown, but Π^{**} is known, β can be found from (6) except for a factor of proportionality. In turn γ^* can be deduced from (5). The condition[3] that (2) be identified by zeros is that (6) can be solved for β. In terms of Π^{**} the condition is that its rank be $H - 1$.

If Ω, the matrix of variances and covariances of the v_t' of (3), is known, the formula

$$\sigma^2 = \beta \Omega \beta' \tag{7}$$

gives the variance of u_t. The normalization of β can be written as

$$\beta \Xi \beta' = 1 , \tag{8}$$

where Ξ may be Ω or may be some known constant matrix.

The basic idea of the "reduced-form method" is that if the rel-

---

[1] See the corollary to Theorem 2.2.1 in [II].

[2] See equation (3.11) in [II]. If $H < G$, the present equations (3) form a subset of equations (3.11) in [II].

[3] See Theorem 2.2.1 in [II].

evant part of the reduced form (3) is known together with the matrix Ω, the coefficients β and γ^* and the variance σ^2 can be deduced from (5), (6), (7), and (8).

If $K^{**} = H - 1$, then we can solve for β from the sample equivalent of (6). However, when $K^{**} \geq H - 1$, the (unrestricted) estimate of Π^{**} will be of rank H, and then the sample form of (6) cannot be solved for a nontrivial β. In the next section this difficulty is met by obtaining a maximum-likelihood estimate of β (under disregard of a part of the a priori information) which also estimates Π^{**} with rank $H - 1$.

3. *The Maximum-Likelihood Estimates*

To derive estimates of the vectors β and γ^* on the basis of maximum likelihood we assume that all the disturbances in the system are normally distributed with mean zero and covariance matrix independent of t and of the exogenous variables and that the disturbances are serially independent. Knowledge of all the predetermined variables in the system is assumed. The only restriction on the system that is taken into account is that the coefficients of z^{**} and possibly of some jointly dependent variables (if $H < G$) in the specified structural equation shall be zero. The likelihood function is

$$L = (2\pi)^{-\frac{1}{2}TH}(\det \Omega)^{-\frac{1}{2}T}$$

(9)

$$\times \exp\Big\{-\frac{1}{2}\sum_{t=1}^{T}(y_t - z_t^*\Pi^{*\prime} - z_t^{**}\Pi^{**\prime})\Omega^{-1}(y_t' - \Pi^* z_t^{*\prime} - \Pi^{**} z_t^{**\prime})\Big\}.$$

The restrictions on the parameters are

$$\beta\,\Pi^{**} = 0, \tag{10}$$

$$\beta' \mathrel{\underline{\underline{\Omega}}} \beta = 1. \tag{11}$$

To define the estimates of β, γ^*, and σ^2 we need the following statistics:

$$\frac{1}{T}\sum_{t=1}^{T} y_t' z_t = M_{yz} = [\, M_{yz^*} \quad M_{yz^{**}} \,], \tag{12}$$

the matrix of mean cross-moments of y and z, partitioned according to whether or not the coordinates of z are in the equation (2);

$$(13) \qquad \frac{1}{T}\sum_{t=1}^{T} z_t' z_t = M_{zz} = \begin{bmatrix} M_{z^*z^*} & M_{z^*z^{**}} \\ M_{z^{**}z^*} & M_{z^{**}z^{**}} \end{bmatrix},$$

the matrix of mean second-order moments of z;

$$(14) \qquad P = M_{yz} M_{zz}^{-1},$$

the matrix of sample regression coefficients which is partitioned as

$$(15) \qquad P = [\, P^* \quad P^{**} \,];$$

$$(16) \qquad \tilde{P}^* = M_{yz^*} M_{z^*z^*}^{-1},$$

the matrix of sample regressions of y on z^*;

$$(17) \qquad M_{z^\circ z^\circ} = M_{z^{**}z^{**}} - M_{z^{**}z^*} M_{z^*z^*}^{-1} M_{z^*z^{**}},$$

the mean moments of the residuals (z°) calculated from the regression of z^{**} on z^*;

$$(18) \qquad W = M_{yy} - M_{yz} M_{zz}^{-1} M_{zy},$$

which is T^*/T times the least-squares estimate of Ω (where $T^* = T - K$). Any of these quantities will be underlined to denote multiplication by T (for example $\underline{M}_{z^\circ z^\circ} = T M_{z^\circ z^\circ}$).

The second step in obtaining the estimates is to find the smallest root of the determinantal equation

$$(19) \qquad | P^{**} M_{z^\circ z^\circ} P'^{**} - \nu W | = 0$$

and the corresponding characteristic vector b defined by

$$(20) \qquad (P^{**} M_{z^\circ z^\circ} P'^{**} - \nu W) b = 0,$$

(21) $$b\, W\, b' = 1.$$

The population quantities corresponding to ν and b are 0 and $\beta\{(\beta\, \Xi\, \beta')/(\beta\, \Omega\, \beta')\}$.

The estimates of β, γ^*, and σ^2 are

(22) $$\hat{\beta} = \frac{b}{(b\, \Xi\, b')^{1/2}},$$

(23) $$\hat{\gamma}^* = -\hat{\beta}\, \tilde{P}^*,$$

(24) $$\hat{\sigma}^2 = \frac{(1+\nu)}{(b\, \Xi\, b')}.$$

For normalization $\beta\, \Omega\, \beta' = 1$, the above equation can be simplified. In this case σ^2 is unity and (22) is

(25) $$\hat{\beta} = \frac{b}{(1+\nu)^{1/2}}.$$

4. *Asymptotic Properties of the Estimates*

4.1. Consistency. Although the usual theorems concerning maximum-likelihood estimates cannot be applied in general to the estimates of section *3* because successive observations may not be independent, most of the usual asymptotic properties hold. In particular when the assumptions of sections *2* and *3* are true (i.e., assumptions for deriving maximum-likelihood estimates), the estimates are consistent and are asymptotically normally distributed. However, stronger results can be given. For certain weaker conditions on the system, asymptotic normality of the estimates holds, and under still weaker conditions, the estimates are consistent. We shall consider consistency under three sets of conditions where neither normality nor a constant covariance matrix is required.

(i) The estimates are consistent if (a) the stochastic process is stable, (b) the limit in probability of the matrix M_{zz} is nonsingular, (c) the ratio of the largest characteristic root of W_{yy}

to the samllest is bounded in probability, and (d)

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} u_t' z_t = 0, \tag{26}$$

where u_t is the row vector of disturbances.

(ii) The estimates of β and γ^* may be consistent even though not all of z^{**} in the system is used. Of course, at least $H - 1$ coordinates must be used.

(iii) In certain nonlinear systems, a linear equation (1) can be consistently estimated by this method. The important condition is that the part of the reduced form that is not a linear function of z^* has a population regression on some z^{**} that is of rank $H - 1$. There are, of course, certain identified nonlinear systems for which this does not hold.

4.2. *Asymptotic normality.* The requirements for asymptotic normality are stronger. Here, we require of a linear, stable system that the $(4 + \varepsilon)$th moments (for any $\varepsilon > 0$) of the disturbances be uniformly bounded in t. In addition there are certain requirements on z.

It should be pointed out that in most cases for the kind of linear systems that an economist is likely to postulate for his problems, these theorems are adequate for the desired asymptotic properties.

5. *Small-Sample Confidence Regions*

If all of the predetermined variables in the system are exogenous (that is, for purposes of obtaining distributions they can be treated as fixed in repeated samples), then confidence regions for the structural coefficients can be found from small-sample theory based on the normality of the disturbance in the specified equation. The regions are based on the idea that given a sequence of observations y_t, z_t^*, z_t^{**} $(t = 1, 2, \ldots, T)$ from a complete system containing (1), the numbers

$$b\, y_t' + c\, z_t^{*\prime} = w_t \tag{27}$$

will behave like observations from a normal distribution if

(28) $$b = \beta,$$

(29) $$c = \gamma^*.$$

Roughly speaking, if b differs from β and c from γ^* by small amounts, then the sequence $\{w_t\}$ will not be much different from $\{u_t\}$. The regressions of w_t on z_t,

(30) $$b\, M_{yz}\, M_{zz}^{-1} + c\,(M_{z^*z})M_{zz}^{-1},$$

are normally distributed if (28) and (29) hold true. Then

(31) $$b\, \underset{\sim}{M}_{yz^{**}}\, \underset{\sim}{M}_{z^{**}z^{**}}^{-1}\, \underset{\sim}{M}_{z^{**}z^{\circ}}\, \underset{\sim}{M}_{z^{\circ}z^{**}}^{-1}\, \underset{\sim}{M}_{z^{**}y}\, b' = X_{K^{**}}$$

has the χ^2-distribution with K^{**} degrees of freedom if

(32) $$\beta\, \Omega\, \beta' = 1.$$

Similarly,

(33) $$(b\, \underset{\sim}{M}_{yz} + c\, \underset{\sim}{M}_{z^*z})\, \underset{\sim}{M}_{zz}^{-1}\, (b\, \underset{\sim}{M}_{yz} + c\, \underset{\sim}{M}_{z^*z})' = X_K$$

has the χ^2-distribution with K degrees of freedom and

(34) $$b\, \underset{\sim}{M}_{yy}\, b' - b\, \underset{\sim}{M}_{yz}\, \underset{\sim}{M}_{zz}^{-1}\, \underset{\sim}{M}_{zy}\, b' = X_{T^*}$$

has the χ^2-distribution with T^* degrees of freedom and is independent of either (31) or (33). If the normalization of β is

(35) $$\beta\, \Xi\, \beta' = 1$$

for fixed Ξ, the three quantities given above must be multiplied by the reciprocal of $\beta\, \Omega\, \beta'$ in order to obtain statistics with χ^2-distributions.

For normalization (32), a confidence region for β of confidence $\varepsilon_1\, \varepsilon_2$ is given by the vectors b for which

(36) $$X_{K^{**}} \leq \chi^2_{K^{**}}(\varepsilon_1),$$

$$\underline{\chi}^2_{T^*}(\varepsilon_2) \leq X_{T^*} \leq \bar{\chi}^2_{T^*}(\varepsilon_2) , \tag{37}$$

where the limits are chosen so that the probability of (36) is ε_1, and that of (37) is ε_2 and

$$\underline{\chi}^2_{T^*}(\varepsilon_2) \leq T^* \leq \bar{\chi}^2_{T^*}(\varepsilon_2) . \tag{38}$$

To derive a region for β and γ^* we substitute

$$X_K \leq \chi^2_K(\varepsilon_1) \tag{39}$$

for (36).

If the normalization is (35) a region of size ε for β is (35) and

$$\frac{T^*}{K^{**}} \frac{X_{K^{**}}}{X_{T^*}} \leq F_{K^{**}\,T^*}(\varepsilon) , \tag{40}$$

where $F_{K^{**}\,T^*}(\varepsilon)$ is chosen from Snedecor's F-table [Snedecor, pp. 88 - 91] so that the probability of (40) is ε. A region for β and γ^* simultaneously is (35) and

$$\frac{T^*}{K} \frac{X_K}{X_{T^*}} \leq F_{K\,T^*}(\varepsilon) . \tag{41}$$

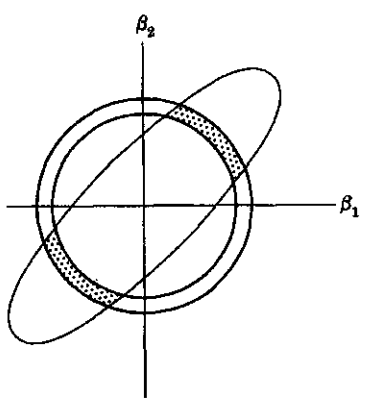


Figure 1.

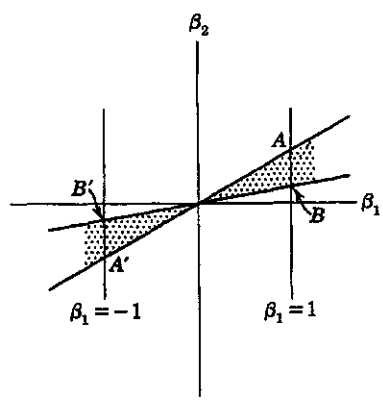


Figure 2.

The first-mentioned region for β of two coordinates is illustrated in Figure 1. The region for β of two coordinates when the normalization is

$$\beta_1^2 = 1 \tag{42}$$

is given in Figure 2 as the two segments $A'B'$ and AB.

The regions suggested converge in probability to the parameter points as $T \to \infty$ in very general systems. One type of condition is that the covariance matrix of successive disturbance vectors be such that the average of the matrices converges.

6. *Statistical Tests of the Assumption of Zero Coefficients*

6.1. The likelihood-ratio test. The hypothesis that the coefficient vector of z^{**} is zero is fundamental for the proposed method. If there are more than $H - 1$ zero coefficients prescribed, we can test the hypothesis that all of these coefficients actually are zero. In a sense we are testing information beyond that which is necessary for identification.

If the coefficient vector of z^{**} is zero in the specified equation, the rank of Π^{**} cannot be greater than $H - 1$. Hence, we shall test the hypothesis that the rank of Π^{**} is $H - 1$ against the set of alternative hypotheses that the rank is H. Of course, this test is unnecessary if $K^{**} < H$.

The likelihood-ratio criterion for this hypothesis is

$$(1 + \nu)^{-\frac{1}{2}T}, \tag{43}$$

where ν is the smallest root of (19). The asymptotic distribution of

$$T \log (1 + \nu) \tag{44}$$

is the χ^2-distribution with $K^{**} - H + 1$ degrees of freedom if the conditions for asymptotic normality of the estimates $\hat{\beta}$, $\hat{\gamma}^*$ are met.

6.2. A test based on small-sample theory. If $\beta\,\Pi^{**}$ is not zero, the quantities given in section 5 do not have the χ^2- and F-distributions as stated. For example, the inequality (40) is satisfied with probability less than ε if $\beta\,\Pi^{**}$ is different from zero. It can be shown that if the number of observations is made large enough the probability of (40) can be made as small as we please.

The minimum of the ratio $(X_{K**}/K^{**})/\ (X_{T*}/T^{*})$ for variations of b is given when b is set equal to a solution of (20). Then the value of the ratio is the smallest root of

$$\left| \frac{P^{**} M_{z^0 z^0} P'^{**}}{K^{**}} - \lambda \frac{W}{T^{*}} \right| = 0 . \tag{45}$$

If the inequality (40) is not satisfied by the minimum value of the ratio, it is not satisfied by the ratio for any choice of b. The probability, say δ, that

$$\lambda \geq F_{K^{**} T^{*}}(\varepsilon) \tag{46}$$

is greater than $1 - \varepsilon$ for $\beta\ \Pi^{**}$ equal to zero. The inequality (46) constitutes a test for the hypothesis that Π^{**} is of rank $H - 1$. However, the significance level, δ, is not exactly known. The inequality $\delta \geq 1 - \varepsilon$ shows that if the test is used as if it were at significance level $1 - \varepsilon$ it will be conservative.

If one attempts to compute a confidence region when there is disagreement with (46), one will find the confidence region given by (46) imaginary.

7. Advantages and Disadvantages of the Limited-Information Maximum-Likelihood Method

The significant advantage of the limited-information or reduced form method is the saving in computation when the coefficients in only one equation (or several, but not all, equations) out of a system are to be estimated. In this case there would be a greater computational expense to estimate the many more coefficients in the entire system. Moreover, simultaneous consideration of all restrictions in the system leads to more complicated formulas. It is possible to estimate the complete system, equation by equation, with the limited-information method. Then the laborious methods of [II–4] are not used.

In systems in which there are restrictions depressing the likelihood function in addition to the condition that Π^{**} be of rank $H - 1$, the limited-information method does not take into account all of the restrictions on the system while estimating the coefficients of one equation. This fact leads us to believe that the

limited-information method is not as efficient as the maximum-likelihood method, which utilizes restrictions on the entire system.

Some systems are restricted by bilinear restrictions (i.e., restrictions connecting two equations). If bilinear restrictions are imposed on an equation, then it may not be possible to estimate it by the limited-information method. On the other hand, all linear restrictions can easily be put into the form of prescribed zero coefficients.

X. SOME COMPUTATIONAL DEVICES

BY HAROLD HOTELLING

In solving normal equations it should be understood that it is highly desirable first to obtain the inverse matrix. This fact is often overlooked, but the necessity of computing standard errors and the frequent desirability of other uses of the inverse matrix – for example, adding to or taking from the set of predictors – are cogent arguments for adopting a standing rule that no large system of linear equations should ever be solved excepting by first finding the inverse of the matrix of their coefficients.

Direct calculation by methods such as those of Doolittle and Dwyer requires labor of the order of p^3, where p is the number of rows or of unknowns. Thus inverting a matrix of 50 rows requires only about one-eighth as much work as inverting a matrix of 100 rows. This consideration contributes interest to methods of inverting a matrix by partitioning it and inverting submatrices. A method of doing this has been set forth[1], involving four inversions of matrices of half the order of the given one. An improvement given by Waugh involves only two inversions of submatrices. In each case there are also multiplications and additions of submatrices, but these are more straightforward operations than inversion, even though the labor of multiplying two matrices of order p is also of the order of p^3. In partitioning a matrix for this purpose there is an advantage in dividing the rows into *equally numerous* groups, since when the sum of two positive numbers is fixed, the sum of their cubes is a minimum when they are equal.

In special cases there are further advantages in partitioning. Thus there may be whole blocks of zeros, or there may be triangles of zeros that make it easy to invert particular submatrices.

Another important method in matrix inversion is iteration. Several iterative methods are discussed in the paper cited [Hotelling, 1943 – 1]. The method recommended when a fairly good approximation C_0 to A^{-1} has been reached is to use

$$C_{m+1} = C_m(2 - A\,C_m)$$

[1] [Hotelling, 1943 - A]. Further results on matrix calculation are given in [Hotelling, 1943 - B and 1949] and [Ullman, 1944].

repeatedly to compute successively C_1, C_2, ..., while considering at each state the matrix of errors

$$D_m = 1 - A C_m .$$

The *norm* of a real matrix is the square root of the sum of the squares of the elements. It is an extremely useful means of setting an upper bound for the errors. Let the norm of D_0, denoted by $N(D_0)$, be k. Then if $k < 1$, we have the following upper bound for the errors in the approximations:

$$N(C_m - A^{-1}) \leq N(C_0) \frac{k^{2^m}}{1 - k} .$$

This formula shows two advantages of this particular method of iteration. One is that the limit of error is expressed only in terms of known numbers, without involving things calculated with a degree of accuracy not previously determined. The other advantage is that this limit of error decreases with great speed as m increases, because of the exponential of the exponential appearing in it. If k happens to be greater than unity, the method will sometimes converge nevertheless, and a new $k < 1$ will emerge at a later stage. If the method diverges, a better first approximation must be found in some other way.

Both iteration and partitioning are useful in least-squares problems with constraints. These problems are of two main kinds. The first is typified in surveying, where the Euclidean value for the sum of the angles of a triangle provides a set of side conditions that modify the solution.

The second kind is represented by the regression treatment of the analysis of variance with disproportionate class frequencies. Here the side conditions are introduced solely for convenience, to make definite the selection of a particular solution among an infinity of possible solutions. When this fact is realized, the arbitrary character of the side conditions becomes evident, and it will be recognized that essentially the same final results will be reached even if the usual simple conditions are considerably altered.

It is customary to take side conditions asserting that the simple sums of certain of the regression coefficients are zero. The labor of calculation is cut down if these are altered by intro-

ducing the marginal total frequencies as coefficients.

The whole system of normal equations and side conditions in such cases is treated in a new paper by the author with the help of partitioned matrices, which clarify many of the relationships. The matrix of the whole system has an inverse, of which a particular submatrix plays a part in this theory closely analogous to that of the inverse of the matrix of the normal equations in the ordinary nonsingular case. This analogy includes both the computational and the probability aspects. Furthermore, the matrix of the whole system can be inverted with the help of the iterative method set forth above. This will be advantageous whenever a good first approximation is available. A suitable source to which we may look for such a first approximation is the familiar treatment of proportionate frequencies.

PART TWO

PROBLEMS SPECIFIC TO TIME SERIES

XI. VARIABLE PARAMETERS IN STOCHASTIC PROCESSES: TREND AND SEASONALITY

BY LEONID HURWICZ[1]

1. When an economic time series is of any considerable length, the existence of *trend* is often suspected. The customary fashion of dealing with such a situation is either to eliminate the trend component by one of the several well-known methods or to treat *time* as a fixed variate entering the stochastic equations (or the regression equations) in the form of some given function, usually a polynomial.

On the other hand, when the data cover periods shorter than a year (days, months, quarters, etc.), *seasonal* behavior is often suspected.

In meteorological series there also exists, in addition to trends and seasonals, the diurnal component when, e.g., hourly data are available. The methods used in this case are analogous to those in treating trend – elimination (by subtraction or division) and fixed-variate regression.

These methods, presupposing as they do an additive or multiplicative effect of trend and seasonality,[2] are bound to fail in all but the simplest cases.

The experience in dealing with many time series, especially those in meteorology, indicates the need for a more flexible and general approach. In the present note the essentials of such an approach are sketched. The note is incomplete in many respects: among its deficiencies is the lack of proof of the consistency of the maximum-likelihood estimates used. For the sake of simplicity the discussion is restricted to time series in one variable only.

---

[1]Part of the work on this paper was done in 1945 - 46 during the author's tenure of the Guggenheim Memorial Fellowship. Some of the problems considered arose in connection with the author's research at the Institute of Meteorology at the University of Chicago.

[2]Diurnal effects are fully analogous to seasonals and will not be discussed separately.

The generalization to several variables is straightforward.

2. Let there be an infinite time series

$$x_\infty = (\ldots, x_{t_0-1}, x_{t_0}, x_{t_0+1}, \ldots) \tag{2.1}$$

such that every finite segment x of x_∞ has the probability density function $f(x)$. The function f is at this stage assumed to be *factorable*[1] (this assumption will later be relaxed); thus there exists p such that

$$f(x) = \prod_\tau f_\tau(x_\tau^{(p)}), \qquad \tau = \ldots, t_0 - 1, t_0, t_0 + 1, \ldots, \tag{2.2}$$

where

$$x_\tau^{(p)} = (x_\tau, x_{\tau+1}, \ldots, x_{\tau+p}), \qquad 1 \leq p < \infty. \tag{2.3}$$

3. We shall now indicate the possibility of imposing restrictions on the nature of

$$f = (\ldots, f_{t_0-1}, f_{t_0}, f_{t_0+1}, \ldots). \tag{3.1}$$

In the very simplest case we might have

$$f_\tau = f_0 \qquad \text{for all } \tau. \tag{3.2}$$

But the case of interest is when f_τ changes with τ. If the change is strictly *periodic*, so that

$$f_\tau = f_{\tau+k} \qquad \text{for all } \tau, \tag{3.3}$$

where k is some integer, we shall speak of *seasonal* fluctuations of f_τ.[2] Clearly, (3.2) is a special case of (3.3) with $k = 1$.

On the other hand, there may be *nonperiodic* variations in f_τ; these may be referred to as *trend*, although, in the narrow sense

---

[1] This implies nonautocorrelated disturbances.

[2] More generally, f_τ might satisfy some given functional equation, say a difference equation of finite order.

of the word, "trend" usually implies some uniform change in the value of $\mathcal{E}(x_t)$.

4. In what follows we shall be concerned with the parametric case and assume that

$$f_\tau \equiv f(x_\tau^{(p)} \mid \theta_\tau^{(s)}) \tag{4.1}$$

where $\theta_\tau^{(s)} \equiv (\theta_{\tau 1}, \ldots, \theta_{\tau s})$ is a parameter vector variable in time. It is assumed that the form of f does not change with time. [$x_\tau^{(p)}$ was defined in (2.3).]

5. Let the sample $O_T = (x_1, \ldots, x_T) = x_1^{(T-1)}$, where $T > p$, be drawn from the universe given by (2.2) and (4.1) after the initial $p - 1$ values of x have been fixed.[1]

Then the likelihood function Φ_T of the sample is

$$\Pr(O_T) = \prod_{\tau=0}^{T-p} f(x_{\tau+p}^{(p)} \mid \theta_\tau^{(s)};\ x_1^{(p-2)}) \tag{5.1}$$

with $x_1^{(p-2)}$ fixed.

