PSYCHOMETRIKA—VOL. 28, NO. 1
MARCH, 1963

THE USE OF FACTOR ANALYSIS IN THE STATISTICAL ANALYSIS OF MULTIPLE TIME SERIES*

T. W. ANDERSON
COLUMBIA UNIVERSITY

Factor analysis has been proposed and used as a method of statistical analysis of several measurements made on one individual repeatedly over time. This paper discusses some difficulties in applying factor analysis to multiple time series and attempts to indicate to what extent such methods can accomplish the goals of time series analysis. Some other methods are suggested.

1. *Introduction*

Factor analysis consists of a set of statistical techniques with related mathematical models that has been developed particularly for analyzing the observed scores of many individuals on a large number of psychological tests. These methods have been applied to other sets of measurements of individuals and of other entities. Its use has been further extended to analyze a set of variables referring to one individual and observed at several occasions. This is the use of factor analysis in the statistical analysis of multiple time series.

Factor analysis techniques have been used largely by psychologists, and their application to time series data has been mostly at the hands of psychologists. Cattell [4, 5] particularly has discussed this use (called *P*-technique by him). The present paper arose from comments made by the author at a symposium sponsored by the American Psychological Association, which included a general survey of this use of factor analysis by Cattell [6], a study of psychological, psychiatric, physiological and chemical measurements of a schizophrenic patient by Mefferd, Moran, and Kimble [13], and a study of biochemical characteristics of a cancer patient by Damarin [7]. (The symposium was "Multivariate Analysis of Repeated Measurements on the Same Individual," Washington, D. C., Division 5 of the American Psychological Association, September 3, 1958.) A summary of these two studies will be given in Section 2 as examples of the use of factor analysis in the statistical analysis of multiple time series. The general problem will be stated formally (Section 3), and a mathematical statement of factor analysis will be given (Section 4).

*Research sponsored by contract AF41(657)-214 between the USAF School of Aerospace Medicine and Teachers College, Columbia University.

The primary purpose of this paper is to outline the possible functions of factor analysis in time series analysis (Section 5), discuss some of the problems and difficulties that arise, and point out limitations on its usefulness (Sections 6 and 7). Since psychologists seem to be particularly favorably disposed to using factor analytic techniques on data in the form of time series, this paper is addressed especially to them. Economists, among others, have been analyzing time series for many decades. Although some of the techniques they apply are quite specialized, they have also considered techniques based on a mathematical model quite similar to that of factor analysis. Such techniques, however, have been largely displaced by other methods based on different models such as stochastic difference equations. A brief description of these methods is given (Sections 8 and 9). Although some aspects of the analysis of repeated measurements of psychological and physiological measurements are rather unique, it should be noted that there are many aspects common to time series analysis as developed in other disciplines.

2. *Two Examples of the Factor Analysis of Time Series Data*

Two substantive studies were reported to the symposium which lead to the preparation of this paper. In one study a schizophrenic patient was observed over a period of 261 days, during which time several types of shock treatment were given. Daily measurements of the patient's characteristics included 30 chemical constituents of urine and 6 physiological quantities (body weight, temperature, pulse and respiration rates, as well as systolic and diastolic blood pressures). The daily scores on a psychiatric rating scale were considered as not sufficiently variable to include in the analysis. During the last 75 days daily scores were obtained on alternate forms of six psychological tests (copying, number facility, four-letter words, digit span, and two time perceptions). During this latter period whether or not the treatment was given was indicated by a coded variable. Four environmental variables were maximum outside temperature, average outside temperature, average wind speed, and sunspot activity. The variables subjected to analysis also included the day number (furnishing a linear trend) and three other synthetic (or dummy) variables to provide cycles of two and four days. The correlations between these 51 variables were computed over the last 75 days, during which time only one type of treatment was given. The factor analysis of these correlations yielded eight orthogonal factors, which were then rotated. The rotated factors were then "identified" by the values of the loadings of the observed variables. Another factor analysis was carried out over an earlier period when a different treatment was given, and a third factor analysis was made on the data in the two periods.

The other study was made of a patient with advanced cancer of the prostate. During and after hormonal therapy, measurements were made on Mondays, Wednesdays, and Fridays over a period of about 63 weeks. The rate of tumor growth in the prostate was indicated by the titer of an enzyme (acid phosphatase), and the rate of repair of bony tissue damaged by metastacism was indicated by the titer of another enzyme (alkaline phosphatase). Other physiological, biochemical, and metabolic variables (including strength of medication for pain) were also included, as well as climatic variables and two synthetic variables to represent a linear trend and a weekly cycle. The period of study between initial therapy with estrogen and death was divided into three stages, Improvement, Remission, and Relapse, with measurements on 63 occasions in each period and on 37, 44, and 47 variables in the three periods, respectively. A factor analysis of the matrix of correlations over the 63 time points in each phase yielded 16 oblique factors. Within each phase the 16 factors were identified. Of these, 4 identified factors seemed to be common to the three phases, 5 factors appeared in pairs of phases, and the other 7 factors were distinct in each phase. The investigator related the appearance and disappearance of factors over the three stages to the course of the disease.

3. *The Data and Goals of Time Series Analysis*

To state our problem in general terms, we suppose that some p measurements $(y_{1t}, y_{2t}, \cdots, y_{pt})$ are made on a given individual at each of T successive times $(t = 1, 2, \cdots, T)$. The set of data on the individual constitutes a $p \times T$ matrix

$$(1) \qquad \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1T} \\ y_{21} & y_{22} & \cdots & y_{2T} \\ \vdots & \vdots & & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pT} \end{bmatrix}$$

We shall suppose the successive times are equally spaced, such as measurements made on successive days. (Equal spacing is advantageous in order to benefit by the time scale.) For example, in one case of the Mefferd-Moran-Kimble study $y_{24,t}$ is the score on the test of number facility on the tth day and $y_{40,t}$ is the systolic blood pressure on the tth day. These p variables characterize the individual in such a study and are, therefore, of primary concern to the investigator.

There may be a related set of q variables that we may term environmental, which are also measured at each time point t. These form another matrix

$$
\begin{bmatrix}
z_{11} & z_{12} & \cdots & z_{1T} \\
z_{21} & z_{22} & \cdots & z_{2T} \\
\vdots & \vdots & & \vdots \\
z_{q1} & z_{q2} & \cdots & z_{qT}
\end{bmatrix}. \tag{2}
$$

Some are strictly environmental, such as average daily temperature and wind speed; others may describe a treatment given an individual; and some may be synthetic or dummy variables such as time trend (that is, $z_{it} = t$) and cyclic or periodic variables (such as $z_{it} = (-1)^t$). This second set of variables describes the conditions acting on the individual.

Some economists have used the terms *endogeneous* and *exogenous* for the corresponding sets of time series in economic studies. (Cattell ([5], p. 594) uses these terms and alternatively *spontaneous* and *situational.*) The endogenous variables form the set that the system is supposed to generate; we are trying to "explain" these. For example, in an investigation of the United States economy, national income might be an endogenous variable. The exogenous variables are the set that acts on the system; these are taken as given. For instance, the size of the population of the United States might be taken as exogenous.

