CONTINUOUS TIME STOCHASTIC MODELS
AND ISSUES OF AGGREGATION OVER TIME

A. R. BERGSTROM

University of Essex

Contents

1. Introduction 1146
2. Closed first-order systems of differential and integral equations 1150
   2.1. Stochastic limit operations and stochastic differential equations 1150
   2.2. Random measures and systems with white noise disturbances 1157
   2.3. Estimation 1171
3. Higher order systems 1193
4. The treatment of exogenous variables and more general models 1202
5. Conclusion 1209
References 1210
1. Introduction

Since the publication of the influential articles of Haavelmo (1943) and Mann and Wald (1943) and the subsequent work of the Cowles Commission [see, especially, Koopmans (1950a)], most econometric models of complete economies have been formulated as systems of simultaneous stochastic difference equations and fitted to either quarterly or annual data. Models of this sort, which are discussed in Chapter 7 of this Handbook, can be written in either the structural form:

\[ \Gamma y_t + B_0 x_t + \sum_{r=1}^{k} B_r y_{t-r} = u_t, \]

or the reduced form:

\[ y_t = \Pi_0 x_t + \sum_{r=1}^{k} \Pi_r y_{t-r} + v_t, \]

where \( y_t \) is an \( n \times 1 \) vector of observable random variables (endogenous variables), \( x_t \) is an \( m \times 1 \) vector of observable non-random variables (exogenous variables), \( u_t \) is a vector of unobservable random variables (disturbances), \( \Gamma \) is an \( n \times n \) matrix of parameters, \( B_0 \) is an \( n \times m \) matrix of parameters, \( B_1, \ldots, B_k \) are \( n \times n \) matrices of parameters, \( \Pi_r = -\Gamma^{-1} B_r, \ r = 0, \ldots, k, \) and \( v_t = \Gamma^{-1} u_t \). It is usually assumed that \( E(u_t) = 0, \ E(u_s u'_t) = 0, \ s \neq t, \) and \( E(u_t u'_t) = \Sigma \), implying that \( E(v_t) = 0, \ E(v_t v'_t) = 0, \ s \neq t, \) and \( E(v_t v'_t) = \Omega = \Gamma^{-1} \Sigma \Gamma^{-1} \).

The variables \( x_{1t}, \ldots, x_{mt}, y_{1t}, \ldots, y_{nt} \) will usually include aggregations over a quarter (or year) of flow variables such as output, consumption, exports, imports and investment as well as levels at the beginning or end of the quarter (or year) of stock variables representing inventories, fixed capital and financial assets. They will also include indices of prices, wages and interest rates. These may relate to particular points of time, as will usually be the case with an index of the yield on bonds, or to time intervals as will be the case with implicit price deflators of the components of the gross national product.

Although the reduced form (2) is all that is required for the purpose of prediction under unchanged structure, the structural form (1) is the means through which a priori information derived from economic theory is incorporated in the model. This information is introduced by placing certain a priori restrictions on the matrices \( B_0, \ldots, B_k \) and \( \Gamma \). The number of these restrictions is, normally, such as to imply severe restrictions on the matrices \( \Pi_0, \ldots, \Pi_k \) of reduced form coefficients. Because of the smallness of the samples available to the
econometrician, these restrictions play a very important role in reducing the variances of the estimates of these coefficients and the resulting variances of the predictions obtained from the reduced form. The most common form of restriction on the matrices $B_0, \ldots, B_k$ and $\Gamma$ is that certain elements of these matrices are zero, representing the assumption that each endogenous variable is directly dependent (in a causal sense) on only a few of the other variables in the model. But $\Gamma$ is not assumed to be a diagonal matrix.

The simultaneity in the unlagged endogenous variables implied by the fact that $\Gamma$ is not a diagonal matrix is the distinguishing feature of this set of models as compared with models used in the natural sciences. It is necessary in order to avoid the unrealistic assumption that the minimum lag in any causal dependency is not less than the observation period. But there are obvious difficulties in interpreting the general simultaneous equations model as a system of unilateral causal relations in which each equation describes the response of one variable to the stimulus provided by other variables. For this reason Wold (1952, 1954, 1956) advocated the use of recursive models, these being models in which $\Gamma$ is a triangular matrix and $\Sigma$ is a diagonal matrix.

One way of interpreting the more general simultaneous equations model, which is not recursive, is to assume that the economy moves in discrete jumps between successive positions of temporary equilibrium at intervals whose length coincides with the observation period. We might imagine, for example, that on the first day of each quarter the values of both the exogenous variables and the disturbances affecting decisions relating to that quarter become known and that a new temporary equilibrium is established, instantaneously, for the duration of the quarter. But this is clearly a very unrealistic interpretation. For, if it were practicable to make accurate daily measurements of such variables as aggregate consumption, investment, exports, imports, inventories, the money supply and the implicit price deflators of the various components of the gross national product, these variables would undoubtedly be observed to change from day to day and be approximated more closely by a continuous function of time than by a quarterly step function.

A more realistic interpretation of the general simultaneous equations model is that it is derived from an underlying continuous time model. This is a more basic system of structural equations in which each of the variables is a function of a continuous time parameter $t$. The variables in this model will, therefore, be $y_1(t), \ldots, y_n(t), x_1(t), \ldots, x_m(t)$ where $t$ assumes all real values. The relation between each of these variables and the corresponding variable in the simultaneous equations model will depend on the type of variable. If the variable is a flow variable like consumption, in which case $y_f(t)$ is the instantaneous rate of consumption at time $t$, then the corresponding variable in the simultaneous equations model is the integral of $y_f(t)$ over an interval whose length equals the observation period, so that, if we identify the unit of time with the observation
period, we can write \( y_{it} = \int_{t-1}^{t} y_i(r) \, dr \). If \( y_i(t) \) is a stock variable like the money supply then the corresponding variable in the simultaneous equations model will be the value of \( y_i(t) \) at an integer value of \( t \) so that we have \( y_{it} = y_i(t) \), \( t = 1, 2, \ldots, T \).

It is intuitively obvious that, if the simultaneous equations system (1) is derived from an underlying continuous time model it will, generally, be no more than an approximation, even when the continuous time model holds exactly. One of the main considerations of this chapter will be the consequences, for estimation, of this sort of approximation and what is the best sort of approximation to use. This involves a precise specification of the continuous time model and a rigorous study of the properties of the discrete vector process generated by such a model.

If the underlying continuous time model is a system of linear stochastic differential equations with constant coefficients, and the exogenous variables and disturbances satisfy certain conditions, then, as we shall see, the discrete observations will satisfy, exactly, a system of stochastic difference equations in which each equation includes the lagged values of all variables in the system, and not just those which occur in the corresponding equation of the continuous time model. The disturbance vector in this exact discrete model is generated by a vector moving average process with coefficient matrices depending on the structural parameters of the continuous time model. A system such as this will be satisfied by the discrete observations whether they are point observations, integral observations or a mixture of the two, as they will be if the continuous time model contains a mixture of stock and flow variables. If there are no exogenous variables, so that the continuous time model is a closed system of stochastic differential equations, then the exact discrete model can be written in the form

\[
\begin{align*}
    y_t &= \sum_{r=1}^{k} F_r(\theta) y_{t-r} + \eta_t, \\
    \eta_t &= \sum_{r=0}^{l} C_r(\theta) \epsilon_{t-r}, \\
    \mathbb{E}(\epsilon_t) &= 0, \quad \mathbb{E}(\epsilon_t \epsilon_s') = K(\theta), \quad \mathbb{E}(\epsilon_s \epsilon_t') = 0, \quad s \neq t,
\end{align*}
\]

where the elements of the matrices \( F_1(\theta), \ldots, F_k(\theta), C_1(\theta), \ldots, C_l(\theta) \) and \( K(\theta) \) are functions of a vector \( \theta \) of structural parameters of the continuous time model.

It is a remarkable fact that the discrete observations satisfy the system (3) even though neither the integral \( \int_{t-1}^{t} y(r) \, dr \) nor the pair of point observations \( y(t) \) and \( y(t-1) \) conveys any information about the way in which \( y(t) \) varies over the interval \( (t-1, t) \) and the pattern of variation of \( y(t) \) over a unit interval varies both as between different realizations (corresponding to different elementary events in the space on which the probability measure is defined), for a given
interval, and as between different intervals for a given realization. Moreover, the form of the system \((3)\) does not depend on the observation period, but only on the form of the underlying continuous time model. That is to say the integers \(k\) and \(l\) do not depend on the observation period, and the matrices \(F_1(\theta), \ldots, F_k(\theta), C_1(\theta), \ldots, C_l(\theta)\) and \(K(\theta)\) depend on the observation period only to the extent that they will involve a parameter \(\delta\) to represent this period, if it is not identified with the unit of time. The observation period is, therefore, of no importance except for the fact that the shorter the observation period the more observations there will be and the more efficient will be the estimates of the structural parameters.

The exact discrete model \((3)\) plays a central role in the statistical treatment of continuous time stochastic models, for two reasons. First, a comparison of the exact discrete model with the reduced form of an approximating simultaneous model provides the basis for the study of the sampling properties of parameter estimates obtained by using the latter model and may suggest more appropriate approximate models. Secondly, the exact discrete model provides the means of obtaining consistent and asymptotically efficient estimates of the parameters of the continuous time model.

For the purpose of predicting future observations, when the structure of the continuous time model is unchanged, all that we require is the system \((3)\). But, even for this purpose, the continuous time model plays a very important role. For it is the means through which we introduce the a priori restrictions derived from economic theory. Provided that the economy is adjusting continuously, there is no simple way of inferring the appropriate restrictions on \((3)\) to represent even such a simple implication of our theory as the implication that certain variables have no direct causal influence on certain other variables. For, in spite of this causal independence, all of the elements in the matrices \(F_1, \ldots, F_k\) in the system \((3)\) will generally be non-zero. In this respect forecasting based on a continuous time model derived from economic theory has an important advantage over the methods developed by Box and Jenkins (1970) while retaining the richer dynamic structure assumed by their methods as compared with that incorporated in most discrete time econometric models.

For a fuller discussion of some of the methodological issues introduced above and an outline of the historical antecedents [among which we should mention Koopmans (1950b), Phillips (1959) and Durbin (1961)] and development of the theory of continuous time stochastic models in relation to econometrics, the reader is referred to Bergstrom (1976, ch. 1). Here we remark that the study of these models has close links with several other branches of econometrics and statistics. First, as we have indicated, it provides a new way of interpreting simultaneous equations models and suggests a more careful specification of such models. Secondly, it provides a further contribution to the theory of causal chain models as developed by Wold and others. Finally, as we shall see, it provides a
potentially important application of recent developments in the theory of vector
time series models. To some extent these developments have been motivated by
the needs of control engineering. But it seems likely that their most important
application in econometrics will be to continuous time models.