Before we proceed with estimation of the $\theta_\tau^{(s)}$, we shall introduce assumptions concerning the behavior of $\theta_\tau^{(s)}$ in time. (The superscript s will be omitted where no danger of confusion exists.)

In the *seasonal* (periodic) case, corresponding to (3.3) we would have

$$\theta_\tau \equiv \theta_{\tau+k}, \tag{5.2}$$

which implies that all θ's can be expressed in terms of the first k θ's, i.e., $(\theta_{\tau_0}, \theta_{\tau_0+1}, \ldots, \theta_{\tau_0+k-1})$, or, say, $\alpha \equiv (\alpha_1, \ldots, \alpha_q)$, where there is a one-to-one correspondence between the components of of α_i and of the first k θ's. Thus the problem is reduced to that of estimating the components of α.

Similarly in the case of *trend* (in the broad sense given to this term earlier) it is assumed that the components of θ_τ are

---

[1] Alternatively, the initial value vector $x_1^{(p-2)}$ might be considered stochastic and assigned some given probability density function. This course, while more realistic, is less expedient since it does not lead to least-squares estimates when the maximum-likelihood criterion is used. Cf. [VI].

given functions of time (e.g., polynomials), say,

$$\theta_\tau \equiv \theta(\tau; \alpha), \tag{5.3}$$

and the problem is again reduced to that of estimating α.

Thus both in the case of seasonals (5.2) and trend (5.3), the likelihood function Φ_T given in (5.1) may be expressed in terms of α and τ and written

$$\Pr(O_T) = \prod_{\tau=0}^{T-p} f(x_{\tau+p}^{(p)} \mid \tau; \alpha; x_1^{(p-2)}), \tag{5.4}$$

and, if the maximum-likelihood criterion is used, $\Pr(O_T)$ is to be maximized with respect to α.[1]

6. As an example, consider the stochastic difference equation

$$x_t = \sum_{j=1}^{p} \theta_{jt} x_{t-j} + \theta_{0t} + \theta_{p+1,t} u_t, \tag{6.1}$$

where the u's are independently and normally distributed [in view of the presence of θ_{0t} and $\theta_{p+1,t}$, u_t may, without any loss of generality, be assumed to have the normal distribution $N(0,1)$ with zero mean and unit variance]; the behavior of the θ's is given by

$$\theta_{it} = \theta_{i,t+k}, \qquad i = 0, 1, \ldots, p+1, \tag{6.2}$$

where k may be either a year or a day (24 hours).

This type of equation is strongly suggested by meteorological data which exhibit seasonal (and diurnal) fluctuations not only in their "normal (i.e., mean) values," but also in their variances and lag-correlation coefficients.

Thus, for instance, in a series consisting of many years of monthly data it is found that the lag correlation of, say, January with February may differ from that of February with March, etc. Moreover, even for the 12-month lag the correlation of, say, the Januaries of two successive years differs from the corresponding 12-month lag correlation for February. This suggests an equation of the type (6.1) with $p > 1$. (For $p = 1$ the 12-month lag corre-

[1] The alternative procedure would be to apply Lagrange's method and to maximize $\Pr(O_T)$ subject to restrictive side relations.

lations do not vary from month to month.)

7. The maximum-likelihood equations can be set up easily, provided the variable coefficients θ_τ are linear in the α's. We shall show this in two examples: one of trend and one of seasonal fluctuations.

7.1. Trend. Let the behavior of the time series be described by a first-order stochastic difference equation

$$x_t = (\alpha_0 + \alpha_1 t)x_{t-1} + \alpha_2 + u_t , \tag{7.1}$$

where the u's are independently normally distributed with a zero mean and a common variance. (It will be observed that in this case the trend affects lag regression and not the expected value of x_t.) This may be considered as a special case of a one-equation stochastic system with one dependent variable x_t, and x_{t-1}, tx_{t-1}, and 1 as "predetermined" variables.[1]

Thus let

$$\begin{aligned} z_{1t} &\equiv x_{t-1}, & \alpha_0 &= \gamma_1 , \\ z_{2t} &\equiv t x_{t-1}, & \alpha_1 &= \gamma_2 , \\ z_{3t} &\equiv 1 , & \alpha_2 &= \gamma_3 . \end{aligned} \tag{7.2}$$

Then the likelihood function of the sample is

$$\Pr(O_T) = \prod_{t=2}^{T} f\{x_t - (\gamma_1 z_{1t} + \gamma_2 z_{2t} + \gamma_3 z_{3t})\} , \tag{7.3}$$

where f is normal with a variance σ^2 and a zero mean. Hence

$$\Pr(O_T) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^T \exp\{-\frac{1}{2\sigma^2}\sum_t (x_t - \sum_{i=1}^{3} \gamma_i z_{it})^2\} , \tag{7.4}$$

and the maximum-likelihood estimates $\hat{\gamma}$ of γ are obtained by minimizing the sum of squares

---

[1] Cf. [II-1.8].

$$S = \sum_t (x_t - \sum_i \gamma_i z_{it})^2 . \tag{7.5}$$

This is an example of equivalence of the maximum-likelihood and least-squares method. (The equivalence would have been absent if x_1 had not been assumed fixed!)

7.2. *Seasonal.* A strictly analogous seasonal situation is given by

$$x_t = (\alpha_0 + \alpha_1 \sin t) x_{t-1} + \alpha_2 + u_t , \tag{7.6}$$

where the period is 2π.

Here let

$$\begin{aligned} z_{1t} &\equiv x_{t-1}, & \alpha_0 &= \gamma_1 , \\ z_{2t} &\equiv (\sin t) x_{t-1}, & \alpha_1 &= \gamma_2 , \\ z_{3t} &\equiv 1, & \alpha_2 &= \gamma_3 , \end{aligned} \tag{7.7}$$

and the procedure is exactly the same as in the preceding example: the maximum-likelihood criterion again leads to the least-squares procedure.

Clearly no obstacles will be encountered as long as the equations remain *linear in the unknown coefficients.*

One problem which may arise is that of extending the available proofs of consistency of the maximum-likelihood estimates of the α's so as to cover the "explosive" (or, in general, undamped) situations. For a very special case this has been accomplished in [XIV].

8. In this section we consider a stochastic difference equation of order p whose coefficients are periodic with period k, and the same is true of the variance of the normally distributed "disturbance." This is the case described by (6.1).

Assuming for simplicity that $T = nk$, where n is an integer, we obtain the likelihood function

$$\Pr(O_T) = (2\pi)^{-\frac{1}{2}T} \prod_{j=1}^{k} \sigma_j^{-n} \exp\{ - \frac{1}{2} \sum_{j=1}^{k} q_j \} , \tag{8.1}$$

where

$$(8.2) \qquad q_j = \sigma_j^{-2} \sum_{r=0}^{n-1} (x_{j+rk} - \sum_{i=1}^{p} \alpha_{ji}\, x_{j+rk-i})^2 .$$

In the case of monthly observations $k = 12$. Let $j = 1$ correspond to January, $j = 2$ to February, etc. Then the *estimates of the* α*'s* in (8.2) would be the least-squares estimates of the regression coefficients in the regression with, say, the ith month as the "dependent" variable and some or all of the remaining eleven months as the "independent" variables.

Thus denoting the January observation in the μth year of the sample by $y_{1\mu}$, that for February by $y_{2\mu}$, etc., we proceed as follows: Minimizing with regard to the α's the jth sum of squares

$$(8.3) \qquad s_j^2 = \sum_{\mu=1}^{n} (y_{1\mu} - \sum_{s=1}^{p} \alpha_{js}\, y_{j-s,\mu})^2, \quad y_{j-s,\mu} \equiv y_{j-s+12,\mu} ,$$

we obtain the maximum-likelihood estimates $\hat{\alpha}_{js}$ for the given j and $s = 1, 2, \ldots, p$. When the $\hat{\alpha}_{js}$ are substituted in (8.3) we obtain $n\hat{\sigma}_j^2$.

This procedure is then simply that of estimating the regression of each month on its predecessors, and the 12 such regressions can be computed quite independently. This has been done in the past, although it is not easy to find justification for the methods used.

9. In practice it frequently occurs that the economist (or the meteorologist) has at his disposal data extending over a period of, say, 15 or 20 years and the number of parameters to be estimated is 3 or more. The significance of the results obtained from the use of annual figures cannot then be too reassuring. Therefore it becomes natural to attempt to utilize monthly (or quarterly) observations. Clearly, this means that seasonal phenomena (absent from annual figures) must now be taken into account. The simplest way to do this is to take the stochastic equations of the model used for annual figures and endow their parameters with seasonal variations.

Let, for instance, the "annual model" be

$$(9.1) \qquad x_t = \alpha_1\, x_{t-k} + \alpha_0 + \alpha_2\, u_t ,$$

where x_t is the annual average of x for the tth year, the time unit is one year, and the properties of u_t are given in section 6 [following equation (6.1)].

Then the simplest way of setting up a "seasonal model" would be to assume periodic behavior of the α's; thus

$$x_t = \alpha_{1t} x_{t-q} + \alpha_{0t} + \alpha_{2t} u_t , \tag{9.2}$$

where q is, say, one-twelfth of the time unit, i.e., a month, x_t is the monthly average, and u_t has the same properties as before.

If this were the case, it would be possible to apply the theory developed in earlier sections and no serious difficulties would arise. Unfortunately, however, such a passage from the annual to the seasonal model would not, in general, be justified. The crucial assumption, which might hold in (9.1)[1] but not in (9.2), is that of independence (or *nonautocorrelation*)[2] of the successive disturbances u_t.

To see why this should be the case it will suffice to remember that the disturbance itself is a stochastic process and it is not unreasonable to assume that its properties can be described by a stochastic difference equation (for simplicity, of first-order)

$$u_t = \beta u_{t-r} + v_t , \tag{9.3}$$

where the time unit r is small compared with q – say, a day. v_t, might be called the *secondary* disturbance (while u_t is *primary*), is a stochastic variable with a constant mean and standard deviation and is nonautocorrelated.

It can easily be seen that the autocorrelation $\rho(u_t, u_{t-\theta})$ of u_t tends to zero as the lag θ increases. It might still be appreciable for lags of one month but negligible for one-year lags.

Hence it may be quite legitimate to consider u_t as nonautocorrelated in the annual model (9.1), but the same assumption might be highly unrealistic for the (monthly) seasonal model (9.2).

It then becomes necessary to generalize the theory of seasonal models as developed in the earlier sections of this note by considering disturbances of the type (9.3) instead of the nonautocorrelated ones.

---

[1] Although even here it would not be very realistic. Cf. [Hurwicz, 1944].

[2] As long as we deal with normally distributed disturbances, the two terms are synonymous.

10.1. Before investigating the properties of models of the type (9.2) with u_t satisfying (9.3), we shall first consider the nonseasonal case, viz.,

$$x_t = \sum_{\tau=1}^{p} \alpha_\tau x_{t-\tau} + u_t \,, \tag{10.1}$$

where x_t is now the *instantaneous* ("daily") value of x at the time t, and u_t the *instantaneous* value of u at time t.

10.2. It is important to see that autocorrelation of the disturbance does not necessarily imply the loss of identification, provided the nature of the autocorrelation is properly specified. (If the autocorrelation function of the disturbance is unrestricted, no identification is, in general, possible.) We distinguish two cases, depending on whether the autocorrelation of the disturbance is of the autoregressive or moving-average type [Wold].

10.2.1. Autoregressive disturbance. Let

$$x_t + \alpha x_{t-1} = u_t \,, \tag{10.2}$$

and

$$u_t + \rho u_{t-1} = v_t \,, \tag{10.3}$$

where v_t is random (i.e., nonautocorrelated) and has a constant variance. Then we have

$$x_t + (\alpha+\rho) x_{t-1} + \alpha\rho x_{t-2} = v_t \,. \tag{10.4}$$

This may also be written as

$$x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} = v_t \,, \tag{10.5}$$

where

$$\begin{aligned} \beta_1 &= \alpha + \rho \,, \\ \beta_2 &= \alpha\rho \,. \end{aligned} \tag{10.6}$$

Now since the β's are functionally independent, it is possible to obtain α and ρ from them.[1] But (10.5) is an ordinary difference

[1] This is a case of *multiple* identification; cf. [IV-5.4.1] and [II-2.4.4].

equation with a random disturbance; hence there is no difficulty in obtaining the β's as well as the variance of v_t. Thus (10.2) is identifiable[1] in spite of the autocorrelation in u_t.

More generally, we have

$$\sum_{j=0}^{J} \alpha_j x_{t-j} = u_t, \qquad \alpha_0 = 1, \tag{10.7}$$

and

$$\sum_{i=0}^{I} \rho_i u_{t-i} = v_t, \qquad \rho_0 = 1, \tag{10.8}$$

where v_t has the same properties as before.

Then we obtain

$$\sum_i \rho_i \sum_j \alpha_j x_{t-i-j} = v_t, \tag{10.9}$$

or

$$\sum_{s=0}^{S} \beta_s x_{t-s} = v_t, \qquad \beta_0 = 1, \qquad S = I + J, \tag{10.10}$$

where

---

[1] Identification would not have been lost even if in addition to the "disturbance in the equation" u we also had had to deal with a "disturbance in the variable" (e.g., error of observation) w. Thus, let x_t in the equation above denote the "true" value while x_t^o is the observed value, so that

(a) $$x_t^o = x_t + w_t$$

where the w_t's are nonautocorrelated. Then

(b) $$x_\tau^o = w_\tau + v_\tau + \gamma_1 v_{\tau-1} + \gamma_2 v_{\tau-2} \quad \cdots;$$

also from (10.5) we have

(c) $$x_t^o + \beta_1 x_{t-1}^o + \beta_2 x_{t-2}^o = v_t + w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2}.$$

By multiplying (b) and (c) and taking the expectations of both sides, enough independent equations are available to determine all the unknown parameters in terms of the lag moments of x_t^o. Cf. [T. W. Anderson and Hurwicz].

$$\beta_s = \varphi_s(\rho_1, \ldots, \rho_I; \alpha_1, \ldots, \alpha_J), \qquad s = 1, 2, \ldots, S. \tag{10.11}$$

Explicitly,

$$\begin{aligned} \beta_0 &= \alpha_0 \rho_0 = 1, \\ \beta_1 &= \rho_0 \alpha_1 + \rho_1 \alpha_0, \\ \beta_2 &= \rho_0 \alpha_2 + \rho_1 \alpha_1 + \rho_2 \alpha_0, \\ &\vdots \end{aligned} \tag{10.12}$$

Since there are as many β's as there are ρ's and α's together, a necessary and sufficient condition for the identification of (10.7) is that the Jacobian of (10.12) should not vanish identically. This was seen to be the case for $I = J = 1$, where the β's were obviously independent, but it can also be proved more generally.

10.2.2. Moving-average disturbance. In the previous two sections of this note the disturbance u was assumed to possess autocorrelation of the *autoregressive* type, expressed in (10.8). Another case of considerable interest is autocorrelation of the *moving-average* type where

$$u_t = \sum_{\tau=0}^{R} \gamma_\tau v_{t-\tau}, \qquad \gamma_0 = 1, \tag{10.13}$$

with random v's. In this case we have

$$\operatorname{cov}(u_t, u_{t-R-k}) = 0, \qquad k \geq 1, \tag{10.14}$$

where $\operatorname{cov}(x,y)$ denotes the covariance of x and y.

Using the fact that

$$x_t = u_t + \delta_1 u_{t-1} + \cdots, \tag{10.15}$$

where the δ's are easily obtainable functions of the α's, and writing

$$\text{(10.16)} \qquad \operatorname{cov}(x_t, x_{t-p}) = \mu_p ,$$

we easily find [with the help of (10.14)] that

$$\text{(10.17)} \qquad \sum \alpha_j \mu_{p-j} = 0 , \qquad p \geq p_0 ,$$

for p_0 sufficiently large in relation to R.

Since the μ's are given if the distribution of the x's is known, the α's can always be determined. Similarly the lower μ's serve to evaluate the weights γ in (10.13), although a proof of functional independence should again be supplied to make this statement rigorous. Hence the moving-average type of autocorrelated disturbance does not destroy identifiability.

So far we have confined ourselves to single equations in one unknown. However, the situation does not seem to be fundamentally different in *simultaneous systems*. As an example we use the system given by Koopmans and Rubin [II-2,5,6].

Simplifying the notation, we have

$$\text{(10.18)} \qquad \begin{aligned} L'_t &\equiv x'_t + \alpha_1 x'_{t-1} + \alpha_2 x''_{t-1} = u'_t , \\ L''_t &\equiv \beta x'_t + x''_t = u''_t \end{aligned}$$

which is undoubtedly identifiable when the u's are random.

Now let

$$\text{(10.19)} \qquad u_t^{(i)} + \rho^{(i)} u_{t-1}^{(i)} = v_t^{(i)} , \qquad i = ', \ '',$$

(v's random) and

$$\text{(10.20)} \qquad L_t^{(i)*} = L_t^{(i)} + \rho^{(i)} L_{t-1}^{(i)} = v_t^{(i)} .$$

In terms of the original parameters we have

$$\text{(10.21)} \qquad \begin{aligned} x'_t + (\alpha_1 + \rho')x'_{t-1} + \rho'\alpha_1 x'_{t-2} + \rho'\alpha_2 x''_{t-2} &= v'_t , \\ \beta x'_t + x''_t + \rho'\beta x'_{t-1} &= v''_t , \end{aligned}$$

or

$$
\begin{aligned}
x'_t \quad &+ \varepsilon_1 x'_{t-1} + \varepsilon_2 x''_{t-1} + \varepsilon_3 x'_{t-2} + \varepsilon_4 x''_{t-2} = v'_t, \\
\eta_1 x'_t + x''_t + \eta_2 x'_{t-1} &+ \eta_3 x''_{t-1} \qquad\qquad = v''_t.
\end{aligned} \tag{10.22}
$$

Clearly, the system (10.22) is identified, even without taking into account the additional restriction on the ε's and η's which follow from (10.21).

It should be emphasized that both (10.18) and (10.19) are very special cases, but it is not unreasonable to conjecture that in a large class of cases identification is not destroyed by introducing autocorrelated disturbances.

10.3. Now define the "monthly" average of x_t

$$
X_t = \frac{1}{H} \sum_{\tau=t-H+1}^{t} x_\tau \tag{10.23}
$$

and of u_t

$$
U_t = \frac{1}{H} \sum_{\tau=t-H+1}^{t} u_\tau, \tag{10.24}
$$

where H is the number of "days" in the "month."

We assume, as is usually the case in practice, that only the X_t but not the x_t are known from observation. These are given for non-overlapping time intervals of H "days" each:

$$
\ldots, X_{t_0}, X_{t_0+H}, X_{t_0+2H}, \ldots, \tag{10.25}
$$

which may be denoted by

$$
\ldots, Y_{t_0}, Y_{t_0+1}, Y_{t_0+2}, \ldots, \tag{10.26}
$$

respectively.

Now in order to estimate the unknown parameters (the mean and variance of v_t, β, and the α's) by the maximum-likelihood method, we must obtain the likelihood function for the Y's.[1] Since in the

[1] In general, of course, the system may be unidentified. This can be remedied by assuming a priori restrictions on the unknown parameters or by introducing, say, a fixed variate in the right-hand member of (10.1). In

case of normal v_t all the other variables are also normal, the likelihood function is specified if the means and the lag covariance matrix of Y_t are obtained. Making the means of v_t zero, we are left only with the problem of finding the lag covariance matrix of Y_t. (The case of a nonzero mean may be treated in an analogous manner.)

11. We first observe that

$$X_t = \sum_{\tau=1}^{p} \alpha_\tau X_{t-\tau} + U_t \tag{11.1}$$

so that

$$X_t = \sum_{\tau=0}^{\infty} \gamma_\tau U_{t-\tau}, \tag{11.2}$$

where the γ's are functions of the α's and may be obtained from the identity

$$\sum_{\tau=0}^{\infty} \gamma_\tau U_{t-\tau} \equiv \sum_{\tau=1}^{p} \alpha_\tau \sum_{\tau'=0}^{\infty} \gamma_{\tau'} U_{t-\tau-\tau'}. \tag{11.3}$$

Thus

$$\begin{aligned} \mathcal{E}(X_t X_{t-k}) &= \mathcal{E}\left[\left(\sum_{\tau=1}^{\infty} \gamma_\tau U_{t-\tau}\right)\left(\sum_{\tau'=0}^{\infty} \gamma_{\tau'} U_{t-k-\tau'}\right)\right] \\ &= \mathcal{E}\left[\sum_{\tau,\tau'}^{\infty} \gamma_\tau \gamma_{\tau'} U_{t-\tau} U_{t-k-\tau'}\right] \qquad (11.4) \\ &= \sum_{\tau,\tau'}^{\infty} \gamma_\tau \gamma_{\tau'} \mu_{t-\tau,\, t-k-\tau'}, \end{aligned}$$

where

$$\mu_{t-\tau,\, t-k-\tau'} = \mathcal{E}(U_{t-\tau} U_{t-k-\tau'}). \tag{11.5}$$

The μ's depend on σ^2 of v and β (as well as H) and may easily be evaluated with the help of (9.3) and (10.24).

---

the latter case the procedure would be essentially of the same type as, though more complex algebraically than, that given in the text.

It remains to obtain the lag moments of the Y's, that is, of every Hth X_t. One will note that the Y's satisfy a stochastic difference equation

$$Y_t = \sum_{\tau=1}^{p} \alpha'_\tau Y_{\tau - H\tau} + U'_t , \tag{11.6}$$

where U' is a linear combination of lagged values[1] of the U_t. The coefficients of this linear combination are functions of the α_τ and so are the α'_τ in (11.6).

These facts make it possible to obtain the lag moments of Y_t and, because of normality and the zero-means assumptions, the likelihood function of the observed variates is obtained.

12. The procedure carried out in sections *10* and *11* was based on the assumption that the parameters were constant. Now, in order to take account of *seasonal fluctuations,* we may make σ^2, β, or the α's vary with time t but subject to the restriction of periodicity. This will change the likelihood function, but its derivation follows the same pattern. Once the likelihood function has been obtained the unknown parameters may be estimated provided the system is identified.

13. The models of stochastic processes used so far have all been of the discrete type. It is not difficult to construct continuous models with properties analogous to those presented above. However, even for the nonseasonal case the study of estimation in continuous systems presents serious difficulties. A treatment of this problem is not yet available.[2]

Nevertheless, for the sake of completeness, an example will be given which is the continuous counterpart of (10.1) and its seasonal version.

This model may be written as an integral equation

$$x(t) = \int_{-\infty}^{t} \varphi(t - \tau)\, x(\tau)\, d\tau + u(t) , \tag{13.1}$$

where $u(t)$ has an autocorrelation function tending to zero as the

[1] Hence (11.6) is a difference equation with an autocorrelated disturbance of the moving-average type. See above, section *10.2.*

[2] Cf. [XVI].

lag increases.

To introduce seasonality we use a modified kernel, thus obtaining

$$x(t) = \int_{-\infty}^{t} \psi(t,\ t-\tau)\, x(\tau)\, d\tau + u(t), \tag{13.2}$$

where

$$\psi(t,\ s) = \psi(t+k,\ s). \tag{13.3}$$

To make the mathematics of (13.3) more manageable one might assume

$$\psi(t,\ s) = \psi_1(t)\, \psi_2(s). \tag{13.4}$$

XII. NONPARAMETRIC TESTS AGAINST TREND[1]

BY HENRY B. MANN

In testing against trend we are testing the hypothesis that the members of a certain sequence of random variables $x_1, \ldots, x_n, \ldots$ are distributed independently of each other, each with the same distribution, which we shall throughout this paper assume to be continuous. Although the null hypothesis is thus well defined, we are not in the same happy position with respect to the alternatives that we wish to admit. We have some rather vague ideas of a sequence where the variables tend to decrease, but no clear notion of how this decrease should be expressed in terms of distribution functions has as yet been advanced. It seems advantageous to eliminate from such a definition any tendency to decrease which has its roots in the dependence between successive observations. We, therefore, propose the following definition for a downward trend:

A sequence of random variables $x_1, x_2, \ldots$ will be said to have a downward trend if the variables are independently distributed so that x_i has the cumulative distribution function f_i and $f_i(x) < f_j(x)$ for every $i < j$ and every x.

An upward trend is similarly defined with $f_i(x) > f_j(x)$ for $i < j$.