This distinction is not absolutely clean cut. Some variables may describe the treatment given a patient, such as the amount of a drug. The treatment acts on the individual and in that sense is environmental. However, if during the period of study the course of treatment is modified because of the development of the variables characterizing the individual, these treatment variables can hardly be regarded as exogenous. In such cases, the sequence of values of the treatment variables cannot be taken as given; it becomes part of the system under study.

The classification of dummy variables may also be questionable. In an economic study a combination of trend terms such as $z_{it} = t$ is usually taken to represent a collection of variables not treated explicitly, such as measures of technological development; it is supposed that all of the variables characterizing the system are adequately represented by the endogenous variables in the study and hence that the synthetic variables are in lieu of unspecified exogenous variables. In any case, of course, it is clear that such variables are exogenous in the sense that they are determined outside the system studied and must be taken as given. It is the interpretation that may be difficult; in psychological and physiological studies of individuals it may not be clear whether a trend term represents environmental conditions, internal quantities, or both.

A modest goal of time series analysis is to give a succinct description of the variables of interest as observed over the assigned period of time. Data on one variable at one time for several individuals are often summarized by quantities like the mean and variance. The summarization of data on several variables may include correlation coefficients. Data in time may involve other modes of summarization because observations in a time series are ordered. Some such summarization statistics are serial correlation coefficients (also termed auto-correlations or lag correlations). For example, the first-order serial correlation coefficient is the correlation between adjacent elements in a time series (that is, the pair x_t , x_{t-1} replaces y_t , z_t in the computation).

Another goal of the study of a time series is prediction. Given the sequence of one or several variables, the investigator wants to say something about the value of a variable anticipated at some point in the future. The economist may want to predict national income next year; the psychologist may want to predict the degree of elation of a patient tomorrow. Prediction may be done on an ad hoc basis, or it may involve a deeper analysis.

A more encompassing goal of time series analysis is an understanding of the mechanism generating the characteristics of interest; in other terms one wants to "explain" the sequence of observed quantities. This is usually the goal of scientific inquiry. A mathematical statistical model to specify the information of the variables under given conditions is to be desired.

In some cases an explanation or determination of a time series of a single variable can be given in terms of itself. For instance, the height of a swinging pendulum at each point in time is a function (trigonometric) of time which satisfies a second-order differential equation; that is, its position at a future time is determined by its present position, velocity, and acceleration (which are equivalent to the information given by its position at three different times). It may be that time series of some psychological variable might be explained by that series alone; perhaps a characteristic of mood is at least roughly explained by its earlier fluctuations.

In other cases several variables are tied together and the entire complex may be explained in terms of its earlier history. Given a stable environment a certain set of psychological or physiological quantities may interact in a system, the development of one quantity affecting another. In such cases the investigator seeks a mechanism explaining the entire constellation.

Often there are external factors that influence the individual characteristics. Sometimes a time series analysis may only be relating the variable of interest to the changing environmental condition. The most complex system includes exogenous variables which affect the set of endogenous variables which are themselves interacting.

4. *Factor Analysis*

Factor analysis can be considered as the analysis of data in terms of a particular mathematical model. We shall state the model mathematically for a general set of variables x_{it} , $i = 1, \cdots, h$, and $t = 1, \cdots, T$. (The variables noted in Section 3 fit into this framework with $x_{it} = y_{it}$, $i = 1, \cdots, p$, and $x_{p+j,t} = z_{jt}$, $j = 1, \cdots, q$, with $h = p + q$.) The factor analysis model is

$$x_{it} = \xi_{it} + u_{it}, \tag{3}$$

$$\xi_{it} = \mu_i + \sum_{\alpha=1}^{m} \lambda_{i\alpha} f_{\alpha t}. \tag{4}$$

The ξ_{it} are the *systematic parts* of the x_{it}. The u_{it} are *errors*; each is composed of a specific factor, which is special to the variable (that is, i), and an error of measurement. The μ_i are constants; the $\lambda_{i\alpha}$ are common factor loadings; and the $f_{\alpha t}$ are common factor scores. The basic (unobservable) variables $(u_{1t}, \cdots, u_{ht}, f_{1t}, \cdots, f_{mt})$, $t = 1, \cdots, T$, are considered to be T independent sets of observations from some population; (3) and (4) indicate how these are combined to form the observable variables. (Sometimes the set of factor scores $f_{\alpha t}$ for T particular individuals or times are considered as fixed or nonrandom variables, called *incidental parameters*, while the errors u_{it} are treated as random observations.) The essential mathematical assumptions are that the errors u_{it} are mutually uncorrelated and that the factors $f_{\alpha t}$ are uncorrelated with the errors u_{it}. The population mean values of the errors and factors can be taken as 0, and then the population mean value of x_{it} is μ_i. The observed variables are correlated (except for special choices of the factor loadings $\lambda_{i\alpha}$), and the correlation is due to the common factors.

The information in the observations is summarized by the sample means, $\bar{x}_i = \sum_t x_{it}/T$, sample variances $s_{ii} = \sum_t (x_{it} - \bar{x}_i)^2/(T - 1)$, and sample covariances $s_{ij} = \sum_t (x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j)/(T - 1)$. If the factor scores and errors are normally distributed, these quantities form a sufficient set of statistics. The sample covariance matrix is "factored" to obtain estimates of the factor loadings. This involves approximating the observed covariance matrix (s_{ij}) by an estimated theoretical covariance matrix $\sum_\alpha \lambda^*_{i\alpha}\lambda^*_{j\alpha} + \sigma^{*2}_i$ so the difference between these two matrices consists of elements close to 0. Here σ^{*2}_i is the estimate of the variance of u_{it} (the uniqueness), and the factors are taken orthogonal. Usually the procedure includes the determination of m (the number of factors) from the sample. Since there is indeterminacy in the specification of the loadings $\lambda_{i\alpha}$ and factor scores $f_{\alpha t}$, the factoring yields a set of estimates, and one of these is selected (for example, by "rotation" to simple structure) to give most meaningful factors.

Given the factor loadings one can estimate the factor scores $f_{\alpha t}$ corresponding to observed variables x_{it} for each t. The estimated factor score is usually a linear function of the observed x_{it} for that t. (If the factor scores and errors are normally distributed the linear functions may be the regression of the factor scores on the observed variables; if the factor scores are taken as incidental parameters the estimated factor scores can be maximum likelihood estimates under the assumptions that the loadings are given and the errors are normally distributed.)

There are different statistical methods of carrying out the determination of m and estimating the loadings and factor scores. Frequently, the covariances are converted to correlations and the centroid method of factoring is applied. A survey and analysis of the different techniques from the mathematical statistical point of view has been given elsewhere [3] and will not be repeated here. The primary interest in this paper is the usefulness in the statistical analysis of multiple time series of any factor analysis methods based on the above model.