In the following section we shall deal fairly throughly with closed first order
systems of linear stochastic differential or integral equations, proving a number of
basic theorems and discussing various methods of estimation. We shall deal with
methods of estimation based on both approximate and exact discrete models and
their application to both stock and flow data. The results and methods discussed
in this section will be extended to higher order systems in Section 3. In Section 4
we shall discuss the treatment of exogenous variables and more general forms of
continuous time stochastic models.

2. Closed first-order systems of differential and integral equations

2.1. Stochastic limit operations and stochastic differential equations

Before getting involved with the problems associated with stochastic limit opera-
tions, it will be useful to consider the non-stochastic differential equation:

\[ Dy(t) = ay(t) + b + \phi(t), \]  

(4)

where \( D \) is the ordinary differential operator \( \frac{d}{dt} \), \( a \) and \( b \) are constants and \( \phi(t) \)
is a continuous function of \( t \) (time). It is easy to see that the solution to (4) subject
to the condition that \( y(0) \) is a given constant is:

\[ y(t) = \int_0^t e^{a(t-r)} \phi(r) \, dr + \left( y(0) + \frac{b}{a} \right) e^{at} - \frac{b}{a}. \]  

(5)

For, by differentiating (5) we obtain:

\[ Dy(t) = \frac{d}{dt} \left[ e^{at} \int_0^t e^{-ar} \phi(r) \, dr + \left( y(0) + \frac{b}{a} \right) e^{at} - \frac{b}{a} \right] \]

\[ = ae^{at} \int_0^t e^{-ar} \phi(r) \, dr + \phi(t) + a \left( y(0) + \frac{b}{a} \right) e^{at} \]

\[ = ay(t) + b + \phi(t). \]
From (5) we obtain:

\[
y(t) = \int_0^t e^{a(t-r)}\phi(r) \, dr + \left( y(0) + \frac{b}{a} \right) e^{at} - \frac{b}{a} + \int_0^t e^{a(t-r)}\phi(r) \, dr
\]

\[
= e^a \int_0^{t-1} e^{a(t-1-r)}\phi(r) \, dr + e^a \left( y(0) + \frac{b}{a} \right) e^{a(t-1)} - \frac{b}{a}
\]

\[
+ \left( e^a - 1 \right) \frac{b}{a} + \int_{t-1}^t e^{a(t-r)}\phi(r) \, dr
\]

\[
= e^a y(t-1) + \left( e^a - 1 \right) \frac{b}{a} + \int_{t-1}^t e^{a(t-r)}\phi(r) \, dr.
\]

We have shown, therefore, that the solution to the differential equation (4) satisfies the difference equation:

\[
y(t) = fy(t-1) + g + \psi_t,
\]

where

\[
f = e^a, \quad g = \left( e^a - 1 \right) \frac{b}{a}
\]

and

\[
\psi_t = \int_{t-1}^t e^{a(t-r)}\phi(r) \, dr.
\]

In order to apply the above argument to an equation in which \( \phi(t) \) is replaced by a random disturbance function it is necessary to define stochastic differentiation and integration. We can do this by making use of the concept of convergence in mean square. The sequence \( \xi_n, n = 1, 2, \ldots \), of random variables is said to converge in mean square to a random variable \( \eta \) if:

\[
\lim_{n \to \infty} E(\xi_n - \eta)^2 = 0.
\]

In this case \( \eta \) is said to be the limit in mean square of \( \xi_n \) and we write:

\[
l.i.m. \xi_n = \eta.
\]

Suppose now that \( \{ \xi(t) \} \) is a family of random variables, there being one member of the family for each value of \( t \) (time). We shall call \( \{ \xi(t) \} \) a continuous time random process if \( t \) takes on all real values and a discrete time random process if \( t \)
R. Bergstrom takes on all integer values, the words "continuous time" or "discrete time" being omitted when it is clear from the context which case is being considered. A random process \( \{ \xi(t) \} \) is said to be mean square continuous on an interval \([c, d]\) if:

\[
E[\xi(t) - \xi(t - h)]^2 \to 0
\]

uniformly in \( t \), on \([c, d]\), as \( h \to 0 \). And it is said to be mean square differentiable if there is a random process \( \{ \eta(t) \} \) such that:

\[
\lim_{h \to 0} E \left\{ \frac{\xi(t + h) - \xi(t) - \eta(t)}{h} \right\}^2 = 0.
\]

In the latter case we shall write:

\[
D\xi(t) = \frac{d\xi(t)}{dt} = \eta(t),
\]

and call \( D = d/dt \) the mean square differential operator.

In order to define integration we can follow a procedure similar to that used in defining the Lebesgue integral of a measurable function [see Kolmogorov and Fomin (1961, ch. 7)]. We start by considering a simple random process which can be integrated by summing over a sequence of measurable sets. The random process \( \{ \xi(t) \} \) is said to be simple on an interval \([c, d]\) if there is a finite or countable disjoint family of measurable sets \( \Delta_k, k = 1, 2, \ldots \), whose union is the interval \([c, d]\) and corresponding random variables \( \xi_k, k = 1, 2, \ldots \), such that:

\[
\xi(t) = \xi_k, \quad t \in \Delta_k, \quad k = 1, 2, \ldots.
\]

Let \( |\Delta_k| \) be the Lebesgue measure of \( \Delta_k \). Then the simple random process \( \{ \xi(t) \} \) is said to be integrable in the wide sense on \([c, d]\) if the series \( \sum_{k=1}^{\infty} \xi_k |\Delta_k| \) converges in mean square. The integral of \( \{ \xi(t) \} \) over \([c, d]\) is then defined by:

\[
\int_{c}^{d} \xi(t) \, dt = \sum_{k=1}^{\infty} \xi_k |\Delta_k| = \text{l.i.m.} \sum_{n=1}^{\infty} \xi_k |\Delta_k|.
\]

We turn now to the integration of an arbitrary random process. We say that a random process \( \{ \xi(t) \} \) is integrable in the wide sense on the interval \([c, d]\) if there exists a sequence \( \{ \xi_n(t) \}, n = 1, 2, \ldots \), of simple integrable processes which converges uniformly to \( \{ \xi(t) \} \) on \([c, d]\), i.e.

\[
E[\xi(t) - \xi_n(t)]^2 \to 0,
\]
as \( n \to \infty \), uniformly in \( t \), on \([c, d]\). It can be shown [see Rozanov (1967, p. 11)]
that, in this case, the sequence \( \int_c^d x_n(t) \, dt, \, n = 1, 2, \ldots \), of random variables has a
limit in mean square which we call the integral of \( \{x(t)\} \) over \([c, d]\). We can then write:

\[
\int_c^d x(t) \, dt = \lim_{n \to \infty} \int_c^d x_n(t) \, dt.
\]

If the random process \( \{x(t)\} \) is mean square continuous over a finite interval
\([c, d]\), then it is integrable over that interval. For we can then define the simple
random process \( \{x_n(t)\} \) by dividing \([c, d]\) into \( n \) equal subintervals \( \Delta_{n1}, \ldots, \Delta_{nn} \),
letting \( x_{n1}, \ldots, x_{nn} \) be the random variables defining \( x(t) \) at, say, the midpoints of
\( \Delta_{n1}, \ldots, \Delta_{nn} \), respectively, and putting:

\[
x_n(t) = x_{nk}, \quad t \in \Delta_{nk}, \quad k = 1, \ldots, n.
\]

The simple random process \( \{x_n(t)\} \) is obviously integrable, its integral being
\( \sum_{k=1}^n x_{nk} |\Delta_{nk}| \). Moreover, it follows directly from the definition of mean square
continuity and the fact that the length of the intervals \( \Delta_{nk} \) tends to zero as \( t \to \infty \)
that the sequence \( \{x_n(t)\}, \, n = 1, 2, \ldots \), of simple random processes converges,
uniformly on \([c, d]\), to \( \{x(t)\} \). We have shown, by this argument, that if a
random process is mean square continuous, then it is integrable over an interval,
not only in the wide sense defined above, but also in a stricter sense correspond-
ing to the Riemann integral. A much weaker sufficient condition for a random
process to be integrable in the wide sense is given by Rozanov (1967, theorem
2.3).

It is easy to show that the integral defined above has all the usual properties. In
particular, if \( \{x_1(t)\} \) and \( \{x_2(t)\} \) are two integrable random processes and \( a_1 \) and
\( a_2 \) are constants, then:

\[
\int_c^d [a_1 x_1(t) + a_2 x_2(t)] \, dt = a_1 \int_c^d x_1(t) \, dt + a_2 \int_c^d x_2(t) \, dt.
\]

And if \( \{x(t)\} \) is mean square continuous in a neighbourhood of \( t \), then:

\[
\frac{d}{dt} \int_{t_1}^t x(r) \, dr = x(t),
\]

where \( d/dt \) is the mean square differential operator.

In addition to the various limit operations defined above we shall use the
assumption of stationarity. A random process \( \{x(t)\} \) is said to be stationary in the
wide sense if it has an expected value \( \mathbb{E}[x(t)] \) not depending on \( t \) and a correlation
function $B(t, s)$ defined by:

$$B(t, s) = E[\xi(t)\xi(s)],$$

depending only on the difference $t - s$, so that we can write:

$$B(t, s) = \gamma(t - s).$$

A wide sense stationary process is said to be **ergodic** if the time average $(1/T)\int_0^T \xi(t) \, dt$ converges in mean square to the expected value $E[\xi(t)]$ as $T \to \infty$. A random process $\{\xi(t)\}$ is said to be **strictly stationary** if, for any numbers $t_1, \ldots, t_k$, the joint distribution of $\xi(t_1 + r), \ldots, \xi(t_k + r)$ does not depend on $r$.

A necessary and sufficient condition for a wide sense stationary process to be mean square continuous is that its correlation function is a continuous function of $(t - s)$ at the origin (i.e. when $t = s$). For we have:

$$E[\xi(t) - \xi(t - h)]^2 = E[\xi(t)]^2 + E[\xi(t - h)]^2 - 2E[\xi(t)\xi(t - h)]$$

$$= 2E[\xi(t)]^2 - 2E[\xi(t)\xi(t - h)]$$

$$= 2[B(t, t) - B(t, t - h)]$$

$$= 2[\gamma(0) - \gamma(h)].$$

We shall now consider the **stochastic differential equation**:

$$Dy(t) = ay(t) - b + \xi(t), \quad (7)$$

where $D$ is the mean square differential operator and $\{\xi(t)\}$ is a mean square continuous wide sense stationary process. Our ultimate aim is to consider the estimation of the parameters $a$ and $b$ from a sequence $y(1), y(2), \ldots$, of observations when $\xi(t)$ is an unobservable disturbance. For this purpose it will be necessary to place further restrictions on $\{\xi(t)\}$. But we shall not concern ourselves with these at this stage. Our immediate aim is to find a solution to (7) and show that this solution satisfies a difference equation whose relation to (7) is similar to that of the non-stochastic difference equation (6) to the differential equation (4).