If we are justified in restricting the null hypothesis to the form

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\,dy \tag{1}$$

and in restricting the alternatives to expressions of the type

$$f_i(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{x} e^{-\frac{(y-\mu_i)^2}{2\sigma^2}}\,dy\ , \tag{2}$$

[1]This is an abstract of a paper published in *Econometrica* under the same title [Mann].

where $\mu_i = P(ih)$ and P is a fixed polynomial of known degree with monotonically decreasing values $P(ih)$ for $i > 0$, $h > 0$, then the test against trend may be carried out in the usual fashion. That is to say, we can test the significance of the regression coefficients and this test is known to possess certain desirable properties. It is, for instance, in the case where P is a linear function, a uniformly most powerful test with respect to the alternatives considered. However, if we are not justified in assuming the null hypothesis to be of the form (1), then the customary procedure will not give the correct size of the critical region. Even if we are justified in assuming the null hypothesis to be of the form (1), but wish to admit as alternatives all downward trends as defined before, then we can not state that the usual procedure of testing the significance of the regression coefficients is in any respect superior to other procedures. In fact, we shall describe a test, based on ranks, which has maximum power with respect to a class of alternatives just as likely to occur in practical work as the alternatives (2). Since we very rarely have a knowledge of the actual distribution, it is of importance to discuss tests for which the size of the critical region does not depend on a priori assumptions about the distributions concerned.

A fairly simple test against trend is the T-test first introduced by Kendall [1938] for the testing of independence in a bivariate distribution. The statistic T counts, in a sample $x_1, \ldots, x_n$, the number of inequalities $x_i < x_j$ for $i < j$. One such inequality will be termed a reverse arrangement. The distribution of T is readily computed from the recursion

$$(3) \quad P_n(T) = P_{n-1}(T) + P_{n-1}(T-1) + \cdots + P_{n-1}(T-n+1),$$

where $P_n(T)$ is the number of permutations of n different values $x_1, \ldots, x_n$, which possess exactly T reverse arrangements. The recursion (3) was derived by M. G. Kendall. He also derived the variance of T and proved ultimate normality. The approach to normality is remarkably rapid.

The odd moments of T around its mean value $n(n-1)/4$ are 0. From the recursion (2) it is possible to derive the following difference equation for the even moments of T:

$$(4) \quad \mathcal{E}_n(x^{2i}) = \mathcal{E}_{n-1}(x^{2i}) + \binom{2i}{2} B_n^{(2)} \mathcal{E}_{n-1}(x^{2i-2}) + \ldots + B_n^{(2i)},$$

where $\mathcal{E}_n(x^{2i})$ is the $2i$th moment of T around its mean in permutations of n variables and

$$B_n^{(2j)} = \frac{1}{n}\sum_{K=0}^{n-1}\left(\frac{n-1}{2} - K\right)^{2j}.$$

From (4) we find

$$\sigma_n^2(T) = \frac{2n^3 + 3n^2 - 5n}{72}$$

and

$$\mathcal{E}_n(x^4) = \frac{100n^6 + 228n^5 - 455n^4 - 870n^3 + 625n^2 + 372n}{43.200}.$$

(4) can also be utilized to give a simple proof for the ultimate normality of T. This can be done in showing by mathematical induction that $\mathcal{E}_n(x^i)/\{\sigma_n(T)\}^i$ converges to the moments of a normal distribution.

To give lower bounds for the power of the T-test, we introduce a parameter λ_n defined by the equation

$$\mathcal{E}_n(T \mid H) - \frac{n(n-1)}{4} = \lambda_n \frac{n(n-1)}{2},$$

where $\mathcal{E}_n(T \mid H)$ is the expectation of T under the alternative H. Denote by $\sigma_n(T \mid H)$ the variance under the alternative H, by σ_{0n} the variance under the null hypothesis, by $P(T \leq \bar{T} \mid H)$ the probability that $T \leq \bar{T}$ if H is the true situation. If H is a trend, then the following inequalities can be derived:

$$(5) \qquad \sigma_n^2(T \mid H) \leq \frac{n(n-1)}{4} + \frac{n(n-1)(n-2)}{6} - \frac{\lambda_n^2 n^2 (n-1)}{4},$$

$$(6) \qquad P(T \leq \bar{T} \mid H) \geq 1 - \frac{\sigma_n^2(T \mid H)}{\left(\lambda_n \frac{n(n-1)}{2} + t_n \sigma_{0n}\right)^2},$$

for $\lambda_n\{n(n-1)/2\} \leq -t_n \sigma_{0n}$, where $\bar{T} = \{n(n-1)/4\} - t_n \sigma_{0n}$

defines t_n. From the ultimate normality of T it follows that t_n converges to a constant t if the size of the critical region is fixed.

For small values of n another inequality gives a better lower bound for the power. It can be shown that

$$P(T \leq \bar{T} \mid H) \geq \frac{-2\lambda_n n(n-1) - 4t_n \sigma_{0n}}{n(n-1) - 4t_n \sigma_{0n}}. \tag{7}$$

From (6) it follows that the T-test is consistent with respect to alternatives for which λ_n is negative and of order larger than $1/\sqrt{n}$. (6) and (7) can also be utilized to derive sufficient conditions for the unbiasedness of the T-test.

An example of a class of alternatives with respect to which the T-test is a most powerful test based on ranks is the following class C of alternatives H.

Let p_{ij} be the probability that x_i is the jth largest of the variables $x_i, x_{i+1}, \ldots, x_n$. If under an alternative H, $p_{ij} = a_i p^{j-1}$ $(p < 1)$, independent of the ranks of $x_1, \ldots, x_{i-1}$, then the T-test will be most powerful with respect to H. It is not known, however, whether there are such alternatives among the trends as defined in the introduction.

Another test that the author considers worth serious consideration is the K-test. This test is carried out as follows. We determine the smallest K for which

$$x_i > x_{i+j+K} \quad \text{for } j \geq 1,\ i = 1, \ldots, n-K. \tag{8}$$

We then fix $\bar{K}$ so that $P(K \leq \bar{K} \mid H_0)$ equals the size of the critical region. If then $K \leq \bar{K}$ we proceed on the hypothesis of a trend. If $K > \bar{K}$ we proceed as if the null hypothesis were true.

$P(K \leq \bar{K} \mid H_0)$ may be found for $K \leq 3$ from the following relations:

$$\begin{gathered} Q_n(1) = 1, \qquad Q_n(2) = Q_{n-1}(2) + Q_{n-2}(2), \\ Q_n(3) = Q_{n-1}(3) + Q_{n-2}(3) + 3Q_{n-3}(3) + Q_{n-4}(3), \end{gathered} \tag{9}$$

where $Q_n(\bar{K})$ is the number of permutations $x_{i1}, \ldots, x_{in}$ satisfying (8).

It is also easy to prove the relations

$$P_n(n - \bar{K}) = P_{2K}(\bar{K}) \quad \text{for } n \geq 2\bar{K}, \tag{10}$$

where $P_n(\bar{K}) = P(K \leq \bar{K})$ under the null hypothesis in samples of size n. It may further be shown that

$$P_n(n - \bar{K}) = \tag{11}$$

$$\sum \{[i_1 + 1][\max(i_1, i_2) + 2] \cdots [\max(i_1, i_2, \ldots, i_K) + \bar{K}]\}^{-1},$$

where Σ denotes summation over all permutations $i_1, \ldots, i_K$ of 1, 2, ..., $\bar{K}$.

Also

$$P_9(4) = \tag{12}$$

$$\sum{}' \{[\bar{\max}(i_1) + 1][\bar{\max}(i_1, i_2) + 2] \cdots [\bar{\max}(i_1, \ldots, i_5) + 5]\}^{-1},$$

where $\bar{\max}(i_1, \ldots, i_e) = \min[\max(i_1, \ldots, i_e), 4]$ and Σ' denotes summation over all permutations $i_1, \ldots, i_5$ of 1, 2, ..., 5 for which 1 precedes 5.

The relations (9), (10), (11), and (12) permit tabulation of $P_n(\bar{K})$ for $n \leq 9$. For higher values of n the following critical regions are available:

$$\begin{aligned}
&P_n(n - 5) = P_{10}(5) = 0.0098, && \text{for } n \geq 10,\\
&P_n(n - 4) = P_8(4) = 0.0284 && \text{for } n \geq 8\\
&P_n(n - 3) = P_6(3) = 0.0792 && \text{for } n \geq 6,\\
&P_n(n - 2) = 0.2083 && \text{for } n \geq 4,\\
&P_n(n - 1) = P_2(1) = 0.5 && \text{for } n \geq 2.
\end{aligned} \tag{13}$$

These are all the critical regions between 0.0098 and 0.5 possible for the K-test for $n \geq 10$. The K-test has the disadvantage

TABLE 1*

PROBABILITY OF OBTAINING A PERMUTATION WITH $T \leqq \bar{T}$ IN PERMUTATIONS OF n VARIABLES.

n \ $\bar{T}$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
3	167	500	833																			
4	042	167	375	625	833	958																
5	008	042	117	242	408	592	758	883	958	992												
6	001	008	028	068	136	235	360	500	500	640	765	864	932	972	992	999						
7	000	001	005	015	035	068	119	191	281	386	500	500	614	719	809	881	932	965	985	995	999	
8	000	000	001	003	007	016	031	054	089	138	199	274	360	452								
9	000	000	000	000	001	003	006	012	022	038	060	090	130	179	238	306	381	460				
10	000	000	000	000	000	000	001	002	005	008	014	023	036	054	078	108	146	190	242	300	364	431
$P(c)$	000	000	000	000	001	001	002	004	006	010	016	025	037	054	076	105	142	186	237	296	360	429

* Tabular values should be divided by 1000.

$$P(c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-c} e^{-x^2/2} dx, \qquad c = \left(\frac{n(n-1)}{4} - T - \frac{1}{2} \right) \Big/ \sqrt{\frac{2n^3 + 3n^2 - 5n}{72}}, \qquad n = 10.$$

TABLE 2

PROBABILITY OF OBTAINING A PERMUTATION WITH $K \leqq \bar{K}$ IN PERMUTATIONS OF n VARIABLES.

n \ $\bar{K}$	1	2	3	4	5	6	7
3	0.1667	0.5000					
4	0.0417	0.2083	0.5000				
5	0.0083	0.0667	0.2083	0.5000			
6	0.0014	0.0181	0.0792	0.2083	0.5000		
7	0.0002	0.0042	0.0246	0.0792	0.2083	0.5000	
8	0.0000	0.0008	0.0066	0.0284	0.0792	0.2083	0.5000
9	0.0000	0.0002	0.0016	0.0086	0.0284	0.0792	0.2083

that the choice of critical regions is rather limited.

It is easy to see that there are trend alternatives for which $P(x_i > x_{i+j+\overline{K}}) = 1$ for $j \geq 0$ $(i = 1, \ldots, n - K)$. The K-test has the power 1 with respect to such alternatives, while other tests, for instance the T-test or the significance tests ordinarily used, may have a considerably lower power. This seems remarkable in view of the fact that the K-test is a test based on ranks and in fact often uses only a fraction of the sample.

XIII. TESTS OF SIGNIFICANCE IN TIME-SERIES ANALYSIS

BY R. L. ANDERSON

1. Suppose we have data on the monthly prices of some product in a given locality over a period of n years. We assume that the prices P_{ij} in the ith month and jth year ($i = 1, 2, \ldots, 12$; $j = 1, 2, \ldots, n$) are independently distributed normal variables with a common variance σ^2 and means μ_{ij} where

$$\mu_{ij} = \mu + \alpha_i + \beta_j \tag{1}$$

and

$$\sum_i \alpha_i = \sum_j \beta_j = 0, \tag{2}$$

so that the likelihood function for the sample is

$$(\sqrt{2\pi}\,\sigma)^{-12n} \exp\{-\frac{1}{2\sigma^2}\sum_{j,\,i}(P_{ij} - \mu_{ij})^2\}. \tag{3}$$

2. We test the hypothesis H_0 given by

$$\alpha_i = 0, \qquad i = 1, 2, \ldots, 12. \tag{4}$$

Thus H_0 specifies that there is no monthly ("seasonal") pattern in the variations of the P_{ij}. The likelihood-ratio criterion leads to the use of the statistic

$$F = \frac{[(12-1)(n-1)]}{(12-1)}\,\frac{\hat{\sigma}'^2 - \hat{\sigma}^2}{\hat{\sigma}^2} = (n-1)\,\frac{\hat{\sigma}'^2 - \hat{\sigma}^2}{\hat{\sigma}^2}, \tag{5}$$

which has Snedecor's distribution with 11 and $(n-1)$ degrees of freedom when H_0 is true.[1] The symbols $\hat{\sigma}'^2$ and $\hat{\sigma}^2$ have the following

[1] For proof, cf., for instance, [Wilks, p.178 ff.].

meaning: $\hat{\sigma}^2$ is the maximum-likelihood estimate of σ^2 obtained by maximizing [subject to (2)] the expression in (3) with regard to the α_i, β_j, and to σ^2 and is given by

$$\hat{\sigma}^2 = \frac{1}{12n} \sum_{i,j} (P_{ij} - m_{ij})^2 , \tag{6}$$

where

$$\begin{aligned} m_{ij} &= \hat{\mu}_{ij} = m + a_i + b_j , \\ m &= \hat{\mu} = \frac{1}{12n} \sum_{i,j} P_{ij} , \\ a_i &= \hat{\alpha}_i = \frac{1}{n} \sum_{j} P_{ij} - m , \\ b_j &= \hat{\beta}_j = \frac{1}{12} \sum_{i} P_{ij} - m . \end{aligned} \tag{7}$$

$\hat{\sigma}'^2$ is the maximum likelihood estimate of σ^2 on the assumption that H_0 is true, and can thus be obtained by maximizing [subject to (2)] the expression in (3) with regard to the β_j and to μ and σ^2 after all the α_i have been set equal to zero.

We then obtain

$$\hat{\sigma}'^2 = \frac{1}{12n} \sum_{i,j} (P_{ij} - m - b_j)^2 , \tag{8}$$

with the values of m and b_j given by (7)

It can then be shown that

$$\hat{\sigma}'^2 - \hat{\sigma}^2 = \frac{1}{12n} \sum_{i,j} a_i^2 = \frac{1}{12} \sum_{i} a_i^2 . \tag{9}$$

Thus (5) becomes[1]

$$F = (n - 1) \frac{\sum_i a_i^2}{\sum_{i,j} (P_{ij} - m_{ij})^2} . \tag{10}$$

[1]The form of the denominator can be simplified for computational purposes.

The analysis of variance table is then as follows:

Source of variation	Degrees of freedom	Sums of squares	Estimate of[1] variation
Months	11	$n \sum_i a_i^2$	$11(\sigma^2 + n\,\sigma_m^2)$
Years	$n - 1$	$12 \sum_j b_j^2$	$(n-1)(\sigma^2 + 12\,\sigma_y^2)$
Residual	$11(n-1)$	$12n\,\hat{\sigma}^2$	$11(n-1)\sigma^2$

[1] $\hat{\sigma}^2$ is a biased estimate of the true error variance σ^2. The unbiased estimate is $12n\,\hat{\sigma}^2 / 11(n-1)$. σ_m^2 and σ_y^2 are the true variances between months and years.

3. The usefulness of the above analysis of variance depends on the extent to which the data fulfill the assumptions made. Some of the problems that merit consideration are:

(a). The simplicity of the analysis of variance depends upon the assumption of no missing prices. Simple methods have been devised for handling a two-way-classification analysis with missing observations, but these methods become more complicated for more than two classifications, especially if there are several missing observations. If many of the observations are missing, or if there is a different number of observations for each category under consideration, the analysis of variance may become quite complicated.

(b). A system of stochastic (difference) equations has been suggested as a proper model for economic phenomena [Haavelmo, 1943, p.1 ff.]. Our problem should be reconsidered from this viewpoint to find whether the likelihood function in (3) is a realistic description of observed phenomena. In particular, equation (1) might include other (economic) variables besides the year and the month; and the presence of lagged values in the equations would invalidate the assumption of independence of successive observations inherent in (3). If it is desired to test for such independence from the residuals, a way should be found to allow for the serial correlation introduced into the residuals by the fact that, as a result of

(7), they add up to zero in all directions.

(c). The assumption of normality usually can be severely violated without seriously affecting the general conclusions; however, the assumption of the same variance σ^2 for every observation cannot be violated to any great extent. A somewhat arbitrary rule often applied in biological experiments is to permit the assumption of constant variance only if no two means differ by more than 50 per cent.

(d). Finally, we should state that more thought should be devoted to the problem of designing social experiments, especially in setting up sampling methods. The analysis of variance set-up has been made much more useful for biological and agricultural research by connecting it with many designs of experiments other than the simple randomized block design, of which our month × year analysis is an example.

XIV. CONSISTENCY OF MAXIMUM-LIKELIHOOD ESTIMATES IN THE EXPLOSIVE CASE

BY HERMAN RUBIN

1. *Introduction*

It was shown by Mann and Wald [1943] that if a temporal stochastic process described by linear difference equations is damped without disturbances, it is stable (the expectations of the squares of all variables are uniformly bounded) with disturbances and the "maximum-likelihood" estimates of the parameters involved are consistent. However, a system which is stable or explosive (the expectations of the squares of some variables being unbounded) without disturbances is explosive with them. Here we prove consistency in a simple example of the explosive case.

Let us consider the equation

$$x_t = \rho x_{t-1} + u_t, \qquad |\rho| \geq 1 \qquad (t = 1, 2, \ldots), \tag{1}$$

where ρ is a real number, the u_t are real stochastic variables independently distributed with mean 0 and variance σ^2, and x_0 is a given real number. (The results derived here hold equally well if ρ, u_t, and x_0 are complex numbers, quaternions, or Cayley numbers.) The maximum-likelihood estimate $\hat{\rho}$ of the parameter ρ is defined as

$$\hat{\rho} = \frac{\sum x_t x_{t-1}}{\sum x_{t-1}^2}, \tag{2}$$

where

$$\sum = \sum_{t=1}^{T}. \tag{3}$$

[In the case of complex numbers, quaternions, or Cayley numbers, replace (2) by

$$\hat{\rho} = \frac{\sum x_t \bar{x}_{t-1}}{\sum x_{t-1} \bar{x}_{t-1}} , \tag{2'}$$

where $\bar{x}$ denotes the conjugate of x.] We shall show that

$$\operatorname*{plim}_{T \to \infty} \hat{\rho} = \rho . \tag{4}$$

2. *Transformation of the Problem*

We can write (2) in the form

$$\begin{aligned} \hat{\rho} &= \frac{\sum (\rho x_{t-1} + u_t) x_{t-1}}{\sum x_{t-1}^2} = \frac{\sum \rho x_{t-1}^2}{\sum x_{t-1}^2} + \frac{\sum u_t x_{t-1}}{\sum x_{t-1}^2} \\ &= \rho + \frac{\sum u_t x_{t-1}}{\sum x_{t-1}^2} . \end{aligned} \tag{5}$$

By the Cauchy-Schwarz inequality,

$$\left| \sum u_t x_{t-1} \right| \leq (\sum u_t^2)^{1/2} (\sum x_{t-1}^2)^{1/2} . \tag{6}$$

Hence

$$\left| \frac{\sum u_t x_{t-1}}{\sum x_{t-1}^2} \right| \leq \frac{(\sum u_t^2)^{1/2}}{(\sum x_{t-1}^2)^{1/2}} . \tag{7}$$

But the u_t^2 are real variables independently distributed with identical distribution functions and means σ^2. Hence [Cramér, Theorem 15]

$$\operatorname{plim} \frac{1}{T} \sum u_t^2 = \sigma^2 . \tag{8}$$

Let us consider the quantities x_t. By (1), we have

$$x_t = \rho^t x_0 + \rho^{t-1} u_1 + \cdots + \rho u_{t-1} + u_t . \tag{9}$$

Dividing (9) by ρ^t, we obtain

$$\rho^{-t} x_t = x_0 + \sum_{\tau=1}^{t} \rho^{-\tau} u_\tau . \tag{10}$$

3. *System Stable Without Disturbances*

In this section let us assume $|\rho| = 1$. Then

$$|\rho^{-\tau} u_\tau| = |u_\tau| . \tag{11}$$

We observe that for any $\delta > 0$ there exists a number $A(\delta)$ such that if $B > A(\delta)$ then

$$\int_{|u_\tau| > B} |\rho^{-\tau} u_\tau|^2 \, dF(u_\tau) = \int_{|u_\tau| > B} |u_\tau|^2 \, dF(u_\tau) < \delta , \tag{12}$$

since $\mathcal{E}(|u_\tau|^2) = \mathcal{E}(u_\tau^2) = \sigma^2 < \infty$. Therefore, if $t > A^2(\delta)/\varepsilon^2$, we have

$$\frac{1}{t} \sum_{\tau=1}^{t} \int_{|u_\tau| > \varepsilon\sqrt{t}} |\rho^{-\tau} u_\tau|^2 \, dF(u_\tau) < \frac{1}{t} \sum_{\tau=1}^{t} \delta = \delta . \tag{13}$$

Hence, for every $\varepsilon > 0$,

$$\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \int_{|u_\tau| > \varepsilon\sqrt{t}} |\rho^{-\tau} u_\tau|^2 \, dF(u_\tau) = 0 . \tag{14}$$

By the central limit theorem [Cramér, Theorem 21a],

$$x = \frac{1}{\sqrt{t}} \sum_{\tau=1}^{t} \rho^{-\tau} u_\tau$$

is asymptotically normally distributed with mean 0 and variance σ^2. Therefore, its cumulative distribution function $G_t(x)$ converges

uniformly[1] to the normal distribution $N(x, 0, \sigma^2)$. Let us now prove the following lemma:

LEMMA 3. *There exists a number* λ *such that for every* $\varepsilon > 0$ *there is a* t_0 *such that if* $t > t_0$ *then for every* b, $P_t(|x-b| < \varepsilon) < \lambda\varepsilon$.

