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 usually 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 distribution 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

$$(1.5) \qquad \sum_{i=1}^{G} \beta_{gi0}\, y_i(t) = \\ -\sum_{i=1}^{G} \sum_{\tau=1}^{\tau^{\square}} \beta_{gi\tau}\, y_i(t-\tau) - \sum_{k=1}^{K} \sum_{\tau=0}^{\tau^{\square}} \gamma_{gk\tau}\, z_k(t-\tau) + u_g(t).$$

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

$$\mathrm{B}_0 = \begin{bmatrix} \beta_{110} & \cdots & \beta_{1G0} \\ \cdot & \cdots & \cdot \\ \beta_{G10} & \cdots & \beta_{GG0} \end{bmatrix} \tag{1.6}$$

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 $\overline{B}$ with elements $\overline{\beta}_{gi} = \sum_{\tau=1}^{\tau^0} \beta_{gi\tau}$, 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 $\overline{B}$ 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,

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

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,

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

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,

$$(1.12) \qquad \sum_{i=1}^{G} \sum_{\tau=0}^{\tau^{\square}} \chi^{(g_r)}_{ri\tau}\, \beta_{g_r i\tau} + \sum_{k=1}^{K} \sum_{\tau=0}^{\tau^{\square}} \psi^{(g_r)}_{rk\tau}\, \gamma_{g_r k\tau} = 0,$$

$$r = 1,\ 2,\ \dots,\ R^{(1)}_{\alpha}\ .$$

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 $\chi^{(g_r)}_{ri\tau}$ and $\psi^{(g_r)}_{rk\tau}$ 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_r} and $y_{i'_r}$, respectively, are required to have the same ratio in two different structural equations (numbered g_r and g'_r):

$$\begin{bmatrix} \beta_{g_r i_r \tau_r} & \beta_{g_r i'_r \tau_r} \\ \beta_{g'_r i_r \tau_r} & \beta_{g'_r i'_r \tau_r} \end{bmatrix} = 0. \tag{1.13}$$

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_r} is supposed to occur with the same lag distribution in two equations numbered g_r and g'_r, this leads to a restriction of the type

$$\begin{bmatrix} \beta_{g_r i_r \tau_r} & \beta_{g_r i_r \tau'_r} \\ \beta_{g'_r i_r \tau_r} & \beta_{g'_r i_r \tau'_r} \end{bmatrix} = 0. \tag{1.14}$$

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, ..., 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

$$(1.15) \qquad \Sigma = \begin{bmatrix} \Sigma_1 & 0 & \dots & 0 \\ 0 & \Sigma_2 & \dots & 0 \\ \cdot & \cdot & \dots & \cdot \\ 0 & 0 & \dots & \Sigma_N \end{bmatrix},$$

where each of the matrices Σ_1, ..., Σ_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 $[\, \mathrm{B} \quad \Gamma \,]$. 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:

$$(1.17) \begin{cases} (1.17\text{a}) \qquad \beta_{g\,i_g 0} = 1, \qquad g = 1,\ 2,\ \dots,\ G, \\ (1.17\text{b}) \qquad \sigma_{gg} = 1, \qquad g = 1,\ 2,\ \dots,\ G. \end{cases}$$

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 $\sigma^{1/2}_{gg}$, 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:

$$(2.1) \qquad y(t) \equiv [y_1(t) \quad \cdots \quad y_G(t)],$$

...

$$z(t) \equiv \begin{bmatrix} y_1(t-1) & \cdots & y_G(t-\tau^{\square}) & z_1(t) & \cdots & z_K(t-\tau^{\square}) \end{bmatrix},$$

$$u(t) \equiv \begin{bmatrix} u_1(t) & \cdots & u_G(t) \end{bmatrix},$$

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

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

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:

$$\mathrm{B}y'(t) + \Gamma z'(t) = u'(t), \tag{2.2}$$

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 $\mathrm{A}x' = u'$, where $\mathrm{A} = [\ \mathrm{B} \quad \Gamma\ ]$, $x = [\ y \quad z\ ]$.

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

$$(2.4)\qquad \begin{gathered} \left| du_1(1)\, du_2(1) \ \cdots \ du_G(T) \right| = J_T \left| dy_1(1)\, dy_2(1) \ \cdots \ dy_G(T) \right|, \\ J_T \equiv \left| \frac{\partial\{u_1(1), \ldots, u_G(1); u_1(2), \ldots, u_G(2); \ldots ; u_1(T), \ldots, u_G(T)\}}{\partial\{y_1(1), \ldots, y_G(1); y_1(2), \ldots, y_G(2); \ldots ; y_1(T), \ldots, y_G(T)\}} \right| . \end{gathered}$$

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

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

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

$$(2.6) \qquad \begin{bmatrix} \mathrm{B} & & & \\ 0 & \mathrm{B} & & \\ \cdot & \cdot & \cdots & \\ 0 & 0 & \cdots & \mathrm{B} \end{bmatrix},$$

if, as before, $\mathrm{B} = [\ \beta_{gi0}\ ]$. 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

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

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

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

In all integrations over the whole or part of the sample space, this function must, of course, be multiplied with the volume element $d y_1(1) \cdots d y_G(T)$.

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

$$\det{}^{T}(\mathrm{B})\prod_{t=1}^{T} f\{\,\mathrm{B}\,y'(t) + \Gamma\,z'(t)\} \equiv \tag{2.9}$$
$$\det{}^{T}(\mathrm{B}^{\oplus})\prod_{t=1}^{T} f^{\oplus}\{\,\mathrm{B}^{\oplus}\,y'(t) + \Gamma^{\oplus}\,z'(t)\}.$$

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.

$$\det \mathrm{B} \neq 0.$$

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

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

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

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

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

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

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:

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

Consequently

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

The square matrix of order G

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

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

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

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 $\mathrm{B}^{\oplus}$, (2.14), and

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

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:

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

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*

$$(2.19)\qquad \begin{aligned} \alpha^{\oplus}_{g_0 k} &= \upsilon_{g_0 g_0}\,\alpha_{g_0 k}, \qquad \sigma^{\oplus}_{g_0 g_0} = \upsilon^2_{g_0 g_0}\,\sigma_{g_0 g_0}, \\ \sigma^{\oplus}_{g_0 g} &= \upsilon_{g_0 g_0}\sum_h \sigma_{gh}\,\upsilon_{gh}, \qquad\qquad g \neq g_0, \end{aligned}$$

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 $(\mathrm{A}^{\oplus}, \Sigma^{\oplus})$ 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 $\bar{R}_0 \equiv 0, \quad \bar{R}_G \equiv R_1 + R_2 + \cdots + R_G \equiv R_\alpha^{(1)}$.

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_0th *structural equation in* (2.3) *by the a priori restrictions* (2.20), *is that the matrix*

$$(2.21)\qquad A^{(g_0)} \equiv [\ \alpha_{gk_r}\ ],$$

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

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, $A^{(g_0)}$ cannot have a rank higher than $G-1$ because its g_0th row consists of zeros only. Stated in other words, the condition in Theorem 2.2.1 requires that from the g_0th row of the matrix $A = [\ \alpha_{gk}\ ]$ 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_0th *structural equation by the a priori restrictions* (2.20) *is that the number* R_{g_0} *of these restrictions involving coefficients of the* g_0th *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

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

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 $\upsilon_{g_0 h}$ 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

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

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_0th structural equation.

That this condition is also necessary is seen if we now assume that the g_0th 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 $\upsilon_{g_0 h}^{(1)}$ and $\upsilon_{g_0 h}^{(2)}$, say, of which the first can be taken to satisfy (2.23). The more general solution $\upsilon_{g_0 h} = \lambda_1 \upsilon_{g_0 h}^{(1)} + \lambda_2 \upsilon_{g_0 h}^{(2)}$ 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:

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

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 "χ" 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

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

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

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

By a repetition of the previous reasoning, it is seen that the g_0th 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_0th 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_0th 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

$$(2.27) \qquad \Phi'_{g_0}\lambda'_{g_0} \neq 0.$$

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

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

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

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

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

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

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

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

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 applicablþ 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]

$$(2.35) \qquad \rho(\Phi \quad \Psi) < \rho(\Phi) + \rho(\Psi),$$

then there exist two nonvanishing vectors λ and μ such that

$$(2.36) \qquad \Phi\lambda' + \Psi\mu' = 0, \qquad \Phi\lambda' \neq 0.$$

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

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

(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),

$$(2.38) \qquad \rho(\Phi \quad \Psi) < c(\Phi \quad \Psi).$$

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

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

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 $\bar{\Phi}$ and $\bar{\Psi}$ respectively such that

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

and, from (2.35) and (2.40),

$$(2.41) \qquad \rho(\bar{\Phi} \quad \bar{\Psi}) \leq \rho(\Phi \quad \Psi) < \rho(\bar{\Phi}) + \rho(\bar{\Psi}).$$

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

$$(2.42) \qquad \bar{\Phi}\,\bar{\lambda}' + \bar{\Psi}\,\bar{\mu}' = 0, \qquad \bar{\Phi}\,\bar{\lambda}' \neq 0.$$

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

$$(2.43) \qquad \rho(A_0 \cdot \Phi_g') = \rho(\Phi_g').$$

As before, the restriction matrix Φ_g is so chosen that

$$(2.44) \qquad \rho(\Phi_g') = c(\Phi_g') = R_g$$

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)

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

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

$$(2.46) \qquad R_g < G - 1 \equiv K_y - 1,$$

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

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

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

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

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* $\overline{\Phi}_g^{(0)}$ *defining*

$$(2.49) \qquad \overline{R}_g = G - 1 - R_g$$

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

$$\rho({}_g\mathrm{A}_0 \cdot \overline{\Phi}'(g)^{(0)}) = G - 1 = K_y - 1 = \rho(\overline{\Phi}'(g)^{(0)}). \tag{2.51}$$

The first step is to choose an arbitrary orthogonal complement of ${}_g\mathrm{A}_0$, 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

$${}_g\mathrm{A}_0 \cdot \overline{\overline{\Phi}}'(g)^{(0)} = 0. \tag{2.53}$$

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),

$${}_g\mathrm{A} \cdot \Phi_g' \cdot \lambda_g' = 0, \tag{2.56}$$

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

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

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

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

$$(2.58)\qquad \rho({}_g\mathrm{A}_0 \cdot \bar{\bar{\Phi}}{}'(g)^{(0)}\ ) = \rho({}_g\mathrm{A}_0),$$

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

$$(2.59)\qquad \rho({}_g\mathrm{A}_0 \cdot \bar{\bar{\Phi}}{}'(g)^{(0)}) = \rho({}_g\mathrm{A}_0 \cdot \bar{\Phi}{}'(g)^{(0)}\ ).$$

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

$$(2.60)\qquad \rho(\ \mathrm{A}_0) = K_y = 1 + \rho({}_g\mathrm{A}_0).$$

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

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

in a neighborhood ${}_gN_\alpha$ of the point ${}_g\mathrm{A}_0$ in the space of the parameters ${}_g\mathrm{A}$ 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 coordinates ${}_gA$ 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,

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

in which the vector

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

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

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

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

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

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 $(\overline{A}^{\oplus}, \overline{\Sigma}^{\oplus})$ that is obtainable from (A, Σ) by a linear transformation (2.62) and satisfies the combined (a priori and dummy) restrictions is obtainable from $(\overline{A}, \overline{\Sigma})$ 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 $(\overline{A}, \overline{\Sigma})$ 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 $\overline{A}$ only as regards the gth row $\alpha(g)$ or

(2.67) $$ {}_gA = {}_g\overline{A}. $$

This means that we can confine ourselves to transformations

(2.68) $$ A = \Upsilon^{-1}\,\overline{A}, \qquad \Sigma = \Upsilon^{-1}\,\overline{\Sigma}\,\Upsilon'^{-1}, $$

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

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

For the gth row of A this gives us in particular

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

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

$$(2.71)\left\{\begin{array}{lrl}(2.71r) & 0 = & -\ {}_g\upsilon(g)\cdot{}_g\overline{A}\cdot\Phi_g'^{(0)}\,, \\ (2.71\overline{r}) & \alpha(g)\cdot\overline{\Phi}_g'^{(0)} = & -\ {}_g\upsilon(g)\cdot{}_gA\cdot\overline{\Phi}_g'^{(0)}\,.\end{array}\right.$$

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$\overline{r}$) 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* $(\overline{A}, \overline{\Sigma})$ *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* $(\overline{A}, \overline{\Sigma})$ *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 further 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.

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

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:

$$
\begin{array}{lll}
 & \begin{array}{l}(2.74l)\\ (R_\alpha^{(1)} + G\\ \text{restrictions})\end{array} & \left\{\begin{array}{lll}(2.74lh) & \upsilon(g)\cdot\mathrm{A}\cdot\Phi_g' = 0, & R \text{ restrictions},\\ (2.74ln) & \upsilon(g)\cdot\mathrm{A}\cdot\iota(i_g) = 1, & \text{one restriction},\\ & & g = 1, \ldots, G.\end{array}\right.\\
(2,74) & & \\
 & \begin{array}{l}(2.74b)\\ (R_\alpha^{(2)} + R_\sigma\\ \text{restrictions})\end{array} & \left\{\begin{array}{ll}(2.74b\alpha) & \begin{bmatrix}\upsilon(g_r)\cdot\mathrm{A}\cdot\iota'(k_r) & \upsilon(g_r)\cdot\mathrm{A}\cdot\iota'(l_r)\\ \upsilon(h_r)\cdot\mathrm{A}\cdot\iota'(k_r) & \upsilon(h_r)\cdot\mathrm{A}\cdot\iota'(l_r)\end{bmatrix} = 0,\\ & \qquad r = 1, \ldots, R_\alpha^{(2)},\\ (2.74b\sigma) & \upsilon(g_r)\cdot\Sigma\cdot\upsilon'(h_r) = 0,\\ & \qquad r = R_\alpha^{(2)} + 1, \ldots, R_\alpha^{(2)} + R_\sigma.\end{array}\right.
\end{array}
$$

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

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

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 unidentifiability. 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

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

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 $S_1 S_2$ 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 (2.74*b*) 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 $S_3 = S_2 - S_1 S_2$, 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.