**Theorem 1**

If $\{\xi(t)\}$ is a mean square continuous wide sense stationary process, then, for any given $y(0)$, (7) has a solution:

$$y(t) = \int_0^t e^{a(t-r)}\xi(r) \, dr + \left[ y(0) + \frac{b}{a} \right] e^{at} - \frac{b}{a}, \quad (8)$$
and this solution satisfies the stochastic difference equation:

\[ y(t) = fy(t - 1) + g + \varepsilon_t, \tag{9} \]

where

\[ f = e^a, \quad g = \left( e^a - 1 \right) \frac{b}{a} \]

and

\[ \varepsilon_t = \int_{t-1}^{t} e^{a(t-r)} \xi(r) \, dr. \]

**Proof**

The random process \( \{e^{at}\xi(t)\} \) is mean square continuous on any finite interval. For:

\[
E\left[ e^{at}\xi(t) - e^{a(t-h)}\xi(t-h) \right]^2 = E\left\{ e^{at}\left[ \xi(t) - \xi(t-h) \right] \right. \\
+ e^{at}(1-e^{-ah})\xi(t-h) \left. \right\}^2 \\
= e^{2at}E\left[ \xi(t) - \xi(t-h) \right]^2 + e^{2at}(1-e^{-ah})^2 \gamma(0) \\
+ 2e^{2at}(1-e^{-ah})[\gamma(h) - \gamma(0)].
\]

And, since \( e^{2at} \) is bounded on any finite interval while \( E[\xi(t) - \xi(t-h)]^2 \to 0 \), uniformly in \( t \), as \( h \to 0 \), the right-hand side of the last equation converges to zero, uniformly in \( t \), as \( h \to 0 \).

It follows that the integral \( \int_0^t e^{-ar}\xi(r) \, dr \) exists. Moreover:

\[
\int_0^t e^{a(t-r)} \xi(r) \, dr = e^{at} \int_0^t e^{-ar}\xi(r) \, dr
\]

and

\[
\frac{d}{dt} \left[ \int_0^t e^{-ar}\xi(r) \, dr \right] = e^{-at}\xi(t).
\]

All the operations that were performed in showing that (5) is a solution to (4) and that this solution satisfies (6) are valid in the mean square sense, therefore, when \( \phi(t) \) is replaced by \( \xi(t) \). It follows that (8) is a solution to (7) and that this solution satisfies (9). \( \blacksquare \)
We turn now to a preliminary consideration of the problem of estimation. It would be very convenient if, in addition to being wide sense stationary and mean square continuous the random process \( \{ \xi(t) \} \) had the property that its integrals over any pair of disjoint intervals were uncorrelated. For then the disturbances \( \epsilon_i, \ t = 1, 2, \ldots, \) in (9) would be uncorrelated and, provided that they satisfied certain additional conditions (e.g. that they are distributed independently and identically), the least squares estimates \( f^* \) and \( g^* \) of the constants in this equation would be consistent. We could then obtain consistent estimates of \( a \) and \( b \) from

\[
\begin{align*}
a^* &= \log f^* \quad \text{and} \quad b^* = \frac{a^* g^*}{f^* - 1}.
\end{align*}
\]

But it is easy to see that, if \( \{ \xi(t) \} \) is mean square continuous, it cannot have the property that its integrals over any pair of disjoint intervals are uncorrelated. For the integrals \( \int_{t-h}^t \xi(r) \, dr \) and \( \int_t^{t+h} \xi(r) \, dr \) will obviously be correlated if \( h \) is sufficiently small. The position is worse than this, however. We shall now show that there is no wide sense stationary process (whether mean square continuous or otherwise) which is integrable in the wide sense and whose integrals over every pair of disjoint intervals are uncorrelated.

Suppose that the wide sense stationary process \( \{ \xi(t) \} \) is integrable in the wide sense and that its integrals over every pair of disjoint intervals are uncorrelated. We may assume, without loss of generality, that \( E[\xi(t)] = 0 \). Let \( E[\int_{t-h}^t \xi(r) \, dr]^2 = c \) and let \( h = 1/n \), where \( n \) is an integer greater than 1. We shall consider the set of \( n \) integrals:

\[
\int_{t-h}^t \xi(r) \, dr, \int_{t-2h}^{t-h} \xi(r) \, dr, \ldots, \int_{t-1}^{t-(n-1)h} \xi(r) \, dr.
\]

By hypothesis these integrals are uncorrelated, and by the assumption of stationarity their variances are equal. It follows that:

\[
E\left[\int_{t-h}^t \xi(r) \, dr\right]^2 = ch,
\]

and hence:

\[
E\left[\frac{1}{h} \int_{t-h}^t \xi(r) \, dr\right]^2 = \frac{c}{h} \to \infty, \quad \text{as} \ n \to \infty,
\]

i.e. the variance of the mean value, over an interval, of a realization of \( \xi(t) \) tends to infinity as the length of the intervals tends to zero. But this is impossible since, for any random process which is integrable in the wide sense, the integrals must
satisfy [see Rozanov (1967, p. 10)] the inequality:

\[
E\left[ \int_{t-h}^{t} \xi(r) \, dr \right]^2 \leq \left\{ \int_{t-h}^{t} \left[ E(\xi(r))^2 \right]^{1/2} \, dr \right\}^2.
\] (10)

And, if the process is stationary in the wide sense, the right-hand side of (10) equals \(\gamma(0)h^2\). It follows that:

\[
E\left[ \frac{1}{h} \int_{t-h}^{t} \xi(r) \, dr \right]^2 \leq \gamma(0).
\]

This contradiction shows that the integrals over sufficiently small adjacent intervals must be correlated.

Although it is not possible for the integrals over every pair of disjoint intervals, of a wide sense stationary process, to be uncorrelated, their correlations can, for intervals of given length, be arbitrarily close to zero. They will be approximately zero if, for example, the correlation function is given by:

\[
\gamma(t-s) = \sigma^2 e^{-\beta|t-s|},
\]

where \(\sigma^2\) and \(\beta\) are positive numbers and \(\beta\) is large. A stationary process with this correlation function does (as we shall see) exist and, because the correlation function is continuous at the origin, it is mean square continuous and hence integrable over a finite interval. If \(\beta\) is sufficiently large the disturbances \(\epsilon_t, t=1,2,\ldots,\) in (9) can, for all practical purposes, be treated as uncorrelated, and we may expect the least squares estimates \(f^*\) and \(g^*\) to be approximately consistent.

Heuristically we may think of an improper random process \(\{\xi(t)\}\) called "white noise" which is obtained by letting \(\beta \to \infty\) in a wide sense stationary process with the above correlation function. For practical purposes we may regard white noise as indistinguishable from a process in which \(\beta\) is finite but large. But this is not a very satisfactory basis for the rigorous development of the theory of estimation. For this purpose we shall need to define white noise more precisely.

2.2. Random measures and systems with white noise disturbances

A precise definition of white noise can be given by defining a random set function which has the properties that we should expect the integral of \(\xi(t)\) to have under our heuristic interpretation. That is to say we define a random set function \(\xi\) which associates with each semi-interval \(\Delta = [s, t)\) (or \(\Delta = (s, t]\)) on the real line a
random variable $\xi(\Delta)$ and has the properties:

$$\xi(\Delta_1 \cup \Delta_2) = \xi(\Delta_1) + \xi(\Delta_2),$$

when $\Delta_1$ and $\Delta_2$ are disjoint, i.e. it is additive,

$$E[\xi(\Delta)]^2 = \sigma^2[t - s],$$

where $\sigma^2$ is a positive constant,

$$E[\xi(\Delta_1)\xi(\Delta_2)] = 0,$$

when $\Delta_1$ and $\Delta_2$ are disjoint.

A set function with these properties is a special case of a type of set function called a random measure. The concept of a random measure is of fundamental importance, not only in the treatment of white noise, but also (in the more general form) in the spectral representation of a stationary process, which will be used in Section 4. We shall now digress, briefly, to discuss the concept more generally and define integration with respect to a random measure and properties of the integral which we shall use in the sequel.

Let $R$ be some semiring of subsets of the real line (e.g. the left closed semi-intervals, or the Borel sets, or the sets with finite Lebesgue measure [see Kolmogorov and Fomin (1961, ch. 5)]. And let $\Phi$ be a random set function which associates with any subset $\Delta \in R$ a random variable $\Phi(\Delta)$ (generally complex valued) and has the properties:

$$\Phi(\Delta_1 \cup \Delta_2) = \Phi(\Delta_1) + \Phi(\Delta_2),$$

if $\Delta_1$ and $\Delta_2$ are disjoint, i.e. it is additive:

$$E|\Phi(\Delta)|^2 = F(\Delta) < \infty,$$

$$E[\Phi(\Delta_1)\Phi(\Delta_2)] = 0,$$

when $\Delta_1$ and $\Delta_2$ are disjoint. Then $\Phi$ is said to be a random measure. If, in addition,

$$\Phi(\Delta) = \sum_k \Phi(\Delta_k)$$

for every $\Delta \in R$ which is the union of disjoint subsets $\Delta_k$ and the series on the right-hand side converges in mean square, then the random measure $\Phi$ is said to be $\sigma$-additive.
It can be shown [Rozanov (1967, theorem 2.1)] that a σ-additive random measure defined on some semiring on the real line can be extended to the ring of all measurable sets in the σ-ring generated by that semiring. This implies that if we give the random measure \( \xi \), defined above, the additional property of σ-additivity so that:

\[ \xi(\Delta) = \sum_k \xi(\Delta_k), \]

whenever the semi-interval \( \Delta \) is the union of disjoint semi-intervals \( \Delta_k \) then it can be extended to the ring of all Borel sets on the real line with finite Lebesgue measure. We shall define white noise by the random measure \( \xi \) extended in this way to the measurable sets on the real line representing the time domain.