Since $G_t(x)$ converges uniformly to $N(x, 0, \sigma^2)$, we see that for $\delta\varepsilon > 0$ there is a t_0 such that if $t > t_0$ then

$$|G_t(x) - N(x, 0, \sigma^2)| < \delta\varepsilon. \tag{15}$$

Then

$$G_t(b+\varepsilon) - G_t(b-\varepsilon) < N(b+\varepsilon, 0, \sigma^2) - N(b-\varepsilon, 0, \sigma^2) + 2\delta\varepsilon . \tag{16}$$

But $N(b+\varepsilon, 0, \sigma^2) - N(b-\varepsilon, 0, \sigma^2) < 2\varepsilon / \sqrt{2\pi}\sigma$, since $1/\sqrt{2\pi}\sigma$ is the maximum of dN/dx. Therefore

$$P_t(|x-b| < \varepsilon) = G_t(b+\varepsilon) - G_t(b-\varepsilon) < \varepsilon\left(\frac{2}{\sqrt{2\pi}\,\sigma} + 2\delta\right) \tag{17}$$

for $t > t_0$. Take b in the foregoing lemma to be $-x_0/\sqrt{t}$. Then we obtain from (10)

$$P\left(\left|\frac{\rho^{-t} x_t}{\sqrt{t}}\right| < \varepsilon\right) < \lambda\varepsilon , \tag{18}$$

or, since $|\rho| = 1$,

$$P(|x_t| < \varepsilon\sqrt{t}) < \lambda\varepsilon . \tag{19}$$

From (1) we see that

$$\big||x_{t+1}| - |x_t|\big| = \big||x_{t+1}| - |\rho x_t|\big| \le |x_{t+1} - \rho x_t| = |u_{t+1}| . \tag{20}$$

In the Tchebycheff inequality,

$$P(|u| > \theta\sigma) < \frac{1}{\theta^2} , \tag{21}$$

---

[2] Let F_n be a sequence of distribution functions converging to a continuous distribution function F. Then F_n converges uniformly to F.

let us take

$$\theta = \frac{\varepsilon\sqrt{t}}{2\mu\sigma}. \tag{22}$$

Then

$$P(|u_\tau| > \frac{\varepsilon\sqrt{t}}{2\mu}) < \frac{4\mu^2\sigma^2}{\varepsilon^2 t}. \tag{23}$$

Therefore

$$P(\{|u_{t+1}| + |u_{t+2}| + \cdots + |u_{t+\mu}|\} > \frac{\varepsilon\sqrt{t}}{2}) < \frac{4\mu^3\sigma^2}{\varepsilon^2 t}, \tag{24}$$

since if $|u_{t+1}| + \cdots + |u_{t+\mu}| > \varepsilon\sqrt{t}/2$, at least one $|u_{t+i}|$ is greater than $\varepsilon\sqrt{t}/2\mu$, and

$$P(\text{at least one } u_{t+i} > \frac{\varepsilon\sqrt{t}}{2\mu}) \leq \sum_{i=1}^{\mu} P(|u_{t+i}| > \frac{\varepsilon\sqrt{t}}{2\mu}). \tag{25}$$

On the other hand it follows that if $k \leq \mu$,

$$|x_{t+k}| \geq |x_t| - \sum_{\tau=1}^{k} |u_{t+\tau}| \geq |x_t| - \sum_{\tau=1}^{\mu} |u_{t+\tau}|. \tag{26}$$

Therefore

$$P(\min\{|x_t|, |x_{t+1}|, \ldots, |x_{t+\mu}|\} < \frac{\varepsilon\sqrt{t}}{2}) < \lambda\varepsilon + \frac{4\mu^3\sigma^2}{\varepsilon^2 t}. \tag{27}$$

Hence we see that

$$P(\sum_{\tau=t}^{t+\mu} x_\tau^2 < \frac{\varepsilon(\mu+1)t}{4}) < \lambda\varepsilon + \frac{4\mu^3\sigma^2}{\varepsilon^2 t}. \tag{28}$$

But

$$\sum_{\tau=0}^{t+\mu} x_\tau^2 \geq \sum_{\tau=t}^{t+\mu} x_\tau^2. \tag{29}$$

Take $\mu = [4A\nu/\varepsilon^2]$, $\nu > 1$, where $[x]$ denotes the greatest integer less than or equal to x. Then

$$(30)\qquad P\Big(\sum_{\tau=0}^{t+[4A\nu/\varepsilon^2]} x_\tau < A\nu t\Big) < \lambda\varepsilon + \frac{256\,\sigma^2 A^3\,\nu^3}{\varepsilon^8\,t}\,.$$

And for any $t > ([4A\nu/\varepsilon^2] + 1)/(\nu - 1)$ we have

$$(31)\qquad t + \left[\frac{4A\nu}{\varepsilon^2}\right] < \nu t - 1\,.$$

Let $\nu t = T$. Since

$$(32)\qquad \sum_{\tau=0}^{t+[4A\nu/\varepsilon^2]} x_\tau^2 < \sum_{\tau=1}^{T} x_{\tau-1}^2\,,$$

it follows that

$$(33)\qquad P\Big(\sum x_{t-1}^2 < A\,T\Big) < \lambda\varepsilon + \frac{256\,\sigma^2 A^3\,\nu^4}{\varepsilon^8\,T}$$

for $T > \max\{t_0,\ ([4A\nu/\varepsilon^2] + 1)/(\nu - 1)\}$, with t_0 defined as in the preceding lemma. Therefore

$$(34)\qquad \operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum x_{t-1}^2 \geq A$$

for every $A > 0$. Hence

$$(35)\qquad \operatorname*{plim}_{T\to\infty} \frac{\sum u_t^2}{\sum x_{t-1}^2} = \operatorname*{plim}_{T\to\infty} \frac{\frac{1}{T}\sum u_t^2}{\frac{1}{T}\sum x_{t-1}^2} \leq \frac{\sigma^2}{A}\,.$$

Therefore

$$(36)\qquad \operatorname*{plim}_{T\to\infty} \frac{(\sum u_t^2)^{1/2}}{(\sum x_{t-1}^2)^{1/2}} = 0\,,$$

and, from (5),

$$\operatorname*{plim}_{T\to\infty} \hat{\rho} = \rho . \tag{37}$$

4. *System Explosive Without Disturbances*

Let us assume now that $|\rho| > 1$. Denote by $F*G$ the convolution

$$F*G(x) = \int_{-\infty}^{+\infty} F(x-t)\,dG(t) \tag{38}$$

of the distributions F and G. If the distribution function of u_τ is $F(x)$, then the distribution function of $\rho^{-\tau} u_\tau$ is $F(\rho^\tau x)$. Since the mean of $\rho^{-\tau} u_\tau$ is 0 and the variance of $\rho^{-\tau} u_\tau$ is $\rho^{-2\tau}\sigma^2$, it follows [Wintner, Theorem 7.1] that if

$$F(x)*F(\rho x)*\cdots*F(\rho^n x) = G_n(x) , \tag{39}$$

then $\lim_{n\to\infty} G_n(x)$ exists and is a distribution function $G(x)$.

Let us now prove the following lemma:

LEMMA 4. *$G(x)$ is continuous.*

We shall consider three cases, for although the proof for the third case applies also to the other two, it is the most complicated.

Case 1. $F(x)$ is continuous. Then [Cramér, p. 37, equation (40)] it follows that G is continuous.

For any distribution function $H(x)$ denote by the functional $J\{H(x)\}$ the sum of all jumps of $H(x)$. We observe that $J\{H(\lambda x)\} = J\{H(x)\}$, and that $J\{H(x)\} = 0$ if and only if $H(x)$ is continuous. If $J\{H(x)\} < 1$, we shall say $H(x)$ is partly continuous, and if $J\{H(x)\} = 1$, we shall say that $H(x)$ is discrete.

Case 2. $H(x)$ is partly continuous. Let $J\{H(x)\} = \delta < 1$. From (37) we have

$$G_n(x) = F(x) * G_{n-1}(\rho x) . \tag{40}$$

But

$$J(H*K) = J(H)\,J(K). \tag{41}$$

Therefore

$$J\{G_n(x)\} = \delta^n. \tag{42}$$

We may derive from (37)

$$G(x) = G_n(x) * G(\rho^n x)\,. \tag{43}$$

Using (39) and (40) we obtain

$$J\{G(x)\} = \delta^n\, J\{G(x)\} \leq \delta^n\,. \tag{44}$$

Therefore $J\{G(x)\} = 0$, and $G(x)$ is continuous.

Case 3. $H(x)$ is discrete. Denote by $L\{H(x)\}$ the maximum jump of $H(x)$. We see that $L\{H(\lambda x)\} = L\{H(x)\}$. We have $L\{F(x)\} = \varphi < 1$. For if $L\{F(x)\} = 1$, then

$$\begin{aligned} P(y \leq x) &= 0 \qquad x < x^0\,, \\ P(y \leq x) &= 1 \qquad x \geq x^0\,, \end{aligned} \tag{45}$$

and the variance of u_τ would be 0, which contradicts its being $\sigma^2 > 0$. Using the fact that [Wintner, Theorem 7.6] $J\{G(x)\} > 0$ if and only if

$$\prod_{k=1}^{\infty} L\{F(\rho^k x)\} > 0\,,$$

we see from

$$\prod_{k=1}^{\infty} L\{F(\rho^k x)\} = \prod_{k=1}^{\infty} \varphi = 0 \tag{46}$$

that $G(x)$ is continuous.

Since the $G_n(x)$ approach $G(x)$, it follows that for any $\varepsilon > 0$, there are a δ and a t_0 such that if $n > t_0$ then

$$P_n(|x + x_0| < \delta) < \varepsilon\,. \tag{47}$$

We observe that $\rho^{-t}x_t - x_0$ has the distribution G_t. Therefore, for $t > t_0$, we have

$$P(|\rho^{-t} x_t| < \delta) < \varepsilon . \tag{48}$$

We further observe that for any $\delta > 0$, and for any K, there is a t_1 such that, if $t > t_1$, then

$$\frac{|\rho|^t \delta}{\sigma\sqrt{t}} > K. \tag{49}$$

From (46), we have for $t > t_0$,

$$P(|x_t| < |\rho|^t \delta) < \varepsilon. \tag{50}$$

Therefore, for $t > \max(t_0, t_1)$, we see that

$$P(|x_t| > K\sigma\sqrt{t}) > 1 - \varepsilon . \tag{51}$$

Then

$$\operatorname*{plim}_{T\to\infty} \frac{\sum u_t^2}{\sum x_{t-1}^2} \leq \operatorname*{plim}_{T\to\infty} \frac{\frac{1}{T}\sum u_t^2}{\frac{1}{T}x_{T-1}^2} \leq \lim_{T\to\infty} \frac{\sigma^2}{K^2\sigma^2\frac{T-1}{T}} = \frac{1}{K^2} . \tag{52}$$

Therefore

$$\operatorname*{plim}_{T\to\infty} \frac{(\sum u_t^2)^{1/2}}{(\sum x_{t-1}^2)^{1/2}} = \left(\operatorname*{plim}_{T\to\infty} \frac{\sum u_t^2}{\sum x_{t-1}^2}\right)^{1/2} = 0 \tag{53}$$

and

$$\operatorname*{plim}_{T\to\infty} \hat{\rho} = \rho . \tag{54}$$

XV. LEAST-SQUARES BIAS IN TIME SERIES[1]

BY LEONID HURWICZ

1.1. In this paper it is shown that there exist cases where the least-squares and the maximum-likelihood estimates of the regression and structural coefficients are biased[2] for any finite-sized sample[3] drawn from a population defined by a noncircular stochastic difference-equation system. This bias is evaluated for certain special cases. It is found that for very small samples the bias may amount to as much as 25 per cent of the true value of the parameter[4], while for medium sized samples (say of 20 observations) the bias is still almost 10 per cent (the numerical value of the expectation of the estimate always being below the true parameter value[5]). The relative bias seems to tend to zero, although rather slowly, as the damping of the system becomes weaker.

1.2. The initial objective of this paper was to prove that the least-squares method yields biased estimates of regression coeffi-

---

[1]Part of the work on this paper was done in 1945-46 during the author's tenure of the Guggenheim Memorial Fellowship. Some of the problems considered arose in connection with the author's research at the Institute of Meteorology at the University of Chicago in 1944.

[2]I.e., the mathematical expectations of the estimates are not identically equal to the true parameter values. *Bias* is defined as the difference between the expectation of the estimate and the true parameter value. *Relative bias* is the ratio of this difference to the true parameter. The quantity N_T appearing in the formulae below is the relative bias plus 1 in samples of T observations; hence N_T equals the ratio of the expectation of the estimate to the true parameter, so that when N tends to 1 the relative bias tends to zero.

[3]Containing more than two observations.

[4]Highest relative bias, found in samples of four observations, is $26\,^2/_3$ per cent.

[5]I.e., the bias, where known, is always negative for positive parameter values and vice versa.

cients in autoregressive[1] noncircular time series. It was generally realized that the usual proofs of the Markoff Theorem [David and Neyman] were not valid for this case[2], but the author does not know of any proof of the actual existence of a bias[3].

It was possible to obtain explicit formulae for the bias in very small samples (three or four observations). This was done for two types of assumptions with regard to the initial value: (1) *fixed initial-value* case – where the initial observation may be regarded as a fixed variate (the initial value chosen being zero) see section *3.1.2* below; (2) *stochastic initial-value* case – where the initial value is a stochastic variable whose distribution is the marginal distribution of the later observations, see section *2.4.2* below. The author regards the latter assumption as being more realistic, but since the *exact* equivalence of least-squares and maximum-likelihood criteria of estimation of the regression coefficients applies only to the former (asymptotically to both), it is the existence of bias in the fixed initial-value case that constitutes a proof of *bias of the maximum-likelihood* criterion as well. Moreover, the case treated is one where the regression coefficient is identically equal to the corresponding structural coefficient. Hence, the *bias* of least-squares and maximum-likelihood methods exists in *structural* as well as *predictive* estimation.

Once the *existence* of the bias had been shown it was of interest to investigate its *magnitude* for samples of various sizes. The first three terms of a series expansion for the bias in a sample of arbitrary size were obtained; see equations (4.6) to (4.10) below.

Because of the equivalence of the structural and regression coefficients in the cases treated, the results described may be regarded as a first (and very modest) step in the small-sample theory of the maximum-likelihood estimates of the structural coefficients in (noncircular) stochastic difference-equation systems.

The results thus far obtained indicate the importance of the

---

[1] I.e., generated by a system of stochastic difference equations, some or all of which contain *lagged* values of *endogenous* variables; autoregressive time series are to be distinguished from those composed *additively* of a given function of time (often a polynomial or a Fourier series) and a stochastic ("error") term.

[2] They *are* valid in the *lagless* case, as shown in [VI].

[3] In this note we are only concerned with "stable" systems, i.e., those satisfying Assumption IV_2 in [Mann and Wald, p. 192]; for the case here treated this assumption is equivalent to postulating $|\alpha| < 1$ in (2.1) below.

bias. There is urgent need for more intensive study of the small-sample properties of estimates in autoregressive time series.

2.1. Let a sample be given consisting of T observations on a stochastic variable X_t. It will be assumed that the joint cumulative distribution function of $(X_1, \ldots, X_T)$, to be written as $F(X_1, \ldots, X_T)$, has the following two properties:

$$\mathcal{E}(X_t \mid X_{t-1}) = \alpha X_{t-1}, \qquad t = 2, 3, \ldots, T, \tag{2.1}$$

$$\sigma^2(X_t \mid X_{t-1}) = \text{const.} \tag{2.2}$$

Thus the regression of any observation on its predecessor is linear[1] and homoscedastic.

The least-squares estimate $\tilde{\alpha}_T$ of α is in this case given by

$$\tilde{\alpha}_T = \frac{\sum_{t=2}^{T} X_t X_{t-1}}{\sum_{t=2}^{T} X_{t-1}^2}. \tag{2.3}$$

2.2. An estimate a_T of α will be said to be (conditionally) *unbiased with regard to a family* $\mathfrak{F}$ *of cumulative distribution functions* F *if all* F *in* $\mathfrak{F}$ *satisfy* (2.1) *and* (2.2) *and if*

$$\mathcal{E}(a_T) \equiv \alpha \quad \text{for all } \alpha, \quad \text{for all } F \in \mathfrak{F}, \quad \text{and for all } T. \tag{2.4}$$

An estimate conditionally unbiased with regard to *all* cumulative distribution functions is said to be *absolutely unbiased*.[2]

Clearly, if a family $\mathfrak{F}_0$ can be found with regard to which a_T is not unbiased, a_T cannot be absolutely unbiased.

2.3. In this note we show that there exists a family[3] $\mathfrak{F}_0$, in

---

[1] The condition (2.1) is stronger than linearity, but for the purposes of this paper no loss of generality is involved.

[2] These definitions are special cases of those given in [VI].

[3] This family, defined by (2.1) and (2.2) with the additional assumption of normality, is divided into two "branches" depending on whether the

fact one of considerable practical importance, with regard to which the least-squares estimate $\tilde{a}_T$ of α is not unbiased. This shows that $\tilde{a}_T$ is not absolutely unbiased. It is not known whether there exists a family, say $\mathfrak{F}_1$, such that $\tilde{a}_T$ is unbiased with regard to $\mathfrak{F}_1$.

However, an unbiased estimate of α does exist. Thus in the sample $(X_0, X_1, \ldots, X_T)$ where X_0 is fixed and different from zero, the ratio X_1/X_0 is an unbiased, though in general very inefficient, estimate of α. For, assuming that

$$X_t = \alpha X_{t-1} + u_t, \qquad t = 1, 2, \ldots, T, \tag{2.5}$$

where the u's have zero means, we have

$$\mathcal{E}\left(\frac{X_1}{X_0}\right) = \mathcal{E}\left(\frac{\alpha X_0 + u_1}{X_0}\right) = \alpha + \frac{1}{X_0}\,\mathcal{E}(u_1) = \alpha. \tag{2.6}$$

It may be remarked that for $T = 2$, the ratio X_1/X_0 is a least-squares estimate. In fact, $T = 2$ is the only known case among finite-sized samples where the least-squares estimate $\tilde{a}_T$ is unbiased.

On the other hand, in the sample $(X_1, \ldots, X_T)$, where X_1 is stochastic and the likelihood function is given by (2.9), the expectation $\mathcal{E}(X_2/X_1)$ exists only in the Cauchy principal-value sense. In that sense, however, X_2/X_1 is an unbiased estimate of α, since it has the Cauchy distribution with a mode at zero. In fact, with the u's independent of each other, the mean

$$\frac{1}{T-1}\sum_{t=2}^{T}\frac{X_t}{X_{t-1}}$$

would also be unbiased in the Cauchy sense, although no more efficient than X_2/X_1. One might conjecture that the median of the ratios X_t/X_{t-1}, $t = 2, \ldots, T$, would be a more efficient estimate of α and perhaps an unbiased one.

2.4.1. In the following sections of the paper the proof of existence of the bias will be given and its magnitude evaluated for first-order stochastic difference equations with the initial value

initial value is assumed stochastic or fixed. The bias exists in both cases. The respective likelihood functions are given by (2.9) and (3.15).

stochastic and fixed.

For samples of three and four observations the size of relative bias[1] is given for all $|\alpha| < 1$ in Table 1 and Figure 1; these are based on the formulae (3.12), (3.19), and (3.37). The derivations and comments are to be found in section *3* of this paper. The case of fixed initial value (chosen as zero) is worked out for a sample of three observations only.

For larger samples, as well as the small ones, the limiting value of bias[1] as $|\alpha| \to 0$ is given in Table 2 and Figure 2; these are based on (4.4). The latter formula is valid for both the stochastic and fixed (zero) initial-value case.

Equations (4.7) to (4.10) give a more general result, viz., the first three terms of the Maclaurin expansion of the relative bias. The derivations are given in section *4*. Sections *4.2* and *4.3* contain some conjectures with regard to the nature of the approximation provided by the expansion and with regard to the nature of relative bias in stochastic difference equations in general, as well as some suggestions for further research.

2.4.2. Define a stochastic process by

$$X_t = \alpha X_{t-1} + u_t, \qquad t = 2, 3, \ldots, T, \tag{2.7}$$

where the u's are independently normally distributed with zero means and unit variances, and where $|\alpha| < 1$. Hence, if the process is stationary[2],

$$\mathcal{E}(X_t) = 0, \qquad \mathcal{E}(X_t^2) = \frac{1}{1-\alpha^2}, \qquad t = 1, 2, \ldots, T. \tag{2.8}$$

Given a sample of size T, its likelihood function will be

$$(1-\alpha^2)^{1/2}(2\pi)^{-T/2}\exp\{-\frac{1}{2}[(1-\alpha^2)X_1^2 + \sum_{t=2}^{T}(X_t - \alpha X_{t-1})^2]\}, \tag{2.9}$$

which is equivalent to equation (9) in [Koopmans, 1942]. It can be

---

[1]The formulae, tables, and figures give the values of $N_T \equiv \mathcal{E}(\tilde{a}_T)/\alpha$, i.e., the relative bias plus 1.

[2]This specifies the stochastic initial-value case. For fixed initial-value case see below, section *3.1.2*.

seen that

$$\mathcal{E}(X_t \mid X_{t-1}) = \alpha X_{t-1}, \qquad t = 2, 3, \ldots, T, \tag{2.10}$$

so that (2.1) is satisfied. (2.2) is also satisfied with a unit variance.

3.1.1. To show the existence of bias it will suffice to find a T for which (2.4) does not hold. The following proof demonstrates that (2.4) does not hold for $T = 3$.

The proof consists in finding the expectation of $\tilde{a}_T$ for $T = 3$, where

$$\tilde{a}_3 = \frac{X_1 X_2 + X_2 X_3}{X_1^2 + X_2^2}. \tag{3.1}$$

Following the procedure used in [Williams] and in [Dixon], we write

$$\tilde{a}_3 = \frac{A_3}{B_3}, \qquad A_3 = X_1 X_2 + X_2 X_3, \qquad B_3 = X_1^2 + X_2^2, \tag{3.2}$$

then find the characteristic function $\varphi_3(t_1, t_2)$ of A_3 and B_3, and, finally, obtain the expectation of $\tilde{a}_3$ from

$$\mathcal{E}(\tilde{a}_3) = \int_{-\infty}^{0} \left[\frac{\partial \varphi_3(t_1, t_2)}{\partial t_1} \right]_{t_1 = 0} dt_2 . \tag{3.3}$$

We have

$$(1 - \alpha^2)\, \varphi_3^{-2}(t_1, t_2) = \begin{vmatrix} y & b & 0 \\ b & z & b \\ 0 & b & 1 \end{vmatrix} = \begin{vmatrix} y & f & 0 \\ 1 & z & f \\ 0 & 1 & 1 \end{vmatrix}, \tag{3.4}$$

where

$$\begin{aligned} y &= -2t_2 + 1, \\ f &= b^2 = (t_1 + \alpha)^2, \\ z &= y + \alpha^2 = -2t_2 + 1 + \alpha^2. \end{aligned} \tag{3.5}$$

Denoting the determinant in (3.4) by $C^{(3)}$, we have

$$\mathcal{E}(\tilde{a}_3) = -\frac{\sqrt{1-\alpha^2}}{2} \int_{-\infty}^{0} \frac{1}{C_0^{(3)}\sqrt{C_0^{(3)}}} \left.\frac{\partial C^{(3)}}{\partial t_1}\right|_{t_1=0} dt_2 , \tag{3.6}$$

where

$$C_0^{(3)} \equiv \left. C^{(3)} \right|_{t_1=0} = 4\left(t_2^2 - t_2 + \frac{1-\alpha^2}{4}\right) \tag{3.7}$$

and

$$\left.\frac{\partial C^{(3)}}{\partial t_1}\right|_{t_1=0} = 4\alpha(t_2 - 1). \tag{3.8}$$

The integral in (3.6) may now be evaluated [Pierce, formula 200] and we obtain

$$\mathcal{E}(\tilde{a}_3) = \frac{\alpha}{2}\left(1 + \frac{1-\sqrt{1-\alpha^2}}{\alpha^2}\right). \tag{3.9}$$

It can be seen that

$$\begin{aligned} \mathcal{E}(\tilde{a}_3) &\to \frac{3}{4}\alpha \quad \text{for } |\alpha| \to 0, \\ \mathcal{E}(\tilde{a}_3) &\to \alpha \quad \text{for } |\alpha| \to 1. \end{aligned} \tag{3.10}$$

Thus $\tilde{a}_3$ is a biased estimate of α.

Writing $\beta = \alpha^2$ and

$$N_T \equiv N_T(\beta) = \frac{\mathcal{E}(\tilde{a}_T)}{\alpha} , \tag{3.11}$$

we may state the above results as

$$N_3(\beta) = \frac{1}{2}\left(1 + \frac{1-\sqrt{1-\beta}}{\beta}\right) \tag{3.12}$$

with $N_3(\beta)$ varying from[1] $N_3(0) = 0.75$ to $N_3(1) = 1$. The values of

[1] $N_T(0) \equiv \lim_{\beta \to 0} N_T(\beta)$; $N_T(1) \equiv \lim_{\beta \to 1} N_T(\beta)$.

$N_3(\beta)$ are shown below in Table 1 and plotted in Figure 1.

The convergence of N_3 to 1 as $|\alpha|$ tends to 1 is very slow. The relative bias is still 12½ per cent (i.e., $N_3 = 0.875$) for $|\alpha| = 0.94$, so that it takes 94 per cent of the range of $|\alpha|$ to remove one-half of the relative bias!

3.1.2.[1] By a similar procedure we can evaluate the bias for the case where the initial value is a fixed variate, here chosen as zero. Let

$$X_0 = \text{fixed}, \tag{3.13}$$

$$X_t = \alpha X_{t-1} + u_t, \qquad t = 1, 2, \ldots, T, \tag{3.14}$$

where α and the u's have the same properties as before. T now denotes the number of *stochastic* observations. The likelihood function becomes

$$(2\pi)^{-T/2} \exp\{-\frac{1}{2}\sum_{t=1}^{T}(X_t - \alpha X_{t-1})^2\}. \tag{3.15}$$

Then the *least-squares*[2] estimate $\tilde{\alpha}_T^*$ of α, which in this case [unlike that specified by (2.9)] is also a *maximum-likelihood* estimate, is given by

$$\tilde{\alpha}_T^* = \frac{\sum_{t=1}^{T} X_t X_{t-1}}{\sum_{t=1}^{T} X_{t-1}^2}. \tag{3.16}$$

Now consider the special case where

$$X_0 = 0. \tag{3.17}$$

Here we find that $\tilde{\alpha}_T^*$ as defined in (3.16) equals $\tilde{\alpha}_T$ as defined in (2.3).