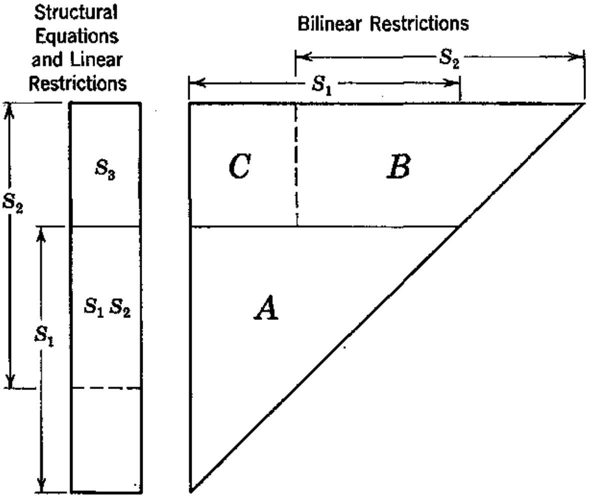

Figure 2.4.6

*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

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

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

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

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:

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

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

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

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

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

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

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

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

so that Υ partitions as follows

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

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 $\Upsilon_{\mathrm{I\ II}}$ 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

$$(2.87) \qquad \mathrm{A} = \begin{bmatrix} \mathrm{A}_{\mathrm{I}1} & \mathrm{A}_{\mathrm{I}2} \\ 0 & \mathrm{A}_{\mathrm{II}2} \end{bmatrix}$$

of the rectangular matrix A, provided $\mathrm{A}_{\mathrm{I}1}$ is square and nonsingular. The case (2.82) in which $\mathrm{A}_{\mathrm{I}1}$ 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 relations 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 critèria 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 mathematical 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

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

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

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

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 Β, Γ, Σ, known as the likelihood function. The maximum-likelihood estimates here considered are those values

$$(3.2) \qquad B, \ C, \ S$$

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

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

and the equation system (1.1) in this notation,

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

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

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

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]

$$\begin{aligned}\frac{1}{T}\log F \equiv L \equiv L(\mathrm{A},\Sigma) = -\frac{1}{2}K_y \log 2\pi + \log\det \mathrm{B} \\ -\frac{1}{2}\log\det\Sigma - \frac{1}{2}\operatorname{tr}\Sigma^{-1}\mathrm{A}M_{xx}\mathrm{A}'.\end{aligned} \tag{3.6}$$

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.

$$(3.7) \qquad M_{xx} = \frac{1}{T}\left[\sum_{t=1}^{T} x_k(t)\ x_l(t)\right] = \frac{1}{T}\sum_{t=1}^{T} x'(t)\ x(t), \qquad k,\ l = 1, \ldots, K_x,$$

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

$$(3.9) \qquad A^{\oplus} = \Upsilon\ A, \qquad \Sigma^{\oplus} = \Upsilon\ \Sigma\ \Upsilon'.$$

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 $\Upsilon = B^{-1}$ in (3.9), which makes

$$\mathrm{B}^{\oplus} = I, \tag{3.10}$$

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 way in what has been called the *reduced form* [Mann and Wald, p. 201, equation (85)] :

$$\begin{gathered} y_i(t) = \sum_{k=K_y+1}^{K_x} \pi_{ik}\, z_{k-K_y}(t) + v_i(t), \qquad i = 1, \ldots, K_y, \\ \text{or} \\ y'(t) - \Pi z'(t) = v'(t), \end{gathered} \tag{3.11}$$

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 $\mathrm{A}^{\oplus}$ in (3.9) corresponding to (3.10), and we shall write Ω for the corresponding value of $\Sigma^{\oplus}$,

$$\begin{gathered} \omega_{kl} = \mathcal{E}\, v_k(t)\cdot v_l(t), \qquad k,\ l = 1, \ldots, K_y, \\ \text{or} \\ \Omega = \mathcal{E}\, v_y'(t)\cdot v_y(t). \end{gathered} \tag{3.12}$$

It follows that Ω, and therefore also Ω^{-1}, are symmetric and positive definite, since B is nonsingular. In this notation, (3.9) becomes

$$[\, I \quad -\Pi\,] = \mathrm{B}^{-1}\mathrm{A}, \qquad \Omega^{-1} = \mathrm{B}'\Sigma^{-1}\mathrm{B}. \tag{3.13}$$

Since

$$\log\det \mathrm{B} - \frac{1}{2}\log\det\Sigma = \frac{1}{2}\log\det(\mathrm{B}'\Sigma^{-1}\mathrm{B}), \tag{3.14}$$

the function L in (3.6) can now be written as

$$\begin{aligned} L = L(-\Pi, \Omega) = -\frac{1}{2}G\log 2\pi + \frac{1}{2}\log\det\Omega^{-1} \\ - \frac{1}{2}\operatorname{tr}\{\Omega^{-1}(M_{yy} - M_{yz}\Pi' - \Pi M_{zy} + \Pi M_{zz}\Pi')\}. \end{aligned} \tag{3.15}$$

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,

$$(3.16) \qquad \frac{d}{d\xi}\operatorname{tr}(XY') = \operatorname{tr}\left(\frac{dX}{d\xi}Y'\right).$$

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

$$(3.17) \qquad \frac{d}{d\xi}\log\det X = \frac{\frac{d}{d\xi}\det X}{\det X} = \frac{\sum\limits_{m,n} X^{mn}\frac{dx_{mn}}{d\xi}}{\det X} = \operatorname{tr} X^{-1}\frac{dX}{d\xi}.$$

An expression for $\frac{dX^{-1}}{d\xi}$ is derived as follows:

$$(3.18) \qquad XX^{-1} = I, \qquad \frac{dX}{d\xi}X^{-1} + X\frac{dX^{-1}}{d\xi} = 0, \qquad \frac{dX^{-1}}{d\xi} = -X^{-1}\frac{dX}{d\xi}X^{-1}.$$

Therefore, if Y is also square, of the same order, and constant,

$$(3.19) \qquad \frac{d}{d\xi}\operatorname{tr}(X^{-1}Y') = -\operatorname{tr}\left(X^{-1}\frac{dX}{d\xi}X^{-1}Y'\right).$$

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

$$\Pi = P + \varepsilon \underset{\to}{P} \tag{3.20}$$

in (3.15), we have

$$\begin{aligned} \frac{dL(-\Pi,\ \Omega)}{d\varepsilon} &= -\frac{1}{2}\operatorname{tr}\{\Omega^{-1}(M_{yz}\underset{\to}{P}' + \underset{\to}{P}M_{zy} - \Pi M_{zz}\underset{\to}{P}' - \underset{\to}{P}M_{zz}\Pi')\} \\ &= -\ \operatorname{tr}\{\Omega^{-1}(M_{yz} - \Pi M_{zz})\underset{\to}{P}'\} \end{aligned} \tag{3.21}$$

in connection with the symmetry[1] of M_{xx} and Ω^{-1}. Furthermore

$$\frac{d^2L}{d\varepsilon^2} = -\ \operatorname{tr}(\Omega^{-1}\underset{\to}{P}M_{zz}\underset{\to}{P}') \tag{3.22}$$

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 $\underset{\to}{P}$. 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 $\underset{\to}{P}$. 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 $\operatorname{tr} XY' = \operatorname{tr} Y'X = \operatorname{tr} YX'$ 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]

$$(3.25) \qquad \frac{d^2L}{d\varepsilon^2} = -\ \mathrm{tr}\{(\Psi \underset{\rightarrow}{P} R)(\Psi \underset{\rightarrow}{P} R)'\}.$$

The reader will have noticed that the elements

$$(3.26) \qquad \hat{\pi}_{ik} = \sum_{l=K_y+1}^{K_z} m_{il}\, m^{lk}_{(zz)}, \qquad \text{with} \qquad [m^{lk}_{(zz)}] = M^{-1}_{zz},$$

$$l,\ k = K_y + 1,\ \ldots,\ K_x,$$

of the ith row of the matrix P represent the coefficients of the *elementary regression* of the dependent variable $y_i(t)$, $1 \leq i \leq K_y$, 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

$$(3.27) \qquad L(\Omega) \equiv -\frac{1}{2}K_y \log 2\pi \ + \ \frac{1}{2}\log \det \Omega^{-1} \ - \ \frac{1}{2}\mathrm{tr}(\Omega^{-1}\cdot {}^{z}M_{yy}),$$

where

$$(3.28) \qquad {}^{z}M_{yy} \equiv M_{yy} \ - \ M_{yz}\cdot M^{-1}_{zz}\cdot M_{zy}$$

is again positive definite, because it is the moment matrix of the "residuals" $v_i^{\oplus}(t) \equiv y_i(t) \ - \ \sum_{l=1}^{K_z} \hat{\pi}_{i,K_y+l}\cdot z_l(t)$, $i = 1, \ldots, K_y$, 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

$$(3.29) \qquad \Omega^{-1} = V \ + \ \eta \underset{\rightarrow}{V}, \qquad V \equiv W^{-1}, \qquad \underset{\rightarrow}{V} = \underset{\rightarrow}{V}',$$

[1] Use is made of the properties $(XY')' = YX'$ and $\mathrm{tr}(XX') = \sum_{m,n}(x_{mn})^2$.

the assumed symmetry of $\underset{\to}{V}$ 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 $\underset{\to}{V}$. For, if ω_1^{-1} and w_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 $|v\,\underset{\to}{V}v'|$, we have $w_1^{-1} > 0$, $\omega_1^{-1} \geq w_1^{-1} - |\eta|\cdot|\,\eta_1| > 0$ for sufficiently small values of $|\eta|$. Using (3.16) and (3.17), we have from (3.27)

$$(3.30) \qquad \frac{dL(\Omega)}{d\eta} = \frac{1}{2}\ \mathrm{tr}(\Omega\ \underset{\to}{V}) - \frac{1}{2}\,\mathrm{tr}(\underset{\to}{V}\cdot{}^{z}M_{yy}) = \frac{1}{2}\ \mathrm{tr}\{(\Omega - {}^{z}M_{yy})\,\underset{\to}{V}\}\ ,$$

and, using (3.19),

$$(3.31) \qquad \left(\frac{d^2 L}{d\eta^2}\right)_{\eta=0} = -\frac{1}{2}\ \mathrm{tr}(W\ \underset{\to}{V}\ W\ \underset{\to}{V}).$$

From (3.30) we see that $L(\Omega)$ has one stationary value, which is reached if Ω equals

$$(3.32) \qquad W \equiv {}^{z}M_{yy}\ .$$

This stationary value is a maximum because the quadratic form (3.31) can be shown to be negative definite in the elements of $\underset{\to}{V}$, by an argument similar to that used in the case of (3.22), and using the symmetry of $\underset{\to}{V}$.

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

$$(3.33) \qquad \Omega^{-1}(\theta) \equiv \theta\ \Omega^{-1}(1)\ +\ (1-\theta)W^{-1}, \qquad 0 \leq \theta \leq 1,$$

is easily shown to belong to the same space. The function

$$(3.34) \qquad \bar{L}(\theta) \equiv L\{\ \Omega^{-1}(\theta)\}$$

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

$$(3.38) \qquad \begin{aligned} \mathrm{B} &= \text{any nonsingular square matrix of order } K_y, \\ \Gamma &= -\,\mathrm{B}P\,, \\ \Sigma &= \mathrm{B}W\mathrm{B}, \end{aligned}$$

of the set of points equivalent to the point

$$(3.39) \qquad \mathrm{B} = I, \quad \Gamma = -P = -M_{yz}M_{zz}^{-1}, \quad \Sigma = W = M_{yy} - M_{yz}M_{zz}^{-1}M_{zy}.$$

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)

$$L = -\frac{1}{2}K\log 2\pi + \frac{1}{2}\log\det\Omega^{-1} - \frac{1}{2}\operatorname{tr}\Big(\Omega^{-1}\{W + (P-\Pi)M_{zz}(P-\Pi)'\}\Big),$$

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:

$$\text{(3.41}lh\text{)} \qquad \beta(g)[-I \quad P\,]\,\Phi_g' = 0,$$

$$g = 1, \ldots, K_y,$$

$$\text{(3.41}ln\text{)} \qquad \beta(g)[-I \quad P\,]\,\iota'(i_g) = 1,$$

$$\text{(3.41}b\alpha\text{)} \qquad \begin{bmatrix} \beta(g_r)[-I \quad P\,]\,\iota'(k_r) & \beta(g_r)[-I \quad P\,]\,\iota'(l_r) \\ \beta(h_r)[-I \quad P\,]\,\iota'(k_r) & \beta(h_r)[-I \quad P\,]\,\iota'(l_r) \end{bmatrix} = 0,$$

$$r = 1, \ldots, R_\alpha^{(2)},$$

$$(3.41b\sigma) \qquad \beta(g_r)\, W\, \beta'(h_r) = 0, \qquad r = R^{(2)}_{\alpha} + 1, \ \ldots, \ R^{(2)}_{\alpha} + R_{\sigma} ,$$

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

$$(3.42) \qquad [-I \quad \Pi\,] = -\mathrm{B}^{-1}[\,\mathrm{B} \quad \Gamma\,] = -\mathrm{B}^{-1}\mathrm{A},$$

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 $\mathrm{B} \neq 0$).

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, $g \doteq 1, \ldots, K_y$, *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

$$(3.43) \qquad \varphi_r(\mathrm{A}, \Sigma) = 0, \qquad r = 1, \ldots, R.$$

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*

$$(3.44) \qquad \Pi = -\mathrm{B}^{-1}\Gamma, \qquad \Omega = \mathrm{B}^{-1}\Sigma\mathrm{B}'^{-1}, \qquad \det\Omega \neq 0,$$

subject to such restrictions,

$$(3.45) \qquad \psi_r(\Pi, \Omega) = 0,$$

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

$$(3.46)\qquad \Pi = P, \qquad \Omega = W,$$

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 $N_{\mathfrak{m}}$ 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

$$(3.47)\qquad \begin{aligned} \Phi_S \cdot \delta\theta_S' + \Phi_{-S} \cdot \delta\theta_{-S}' &= 0, \\ \mathrm{H}_S \cdot \delta\theta_S' + \mathrm{H}_{-S} \cdot \delta\theta_{-S}' &= \delta\eta', \end{aligned}$$

by elimination of $\delta\theta_{-S}$. Here

$$(3.48)\qquad \Phi_S \equiv \begin{bmatrix} \dfrac{\partial\varphi_1}{\partial\theta_1} & \cdots & \dfrac{\partial\varphi_1}{\partial\theta_{P_S}} \\ \cdot & \cdots & \cdot \\ \dfrac{\partial\varphi_R}{\partial\theta_1} & \cdots & \dfrac{\partial\varphi_R}{\partial\theta_{P_S}} \end{bmatrix}, \qquad \Phi_{-S} \equiv \begin{bmatrix} \dfrac{\partial\varphi_1}{\partial\theta_{P_S+1}} & \cdots & \dfrac{\partial\varphi_1}{\partial\theta_P} \\ \cdot & \cdots & \cdot \\ \dfrac{\partial\varphi_R}{\partial\theta_{P_S+1}} & \cdots & \dfrac{\partial\varphi_R}{\partial\theta_P} \end{bmatrix},$$

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

$$(3.49)\qquad \rho\begin{pmatrix} \Phi_S & \Phi_{-S} \\ \mathrm{H}_S & \mathrm{H}_{-S} \end{pmatrix} = \rho(\ \Phi_S \quad \Phi_{-S}\ ) + \rho(\ \mathrm{H}_S \quad \mathrm{H}_{-S}\ ).$$

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

$$(3.54) \qquad \rho\begin{pmatrix} \Phi_S \\ \mathrm{H}_S \end{pmatrix} = P_S ,$$

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

$$(3.55) \qquad \varphi(\theta_S) = 0$$

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

$$(3.56) \qquad [\ \varphi_S \quad 0_{-S}\ ] \equiv \left[\ \frac{\partial\varphi}{\partial\theta_1} \quad \cdots \quad \frac{\partial\varphi}{\partial\theta_p} \quad 0_{P_S+1} \quad \cdots \quad 0_{P_S}\ \right]$$

is added to the matrix $[\ \Phi_S \quad \Phi_{-S}\ ]$, the system (3.47) so enlarged or

$$(3.57) \qquad \begin{aligned} \varphi_S\cdot\delta\theta_S' &= 0, \\ \Phi_S\cdot\delta\theta_S' + \Phi_{-S}\cdot\delta\theta_{-S}' &= 0, \\ \mathrm{H}_S\cdot\delta\theta_S' + \mathrm{H}_{-S}\cdot\delta\theta_{-S}' &= \delta\eta' , \end{aligned}$$

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

$$(3.58) \qquad \rho\begin{pmatrix} \Phi_S \\ \\ H_S \end{pmatrix} = P = \rho\begin{pmatrix} \varphi_S \\ \Phi_S \\ H_S \end{pmatrix}.$$

From (3.52) applied to both members of (3.58) follows

$$(3.59) \qquad \rho\begin{pmatrix} \Phi_S & \Phi_{-S} \\ \\ H_S & H_{-S} \end{pmatrix} = \rho\begin{pmatrix} \varphi_S & 0 \\ \Phi_S & \Phi_{-S} \\ H_S & H_{-S} \end{pmatrix},$$

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)

$$(3.60) \qquad \rho(\ \Phi_S \quad \Phi_{-S}\ ) = \rho\begin{pmatrix} \varphi_S & 0 \\ \\ \Phi_S & \Phi_{-S} \end{pmatrix}.$$

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

$$(3.61) \qquad \rho(\ \Phi_S \quad \Phi_{-S}\ ) = \rho\begin{pmatrix} \varphi_S & 0 \\ \\ \Phi_S & \Phi_{-S} \end{pmatrix} - 1\ .$$