We turn now to the definition of the integral of a (non-random but, generally, complex valued) measurable function \( f(x) \) with respect to a random measure \( \Phi \) which is defined on the Borel sets of some interval \([c, d]\) (where \( c \) and \( d \) may have the values \(-\infty \) and \( \infty \), respectively). We start by defining the integral of a simple function. The measurable function \( f(x) \) is said to be simple on the interval \([c, d]\) if it assumes a finite or countable set of values on disjoint sets \( A \), whose union is \([c, d]\). And a simple function is said to be integrable with respect to the random measure \( \Phi \) on the interval \([c, d]\) if the series \( \sum_k f_k(\Phi(\Delta_k)) \) converges in mean square. The integral of \( f(x) \) with respect \( \Phi \) over \([c, d]\) is then defined as the limit in mean square to which this sum converges and we write:

\[ \int_c^d f(x) \Phi(dx) = \sum_{k=1}^{\infty} f_k(\Phi(\Delta_k)). \]

An arbitrary measurable function \( f(x) \) is said to be integrable with respect to \( \Phi \) on the interval \([c, d]\) if there is a sequence \( f_n(x), n = 1, 2, \ldots, \) of simple integrable measurable functions which converges in mean square to \( f(x) \) on \([c, d]\), i.e.

\[ \int_a^b |f_n(x) - f(x)|^2 F(dx) \to 0, \]

as \( n \to \infty \), where the integral is defined in the Lebesgue–Stieltjes sense [see Cramér (1951, ch. 7)]. It can be shown [Rozanov (1967, p. 7)], that, in this case, the sequence \( \int_c^d f_n(x) \Phi(dx), n = 1, 2, \ldots, \) has a limit in mean square which we call the integral of \( f(x) \) over \([c, d]\). We can then write:

\[ \int_c^d f(x) \Phi(dx) = \lim_{n \to \infty} \int_c^d f_n(x) \Phi(dx). \]
If $\Phi(\Delta)$ is undefined on $[-\infty, \infty]$ we can define $\int_{-\infty}^{\infty} f(x) \Phi(dx)$ by:

$$
\int_{-\infty}^{\infty} f(x) \Phi(dx) = \lim_{c \to -\infty} \lim_{d \to \infty} \int_{c}^{d} f(x) \Phi(dx).
$$

provided that the limit on the right-hand side of the equation exists.

A necessary and sufficient condition for the existence of the integral $\int_{c}^{d} f(x) \Phi(dx)$, where $f(x)$ is an arbitrary measurable function (and $c$ and $d$ may assume the values $-\infty$ and $\infty$, respectively), is:

$$
\int_{c}^{d} |f(x)|^2 F(dx) < \infty.
$$

If this condition is satisfied, then [Rozanov (1967, p. 7)]:

$$
\mathbb{E}\left(\int_{c}^{d} f(x) \Phi(dx)\right)^2 = \int_{c}^{d} |f(x)|^2 F(dx).
$$

And, if the measurable functions $f(x)$ and $g(x)$ each satisfy condition (11), then [Rozanov (1967, p. 7)]:

$$
\mathbb{E}\left\{\int_{c}^{d} f(x) \Phi(dx) \int_{c}^{d} g(x) \Phi(dx)\right\} = \int_{c}^{d} f(x) g(x) F(dx).
$$

When $\Phi$ is the random measure, $\xi$, by which we have defined white noise, $F(\Delta)$ has the simple form $\sigma^2 |\Delta|$, where

$$
\sigma^2 = \mathbb{E}\left\{\int_{t-1}^{t} \xi(dr)\right\}^2,
$$

and $|\Delta|$ is the Lebesgue measure of $\Delta$. A necessary and sufficient condition for the existence of the integral $\int_{c}^{d} f(r)\xi(dr)$, where $f(r)$ is a measurable function (and $c$ and $d$ may assume the values $-\infty$ and $\infty$, respectively), is:

$$
\int_{c}^{d} \{f(r)\}^2 dr < \infty,
$$

the integral in (14) being the ordinary Lebesgue integral [which will be equal to the Riemann integral if $f(r)$ is a continuous function]. If this condition is satisfied, then [as a special case of (12)]:

$$
\mathbb{E}\left(\int_{c}^{d} f(r)\xi(dr)\right)^2 = \sigma^2 \int_{c}^{d} \{f(r)\}^2 dr.
$$
And, if the measurable functions \( f(x) \) and \( g(x) \) each satisfy condition (14), then [as a special case of (13)]:

\[
E\left( \int_c^d f(r) \xi(dr) \int_c^d g(r) \xi(dr) \right) = \sigma^2 \int_c^d f(r) g(r) dr.
\] (16)

We note incidentally that \( \int_c^d f(r) \xi(dr) \) is a random process whose increments are uncorrelated, i.e. a random process with orthogonal increments [see Doob (1953, ch. 9) for a full discussion of such processes].

Before applying the above results in the treatment of stochastic differential equations with white noise disturbances, we shall illustrate their application by proving the existence of a wide sense stationary process with the correlation function \( \sigma^2 e^{-\beta|t-s|} \), as assumed in the heuristic introduction to the concept of white noise given at the end of the last subsection. The function \( f(r) \), defined by

\[
f(r) = \sigma(2\beta)^{1/2} e^{-\beta(t-r)},
\]

is integrable, with respect to \( \xi \), over the interval \([-\infty, t]\), since

\[
\int_{-\infty}^t \left[ \sigma(2\beta)^{1/2} e^{-\beta(t-r)} \right]^2 dr = \sigma^2,
\]

i.e. condition (14) is satisfied. Using (15):

\[
E\left[ \int_{-\infty}^t \sigma(2\beta)^{1/2} e^{-\beta(t-r)} \xi(dr) \right]^2 = \sigma^2.
\]

And, if \( s < t \):

\[
E\left[ \int_{-\infty}^s \sigma(2\beta)^{1/2} e^{-\beta(s-r)} \xi(dr) \int_{-\infty}^t \sigma(2\beta)^{1/2} e^{-\beta(t-r)} \xi(dr) \right]
= 2\beta \sigma^2 E\left[ \int_{-\infty}^s e^{-\beta(s-r)} \xi(dr) \int_{-\infty}^s e^{-\beta(s-r)} \xi(dr) \right]
+ 2\beta \sigma^2 E\left[ \int_{-\infty}^s e^{-\beta(s-r)} \xi(dr) \int_{s}^t e^{-\beta(t-r)} \xi(dr) \right]
- 2\beta \sigma^2 e^{-\beta(t-s)} E\left[ \int_{-\infty}^s e^{-\beta(s-r)} \right]^2
= \sigma^2 e^{-\beta(t-s)}.
\]

It follows that \( \{ \xi(t) \} \), where

\[
\xi(t) = \int_{-\infty}^t \sigma(2\beta)^{1/2} e^{-\beta(t-r)} \xi(dr)
\]
is a wide sense stationary process with the correlation function \( \sigma^2 e^{-\beta|t-s|} \).

A stochastic differential equation with a white noise disturbance is sometimes written like (7) with \( Dv(t) \) (or \( dy(t)/dt \)) on the left-hand side and \( \xi(t) \) in place
of $\xi(t)$ on the right-hand side. It is then understood that $\xi(t)$ is defined only by the properties of its integral and that $y(t)$ is not mean square differentiable. We shall not use that notation in this chapter since we wish to reserve the use of the operator $D$ for random processes which are mean square differentiable, as is $y(t)$ in (7). A first-order, linear stochastic differential equation with constant coefficients and a white noise disturbance will be written:

$$\frac{dy(t)}{dt} = (\alpha(t) + b)dt + \xi(dt),$$

which will be interpreted as meaning that the random process $y(t)$ satisfies the stochastic integral equation:

$$y(t) - y(0) = \int_{0}^{t} [\alpha(r) + b] \, dr + \int_{0}^{t} \xi(dr),$$

for all $t$.

Equation (17) is a special case of the equation

$$\frac{dy(t)}{dt} = m(t, y(t))dt + \sigma(t, y(t))\xi(dt),$$

in which the functions $m(t, y)$ and $\sigma(t, y)$ can be non-linear in $t$ and $y$, and which is interpreted [see Doob (1953, ch. 6)] as meaning that $y(t)$ satisfies the stochastic integral equation:

$$y(t) - y(0) = \int_{0}^{t} m(r, y(r)) \, dr + \int_{0}^{t} \sigma(r, y(r)) \xi(dr),$$

for all $t$ on some interval $[0, T]$. It has been shown [see Doob (1953, ch. 6), which modifies the work of Ito (1946) and (1950)] that, under certain conditions, there exists a random process $y(t)$ satisfying (20), for all $t$ on an interval $[0, T]$, and that, for any given $y(0)$, this solution is unique in the sense that, if $\hat{y}(t)$ is any other solution:

$$P[y(t) - \hat{y}(t) = 0] = 1, \quad 0 \leq t \leq T.$$  \hspace{1cm} (21)

The conditions, assumed in Doob (1953), are that the process $\{\int_{0}^{t} \xi(dr)\}$ is Gaussian and that the functions $m$ and $\sigma$ satisfy a Lipschitz condition and certain other conditions on $[0, T]$. A random process $\{\xi(t)\}$ is said to be Gaussian if the random variables $\xi(t)$ are normally distributed. The assumption that $\{\int_{0}^{t} \xi(dr)\}$ is Gaussian implies, therefore, that $\xi(\Delta_1)$ and $\xi(\Delta_2)$ are independent if $\Delta_1$ and $\Delta_2$ are disjoint and identically distributed if $|\Delta_1| = |\Delta_2|$.

In discussing the solution to (18) we shall not need to assume that $\{\int_{0}^{t} \xi(dr)\}$ is Gaussian. We shall now show that this equation has a solution, which will be given explicitly, that this solution is unique, in the sense of (21), in the class of mean square continuous processes, and that it satisfies a difference equation with serially uncorrelated disturbances.
**Theorem 2**

If $\xi$ is a random measure, defined on all subsets of the line $-\infty < t < \infty$ with finite Lebesgue measure, such that:

$$E[\xi(dt)] = 0, \quad E[\xi(dt)]^2 = \sigma^2 dt,$$

then

(a) for any given $y(0)$ (18) has a solution:

$$y(t) = \int_0^t e^{a(t-r)}\xi(dr) + \left[y(0) + \frac{b}{a}\right]e^{at} - \frac{b}{a}; \quad (22)$$

(b) the solution (22) is unique in the class of mean square continuous processes, i.e. if $\hat{y}(t)$ is any other mean square continuous process satisfying (18) and $\hat{y}(0) = y(0)$, then (21) holds for any interval $[0, T]$;

(c) the solution (22) satisfies the stochastic difference equation:

$$y(t) = fy(t-1) + g + \epsilon_t, \quad (23)$$

where

$$f = e^a, \quad g = (e^a - 1)\frac{b}{a}, \quad E(\epsilon_t) = 0,$$

$$E(\epsilon^2_t) = \frac{\sigma^2}{2a}(e^{2a} - 1), \quad E(\epsilon_s\epsilon_t) = 0, \quad s \neq t.$$

**Proof**

(a) We first note that the integral $\int_0^t e^{a(t-r)}\xi(dr)$ exists, since condition (14) is satisfied. Now let $y(t)$ be defined by (22) and let $h$ be any positive number. Then

$$y(t) - y(t-h) = \int_0^t e^{a(t-r)}\xi(dr) - \int_0^{t-h} e^{a(t-r)}\xi(dr)$$

$$\quad + \left[y(0) + \frac{b}{a}\right](e^{at} - e^{a(t-h)}) + \int_{t-h}^t e^{a(t-r-h)}\xi(dr)$$

$$= (1 - e^{-ah})y(t) + (1 - e^{-ah})\frac{b}{a} + e^{-ah}\int_{t-h}^t (e^{a(t-r)} - 1)\xi(dr)$$

$$+ (e^{-ah} - 1)\int_{t-h}^t \xi(dr) + \int_{t-h}^t \xi(dr)$$

$$= ahy(t) + bh + \int_{t-h}^t \xi(dr) + u(t, h), \quad (24)$$
where

\[ u(t, h) = \left( -\frac{a^2 h^2}{2} + \frac{a^3 h^3}{6} - \cdots \right) \left( y(t) + \frac{b}{a} \right) \]

\[ + e^{-ah} \int_{t-h}^{t} (e^{a(t-r)} - 1) \xi(dr) + \left( -ah + \frac{a^2 h^2}{2} - \cdots \right) \int_{t-h}^{t} \xi(dr). \]