---

[1] T. W. Anderson has made helpful suggestions in connection with the problem treated in this section.

[2] The asterisk in $\tilde{\alpha}_T^*$ indicates that $\tilde{\alpha}_T^*$ is the least-squares estimate of α for the sample $(X_0, X_1, \ldots, X_T)$ while $\tilde{\alpha}_T$ is the least-squares estimate for the sample $(X_1, \ldots, X_T)$.

Denote by $\mathcal{E}^*(\tilde{a}_T)$ the expectation of $\tilde{a}_T$ evaluated on the basis of (3.15) with $X_0 = 0$; also, write

$$N_T^* \equiv N_T^*(\beta) \equiv \frac{\mathcal{E}^*(\tilde{a}_T)}{\alpha} . \tag{3.18}$$

Then we find that

$$N_3^*(\beta) = \frac{3 + \beta}{4 + \beta} . \tag{3.19}$$

Thus $N_3^*(0) = 0.75 = N_3(0)$;[1] but while $N_3(1) = 1$, we have here $N_3^*(1) = 4/5$: there is a bias even at $|\alpha| = 1$!

The case where $X_0 \neq 0$ has not been treated, but one might conjecture that, for given T and β, $N_T^*(\beta) \to 1$ as X_0 becomes numerically large.

3.2.1. When $T = 4, 5$ the integrals to be evaluated are of the elliptic type. For $T \geq 6$ they are hyperelliptic. Only for the case of $T = 4$ has the bias been evaluated in closed form.

To perform the integration in the elliptic case it is necessary to factor the T-rowed determinant $C_0^{(T)}$ where[2] [cf. (3.4) above]

$$C^{(T)} = (1 - \beta)\, \varphi^{-2}(t_1, t_2) = \begin{vmatrix} y & f & 0 & \dots & 0 & 0 & 0 \\ 1 & z & f & \dots & 0 & 0 & 0 \\ 0 & 1 & z & \dots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 & z & f \\ 0 & 0 & 0 & \dots & 0 & 1 & 1 \end{vmatrix}_{(T)} \tag{3.20}$$

in a sample of T observations and

$$C_0^{(T)} = C^{(T)}\Big|_{t_1=0} . \tag{3.21}$$

We may write

---

[1] This is an example of a more general phenomenon: $N_T^*(0) = N_T(0)$ for all T, cf. (4.4).

[2] For the definitions of y, z, f, see (3.5).

$$(3.22)\qquad C_0^{(T)} = \begin{vmatrix} z-\beta & \beta & 0 & \dots & 0 & 0 & 0 \\ 1 & z & \beta & \dots & 0 & 0 & 0 \\ 0 & 1 & z & \dots & 0 & 0 & 0 \\ . & . & . & \dots & . & . & . \\ 0 & 0 & 0 & \dots & 1 & z & \beta \\ 0 & 0 & 0 & \dots & 0 & 1 & 1 \end{vmatrix}_{(T)} .$$

For $\mu = 2, 3, \dots$ we have

$$(3.23)\qquad \begin{aligned} C_0^{(2\mu+1)} &= R_\mu^2 - \beta R_{\mu-1}^2 = (R_\mu - \alpha R_{\mu-1})(R_\mu + \alpha R_{\mu-1}), \\ C_0^{(2\mu)} &= R_{\mu-1}(z R_{\mu-1} - 2\beta R_{\mu-2}), \end{aligned}$$

where

$$(3.24)\qquad \begin{aligned} R_\mu &= \begin{vmatrix} z & \beta & 0 & \dots & 0 & 0 & 0 \\ 1 & z & \beta & \dots & 0 & 0 & 0 \\ 0 & 1 & z & \dots & 0 & 0 & 0 \\ . & . & . & \dots & . & . & . \\ 0 & 0 & 0 & \dots & 1 & z & \beta \\ 0 & 0 & 0 & \dots & 0 & 1 & 1 \end{vmatrix}_{(\mu+1)} \\ &= \begin{vmatrix} z-\beta & \beta & 0 & \dots & 0 & 0 & 0 \\ 1 & z & \beta & \dots & 0 & 0 & 0 \\ 0 & 1 & z & \dots & 0 & 0 & 0 \\ . & . & . & \dots & . & . & . \\ 0 & 0 & 0 & \dots & 1 & z & \beta \\ 0 & 0 & 0 & \dots & 0 & 1 & z \end{vmatrix}_{(\mu)} \end{aligned}$$

with

$$R_0 = 1, \qquad R_1 = \begin{vmatrix} z & \beta \\ 1 & 1 \end{vmatrix}. \tag{3.25}$$

Equations (3.23) make it possible to find by radicals the roots of the determinant $C_0^{(T)}$ for $T \leq 9$. For example,

$$\begin{aligned} C_0^{(4)} &= (z-\beta)[z(z-\beta) - 2\beta], \\ C_0^{(5)} &= \{[z(z-\beta) - \beta - \alpha(z-\beta)]\}\{[z(z-\beta) - \beta] + \alpha(z-\beta)\}, \\ C_0^{(6)} &= [z(z-\beta) - \beta]\{z[z(z-\beta) - \beta] - 2\beta(z-\beta)\}, \end{aligned} \tag{3.26}$$

etc.

3.2.2. Now in order to obtain $N_4(\beta)$ we follow a procedure similar to that for $T = 3$. We have

$$\mathcal{E}(\tilde{a}_4) = -\frac{1}{2}\sqrt{1-\beta} \int_{-\infty}^{0} [C_0^{(4)}]^{-3/2} \left.\frac{\partial C^{(4)}}{\partial t_1}\right|_{t_1=0} dt_2, \tag{3.27}$$

where[1]

$$C^{(4)} = \begin{vmatrix} y & f & 0 & 0 \\ 1 & z & f & 0 \\ 0 & 1 & z & f \\ 0 & 0 & 1 & 1 \end{vmatrix}, \tag{3.28}$$

so that

$$C_0^{(4)} = y(y^2 + \beta y - 2\beta), \tag{3.29}$$

$$\left.\frac{\partial C^{(4)}}{\partial t_1}\right|_0 = -2\alpha[(y^2 + \beta y - 2\beta) + (2y + \beta)]. \tag{3.30}$$

Hence, substituting these values into (3.27) and dividing both

[1]The symbols used here are defined in (3.5).

sides by α we have

$$N_4(\beta) = \sqrt{1-\beta} \int_{-\infty}^{0} \frac{(y^2+\beta y-2\beta)+(2y+\beta)}{[y(y^2+\beta y-2\beta)]^{3/2}}\, dt_2 . \tag{3.31}$$

Splitting into partial fractions and making the substitutions

$$y = -2t_2 + 1, \tag{3.32}$$

$$w^2 = \frac{y-y_1}{y}, \qquad y_1 = \frac{1}{2}(\sqrt{\beta^2+8\beta}-\beta), \tag{3.33}$$

and, finally,

$$u = \mathrm{sn}^{-1} w, \tag{3.34}$$

we obtain

$$\begin{aligned}\frac{4\sqrt{\beta^2+8\beta}}{\sqrt{1-\beta}} N_4(\beta) &= \frac{1}{y_1^2}\int \mathrm{cs}^2 u\, du + \frac{1}{2y_1}\int \mathrm{cn}^2 u\, du \\ &\quad + \frac{1}{y_3(y_1-y_3)}\int \mathrm{cd}^2 u\, du ,\end{aligned} \tag{3.35}$$

$$y_1,\ y_3 = \frac{1}{2}(\pm\sqrt{\beta^2+8\beta}-\beta), \tag{3.36}$$

where Glaisher's notation is used ($\mathrm{cs}\,u \equiv \mathrm{cn}\,u/\mathrm{sn}\,u$, $\mathrm{cd}\,u \equiv \mathrm{cn}\,u/\mathrm{dn}\,u$); the upper limit of integration is $\mathrm{sn}^{-1}(1)$ and the lower limit $\mathrm{sn}^{-1}[(1-y_1)^{1/2}]$.

With the help of formulae given by [Whittaker and Watson, 22.72, Ex. 3], we perform the integration, thus obtaining, after simplification,

$$N_4(\beta) = \frac{1}{2\beta}\left[(1+\beta) - \frac{\sqrt{1-\beta}}{4\sqrt{\beta^2+8\beta}}\,\Delta F\right], \tag{3.37}$$

where

$$\Delta F = F_{90^\circ}(\theta) - F_{\varphi}(\theta) . \tag{3.38}$$

$F_{\varphi}(\theta)$ is the incomplete elliptic integral of the first kind with modular angle θ and amplitude φ; $F_{90^\circ}(\theta)$ is the corresponding complete integral. The angles θ and φ are given by

$$\sin \varphi = \sqrt{1 - y_1} , \qquad \sin \theta = k , \quad 2k^2 = 1 + \frac{\beta}{\sqrt{\beta^2 + 8\beta}} . \tag{3.39}$$

3.2.3. It is easily seen that

$$N_4(1) \equiv \lim_{\beta \to 1} N_4(\beta) = 1 . \tag{3.40}$$

From this and the behavior of $N_3(1)$, one might conjecture that $N_T(1) = 1$ for all T.

It will be shown later in (4.4) that $N_4(0) \equiv \lim_{\beta \to 0} N_4(\beta) = 11/15 < N_3(0)$. As can be seen in Table 1 and Figure 1, $N_4(\beta)$ is a monotonic function of β. Whether this is the property of $N_T(\beta)$ for all T is not known.

3.2.4. The numerical values[1] computed from (3.37) are given in Table 1 and plotted in Figure 1. It is of interest to note that $N_3(\beta) > N_4(\beta)$ for all β. It will be seen later [from (4.4)] that $N_T(0)$ has a minimum for $T = 4$ (if T is an integer).[2] The convergence of $N_4(\beta)$ to 1 as $\beta \to 1$ is very slow, as was also the case for $N_3(\beta)$. To reduce the relative bias to one-half of its maximal value (so that $N_4 = 0.866$) we must have $|\alpha| = 0.95$. Whether the situation is quite as serious for larger samples is not known.

4.1. By expanding the integrand of (3.27), with T replacing

[1] Obtained with the aid of Miss Estelle Mass.

[2] For $T \geq 2$ and real, but not necessarily an integer, the minimum is at $T = 2 + \sqrt{3}$; $N_{2+\sqrt{3}}(0) = \sqrt{3} - 1 = 0.7321$.

TABLE 1.

Ratio of Estimate Expectation to True Parameter Value for Samples of 3 and 4 Observations

α	$\beta \equiv \alpha^2$	$N_3 \equiv \mathcal{E}(\tilde{a}_3)/\alpha$	$N_3^* \equiv \mathcal{E}^*(\tilde{a}_3)/\alpha$	$N_4 \equiv \mathcal{E}(\tilde{a}_4)/\alpha$
0	0	0.7500	0.7500	0.7333
0.1	0.01	0.7506	0.7506	0.7340
0.2	0.04	0.7526	0.7525	0.7359
0.3	0.09	0.7559	0.7555	0.7388
0.4	0.16	0.7609	0.7596	0.7434
0.5	0.25	0.7679	0.7647	0.7501
0.6	0.36	0.7778	0.7706	0.7595
0.7	0.49	0.7918	0.7773	0.7730
0.8	0.64	0.8125	0.7845	0.7938
0.9	0.81	0.8482	0.7921	0.8307
0.95	0.9025	0.8810	0.7960	0.8656
0.99	0.9801	0.9382	0.7992	0.9289
1.00	1.0000	1.0000	0.8000	1.0000

N and N^* refer to stochastic and fixed (zero) initial-value cases, respectively. The subscript indicates the number of (stochastic) observations in the sample.

TABLE 2.

The Limit (for Small Parameter Values) of the Ratio of Estimate Expectation to True Parameter Value: Stochastic Initial-Value Case

Sample Size T	$N_T(0) \equiv \lim_{\alpha \to 0} \mathcal{E}(\tilde{a}_T)/\alpha$	Sample Size T	$N_T(0) \equiv \lim_{\lvert\alpha\rvert \to 0} \mathcal{E}(\tilde{a}_T)/\alpha$
22	1.0000	13	0.8690
3	0.7500	14	0.8769
4	0.7333	15	0.8839
5	0.7500	16	0.8902
6	0.7714	17	0.8958
7	0.7917	18	0.9009
8	0.8095	19	0.9056
9	0.8250	20	0.9098
10	0.8384	50	0.9616
11	0.8500	100	0.9804
12	0.8601	500	0.9960

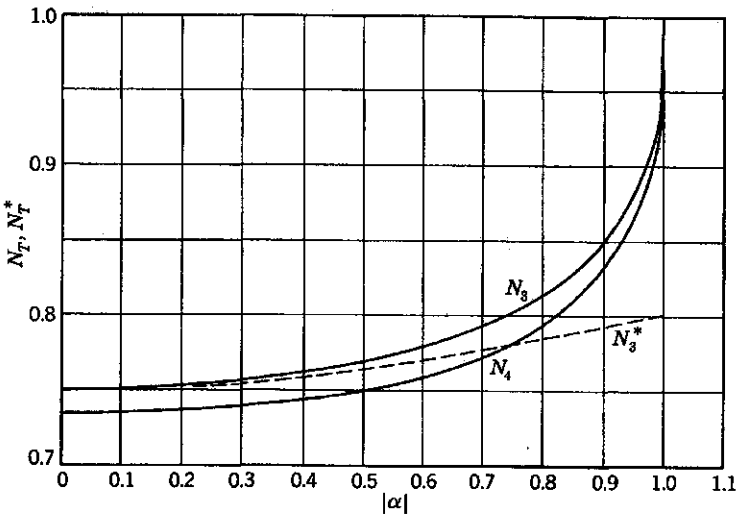


FIGURE 1. *Ratio of Estimate Expectation to True Parameter Value for Samples of 3 and 4 Observations. (For explanation of symbols see Table 1.)*

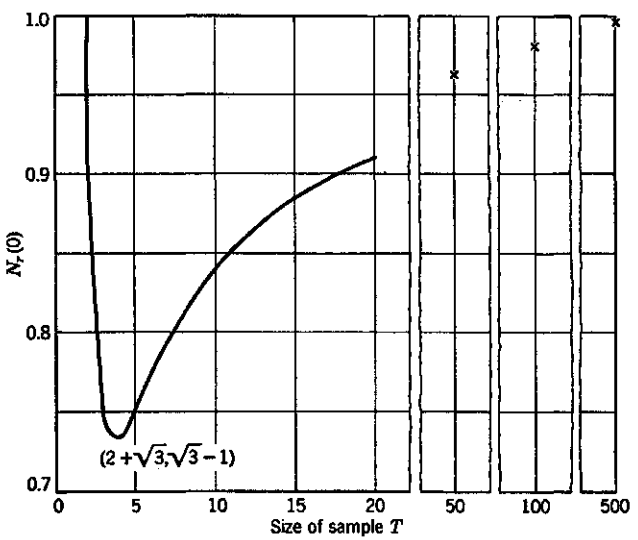


FIGURE 2. *The Limit (for Small Parameter Values) of the Ratio of Estimate Expectation to True Parameter Value: Stochastic Initial-Value Case. (For explanation of symbols see Table 2.)*

the affix 4, in a Maclaurin series and integrating termwise, it is possible to obtain an expansion[1] for $N_T(\beta)$ in powers of β.

It will now be shown how the first term of the expansion is obtained.[2] We first find

$$\lim_{\beta \to 0} C_0^{(T)} = y^{T-1}, \tag{4.1}$$

$$\lim_{\beta \to 0} \frac{1}{\alpha} \frac{\partial C^{(T)}}{\partial t_1}\bigg|_{t_1=0} = -2y^{T-3}[y + (T-2)]_0 . \tag{4.2}$$

Hence

$$N_T(0) \equiv \lim_{\beta \to 0} \frac{\mathcal{E}(\tilde{a}_T)}{\alpha} = \int_{-\infty}^{0} \frac{y^{T-3}[y + (T-2)]}{(y^{T-1})^{3/2}} dt_2 , \tag{4.3}$$

and, upon evaluation, this yields[3]

$$N_T(0) = \frac{T^2 - 2T + 3}{(T-1)(T+1)} . \tag{4.4}$$

Hence the bias exists for all finite-sized samples except $T = 2$. The values of $N_T(0)$ for some T are given in Table 2 and plotted in Figure 2.

It will be noted that the relative bias is 9 per cent for a sample of 20 observations and 2 per cent for a sample of 100 observations.

The second and third terms of the expansion of $N_T(\beta)$ in powers of β are obtained by similar methods, although the procedure becomes quite laborious. Writing

---

[1] As before, $N_T(\beta) \equiv \mathcal{E}(\tilde{a}_T)/\alpha$ and $\tilde{a}_T = \sum_{t=2}^{T} X_t X_{t-1} / \sum_{t=2}^{T} X_{t-1}^2$ is the least-squares estimate of α.

[2] Valuable suggestions in connection with this problem were made by Professor John von Neumann, Institute for Advanced Study.

[3] The same formula holds for $N_T^*(0)$, i.e., for the fixed initial-value case with $X_0 = 0$; see section 3.1.2 above.

$$C_{ij}^{(T)} = \left[\frac{\partial^j}{\partial\beta^j}\left(\frac{\partial^i C^{(T)}}{\partial f^i}\right)_{f=\beta}\right]_{\beta=0}, \tag{4.5}$$

we have

$$\begin{aligned}
C_{00}^{(T)} &= y^{T-1}, \\
C_{01}^{(T)} &= y^{T-3}[(T-3)y - (T-2)], \\
C_{02}^{(T)} &= (T-3)(T-4)\,y^{T-5}(y^2 - 2y + 1), \\
C_{10}^{(T)} &= -y^{T-3}[y + (T-2)], \\
C_{11}^{(T)} &= -(T-3)\,y^{T-5}[y^2 + (T-5)y - (T-4)], \\
C_{12}^{(T)} &= -(T-4)\,y^{T-7}[(T-3)y^3 + (T-4)(T-7)y^2 \\
&\qquad - (T-5)(2T-11)y + (T-5)(T-6)],
\end{aligned} \tag{4.6}$$

where the expression for $C_{12}^{(T)}$ is valid only for $T > 3$. With the help of (4.6) we find that, for $T > 3$, the first three coefficients of the expansion

$$N_T(\beta) = N_T(0) + N_T'(0)\,\beta + \frac{1}{2}N_T''(0)\,\beta^2 + \cdots \tag{4.7}$$

are given by

$$N_T(0) = \frac{T^2 - 2T + 3}{(T-1)(T+1)}, \tag{4.8}$$

$$N_T'(0) = \frac{2(T^2 - 8T + 21)}{(T-1)(T+1)(T+3)(T+5)}, \tag{4.9}$$

$$N_T''(0) = \frac{4(T^4 + 24T^3 + 98T^2 - 264T - 99)}{(T-1)(T+1)(T+3)(T+5)(T+7)(T+9)}. \tag{4.10}$$

Equations (4.8) and (4.9) are valid for $T \geq 2$ and $T \geq 3$, respectively, equation (4.10) for $T > 3$ only.

4.2. It is quite evident that the first three terms of the expansion for $N_T(\beta)$ do not give a very good approximation for values of β near 1. For example, $N_4(1) = 1$, while the first three terms of the expansion give only 0.823. For $|\alpha| = 0.5$, however, while[1] $N_4(0.25) = 0.7501$, the first three terms of the expansion give 0.7497 which is correct to the third decimal digit. If the same phenomenon should exist for $T > 4$, which remains to be proved, the first three terms of the expansion could not only be used successfully for values of $|\alpha|$ below 0.5 but also for values of $|\alpha|$ somewhat (though not too much) above 0.5.

Moreover, it would seem that these three expansion terms will always give values lower than the true $N_T(\beta)$. It would be desirable to examine the correctness of this conjecture and also to obtain an upper bound for $N_T(\beta)$ in terms of the expansion. It seems possible that such a bound is given by the expression

$$1 - (1 - \beta)N_T'(0) - (1 - \beta^2)\frac{1}{2}N_T''(0) \ . \tag{4.11}$$

However, even if proved correct, this would only be useful for values of β in the neighborhood of 1. For example, for $T = 4$ and $|\alpha| = 0.9$, the upper bound would be given by 0.978, while the correct value is only 0.831. It may be observed that the first three terms of the expansion give a much better approximation, namely 0.801, despite the high value of $|\alpha|$.

4.3. The following more general propositions concerning first-order stochastic linear difference equations still await proof:

$$N_T(\beta) \leq 1 \quad \text{for} \quad \beta < 1 \quad \text{and all} \quad T, \tag{4.12}$$

$$\frac{\partial N_T(\beta)}{\partial \beta} > 0 \quad \text{for} \quad \beta < 1 \quad \text{and all} \quad T, \tag{4.13}$$

and hence

$$\lim_{\beta \to 1} N_T(\beta) = 1 \quad \text{for all} \quad T. \tag{4.14}$$

[1] It will be remembered that the argument of $N_T(\ )$ is β, not $|\alpha|$.

It would also be desirable to investigate analogous problems first for higher-order difference equations and then for equation systems.[1] It would be of interest to see whether the following two propositions are generally true: (1) that a system becomes more strongly damped when estimate expectations are substituted for the structural parameters, and (2) that the relative bias is lower in systems with stronger damping and tends to zero as the characteristic roots of the system approach 1 in absolute value.

These investigations should be carried out for both the stochastic and fixed-variate initial-value case, without the fixed initial value being necessarily zero.

Finally, higher moments of the sampling distributions of the estimates should be investigated. It is probable that these problems will have to be studied with the help of more powerful tools, especially that of the approximate sampling distributions,[2] provided the upper bound of the approximation error can be determined.

---

[1] At this point one would have to distinguish between the bias of the (maximum-likelihood) estimates of the structural coefficients and the bias of the (least-squares equivalent to maximum-likelihood for the fixed-variate initial-value case) estimates of the regression coefficients. The distributions of the conditional variances and of the disturbance covariance matrix also deserve attention in small samples.

[2] See [Koopmans, 1942], [Dixon], and [Leipnik].

XVI. MODELS INVOLVING A CONTINUOUS TIME VARIABLE

BY TJALLING C. KOOPMANS

1. In the opinion of this author, the simultaneous-equations method to which the first part of this volume is devoted constitutes an important advance over single-equation methods in the measurement of economic relations. Reasons for this opinion have been stated elsewhere [Koopmans, 1945]. In this note it is my intention to point out certain shortcomings of the new methods in their present stage.

Limitations to the usefulness of the new methods arise from the combination of the following two aspects in the specification of the distribution of the variables:

(a) the treatment of time as a discrete variable,

(b) the assumption that disturbances at different points in time are independent.

These aspects are also found in the single-equation methods that have been generally used, and some of the limitations to be mentioned below therefore apply equally to simultaneous-equations and single-equation methods. However, the development of the simultaneous-equations method has brought to light further disadvantages of the specifications (a) and (b) above, and thus has made their revision even more desirable than before.

2. It was found in another article[1] that in the simultaneous-equations method the procedure for estimating the coefficient α_{11} of the endogenous variable x_1 in the first structural equation, say, is different according to whether that variable is "predetermined" or is one of the "jointly dependent variables." A given variable may be in one category or the other depending solely on the timing of that variable in the equation concerned. If an endogenous variable x_1 appears in the first equation only with a lag of one unit behind the most recent timing of that variable in the

[1] [II–*3.1.3*]. For the definition of "endogenous," "predetermined," "dependent," variables, see also [XVII].

equation system as a whole, then in the first equation it will be classified as predetermined. If the time lag is reduced to zero, the variable is in general to be treated as one of the jointly dependent variables. This situation can be reduced *ad absurdum* by making the time unit of measurement smaller and smaller and at the same time reducing the time lag to zero. Then suddenly at the moment the time lag reaches zero the status of the variable x_1 in the first equation is changed, and with it the "unbiased" estimate of the coefficient in question is changed. The solution of this paradox is, of course, that such a procedure is illegitimate. The assumption of independent disturbances in successive observations can be maintained only if the size of the time unit to which these observations refer is not made too small. Therefore, the independence assumption makes the distinction between predetermined and dependent variables appear as absolute instead of a matter of degree, which it would be in a more refined model.