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

$$(3.62) \qquad \rho\begin{pmatrix} \varphi_S & 0 \\ \Phi_S & \Phi_{-S} \\ H_S & H_{-S} \end{pmatrix} = \rho\begin{pmatrix} \varphi_S & 0 \\ \Phi_S & \Phi_{-S} \end{pmatrix} + \rho(\; H_S \quad H_{-S}\;) - 1,$$

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

$$\text{(3.63)} \qquad \mathrm{B} = \begin{bmatrix} \mathrm{B}_{\mathrm{I\,I}} & \mathrm{B}_{\mathrm{I\,II}} \\ 0 & \mathrm{B}_{\mathrm{II\,II}} \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{\mathrm{I\,I}} & 0 \\ 0 & \Sigma_{\mathrm{II\,II}} \end{bmatrix},$$

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

$$\text{(3.64)} \qquad L(\mathrm{A},\ \Sigma) = L_{\mathrm{I}}(\mathrm{A}_{\mathrm{I}},\ \Sigma_{\mathrm{II}}) + L_{\mathrm{II}}(\mathrm{A}_{\mathrm{II}},\ \Sigma_{\mathrm{II\,II}}),$$

where

$$\text{(3.65)} \qquad \begin{aligned} L_{\mathrm{I}} &= -\frac{1}{2}K_{\mathrm{I}} \log 2\pi + \log \det \mathrm{B}_{\mathrm{I\,I}} - \frac{1}{2} \log \det \Sigma_{\mathrm{I\,I}} \\ &\quad - \frac{1}{2} \operatorname{tr}(\Sigma_{\mathrm{I\,I}}^{-1} \cdot \mathrm{A}_{\mathrm{I}} \cdot M_{xx} \cdot \mathrm{A}_{\mathrm{I}}'), \\ L_{\mathrm{II}} &= -\frac{1}{2}K_{\mathrm{II}} \log 2\pi + \log \det \mathrm{B}_{\mathrm{II\,II}} - \frac{1}{2} \log \det \Sigma_{\mathrm{II\,II}} \\ &\quad - \frac{1}{2} \operatorname{tr}(\Sigma_{\mathrm{II\,II}}^{-1} \cdot \mathrm{A}_{\mathrm{II}} \cdot M_{xx} \cdot \mathrm{A}_{\mathrm{II}}') . \end{aligned}$$

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 $M_{y_{\text{I}}x}$ 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*

$$(3.66)\qquad \det\left[\sum_{\tau=0}^{\tau^{\square}} \mathrm{B}(\tau)\ \rho^{\tau^{\square}-\tau}\right] = \det\left[\sum_{\tau=0}^{\tau^{\square}} \beta_{gi\tau}\ \rho^{\tau^{\square}-\tau}\right] = 0, \qquad g,\ i = 1,\ \ldots,\ K_y,$$

satisfy

$$(3.67)\qquad |\rho| < 1.$$

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 $y(t)$, $t \leq 0$, of endogenous variables, with a timing preceding the period $1 \leq t \leq T$ 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

$$(3.70) \qquad F(M_{xx}, A, \Sigma) = (2\pi)^{-\frac{1}{2}\cdot K_y T} \cdot \det{}^{T} B \cdot \det{}^{-\frac{1}{2}\cdot T} \Sigma \cdot \exp\{-\tfrac{1}{2}\mathrm{tr}(\Sigma^{-1} A\, M_{xx}\, A')\}.$$

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*

$$(3.73) \qquad M_{xx} - \mathrm{M}_{xx}$$

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),

$$(3.74) \qquad L(M, \bar{\theta}) = \frac{1}{T} \log F(M, \bar{\theta})$$

as written out in (3.6), is linear in the moment matrix M. Its expected value therefore equals

$$(3.75) \qquad \mathcal{E} L(M, \bar{\theta}) = L(\mathrm{M}, \bar{\theta}),$$

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 $dx_1(1) \quad \cdot \cdot \cdot \quad dx_{K_y}(T)$, 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

$$\varphi(\theta) \equiv \begin{bmatrix} \varphi_1(\theta) & \cdots & \varphi_R(\theta) \end{bmatrix} = 0\,. \tag{3.80}$$

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 $\mathcal{N}$ of the restricted parameter space if it is constant, within $\mathcal{N}$, on any set of mutually equivalent points.*

Let $\eta(\theta) \equiv \begin{bmatrix} \eta_1(\theta) & \cdots & \eta_{P^{\oplus}}(\theta) \end{bmatrix}$ 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* ${}^rN_\theta$ *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* ${}^rN_\theta$, *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*

$$R^{\oplus} \equiv R - R^{\oplus\oplus} \tag{3.84}$$

independent restrictions

$$\psi(\eta) \equiv \begin{bmatrix} \psi_1(\eta) & \cdots & \psi_{R^{\oplus}}(\eta) \end{bmatrix} = 0, \qquad \rho\left(\frac{d\psi}{d\eta'}\right) = R^{\oplus}, \tag{3.85}$$

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,

$$\rho\left(\frac{d\eta}{d\theta'}\right) = P^{\oplus}, \tag{3.86}$$

in any region of the parameter space (excluding, of course, the points with $\det \mathrm{B} = 0$). It follows from (3.81), (3.82), (3.84), and (3.86) that

$$0 \leq R^{\oplus\oplus} \leq R, \qquad \text{so} \qquad 0 \leq R^{\oplus} \leq R. \tag{3.87}$$

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 $\psi(\eta, \varphi) \equiv \begin{bmatrix} \psi_1(\eta, \varphi) & \cdots & \psi_{R^{\oplus}}(\eta, \varphi) \end{bmatrix}$ such that in N_{θ}

$$\psi\{\eta(\theta), \varphi(\theta)\} = 0, \qquad \rho\begin{pmatrix} \dfrac{\partial\psi}{\partial\eta'} \\ \\ \dfrac{\partial\psi}{\partial\theta'} \end{pmatrix} = R^{\oplus}. \tag{3.88}$$

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 $\varphi(\theta)$,

$$\rho\left(\frac{\partial\psi}{\partial\eta'}\right) = \rho\left(\frac{\partial\psi}{\partial\varphi'}\right) = R^{\oplus}, \tag{3.89}$$

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

$$\begin{bmatrix} \dfrac{d\eta}{d\theta'} & \dfrac{d\varphi}{d\theta'} \end{bmatrix} \begin{bmatrix} \dfrac{\partial\psi}{\partial\eta'} \\[2ex] \dfrac{\partial\psi}{\partial\varphi'} \end{bmatrix} = 0 \tag{3.91}$$

obtained from (3.88) by differentiation with respect to θ' would possess a linear combination

$$\begin{bmatrix} \dfrac{d\eta}{d\theta'} & \dfrac{d\varphi}{d\theta'} \end{bmatrix} \begin{bmatrix} \dfrac{\partial\psi}{\partial\eta'} \\[2ex] \dfrac{\partial\psi}{\partial\varphi'} \end{bmatrix} \varkappa' = \frac{d\varphi}{d\theta'}\bar{\varkappa}' = 0\,, \qquad \bar{\varkappa}' \neq 0\,, \tag{3.92}$$

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)

$$\rho\begin{pmatrix} \dfrac{d\eta}{d\theta'}\dfrac{d\psi(\eta)}{d\eta'} & \dfrac{d\varphi}{d\theta'} \end{pmatrix} = \rho\begin{pmatrix} -\dfrac{d\varphi}{d\theta'}\dfrac{\partial\psi(\eta,\varphi)}{\partial\varphi'} & \dfrac{d\varphi}{d\theta'} \end{pmatrix} = \rho\begin{pmatrix} \dfrac{d\varphi}{d\theta'} \end{pmatrix}. \tag{3.94}$$

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*

$$\rho(\,\Xi \quad \Phi\,) = \rho(\Xi) + \rho(\Phi), \qquad \rho(\Xi) = c(\Xi), \qquad \rho(\Phi) = c(\Phi), \tag{3.99}$$

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 $\eta(\theta)$:

$$(3.104) \qquad \zeta(\theta) = \xi\{\eta(\theta)\} \qquad \text{whenever} \qquad \varphi(\theta) = 0 .$$

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 $\varphi(\theta) = 0$,

$$(3.105) \qquad \rho\left(\left(\frac{d\zeta}{d\theta'} - \frac{d\eta}{d\theta'}\frac{d\xi}{d\eta'}\right) \qquad \frac{d\varphi}{d\theta'}\right) = \rho\left(\frac{d\varphi}{d\theta'}\right) = R .$$

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

$$(3.106) \qquad \rho\left(\frac{d\eta}{d\theta'}\frac{d\xi}{d\eta'} \qquad \frac{d\varphi}{d\theta'}\right) = Q + R .$$

Since this is the maximum possible rank for a matrix of $Q + R$ columns, we also have in ${}^{r}N_{\theta}$

$$(3.107) \qquad \rho\left(\frac{d\eta}{d\theta'}\frac{d\xi}{d\eta'}\right) = Q .$$

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 ${}^{r}N_{\theta}$

$$(3.108)\qquad \rho\left(\frac{d\eta}{d\theta'}\begin{bmatrix}\dfrac{d\xi}{d\eta'} & \dfrac{d\psi}{d\eta'}\end{bmatrix}\right) = \rho\begin{pmatrix}\dfrac{d\eta}{d\theta'}\dfrac{d\xi}{d\eta'} & \dfrac{d\eta}{d\theta'}\dfrac{d\psi}{d\eta'}\end{pmatrix} = Q + R^{\oplus}.$$

Since the rank of a matrix product does not exceed the rank of either of the two factors, we must have

$$(3.109)\qquad \rho\begin{pmatrix}\dfrac{d\xi}{d\eta'} & \dfrac{d\psi}{d\eta'}\end{pmatrix} \geq Q + R^{\oplus}$$

in the point set ${}^{r}N_{\eta}$ on which ${}^{r}N_{\theta}$ 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 ${}^{r}N_{\eta}$

$$(3.110)\qquad \rho\begin{pmatrix}\dfrac{d\xi}{d\eta'} & \dfrac{d\psi}{d\eta'}\end{pmatrix} = c\begin{pmatrix}\dfrac{d\xi}{d\eta'} & \dfrac{d\psi}{d\eta'}\end{pmatrix} = Q + R^{\oplus},$$

and, because of the continuity of the functions involved, (3.110) holds also in a neighborhood N_{η}, of the point set ${}^{r}N_{\eta}$, 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 $\chi(\eta) \equiv \begin{bmatrix}\chi_1(\eta) & \cdots & \chi_{S^{\oplus}}(\eta)\end{bmatrix}$ in N_{η} having $S^{\oplus}$ elements, with continuous third derivatives, such that in N_{η}

$$(3.112)\qquad \rho\begin{pmatrix}\dfrac{d\xi}{d\eta'} & \dfrac{d\psi}{d\eta'} & \dfrac{d\chi}{d\eta'}\end{pmatrix} = Q + R^{\oplus} + S^{\oplus} = P^{\oplus} .$$

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 $L(M, \bar{\theta})$ reaches its unrestricted absolute maximum in the true parameter point θ. Since $\eta = \bar{\eta}(\theta)$, it follows that

$$\Lambda^{\oplus} \equiv \left(\frac{\partial^2 L^{\oplus}(M, \bar{\eta})}{\partial\bar{\eta}'\,\partial\bar{\eta}}\right)_{\bar{\eta}=\eta} \tag{3.121}$$

is negative definite. We shall now prove that in consequence

$$\Lambda^{\oplus\oplus} \equiv \left(\frac{\partial^2 L^{\oplus\oplus}(M, \bar{\zeta})}{\partial\bar{\zeta}'\,\partial\bar{\zeta}}\right)_{\bar{\zeta}=\zeta} \tag{3.122}$$

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

$$(3.123)\qquad \bar{\eta} \equiv \bar{\eta}\{\bar{\zeta},\ 0,\ \hat{\chi}(\mathrm{M},\ \bar{\zeta})\} \equiv \eta(\bar{\zeta})\,,$$

say, of $\bar{\zeta}$ alone, which possesses continuous first and second derivatives, the first being

$$(3.124)\qquad \left(\frac{\partial\bar{\eta}}{\partial\bar{\zeta}'} + \frac{\partial\hat{\chi}}{\partial\bar{\zeta}'}\frac{\partial\bar{\eta}}{\partial\bar{\chi}'}\right)_{\bar{\psi}=0,\ \bar{\chi}=\bar{\chi}(\mathrm{M},\bar{\zeta})} \equiv \frac{d\bar{\eta}}{d\bar{\zeta}'}\,,$$

say. In particular, owing to Theorem 3.5.5,

$$(3.125)\qquad \bar{\eta}(\zeta) = \eta\,.$$

Differentiating (3.118) with respect to $\bar{\zeta}$, after substituting M for *M*, we have,

$$(3.126)\qquad \frac{\partial L^{\oplus\oplus}(\mathrm{M},\ \bar{\zeta})}{\partial\bar{\zeta}'} = \frac{d\bar{\eta}}{d\bar{\zeta}'}\left(\frac{\partial L^{\oplus}(\mathrm{M},\ \bar{\eta})}{\partial\bar{\eta}'}\right)_{\bar{\eta}=\bar{\eta}(\bar{\zeta})}\,.$$

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),

$$(3.127)\qquad \Lambda^{\oplus\oplus} = \mathrm{H}\,\Lambda^{\oplus}\,\mathrm{H}'\,,$$

where

$$(3.128)\qquad \mathrm{H} \equiv \left(\frac{d\bar{\eta}}{d\bar{\zeta}'}\right)_{\bar{\zeta}=\zeta}\,.$$

From (3.127) we conclude that the quadratic form

$$(3.129)\qquad z\,\Lambda^{\oplus\oplus}\,z' = z\,\mathrm{H}\,\Lambda^{\oplus}\,\mathrm{H}'\,z' = y\,\Lambda^{\oplus}\,y'\,,$$

say, is equal to a negative definite quadratic form, in which, owing to (3.112), $y = z\,\mathrm{H}$ 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

$$\begin{aligned} 0 = \overset{*}{l}{}^{\oplus\oplus}(M, \hat{\zeta}) - l^{\oplus\oplus}(\mathrm{M}, \zeta) = &\left[\frac{\partial}{\partial\bar{\zeta}} \operatorname{tr}\left\{(M - \mathrm{M}) \frac{\partial L^{\oplus\oplus}(\mathrm{M}, \bar{\zeta})}{\partial \mathrm{M}'}\right\}\right]_{\bar{\zeta}=\zeta} \\ &+ (\hat{\zeta} - \zeta)\Lambda^{\oplus\oplus} + \cdots \end{aligned} \tag{3.131}$$

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 $T^{-½}$ 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

$$(3.133) \qquad [\bar{\zeta} \quad \bar{\chi}] \equiv \bar{\omega},$$

we define, analogously to (3.114),

$$(3.134) \qquad F(M, \bar{\theta}) = F^{\oplus}(M, \bar{\eta})$$

and, somewhat differently from (3.118),

$$(3.135) \qquad F^{\oplus}\{M, \bar{\eta}(\bar{\zeta}, 0, \bar{\chi})\} \equiv \tilde{F}(M, \bar{\omega}) ,$$

with similar formulae in terms of $L = (1/T)\log F$. Finally, we define

$$(3.136) \qquad \tilde{l}(M, \bar{\omega}) \equiv \frac{\partial}{\partial \bar{\omega}} \tilde{L}(M, \bar{\omega}) .$$

Then, in the point $\bar{\omega} = \omega \equiv [\zeta \quad \chi]$,

$$\begin{aligned}
\mathcal{E}\, \tilde{l}'(M,\omega)\, \tilde{l}(M,\omega) &= T^{-2} \int \tilde{F}(M,\omega) \frac{\partial \log \tilde{F}(M,\omega)}{\partial \omega'} \frac{\partial \log \tilde{F}(M,\omega)}{\partial \omega}\, dx \\
&= T^{-2} \int \frac{\partial \tilde{F}(M,\omega)}{\partial \omega'} \frac{\partial \log \tilde{F}(M,\omega)}{\partial \omega}\, dx \\
&= T^{-2} \int \left[\frac{\partial}{\partial \omega'} \left\{ F(M,\omega) \frac{\partial \log \tilde{F}(M,\omega)}{\partial \omega} \right\} \right. \\
(3.137) \qquad &\qquad \left. - \tilde{F}(M,\omega) \frac{\partial^2 \log \tilde{F}(M,\omega)}{\partial \omega'\, \partial \omega} \right] dx \\
&= T^{-2} \frac{\partial}{\partial \omega'} \int \tilde{F}(M,\omega) \frac{\partial \log \tilde{F}(M,\omega)}{\partial \omega}\, dx \\
&\qquad - T^{-2} \int \tilde{F}(M,\omega) \frac{\partial^2 \log \tilde{F}(M,\omega)}{\partial \omega'\, \partial \omega} dx \\
&= 0 - T^{-1} \mathcal{E} \frac{\partial^2 \tilde{L}(M,\omega)}{\partial \omega'\, \partial \omega}
\end{aligned}$$

$$= -T^{-1}\frac{\partial^2 L(\mathrm{M},\omega)}{\partial\omega'\,\partial\omega} \equiv -T^{-1}\tilde{\Lambda},$$

say.