But

\[
E \left[ \int_{t-h}^{t} (e^{a(t-r)} - 1) \xi(dr) \right]^2 = \int_{t-h}^{t} (e^{a(t-r)} - 1)^2 dr \\
= \frac{1}{2a} (e^{2ah} - 1) - \frac{2}{a} (e^{ah} - 1) + h \\
= O(h^3).
\]

Therefore

\[ E[u(t, h)]^2 = O(h^3). \]

Now let \( h = t/n \), where \( n \) is a positive integer. Then

\[
y(t) - y(0) = \left( y(t) - y(t-h) \right) + \left( y(t-h) - y(t-2h) \right) + \cdots \\
+ \left[ y(t - (n-1)h) - y(0) \right] \\
= a \left[ \frac{t}{n} \sum_{r=1}^{n} y \left( \frac{rt}{n} \right) \right] + bt \int_{0}^{t} \xi(dr) + \sum_{r=1}^{n} u \left( \frac{rt}{n}, \frac{t}{n} \right). \tag{25}
\]

Now it is clear from (24) that \( y(t) \) is mean square continuous, and hence:

\[
\text{l.i.m.} \left[ \frac{t}{n} \sum_{r=1}^{n} y \left( \frac{rt}{n} \right) \right] = \int_{0}^{t} y(r) dr.
\]

Moreover, since

\[
E \left[ u \left( \frac{rt}{n}, \frac{t}{n} \right) \right]^2 = O \left( \frac{1}{n^3} \right),
\]

\[
\text{l.i.m.} \sum_{r=1}^{n} u \left( \frac{rt}{n}, \frac{t}{n} \right) = 0.
\]

Since (25) holds for all positive integers \( n \), it must hold when each term is
replaced by its limit in mean square as \( n \to \infty \), i.e.
\[
y(t) - y(0) = \int_0^t [ay(t) + b] \, dr + \int_0^t \xi(dr).
\]

(b) Let \( \hat{y}(t) \) be any other mean square continuous process satisfying (18), on \([0, T]\), and \( \hat{y}(0) = y(0) \). Define:
\[
\xi(t) = \hat{y}(t) - y(t).
\]

Then
\[
\xi(t) = a \int_0^t \xi(r) \, dr, \quad 0 \leq t \leq T
\] (26)

Since \( y(t) \) and \( \hat{y}(t) \) are mean square continuous, \( \xi(t) \) is mean square continuous. Therefore \( E[\xi(t)]^2 \) is a continuous function of \( t \), since
\[
\left| E[\xi(t)]^2 - E[\xi(t-h)]^2 \right| \leq E[|\xi(t)]^2 - [\xi(t-h)]^2 | \leq E[|\xi(t) + \xi(t-h)|] \leq \{E[\xi(t)]^2 E[\xi(t-h)]^2\}^{1/2} \to 0, \quad \text{as } h \to 0.
\]

Let \( n \) be a positive integer such that \( \tau = T/n < 1/a \). Since \( E[\xi(t)]^2 \) is continuous it has a maximum \( E[\xi(\tau_1)]^2 \) on the closed interval \([0, \tau]\), and \( 0 \leq \tau_1 \leq \tau \leq 1/a \). Therefore, using (10) and (26):
\[
E[\xi(\tau_1)]^2 = a^2 E\left( \int_0^{\tau_1} \xi(r) \, dr \right)^2 \leq a^2 \left( \int_0^{\tau_1} [E(\xi(r))]^2 \, dr \right)^2 \leq a^2 \tau_1^2 E[\xi(\tau_1)]^2.
\]

but \( a^2 \tau_1^2 < 1 \). Therefore:
\[
E[\xi(\tau_1)]^2 = 0.
\]

Therefore:
\[
P(\xi(t) = 0) = 1, \quad 0 \leq t \leq \tau.
\]
Since a similar relation holds for each of the \( n \) intervals, of length \( \tau \), whose union is \([0, T]\) we have:

\[
P(\xi(t) = 0) = 1, \quad 0 \leq t \leq T.
\]

(c) Let \( y(t) \) be the random process defined by (22). Then

\[
y(t) = e^a \int_0^t e^{a(t-1-r)} \xi(dr) + e^a \left[ y(0) + \frac{b}{a} \right] e^{a(t-1)}
- e^a \left( \frac{b}{a} \right) + (e^a - 1) \frac{b}{a} + \int_t^{t-1} e^{a(t-r)} \xi(dr)
= f_y(t-1) + g + \epsilon_t,
\]

where

\[
\epsilon_t = \int_{t-1}^t e^{a(t-r)} \xi(dr).
\] (27)

It is clear from the definition of the integral that:

\[
E(\epsilon_t) = E \left[ \int_{t-1}^t e^{a(t-r)} \xi(dr) \right] = 0.
\]

And, using (15) and (16):

\[
E(\epsilon_t)^2 = E \left[ \int_{t-1}^t e^{a(t-r)} \xi(dr) \right]^2 = \sigma^2 \int_{t-1}^t e^{2a(t-r)} dr
= \sigma^2 \int_0^1 e^{2ar} dr - \frac{\sigma^2}{2a} (e^{2a} - 1),
\]

\[
E(\epsilon_s \epsilon_t) = E \left[ \int_{s-1}^s e^{a(s-r)} \xi(dr) \int_{t-1}^t e^{a(t-r)} \xi(dr) \right] = 0, \quad s \neq t,
\]

where \( s \) and \( t \) are integers. \( \blacksquare \)

In order to prove, in the simplest form, certain results which will be used throughout this chapter, we have dealt very fully with a single stochastic differential equation with a white noise disturbance. But, from the point of view of econometrics, our main interest is with systems of equations. These introduce new problems. For the coefficients of a system of stochastic differential equations, representing a system of causal adjustment relations, will be subject to certain a priori restrictions derived from economic theory, and these will imply certain
restrictions on the coefficients of the derived system of difference equations used for estimation purposes. Because of the complexity of the latter restrictions and the fact that they cannot be inferred directly from economic theory, the continuous time formulation of the model is important, even if our ultimate aim is only to predict the future discrete observations of the variables.

We shall consider the system:

\[
dy(t) = \left[ A(\theta) y(t) + b(\theta) \right] dt + \xi(dt),
\]

where \( y(t) = [y_1(t), \ldots, y_n(t)]' \) is a vector whose elements are random processes, \( A(\theta) \) is an \( n \times n \) matrix whose elements are functions of a vector \( \theta = [\theta_1, \ldots, \theta_p] \) of structural parameters and \( b(\theta) \) is a vector whose elements are functions of \( \theta \). We assume that \( p < n^2 \), so that the matrix \( A \) is restricted. In the simplest case, where the only a priori restrictions are that certain variables have no direct causal influence on certain other variables, the only restrictions on \( A \) are that certain specified elements of this matrix are zero, and \( \theta \) is then the vector of unrestricted elements of \( A \). With regard to the disturbance vector \( \xi(dt) \) we introduce the following assumption.

**Assumption 1**

\[
\xi = [\xi_1, \ldots, \xi_n]
\]

is a vector of random measures defined on all subsets of the line \( -\infty < t < \infty \) with finite Lebesgue measure, such that

\[
E[\xi(dt)] = 0, \quad E[\xi(dt)\xi'(dt)] = (dt)\Sigma,
\]

where \( \Sigma \) is a positive definite matrix.

Equation (28) will be interpreted as meaning that the vector random process \( y(t) \) satisfies the system:

\[
y(t) - y(0) = \int_0^t [A(\theta) y(r) + b(\theta)] dr + \int_0^t \xi(dr),
\]

for all \( t \). With respect to this system, we shall now prove a theorem which generalizes Theorem 2.

**Theorem 3**

If \( \xi \) satisfies Assumption 1, then:

(a) for any given \( n \times 1 \) vector \( y(0) \), the system (29) has a solution:

\[
y(t) = \int_0^t e^{(r-t)A(\theta)}\xi(dr) + e^{tA(\theta)}[y(0) + A^{-1}(\theta)b(\theta)]
\]

\[
- A^{-1}(\theta)b(\theta),
\]

(30)
where, for any matrix $A$, $e^A$ is defined by

$$e^A = I + \sum_{r=1}^{\infty} \frac{1}{r!} A^r;$$

(b) the solution (30) is unique in the class of mean square continuous vector processes, i.e. if $\hat{y}(t)$ is any other vector of mean square continuous processes satisfying (29) and $\hat{y}(0) = y(0)$, then (21) holds on any interval $[0, T]$;

(c) the solution (30) satisfies the system

$$y(t) = F(\theta) y(t-1) + g(\theta) + \varepsilon_t$$

(31)

difference equations, where

$$F = e^A, \quad g = (e^A - I) A^{-1} b, \quad E(\varepsilon_t) = 0,$$

$$E(\varepsilon_t \varepsilon'_t) = \int_0^1 e^{rA} \Sigma e^{rA'} dr = \Omega, \quad E(\varepsilon_s \varepsilon'_t) = 0, \quad s \neq t.$$

Proof

(a) For the purpose of the proof we shall assume that $A$ has distinct eigenvalues, $\lambda_1, \ldots, \lambda_n$, although this is not essential for the validity of the theorem. We then have:

$$A = H^{-1} \Lambda H,$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

and $H$ is a matrix whose columns are eigenvectors of $A$. Now define:

$$z(t) = Hy(t).$$

Then, from (29):

$$z(t) - z(0) = \int_0^t [\Lambda z(r) + Hb] dr + \int_0^t H\xi'(dr). \quad (32)$$
Clearly, $H_\xi$ is a vector of random measures, such that:

$$F \left[ (H_\xi(dt))(H_\xi(dt))' \right] = (dt) H \Sigma H'. $$

Each equation in the system (32) satisfies the conditions of Theorem 2, therefore, and, by a direct application of that theorem to each equation in the system, we obtain the solution:

$$z(t) = \int_0^t e^{(t-r)\Lambda} H_\xi(dr) + e^{\Lambda} \left[ z(0) + \Lambda^{-1} Hb \right] - \Lambda^{-1} Hb. $$  \hspace{1cm} (33)

Then, premultiplying (33) by $H^{-1}$, we obtain:

$$y(t) = \int_0^t H^{-1} e^{(t-r)\Lambda} H_\xi(dr) + H^{-1} e^{\Lambda} \left[ y(0) + H^{-1} \Lambda^{-1} Hb \right] - H^{-1} \Lambda^{-1} Hb. $$ \hspace{1cm} (34)

But:

$$H^{-1} e^{\Lambda} H = H^{-1} \left[ I + t\Lambda + \frac{t^2}{2} \Lambda^2 + \cdots \right] H $$

$$= I + t\Lambda H + \frac{t^2}{2} \Lambda HH^{-1} \Lambda H + \cdots $$

$$= I + tA + \frac{t^2}{2} A^2 + \cdots $$

$$= e^{\Lambda A}. $$

Equation (34) can, therefore, be written as (30).