The main source of disturbances is the erratic element in economic behavior. Some causes of erratic behavior (not already represented by measurable variables), like the weather affecting the amount and direction of consumers' expenditure, may be so variable as to reverse themselves in a few days. Other causes like fads and fashions affecting consumption, confidence or lack of confidence affecting investment, may lead to deviations in the same direction for a whole year or even longer. Therefore, as the time unit of observation is reduced in size, a situation in which serial correlation of the disturbances in a given equation can no longer be neglected is bound to arise at some stage.

Methods based on the independence assumption thus involve a lower bound on the permissible size of the time unit of observation. This precludes adequate treatment of a number of important statistical problems in the measurement of economic relations. The time lags occurring in economic behavior are not always integral multiples of one time unit of a size compatible with the independence assumption. They are almost always distributed lags, with the lower limit to the range of lags sometimes practically equal to zero. There is therefore a need for methods of estimating the parameters that characterize lag distributions.

Another problem that cannot be studied adequately under the independence assumption is that of the most economic time unit of observation. Whether it is best to use annual, quarterly, or monthly, data depends on a comparison of the cost of collection of such data (if not already available), the cost of calculating

the necessary estimates, and the information gained (in the sense of smaller sampling errors of estimated parameters) by a given reduction in the size of the unit period of observation. The latter gain in information is likely to be reduced by serial correlation in the disturbances. For time units below a certain size, this gain in information can therefore not be analyzed theoretically on the basis of the independence assumption.

3. An adequate model for the study of the foregoing problem is obtained by considering the disturbances (and therefore the economic variables) as generated by a stochastic process with a continuous time variable. Let us for simplicity assume that this process is normal and stationary, i.e., the values of the variables $u_g(t)$ at any set of time points $t_1, t_2, \ldots, t_S$ have a joint normal distribution depending only on the differences $t_2 - t_1, \ldots, t_S - t_{S-1}$. In that case the process is entirely characterized [Doob, Theorem 4.3] by the elements

$$\mathcal{E}\, u_g(t)\, u_h(t) = \sigma_{gh}(0) \tag{1}$$

of the covariance matrix $\Sigma(0)$, and by a matrix Ξ, determining the lagged covariance matrix

$$\mathcal{E}\, u_g(t)\, u_h(t+\tau) = \sigma_{gh}(\tau) \tag{2}$$

through the formula

$$\Sigma(\tau) = \Sigma(0)\, e^{-\tau \Xi}, \qquad \tau > 0. \tag{3}$$

If economic variables $x_n(t)$ are considered as determined by equations in which quantities $u_g(t)$ of this nature are the only random elements, it is necessary to define further the way in which these variables are observed. In practice, the method of observation is again a discrete procedure. One method of observation is to make readings

$$x_n(t_0),\ x_n(t_1),\ \ldots,\ x_n(t_S), \qquad t_s - t_{s-1} = h, \qquad s = 1, \ldots, S, \tag{4}$$

at equidistant points in time. Price variables are sometimes observed in this way. Quantities of goods and flows of money are

usually observed through averages

$$\bar{x}_n(t_s) = \int_{t_s - h}^{t_s} x_n(\tau)\, d\tau \tag{5}$$

over a period of observation of length h. For any specified method of obtaining a finite number of observations for each variable in the system, the set of all observations on all variables becomes subject to a joint probability distribution derivable from (3) (or from whatever other process of a continuous time variable t is specified).

Although the mathematical difficulties involved may be considerable, a model of this kind would provide a means of studying the estimation of lag distributions and the choice of the most economic time unit of observation.

4. Perhaps the most important advantage of a continuous treatment of time has not yet been mentioned. It is explained in another article [II-2.5.6] that, in the discrete case, as soon as the independence assumption for successive disturbances is dropped, the problem of identification of the structural equations is greatly complicated. For the removal of that assumption may open up a new group of transformations of the equations (involving shifts along the time axis) that preserve the probability distribution of the variables.

The introduction of a continuous time variable is perhaps the best way to study fully all aspects of the identification problem of relations between economic time series. Consider for instance a system of one linear equation containing only one variable x. In the discrete formulation this equation would be of the type

$$\mathfrak{L}x(t) \equiv x(t) + \alpha_1 x(t-1) + \cdots + \alpha_\tau x(t-\tau) = u(t). \tag{6}$$

Under the independence assumption for $u(t)$,

$$\mathcal{E}u(t)u(t+\theta) = 0 \quad \text{if} \quad \theta \neq 0, \tag{7}$$

there is no identification problem. For any linear combination of the type

$$\lambda_0 \mathfrak{L}x(t) + \lambda_1 \mathfrak{L}x(t-1) + \cdots + \lambda_K \mathfrak{L}x(t-K) = \bar{v}(t) = \sum_{k=0}^{K} \lambda_k u(t-k) \tag{8}$$

introduces serial correlation into $v(t)$ unless all but one of the quantities λ_k vanish.

In the continuous formulation the equation is of the type

$$\mathfrak{L}^*x(t) \equiv x(t) - \int_{-\infty}^{t} \varphi(t-\tau)\, x(\tau)\, d\tau = u(t) \tag{9}$$

where the disturbance process, if normal and stationary, is described by

$$\mathcal{E}u(t)\, u(t+\theta) = f(|\theta|). \tag{10}$$

Now there are infinitely many transformations in the space of the functions φ and f which preserve the form of (9). The simplest of these is obtained by substituting

$$\int_{-\infty}^{t} \varphi(t-\tau)\, x(\tau)\, d\tau + u(t) \tag{11}$$

for $x(t)$ under the integral sign in (9). Another possibility is first to write

$$\varphi(t-\tau) = \varphi_1(t-\tau) + \varphi_2(t-\tau), \tag{12}$$

and to substitute (11) only in one of the two integrals so obtained, etc. The only invariants of all these transformations are the autocovariance function

$$g(\theta) = \mathcal{E}x(t)\, x(t+\theta) \tag{13}$$

of the variable $x(t)$ and all its functions and functionals. These are therefore the only identifiable characteristics of the process (9), and only these are subject to estimation.

The identification difficulties just described are absent from the process

$$x(t) = \int_{-\infty}^{t} \varphi(t-\tau)\, x(\tau)\, d\tau + \alpha y(t) + u(t) \tag{14}$$

containing an observable exogenous variable $y(t)$. They reappear if $y(t)$ occurs in the form

$$(15)\quad x(t) = \int_{-\infty}^{t} \varphi(t-\tau)\,x(\tau)\,d\tau + \int_{-\infty}^{t} \psi(t-\tau)\,y(\tau)\,d\tau + u(t)$$

with unknown lag-distribution functions φ and ψ.

The process (9) becomes a system of equations if $x(t)$ and $u(t)$ are interpreted as column vectors, $\varphi(t-\tau)$ as a matrix. In this case, the identification of individual equations is aided if economic considerations require or permit the specification that all diagonal elements of $\varphi(t-\tau)$ shall vanish for all values of $t-\tau$. In two dimensions this leads to the system

$$(16)\quad \begin{aligned} x_1(t) &= \int_{-\infty}^{t} \varphi_{12}(t-\tau)\,x_2(\tau)\,d\tau + u_1(t), \\ x_2(t) &= \int_{-\infty}^{t} \varphi_{21}(t-\tau)\,x_1(\tau)\,d\tau + u_2(t). \end{aligned}$$

The equations (16) are completely identified, i.e., there is no transformation other than the identity, in the space of the functions φ_{12}, φ_{21} and f_1, f_2 [defined as in (10)] which preserves the form of (16).

PART THREE

SPECIFICATION OF HYPOTHESES

XVII. WHEN IS AN EQUATION SYSTEM COMPLETE FOR STATISTICAL PURPOSES?

BY TJALLING C. KOOPMANS

1. *Endogenous and Exogenous Variables*

In static or dynamic economic theory, the criteria employed in determining whether or not a system of equations is complete are derived from the purpose for which such systems are constructed: the explanation of economic phenomena. In each case a distinction is drawn between the *endogenous* variables that the economist sets out to explain and the *exogenous* variables that he takes as given. The number of equations required for the explanation of the values or, in the dynamic case, of the movements of the endogenous variables, then equals the number of such variables.

2. *Exogenous Variables in Economic Theory*

In determining which variables are set aside as exogenous, two main principles are implicitly or explicitly applied in economic literature. They might be described as the departmental principle and the causal principle. The departmental principle treats as exogenous those variables which are wholly or partly outside the scope of economics, like weather and climate, earthquakes, popula-

tion, technological change, political events. The causal principle, which does not always lead to the same result, regards as exogenous those variables which influence the remaining (endogenous) variables but are not influenced thereby.

The causal principle is often used also if it applies only approximately, that is, if the influence of the endogenous variables on those treated as exogenous is presumed to be small. For instance, in explaining the level of employment in a country which has only a small share in world trade, the shifts in the schedules of foreign demand for its exports and of foreign supply of its imports are sometimes treated as exogenous in first approximation. Another example is found in the formation of quantity and price of a consumers' good that attracts only a small fraction of consumers' expenditure. In such cases, consumers' income is often taken as an exogenous variable, operating at the demand side, although of course consumers' income itself depends on the demand for *all* commodities. In order to distinguish between cases where the causal principle of classification of variables is strictly or only approximately applicable, we shall in what follows speak of the strict causal principle and the approximate causal principle.

There is no sharp line of demarcation between the application of the approximate causal principle and what deserves mention as a third principle or consideration: the purpose of exposition. At a certain stage of the analysis, variables are often treated as exogenous to facilitate understanding of the model studied, reserving for later elaboration their inclusion among the endogenous variables.

3. *Statistical Definition of Exogenous Variables*

One of the main purposes of the present volume is to study the statistical implications of the fact that economic data are governed by a system of simultaneous equations. It is therefore necessary to review the foregoing principles, and such other considerations as may present themselves, in relation to the purpose of statistical estimation of the equations of dynamic economics. The question which, if any, are exogenous variables must be raised afresh from the statistical point of view.

It will be clear that the departmental principle is of no value in this connection. Suppose that rainfall and temperature enter as determining factors in one or more equations of the system. Is it possible that a certain method of estimation, applied to the economic relations with disregard of purely physical relations between

these and other meteorological variables, for that very reason leads to inconsistent and asymptotically biased estimates[1] of the parameters of the economic relations? The answer to this question obviously does not depend on which particular department is in charge of studying such relations.

The causal definition of exogenous variables does permit an answer to this question, with specific reference to the maximum-likelihood method of estimation. The answer is based on a mathematical observation which we shall now formulate. Suppose that

$$(1) \qquad \varphi_n(\alpha_{n1}, \ldots, \alpha_{nQ_n}; x_1, \ldots, x_N) = u_n, \quad n = 1, \ldots, N,$$

represents the complete system of all structural equations (linear or otherwise) between *all* variables, economic or noneconomic, that enter directly or indirectly into the explanation of economic variables. Suppose, for the present, that all these variables enter without time lags, and let the random terms u_n have for each time t a joint continuous distribution

$$(2) \qquad f(u_1, \ldots, u_N)\,du_1 \cdots du_N,$$

these distributions being independent for successive values $t = 1, 2, \ldots$ of t. Suppose finally that both the equations (1) and the variables $x_1, \ldots, x_N$, can be separated into two sets, with numbering $n = 1, \ldots, G$ and $n = G + 1, \ldots, N$ for the first and second sets, respectively, such that the following three assumptions are satisfied.

ASSUMPTION 3.1. *The first set of "endogenous variables" does not occur in the second set of equations:*

$$(3) \begin{cases} (3a) \quad \varphi_n(\alpha_{n1}, \ldots, \alpha_{nQ_n}; x_1, \ldots, x_G; x_{G+1}, \ldots, x_N) = u_n, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n = 1, \ldots, G, \\ (3b) \quad \varphi_n(\alpha_{n1}, \ldots, \alpha_{nQ_n}; x_{G+1}, \ldots, x_N) = u_n, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n = G + 1, \ldots, N. \end{cases}$$

---

[1] An estimate of a parameter α, derived from a sample of size T, is called consistent if, for any $\varepsilon > 0$, $\lim_{T \to \infty} P(|a - \alpha| > \varepsilon) = 0$ if $P(E)$ denotes the

ASSUMPTION 3.2. *The distribution function in* (2) *can be factorized as follows:*

$$f(u_1, \ldots, u_N) = f_1(u_1, \ldots, u_G)\, f_2(u_{G+1}, \ldots, u_N). \tag{4}$$

ASSUMPTION 3.3. *In any point* $x_1, \ldots, x_N$ *permitted by the distribution function* (2), *the Jacobian* $\partial(u_1, \ldots, u_N)/\partial(x_1, \ldots, x_N)$ *of the "transformation"* (3) *differs from zero everywhere in the space of the parameters* α_{nq}, $q = 1, \ldots, Q_n$; $n = 1, \ldots, N$.

The economic meaning of these assumptions is that (Assumption 3.1) we isolate in (3b) the equations that connect only exogenous variables, that (Assumption 3.2) we assume that the random elements entering those equations – and hence the (exogenous) variables of the second set themselves – are distributed independently from the random disturbances in the equations (3a) explaining the endogenous variables, and that (Assumption 3.3) no situation can arise in which small changes in the disturbances $u_1, \ldots, u_N$ lead to very large changes in the variables $x_1, \ldots, x_N$.

Since the equation system (3) together with (4) completely specifies the joint distribution of the variables $x_1, \ldots, x_N$, maximum-likelihood estimation based on (3) and (4) may be expected to lead to consistent and asymptotically unbiased estimates of all identifiable[1] parameters.[2] Assuming this to be the case, we shall now prove that in consequence *the identifiable parameters of the equation* (3a) *can be estimated, by the method of maximum likelihood, i.e., consistently and without bias in large samples, from those equations only, treating the "exogenous" variables* $x_{G+1}, \ldots, x_N$ *as if they were fixed in repeated samples.* To prove this point, let us regard (3a) and (3b) as one system of equations (3) and write down the joint distribution function h of a complete set of observations for one value of t. This function is given by

---

probability of an event E. The estimate a is called asymptotically unbiased if $\lim_{T\to\infty} \mathcal{E}a = \alpha$, where $\mathcal{E}$ denotes the mathematical expectation.

[1] For the concept of identifiability, see [II–2], also [IV].

[2] This has been proved rigorously for *linear* systems; see [II–3.3].

$$h(\alpha_{11}, \ldots, \alpha_{NQ_N}; x_1, \ldots, x_N)\, dx_1 \cdots dx_N$$

(5)

$$= f(u_1, \ldots, u_N)\, \frac{\partial(u_1, \ldots, u_N)}{\partial(x_1, \ldots, x_N)}\, dx_1 \cdots dx_N ,$$

in which the quantities $u_1, \ldots, u_N$ are regarded as functions (1) of $x_1, \ldots, x_N$. Because of the assumed form of (3b), the Jacobian in (5) factorizes as follows:

(6)

$$\frac{\partial(u_1, \ldots, u_N)}{\partial(x_1, \ldots, x_N)} = \begin{vmatrix} \frac{\partial u_1}{\partial x_1} & \cdots & \frac{\partial u_G}{\partial x_1} & 0 & \cdots & 0 \\ \cdot & \cdots & \cdot & \cdot & \cdots & \cdot \\ \frac{\partial u_1}{\partial x_G} & \cdots & \frac{\partial u_G}{\partial x_G} & 0 & \cdots & 0 \\ \frac{\partial u_1}{\partial x_{G+1}} & \cdots & \frac{\partial u_G}{\partial x_{G+1}} & \frac{\partial u_{G+1}}{\partial x_{G+1}} & \cdots & \frac{\partial u_N}{\partial x_{G+1}} \\ \cdot & \cdots & \cdot & \cdot & \cdots & \cdot \\ \frac{\partial u_1}{\partial x_N} & \cdots & \frac{\partial u_G}{\partial x_N} & \frac{\partial u_{G+1}}{\partial x_N} & \cdots & \frac{\partial u_N}{\partial x_N} \end{vmatrix}$$

$$= \frac{\partial(u_1, \ldots, u_G)}{\partial(x_1, \ldots, x_G)} \cdot \frac{\partial(u_{G+1}, \ldots, u_N)}{\partial(x_{G+1}, \ldots, x_N)} .$$

Combining corresponding factors in (6) and (4), we obtain the following factorization for h:

$$(7) \qquad h(\alpha_{11}, \ldots, \alpha_{NQ_N}; x_1, \ldots, x_N)$$

$$= h_1(\alpha_{11}, \ldots, \alpha_{GQ_G}; x_1, \ldots, x_N)\, h_2(\alpha_{G+1,1}, \ldots, \alpha_{NQ_N}; x_{G+1}, \ldots, x_N),$$

in which the parameters α_{nq} occurring in each factor of (6) have been determined from (3).

What happens if the likelihood function

$$(8) \qquad F \equiv \prod_{t=1}^{T} h\{\alpha_{11}, \ldots, \alpha_{NQ_N}; x_1(t), \ldots, x_N(t)\}$$

following from (7) is maximized with respect to the parameters α_{nq}? To answer this question, it is necessary to remember that the factors f_1 and f_2 in (4) are nonnegative, and are positive somewhere within their range, in view of their nature as probability densities. Since the nonvanishing Jacobians in (6) enter in (5) and hence in (7) only with their absolute values, the factors h_1 and h_2 in (4) are likewise nonnegative and somewhere positive. Furthermore, the two factors $F_1 = \prod_t h_1$ and $F_2 = \prod_t h_2$ of (8) depend on entirely different sets of parameters, $\alpha_{11}, \ldots, \alpha_{GQ_G}$ and $\alpha_{G+1,1}, \ldots, \alpha_{NQ_N}$, respectively. It follows that, for the product $F = F_1 F_2$ to be as large as possible, it is necessary and sufficient that each of the factors F_1 and F_2 separately be as large as possible.

The problem of (asymptotically unbiased) estimation of the parameters by the method of maximum likelihood has thereby been split into two separate estimation problems for the two sets of parameters occurring in (3a) and (3b), respectively. In particular, it is possible to disregard entirely the estimation of the equations (3b) and to estimate only the parameters of (3a) by maximizing F_1. The latter problem is identical with the estimation problem met with if the exogenous variables $x_{G+1}, \ldots, x_N$ in the equations (3a) are regarded not as determined by (3b), but as given in advance and fixed in repeated samples.

It might be observed that although under the present assumptions the maximum-likelihood estimates of the parameters occurring

in (3a) are the same functions of the observations whether the exogenous variables are regarded as subject to a probability distribution or as fixed in repeated samples, a difference arises again between these two cases in the distribution of the maximum-likelihood estimates at least in small samples. However, a procedure is available whereby this difference can be removed. In the case where the exogenous variables are subject to a probability distribution, it is both permissible and useful to consider, for the construction of confidence intervals, only that subclass of all possible samples in which the exogenous variables have values equal to those observed.[1]

The foregoing proposition provides a justification for using the concept of exogenous variables defined according to the causal principle. In equation systems in which all variables enter simultaneously (the only case covered so far), a clear separation has now been obtained between the endogenous variables that must be explained by an equal number of relations (all of which are in principle relevant to the estimation of any one of them) and the exogenous variables that may be accepted without explanation. It is worth stressing again that for purposes of statistical estimation the concept of exogenous variables must be defined more strictly and narrowly than for some purposes of economic theory. Whereas the theorist may at a certain stage choose, for reasons of approximation or exposition, to disregard possible influences exerted by variables "inside" his system on variables regarded by him as "outside," the statistician must be convinced (on a priori grounds or as a result of statistical test) of the absence of such influence before he can declare the "outside" variable to be exogenous in the foregoing sense.

4. *Endogenous and Exogenous Variables in Systems Containing Time Lags*

We must now take into consideration the fact that the action of one variable on another is often subject to a time lag. Suppose therefore that the variables $x_1, \ldots, x_n$ occur in the equations (1) not only with timing t, but also with various time lags τ which are integral nonnegative multiples of the chosen unit of time:.

[1]For a discussion of this device, see [Hotelling, 1940].

$$\varphi_n\{\alpha_{n1}, \ldots, \alpha_{nQ_n};\ x_1(t), \ldots, x_N(t);\ x_1(t-1), \ldots, x_N(t-1);\ \ldots\}$$

(9)

$$= u_n(t), \qquad n = 1, \ldots, N;\quad t = 1, \ldots, T.$$

The joint distribution of all observations $x_n(t)$, $n = 1, \ldots, N$; $t = 1, \ldots, T$, again follows from that of the disturbances $u_n(t)$, $n = 1, \ldots, N$; $t = 1, \ldots, T$, provided we specify how any values $x_n(t-\tau)$ with $t-\tau \leq 0$, occurring in the equations (1) for $t \geq 1$, are distributed. We shall assume such values to be constant in repeated samples and equal to the values observed in the sample at hand.

The first question requiring an answer under the present assumption is whether maximum-likelihood estimation based on the joint distribution function of all observations $x_n(t)$, $n = 1, \ldots, N$; $t = 1, \ldots, T$, still leads to consistent and asymptotically unbiased estimates. This does not follow from the general theory of maximum-likelihood estimation because the observations $x_n(t)$ for successive values of t are no longer stochastically independent. However, Mann and Wald [1943] have proved that for *linear* systems of the form (9), maximum-likelihood estimates are consistent and asymptotically unbiased. Presuming that this result can be extended to nonlinear systems, we shall therefore continue the analysis on the basis of the maximum-likelihood method of estimation.

It is easily seen that the Jacobian

$$J_T \equiv \frac{\partial\{u_1(1), \ldots, u_N(1); \ldots; u_1(T), \ldots, u_N(T)\}}{\partial\{x_1(1), \ldots, x_N(1); \ldots; x_1(T), \ldots, x_N(T)\}} \tag{10}$$

of the transformation from all disturbances to all observations (that are not constants) factorizes as follows

$$J_T = \begin{vmatrix} \dfrac{\partial\{u_1(1), \ldots, u_N(1)\}}{\partial\{x_1(1), \ldots, x_N(1)\}} & \dfrac{\partial\{u_1(2), \ldots, u_N(2)\}}{\partial\{x_1(1), \ldots, x_N(1)\}} & \cdots & \cdot \\ 0 & \dfrac{\partial\{u_1(2), \ldots, u_N(2)\}}{\partial\{x_1(2), \ldots, x_N(2)\}} & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \dfrac{\partial\{u_1(T), \ldots, u_N(T)\}}{\partial\{x_1(T), \ldots, x_N(T)\}} \end{vmatrix}$$

(11)

$$= \prod_{t=1}^{T} \frac{\partial\{u_1(t), \ldots, u_N(t)\}}{\partial\{x_1(t), \ldots, x_N(t)\}} = \prod_{t=1}^{T} J(t) ,$$

say, all entries to the right of the main diagonal in J_T being immaterial. Consequently, the joint distribution function of the observations is again given by the expressions (8) and (5), the only difference being that the arguments $u_1, \ldots, u_N$ of f in (5) are now given by (9) instead of (1).

By going through the previous reasoning after this change in the structural equations, it will be seen that the previous result remains true: *If* (9) *has the form* (3), *where* u_n *now stands for* $u_n(t)$, *and each of the symbols* x_n *stands for a sequence of values* $x_n(t)$, $x_n(t-1)$, ..., *maximum-likelihood estimation of the parameters* $\alpha_{11}, \ldots, \alpha_{GQ_G}$ *can be performed by maximizing only the function*

$$(12) \qquad F_1 = \prod_{t=1}^{T} f_1\{u_1(t), \ldots, u_G(t)\} \left| \frac{\partial\{u_1(t), \ldots, u_G(t)\}}{\partial\{x_1(t), \ldots, x_G(t)\}} \right| ,$$

with $u_g(t)$ *given by* (3a).