On the other hand, we have a Taylor expansion

$$(3.138)\qquad \begin{aligned} -\tilde{l}(M,\omega) &= \tilde{l}(M,\hat{\omega}) - \tilde{l}(M,\omega) \\ &= (\hat{\omega}-\omega)\frac{\partial^2 L(M,\omega)}{\partial\omega'\,\partial\omega} + \cdots, \end{aligned}$$

where the omitted term is of order $T^{-\frac{1}{2}}$ 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

$$(3.139)\qquad \frac{\partial^2 \tilde{L}(M,\omega)}{\partial\omega'\,\partial\omega} \equiv \tilde{L},$$

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,

$$(3.140)\qquad \hat{\omega}-\omega = -\tilde{l}(M,\omega)\tilde{L}^{-1} + \cdots,$$

and from (3.137)

$$(3.141)\qquad \mathcal{E}(\hat{\omega}'-\omega')(\hat{\omega}-\omega) = -T^{-1}\tilde{L}^{-1}\tilde{\Lambda}\tilde{L}^{-1} + \cdots,$$

in which the omitted term is of order $T^{-\frac{1}{2}}$ relative to the term shown.

Owing to Theorem 3.3.4, property a), and the continuity of the relevant derivatives of the likelihood function,

$$(3.142)\qquad \lim_{T\to\infty}\tilde{\Lambda} \equiv \tilde{\Lambda}_{\infty}$$

exists and is finite. Furthermore, because of property b) of the same theorem,

$$(3.143) \qquad \operatorname*{plim}_{T\to\infty} \tilde{L} = \tilde{\Lambda}_{\infty}.$$

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

$$(3.144) \qquad \tilde{\Lambda} \equiv \begin{bmatrix} \tilde{\Lambda}_{\zeta\zeta} & \tilde{\Lambda}_{\zeta\chi} \\ \tilde{\Lambda}_{\chi\zeta} & \tilde{\Lambda}_{\chi\chi} \end{bmatrix}.$$

Our problem then is to evaluate

$$(3.145) \qquad \operatorname*{plim}_{T\to\infty} T\,\mathcal{E}(\hat{\zeta}' - \zeta')(\hat{\zeta} - \zeta) = -(\tilde{\Lambda}_{\infty}^{-1})_{\zeta\zeta}$$

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]

$$(3.146) \qquad (\tilde{\Lambda}^{-1})_{\zeta\zeta} = (\tilde{\Lambda}_{\zeta\zeta} - \tilde{\Lambda}_{\zeta\chi}\tilde{\Lambda}_{\chi\chi}^{-1}\tilde{\Lambda}_{\chi\zeta})^{-1}.$$

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

$$(3.147) \qquad \hat{\chi}(\mathrm{M},\zeta) = \chi.$$

At present, we must also use the property that, owing to the definition of $\hat{\chi}(M,\bar{\zeta})$,

$$(3.148) \qquad \left(\frac{\partial \tilde{L}(\mathrm{M},\bar{\omega})}{\partial \bar{\chi}'}\right)_{\bar{\chi} = \hat{\chi}(\mathrm{M},\bar{\zeta})} = 0.$$

We differentiate (3.148) with respect to $\bar{\zeta}$,

$$\left(\frac{\partial^2 \tilde{L}(\text{M}, \omega)}{\partial \bar{\chi}' \, \partial \bar{\zeta}} + \frac{\partial^2 \tilde{L}(\text{M}, \bar{\omega})}{\partial \bar{\chi}' \, \partial \bar{\chi}} \, \frac{\partial \hat{\chi}(\text{M}, \bar{\zeta})}{\partial \bar{\zeta}}\right)_{\bar{\chi} = \hat{\chi}(\text{M}, \bar{\zeta})} = 0 , \tag{3.149}$$

substitute $\bar{\zeta} = \zeta$ using (3.147), and solve as follows:

$$\frac{\partial \hat{\chi}'(\text{M}, \zeta)}{\partial \zeta} = - \tilde{\Lambda}_{\chi\chi}^{-1} \tilde{\Lambda}_{\chi\zeta} , \tag{3.150}$$

using the nonsingularity of $\tilde{\Lambda}_{\chi\chi}$ which follows from the negative definiteness of $\Lambda^{\oplus}$.

On the other hand, if we write

$$\tilde{L}(\text{M}, \bar{\omega}) \equiv \tilde{L}(\text{M}, \bar{\zeta}, \bar{\chi}) , \tag{3.151}$$

we have, from a comparison with (3.118),

$$L^{\oplus\oplus}(\text{M}, \zeta) = \tilde{L}\{\text{M}, \bar{\zeta}, \hat{\chi}(\text{M}, \bar{\zeta})\} , \tag{3.152}$$

and hence, using (3.148),

$$\frac{\partial L^{\oplus\oplus}(\text{M}, \bar{\zeta})}{\partial \bar{\zeta}} = \left(\frac{\partial \tilde{L}(\text{M}, \bar{\zeta}, \bar{\chi})}{\partial \bar{\zeta}}\right)_{\bar{\chi} = \hat{\chi}(\text{M}, \bar{\zeta})} . \tag{3.153}$$

Differentiating once more with respect to $\bar{\zeta}$ and substituting $\bar{\zeta} = \zeta$, we obtain, using (3.150),

$$\begin{aligned} \Lambda^{\oplus\oplus} = \frac{\partial^2 L^{\oplus\oplus}(\text{M}, \zeta)}{\partial \zeta' \, \partial \zeta} &= \frac{\partial^2 \tilde{L}(\text{M}, \zeta, \chi)}{\partial \zeta' \, \partial \zeta} + \frac{\partial^2 \tilde{L}(\text{M}, \zeta, \chi)}{\partial \zeta' \, \partial \bar{\chi}} \, \frac{\partial \hat{\chi}'(\text{M}, \zeta)}{\partial \zeta} \\ &= \tilde{\Lambda}_{\zeta\zeta} - \tilde{\Lambda}_{\zeta\chi} \tilde{\Lambda}_{\chi\chi}^{-1} \tilde{\Lambda}_{\chi\zeta} . \end{aligned} \tag{3.154}$$

Comparison with (3.145) and (3.146) now yields

$$\underset{T \to \infty}{\text{plim}}\, T\mathcal{E}\,(\hat{\zeta}' - \zeta')(\hat{\zeta} - \zeta) = -\,(\Lambda_{\infty}^{\oplus\oplus})^{-1}. \tag{3.155}$$

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*

$$\text{est}\; T\,\mathcal{E}\,(\hat{\zeta}' - \zeta')(\hat{\zeta} - \zeta) = -\left(\frac{\partial^2 L^{\oplus\oplus}(M, \bar{\zeta})}{\partial\bar{\zeta}'\,\partial\zeta}\right)_{\bar{\zeta} = \hat{\zeta}} \tag{3.156}$$

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 mention 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 Ⓐ 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

$$(4.3)\qquad \begin{aligned} -\frac{1}{2}\,\alpha(g)\cdot M_{xx}\cdot\alpha'(g) = -\frac{1}{2}\{\beta(g)\cdot M_{yy}\cdot\beta'(g) &+ 2\,\beta(g)\cdot M_{yz}\cdot\gamma'(g) \\ &+ \gamma(g)\cdot M_{zz}\cdot\gamma'(g)\}\,, \end{aligned}$$

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

$$(4.6)\qquad L^*(\mathrm{B}) = \text{const} + \log\det\mathrm{B} - \frac{1}{2}\sum_{g=1}^{G}\beta^g\cdot {}^{z}M^g_{yy}\cdot\beta'^g\,,$$

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]

$$(4.9)\qquad d_{0,1}\cdot \mathrm{cof}\ b^{1}_{0,1} - b^{1}_{0,1}\cdot {}^{z}M^{1}_{yy} = 0\ ,$$

where $\mathrm{cof}\ b^{g}_{0,1}$ 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

$$(4.10)\qquad d_{0,1} \equiv \det^{-1} B_{0,1} = (b^{g}_{0,1}\cdot \mathrm{cof}'\ b^{g}_{0,1})^{-1}\ ,$$

according to the Laplace expansion of $\det B_{0,1}$. Since the elements of $\mathrm{cof}\ b^{g}_{0,1}$ 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

$$(4.11) \qquad b^1_{0,1} = d_{0,1}\cdot \mathrm{cof}\ b^1_{0,1}\cdot({}^zM^1_{yy})^{-1}.$$

The one remaining unknown $d_{0,1}$ is found, from (4.10) and (4.11), to satisfy

$$(4.12) \qquad (d_{0,1})^2 = \{\mathrm{cof}\ b^1_{0,1}\cdot({}^zM^1_{yy})^{-1}\cdot \mathrm{cof}'\ b^1_{0,1}\}^{-1},$$

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

$$(4.13) \qquad B_{0,G} = B_1$$

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,

$$(4.14) \qquad b_1(g) = b_{0,g}(g),$$

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

$$(4.15) \qquad b_1^g \cdot {}^zM_{yy}^g = \iota(g) \cdot B_0'^{-1} \cdot \Phi^{g\prime}.$$

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]

$$(4.16) \qquad \alpha(g)\,\Phi_g' = 0, \qquad \rho(\Phi_g) = r(\Phi_g) = R_g, \qquad g = 1, \ldots, G.$$

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

$$(4.17) \qquad \Phi^g\,\Phi_g' = 0, \qquad \rho(\Phi^g) = r(\Phi^g) = Q_g = K_z - R_g.$$

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

$$(4.18) \qquad \Phi(g) \equiv \begin{bmatrix} \Phi^g \\ \Phi_g \end{bmatrix} \equiv [\mathrm{X}(g) \qquad \Psi(g)],$$

[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

$$(4.19) \qquad \theta(g) \equiv \theta(*g)\ \Phi(g) \equiv \theta^g\ \Phi^g + \theta_g\ \Phi_g$$

establishes a one-to-one correspondence between the space of the vector $\theta(g)$ and that of the vector

$$(4.20) \qquad \theta(*g) \equiv [\ \theta^g \quad \theta_g\ ]$$

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

$$(4.21) \qquad \alpha_g\ \Phi_g\ \Phi_g' = 0$$

and hence, from the rank condition in (4.16), that $\alpha_g = 0$. Thus $\alpha(g)$ is expressible as

$$(4.22) \qquad \alpha(g) = \alpha^g\ \Phi^g.$$

Conversely, any vector so expressible satisfies the restrictions (4.16). The components of the G vectors

$$(4.23) \qquad \alpha^g, \qquad g = 1,\ \ldots,\ G,$$

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

$$\theta \equiv \text{vec } \Theta \equiv [\ \theta(1) \ \ \theta(2) \ \cdots \ \theta(K_y)\ ] . \tag{4.28}$$

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

$$(4.30) \qquad \mathrm{tr}\, \Theta\, \mathrm{H}' = \theta \eta', \qquad \mathrm{vec}(\Theta\, M_{xx}) = \theta M, \qquad \mathrm{tr}\, \Theta\, M_{xx}\, \Theta' = \theta M \theta',$$

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

$$(4.32) \qquad \Phi^* \Phi_*' = 0 .$$

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

$$(4.33) \qquad \theta = \theta(*)\,\Phi(*) = \theta^* \Phi^* + \theta_* \Phi_* ,$$

where

$$(4.34) \qquad \theta(*) - [\theta^* \; \theta_*], \qquad \begin{cases} \theta^* = [\,\theta^1 \;\; \theta^2 \;\; \cdots \;\; \theta^G\,], \\ \theta_* = [\,\theta_1 \;\; \theta_2 \;\; \cdots \;\; \theta_G\,]. \end{cases}$$

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

$$(4.35) \qquad \alpha \Phi_*' = 0,$$

we must have

$$(4.36) \qquad \alpha^* = [\,\alpha^1 \;\; \alpha^2 \;\; \cdots \;\; \alpha^G\,], \qquad \alpha_* = 0 .$$

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

$$(4.37) \qquad \alpha = \alpha^* \Phi^* .$$

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

$$(4.38) \qquad \theta(*) \equiv \mathrm{vec}*\Theta, \qquad \theta^* \equiv \mathrm{vec}^*\Theta, \qquad \theta_* \equiv \mathrm{vec}_*\Theta .$$

In these terms (4.33) runs

$$(4.39) \qquad \theta = \mathrm{vec}\ \Theta = \left(\mathrm{vec}^*\ \Theta\right)\Phi^* + \left(\mathrm{vec}_*\ \Theta\right)\Phi_* ,$$

from which we can solve for vec* Θ through postmultiplication by $\Phi'^*(\Phi^*\ \Phi'^*)^{-1}$ using (4.32), thus obtaining

$$(4.40) \qquad \mathrm{vec}^*\Theta = \left(\mathrm{vec}\ \Theta\right)\Phi'^*(\Phi^*\ \Phi'^*)^{-1} .$$

Conversely, we define

$$(4.41) \qquad \Theta \equiv \mathrm{mat}\ \theta \equiv \mathrm{mat}*\theta(*) \equiv \mathrm{mat}*\mathrm{vec}*\Theta ,$$

and

$$(4.42) \qquad \mathrm{mat}*\theta^* \equiv \mathrm{mat}*[\theta^* \quad 0_*], \qquad \mathrm{mat}*\theta_* \equiv \mathrm{mat}*[0^* \quad \theta_*] .$$

The decomposition (4.39) can thus be written in matrix coordinates as

$$(4.43) \qquad \Theta = \mathcal{Q}\Theta + \mathcal{R}\Theta ,$$

where

$$(4.44) \qquad \mathcal{Q}\Theta \equiv \mathrm{mat}*\ \mathrm{vec}^*\Theta , \qquad \mathcal{R}\Theta \equiv \mathrm{mat}*\ \mathrm{vec}_*\ \Theta .$$

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

$$(4.45) \qquad A = \mathcal{Q}A , \quad \text{or} \quad \mathcal{R}A = 0 .$$

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

$$(4.46) \qquad \operatorname{tr} \Theta \mathrm{H}' = \theta \eta' = (\theta^* \Phi^* + \theta_* \Phi_*) \Phi'^* \eta'^* = \theta^* \Phi^* \Phi'^* \eta'^*$$

because of (4.32) does not depend on θ_*. In particular we have:

LEMMA 4.2.4. *A necessary and sufficient condition that* $\operatorname{tr} \Theta \mathrm{H}' = 0$ *for all values of* H *satisfying the a priori restrictions* $\eta_* = \operatorname{vec}_* \mathrm{H} = 0$ *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:

$$(4.47) \qquad L(\mathrm{A}) = \text{const} + \log \det \mathrm{B} - \frac{1}{2} \operatorname{tr}(\mathrm{A}\, M_{xx}\, \mathrm{A}') .$$

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)

$$(4.48) \qquad \mathrm{A} = A_0 + \delta A_0 , \qquad \operatorname{vec}_* \delta A_0 = 0,$$

which insures that A again satisfies the restrictions. The Taylor expansion of $L(\mathrm{A})$ 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),

$$(4.49) \qquad \begin{aligned} L(\mathrm{A}) &= L(A_0) + \operatorname{tr}\{B_0'^{-1} (\delta B_0)'\} - \operatorname{tr}\{A_0\, M_{xx}\, (\delta A_0)'\} + \cdots \\ &= L(A_0) + \operatorname{tr}\{ (B_0'^{-1}\, I_{[K_y\, K_x]} - A_0\, M_{xx})(\delta A_0)' \} + \cdots . \end{aligned}$$

(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

$$(4.50)\quad \begin{cases} (4.50q) \qquad \operatorname{vec}^* (B'^{-1}\, I_{[K_y\,K_x]} - A\, M_{xx}) = 0, \\ (4.50r) \qquad a_* \equiv \operatorname{vec}_* A = 0. \end{cases}$$

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 $\mathfrak{Q}$ 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

$$(4.51)\quad \begin{cases} (4.51y) \qquad B'^{-1} - A\, M_{xy} = 0, \\ (4.51z) \qquad A\, M_{xz} = 0. \end{cases}$$

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 $\mathbb{P}_1$, $\mathbb{P}_h$, and $\mathbb{P}_{h_n}$.