(b) It follows from Theorem (2) that (33) is a unique solution to (32), and, hence, (34) or (30) is a unique solution to (29) in the class of mean square continuous vector processes.

(c) Let $z(t)$ be the vector random process defined by (33). Then

$$z(t) = e^{\Lambda} \int_0^t e^{(t-r)\Lambda} H_\xi(dr) + e^{\Lambda} e^{(t-1)\Lambda} \left[ z(0) + \Lambda^{-1} Hb \right] $$

$$- e^{\Lambda} \Lambda^{-1} Hb + (e^{\Lambda} - I) \Lambda^{-1} Hb + \int_{t-1}^t e^{(t-r)\Lambda} H_\xi(dr) $$

$$= e^{\Lambda} z(t-1) + (e^{\Lambda} - I) \Lambda^{-1} b + \int_{t-1}^t e^{(t-r)\Lambda} H_\xi(dr). $$  \hspace{1cm} (35)
Premultiplying (35) by $H^{-1}$, we obtain:

$$y(t) = Fy(t-1) + g + \epsilon_t,$$

where

$$\epsilon_t = \int_{t-1}^{t} e^{(t-r)A}\xi(dr).$$  \hfill (36)

It is clear from the definition of the integral that:

$$E(\epsilon_t) = E\left[\int_{t-1}^{t} e^{(t-r)A}\xi(dr)\right] = 0.$$

And, using the generalizations of (15) and (16) for a vector random measure, we obtain:

$$E(\epsilon_t\epsilon'_t) = E\left[\int_{t-1}^{t} e^{(t-r)A}\xi(dr)\int_{t-1}^{t} e^{(t-r)A}\xi(dr)\right]$$

$$- \int_{t-1}^{t} \left[e^{(t-r)A}\Sigma e^{(t-r)A'}\right]dr$$

$$= \int_{0}^{1} e^{rA}\Sigma e^{rA'}dr = \Omega,$$

$$E(\epsilon_s\epsilon'_t) = E\left[\int_{s-1}^{s} e^{(s-r)A}\xi(dr)\int_{t-1}^{t} e^{(t-r)A}\xi(dr)\right]$$

$$= 0, \quad s \neq t. \quad \Box$$

We shall refer to the system (31) as the exact discrete model corresponding to the continuous time model (29). It should be emphasized that, unlike the continuous time model from which it is derived, the exact discrete model is not a system of structural relations. It cannot be interpreted as a system of causal relations in which each equation describes the direct response of one variable to the stimulus provided by other variables in the system. For each coefficient in the matrix $F$ will reflect the interaction between all variables in the system during the observation period. Even if the only a priori restrictions on the matrix $A$ are that certain elements of this matrix are zero, in which case $\theta$ is a vector whose elements are the unrestricted elements of $A$, the elements of $F$ will be complicated transcendental functions of the elements of $\theta$ and will, generally, be all non-zero. And, even if $\Sigma$ is a diagonal matrix, the elements of $\Omega$ will, generally, be all non-zero.

The relation of the exact discrete model (31) to the continuous time model (29) is rather similar, therefore, to that of the reduced form of a simultaneous
equations model to the structural form of the model. And, as we shall see, the relation between the exact discrete model of a continuous time model and the reduced form of a simultaneous equations model, used to approximate the continuous time model, plays an important role in the analysis of the properties of various estimators.

2.3. Estimation

It is easy to see that a necessary and sufficient condition for the identifiability of the parameter vector $\theta$ in the model (29) is that the correspondence between $\theta$ and $[F(\theta), g(\theta)]$ is one to one. But this condition is more restrictive than it might, at first sight, appear to be. It is more restrictive than the condition that the correspondence between $\theta$ and $[A(\theta), b(\theta)]$ is one to one. For the equation

$$e^A = F \quad (37)$$

will, generally, not have a unique solution unless $A$ is restricted. This is because, if $A$ is a matrix satisfying (37) and some of its eigenvalues are complex, then by adding to each pair of conjugate complex eigenvalues the imaginary numbers $2i\pi n$ and $-2i\pi n$, respectively, where $n$ is an integer, we obtain another matrix satisfying (37). For identifiability the restrictions on $A$ must be sufficient to exclude any other matrix obtained in this way.

The real problem here is that, unless our model incorporates sufficient a priori restrictions we cannot distinguish between structures generating oscillations whose frequencies differ by integer multiples of the observation period. This phenomenon is known as aliasing. The identification problem is more complicated for continuous time models, therefore, than it is for discrete time models. For a fuller discussion of the identification problem the reader is referred to Phillips (1973) who derives a rank condition for identifiability in the case in which each a priori restriction on $A$ is a linear homogeneous relation between the elements of a row of $A$.$^1$ We shall assume throughout the rest of this section that $\theta$ is identifiable.

In the discussion of estimation methods we shall assume, initially, that the sample is of the form $y(1), \ldots, y(T)$ as it would be if all variables were stock variables or prices at points of time. The complications arising when some or all of the variables are observable only as integrals will be discussed later.

The problem of estimating $\theta$ is equivalent to the problem of estimating $[F, g]$ subject to the restriction that this matrix can be written as $[F(\theta), g(\theta)]$ for some vector $\theta$ in $p$-dimensional space (or the subset of this space to which $\theta$ is required to belong). As we have seen this restriction is very complicated, even in the

$^1$See also the recent contributions of Hansen and Sargent (1981, 1983).
simplest cases, and the computational problem of obtaining a consistent estimate of $\theta$ in a large model is such that it is worth considering methods based on an approximate discrete model. Such methods are likely to be useful in any research programme, at least for the preliminary screening of hypotheses.

An obvious approximation can be obtained from (29) by using $\frac{1}{2}[y(t)+y(t-1)]$ as an approximation for $\int_{t-1}^{t} y(r) \, dr$. This gives the approximate simultaneous equations model:

$$y(t) - y(t-1) = \frac{1}{2} A(\theta) [y(t) + y(t-1)] + b(\theta) + u_t,$$

$$E(u_t) = 0, \quad E(u_t u'_s) = \Sigma, \quad E(u_t u'_s) = 0, \quad s \neq t, \quad s, t = 1, 2, \ldots.$$ 

The model is approximate because, if $u_t$ is defined in such a way that (38) holds exactly, then the condition $E(u_t u'_s) = 0, s \neq t$, will be only approximately satisfied.

We can write the model (38) in the reduced form:

$$y(t) = \Pi_1(\theta) y(t-1) + \Pi_0(\theta) + v_t,$$

where

$$\Pi_1 = [I - \frac{1}{2} A]^{-1} [I + \frac{1}{2} A],$$

$$\Pi_0 = [I - \frac{1}{2} A]^{-1} b,$$

$$v_t = [I - \frac{1}{2} A]^{-1} u_t,$$

so that

$$E(v_t) = 0, \quad E(v_t v'_s) = [I - \frac{1}{2} A]^{-1} \Sigma [I - \frac{1}{2} A']^{-1},$$

$$E(v_t v'_s) = 0, \quad s \neq t, \quad s, t = 1, 2, \ldots.$$ 

The use of the approximate simultaneous equations model (38) is particularly convenient when the elements of $A(\theta)$ are linear functions of $\theta$. For then we can estimate $\theta$ by applying a non-iterative procedure, such as two-stage least squares or three-stage least squares, to this model, as if it were the true model. But even the application of the full information quasi-maximum likelihood procedure to (38) is computationally simpler than the application of the same procedure to the exact discrete model (31). Estimates obtained by any of these methods will, of course, be asymptotically biased because of the error of specification in the model (38). It is important, therefore, that we should investigate the sampling properties of these estimators when the data have been generated by the continuous time model (29) or, equivalently, by the exact discrete model (31).
Such an investigation was undertaken by Bergstrom (1966). The central idea which was put forward in this article, and further discussed in Bergstrom (1967, ch. 9), is that the restrictions on the matrix \([\Pi_1, \Pi_0]\) of reduced form coefficients of the approximate simultaneous equations model can be regarded as convenient approximations to the restrictions on the matrix \([F, g]\) of coefficients of the exact discrete model. In particular, if the elements of \([A, b]\) are linear functions of \(\theta\), then the elements of \([\Pi_1, \Pi_0]\) will be rational functions of \(\theta\) whereas the elements of \([F, g]\) will be complicated transcendental functions of \(\theta\). Some idea of the goodness of the approximation can be obtained by comparing the power series expansions:

\[
\Pi_1 = \left[ I + \frac{1}{2} A + \frac{1}{4} A^2 + \frac{1}{6} A^3 + \ldots \right] \left[ I + \frac{1}{2} A \right] \\
= I + A + \frac{1}{2} A^2 + \frac{1}{4} A^3 + \ldots
\]

and

\[
F = I + A + \frac{1}{2} A^2 + \frac{1}{6} A^3 + \ldots.
\]

It should be noted, however, that whereas the power series expansion of \(F\) is convergent for any matrix \(A\) that of \(\Pi_1\) is convergent only if the eigenvalues of \(A\) lie within the unit circle [see Halmos (1958, ch. 4)].

We shall now introduce two more assumptions.

**Assumption 2**

The vector process \(\int_0^t \xi(d\tau)\) is Gaussian.

**Assumption 3**

The eigenvalues of \(A(\theta^0)\) (where \(\theta^0\) is the true parameter vector) have negative real parts.

Assumption 2 is introduced in order to ensure that the disturbance vectors \(e_t, t = 1, 2, \ldots, \) in the system (31) are independently and identically distributed. The fact that it implies that they are normally distributed is incidental. Once we have assumed that the orthogonal increments (corresponding to a sequence of intervals of equal length) in the process \(\int_0^t \xi(d\tau)\) are independently and identically distributed we are committed to assuming that they are normal. This can be seen by dividing the interval \([0, t]\) into \(n\) equal subintervals and applying the Lindberg-Levy central limit theorem [see Cramér (1951, p. 215)] to the sum \(\sum_{i=1}^n \int_{(i-1)/n}^{i/n} \xi(d\tau)\), when \(n \to \infty\).