5. *Three Possible Functions of Factor Analysis in the Study of Time Series*

5.1 *Measurement of Underlying Traits*

Factor analysis in the sense of an analysis of the model presented in Section 4 has been developed primarily as a means of ascertaining underlying traits or latent variables (the factors); these are characteristics of individuals that cannot be measured directly but which do affect suitably chosen variables that can be measured or observed. The model is widely accepted as adequately representing reality in certain areas of psychology. This is particularly true in the case where the observed variables are scores on various kinds of intelligence tests and the unobserved variables are factors representing mental abilities. In such cases, the observed data together with a statistical analysis appropriate to the model may yield measurements of the underlying traits. It might be pointed out that in using this approach to evaluate underlying traits the psychologist designs and chooses the instruments (tests, questionnaires, etc.) to tap the traits in which he is interested.

The estimate of a factor score or measurement of a trait of a given individual at a given time comes out as a linear combination of the observed test scores of that individual at that time. A time sequence of the factor scores of one individual can be estimated by taking this linear combination of the test scores observed at a sequence of time points. This time series of estimated factor scores can be analyzed by itself, or it can be included in an analysis with other time series. This use of a combination of other variables is analogous to the economist's use of an index, such as a price index. As is known to econometricians, the use of an index poses its own statistical problems, but this is not the place for a discussion of the indices used in economics.

The selection of the traits or factors and the determination of the linear combinations of observations to estimate them can be made from repeated observations on one individual or from observations at one time on many individuals. If certain specified linear combinations of scores of certain specified tests are accepted a priori as measuring certain traits, it would seem reasonable to use such measures in a time series analysis. Of course, this recommendation is qualified by the requirement that the instruments are not substantially affected by time (practice, fatigue, etc.), but this difficulty will arise with any repeated use of such instruments. Cattell ([5], p. 896, for example) has suggested the use of "marker variables" in studies involving factor analysis; these variables are tests which are taken to have high loadings on certain desired factors and hence assist in obtaining measurements of specified factors.

A set of tests or other instruments observed for one individual over time may not show the same pattern of variation as when observation is made on many individuals at one time. Cattell has distinguished factors found from the two types of analysis by calling an underlying characteristic varying over individuals as a trait and an underlying characteristic varying over time as a state. In some cases the former may give more relevant measures, but not necessarily. It may be that an important part of the time series analysis is to find an individual's time pattern of a measurement specified in advance.

The use of factor analytic methods to establish the measurement of an underlying characteristic is only a preliminary step in a time series analysis; it simply determines a means of converting a time series of multiple scores on certain instruments into a time series of quantities considered more fundamental. The development of these quantities in time and their relation to other variables is then the subject of further inquiry.

5.2 *Exploration*

The term "factor analysis" is often used to mean any kind of analysis of a large set of variables which leads to the selection or formation of a smaller set of variables for more intensive study. This is particularly true of statisticians not well acquainted with statistics as used in psychology. In this loose sense factor analysis does not necessarily mean a procedure based on the model of Section 4. In particular, principal component analysis (see Chapter 11 of [1], for example) can be viewed as a method of exploration.

In many situations where a large number of measurements are made there is little or no knowledge of the variables involved or the underlying mechanism. In these cases it may be appropriate to use an omnibus statistical technique to explore the data, that is, to find the variables or functions of variables that show variability or that are related. In these instances the exploratory study is only a preliminary stage of investigation; it points the way to more basic study. In many types of inquiries concerning individuals there is the possibility of using the results of one study in the study of other individuals; the psychologist can search the data of one individual for apparently relevant variables and then study these variables for other individuals. This feature is in constrast to the economist studying the entire economy of the United States and hence being limited to one set of data.

Besides the various forms of factor analysis there are other methods of exploration. If two distinct sets of variables are under consideration, canonical correlation analysis may be more appropriate (see Chapter 12 of [1], for example). If some form of factor analysis is used, care must be exercised in selecting the procedure. There are questions whether to use correlations or covariances and how to treat communalities. Some methods are based on the idea that factors account for the dependence between observed variables as indicated in Section 4. Some methods are based on the idea that "factors" account for the variances, as with principal component analysis. One method may give results quite different from those given by another method.

As shall be noted in more detail later, factor analysis does not take into account the time relationships of the variables. If factor analysis or principal component analysis is used for exploration, there is a danger of missing important and interesting characteristics which are significant because of their development in time though not because of their variability or high relationship to other variables.

If factor analysis is used in exploration, it may be reasonable that it be applied only to a set of data of the same kind, say intelligence test scores, or personality variables, or physiological measurements. If the investigator wants to study several sets together, he can explore one and then another or he can use another method, such as canonical correlations.

If it is known that a particular variable is of special importance to the investigation, it should not be included in an exploratory factor analysis, because whatever is the outcome of the exploration, this particular variable must be included in the ensuing time series analysis. For example, in Damarin's study of cancer therapy [7] the treatment variable had to be studied, and since acid phosphatase enzyme titer is a generally accepted index of tumor growth it has to enter the complete analysis, too.

An exploratory factor analysis is only a preliminary phase of a time series analysis. Once the investigator has decided what the important variables are, he will want to know the interrelations among them and the mechanism of their formation.

5.3 *Analysis*

Measurement and exploration are only introductions to understanding processes in time. The scientist wants to establish a mechanism (which may be a mathematical model) which explains or describes the generation of the phenomena of interest. In considering characteristics of individuals over time, the investigator would like to discover how one phase leads to another or how one sequence of variables affects another sequence. In fact, eventually he wants to be able to say how the development of the variables is brought about. Wold [14] discusses these goals in terms of the analysis of multiple economic time series.

In the next two sections we consider the questions of the extent to which factor analysis can be used to elucidate the formation of the sequence of variables and the problems that arise in this use.

6. *Some Problems in the Factor Analysis of Time Series Data*

6.1 *Independence of Errors*

The error terms u_{it} in (3) are assumed to be statistically independent; that is, the discrepancies between the observations x_{it} and the systematic parts ξ_{it} are assumed uncorrelated. If one were able to make repeated measurements on the entire matrix, there would be no (population) correlation in the repeated measurements on the pair (u_{it}, u_{jt}), the errors in two different variables at one time, or in the repeated measurements on the pair (u_{it}, u_{is}), the errors in one variable at two different times, or in the repeated measurements in the pair (u_{it}, u_{js}), the errors in two different variables at two different times. This kind of assumption is reasonable if the discrepancies are really errors of measurement, such as might occur if one made physically independent measurements with a tape measure. In the situations we are considering, however, it seems likely that the discrepancies would include irregularities due to a multitude of influences and effects that are not taken explicitly into account. Some of these, the daily weather, for example, might tend to correlate (u_{it}, u_{jt}) by acting similarly on several variables from day to day. The effect might be, however, to correlate (u_{it}, u_{is}) because the effect is spread over several days. The day-to-day correlation may be of no greater disadvantage than making the information in the series of T times less than if the observations were independent, but the variable-to-variable correlation may be serious enough to bias critically the analysis, because the entire principle of factor analysis is that only the common factors tie the observed variables together.