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:

$$(4.57)\left\{\begin{array}{ll} (4.57^*) & \operatorname{vec}^*(\Delta A_0 \cdot M_{xx}) = \operatorname{vec}^*(B_0'^{-1} I_{[K_y\, K_x]} - A_0 M_{xx}), \\ (4.57_*) & \Delta a_{0*} = \operatorname{vec}_* \Delta A_0 = 0\,. \end{array}\right.$$

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 $\mathbb{P}_1$. 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 $\mathbb{P}_1$, is equivalent to the process demonstrated in a special case in section *4.1*. For that choice of h, A_1 is defined by

$$(4.60)\quad\begin{cases}(4.60^*) & \operatorname{vec}^*(A_1\, M_{xx}) = \operatorname{vec}^*[B_0'^{-1} \quad 0]\,,\\ (4.60_*) & a_{1*} = \operatorname{vec}_* A_1 = 0\,.\end{cases}$$

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:

$$(4.61)\quad\begin{cases}(4.61\mathcal{Q}) & \mathcal{Q}(A_1\, M_{xx}) = \mathcal{Q}[B_0'^{-1} \quad 0],\\ (4.61\mathfrak{R}) & \mathfrak{R}A_1 = 0.\end{cases}$$

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$\mathfrak{R}$). In this case, therefore, if the partitioning of Θ corresponding to $\mathrm{A} = [\mathrm{B} \quad \Gamma]$ is denoted momentarily by $\Theta = [\mathrm{H} \quad \mathrm{Z}]$, the definition of $\mathcal{Q}$ can be extended to submatrices of Θ through

$$(4.62)\qquad \mathcal{Q}\Theta \equiv [\mathcal{Q}\mathrm{H} \quad \mathcal{Q}\mathrm{Z}],$$

and (4.61$\mathcal{Q}$) partitions into

$$(4.63\mathcal{Q})\quad\begin{cases}(4.63\mathcal{Q}\mathrm{y}) & \mathcal{Q}(A_1\, M_{xy}) = \mathcal{Q}B_0'^{-1},\\ (4.63\mathcal{Q}\mathrm{z}) & \mathcal{Q}(A_1\, M_{xz}) = 0.\end{cases}$$

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 $\mathbb{P}_1$ 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 $\mathbb{P}_{1/2}$. In a special borderline case the choice $h = \frac{1}{2}$ leading to the process $\mathcal{P}_{\frac{1}{2}}$ 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 $\mathcal{P}_{\frac{1}{2}}$ now runs

$$(4.64) \qquad A_1 = \frac{1}{2}(A_0'^{-1}\cdot M_{xx}^{-1} + A_0).$$

Simple calculations will show that this process possesses the property

$$(4.65) \qquad \begin{aligned} &A_1\, M_{xx}\, A_1' - I = \\ &\frac{1}{4}(A_0\, M_{xx}\, A_0' - I)(A_0\, M_{xx}\, A_0')^{-1}(A_0\, M_{xx}\, A_0' - I). \end{aligned}$$

Since under the present assumptions the maximum-likelihood equations (4.51) to be solved iteratively are equivalent to

$$(4.66) \qquad A\, M_{xx}\, A' = I,$$

(4.65) implies a high rate of convergence of $\mathcal{P}_{\frac{1}{2}}$, 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 $\mathcal{P}_{\frac{1}{2}}$ possesses this very desirable property, a process $\mathcal{P}_h$ 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,

$$(4.67) \qquad a_1 = \frac{1}{2}(\frac{m^{-1}}{a_0} + a_0),$$

to obtain the square root of a scalar m^{-1}.

4.3.2.4... The process $\mathbb{P}_{h_n}$. 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

$$\begin{aligned} L(A_1) - L(A_0) &= h \operatorname{tr}\{(B_0'^{-1} I_{[K_y\, K_x]} - A_0 M_{xx})(\Delta A_0)'\} \\ (4.68) \qquad &\quad + \frac{1}{2}h^2 \operatorname{tr}\{-B_0'^{-1}(\Delta B_0)' B_0'^{-1}(\Delta B_0)' \\ &\quad - (\Delta A_0) M_{xx} (\Delta A_0)'\} + \cdots , \end{aligned}$$

and using, as in (4.58), the definition of ΔA_0 given by (4.57), we find that the value,

$$(4.69) \qquad h_0 = \frac{\operatorname{tr} \Delta A_0 \cdot M_{xx} \cdot (\Delta A_0)'}{\operatorname{tr}\left(\{B_0'^{-1}(\Delta B_0)'\}^2 + \Delta A_0 \cdot M_{xx} \cdot (\Delta A_0)'\right)} ,$$

of h satsifies the criterion mentioned. This procedure, which we denote by $\mathbb{P}_{h_n}$, 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 $\mathbb{P}_1$, $\mathbb{P}_h$, $\mathbb{P}_{h_n}$

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

$$(4.70) \qquad \Phi(*)\,\Phi'(*) = I, \qquad \text{so} \qquad \Phi^{-1}(*) = \Phi'(*)\,.$$

The relationships (4.30) can then be extended in a simple way to the θ-space. We register for future use:

$$(4.71) \qquad \begin{aligned} \operatorname{tr}(\Theta\,\mathrm{H}') &= \theta(*)\,\eta'(*)\,, \qquad \operatorname{vec}*(\Theta\,M_{xx}) = \theta(*)\,M(*)\,,\\ \operatorname{tr}(\Theta\,M_{xx}\,\Theta') &= \theta(*)\,M(*)\,\theta'(*)\,, \end{aligned}$$

where

$$(4.72) \qquad M(*) = \Phi(*)\,M\,\Phi'(*) = M'(*) = \begin{bmatrix} M^{**} & M^{*}{}_{*} \\ M_{*}{}^{*} & M_{**} \end{bmatrix}.$$

In particular we have, if H satisfies the restrictions $\eta_* = 0$, the important identities

$$(4.73) \qquad \begin{aligned} \operatorname{tr}\,\Theta\,\mathrm{H}' &= \theta^*\,\eta'^*\\ \operatorname{vec}^*(\mathrm{H}\,M_{xx}) &= \eta^*\,M^{**}\,,\\ \operatorname{tr}(\mathrm{H}\,M_{xx}\,\mathrm{H}') &= \eta^*\,M^{**}\,\eta'^*\,. \end{aligned}$$

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

$$(4.74) \qquad A_n = A + \bar{A}_n\,, \qquad \bar{a}_{n*} = \operatorname{vec}_* \bar{A}_n = 0, \qquad n = 0,\ 1,\ \ldots\ .$$

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

$$(4.75) \qquad \bar{A}_1 = \bar{A}_0 + h(\Delta\bar{A}_0)\,,$$

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),

$$(4.76)\qquad \text{vec}^*\,(\Delta\bar{A}_0 M_{xx}) = \text{vec}^*\,(-B'^{-1}\bar{B}_0'\,B'^{-1}I_{[K_yK_x]} - \bar{A}_0\,M_{xx}) + \cdots .$$

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

$$(4.77)\qquad \begin{aligned} \frac{1}{2}\,L_{(2)}(A_0) &\equiv \frac{1}{2}\,\text{tr}(-B'^{-1}\bar{B}_0'\,B'^{-1}\,\bar{B}_0' - \bar{A}_0\,M_{xx}\,\bar{A}_0') \\ &\equiv \frac{1}{2}\,\bar{a}_0^*\,L^*\,\bar{a}_0'^*, \end{aligned}$$

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

$$(4.78)\qquad \begin{aligned} \frac{1}{2}\,\frac{dL_{(2)}}{d\bar{a}_0^*} &= \frac{\frac{1}{2}\,\text{tr}\left(\frac{dL_{(2)}}{d\bar{A}_0}\,d\bar{A}_0'\right)}{d\bar{a}_0^*} \\ &= \frac{\text{tr}\{(-B'^{-1}\,\bar{B}_0'\,B'^{-1}\,I_{[K_yK_x]} - \bar{A}_0\,M_{xx})\,d\bar{A}_0'\}}{d\bar{a}_0^*} \\ &= \frac{\text{vec}^*\,(-B'^{-1}\,\bar{B}_0'\,B'^{-1}I_{[K_yK_x]} - \bar{A}_0\,M_{xx})\,d\bar{a}_0^*}{d\bar{a}_0^*} \\ &= \text{vec}^*\,(-B'^{-1}\,\bar{B}_0'\,B'^{-1}\,I_{[K_yK_x]} - \bar{A}_0\,M_{xx}). \end{aligned}$$

Comparing this with the result of differentiating the last member of (4.77), we find that (4.76) is equivalent to

$$(4.79) \qquad \Delta\bar{a}_0^* M^{**} = \bar{a}_0^* L^* + \cdots ,$$

in which the first member is obtained through the second relation (4.73). Through one further transformation

$$(4.80) \qquad \begin{aligned} M^{**} &\equiv R^* R'^*, & L^\dagger &\equiv (R^*)^{-1} L^* (R^*)'^{-1}, \\ \Delta a_0^\dagger &\equiv \Delta\bar{a}_0^* R^*, & \bar{a}_0^\dagger &\equiv \bar{a}_0^* R^*, \end{aligned}$$

this goes over into

$$(4.81) \qquad \Delta\bar{a}_0^\dagger = \bar{a}_0^\dagger L^\dagger + \cdots ,$$

so that

$$(4.82) \qquad \bar{a}_1^\dagger = \bar{a}_0^\dagger(I + hL^\dagger) + \cdots .$$

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

$$(4.83) \qquad K^\dagger \equiv I + hL^\dagger .$$

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

$$(4.84) \qquad k_q = 1 + hl_q .$$

We are free to arrange the values (4.84) in descending order of algebraic magnitude[1] through suitable choice of R^*. The elementary vectors

$$(4.85) \qquad \iota(q) = [0_1 \quad \cdots \quad 0_{q-1}\ 1_q \quad 0_{q+1} \cdots \quad 0_Q], \quad q = 1, \ldots, Q,$$

[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.,

$$(4.86)\qquad \iota(q)K^{\dagger} = k_q\,\iota(q), \qquad \iota(q)L^{\dagger} = l_q\,\iota(q),$$
$$\iota(q_1)\ \iota'(q_2) = \begin{cases} 0 & \text{if } q_1 \neq q_2, \\ 1 & \text{if } q_1 = q_2. \end{cases}$$

If

$$(4.87)\qquad \bar{a}_n^{\dagger} = \sum_{q=1}^{Q} \{\bar{a}_n^{\dagger}\ \iota'(q)\}\ \iota(q)\ ,$$

the iterative process (4.82) consists – to the first order of magnitude – of a multiplication

$$(4.88)\qquad \bar{a}_1^{\dagger}\ \iota'(q) = k_q\{\bar{a}_0^{\dagger}\ \iota'(q)\}$$

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,

$$(4.89) \qquad \mathrm{M}_{xx} = \mathcal{E}M_{xx},$$

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

$$(4.90) \qquad \operatorname{plim}(M_{xx} - \mathrm{M}_{xx}) = 0.$$

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 $\bar{a}^\dagger$,

$$(4.91)\qquad L_{(2)} = \bar{a}^\dagger\, L^\dagger\, \bar{a}'^\dagger = \mathrm{tr}\{-(B'^{-1}\,\bar{B}')^2 - \bar{A}\,M_{xx}\,\bar{A}'\},$$

under the restrictions on $\bar{a}^\dagger$,

$$(4.92)\qquad 1 = \bar{a}^\dagger\,\bar{a}'^\dagger = \mathrm{tr}\,\bar{A}\,M_{xx}\,\bar{A}'.$$

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,

$$(4.93)\qquad H \equiv \begin{bmatrix} A \\ D \end{bmatrix} \equiv \begin{bmatrix} B & C \\ 0 & F \end{bmatrix}, \qquad \text{where} \qquad F'\,F = M_{zz}^{1},$$

which is such that

$$(4.94)\qquad H\,M_{xx}\,H' = I, \qquad \text{or} \qquad M_{xx} = H^{-1}\,H'^{-1}.$$

If we now transform $\bar{A}$ uniquely by

$$(4.95)\qquad \bar{A} \equiv \tilde{A}\,H \equiv \tilde{B}\,A + \tilde{C}\,D, \qquad \text{or} \qquad \tilde{A} \equiv \bar{A}\,H^{-1} \equiv [\tilde{B} \quad \tilde{C}],$$

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

$$(4.96)\qquad L_2 = \mathrm{tr}(-\tilde{B}'^2 - \tilde{A}\,\tilde{A}') = -\sum_{g,h=1}^{K_y} \tilde{a}_{gh}\,\tilde{a}_{hg} - \sum_{g=1}^{K_y}\sum_{k=1}^{K_x} \tilde{a}^2_{gk},$$

subject to the $\tilde{A}$-restrictions and the additional restriction

$$(4.97)\qquad \mathrm{tr}\,\tilde{A}\,\tilde{A}' = \sum_{g=1}^{K_y}\sum_{k=1}^{K_x} \tilde{a}^2_{gk} = 1,$$

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

$$(4.98)\qquad \frac{1}{2}\,\frac{d}{d\tilde{A}}(L_{(2)} - l\,\mathrm{tr}\,\tilde{A}\,\tilde{A}') = -\tilde{B}'\,I_{[K_y K_x]} - (1 + l)\tilde{A} = 0.$$

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" 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

$$(4.100)\qquad -2 \le L_{(2)} \le 0 \quad \text{if} \quad \mathrm{tr}\,\tilde{A}\,\tilde{A}' = 1.$$

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

$$(4.101) \qquad -2 \leq l_q \leq 0, \qquad 1-2h \leq k_q \leq 1, \qquad q = 1, \ldots, Q.$$

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

$$(4.102) \qquad \mathrm{A}^{\oplus} = \Upsilon\, \mathrm{A}$$

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 $A + \tilde{A}$ in a sufficiently small neighborhood of A are excluded by the restrictions on $A + \tilde{A}$. 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).

$$(4.106) \qquad -2 \leq l_q < 0, \qquad 1-2h \leq k_q < 1, \qquad q = 1, \ldots, Q.$$

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

$$(4.107) \qquad A + \bar{A} = (I + \tilde{B})A, \qquad (\tilde{C} = 0),$$

permitted by the restrictions on $A + \bar{A}$ are such that

$$(4.108) \qquad \tilde{B} \text{ is diagonal.}$$

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,

$$(4.109) \qquad l_q = -2, \qquad q = Q - K_y + 1, \ldots, Q.$$

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

$$(4.110) \qquad -2 < l_q < 0, \qquad 1-2h < k_q < 1, \qquad q = 1, \ldots, Q - K_y,$$

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 $\mathcal{P}_1$ 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 $\mathcal{P}_{1/2}$ instead of $\mathcal{P}_1$, 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 $\mathcal{P}_1$ in the circumstances at present assumed:

(a) The cost of computation per iteration for $\mathcal{P}_1$ is below that for any other constant value of h, and considerably below that for the process $\mathcal{P}_{h_n}$ 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 accumulated 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 $\mathbb{P}_1$ with an occasional application of $\mathbb{P}_{1/2}$, 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 $\mathbb{P}_1$ 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 $\mathbb{P}_h$ with a constant value of $h < 1$, say $h = 3/4$, or one may insert $\mathbb{P}_{1/2}$ at regular intervals. One initial application of $\mathbb{P}_{1/2}$ 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 $\mathbb{P}_1$, but the presence of higher-order terms in (4.88) may reintroduce those components to some extent, thus requiring that $\mathbb{P}_{1/2}$ now be inserted with regularity.