Assumption 3 implies that the eigenvalues of \(F(\theta^0)\) lie within the unit circle. It follows, by applying the results of Mann and Wald (1943) to the system (31), that, under Assumptions 2 and 3, the sample mean vector \((1/T) \sum_{t=1}^T y(t)\) and sample
moment matrices \( (1/T) \sum_{t=1}^{T} y(t)y'(t) \) and \( (1/T) \sum_{t=1}^{T} y(t)y'(t-1) \) converge in probability, as \( T \to \infty \), to limits which do not depend on \( y(0) \) and that \( (1/\sqrt{T}) \sum_{t=1}^{T} y(t) \varepsilon_i' \) has a limiting normal distribution. In establishing these results Mann and Wald assumed that \( \varepsilon_1, \varepsilon_2, \ldots \) have finite fourth moments. Although Assumption 2 ensures that this condition is satisfied it is now known to be unnecessary [see Anderson (1959) and Hannan (1970, ch. 6)].

Since the probability limits of the sample moments of \( y(t) \) can be expressed as functions of \( F, g \) and \( \Omega \), and hence as functions of \( \theta \) and \( \Sigma \), we can, in principle, find a formula for the asymptotic bias of any estimator of \( \theta \) which can be expressed as a vector of rational functions of the sample moments. This is the case with the estimator obtained by applying two-stage least squares or three-stage least squares to the approximate simultaneous equations model (38). The formula would express the asymptotic bias of such an estimator in terms of the parameters of the continuous time model, i.e. the elements of \( \theta \) and \( \Sigma \). It would, of course, be very cumbersome if written out explicitly. But it is implicit in the calculations of Bergstrom (1966) who derives the asymptotic bias and approximate sampling variances of the estimates obtained by applying three-stage least squares to the approximate simultaneous equations model when the data are generated by a three equation continuous time model of the form (29).

In this example it is assumed that \( b = 0, \Sigma = I \) and that the only restrictions on \( A \) are that three of its elements are zero so that \( \theta \) is a vector of the unrestricted elements of \( A \). The assumed matrix \( A \) and derived matrix \( F \) are:

\[
A = \begin{bmatrix}
-1.0 & 0.8 & 0.0 \\
0.0 & -0.5 & 0.2 \\
0.1 & 0.0 & -0.2
\end{bmatrix},
\]

\[
F = \begin{bmatrix}
0.369 & 0.382 & 0.046 \\
0.006 & 0.608 & 0.142 \\
0.056 & 0.023 & 0.820
\end{bmatrix}.
\]

The interpretation of \( A \), assuming that the time unit is 3 months, is that \( y_1 \) is causally dependent on \( y_2, y_2 \) on \( y_3 \) and \( y_3 \) on \( y_1 \) with mean time lags of 3 months, 6 months and 15 months, respectively. The probability limits of the estimators \( \hat{A} \) and \( \hat{\Pi}_1 \) obtained by applying three-stage least squares to the approximate simultaneous equations model of the form (38) are:

\[
\text{plim} \hat{A} = \begin{bmatrix}
-0.922 & 0.710 & 0.000 \\
0.000 & -0.488 & 0.193 \\
0.098 & 0.000 & -0.199
\end{bmatrix},
\]

\[
\text{plim} \hat{\Pi}_1 = \begin{bmatrix}
0.370 & 0.391 & 0.034 \\
0.005 & 0.609 & 0.141 \\
0.061 & 0.017 & 0.821
\end{bmatrix}.
\]
It is interesting to note that the estimated reduced form matrix \( \hat{\mathbf{I}}_1 \) provides a remarkably good estimator of the matrix \( \mathbf{F} \) of coefficients in the exact discrete model, whereas \( \hat{\mathbf{A}} \) is a somewhat less satisfactory estimator of \( \mathbf{A} \). A heuristic explanation of this is that, even if there were no a priori restrictions on \( \mathbf{A} \), \( \hat{\mathbf{A}} \) would be an asymptotically biased estimator of this matrix whereas \( \hat{\mathbf{I}}_1 \) would, in this case, be identical with the least squares estimator \( \mathbf{F}^* \) and, therefore, a consistent estimator of \( \mathbf{F} \). [See Bergstrom (1966, 1967) for a further discussion of this point and a proposed two stage estimator of \( \mathbf{A} \) based on \( \hat{\mathbf{I}}_1 \).]

Since it is the matrix \( \mathbf{F} \) which is of interest for the purpose of predicting future discrete observations, it is important to consider the question of whether or not it would be better, for this purpose, to use the least squares estimator \( \mathbf{F}^* \) when \( \mathbf{A} \) is restricted. Since \( \mathbf{F}^* \) is a consistent estimator of \( \mathbf{F} \) while \( \hat{\mathbf{I}}_1 \) is not, it would always be better to use \( \mathbf{F}^* \) rather than \( \hat{\mathbf{I}}_1 \) if the sample size were sufficiently large. But with smaller samples the bias in any element of \( \hat{\mathbf{I}}_1 \) (as an estimator of the corresponding element of \( \mathbf{F} \)) will be more than outweighed by its lower variance as compared with the variance of the corresponding element of \( \mathbf{F}^* \). Calculations presented in Bergstrom (1966) show that for the above example with a sample of 100 observations the reduction in the variance obtained by using \( \hat{\mathbf{I}}_1 \) rather than \( \mathbf{F}^* \) heavily outweighs the squared asymptotic bias in any element of \( \hat{\mathbf{I}}_1 \).

The results of the above study suggest that the simultaneous equations model (38) is likely to be a useful approximation for the purpose of estimating the parameters of the underlying continuous time model from quarterly observations, and that the predictions obtained from the reduced form of this model, when the structural parameters are estimated by three-stage least squares, are likely to be better than those obtained from the ordinary least squares estimates of the coefficient of the exact discrete model ignoring the a priori restrictions. But there is, clearly, a need for a more general study, comparing the sampling properties of various estimators, applied to various approximate discrete models. An important step in this direction was taken by Sargan (1974, 1976). He generalizes the model (29) by including exogenous variables and considers the asymptotic bias of estimators obtained by applying the methods of two-stage least squares, three-stage least squares and full information maximum likelihood to the approximate simultaneous equations model (38), extended to include exogenous variables. He shows, in particular, that the proportional asymptotic bias of all of these estimators is of the same order of smallness as the square of the observation period as this tends to zero.

The econometrician cannot, of course, obtain observations of macroeconomic variables at arbitrary small intervals of time. He must, generally, do the best that he can with quarterly observations of such variables as the gross national product and its components. But the results of the study by Bergstrom (1966), which assumes a realistic pattern of time lags and quarterly observations, suggest that Sargan's criterion may, nevertheless, be useful for the ranking of various estima-
tors and various approximate discrete models. Since Sargan uses only one
approximate discrete model, and the asymptotic bias of each of the three
estimators considered by him is of the same order of smallness, the significance of
his results could, easily, be overlooked. Before proving his basic result, therefore,
we shall apply his method to an even simpler approximate discrete model, which
has been more widely used than (38). This is the model:

\[ y(t) - y(t - 1) = A(\theta) y(t - 1) + b(\theta) + u_t, \quad (40) \]

\[ E(u_t) = 0, \quad E(u_t u_t') = \Sigma, \quad E(u_s u_t') = 0, \quad s \neq t, \quad s, t = 1, 2, \ldots. \]

We shall show that estimates obtained from the model (40) will be inferior to
those obtained from (38) if the observation period is sufficiently short and the
data are generated by (29).

We assume, for this purpose, that \( b = 0 \) and that the only other a priori
restrictions are that certain elements of \( A \) are zero so that \( \theta \) is the vector of
unrestricted elements of \( A \). The continuous time model (28) can then be written:

\[ dy_i(t) = \theta^{(i)}(t)r_i(t) dt + \xi_i(t) dt, \quad i = 1, \ldots, n, \quad (41) \]

where \( y_i(t) \) is the \( i \)th element of \( y(t) \), \( \theta^{(i)} \) is the vector of unrestricted elements
of the \( i \)th row of \( A \) and \( y^{(i)}(t) \) is a vector of the corresponding elements of \( y(t) \).
The system (29), by which we give a precise interpretation of (28), can be written:

\[ y_i(t) - y_i(0) = \int_0^t \theta^{(i)}(r)y^{(i)}(r) dr + \int_0^t \xi_i(t) dr, \quad i = 1, \ldots, n. \quad (42) \]

Following Sargan we shall keep the time unit constant and denote the observation
period by \( \delta \) so that we can consider the behaviour of our estimators as \( \delta \to 0 \)
while keeping the elements of \( \theta \) constant. Then, defining \( y_r = y(r \delta) \), the exact
discrete model is:

\[ y_t = e^{\delta A} y_{t-1} + \epsilon_t, \quad (43) \]

\[ E(\epsilon_t) = 0, \quad E(\epsilon_t \epsilon_t') = \int_0^\delta e^{rA} \Sigma e^{rA'} dr, \]

\[ E(\epsilon_s \epsilon_t') = 0, \quad s \neq t, \quad s, t = 1, 2, \ldots. \]

The approximate discrete model (40) can be written:

\[ \frac{y_{it} - y_{i,t-1}}{\delta} = \theta^{(i)}(0)y_{i,t-1} + u_{it}, \quad i = 1, \ldots, n. \quad (44) \]

We can now show that the asymptotic bias of the estimator \( \theta^* \) obtained by
applying ordinary least squares to each equation of the system (44) is $O(\delta)$ as $\delta \to 0$.