The model is set up with interest centered in the systematic parts, the random irregularities being added to them. In many cases, however, it may be more realistic to think that the random irregularities are absorbed into the quantity of interest and, hence, affect the observations at a later time. This point will be clarified in Section 9 by displaying a different kind of model. The usual assumption that f_{it} and f_{is}, $t \neq s$, are uncorrelated would not be fulfilled in time series data, but this effect would probably not be serious.

An effect of time that is more peculiar to psychological studies is the effect of repeated uses of the same instruments. The individual learns to do certain tasks, for instance, if he is asked to do them again and again.

This challenge to the psychologist arises, of course, because of the time aspect, not because of the use of factor analysis.

6.2 *Symmetrical Treatment of Variables*

The method of factor analysis treats all observed variables symmetrically; this feature appears even in the statement of the model. As indicated in Section 3, however, the observed variables can usually be divided into two classes, those which the system is to "explain" and those which are to be taken as given. Although there are some difficulties of classification, the distinction is meaningful. The latter set may affect the former, but not vice versa. In the Mefferd-Moran-Kimble study, for example, it is possible that sunspots affect the patient's performance on psychological tests, but the reverse does not seem likely. A deficiency of factor analysis here is that it does not take into account the different statuses or functions of different variables; in particular, a direction of possible causation is missing.

Another classification of variables is according to whether they are measured with or without actual error. For example, a synthetic variable has no irregular component, whereas a psychological test may have considerable error. It would seem more realistic in the model to set the formal "errors" $u_{it} = 0$ for variables with no irregularity. The method of factor analysis as customarily used does not allow for this difference. The anomaly is that a variable which has no real error shows up with a formal "error." How can one interpret an error in time ($x_{it} = t$) or a dichotomous variable ($x_{jt} = 0$ or 1) indicating absence or presence of treatment? This occurs, for example, in the Mefferd-Moran-Kimble study. The discrepancy between such a variable as observed and as it enters mathematically into the fitted model is properly considered as an error in the model, but such an error does not have the statistical properties assumed by the factor analysis model.

The anomaly of obtaining a formal error u_{it} for a variable that has no error of measurement and cannot be considered as having a distribution can be avoided by setting $u_{it} = 0$ for the variable and adjusting the statistical procedure appropriately. This matter will be discussed in Section 7. Essentially these restrictions force the use of regression procedures for such variables without distributions rather than the correlational procedures of factor analysis.

Besides the two classifications of the variables mentioned above, there may be natural groupings of variables. For example, the Mefferd-Moran-Kimble study includes scores on psychological tests and constituents of urine. These are different, at least in a formal sense. The biochemical measurements are made in physical units such as density or percentage, while the units of psychological scores may be quite arbitrary. Measurements made in the same units are directly comparable and permit statistical techniques that take account of the origins and units of measurement.

6.3 *Interpretation of Factors*

The factor analysis yields the representation

$$x_{it} = \sum_{\alpha=1}^{m} \lambda^*_{i\alpha} f^*_{\alpha t} + \mu^*_i + u^*_{it} , \tag{5}$$

where now the $\lambda^*_{i\alpha}$, μ^*_i and $f^*_{\alpha t}$ are estimates based on the observed data and u^*_{it} is the residual. "Rotation" to simple structure may have been done in hopes of giving the factors an interpretation. Do these factors have any meaning other than what is given by the formal statistical technique applied? It is desired that these $f^*_{\alpha t}$ represent some fundamental quantities, just as in the factor analysis of intelligence tests scores the factors are considered as primary intellectual abilities.

The loadings for a given factor may be such that the factor relates (i) only to the observed individual characteristics, (ii) only to the environmental characteristics, or (iii) to both the individual and the environmental characteristics. In the first case the factor would be interpreted as a quantity fundamental to the individual and in the second case to the environment. In the last case is the factor considered as fundamental in both kinds of variables? In some instances this may be satisfactory, in others not. Mefferd, Moran, and Kimble, for example, found that Factor I (in Case *D*-1) has high loadings for sunspot activity and for some biochemical measurements. This fact may imply that sunspot activity is correlated with biochemical measurements, but it does not yield any more basic explanation.

When all pairs of factors f_{it} and f_{jt} $(i \neq j)$ are uncorrelated as assumed in Section 4, they are called *orthogonal.* If these conditions of being uncorrelated are not imposed, the factors are called *oblique.* In the case of orthogonal factors, alternative sets are related by orthogonal transformations; alternative sets of oblique factors are linear transformations of each other. In either case such a transformation is called a *rotation.* (The use of the term rotation for all nonsingular linear transformations is an undesirable conflict with mathematical and everyday usage of the term rotation.) In the case of an orthogonal factor system, factors that relate only to environment are really irrelevant because these factors would necessarily be uncorrelated with the individual characteristics upon which the study is presumably focussed. If the orthogonal factor system consists of some factors involving only individual characteristics and other factors involving only environmental characteristics, the conclusion is that these two sets of variables are independent. As will be pointed out in more detail in Section 9.2, however, factor analysis is not needed for reaching such a conclusion and is not the most efficient statistical procedure for ascertaining independence. If the two sets are independent, then the set of environmental variables can be dropped from the study because they do not affect the set of individual variables in which the investigator is primarily interested.

Another possibility to consider is that of an oblique factor system with some factors relating to individual characteristics and some relating to environment. The correlations between the factors of two types will connect the two sets of observed variables. The pattern of factor correlations may give the investigator insight into the relationship between individual and environmental characteristics. (If the investigator analyzes these first-order factors into a set of second-order factors, the problems all arise again at the level of second-order factors.)

The consideration of factors in studies involving two or more different kinds of variables suggests that the determination of the best factor system might take this feature into account. "Rotation to simple structure" involves transforming a set of estimated loadings $\lambda_{i\alpha}^*$ by a transformation

$$\sum_{\alpha=1}^{m} \lambda_{i\alpha}^* a_{\alpha\beta} = \lambda_{i\beta}^{**},$$

say, and simultaneously $f_{\alpha t}^*$ by

$$\sum_{\alpha=1}^{m} a^{\beta\alpha} f_{\alpha t}^* = f_{\beta t}^{**},$$

say, where the matrix $(a^{\beta\alpha})$ is inverse to $(a_{\alpha\beta})$. The matrix $(a_{\alpha\beta})$ is chosen to give a matrix of loadings $(\lambda_{i\beta}^{**})$ that is meaningful in the sense that the implied factors can be interpreted substantively and is basic in the sense of invariance under selection. (Factorial invariance means that the definition of $\lambda_{i\beta}^{**}$ would not be changed if x_{jt}, $j \neq i$, were deleted or added or if the range of measurements (t here) were modified.) One criterion for simple structure is to make as many $\lambda_{i\beta}^{**}$ equal to zero as possible (or at least near zero). When there are two different sets of variables the investigator might obtain a more meaningful interpretation of the factors if for some β's he tried to make $\lambda_{i\beta}^{**}$ close to zero for the i corresponding to one set of variables and for other β's he tried to make $\lambda_{i\beta}^{**}$ close to zero for i corresponding to the other set.