4.3.3.10. Asymptotic properties of $\mathbb{P}_{h_n}$. Alternatively, whether nontrivial characteristic values $l = -2$ are present or not, one may pay the higher cost per iteration of the process $\mathbb{P}_{h_n}$ 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.

$$h_0 = -\frac{\Delta\bar{a}_0^\dagger \cdot \Delta\bar{a}_0'^\dagger}{\Delta\bar{a}_0^\dagger \cdot L^\dagger \cdot \Delta\bar{a}_0'^\dagger} = -\frac{\bar{a}_0^\dagger\ (L^\dagger)^2\ \bar{a}_0'^\dagger}{\bar{a}_0^\dagger\ (L^\dagger)^3\ \bar{a}_0'^\dagger} + \cdots$$

$$(4.113) \qquad = -\frac{\sum_{q=1}^{Q} l_q^2\, \{\bar{a}_0^\dagger\ \iota'(q)\}^2}{\sum_{q=1}^{Q} l_q^3\, \{\bar{a}_0^\dagger\ \iota'(q)\}^2} + \cdots .$$

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

$$\bar{a}_1^\dagger\ \bar{a}_1'^\dagger = \sum_{q=1}^{Q} \{\bar{a}_1^\dagger\ \iota'(q)\}^2$$

$$(4.114) \qquad = \sum_{q=1}^{Q} \{1 + h_0\, l_q\}^2 \{\bar{a}_0^\dagger\ \iota'(q)\}^2 + \cdots ,$$

in keeping with the principle from which the process $\mathbb{P}_{h_n}$ is derived.

Another interesting property of $\mathbb{P}_{h_n}$ 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,

$$(4.115) \qquad \Delta\bar{a}_1^\dagger = \Delta\bar{a}_0^\dagger - \frac{\Delta\bar{a}_0^\dagger \cdot \Delta\bar{a}_0'^\dagger}{\Delta\bar{a}_0^\dagger \cdot L^\dagger \cdot \Delta\bar{a}_0'^\dagger} \cdot \Delta\bar{a}_0^\dagger \cdot L^\dagger + \cdots .$$

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 $\mathbb{P}_{h_n}$ are homogeneous of degree zero in the characteristic values l_q of $L^\dagger$. It fol-

[1] Taking $n = 0$ and 1, respectively.

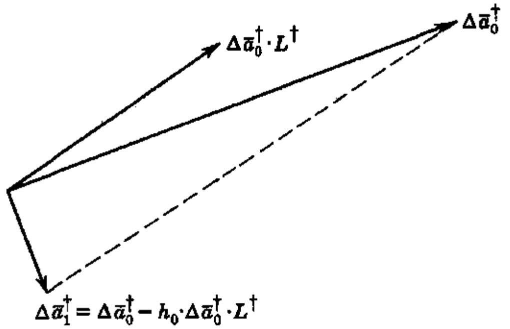
$\Delta\bar{a}_0^\dagger\cdot L^\dagger$
$\Delta\bar{a}_0^\dagger$
$\Delta\bar{a}_1^\dagger = \Delta\bar{a}_0^\dagger - h_0\cdot\Delta\bar{a}_0^\dagger\cdot L^\dagger$

Figure 4.3.3.10

lows that only the ratios of the values l_q affect the asymptotic speed of convergence of $\mathbb{P}_{h_n}$. 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 $\mathbb{P}_h$ and $\mathbb{P}_{h_n}$ 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 $\mathbb{P}_h$ and $\mathbb{P}_{h_n}$. Experience with actual data is needed to determine whether the greater speed of convergence that might be expected of $\mathbb{P}_{h_n}$ as compared with $\mathbb{P}_h$ for any constant h justifies the greater cost of computation per iteration. One would expect $\mathbb{P}_{h_n}$ to be particularly economical if the ratios of the characteristic values l_q are close to unity. In any case, $\mathbb{P}_{h_n}$ 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:

$$
\begin{aligned}
\tilde{A} &= [\tilde{B} \quad \tilde{C}] \\
&= [\tilde{B}_{\text{dia}} \quad 0] + [\tilde{B}_{\text{sym}} \quad 0] + [\tilde{B}_{\text{ant}} \quad 0] + [0 \quad \tilde{C}] \\
&= \tilde{B}_{\text{dia}}\, I_{[K_y\, K_x]} + \tilde{B}_{\text{sym}}\, I_{[K_y K_x]} + \tilde{B}_{\text{ant}}\, I_{[K_y K_x]} + \tilde{C}\, I'_{[K_x\, K_z]} \;,
\end{aligned}
\tag{4.116}
$$

$\tilde{B}_{\text{dia}}$ 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 restrictions 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

$$(4.117) \qquad C = 0, \quad \text{and hence} \quad \bar{C} = 0\,.$$

From (4.93) and (4.95) the corresponding $\tilde{A}$-restrictions are seen to be

$$(4.118) \qquad \tilde{C} = 0\,.$$

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 $\wp_{1/2}$ 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

$$\tilde{a}_{gh} = \bar{a}(g)\, H^{-1}\, \iota'(h) = \bar{a}(g) \begin{bmatrix} B^{-1} \\ 0 \end{bmatrix} \iota'(h) \tag{4.119}$$

$$= \bar{a}^{g}\, \mathrm{X}^{g}\, B^{-1}\, \iota'(h) = 0 \qquad \text{for} \qquad g \neq h, \quad g, h \leq G,$$

whatever $\bar{a}^{g}$, or that

$$\mathrm{X}^{g}\, B^{-1}\, \iota'(h) = 0 \qquad \text{for} \qquad g \neq h\,. \tag{4.120}$$

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 Λ_{yy} $\mathrm{B}^{\oplus}$ of an unknown nonsingular diagonal matrix Λ_{yy} with a known nonsingular matrix $\mathrm{B}^{\oplus}$. Hence a known nonsingular transformation,

$$y' = \mathrm{B}^{\oplus} y'^{\oplus}, \tag{4.123}$$

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 $\tilde{B}_{\mathrm{sym}} I_{[K_y K_x]}$ and $\tilde{B}_{\mathrm{ant}} I_{[K_y K_x]}$ leads to a trivial case which can be treated by more elementary methods. In fact, one single application of $\mathbb{P}_1$, 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 $\mathbb{P}_h = \mathbb{P}_{h_n}$ *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 $\mathcal{P}_1$, $\mathcal{P}_h$, $\mathcal{P}_{h_n}$

4.3.4.1. A constructed example for numerical illustration of $\mathcal{P}_h$, $\mathcal{P}_{h_n}$ *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

$$\mathrm{A} = \begin{bmatrix} 0 & 1 & 4 & 1 & 0 & 0 \\ 1 & 0 & -3 & 0 & 1 & 0 \\ -2 & 1 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\Sigma = \begin{bmatrix} 0.2 & 0.1 & 0.0 \\ 0.1 & 0.2 & 0.1 \\ 0.0 & 0.1 & 0.3 \end{bmatrix}, \tag{4.125}$$

$$M_{zz} = I,$$

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):

$$M_{yx} = \mathrm{M}_{yx} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{E} y'(t)\, x(t) \tag{4.126}$$

$$= \mathrm{B}^{-1}\, \Sigma\, \mathrm{B}'^{-1}\, I_{[K_y K_x]} - \mathrm{B}^{-1}\, \Gamma\, M_{zz} \left[-\Gamma'\, \mathrm{B}'^{-1} \qquad I_{[K_z]}\right]$$

$$= \begin{bmatrix} 0.417 & 0.484 & -0.021 & -0.300 & -0.400 & 0.300 \\ 0.484 & 1.568 & -0.192 & -0.600 & -0.800 & -0.400 \\ -0.021 & -0.192 & 0.073 & -0.100 & 0.200 & 0.100 \end{bmatrix}.$$

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

$$\begin{aligned} \mathcal{Q}(\Delta A_0\, M_{xx}) &= \mathcal{Q}(B_0'^{-1}\, I_{[K_y K_x]} - A_0\, M_{xx}), \\ \mathcal{Q}\, \Delta A_0 &= \Delta A_0, \end{aligned} \tag{4.127}$$

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 $\mathfrak{P}_1$ is applied,

$$(4.129) \qquad \mathfrak{D}(A_n M_{xz}) = 0 , \qquad \mathfrak{D}(\Delta A_n M_{xz}) = 0$$

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 $\mathfrak{D}\tilde{C}_n$ rather than its general reduction in successive iterations. It was already recognized in section *4.1.5* in connection with $\mathfrak{P}_1$ 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 $\mathfrak{D}B_n$ 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 $\mathfrak{D}\Delta C_n$ and $\mathfrak{D}C_n$ from (4.129) and substituting in (4.127), we have as the definition of $\mathcal{Q}\,\Delta B_0$ in that notation,

$$(4.130) \qquad \Delta b_0^g \cdot {}^zM_{yy}^g = \iota(g)\cdot B_0'^{-1}\cdot \Phi'^g - b_0^g\cdot {}^zM_{yy}^g ,$$

from which we solve for Δb_0^g for computational purposes,

$$(4.131) \qquad \Delta b_0^g = \iota(g)\; B_0'^{-1}\; \Phi'^g\,({}^zM_{yy}^g)^{-1} - b_0^g .$$

If $\mathbb{P}_{h_n}$ is applied, the value (4.69) of h is obtained most conveniently from

$$(4.132) \qquad h_0 = \frac{\sum_g \Delta b_0^g\cdot {}^zM_{yy}^g\cdot \Delta b_0'^g}{\operatorname{tr}\{(B_0'^{-1}\cdot \Delta B_0'^{-1})^2 + \sum_g \Delta b_0^g\cdot {}^zM_{yy}^g\cdot \Delta b_0'^g\}}$$

in which the first term in the denominator is not transformed to vector coordinates.

First the K_y matrices ${}^zM_{yy}^g$ 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 $({}^zM_{yy}^g)^{-1}$ represents a least-squares regression as a possible choice of initial vector b_0^g. The initial vectors are then normalized by

$$(4.133) \qquad b_0^g\cdot {}^zM_{yy}^g\cdot b_0'^g = 1,$$

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 $\mathbb{P}_{h_n}$ with the help of (4.132), and the next iteration can proceed.

If the process is terminated after the nth iteration, $\mathcal{Q}C_n$ is obtained from $\mathcal{Q}B_n$ by (4.5) or

$$(4.134) \qquad c_n^g = -\,b_n^g\cdot M_{yz}^g\cdot (M_{zz}^g)^{-1} .$$

4.3.4.4. Application of $\mathbb{P}_1$, $\mathbb{P}_{3/4}$, $\mathbb{P}_{h_n}$ *to the constructed example.* Table 4.3.4.4 gives the results of applying $\mathbb{P}_1$, $\mathbb{P}_{3/4}$,

TABLE 4.3.4.4. Numerical illustration of $\mathcal{P}_1$,

Method	Iteration $n =$	Ratios of elements of $b_n(g)$			Scales: $m\{a(g)\} \equiv \{a_n(g)\ M_{xx}\ a_n'(g)\}^{1/2}$		
		$-\frac{b^n_{12}}{b^n_{13}}$	$-\frac{b^n_{21}}{b^n_{23}}$	$-\frac{b^n_{32}}{b^n_{31}}$	$m\{a_1(g)\}$	$m\{a_2(g)\}$	$m\{a_3(g)\}$
(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

$\wp_{3/4}$, and $\wp_{h_n}$ when Σ is required to be diagonal.

$\wp_{h_n}$							
$h_n = 1.06540$	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 $\wp_1$ 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 $\mathbb{P}_{h_n}$ 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 $\mathbb{P}_1$ is superior to $\mathbb{P}_{3/4}$, and even slightly better than $\mathbb{P}_{h_n}$, with respect to the most essential property: the speed of convergence of the ratios of elements of each $a_n(g)$. $\mathbb{P}_1$ shows the characteristic alternation in successive values $m\{a_n(g)\}$, $n = 0, 1, \ldots$. These values converge gradually in $\mathbb{P}_{3/4}$ and slightly faster in $\mathbb{P}_{h_n}$. 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

$$(4.135) \qquad \Delta a_0^* M^{**} = \mathrm{vec}[B_0'^{-1} \quad 0]\,\Phi'^* - a_0^* M^{**}.$$

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

$$(4.136)\qquad \Delta a_0^* = \mathrm{vec}[B_0'^{-1} \quad 0]\, \Phi'^* (M^{**})^{-1} - a_0^* ,$$

to which we again apply (4.37) to obtain

$$(4.137)\qquad \Delta a_0 = \mathrm{vec}[B_0'^{-1} \quad 0]\, \Phi'^* (M^{**})^{-1}\, \Phi^* - a_0 .$$

Defining

$$(4.138)\qquad P^{(g)} \equiv \Phi'^g (\Phi^g M \Phi'^g)^{-1}\, \Phi^g \equiv P_{xx}^{(g)} \equiv \begin{bmatrix} P_{yy}^{(g)} & P_{yz}^{(g)} \\ P_{zy}^{(g)} & P_{zz}^{(g)} \end{bmatrix} ,$$

we have

$$(4.139)\qquad P \equiv \Phi'^* (M^{**})^{-1}\, \Phi^* = \begin{bmatrix} P^{(1)} & 0 & \dots & 0 \\ 0 & P^{(2)} & \dots & 0 \\ \cdot & \cdot & \cdot\,\cdot\,\cdot & \cdot \\ 0 & 0 & \dots & P^{(G)} \end{bmatrix} .$$

Therefore, (4.137) is equivalent to

$$(4.140)\qquad \Delta a_0(g) = \iota(g)\, B_0'^{-1} P_{yx}^{(g)} - a_0(g) , \qquad g = 1, \dots, G .$$

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

$$Q_g \equiv \rho(\bar{\mathrm{X}}^g) + Q_g^{\mathrm{III}} \equiv \rho(\bar{\Psi}^g) + Q_g^{\mathrm{I}} \equiv Q_g^{\mathrm{I}} + Q_g^{\mathrm{II}} + Q_g^{\mathrm{III}}, \tag{4.142}$$

then there exists a nonsingular transformation matrix[1] Ω *such that*

$$\Omega\,\bar{\Phi}^g = \Phi^g = [\mathrm{X}^g \quad \Psi^g] = \begin{bmatrix} \mathrm{X}_{\mathrm{I}}^g & 0 \\ \mathrm{X}_{\mathrm{II}}^g & \Psi_{\mathrm{II}}^g \\ 0 & \Psi_{\mathrm{III}}^g \end{bmatrix}, \tag{4.143}$$

with

$$\begin{aligned} \rho(\mathrm{X}_{\mathrm{I}}^g) &= r(\mathrm{X}_{\mathrm{I}}^g) = Q_g^{\mathrm{I}}, \\ \rho(\Psi_{\mathrm{III}}^g) &= r(\Psi_{\mathrm{III}}^g) = Q_g^{\mathrm{III}}, \\ \rho(\Phi_{\mathrm{II}}^g) &= r(\Phi_{\mathrm{II}}^g) = Q_g^{\mathrm{II}}, \end{aligned} \tag{4.144}$$

and there exists no nonsingular transformation matrix Ω such that in (4.143) $\mathrm{X}_{\mathrm{I}}^g$ 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

$$(4.146) \qquad \Omega_{III}\,\bar{X}^{g} = 0, \qquad \Omega_{I}\,\bar{\Psi}^{g} = 0.$$