Thus the distinction between endogenous and exogenous variables

is maintained equally well in systems containing time lags. Further analysis needs to be concerned only with the function (12) which represents the conditional probability density in the space of all values of the endogenous variables $x_1, \ldots, x_G$ within the period $t = 1, \ldots, T$, for given values of the exogenous variables $x_{G+1}, \ldots, x_N$ within that period, and of all variables $x_1, \ldots, x_N$ previous to that period $(t \leq 0)$. Distribution theory of the estimates of the parameters α_{gq}, $g = 1, \ldots, G$, can also be based entirely on the conditional distribution function (12). It is therefore unnecessary from now on to carry the "given" values of the exogenous variables $x_{G+1}, \ldots, x_N$ along in all formulae, and we shall rewrite (3a) as

$$\varphi_g\{\alpha_{g1}, \ldots, \alpha_{gQ_g};\ x_1(t), \ldots, x_G(t);\ x_1(t-1), \ldots, x_G(t-1); \ldots\} = u_g(t), \tag{13}$$

$$g = 1, \ldots, G; \quad t = 1, \ldots, T.$$

5. *The Nature of Exogenous Variables*

Before continuing the analysis on the basis of (12) and (13), the question is in order as to the nature of the variables so set aside as "exogenous." Which factors in man's physical and historical environment are not influenced by his economic activity? If the question is put in this way one can think of little else besides changes in weather, climate, geology, and geography that are brought about by natural causes. There remains a host of sociological, political, and psychological factors that are in continuous interaction with economic activity, and therefore cannot, on any grounds so far adduced, be accepted as they come without incorporating the explanation of their fluctuations in the system of equations.

6. *Predetermined Variables and Jointly Dependent Variables*

A further delimitation of the area that needs to be covered by an equation system, for purposes of unbiased estimation of parameters in large samples, can sometimes be obtained through closer study of the manner in which time lags occur in the structural equations (13).

The assumption that the variables $x_1, \ldots, x_G$ are endogenous implies that each variable occurs without time lag in at least one equation, because otherwise the equation system (13) would not describe the determination of each variable. The earlier values $x_i(t-1), x_i(t-2), \ldots, i = 1, \ldots, G$, of the endogenous variables may accordingly be called predetermined variables. At time t their values are already given by the equations (13) for earlier values of t. Any possibility of again splitting off from the system (13) a subset of equations for separate treatment depends, as before, on further factorizing properties of the Jacobian $J(t)$ referring to one single time point t as defined in the last line of (11). This leads immediately to the following extension of the observations made above. Suppose that both the equations (13) and the (endogenous) variables $x_i(t)$, $i = 1, \ldots, G$, can again be separated into two sets, (13a) and (13b), such that instead of the previous Assumptions 3.1, 3.2, 3.3 the following assumptions are satisfied:

ASSUMPTION 6.1. *The first set of variables* $x_i(t)$, $i = 1, \ldots, G_1$, *does not occur in the second set* $(g = G_1 + 1, \ldots, G)$ *of equations* (13) *except with a time lag* $\tau \geq 1$:

(13)

$$\varphi_g\{\alpha_{g1}, \ldots, \alpha_{gQ_g}; x_1(t), \ldots, x_G(t);$$

(13a) $$x_1(t-1), \ldots; x_G(t-1); \ldots\} = u_g(t),$$

$$g = 1, \ldots, G_1; \quad t = 1, \ldots, T,$$

$$\varphi_g\{\alpha_{g1}, \ldots, \alpha_{gQ_g}; x_{G_1+1}(t), \ldots, x_G(t);$$

(13b) $$x \quad x_1(t-1), \ldots, x_G(t-1); \ldots\} = u_g(t),$$

$$g = G_1+1, \ldots, G; \quad t = 1, \ldots, T,$$

ASSUMPTIONS 6.2 and 6.3. *The previous Assumptions* 3.2 *(factorizing of the distribution function of disturbances) and* 3.3 *(non-vanishing Jacobian), respectively, are again satisfied for the new subdivisions of variables and equations.*

Then it follows that *the parameters of the two sets of equations can again be estimated by the maximum-likelihood method in two separate steps*. In particular, the first set of equations can be estimated without any knowledge of the second set beyond that expressed by the Assumptions 6.1, 6.2, 6.3.

If the conditions of this proposition are satisfied, the endogenous variables of the second set $x_{G_1+1}(t), \ldots, x_G(t)$ will be called the predetermined variables even where they occur without time lags. For even their values $x_g(t)$ without lags are determined by *earlier* values of the variables of both sets and by disturbances independent of the *simultaneous* values of the variables of the first set. These characteristics permit the predetermined variables to be treated in the first set of equations *as if* they were exogenous variables for the purpose of the separate determination of maximum-likelihood estimates of the parameters α_{gq}, $q = 1, \ldots, Q_g$, $g = 1, \ldots, G$, of the first set of equations.

With regard to the distribution of these estimates the position is slightly more complicated than in the case where only the exogenous variables and the equations explaining these variables are split off. In the present case the small-sample distribution of the maximum-likelihood estimates a_{gq} of the parameters α_{gq}, $g = 1, \ldots, G_1$, of the first set of equations cannot be made independent of the parameters α_{gq}, $g = G_1 + 1, \ldots, G$, of the second set of equations by any such construction or restriction as was applied in the case of exogenous variables. The asymptotic theory of maximum-likelihood estimation for stochastic difference equations, reviewed in [II–3.3], implies, however, that at least for linear systems with normally distributed disturbances, the limiting distribution of the a_{gq}, $g = 1, \ldots, G_1$, for infinitely large samples is independent of the parameters α_{gq}, $g = G_1 + 1, \ldots, G$, and is such as to make the estimates involved both consistent and asymptotically unbiased.

The distinction of predetermined variables thus makes possible a great further reduction in the size and scope of the equation systems needed for statistical purposes in cases where large-sample approximations are adequate.[1] In order to be "complete" in this "large-sample" statistical sense, the equation system then only needs to cover with reasonable completeness the formation of those

[1] See, however, [XV] where L. Hurwicz discusses in a simple case the bias involved in the application of large-sample approximations to samples of moderate size.

economic (and possibly some noneconomic) variables that are connected through a process of *instantaneous* interactions. The variables between which these instantaneous interactions occur will be called the *jointly dependent variables*, or, briefly, the *dependent variables*. Any further determining variables affecting this process of interaction, which variables themselves depend only on earlier values of the endogenous variables and on chance variation independent of that process, need not be "explained" through additional equations, but can be treated as predetermined variables. In particular, slow changes in environment affecting consumers' preferences, technological changes, and other factors usually described as "trends" can probably be treated as predetermined variables in this sense with a reasonable degree of accuracy. Likewise, the interactions between economic variables and political developments is often subject to sizable time lags, and in such cases need not be taken into account for the purposes here considered.

7. *Summary of the Classification of Variables*

The foregoing distinctions are summarized in the following table[1] which also provides a comparison of the notation of the present article with that introduced in two different places in [II].

8. *Insufficiency of the "Approximate" Causal Principle*

Further comment is needed on some of the principles underlying this classification. Both the distinction between exogenous and endogenous variables and that between predetermined and dependent variables are based on a subdivision of the complete set of equations "explaining" the formation of all variables into two subsets of equations. In both cases it is necessary to stipulate (see Assumption 3.3) that the disturbances affecting equations thereby placed into different subsets should be independently distributed. The necessity of this assumption definitely limits the extent to which equation systems can be reduced in size without sacrificing "completeness" in the statistical sense. In particular, this assumption frequently prohibits the application of the "approximate" causal principle in cutting down the size of the system. This point is of great importance, for instance, to the statistical measurement of demand and supply curves for individual commodities. For this reason, and because of the need for further mathematical demonstration of the insufficiency of the "approximate" causal

[1] See p. 406.

TABLE 7.1.

CLASSIFICATION OF VARIABLES WITH NOTATION USED IN THIS ARTICLE

Notation introduced in II-1.3	Distinction independent of timing of variable	Distinctions dependent on timing of variable in the equations			Notation introduced in II-2.1.2
$y_i(t-\tau)$, $\tau \geq 0$, $i = 1, \ldots, G$	I. Endogenous variables (affected by variables under I and II), x_n, $n = 1, \ldots, G$	A. occurring without time lag, $x_i(t)$, $i = 1, \ldots, G$	a. in instantaneous interaction with each other, $x_i(t)$, $i = 1, \ldots, G_1$	Jointly dependent variables	$y(t)$
			b. affected only by variables II and IB and by random disturbances independent of variables IAa, $x_i(t)$, $i = G_1+1, \ldots, G$	Predetermined variables	$z(t)$
		B. occurring with a time lag, $x_i(t-\tau)$, $\tau \geq 1$, $i = 1, \ldots, G$		Predetermined variables	$z(t)$
$z_k(t-\tau)$, $\tau \geq 0$, $k = 1, \ldots, K$	II. Exogenous variables (not affected by the variables under I), x_n, $n = G+1, \ldots, N$			Predetermined variables	$z(t)$

principle, a separate discussion of this problem will be given elsewhere in due course.

9. *Time Lags as a Criterion of Classification*

The question must be raised whether the concept of a time lag is sharp enough to serve as the basis for a classification of variables that determines their treatment in the estimation procedure. After all, time lags may occur in all sizes, varying continuously down to zero.

It is possible to give a provisional answer to this objection which safeguards the statistical procedures developed in this volume as logically defensible and consistent. It has been assumed in those procedures that disturbances in successive observations (that is, in successive time points or time intervals) are independent. That assumption cannot continue to be valid if the unit of observation is made smaller and smaller. Therefore, the term "time lag" in the previous discussion means a time lag not smaller than the smallest time unit of observation for which the independence assumption is still tenable as a fair approximation to reality.

Nevertheless, such a discrete treatment of what is a continuous time variable should not be accepted as the final word of statistical theory applied to economic time series. In this writer's opinion, an extension of the "discrete" methods of this volume to permit continuously distributed lags, ranging down to zero wherever appropriate, is the next important refinement needed in the development of statistical methods adapted to the analysis of relations between economic time series. In [XVI] the need for and the nature of such a generalization are discussed further.

10. *Identification Problems in Systems Containing Exogenous or Other Predetermined Variables*

In the foregoing discussion we have considered two successive reductions of the size of the original equation system (3), where x_n now, as in section 4, stands for a sequence of values $x_n(t)$, $x_n(t-1)$, ... In the first reduction the equations (3b) connecting only exogenous variables are omitted, and the equations (3a) or (13) are retained. In the seconed reduction the subsystem (13b) "explaining" the predetermined variables is omitted, and the remain-

ing equations (13a) are retained. In order to provide a link with the discussion of identification problems in II–2, the following question should be raised: Could any equation of (13a), or any parameter of such an equation, which appears to be identifiable on the basis of a study of transformations in the space of the parameters occurring in (13a) only, be found nonidentifiable if possible transformations in the space of all parameters entering in the complete system (3), or in its first subsystem (13), were taken into consideration? The answer is that, under the various independence assumptions stated above, this cannot occur as long as the distribution of the disturbances $u_g(t)$ is nonsingular, in the sense that no exact relationship of the type

$$\psi\{u_{g_1}(t),\ u_{g_2}(t),\ \dots\} = 0 \tag{14}$$

is satisfied by the disturbances of either of the subsystems (3b) or (13b). For, if the distribution of the disturbances is nonsingular, any attempt to combine equations from (3b) or (13b) with those of (13a) to produce new equations of the form (13a) will violate the independence assumptions as between the disturbances in (3a) and (3b) or in (13a) and (13b), respectively. Under the present assumptions, therefore, nonsingularity of the distribution of disturbances is the only condition that needs to be met by the complete system (3) in order that questions of identifiability can be settled by study of the appropriate subsystem (3a) or (13a). In order to visualize the meaning of this condition, let us assume that a relation (14) holds for the disturbances of the equations (3b) "explaining" the exogenous variables. By inserting the left hand members of (3b) for $u_g(t)$ in (14), we find that then an exact relation

$$\chi\{x_{G+1}(t),\ \dots,\ x_M(t)\} = 0 \tag{15}$$

connects the exogenous variables. This case was excluded from the discussion of identification problems in II–2 by the wording of Definition 2.1.2. Whether the presence of a relation (15) does in fact destroy identifiability of otherwise identifiable parameters of (3a) depends on the functional form of (15) in connection with any a priori information about the functional form of the equations (3a). For instance, if (15) is linear, the identifiability of any linear equation in (3a), containing the same variables as (15) with-

out restrictions on their coefficients, is destroyed. The occurrence of such a case in applications should reveal itself, theoretically through infinite, practically through very high calculated, standard errors of the estimates of parameters whose identification is destroyed.

XVIII. SYSTEMS WITH NONADDITIVE DISTURBANCES[1]

BY LEONID HURWICZ

0.1. In this note the need is shown for generalization of the manner in which disturbances are introduced in stochastic equation systems. While at present in the systems treated each equation contains only *one* disturbance, in more realistic systems there might be equations each containing *several* disturbances. Moreover, some of these disturbances might enter in a nonadditive fashion as coefficients of the observed variables.

A model with nonadditive disturbances, while often appropriate on grounds of a priori (economic) knowledge, is much more difficult to handle mathematically. In fact the difficulties arise at an extremely early stage: it has not yet been possible, in general, to derive formulae on which the computation of estimates could be based. The maximum-likelihood estimates have been obtained for a lagless single-equation system, i.e., where all variables but one are fixed.[2] Even in this simplest case the equation for the unknown estimates is, in general, nonlinear.

But when a simultaneous system is considered, even the likelihood function cannot, in general, be obtained except in the form of an integral. Hence, unless more powerful analytical methods can be found, the only practical solution (if the maximum-likelihood principle is to be used) would seem to be to resort to numerical methods[3] in evaluating the likelihood function and locating its maximum.

The same would appear to hold for the method of moments, although it may be that this problem deserves a more careful investigation. One may hope that other criteria for estimation can be found, but if not, the numerical methods applied to the maximum-likelihood criterion of estimation would seem to be the best way out.

---

[1]Part of the work done on this paper was done in 1945-46 during the author's tenure of the Guggenheim Memorial Fellowship.

[2]See [XIX].

[3]Perhaps with the aid of the recently designed electronic computing machines.

The next stage would be that of investigating the sampling properties of the estimates used. This would also be a work of considerable complexity.

0.2. This note is divided into three sections. Section 1 is devoted to the specification of the nonadditive disturbance model. Section 2 provides some examples, mostly from the field of economics, of nonadditive disturbance models. In section 3 the likelihood function is derived for a special and relatively simple case of a two-equation model where one of the equations contains a nonadditive disturbance.

1.1. The simultaneous linear systems treated by methods of statistical analysis as a rule belong to the category[1]

$$\varphi_g(y', z') \equiv \varphi_g(x') = u_g, \qquad g = 1, 2, \ldots, G, \tag{1.1}$$

where $y = (y_1, \ldots, y_{K_y})$, $z = (z_1, \ldots, z_{K_z})$, $G = K_y$; y is stochastic and z a fixed variate. We observe $x \equiv (y, z)$ but not the disturbances u_g. The function (vector) $\Phi = (\varphi_1, \ldots, \varphi_G)$ and the covariance matrix Σ^u of the disturbances are partly specified by a priori knowledge (the identifying restrictions) and are partly to be estimated from the observations on x. In the linear case (1.1) becomes

$$\alpha_g x' = u_g, \qquad g = 1, 2, \ldots, G, \tag{1.2}$$

where α_g is the gth row of the structural coefficient matrix A.

Now (1.1) is but a special case of

$$\psi_g(x', u') = 0, \qquad g = 1, 2, \ldots, G = K_y, \tag{1.3}$$

where $u = (u_1, \ldots, u_{K_u})$. (1.1) can be obtained from (1.3) by assuming

$$K_u = K_y, \tag{1.4}$$

$$\frac{\partial \psi_g}{\partial u_p} = \delta_{gl}, \tag{1.5}$$

[1] y' is the transpose of the row vector y, etc.

where δ_{gl} is the Kronecker symbol.

If (1.5) were abandoned but (1.4) retained, it would still be possible to solve for the u's and rewrite (1.3) as

$$\psi_g^*(x') = u_g, \qquad g = 1, 2, \ldots, G, \tag{1.6}$$

although equations (1.6) would no longer be the original behavior equations. This type of situation, because of its nonlinear nature, would be of considerable mathematical complexity. The Jacobian of the transformation (from u to y) becomes a function of y, rather than a constant as in (1.2). Thus, while it is not difficult to obtain the likelihood function, it will, in general, be quite difficult to find the maximizing values of the unknown parameters.

But when K_u is permitted to exceed K_y $(=G)$ an additional difficulty arises since the distribution of y is then essentially a marginal one; an integration must be performed in order to obtain the likelihood function. The difficulty of this integration is due to the fact that the integrand contains the absolute value of a Jacobian and the latter is a function of the variables of integration (see below, section 3.2).

1.2. Let (1.3) hold with $K_u > K_y$ so that the equations (1.4) and (1.5) are not satisfied. Nevertheless, the disturbances might still enter (1.3) in a purely additive manner. For instance, we might have the system

$$\begin{aligned} y_1 + \alpha_{12} y_2 + u_{11} + \varepsilon u_{12} &= 0, \\ \alpha_{21} y_1 + y_2 + u_{21} &= 0. \end{aligned} \tag{1.7}$$

This, however, is a trivial case and it will henceforth be assumed that any number of additive disturbances is treated as one single disturbance; thus in the above case we would define $u_1^* \equiv u_{11} + \varepsilon u_{12}$ and say that $K_y = K_u = 2$.

1.3. The simplest general case of nonadditive disturbances is given by the equation system

$$(\alpha_g + u_g)x' \equiv (\alpha_g^y + u_g^y)y' + (\alpha_g^z + u_g^z)z' = 0, \tag{1.8}$$

$$g = 1, 2, \ldots, G = K_y,$$

where u_g is the gth row of the disturbance matrix $U \equiv [u_{gs}]$, $g = 1, 2, \ldots, G$, $1 \leq s \leq K_x$; it is assumed, without loss of generality, that $\mathcal{E}(U) = 0$; α_g is, as before, the gth row of the structural coefficient matrix A.

The linear additive disturbance case (1.2) can be obtained from (1.8) by making all the elements of U except one column in u^z corresponding to, say, $z_{K_z} \equiv 1$, vanish.

2.1. Why a nonadditive disturbance model should be the appropriate one cannot be seen without examining the manner of introducing the stochastic element into those models.

Suppose the *undisturbed* model is given by

$$\varphi_g(x'; \theta_g) = 0, \qquad g = 1, 2, \ldots, G. \tag{2.1}$$

To transform this into a *stochastic (disturbed)* model we must consider some component $\theta_g^{(1)}$ of θ_g as stochastic. Just which component of θ_g should become stochastic cannot be decided arbitrarily: it depends on the composition of the universe of objects from which the observation is assumed to be drawn. Thus suppose the theory, as embodied in (2.1), refers to an individual firm, while the observations (say in a cross-section study) describe a group of firms (the universe being some larger aggregate of existing and, possibly, nonexistent firms). Let $\theta_g^{(1)}$ be the component of θ_g which characterizes an individual firm. Then $\theta_g^{(1)}$ may depend 1) on x, or 2) on nonobservable fixed variates, or 3) it may be regarded as stochastic. 1) creates no new problem, 2) may imply loss of identification, 3) converts $\theta_g^{(1)}$ into a disturbance u_g.

Suppose the third possibility holds so that $\theta_g^{(1)}$ may be regarded as stochastic. Then, if $\theta_g^{(1)}$ has more than one component, we are dealing with the case of nonadditive disturbances except for the trivial case mentioned in section 1.2. Such situations are extremely frequent in economic models, although, unfortunately, the second of the above three possibilities is not uncommon.

Some economic models of this type are described in article [I], others in section 2.2 below.

2.2. An example of (1.3) can be found in the estimation of a firm's production function in a cross-section study, as discussed on p. 415. Other examples inserted here in small print may help

clarify the nature of the problem and show the great range of possibilities.

(A) *A game of chance (fixed variates).* Let X_{it}, $i = 1, 2, \ldots, k$, be k numbers arbitrarily chosen by the player at the time t and α_{it} the number given by throwing the ith set of dice. (The α's are independently distributed in time.) The player wins each time the amount

$$Z_t = \sum_i \alpha_{it} X_{it}. \tag{1}$$

We observe the X's and the Z's but not the α's, and we wish to estimate the distribution of the latter, say its first two moments.

(B) *A game of chance (time series; discrete).* Let α_{it}, $i = 0, 1$, be numbers given by throwing the ith set of dice at the time t. (The α's are again independently distributed in time.) X_0 is an arbitrary initial constant, and the player's winnings at time t, denoted by X_t, are determined by

$$X_t = \alpha_{it} X_{t-1} + \alpha_{0t}, \qquad t = 1, 2, \ldots. \tag{2}$$

Here again the problem would consist in estimating the unknown distribution of the α's (which are not observable) given the observations on X_t.

(C) *An economic system (time series; discrete).* Let there be a system of equations

$$\begin{aligned} \alpha_{11t} x_{1t} + \alpha_{12t} x_{2,t-1} + \alpha_{10t} &= u_{1t}, \\ \alpha_{21t} x_{1t} + \alpha_{22t} x_{2t} + \alpha_{20t} &= u_{2t}, \end{aligned} \tag{3}$$

which may be interpreted as supply and demand equations, respectively, the x's representing quantity and price, with a lag in the first equation as in the case of the "hog cycle." The α's have the subscript t since, for instance, they may vary seasonally or exhibit a trend. But it seems quite plausible to consider them as also having variations of a random nature. In that case the problem of estimating the distribution of the α's would again arise.

(D) *Continuous systems.* The above system may also be presented in continuous form. However, in order to simplify exposition, the continuous example given here will correspond to the one-variable discrete case, say to

$$x_t = \sum_{\tau=1}^{\tau_m} \alpha_{\tau t} x_{t-\tau} + \alpha_{0t} + u_t. \tag{4}$$

The corresponding (slightly simplified) continuous situation would be

$$x(t) = \int_0^\infty \varphi(\tau)\, x(t-\tau)\, d\tau \quad + \quad u(t), \tag{5}$$

where $u(t)$ is a continuous normal process with a specified autocorrelation pattern.

The difficulties of mathematical manipulation are not to be underestimated. But in principle it would seem possible to assign stochastic properties to φ as well as (or instead of) u. This would require making φ a function of t as well as of τ; for instance,

$$x(t) = \int_0^\infty \varphi(t,\ \tau)\, x(t-\tau)\, d\tau \quad + \quad u(t), \tag{6}$$

where φ (as a function of t), or u, or both, are specified continuous stochastic processes.

When the production function is of the Cobb-Douglas type it may be written as

$$X_0^{(j)} = \sum_{i=1}^{K} \alpha_i^{(j)}\, X_i^{(j)} \quad + \quad \alpha_0^{(j)}\ . \tag{2.2}$$

Here X_0 is the logarithm of the product and the other X's are logarithms of the factors. The superscript refers to the jth firm.

Now the $\alpha^{(j)}$ (the "productivities") may differ from firm to firm and the firm in existence may be considered as a sample from the universe of all possible firms with $\alpha^{(j)}$ which are stochastic variates.

The current practice does this, but only for $\alpha_0^{(j)}$. It is difficult to see any reason, except that of expediency, for regarding other α's as fixed.

On the other hand, if the α's are random variables, it is reasonable to try to estimate their joint distribution. This, however, seems to be a rather difficult task. In the following section we show what problems are involved in such an estimation procedure. The example treated is highly oversimplified and should be regarded only as illustrating the mathematics involved.

3.1. Let the nonadditive disturbance model be

$$x_1 + u_3 x_2 = u_1\,, \qquad \alpha x_1 + x_2 = u_2\,, \tag{3.1}$$

where the x's are observed, the u's are the disturbances (with u_3 entering in a nonadditive fashion), and α is a (known or unknown) constant.

The problem in general is that of estimating α and the distribution of the u's. This can be accomplished by the maximum-likelihood method if the likelihood function is known. This function will now be obtained for the case of jointly normally distributed disturbances; the sample is assumed to be random so that equation (3.4) holds. [In (3.4), $h(x_1, x_2)$ is the joint probability density function of the observed variates.]

It should be noted that the problem thus defined would in general be indeterminate; without additional *a priori* knowledge only certain relations between the parameters can be estimated. It will be assumed, however, that after the likelihood function is formed, these restrictions will be formulated and taken into account in the procedure of maximizing the likelihood function.

It should also be noted that it may not be very realistic to endow the disturbances with a normal distribution. It might, for instance, be more appropriate to make the marginal distribution of u_3 one of χ^2 type, etc.

3.2. Thus let the joint probability density function of the disturbances be

$$f(u_1, u_2, u_3) = (2\pi)^{-3/2} (\det [\sigma^{ij}])^{1/2} \exp(-\frac{1}{2} Q) \tag{3.2}$$

with

$$Q = \sum_{i,j=1}^{3} \sigma^{ij}(u_i - \alpha_i)(u_j - \alpha_j). \tag{3.3}$$

We shall now proceed to obtain the joint distribution $h(x_1, x_2, u_3)$ and then integrate out u_3. The result of the integration, say $g(x_1, x_2)$ will yield the logarithm of the likelihood function

$$\sum_{t=1}^{T} \log g(x_{1t}, x_{2t}). \tag{3.4}$$

The computation of $h(x_1, x_2, u_3)$ is straightforward.