6.4 *Time Sequence Effect*

As noted in Section 4, any method of factor analysis starts with the mean $\bar{x}_i = \sum_t x_{it}/T$ of each variable, the variances $s_{ii} = \sum_t (x_{it} - \bar{x}_i)^2/(T - 1)$, and the covariances $s_{ij} = \sum_t (x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j)/(T - 1)$; sometimes this information is reduced to the correlations $r_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}$. The inferences about the factor loadings $\lambda_{i\alpha}$, that is, the determination of m and the estimation of the loadings, is done on the basis of the covariance matrix (s_{ij}) or the correlation matrix (r_{ij}). These sums involve products of variables at the same time (x_{it} and x_{jt}) but not at different times, x_{it} and $x_{j,t-1}$, for example. Thus the timing of the variables is lost, and the relations between the variables at different times are absent from the summarization. To put it another way, given the variables x_{it} one could renumber the index t and the analysis on the newly ordered variables would be the same. However, the significant feature of time in time series analysis is the ordering.

The effect of time may come into this analysis indirectly in the sense that some of the x_{it} are closely and obviously related to time. The extreme cases occur when some x_{it} are mathematical functions of time such as t itself or a periodic function (a cycling variable). Thus the timing of two individual characteristics may be seen in part and indirectly in their relationships to t or other functions of t.

The estimated factor loadings may be used to estimate the factor scores $f_{\alpha t}$ (as mentioned in Section 5.1). These estimated factor scores form a multiple time series which show the time sequence and may be submitted to further analysis.

6.5 *Goal of Analysis*

If a factor analysis of some time series data has been done and the "rotation to simple structure" yielded factors that seem to have sensible interpretation, would this complete the analysis of the time series? On the basis of the factor analysis, one "explains" the observed variables by factors, which vary through time. These factors in themselves, however, are not necessarily ultimate. In most cases the investigator will not be satisfied until he has explained, or at least described, the structure of the movement of the factors over time.

In time series analysis we first want a succinct statistical description of the variation, relations and development of the relevant time series data; we want to distinguish the irregular or random manifestations from the systematic relations. (The random effects do not necessarily need to be simply independent additive effects.)

Secondly, we would like to represent the regularities of the data by as simple a model as possible to yield meaningful interpretations in the particular discipline. The desired objective is a statistical or probabilistic model of the generation of the series that is basic enough so that (at least for a time) it may be considered as an "explanation" of such kinds of series.

In the case of a pendulum the mathematical model is a trigonometric function (a sine or cosine of a multiple of t) or a second-order differential equation; this trigonometric function describes the observed phenomena, and the differential equation with the notion of gravity is an explanation. There may be series of psychological measurements that have a similar analysis; the phenomena are roughly periodic but with random disturbances. For instance, a person's mood (elation or depression) may be of this kind, and its fluctuations in time may be described simply (for example, as a second-order stochastic difference equation; see Section 9).

In other cases a somewhat similar but more complicated analysis might be possible. By way of illustration, suppose that Mefferd, Moran, and Kimble had observed the following phenomena: One day a shock treatment was given a patient (say, $z_{1t} = 1$ for treatment and $z_{1t} = 0$ for no treatment); the next day some biochemical measurement of the patient (y_{1t}) increased; the following day a measure of the patient's anxiety (y_{2t}) diminished; and on the next day, performance on a psychological test (y_{3t}) went up. Such a sequence of events would have the simple interpretation that the shock affects the body chemistry, which in turn affects the emotional state, which in turn affects mental performance. Given the multiple time series (y_{1t} , y_{2t} , y_{3t} , z_{1t}), the investigator would like a method of analysis that would discover and emphasize the time relations mentioned above. Further, he would like a simple statistical model to describe these relationships. The interpretation of such a model would be in terms of the physiological and psychological mechanisms.

7. *Some Problems of Adapting Factor Analysis to Apply to Time Series*

7.1 *Observed Variables as Factors, Trends*

In many time series studies functions of time are included, as in the two substantive studies described in Section 2. These functions may be powers of t (yielding polynomials) and cyclical functions (often expressed as sines and cosines). Such variables have no real errors of measurement, and it is hard to give any meaning to the assignment of a specific factor to such a variable (unless the variable itself is a specific factor; that is, uncorrelated with all the other variables). Accordingly, the model defined by (3) and (4) should be correspondingly restricted. For each i that denotes such a variable (measured without error and not decomposable into a systematic part and specific factor) the formal error term u_{it} should be set equal to 0. Equivalently, the variance of such a term is set equal to 0 (that is, in correlation terms the "communalities" are 1). Let us see the implications of such restrictions.

For illustrative purposes, suppose there is one variable, say x_{ht} , without formal error, and set $u_{ht} = 0$. Then

$$x_{ht} = \mu_h + \sum_{\alpha=1}^{m} \lambda_{h\alpha} f_{\alpha t} \, . \tag{6}$$

Solving (6) for f_{mt} , we obtain

$$f_{mt} = \frac{1}{\lambda_{hm}} \left(x_{ht} - \mu_h - \sum_{\alpha=1}^{m-1} \lambda_{h\alpha} f_{\alpha t} \right); \tag{7}$$

that is, one factor is represented by a linear combination of the error-free observed variable and the other factors. Put this expression for f_{mt} in (4) to obtain

$$(8)\qquad \xi_{it} = \left(\mu_i - \frac{\lambda_{im}}{\lambda_{hm}}\mu_h\right) + \sum_{\alpha=1}^{m-1}\left(\lambda_{i\alpha} - \frac{\lambda_{im}\lambda_{h\alpha}}{\lambda_{hm}}\right) f_{\alpha t} + \frac{\lambda_{im}}{\lambda_{hm}} x_{ht}\,,$$

$$i = 1, \cdots, h - 1.$$

This is an appropriate model for the other $h - 1$ observed variables; the systematic part consists of the constant plus a linear combination of the $m - 1$ factors plus a multiple of the error-free observed variable. This last variable, x_{ht}, enters the equations as an independent regression variable.

What is the appropriate statistical analysis for this model? For the sake of statistical analysis we can take a convenient specification of the "rotation" (to eliminate some indeterminacy in the $\lambda_{i\alpha}$ and $f_{\alpha t}$) in finding the estimates of the factor loadings. The estimated factor system can then be transformed to another system to facilitate interpretation. We shall take f_{mt} to be orthogonal to the other factors (or even take all the factors to be orthogonal) and take $\lambda_{h\alpha} = 0$ for $\alpha = 1, \cdots, m - 1$. (From any set of $\lambda_{i\alpha}$ there is a rotation to a new set satisfying these restrictions.) Then $x_{ht} = \mu_h + \lambda_{hm} f_{mt}$; the last factor is simply the last variable with a suitable origin and unit. In (8) x_{ht} is therefore orthogonal to $f_{\alpha t}$, $\alpha = 1, \cdots, m - 1$. Thus

$$\xi_{it} - \frac{\lambda_{im}}{\lambda_{hm}} x_{ht}$$

has a factor structure with $m - 1$ factors. If $\lambda_{im}/\lambda_{hm}$ were known, a factor analysis method could be applied to

$$x_{it} - \frac{\lambda_{im}}{\lambda_{hm}} x_{ht}\,.$$

Not knowing these parameters, one estimates them by taking the (sample) regression of x_{it} on x_{ht}, for (8) indicates the usual regression situation in which x_{ht} is orthogonal to the (unknown) $f_{\alpha t}$. The estimation of the regression coefficient of x_{ht} is then done directly.