We must have

$$(4.147) \qquad \rho\begin{pmatrix}\Omega_{I}\\ \Omega_{III}\end{pmatrix} = \rho(\Omega_{I}) + \rho(\Omega_{III})$$

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

$$(4.148) \qquad \mu = \lambda_{I}\,\Omega_{I} = \lambda_{III}\,\Omega_{III} \neq 0.$$

In that case (4.146) would imply

$$(4.149) \qquad \mu[\bar{X}^{g} \quad \bar{\Psi}^{g}] = \mu\,\bar{\Phi}^{g} = 0, \qquad \mu \neq 0,$$

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

$$(4.150) \qquad \rho(\Omega_{II}) = r(\Omega_{II}) = Q_{g}^{II} \geq 0$$

such that the matrix Ω defined by

$$(4.151) \qquad \Omega \equiv \begin{bmatrix}\Omega_{I}\\ \Omega_{II}\\ \Omega_{III}\end{bmatrix}$$

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 Φ^{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:

$$(4.153) \qquad \begin{aligned} \beta(g) &= \alpha_{\mathrm{I}}^g X_{\mathrm{I}}^g + \alpha_{\mathrm{II}}^g X_{\mathrm{II}}^g , \\ \gamma(g) &= \qquad\qquad \alpha_{\mathrm{II}}^g \Psi_{\mathrm{II}}^g + \alpha_{\mathrm{III}}^g \Psi_{\mathrm{III}}^g . \end{aligned}$$

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,

$$a^g_{\mathrm{III}} = -\,{}_{\mathrm{III}}a^{g}\cdot {}_{\mathrm{III}}M^g_{\mathrm{III}}\cdot(M^g_{\mathrm{III\;III}})^{-1}, \tag{4.161}$$

which is the equivalent of (4.134) under the present form of the Φ^g. From (4.22) and (4.143) we have

$$b_n(g) = {}_{\mathrm{III}}a^g_n\cdot {}_{\mathrm{III}}X^g, \qquad c_n(g) = a^g_{n,\mathrm{II}}\cdot \Psi^g_{\mathrm{II}} + a^g_{n,\mathrm{III}}\cdot \Psi^g_{\mathrm{III}}\,. \tag{4.162}$$

The first of these relations is solved for ${}_{\mathrm{III}}a^g_n$ by

$${}_{\mathrm{III}}a^g_n = b_n(g)\,\Xi, \tag{4.163}$$

where the relation

$${}_{\mathrm{III}}X^g\cdot\Xi = {}_{\mathrm{III\;III}}I \tag{4.164}$$

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

$$\Xi = {}_{\mathrm{III}}X'^g\,({}_{\mathrm{III}}X^g\cdot {}_{\mathrm{III}}X'^g)^{-1}. \tag{4.165}$$

Combining the second relation (4.162) with (4.161) and (4.163), we have

$$
\begin{aligned}
c_n(g) &= b_n(g)\ \Xi\ \{{}_{\mathrm{III}}\Psi^{g} - {}_{\mathrm{III}}M^{g}_{\mathrm{III}}\,(M^{g}_{\mathrm{III\ III}})^{-1}\ \Psi^{g}_{\mathrm{III}}\} \\
&\equiv b_n(g)\ J(g)\,,
\end{aligned}
\tag{4.166}
$$

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 $\mathbb{P}_{h_n}$, 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

$$
h_0 = \frac{\sum\limits_{g=1}^{G} \Delta b_0(g)\cdot M_{yy}(g)\cdot \Delta b_0'(g)}{\operatorname{tr}(B_0'^{-1}\cdot \Delta B_0') + \sum\limits_{g=1}^{G} \Delta b_0(g)\cdot M_{yy}(g)\cdot \Delta b_0'(g)}
\tag{4.167}
$$

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

$$\begin{aligned} L(A_1) - L(A_0) &= \operatorname{tr}\{(B_0'^{-1}\, I_{[K_y\, K_x]} - A_0\, M_{xx})\,\Delta A_0'\} \\ &\quad + \frac{1}{2}\operatorname{tr}\{-(B_0'^{-1}\,\Delta B_0')^2 - \Delta A_0\, M_{xx}\,\Delta A_0'\} + \cdots . \end{aligned} \tag{4.170}$$

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 $\mathbb{P}_h$, $\mathbb{P}_{h_n}$. 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 $\bar{\alpha}^* = \alpha^* - a^*$,

$$(4.174) \qquad L(\alpha^*) - L(a^*) = \frac{1}{2}\,\bar{\alpha}^*\, L^*\, \bar{\alpha}'^* + \cdots ,$$

we have

$$(4.175) \qquad l_0^* \equiv \left(\frac{d\,L(\alpha^*)}{d\bar{\alpha}^*}\right)_{\bar{\alpha}^* = \bar{a}_0^*} = \bar{a}_0^*\, L^* + \cdots .$$

From (4.173) and (4.175) we have the relation

$$(4.176) \qquad \bar{a}_1^* = \bar{a}_0^* - l_0^*\,(L_0^*)^{-1} = \bar{a}_0^*\,\{\, I - L^*\,(L_0^*)^{-1} + \cdots \,\},$$

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

$$l_0^* \equiv \left(\frac{d\,L(\alpha^*)}{d\,\alpha^*}\right)_{\alpha^* = a_0^*} = \operatorname{vec}^* (B_0'^{-1} \; I_{[K_y K_x]} - A_0 \, M_{xx}). \tag{4.178}$$

On the other hand we have, for any $\mathrm{A} = \mathrm{mat} * \alpha^*$, from the definition (4.171) of L_0^*,

$$\begin{aligned} \alpha^* \, L_0 &= \frac{1}{2}\frac{d}{d\alpha^*}(\alpha^* \, L_0^* \, \alpha'^*) \\ &= \operatorname{vec}^* (-B_0'^{-1} \, \mathrm{B}' \, B_0'^{-1} \; I_{[K_y K_x]} - \mathrm{A} \, M_{xx}), \end{aligned} \tag{4.179}$$

and, hence, in particular,

$$(4.180) \qquad a_0^* \, L_0^* = \text{vec}^*(-B_0'^{-1} \, I_{[K_y K_x]} - A_0 \, M_{xx}).$$

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

$$(4.181) \qquad \begin{aligned} -a_1^* \, L_0^* &\equiv \text{vec}^*(B_0'^{-1} \, B_1' \, B_0'^{-1} \, I_{[K_y K_x]} + A_1 \, M_{xx}) \\ &= 2\,\text{vec}^*(B_0'^{-1} \, I_{[K_y K_x]}). \end{aligned}$$

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

$$(4.182) \qquad {}_{\text{III}}\text{X}_{\text{I}}^* \equiv \begin{bmatrix} {}_{\text{III}}\text{X}^1 & 0 & \dots & 0 \\ 0 & {}_{\text{III}}\text{X}^2 & \dots & 0 \\ \cdot & \cdot & \dots & \cdot \\ 0 & 0 & \dots & {}_{\text{III}}\text{X}^G \end{bmatrix}.$$

It is seen from (4.181) [or from the definition (4.171) of L_0^*] that

$$(4.183) \qquad L_0^* = \begin{bmatrix} {}_{\text{III}\,\text{III}}L_0^* & -{}_{\text{III}}M_{\text{III}}^{**} \\ -{}_{\text{III}}M_{\text{III}}'^{**} & -M_{\text{III}\,\text{III}}^{**} \end{bmatrix},$$

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

$$(4.184) \qquad a_1^* \begin{bmatrix} -{}_{III}M^{**}_{III} \\ \\ -M^{**}_{III\,III} \end{bmatrix} = -{}_{III}a_1^* \cdot {}_{III}M^{**}_{III} - a^*_{1,III} \cdot M^{**}_{III\,III} = 0,$$

from which $a^*_{1,III}$ can be solved according to (4.161) and inserted in (4.181) to give

$$(4.185) \qquad {}_{III}a_1^* \cdot {}^{III}L_0^* = {}_{III}\text{vec}^* \, [B_0'^{-1} \qquad 0] \, ,$$

with

$$(4.186) \qquad {}^{III}L_0^* \equiv {}_{III\,III}L_0^* + {}_{III}M^{**}_{III} \cdot (M^{**}_{III\,III})^{-1} \cdot {}_{III}M'^{**}_{III} \, .$$

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 $M^{**}_{III\,III}$ can be carried out for each of its diagonal blocks $M^{g}_{III\,III}$ separately. Then one chooses B_0, calculates its inverse and uses it in calculating both

$$(4.187) \qquad {}_{III}\text{vec}^* \, [B_0'^{-1} \qquad 0] = \text{vec}[B_0'^{-1} \qquad 0] \cdot {}_{III}\Phi'^*$$

and

$$(4.188) \qquad {}_{III\,III}L_0^* = -{}_{III}X^* \cdot K_0 \cdot {}_{III}X'^* - {}_{III\,III}M^{**} \, .$$

The matrix K_0 is defined by

$$(4.189) \qquad \text{tr}(B_0'^{-1} \, B')^2 \equiv {}_{III}a^* \cdot {}_{III}X^* \cdot K_0 \cdot {}_{III}X'^* \cdot {}_{III}a'^* ,$$

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, ${}^{\mathrm{III}}L_0^*$ is put together from (4.186), and ${}_{\mathrm{III}}a_1^*$ solved from (4.185), leading to

$$(4.192) \qquad B_1 = \mathrm{mat}\,({}_{\mathrm{III}}a_1^* \cdot {}_{\mathrm{III}}\mathrm{X}^*) .$$

At the termination of iterations, C_n is obtained from

$$(4.193) \qquad \begin{aligned} C_n &= \mathrm{mat}\,({}_{\mathrm{III}}a_n^* \cdot {}_{\mathrm{III}}\Psi^* + a^*_{n\,\mathrm{III}} \cdot \Psi^*_{\mathrm{III}}) \\ &= \mathrm{mat}\,[{}_{\mathrm{III}}a_n^* \{ {}_{\mathrm{III}}\Psi^* - {}_{\mathrm{III}}M^{**}_{\mathrm{III}} \cdot (M^{**}_{\mathrm{III}\,\mathrm{III}})^{-1} \cdot \Psi^*_{\mathrm{III}})] . \end{aligned}$$

It appears from (4.188) that the matrix ${}_{\mathrm{III}\,\mathrm{III}}L_0^*$ has considerable regularity in its make-up. The problem of how to best utilize those regularities for the inversion of ${}_{\mathrm{III}\,\mathrm{III}}L_0^*$ 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

$$\text{(4.194)}\qquad \begin{aligned} \text{est}\,\mathcal{E}\{(a'^{*} - \alpha'^{*})(a^{*} - \alpha^{*})\} &= -\frac{1}{T}\left[\frac{\partial^2 L(\alpha^*)}{\partial\alpha'^{*}\,\partial\alpha^{*}}\right]^{-1}_{\alpha^* = a^*} \\ &= -\frac{1}{T}\,L^{*-1} \end{aligned}$$

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_{III}^{g} : for any ${}_{\text{III}}a_{n}^{g}$ the corresponding maximizing values of α_{III}^{g} 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

$$(4.197)\qquad \begin{aligned} \alpha^{*}_{[1]} &\equiv [\ \alpha^{1}\cdot\iota'(1) \quad \alpha^{2}\cdot\iota'(1) \quad \ldots \quad \alpha^{G}\cdot\iota'(1)\ ] \\ &= [\ 1 \quad 1 \quad \ldots \quad 1\ ] \end{aligned}$$

to express the normalization rule, and introduce similar notations

$$\Phi^{g} = \begin{bmatrix} \Phi^{g}_{[1]} \\ {}_{[1]}\Phi^{g} \end{bmatrix},$$

$$(4.198)\quad \begin{aligned} \Phi^*_{[1]} &\equiv \begin{bmatrix} \Phi^1_{[1]} & 0 & \cdots & 0 \\ 0 & \Phi^2_{[1]} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \Phi^G_{[1]} \end{bmatrix}, \\ {}_{[1]}\Phi^* &\equiv \begin{bmatrix} {}_{[1]}\Phi^1 & 0 & \cdots & 0 \\ 0 & {}_{[1]}\Phi^2 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & {}_{[1]}\Phi^G \end{bmatrix}, \end{aligned}$$

for the corresponding partitioning of the basic matrices.

4.4.3. Elimination of the parameters Σ. We shall first maximize the likelihood function

$$(4.199)\quad \begin{aligned} L(\mathrm{A},\ \Sigma) = \text{const} + \log\det \mathrm{B} - \frac{1}{2}\log\det\Sigma \\ - \frac{1}{2}\operatorname{tr}(\Sigma^{-1}\,\mathrm{A}\,M_{xx}\,\mathrm{A}') \end{aligned}$$

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

$$(4.200)\quad \Sigma = \hat{\Sigma} \equiv \mathrm{A}\,M_{xx}\,\mathrm{A}' = \hat{\Sigma}'.$$

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),

$$(4.203) \qquad \begin{aligned} L(\mathrm{A}) - L(A_0) &= \operatorname{tr}(B_0'^{-1}\Delta B_0') \\ &\quad - \frac{1}{2}\operatorname{tr}\{S_0^{-1}(A_0\, M_{xx}\, \Delta A_0' + \Delta A_0\, M_{xx}\, A_0')\} + \cdots \\ &= \operatorname{tr}\{B_0'^{-1}\, I_{[K_y K_x]} - S_0^{-1}\, A_0\, M_{xx}\}\,\Delta A_0' + \cdots . \end{aligned}$$

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),

$$\begin{aligned} L(A_1) - L(A_0) = & \\ = h \, \text{tr}\{ & (B_0'^{-1} \, I_{[K_y \, K_x]} - S_0^{-1} A_0 \, M_{xx}) \, \Delta A_0' \} \\ + \frac{1}{2} h^2 \, \text{tr} \{ & - (B_0'^{-1} \, \Delta B_0')^2 + S_0^{-1} (A_0 \, M_{xx} \, \Delta A_0' \\ & + \Delta A_0' \, M_{xx} \, A_0') \, S_0^{-1} A_0 \, M_{xx} \, \Delta A_0' - S_0^{-1} \, \Delta A_0 \, M_{xx} \, \Delta A_0' \} \\ + \cdots . & \end{aligned} \tag{4.207}$$

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)

$$h_0 = \frac{\operatorname{tr}(S_0^{-1}\,\Delta A_0\, M_{xx}\,\Delta A_0')}{\begin{aligned}\operatorname{tr}\{(B_0'^{\,-1}\,\Delta B_0') - S_0^{-1}(A_0\, M_{xx}\,\Delta A_0' \\ + \Delta A_0\, M_{xx}\, A_0)\, S_0^{-1} A_0\, M_{xx}\,\Delta A_0' + S_0^{-1}\,\Delta A_0\, M_{xx}\,\Delta A_0'\}\end{aligned}}\,.$$

We define matrices M_0^* and L_0^* such that the numerator and denominator in (4.208) are identical with those in

$$h_0 = \frac{\Delta a_0^*\, M_0^*\,\Delta a_0'^{*}}{-\Delta a_0^*\, L_0^*\,\Delta a_0'^{*}}\,, \tag{4.209}$$

postponing their explicit evaluation until the discussion of computational arrangements below.