For convenience of notation we may add to (3.1) a third equation

$$x_3 = u_3 \tag{3.5}$$

involving the auxiliary variable x_3 and consider (3.1) with (3.5) as a transformation from (u_1, u_2, u_3) to (x_1, x_2, x_3). In the result u_3 may again be written instead of x_3.

The Jacobian of the transformation is

$$|J| = |1 - \alpha x_3|, \tag{3.6}$$

where $|\ |$ is the absolute-value symbol.

Making the appropriate substitutions we find

$$h(x_1, x_2, x_3) = (2\pi)^{-3/2} (\det[\sigma^{ij}])^{1/2} |1 - \alpha u_3| \times \exp\{-\frac{1}{2}\frac{(u_3 + \psi)^2}{\varphi^2} - \frac{1}{2}\lambda\}, \tag{3.7}$$

where φ, ψ, and λ do not contain u_3.[1]

[1]The values of φ, ψ, and λ are as follows:

$$\begin{aligned} \varphi &= 1/\sqrt{B_2}, \\ \psi &= B_1/B_2, \\ \lambda &= B_0 - B_1^2/B_2, \end{aligned} \tag{f.1}$$

where the B's are given by

$$\begin{aligned} B_2 &= \sigma^{11} x_2^2 + 2\sigma^{13} x_2 + \sigma^{33}, \\ B_1 &= \sigma^{11} w_1 x_2 + \sigma^{12} x_2 w_2 + \sigma^{13}(w_1 - x_2 \alpha_3) \\ &\qquad + \sigma^{23} w_2 - \sigma^{33} \alpha_3, \\ B_0 &= \sigma^{11} w_1^2 + 2\sigma^{12} w_1 w_2 + \sigma^{22} w_2^2 - 2\sigma^{13} w_1 \alpha_3 \\ &\qquad - 2\sigma^{23} w_2 \alpha_3 + \sigma^{33} \alpha_3^2. \end{aligned} \tag{f.2.1}$$

The w's in (f.2.1) have the following meaning:

$$\begin{aligned} w_1 &= x_1 - \alpha_1, \\ w_2 &= \alpha x_1 + x_2 - \alpha_2. \end{aligned} \tag{f.2.2}$$

Hence

$$g(x_1, x_2)$$

$$(3.8) \quad = (2\pi)^{-\frac{3}{2}} (\det [\sigma^{ij}])^{\frac{1}{2}} \; e^{-\frac{1}{2}\lambda} \int_{-\infty}^{\infty} |1 - \alpha u_3| \; e^{-\frac{1}{2}\frac{(u_3+\psi)^2}{\varphi^2}} du_3 .$$

It will be observed that for $\alpha = 0$, this integral may be evaluated in terms of elementary functions. In this case we have[1]

$$(3.9) \qquad g_0(x_1, x_2) = (2\pi)^{-1} (\det [\sigma^{ij}])^{\frac{1}{2}} e^{-\frac{1}{2}\lambda} \varphi .$$

---

[1]When the u's are independent with zero means and unit variances, $g_0(x_1, x_2)$ becomes

$$(f.3) \quad g_{00}(x_1, x_2) = (2\pi)^{-1} (1 + x_2^2)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_1^2 + x_2^2 - \frac{x_1^2 x_2^2}{1 + x_2^2})\} .$$

XIX. NOTE ON RANDOM COEFFICIENTS

BY HERMAN RUBIN

Let us consider an equation of the form

$$y_t = \sum_{k=1}^{K} a_{kt} x_{kt}, \qquad t = 1, \ldots, T, \tag{1}$$

which defines the dependent variable y_t in terms of the fixed variables x_{kt} and the coefficients a_{kt} which are assumed to be normally and independently distributed with mean α_k and variance $\sigma_k^2 \varphi_{kt}$, where the inverse weights φ_{kt} are known functions of the fixed variates.[1] We could estimate the parameters α_k by least squares, i.e., by the formula

$$\bar{\alpha}_k = \sum_{l=1}^{L} m_{yx_l} m^{lk},$$

where

$$m_{yx_l} = \sum_{t=1}^{T} y_t x_{lt},$$

and m^{lk} is the element in the lth row and kth column of

$$(m_{kl})^{-1} = \Big(\sum_{t=1}^{T} x_{kt} x_{lt}\Big)^{-1}.$$

Although this method can be shown, under certain conditions, to be consistent, it would not be efficient.

We shall derive equations defining maximum-likelihood estimates of the α_i and σ_j^2. Let us consider the distribution of the y_t. We see that the quantities

$$z_t = y_t - \sum_{k=1}^{K} \alpha_k x_{kt}, \qquad t = 1, \ldots, T, \tag{2}$$

[1] See also [XVIII].

are normally and independently distributed with mean 0 and variance

$$\sum x_{kt}^2 \sigma_k^2 \varphi_{kt}^2 . \tag{3}$$

Then the likelihood function of the observation becomes

$$L = (2\pi)^{-\frac{T}{2}} \prod_{t=1}^{T} \left(\sum x_{kt}^2 \sigma_k^2 \varphi_{kt}\right)^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} \frac{\left(y_t - \sum_{k=1}^{K} \alpha_k x_{kt}\right)^2}{\sum_{k=1}^{K} x_{kt}^2 \sigma_k^2 \varphi_{kt}}\right\}. \tag{4}$$

Let us maximize $\log L$ with respect to α_i and σ_j^2, subject to the restriction $\sigma_j^2 \geq 0$. We obtain

$$\sum_{t=1}^{T} \frac{x_{it}\left(y_t - \sum_{k=1}^{K} \alpha_k x_{kt}\right)}{\sum_{k=1}^{K} x_{kt}^2 \sigma_k^2 \varphi_{kt}} = 0, \tag{5}$$

$$\sum_{t=1}^{T} \frac{x_{jt}^2 \varphi_{jt}}{\sum_{k=1}^{K} x_{kt}^2 \sigma_k^2 \varphi_{kt}} = \sum_{t=1}^{T} \frac{x_{jt}^2 \varphi_{jt}\left(y_t - \sum_{k=1}^{K} \alpha_k x_{kt}\right)^2}{\left(\sum_{k=1}^{K} x_{kt}^2 \sigma_k^2 \varphi_{kt}\right)^2} - \lambda_j , \tag{6}$$

$$\lambda_j \sigma_j^2 = 0 , \tag{7}$$

$$\sigma_j^2 \geq 0 . \tag{8}$$

It is necessary to introduce the Lagrange multipliers λ_j because the solutions of (5) and (6) with $\lambda_j = 0$ might give negative σ_j^2. We take

that solution of (5), (6), (7), and (8) which gives to (4) its highest value. It should be noted that it is unnecessary to consider solutions having λ_j's $\neq 0$ if there is a solution with λ_j's $= 0$, and, in general, if there is a solution with only $\lambda_{j_1}, \ldots, \lambda_{j_\beta} \neq 0$, it is not necessary to consider solutions with those λ_j's $\neq 0$ and other λ_j's $\neq 0$.

REFERENCES

ANDERSON, R. L., "Use of Variance Components in the Analysis of Hog Prices in Two Markets," *Journal of the American Statistical Association*, Vol. 42, December, 1947, pp. 612-634.

ANDERSON, T. W., and LEONID HURWICZ, "Errors and Shocks in Economic Relationships," Paper presented at the meeting of The Econometric Society, September 6-18, 1947, in Washington, D.C.; abstract in *Econometrica*, Vol. 16, January, 1948, pp. 36-37.

ANDERSON, T. W., and HERMAN RUBIN, "Estimation of the Parameters of a Single Stochastic Difference Equation in a Complete System," *Annals of Mathematical Statistics*, Vol. 20, March, 1949, pp. 46-63 (and to be included in *Cowles Commission Paper, New Series, No. 36*).

_____, "Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations," to be published (and to be included in *Cowles Commission Paper, New Series, No. 36*).

ANDREWS, WILLIAM H., JR., *see* JACOB MARSCHAK and WILLIAM H. ANDREWS, JR.

BENTZEL, H., and H. WOLD, "On Statistical Demand Analysis from the Viewpoint of Simultaneous Equations," *Skandinavisk Aktuarietidskrift*, Vol. 29, Nos. 1-2, 1946, pp. 95-114.

BIRKHOFF, GARRETT, and SAUNDERS MACLANE, *A Survey of Modern Algebra*, New York: The Macmillan Co., 1941, 450 pp.

BROOKNER, RALPH J., "Choice of One Among Several Statistical Hypotheses," *Annals of Mathematical Statistics*, Vol. 16, September, 1945, pp. 221-242.

CHERNOFF, HERMAN, "Gradient Methods of Maximization in Estimating Economic Parameters," Paper presented at the Madison meeting of The Econometric Society, September 7-10, 1948; abstract in *Econometrica*, Vol. 17, January, 1949, pp. 75-76.

CRAMÉR, HARALD, *Random Variables and Probability Distributions*, London: Cambridge, England: The University Press, 1937, 120 pp.

DAVID, F. N., and J. NEYMAN, "Extension of the Markoff Theorem on Least Squares," *Statistical Research Memoirs*, Vol. II, London: Department of Statistics, University of London, University College, 1938, pp. 105-116.

DIXON, WILFRID J., "Further Contributions to the Problem of Serial Correlation," *Annals of Mathematical Statistics*, Vol. 15, June, 1944, pp. 119-144.

DOOB, J. L., "The Elementary Gaussian Processes," *Annals of Mathematical Statistics*, Vol. 15, September, 1944, pp. 229-282.

FRISCH, RAGNAR, 1929, "Correlation and Scatter in Statistical Variables," *Nordic Statistical Journal*, Vol. 1, 1929, pp. 36-102.

_____, 1933, *Pitfalls in the Construction of Statistical Demand Curves,* Veröffentlichungen der Frankfurter Gesellschaft für Konjunkturforschung, Neue Folge, Heft 5, Leipzig: Hans Buske Verlag, 1933, 39 pp.

_____, 1934, *Statistical Confluence Analysis by Means of Complete Regression Systems,* Oslo: Universitetets Økonomiske Institutt, 1934, 192 pp.

_____, 1938, "Statistical Versus Theoretical Relations in Economic Macrodynamics," Memorandum prepared for a conference in Cambridge, England, July 18-20, 1938, to discuss drafts of Tinbergen's League of Nations Publications; mimeographed.

FRISCH, RAGNAR, and BRUCE D. MUDGETT, "Statistical Correlation and the Theory of Cluster Types," *Journal of the American Statistical Association,* Vol. 26, December, 1931, pp. 375-392.

GEARY, R. C., 1942, "Inherent Relations between Random Variables," *Proceedings of the Royal Irish Academy,* Vol. 47, Section A, March, 1942, pp. 63-196.

_____, 1943, "Relations between Statistics: The General and the Sampling Problem when the Samples Are Large," *Proceedings of the Royal Irish Academy,* Vol. 49, Section A, December, 1943, pp. 177-196.

GIRSHICK, M. A., and TRYGVE HAAVELMO, "Statistical Analysis of the Demand for Food: Examples of Simultaneous Estimation of Structural Equations," *Econometrica,* Vol. 15, April, 1947, pp. 79-110 (and reprinted as *Cowles Commission Paper, New Series, No. 24*).

HAAVELMO, TRYGVE, 1943, "The Statistical Implications of a System of Simultaneous Equations," *Econometrica,* Vol. 11, January, 1943, pp. 1-12.

_____, 1944, "The Probability Approach in Econometrics," *Econometrica,* Vol. 12, Supplement, July, 1944, 118 pp. (and reprinted as *Cowles Commission Paper, New Series, No. 4*).

_____, 1947-A, "Methods of Measuring the Marginal Propensity to Consume," *Journal of the American Statistical Association,* Vol. 42, March, 1947, pp. 105-122 (and reprinted as *Cowles Commission Paper, New Series, No. 22*).

_____, 1947-B, "Quantitative Research in Agricultural Economics: The Interdependence Between Agriculture and the National Economy," *Journal of Farm Economics,* Vol. 24, 1947, pp. 910-924 (and included in *Cowles Commission Paper, New Series, No. 27*).

_____, *see also* M. A. GIRSHICK and TRYGVE HAAVELMO.

HARDY, G. H., J. E. LITTLEWOOD, and G. PÓLYA, *Inequalities,* Cambridge, England: The University Press, 1934, 314 pp.

HOTELLING, HAROLD, 1933, "Analysis of a Complex of Statistical Variables in Principal Components," *Journal of Educational Psychology,* Vol. 24, September and October, 1933, pp. 417-444 and 498-520.

_____, 1936-A, "Simplified Calculation of Principal Components," *Psychometrika,* Vol. 1, March, 1936, pp. 27-35.

_____, 1936-B, "Relations Between Two Sets of Variables," *Biometrica,* Vol. 28, December, 1936, pp. 321-377.

_____, 1940, "The Selection of Variates for Use in Prediction with Some Comments on the Problem of Nuisance Parameters," *Annals of Mathematical Statistics,* Vol. 11, September, 1940, pp. 271-283.

_____, 1943-A, "Some New Methods in Matrix Calculation," *Annals of Mathematical Statistics*, Vol. 14, March, 1943, pp. 1-34.

_____; 1943-B, "Further Points on Matrix Calculation and Simultaneous Equations," *Annals of Mathematical Statistics*, Vol. 14, December, 1943, pp. 440-441.

_____, 1949, "Practical Problems of Matrix Calculation," *Proceedings of the Berkeley Symposium on Probability and Statistics*, Jerzy Neyman, ed., Berkeley and Los Angeles: University of California Press, 1949, pp. 275-293.

HURWICZ, LEONID, 1944, "Stochastic Models of Economic Fluctuations," *Econometrica*, Vol. 12, April, 1944, pp. 114-124 (and reprinted as *Cowles Commission Paper, New Series, No. 3*).

_____, 1946, "Theory of the Firm and of Investment," *Econometrica*, Vol. 14, April, 1946, pp. 109-136 (and reprinted as *Cowles Commission Paper, New Series, No. 16*).

_____, 1947, "Some Problems Arising in Estimating Economic Relations," *Econometrica*, Vol. 15, July, 1947, pp. 236-240.

_____, *see also* T. W. ANDERSON and LEONID HURWICZ.

_____, *see also* JACOB MARSCHAK; LEONID HURWICZ; *et al.*

JOHNSON, EVAN, JR., "Estimates of Parameters by Means of Least Squares," *Annals of Mathematical Statistics*, Vol. 11, December, 1940, pp. 453-456.

KENDALL, M. G., "A New Measure of Bank Correlation," *Biometrika*, Vol. 30, June, 1938, pp. 81-93.

KLEIN, LAWRENCE R., 1946-A, "Macroeconomics and the Theory of Rational Behavior," *Econometrica*, Vol. 14, April, 1946, pp. 93-108 (and reprinted as *Cowles Commission Paper, New Series, No. 14*).

_____, 1946-B, "A Post-Mortem on Transition Predictions of National Product," *Journal of Political Economy*, Vol. 54, August, 1946, pp. 289-308 (and reprinted as *Cowles Commission Paper, New Series, No. 18*).

_____, 1947, "The Use of Econometric Models as a Guide to Economic Policy," *Econometrica*, Vol. 15, April, 1947, pp. 111-151 (and reprinted as *Cowles Commission Paper, New Series, No. 23*).

_____, 1950, *Economic Fluctuations in the United States, 1921-1941*, Cowles Commission Monograph No. 11, New York: John Wiley & Sons, 1950, about 170 pp.

KOOPMANS, TJALLING C., 1937, *Linear Regression Analysis of Economic Time Series*, Haarlem: De Erven F. Bohn N. V., 1937, 150 pp.

_____, 1942, "Serial Correlation and Quadratic Forms in Normal Variables," *Annals of Mathematical Statistics*, Vol. 13, March, 1942, pp. 14-33.

_____, 1945, "Statistical Estimation of Simultaneous Economic Relations," *Journal of the American Statistical Association*, Vol. 40, December, 1945, pp. 448-466 (and reprinted as *Cowles Commission Paper, New Series, No. 11*).

_____, 1949, "Identification Problems in Economic Model Construction," *Econometrica*, Vol. 17, April, 1949, pp. 125-144 (and reprinted as *Cowles Commission Paper, New Series, No. 31*).

_____, *see also* JACOB MARSCHAK; LEONID HURWICZ; *et al.*

LEIPNIK, ROY B., "Distribution of the Serial Correlation Coefficient in a Circularly Correlated Universe," *Annals of Mathematical Statistics*, Vol. 18, March, 1947, pp. 80-87 (and included in *Cowles Commission Paper, New Series, No. 21*).

_____, *see also* JACOB MARSCHAK; LEONID HURWICZ; *et al.*

LITTLEWOOD, J. E., *see* G. H. HARDY, J. E. LITTLEWOOD, and G. PÓLYA.

MACDUFFEE, CYRUS COLTON, *Vectors and Matrices*, Menasha, Wisconsin: Mathematical Association of America, 1943, 192 pp.

MACLANE, SAUNDERS, *see* GARRETT BIRKHOFF and SAUNDERS MACLANE.

MADOW, WILLIAM G., "Note on the Distribution of the Serial Correlation Coefficient," *Annals of Mathematical Statistics*, Vol. 16, September, 1945, pp. 308-310.

MANN, HENRY B., "Nonparametric Tests against Trend," *Econometrica*, Vol. 13, July, 1945, pp. 245-259.

MANN, H. B., and A. WALD, "On the Statistical Treatment of Linear Stochastic Difference Equations," *Econometrica*, Vol. 11, July-October, 1943, pp. 173-220.

MARSCHAK, JACOB, 1947-A, "Economic Structure, Path, Policy, and Prediction," *American Economic Review, Papers and Proceedings of the 59th Annual Meeting of the American Economic Association*, Vol. 37, May, 1947, pp. 81-84.

_____, 1947-B, "Statistical Inference from Nonexperimental Observations: An Economic Example," Paper presented at the meeting of the International Statistical Institute and The Econometric Society, September 6-18, 1947, in Washington, D.C.; to be published in 1950 in *Proceedings of the International Statistical Conference* (and to be reprinted as *Cowles Commission Paper, New Series, No. 32*); abstract in *Econometrica*, Vol. 16, January, 1948, pp. 53-55.

MARSCHAK, JACOB, and WILLIAM H. ANDREWS, JR., "Random Simultaneous Equations and the Theory of Production," *Econometrica*, Vol. 12, July-October, 1944, pp. 143-205 (and reprinted as *Cowles Commission Paper, New Series, No. 5*).

MARSCHAK, JACOB; LEONID HURWICZ; TJALLING C. KOOPMANS and ROY B. LEIPNIK; "Estimating Relations from Nonexperimental Observations," Papers presented at the Cleveland meeting of The Econometric Society, January 24-27, 1946; abstracts in *Econometrica*, Vol. 14, April, 1946, pp. 165-172 (and included in *Cowles Commission Paper, New Series, No. 17*)

MARSHALL, ALFRED, *Principles of Economics*, Eighth Edition, London: Macmillan & Co., 1920, 754 pp.

MUDGETT, BRUCE D., *see* RAGNAR FRISCH and BRUCE D. MUDGETT.

NEYMAN, J., "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," *Philosophical Transactions of the Royal Society of London, Series A*, Vol. 236, August 30, 1937, pp. 330-380.

_____, *see also* F. N. DAVID and J. NEYMAN.

PIERCE, B. O., *A Short Table of Integrals*, New York: Ginn and Co., 1929, 156 pp.

PÓLYA, G., *see* G. H. HARDY, J. E. LITTLEWOOD, and G. PÓLYA.

REIERSØL, OLAV, "Residual Variables in Regression and Confluence Analysis,"

Skandinavisk Aktuarietidskrift, Vol. 28, Nos. 3-4, 1945, pp. 201-217.

RUBIN, HERMAN, 1945, "On the Distribution of the Serial Correlation Coefficient," *Annals of Mathematical Statistics,* Vol. 16, June, 1945, pp. 211-215 (and reprinted as *Cowles Commission Paper, New Series, No. 10*).

_____, 1946, "Asymptotic Distribution of Moments from a System of Linear Stochastic Difference Equations," Abstract No. 332 in *Bulletin of the American Mathematical Society,* Vol. 52, September, 1946, pp. 827-828.

_____, 1948, "Some Results on the Asymptotic Distribution of Maximum- and Quasi-Maximum-Likelihood Estimates," Abstract No. 529 in *Bulletin of the American Mathematical Society,* Vol. 54, November, 1948, p. 1080, and in *Annals of Mathematical Statistics,* Vol. 19, December, 1948, p. 598.

_____, 1949, "Properties of Maximum- and Quasi-Maximum-Likelihood Estimates of Parameters of a System of Linear Stochastic Difference Equations with Serially Correlated Disturbances," abstract in *Annals of Mathematical Statistics,* Vol. 20, March, 1949, p. 137.

_____, *see also* T. W. ANDERSON and HERMAN RUBIN.

SNEDECOR, GEORGE W., *Calculation and Interpretation of Analysis of Variance and Covariance,* Ames, Iowa: Collegiate Press, Inc., 1934, 96 pp.

TINBERGEN, J., *Statistical Testing of Business-Cycle Theories:* Vol. I, *A Method and Its Application to Investment Activity;* Vol. II, *Business Cycles in the United States of America, 1919-1932;* Geneva: League of Nations, 1939, 164 and 244 pp.

TINTNER, GERHARD, 1942, "A Contribution to the Nonstatic Theory of Production," in *Studies in Mathematical Economics and Econometrics, in Memory of Henry Schultz,* O. Lange, F. McIntyre, T. O. Yntema, eds., Chicago: The University of Chicago Press, 1942, pp. 92-109.

_____, 1946, "Multiple Regression for Systems of Equations," *Econometrica,* Vol. 14, January, 1946, pp. 5-36.

ULLMAN, JOSEPH, "The Probability of Convergence of an Iterative Process of Inverting a Matrix," *Annals of Mathematical Statistics,* Vol. 15, June, 1944, pp. 205-213.

WALD, ABRAHAM, 1940, "The Fitting of Straight Lines If Both Variables Are Subject to Error," *Annals of Mathematical Statistics,* Vol. 11, September, 1940, pp. 284-300.

_____, 1942, *On the Principles of Statistical Inference,* Notre Dame Mathematical Lectures, No. 1, South Bend, Indiana: University of Notre Dame, 1942, 47 pp.

_____, 1944, "Note on a Lemma," *Annals of Mathematical Statistics,* Vol. 15, September, 1944, pp. 330-333.

_____, *see also* H. B. MANN and A. WALD.

WALD, A., and J. WOLFOWITZ, "Confidence Limits for Continuous Distribution Functions," *Annals of Mathematical Statistics,* Vol. 10, June, 1939, pp. 105-118.

WATSON, G. N., *see* E. T. WHITTAKER and G. N. WATSON.

WAUGH, F. V., "A Note Concerning Hotelling's Method of Inverting a Partitioned Matrix," *Annals of Mathematical Statistics,* Vol. 16, June, 1945, pp. 216-217.

WHITTAKER, E. T., and G. N. WATSON, *A Course of Modern Analysis*, American Edition, Cambridge, England: The University Press; New York: The Macmillan Co., 1945, 608 pp.

WILKS, S. S., *Mathematical Statistics*, Princeton: Princeton University Press, 1943, 284 pp.

WILLIAMS, J. D., "Moments of the Ratio of the Mean Square Successive Differences to the Mean Square Difference in Samples from a Normal Universe," *Annals of Mathematical Statistics*, Vol. 12, June, 1941, pp. 239-241.

WINTNER, A., *Lectures by Aurel Wintner on Asymptotic Distributions and Infinite Convolutions*, Ann Arbor, Michigan: Edwards Brothers, 1938, 54 pp.

WOLD, HERMAN, *A Study in the Analysis of Stationary Time Series*, Uppsala: Almqvist & Wiksells, 1938, 214 pp.

———, *see also* H. BENTZEL and H. WOLD.

WOLFOWITZ, J., *see* A. WALD and J. WOLFOWITZ.

INDEX