In general, the procedure is as follows: Take the ordinary regression of the observed variables with error on the error-free variables (particularly trend variables). Submit the residuals of the observed variables with error from their regression to a factor analysis procedure. The error-free variables are factors—at least formally, and alternative factor systems consist of linear combinations of the error-free variables with the factors resulting from the factor analysis of the residuals.

One alternative procedure is to submit the original set of h variables to a factor analysis without imposing the restrictions that certain u_{it} are 0 and then rotating the factor structure obtained so that some of the estimated factors are approximately the error-free variables. In each of the two investigations reported in Section 2 the rotation to simple structure gave a factor that had a high loading for the time variable (that is, t), the factor being interpreted as the time variable; in the Mefferd-Moran-Kimble study, another factor had a high loading for the cyclic variable of period two (namely, $(-1)^t$). From the point of view of statistical inference, whatever factor analysis method is used in this procedure is inefficient because it is based on the allowance of errors u_{it} which could have been assumed 0. The interpretation of the factors requires the incongruity of the decomposition of time or other trend variables into two parts.

Another alternative is to impose the restrictions that certain u_{it} are 0 and apply factor analysis under these conditions. In this case, the efficiency of statistical estimation would depend on the efficiency of the factor analysis method. The centroid method of factoring would not be as efficient as regression analysis; on the other hand the maximum likelihood method when conditions are imposed and the distributions are normal would be equivalent to taking the regression and applying maximum likelihood to the factor analysis of residuals.

Generally speaking, the more that regression analysis can be used before factor analysis, the better. Methods of regression analysis are simple and efficient. The variables are explicit as opposed to factors which are implicitly or indirectly obtained. The statistical methods are highly developed; for instance, one can test the hypothesis of a regression coefficient being 0, which is equivalent to the hypothesis that an observed variable is independent of the time variable, or one can construct a confidence interval for a coefficient.

The factor analysis model defined by (3) and (4) is formally a model indicating the regression of the observed variables on the factors. The difference between regression analysis and factor analysis is that in the former the effecting variables are known instead of being unknown factors. It seems clear that use of this knowledge must make regression analysis much more efficient. These properties are reflected by the fact that regression procedures are not based on an assumption that errors in different variables are uncorrelated.

Now let us turn to the question of interpretation when some factors are identical with explicit variables (trend variables or individual or environmental characteristics) or are approximately the same. If these are the only factors, then we have "explained" our observations by the effect of these explicit variables. The resulting statistical model is that some observed variables are linear combinations of other observed variables plus errors. This is the usual regression model with the possible difference that after the factor analysis the errors in different variables are uncorrelated. For this model, however, regression techniques are more powerful than factor analysis techniques. One advantage of factor analysis is that its usual procedures can lead directly to establishing this type of model, whereas this model would be more difficult to determine by regression methods if possible explanatory variables were not known in advance.

If these observed variables, which are approximately the same as factors, are genuinely environmental, the investigator should be satisfied with the meaning of the model. If some of the variables are dummy or synthetic variables, the interpretation is less direct and the model is more descriptive and less explanatory.

In the above case the factors can be taken as oblique in order to coincide as well as possible with observed variables. In case some factors are approximately observed variables and some are not, there is a question whether the latter set of factors should be taken orthogonal to the former. If so, the second set of factors "explains" what is not "explained" by the observed variables-factors. This approach has the advantage of exploiting to the fullest the explicit variables. If orthogonality is not enforced, one has the problem of interpreting the relation between the observed variables-factors (at the level of direct observation) and the other factors (at a level of being inferred). In ordinary regression analysis there is difficulty in interpretation when the "independent" variables are correlated. For instance, in the case of two such variables the regression function can be written in terms of the two variables as given, or in terms of the first variable and the part of the second orthogonal to the first, or in terms of the second and the part of the first orthogonal to the second.

7.2 *Selection of Lagged Correlations*

Straightforward factor analysis takes into account only simultaneous relations between variables since it is based on the matrix of variances and covariances s_{ij} . Other summarization statistics that might be considered are the lagged covariances

$$\frac{1}{T}\sum_{t=1}^{T-s}(x_{it}-\bar{x}_i)(x_{j,t+s}-\bar{x}_j), \qquad s = \pm 1, \pm 2, \cdots . \tag{9}$$

(Some variations in the definition involving the range of summation (that is, "end effects") and division by the number of terms summed instead of T are ignored here.) These can be converted into lag or serial correlations. In order to take into account relations between two variables at different times it has been suggested ([5], p. 678, for example) that for each pair of variables one consider the covariances or correlations with different lags and select the numerically largest one; the factor analysis is then carried out on the $h \times h$ matrix consisting of these selected lagged covariances (or correlations).

This procedure takes some account of the timing of variation of the several variables, but only to a limited extent. Only the relationship between two variables at one interval of time is used, while the relationship may exist over several units of time. The most striking aspect of this limitation is that a variable is not even related to its own earlier values. In simple cases of time series analysis (such as the pendulum or perhaps fluctuations in mood) a sensible and useful analysis can be made of a single series where lagged relationships are studied. In multiple time series analysis considering the effect of some variable on another, the effect may depend on the rising or falling of the second; that is, the effect may depend on the rate of change, but this involves the second variable at two relative times. An effect on a variable that involves the physiology of an individual may take several days to wear itself out. Relating two variables at only one interval of time may neglect considerable important information.

A second difficulty in this procedure is that of interpreting the results. The model specified in (3) and (4) is no longer applicable, and in many cases it cannot be simply modified to be applicable. As an example, consider three variables. The model

$$\begin{aligned} x_{1t} &= \lambda_{11} f_{1t} + u_{1t} , \\ x_{2t} &= \lambda_{21} f_{1,t-1} + u_{2t} , \\ x_{3t} &= \lambda_{31} f_{1,t-2} + u_{3t} \end{aligned} \tag{10}$$

might be applicable if the maximum lag correlation for the first two variables is between $x_{1,t-1} = \lambda_{11} f_{1,t-1} + u_{1,t-1}$ and x_{2t} (both having $f_{1,t-1}$), for the first and third is between $x_{1,t-2} = \lambda_{11} f_{1,t-2} + u_{1,t-2}$ and x_{3t} , and for the last two is between $x_{2,t-1}$ and x_{3t} . Such a model, however, will not work if the maximum lag correlations turn out to be between x_{1t} and x_{2t} , $x_{1,t-1}$ and x_{3t} , and $x_{2,t-3}$ and x_{3t} because the patterns of lags do not fit together. The first pair suggests that x_{1t} and x_{2t} involve f_{1t} ; the second pair suggests that x_{3t} involves $f_{1,t-1}$ since $x_{1,t-1}$ involves this; but the third pair suggests that x_{3t} involves $f_{1,t-3}$ since $x_{2,t-3}$ involves this. These involvements are incompatible. This example simply illustrates the difficulties that arise because the proposed statistical procedure is not based on a consistent mathematical model, explicitly stated.