4.4.6. Asymptotic properties of $\mathbb{P}_h$ *and* $\mathbb{P}_{h_n}$. 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),

$$\begin{aligned} &\operatorname{vec}^*(S^{-1}\,\Delta\bar{A}_0\, M_{xx}) + \cdots \qquad (4.210)\\ &= \operatorname{vec}^*\{-B'^{\,-1}\,\bar{B}_0'\, B'^{\,-1}\, I_{[K_y\, K_x]} \\ &\qquad + S^{-1}(A\, M_{xx}\,\bar{A}_0' + \bar{A}_0\, M_{xx}\, A')\, S^{-1}\, A\, M_{xx} - S^{-1}\,\bar{A}_0\, M_{xx}\}\,. \end{aligned}$$

Omitting bars from Δa_n and ΔA_n (see p.174), this can be written as

$$\Delta a_0^*\, M^* + \cdots = \bar{a}_0^*\, L^* + \cdots\,, \tag{4.211}$$

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)

$$(4.212) \qquad H\,M_{xx}\,H' = \begin{bmatrix} S & 0 \\ 0 & I_{[K_z]} \end{bmatrix} = T\,,$$

say. Writing for the moment, instead of (4.95),

$$(4.213) \qquad \bar{A} = \tilde{A}^{\oplus} H\,,$$

we have from (4.212) and (4.202)

(4.214)

$$L_{(2)} \equiv \operatorname{tr}\{-(B'^{-1}\,\bar{B}')^2 + S^{-1}(AM\bar{A}' + \bar{A}M_{xx}A')S^{-1}\,AM_{xx}\bar{A}' - S^{-1}\bar{A}M_{xx}\bar{A}'\}$$

$$= \operatorname{tr}\{-(\tilde{B}'^{\oplus})^2 + (\tilde{B}'^{\oplus} + S^{-1}\tilde{B}^{\oplus}S)\,\tilde{B}'^{\oplus} - S^{-1}(\tilde{B}^{\oplus}S\tilde{B}'^{\oplus} + \tilde{C}^{\oplus}\tilde{C}'^{\oplus}))\}$$

$$= -\operatorname{tr}(S^{-1}\,\tilde{C}^{\oplus}\tilde{C}'^{\oplus})$$

and

$$(4.215) \quad M_{(2)} \equiv \operatorname{tr}(S^{-1}\,\bar{A}\,M_{xx}\,\bar{A}') = \operatorname{tr}\{S^{-1}(\tilde{B}^{\oplus}\,S\,\tilde{B}'^{\oplus} + \tilde{C}^{\oplus}\,\tilde{C}'^{\oplus})\}\,.$$

Through a further transformation

$$(4.216) \qquad S = U\,U'\,, \qquad U^{-1}\,\tilde{A}^{\oplus}\begin{bmatrix} U & 0 \\ 0 & I_{[K_z]} \end{bmatrix} = \tilde{A}\,,$$

it is seen that, in the absence of restrictions on A, the characteristic values l_q are the stationary values of the quadratic form

$$(4.217) \qquad L_{(2)} = -\operatorname{tr}(\tilde{C}\,\tilde{C}') = -\sum_{g=1}^{K_y}\ \sum_{h=K_y+1}^{K_z} \tilde{a}_{gh}^2$$

under the restrictions

$$(4.218) \qquad M_{(2)} = \operatorname{tr}(\tilde{A}\,\tilde{A}') = \sum_{g=1}^{K_y} \sum_{h=1}^{K_x} \tilde{a}_{gh}^2 = 1\,.$$

We record in one formula combining (4.213) and (4.216) the transformation

$$(4.219) \qquad \bar{A} = U\tilde{A}\begin{bmatrix} U^{-1} & 0 \\ 0 & I_{[K_z]} \end{bmatrix} H = U\tilde{B}U^{-1}A + U\tilde{C}FI_{[K_z\,K_x]}$$

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)	(b)	(c)	(d)
Value of l	Value of $k = 1 + h\,l$	Multi- plicity	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,

$$(4.221) \qquad -1 \leq l_q \leq 0, \qquad 1 - h \leq k_q \leq 1, \qquad q = 1, \ldots, Q.$$

Furthermore, under any such restrictions that in addition ensure (as here supposed) complete identification of each structural equation,

$$(4.222) \quad -1 \leq l_q < 0, \quad 1 - h \leq k_q < 1, \quad q = 1, \ldots, Q - K_y,$$

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 $\mathbb{P}_h$ with a constant value of h, $\mathbb{P}_1$ does not have the excellence it possesses in one important case with uncorrelated disturbances. In large samples under valid restrictions, $\mathbb{P}_1$ 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 $\mathbb{P}_1$, now leads to the recommendation of $\mathbb{P}_2$: 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* $\mathbb{P}_h$, $\mathbb{P}_{h_n}$. 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

$$(4.227) \qquad S_n^{-1} \equiv [\ s_n^{gh}\ ],$$

which are again functions of A_n. Combining (4.225) and (4.226), we can alternatively write for M_n^*, using (4.31),

$$(4.228) \qquad M_n^* = \begin{bmatrix} M_n^{11} & \cdots & M_n^{1G} \\ \cdot & \cdots & \cdot \\ M_n^{G1} & \cdots & M_n^{GG} \end{bmatrix},$$

with the further definition

$$(4.229) \qquad M_n^{gh} \equiv s_n^{gh}\ \Phi^g\ M_{xx}\ \Phi'^{\,h} \equiv s_n^{gh}\ V^{gh},$$

say. The symbol V^{gh} is merely an abbreviation for the matrix product it represents. There is nó 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 $\mathbb{P}_h$ 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 remaining 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

$$(4.230) \qquad M_n^{-1} = S_n \otimes M_{xx}^{-1},$$

followed by

$$(4.231) \qquad M_n^{-1}(*) = \Phi'^{-1}(*)\, M_n^{-1}\, \Phi^{-1}(*) .$$

[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

$$(4.232) \qquad \operatorname{vec}^*_{[1]} \Delta A_0 = 0 ,$$

and the first-order condition for a maximum would be expressed by

$$(4.233) \begin{cases} (4.233,-1) \qquad {}_{[1]}\operatorname{vec}^*(B'^{-1} I_{[K_y K_x]} - S^{-1} A\, M_{xx}) = 0 , \\ (4.233,+1) \qquad a^*_{[1]} = [\,1 \;\; 1 \;\; \cdots \;\; 1\,] . \end{cases}$$

It follows from the homogeneity properties of the likelihood function that (4.233) is equivalent to (4.205). However, the analogous iterative procedures using ΔA_0 defined by

$$(4.234)\begin{cases} (4.234,-1) & {}_{[1]}\mathrm{vec}^*(S_0^{-1}\,\Delta A_0\,M_{xx}) \\ & \qquad = {}_{[1]}\mathrm{vec}(B_0'^{-1}\,I_{[K_y K_x]} - S_0^{-1}\,A_0\,M_{xx}), \\ (4.234,+1) & \Delta a^*_{0,\,[1]} = 0\,, \quad \text{or} \end{cases}$$

$$\Delta a_0 = (\Delta\,{}_{[1]}a_0^*)_{[1]}\Phi^*\,,$$

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^g 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 $\mathcal{P}_h$, $\mathcal{P}_{h_n}$ 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 $\mathcal{P}_h$ 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 $\wp_{h}$, $\wp_{h_n}$ differs from that described earlier only in easily perceived details.

4.4.10. Numerical illustrations of $\wp_1$, $\wp_{5/4}$. We have applied $\wp_1$ and $\wp_{5/4}$ without normalization, and $\wp_1$ modified by normalization, to the constructed example discussed before. As initial values we have taken the result of the 6th iteration with $\wp_1$ 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

$$\begin{aligned} &a_1^* = a_0^* + \Delta a_0^*, \\ (4.235) \qquad &(\Delta_{[1]}a_0^*)\,{}_{[1]}L_0^* = -{}_{[1]}l_0, \qquad \Delta a^*_{0,[1]} = 0. \end{aligned}$$

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)

$$(4.236) \quad {}_{[1]}l_0 \equiv \left(\frac{d\,L(\alpha^*)}{d\,{}_{[1]}\alpha^*}\right)_{\alpha^*=a_0^*} = {}_{[1]}\mathrm{vec}^*\{B_0'^{-1}\,I_{[K_y\,K_x]} - S_0^{-1}\,A_0\,M_{xx}\}.$$

TABLE 4.4.10

Numerical illustrations of $\wp_1$ and $\wp_{5/4}$ without restrictions on Σ.

Method	Iter-ation	Row	Matrices B_n			Matrices S_n			Scale Factors[2]
	$n =$	$g =$	$k = 1$	2	3	$h = 1$	2	3	$a_{n;g,g+3}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\wp_1$ without 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	0.99521	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
		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 $\wp_1$ (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)
$\mathfrak{P}_{5/4}$ without normalization[1]	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	1.000C0
		3	−2.04599	0.96553	0.00000	−0.02277	0.08700	0.29509	1.00000
	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
		3	−2.00537	1.00362	0.00000	0.00002	0.09957	0.30200	0.99498
	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 $\mathfrak{P}_{5/4}$ (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)
$\mathbb{P}_1$ 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	0.99331	4.03780	0.19760	0.09666	−0.00542
		2	0.99910	0.00000	−3.00006	0.09666	0.19986	0.09740
		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
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*.

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

$$(4.237)\quad \begin{aligned}(\Delta_{[1]}a_0^*)\;{}_{[1]}L_0^* &= \frac{1}{2}\,\frac{d}{d(\Delta_{[1]}a_0^*)}\,\{(\Delta_{[1]}a_0^*)\;{}_{[1]}L_0^*\;(\Delta_{[1]}a_0'^*)\}\\ &= \frac{1}{2}\,\frac{d}{d(\Delta_{[1]}a_0^*)}\,(\Delta a_0^*\;L_0^*\;\Delta a_0'^*)\,.\end{aligned}$$

Combining (4.235), (4.236), (4.237), and the definition (4.208) – (4.209) of L_0^*, we obtain

$$(4.238)\quad \begin{aligned}(\Delta_{[1]}a_0^*)\;{}_{[1]}L_0^* = {}& {}_{[1]}\mathrm{vec}^*\{-[B_0'^{-1}\,\Delta B_0'\,B_0'^{-1} \qquad 0]\\ &+ S_0^{-1}(A_0\,M_{xx}\,\Delta A_0 + \Delta A_0\,M_{xx}\,A_0')\,S_0^{-1}\,A_0\,M_{xx}\\ &- S_0^{-1}\,\Delta A_0\,M_{xx}\}\\ = {}& {}_{[1]}\mathrm{vec}^*(-B_0'^{-1}\,I_{[K_y\;K_x]} + S_0^{-1}\,A_0\,M_{xx})\end{aligned}$$

as a formulation of the Newton method adapted to computational use. The middle member of (4.238) serves to define ${}_{[1]}L_0^*$ through

$$(4.239)\qquad \Delta A_0 \equiv \mathrm{mat}^*\{(\Delta_{[1]}a_0^*)\;{}_{[1]}\Phi^*\}\,.$$

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 $\mathbb{P}_h$ 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 $\mathbb{P}_1$, 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 $L(A, \Sigma)$ 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

$$\text{(4.242)}\quad \frac{\partial L(A, \Sigma)}{\partial\sigma}\,\delta\sigma' = -\frac{1}{2}\operatorname{tr}\{\Sigma^{-1}\,\delta\Sigma\,(I - \Sigma^{-1} A\, M_{xx}\, A')\}\,.$$

We introduce the notational definition

$$\left(\begin{bmatrix} \dfrac{\partial^2}{\partial_{[1]}\alpha'^*\,\partial_{[1]}\alpha^*} & \dfrac{\partial^2}{\partial_{[1]}\alpha'^*\,\partial\sigma} \\[2ex] \dfrac{\partial^2}{\partial\sigma'\,\partial\alpha^*} & \dfrac{\partial^2}{\partial\sigma'\,\partial\sigma} \end{bmatrix} L(\mathrm{A}\,,\,\Sigma)\right)_{\substack{\mathrm{A}\,=\,A\\ \Sigma\,=\,S}}$$

(4.243)

$$= \begin{bmatrix} {}_{[11]}\hat{L}^* & {}_{[1]}\hat{L}^*_{\sigma} \\[1ex] {}_{[1]}\hat{L}'^*_{\sigma} & \hat{L}^*_{\sigma\,\sigma} \end{bmatrix} = \hat{L}\,.$$

Then, if $\delta a^*_{[1]} = 0$, we have from (4.242) and (4.202), after using (3.16) and (3.19),

$$\delta_{[1]}a^*\cdot{}_{[11]}\hat{L}^*\cdot\delta_{[1]}a'^* = \mathrm{tr}\{-(B'^{-1}\,\delta A')^2 - S^{-1}\,\delta A\,M_{xx}\,\delta A'\}\,,$$

$$\delta_{[1]}a^*\;{}_{[1]}\hat{L}^*_{\sigma}\,\delta s' = \mathrm{tr}\{S^{-1}\,\delta\Sigma\,S^{-1}\,A\,M_{xx}\,\delta A'\},,$$

$$\delta s\;\hat{L}^*_{\sigma\sigma}\,\delta s' = -\frac{1}{2}\,\mathrm{tr}(S^{-1}\,\delta\,\Sigma)^2\,, \tag{4.244}$$

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

$$\text{est}\;\mathcal{E}\begin{bmatrix} ({}_{[1]}a'^* - {}_{[1]}\alpha'^*) \\ (s' - \sigma') \end{bmatrix}\Big[({}_{[1]}a^* - {}_{[1]}\alpha^*) \qquad (s-\sigma)\Big]$$

(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 $\mathbb{P}_h$ and $\mathbb{P}_{h_n}$ 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 a^*_{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. 2. The first term requires for its evaluation the inversion of B_n for each iteration. This term vanishes outside ${}_{\mathrm{III\,III}}L_n^*$. 3. The remaining term $-M_n^*$ is constant in iterations, so that use of ${}^{\mathrm{III}}L_n^*$ permits the elimination of α^*_{III}. The required inversion of $M^{**}_{\mathrm{III\,III}}$ is facilitated by partitioning into diagonal blocks. 4. The inversion of ${}^{\mathrm{III}}L_n^*$ (or other processing in solving for ${}_{\mathrm{III}}a^*_{n+1}$) must be repeated for each iteration. 5. There may be advantages in the regularities of ${}^{\mathrm{III}}L_n^*$ for its inversion or other processing.	1. ${}_{[1]}L_n^*$ contains four terms. 2. Its evaluation requires the inversion of B_n and S_n for each iteration. 3. No submatrix of ${}_{[1]}L_n^*$ remains constant in iterations. Its inversion or other processing must be repeated for each iteration. 4. The advantage from regularities in ${}_{[1]}L_n^*$ for its inversion or other processing are highly uncertain.

The relative cheapness of $\wp_h$ and $\wp_{h_n}$ 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 B I, A II, B II 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^{-1}_{xx} 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^{-1}_{n-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^{-1}_n 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 (4.246 l). The treatment of restrictions of the type (4.246k) has already been demonstrated above. A restriction of the type (4.246 l) 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

$$(4.249) \qquad \det \mathrm{B} = 0$$

divides the space of the elements α_{gh}, $g, h = 1, \ldots, K_y$, into two connected regions. The logarithmic likelihood function

$$(4.250) \qquad L(\mathrm{A}) = \text{const} + \log \det \mathrm{B} - \frac{1}{2} \operatorname{tr} \mathrm{A} M_{xx} \mathrm{A}'$$

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

$$\begin{aligned} L(\mathrm{A}) &= \text{const} + \log\det \mathrm{B} - \frac{1}{2}\log\det \mathrm{A}\, M_{xx}\, \mathrm{A}' \\ (4.251) \qquad &= \text{const} - \frac{1}{2}\log\det \mathrm{B}^{-1}\mathrm{A}\, M_{xx}\, \mathrm{A}'\, \mathrm{B}'^{-1} \\ &= \text{const} - \frac{1}{2}\log\det \begin{bmatrix} -I_{[K_y]} & \Pi_{[K_y K_z]} \end{bmatrix} M_{xx} \begin{bmatrix} -I_{[K_y]} \\ \Pi_{[K_y K_z]} \end{bmatrix}. \end{aligned}$$

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].