Holtzman [10] has further criticized the procedure on the grounds that searching a number of covariances for the largest may lead to spurious or at least distorted relations simply because some large covariances may turn up by chance. It might also be pointed out that lagged covariances can have considerable sampling error.

8. *Models with Systematic Parts and Errors*

The factor analysis model specified by (3) and (4) is mathematically equivalent to a model considered by econometricians and mathematical statisticians. Equations (4) can be considered as defining an m-dimensional hyperplane in the $(p + q)$-dimensional space of the ξ_{it} (one point for each t). For example, if there are 3 measurements and one factor, this is a line in the usual 3-dimensional space. The factor scores are coordinates in the m-dimensional space. One can take $(p + q - m)$ linear functions of the ξ_{it} and eliminate the $f_{\alpha t}$, say

$$(11) \qquad \sum_i \beta_{\nu i}\xi_{it} = \sum_i \beta_{\nu i}\mu_i , \qquad \nu = 1, 2, \cdots , p + q - m.$$

These linear equations also define the m-dimensional hyperplane. For example, if $p + q = 3$ and $m = 1$, the factor analysis model includes

$$(12) \qquad \begin{aligned} \xi_{1t} &= \mu_1 + \lambda_{11}f_{1t} , \\ \xi_{2t} &= \mu_2 + \lambda_{21}f_{1t} , \\ \xi_{3t} &= \mu_3 + \lambda_{31}f_{1t} . \end{aligned}$$

Elimination of f_{1t} can be done to give

$$(13) \qquad \begin{aligned} \lambda_{21}\xi_{1t} - \lambda_{11}\xi_{2t} &= \lambda_{21}\mu_1 - \lambda_{11}\mu_2 , \\ \lambda_{31}\xi_{1t} - \lambda_{11}\xi_{3t} &= \lambda_{31}\mu_1 - \lambda_{11}\mu_3 . \end{aligned}$$

The estimation of the $\beta_{\nu i}$ has sometimes been called "confluence analysis." (See [8].) Considerable research on this problem has been done by statisticians under such terms as "estimating structural relations." Formally, the problem of estimating the factor loadings is equivalent to that of estimating the coefficients of the structural relations; that is, a set of factor loadings determines a set of coefficients (with indeterminacy) and vice versa. However, there have been different points of view in the two lines of research. The number of factors plus the number of structural equations is the number of variables. Factor analysts look for a small number of factors. In confluence analysis attention is given to a small number of equations, usually one. (A small number of equations can be treated only if further conditions are assumed; otherwise the indeterminacy is too great). While the formal equivalence of the two approaches is occasionally acknowledged, little has been done to relate explicitly the multitude of results.

The formal difference between a linear space defined "parametrically" as in (4) and by equations as in (11) reflects a difference in interpretation between psychologists and economists. The latter interpret an equation as describing the behavior of some group in the economy; for instance, one equation may describe the behavior of consumers.

We shall not pursue this comparison further nor indicate further developments along this line because a different approach to time series will be discussed in the next section.

9. *Some General Methods of Time Series Analysis*

9.1 *Single Time Series*

First let us review some methods of univariate time series analysis. In many ways the most satisfactory analysis of a single time series, say x_t , is in terms of itself. A first-order stochastic difference equation (or "autoregressive scheme"),

$$y_t = \alpha y_{t-1} + u_t \, , \tag{14}$$

indicates that the observation at time t is made up of a part from the last observation and a random disturbance, where the random disturbances u_t are uncorrelated. One can think of y_1 as starting the sequence; then y_2 is generated from y_1 and u_2 ; and in turn y_t is generated from y_{t-1} and u_t . The effect of the time sequence comes into this model by each y_{t-1} affecting the next; often the parameter α has a meaningful interpretation; and the irregular part u_t is incorporated into the observed variable. (An error of measurement could be superimposed.)

The second-order stochastic difference equation,

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + u_t \, , \tag{15}$$

is a probabilistic analogue of the second-order differential equation. A time series generated by this model will show oscillations or fluctuations of varying length, the average length depending on α_1 and α_2 . This is the model for the swinging pendulum being given random pushes u_t at regular intervals. This model might be suitable for some psychological characteristic that fluctuates, such as mood.

A stochastic difference equation model is formally similar to a regression model; for example, (14) is similar to

$$y_t = \beta z_t + u_t \, , \tag{16}$$

where y_t is the "dependent" variable and z_t is the "independent" (or fixed) variable. The statistical methods for the difference equation (estimation of α, etc.) are similar to the methods of regression; in fact, the large-sample theory is almost identical. In (16) z_t "explains" y_t , and in (14) y_{t-1} "explains" y_t .

Another kind of model is the moving average. In the simplest case

$$y_t = \alpha_1 u_t + \alpha_2 u_{t-1} \, , \tag{17}$$

where the u_t are unobservable uncorrelated disturbances. The two models can be combined by defining the disturbance of the stochastic difference equation as a moving average. The combined model (with certain restrictions) generates a wide variety of stationary stochastic processes.

A *stationary stochastic process* is a time series in which the statistical properties of observations in one interval of time are the same as in another interval of time of the same length. More precisely, the distribution of y_t is the same as that of y_s ; the joint distribution of y_t and y_s is the same as that of y_{t+r} and y_{s+r} ; etc. In particular, the population mean and variance of y_t do not depend on t, and the covariance (or correlation) between y_t and y_s depends only on $t - s$. A succinct summarization of such a time series is the sample mean, the sample variance and the set of observed serial covariances or correlations, the latter called the *correlogram.* The *periodogram* or *spectral function* is a mathematical (Fourier) transform of the correlogram.

Some stationary time series can be represented as

$$y_t = \sum_{g=1}^{m} (A_g \cos \lambda_g t + B_g \sin \lambda_g t), \tag{18}$$

where the A's and B's are uncorrelated random variables with A_g and B_g having mean 0 and the same variance and the λ's are constants with $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_q < \pi$. The spectral distribution function $F(\lambda)$ is the sum of the variances of the A_g or B_g corresponding to $\lambda_g \leq \lambda$. A broader representation is obtained by replacing the finite sum in (18) by an infinite sum, and a completely general representation is obtained by replacing the sum by an integral. Under certain conditions the spectral distribution function $F(\lambda)$ can be written as the integral of a nonnegative function $f(\lambda)$, which is called the *spectral density.*

The interpretation of the spectral analysis of a time series can perhaps best be clarified by analogy. When light is passed through a prism, it is spread out into a spectrum. If the source is an incandescent element, the spectrum consists of a finite number of bright lines. The mathematical representation of this light is (18) where t is now continuous time. The brightness of the spectral line at frequency λ_i is a function of $A_i^2 + B_i^2$. Another source of light might produce a smooth spectrum with some brightness at all frequencies in the range. The energy over an interval of frequencies can be given as a function of the integral of the spectral density over this interval.

The statistical models presented here (i) are univariate, (ii) are characterized by lagged covariances (or variances and lagged correlations or equivalently the spectral function), and (iii) "explain" the series internally, that is, in terms of a mechanism involving this one quantity. These models have been studied extensively by mathematicians, and statistical procedures have been worked out fairly well, at least on the basis of large samples. Hannan [9] gives a good summary of the methods of statistical analysis at a somewhat sophisticated mathematical level; an introduction [2] requiring less mathematical background is being prepared.

The effect of outside influences can be incorporated into these models.

The first-order stochastic difference equation can include a regression variable

$$y_t = \alpha y_{t-1} + \beta z_t + u_t \,. \tag{19}$$

The endogenous variable at a given time is affected by the next previous value of it and by an exogenous variable. This model can be extended to any number of lags and any number of external variables.

9.2 *Several Time Series*

The univariate models can be generalized by replacing the single variables by vectors. For example (19) can be replaced by

$$y_{it} = \sum_j \alpha_{ij} y_{j,t-1} + \sum_k \beta_{ik} z_{kt} + u_{it} \,, \qquad i = 1, \cdots, p. \tag{20}$$

Here the u_{it} are independent from one time to the next though not necessarily from one variable to the next. The system permits any endogenous variables at $t - 1$ and any exogenous variables to affect each endogenous variable at time t. Some of the coefficients α_{ij} and β_{ik} may be 0, indicating no effect. This model can be generalized further to allow for interactions between variables at the same time as

$$\sum_j \gamma_{ij} y_{jt} = \sum_j \alpha_{ij} y_{j,t-1} + \sum_k \beta_{ik} z_{kt} + u_{it} \,, \qquad i = 1, \cdots, p. \tag{21}$$

When $\gamma_{ii} = 1$ and $\gamma_{ij} = 0, i \neq j$, we obtain (20). Of course, more lags can be included, and z_{kt} can denote a lagged exogenous variable.

These models are flexible and dynamic. Exogenous variables can be included, and the effect of lagged endogenous variables can be extended as far in the past as desired; the models can generate fluctuations. The models are explicit; they indicate relations between directly observed variables. The models are usually easy to interpret, and the coefficients can be given substantive meaning.

The application of these models requires some a priori knowledge of the phenomena being studied. Endogenous variables must be distinguished from exogeneous variables in order to determine the equations (20). Also a limit must be put on the number of lags to be included. Of course, the longer the observed series the more terms can be included. Exclusion of a term in an equation, either $y_{j,t-s}$ for specified j and lag s or z_{kt} for specified k, can be determined by testing the null hypothesis that the coefficient of that variable is 0. Whether the entire set of y's are independent of a given z or a set of z's can be decided by testing whether the corresponding β's are 0.

Econometricians have studied these models considerably and have developed appropriate statistical methods. Unfortunately, there is no straightforward, easily understandable, and relatively complete exposition of these models of multiple time series analysis, but a fairly good introduction with an econometric slant is available in [11]. A more elementary discussion has been given in economic terms by Klein [12]. Many papers have been published in the *Annals of Mathematical Statistics*, *Econometrica*, *Biometrika*, and *Journal of the Royal Statistical Society*, Series B.

Spectral analysis can also be extended to multiple time series. The generalization of (18) is

$$y_{it} = \sum_{g=1}^{m} (A_{ig} \cos \lambda_g t + B_{ig} \sin \lambda_g t), \tag{22}$$

where the sets of random variables $(A_{1g}, \cdots, A_{pg})$ and $(B_{1g}, \cdots, B_{pg})$ are statistically independent with common covariance matrix and are independent of other A's and B's.

9.3 *Time Series of Psychological Quantities*

One aspect of psychological statistics that differs substantially from others is that several observed variables are often used to obtain indirect measurements on underlying latent variables. It is these underlying variables that are considered to be of primary psychological significance. In such a time series analysis, therefore, there are many observed variables. To understand the process a large number of certain kinds of variables must be reduced to a small number. Factor analysis may be useful in this reduction. This feature is similar to that of economics where a number of variables, say prices of different kinds of goods, are to be combined into one representative variable, a price index. The aggregated variables are then subjected to time series analysis. The economic problem of "index numbers" or "aggregation" is different from the psychological problem of factors or primary abilities because the index number is supposed to approximate the total relevant effect of the quantities involved. On the other hand, the factor is considered to be the common part of several observed quantities. The appropriate mathematical models are correspondingly different. The similarity in approach, however, is that the reduction of many variables to a small number is to be carried out before completing a time series study.

REFERENCES

[1] Anderson, T. W. *An introduction to multivariate statistical analysis.* New York: Wiley, 1958.

[2] Anderson, T. W. *An introduction to the statistical analysis of time series.* (In preparation)

[3] Anderson, T. W. and Rubin, H. Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium in Mathematical Statistics and Probability,* Vol. V. Berkeley and Los Angeles: Univer. California Press, pp. 111–150, 1956.

[4] Cattell, R. B. *Factor analysis; an introduction and manual for the psychologist and social scientist.* New York: Harper, 1952.

[5] Cattell, R. B. *Personality and motivation: structure and measurement.* Yonkers-on-Hudson: World Book, 1957.

[6] Cattell, R. B. The potentialities of *P*-technique deduced from research applications. (Unpublished)

[7] Damarin, F. L., Jr. The factor analysis of time series as applied to a problem in cancer chemotherapy. (Unpublished)

[8] Frisch, R. *Statistical confluence analysis by means of complete regression systems.* Oslo: Universitetets Okonomiske Institutt, 1934.

[9] Hannan, E. J. *Time series analysis*. London: Methuen, 1960.

[10] Holtzman, W. H. Methodological issues in *P*-technique. (Unpublished).

[11] Hood, W. C. and Koopmans, T. C. *Studies in econometric method.* Cowles Commission Monograph No. 14. New York: Wiley, 1953.

[12] Klein, L. R. *A textbook of econometrics.* Evanston: Row, Peterson, 1953.

[13] Mefferd, R. B., Moran, L. J., and Kimble, J. P. Use of a factor analytic technique in the analysis of long term repetitive measurements made upon a single schizophrenic patient. (Unpublished)

[14] Wold, H. O. A. Ends and means in econometric model building. *Probability and statistics. The Harald Cramér volume.* New York: Wiley, 1960. Pp. 355–434.

Manuscript received 9/7/62

Revised manuscript received 11/24/62