**Theorem 4**

If $\theta^*$ is the estimator obtained from a sample $y_1, y_2, \ldots, y_T$ [i.e. $y(\delta), y(2\delta), \ldots, y(T\delta)$] of vectors generated by (42) by applying ordinary least squares to each equation of (44), then, under Assumptions 1, 2 and 3:

$$\lim_{T \to \infty} \theta^* - \theta = O(\delta), \quad \text{as} \quad \delta \to 0.$$  

**Proof**

From (43) we obtain:

$$\frac{1}{\delta} (y_t - y_{t-1}) = \left[ A + \frac{\delta}{2} A^2 + \frac{\delta^2}{3!} A^3 + \cdots \right] y_{t-1} + \frac{1}{\delta} \epsilon_t = Ay_{t-1} + Hy_{t-1} + \frac{1}{\delta} \epsilon_t, \quad (45)$$

where

$$H = O(\delta).$$

The system (45) can be written:

$$\frac{y_{it} - y_{i,t-1}}{\delta} = \theta^{(i)} y_{t-1}^{(i)} + h_i' y_{t-1} + \frac{\epsilon_{it}}{\delta}, \quad i = 1, \ldots, n, \quad (46)$$

where $h_i'$ is the $i$th row of $H$ and $\epsilon_{it}$ is the $i$th element of $\epsilon_t$. Then the estimator $\theta^*(i)$ obtained by applying ordinary least squares to the $i$th equation of (44) is:

$$\theta^*(i) = \left[ \frac{1}{T} \sum_{t=1}^{T} \left( \frac{y_{it} - y_{i,t-1}}{\delta} \right) y_{t-1}^{(i)} \right] \left[ \frac{1}{T} \sum_{t=1}^{T} y_{t-1}^{(i)} y_{t-1}^{(i)} \right]^{-1}$$

$$= \left[ \frac{1}{T} \sum_{t=1}^{T} \left( \theta^{(i)} y_{t-1}^{(i)} + h_i' y_{t-1} + \frac{\epsilon_{it}}{\delta} \right) y_{t-1}^{(i)} \right] \left[ \frac{1}{T} \sum_{t=1}^{T} y_{t-1}^{(i)} y_{t-1}^{(i)} \right]^{-1}$$

$$= \theta^{(i)} + \left[ \frac{1}{T} \sum_{t=1}^{T} \left( h_i' y_{t-1} + \frac{\epsilon_{it}}{\delta} \right) y_{t-1}^{(i)} \right] \left[ \frac{1}{T} \sum_{t=1}^{T} y_{t-1}^{(i)} y_{t-1}^{(i)} \right]^{-1}.$$  

But, from the Mann and Wald results:

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} y_{t-1}^{(i)} \epsilon_{it} = 0$$
and
\[ \text{plim} \frac{1}{T} \sum_{t=1}^{T} y_{t-1}y_{t-1}' \text{ exists.} \]

Therefore:
\[ \text{plim} \theta^{* (i)} - \theta^{(i)} = O(h_t) = O(\delta), \quad \text{as } \delta \to 0. \]

We shall consider, next, estimates obtained by using the approximate simultaneous equations model (38). When the continuous time model is (42), the system (38) can be written:

\[ \frac{y_{it} - y_{i,t-1}}{\delta} = \theta^{(i)} \left[ \frac{1}{2} \left( y_t^{(i)} + y_{t-1}^{(i)} \right) \right] + u_{it}, \quad i = 1, \ldots, n. \]  

(47)

We shall prove a theorem which includes, as a special case, Sargan's basic theorem (when there are no exogenous variables).

**Theorem 5**

Let \( \bar{\theta}^{(i)} \) be the instrumental variables estimator, defined by:

\[ \bar{\theta}^{(i)} = \left[ \frac{1}{T} \sum_{t=1}^{T} \left( \frac{y_{it} - y_{i,t-1}}{\delta} \right) z_t^{(i)} \right] \left[ \frac{1}{T} \sum_{t=1}^{T} \frac{1}{2} \left( y_t^{(i)} + y_{t-1}^{(i)} \right) z_t^{(i)} \right]^{-1}, \]

where \( y_1, \ldots, y_T \) [i.e. \( y(\delta), \ldots, y(T\delta) \)] are vectors generated by (42) and \( z_1^{(i)}, \ldots, z_T^{(i)} \) are random row vectors such that:

\[ \text{plim} \frac{1}{T} \sum_{i=1}^{T} e_{it} z_t^{(i)} = 0, \]

while

\[ \text{plim} \frac{1}{T} \sum_{t=1}^{T} y_t^{(i)} z_t^{(i)} \quad \text{and} \quad \text{plim} \frac{1}{T} \sum_{t=1}^{T} y_{t-1}^{(i)} z_t^{(i)} \text{ exist.} \]

Then, under Assumptions 1, 2 and 3:

\[ \text{plim} \bar{\theta}^{(i)} - \theta^{(i)} = O(\delta^2), \quad \text{as } \delta \to 0. \]
Proof

Using (43), we obtain:

\[
\frac{1}{\delta}(y_t - y_t) = \left[A + \frac{\delta}{2}A^2 + \frac{\delta^2}{3!}A^3 + \cdots \right]y_{t-1} + \frac{1}{\delta} \varepsilon_t
\]

\[
- \frac{1}{\delta}Ay_{t-1} + \frac{1}{\delta}A\left[I + \delta A + \frac{\delta^2}{3}A^2 + \cdots \right]y_{t-1} + \frac{1}{\delta} \varepsilon_t
\]

\[
= \frac{1}{\delta}Ay_{t-1} + \frac{1}{\delta}Ae^\delta Ay_{t-1} + Ly_{t-1} + \frac{1}{\delta} \varepsilon_t
\]

\[
= A\left[\frac{1}{\delta}(y_t + y_{t-1})\right] + Ly_{t-1} + \left(\frac{1}{\delta}I - \frac{1}{\delta}A\right)\varepsilon_t,
\]

where

\[
L = O(\delta^2).
\]

Therefore

\[
\frac{y_{it} - y_{i,t-1}}{\delta} = \frac{1}{\delta} \theta^{(i)}(y^{(i)}_t + y^{(i)}_{t-1}) + l'_iy_{t-1} + \frac{1}{\delta} \varepsilon_{it} - \frac{1}{2} \theta^{(i)} \varepsilon_t,
\]

where \(l'_i\) is the \(i\)th row of \(L\). And, hence:

\[
\bar{\theta}^{(i)} = \left[\frac{1}{T} \sum_{t=1}^{T} \left\{ \frac{1}{2} \theta^{(i)}(y^{(i)}_t + y^{(i)}_{t-1}) + l'_iy_{t-1} + \frac{1}{\delta} \varepsilon_{it} - \frac{1}{2} \theta^{(i)} \varepsilon_t \right\} z^{(i)}_t \right]
\]

\[
\times \left[\frac{1}{T} \sum_{t=1}^{T} \frac{1}{2} \left(y^{(i)}_t + y^{(i)}_{t-1}\right) z^{(i)}_t \right]^{-1}
\]

\[
= \theta^{(i)} + \left[\frac{1}{T} \sum_{t=1}^{T} (l'_iy_{t-1} + \frac{1}{\delta} \varepsilon_{it} - \frac{1}{2} \theta^{(i)} \varepsilon_t) z^{(i)}_t \right]\left[\frac{1}{T} \sum_{t=1}^{T} \frac{1}{2} \left(y^{(i)}_t + y^{(i)}_{t-1}\right) z^{(i)}_t \right]^{-1}.
\]

Therefore

\[
\lim_{T \to \infty} \bar{\theta}^{(i)} - \theta^{(i)} = O(\delta^2), \quad \text{as } \delta \to 0.
\]

The two-stage least squares estimator is obtained as a special case of the estimator \(\bar{\theta}^{(i)}\) by putting:

\[
z^{(i)}_t = \frac{1}{2} \left(y^{*^{(i)}}_t + y^{(i)}_{t-1}\right)',
\]

where

\[
y^{*^{(i)}}_t = \left[\frac{1}{T} \sum_{t=1}^{T} y^{(i)}_t y'_{t-1}\right]\left[\frac{1}{T} \sum_{t=1}^{T} y_{t-1} y'_{t-1}\right]^{-1} y_{t-1}.
\]
It is fairly obvious that the above argument can be extended to show that the asymptotic bias of the three-stage least squares estimator of \( \theta \) is \( O(\delta^2) \). Sargan (1976) shows that the asymptotic bias of the full information maximum likelihood estimator [applied to (47)] is also \( O(\delta^2) \) and that the difference between the limits in probability of the three-stage least squares and full information maximum likelihood estimators is \( O(\delta^5) \). He also finds sufficient conditions for these results to hold when the model contains exogenous variables. These conditions will be given in Section 4.

We turn now to the problem of finding consistent and asymptotically efficient estimators of \( \theta \) from discrete data generated by (29). For this purpose the following additional assumptions are introduced.

**Assumption 4**

It is known that \( \theta \) belongs to a compact subset \( \Theta \) of \( p \)-dimensional space.

**Assumption 5**

Let \( \Psi \) be the subset of \( n \times (n + 1) \) matrices which can be written in the form \([F(\theta), g(\theta)]\) for some \( \theta \in \Theta \), where

\[
F(\theta) = e^{A(\theta)}, \quad g(\theta) = [e^{A(\theta)} - I]A^{-1}(\theta)b(\theta)
\]

Then the mapping from \( \Psi \) to \( \Theta \) defined by the inverse of \([F(\theta), g(\theta)]\) is one to one and continuous in a neighbourhood of the true parameter vector \( \theta^0 \); i.e. every sequence \( \theta^n, n = 1, 2, \ldots \), of vectors in \( \Theta \) such that \([F(\theta^n), g(\theta^n)] \rightarrow [F(\theta^0), g(\theta^0)]\) converges to \( \theta^0 \) as \( T \rightarrow \infty \).

**Assumption 6**

The set \( \Theta \) contains a neighbourhood of \( \theta^0 \) in which the derivatives up to the third order of the elements of \([F(\theta), g(\theta)]\) are bounded. The vector \( \theta^0 \) is not a singularity point of \([F(\theta), g(\theta)]\); i.e. there is no set of numbers \( \lambda_1, \ldots, \lambda_p \), not all zero, such that:

\[
\sum_{k=1}^{p} \lambda_k \frac{\partial}{\partial \theta_k} [F(\theta^0), g(\theta^0)] = 0.
\]

We shall consider, first, the minimum distance estimator \( \hat{\theta}^{**} \) which is defined as the vector \( \theta \in \Theta \) that minimises:

\[
\frac{1}{T} \sum_{t=1}^{T} [y(t) - F(\theta)y(t-1) - g(\theta)]'M_{ee}^{*^{-1}}[y(t) - F(\theta)y(t-1) - g(\theta)],
\]
where

\[ M_{ee}^* = \frac{1}{T} \sum_{t=1}^{T} [y(t) - F^* y(t-1) - g^*)(y(t) - F^* y(t-1) - g^*)]' \]

and \( F^* \) and \( g^* \) are the ordinary least squares estimators of \( F \) and \( g \) respectively. The properties of this estimator have been studied by Malinvaud (1970, ch. 9) when applied to the model:

\[ y_t = A(\theta) x_t + \epsilon_t, \]

where \( A \) is a matrix of non-linear functions of the parameter vector and \( x_t, t = 1, 2, \ldots, \) is a sequence of non-random vectors. Since the model (48) contains no lagged dependent variables we cannot rely on Malinvaud's results for inferring the properties of the minimum distance estimator when applied to the model (31). But, by using the Mann and Wald results and modifying Malinvaud's proofs, we can prove Theorems 6 and 7 which, together, correspond to theorem 5 of Malinvaud (1970, ch. 9).

**Theorem 6**

Under Assumptions 1–5:

\[ \lim_{T \to \infty} \theta^{**} = \theta^0 \]

**Theorem 7**

Under Assumptions 1–6, \( \sqrt{T}(\theta^{**} - \theta^0) \) has a limiting normal distribution whose covariance matrix is the limit is probability, as \( T \to \infty \), of the inverse of the matrix whose \((k l)\)th element is:

\[
\text{tr} \left( \frac{\partial}{\partial \theta_k} [F(\theta), g(\theta)]' M_{ee}^* -1 \frac{\partial}{\partial \theta_l} [F(\theta), g(\theta)] \right) \times \left[ \begin{array}{cc}
\frac{1}{T} \sum_{t=1}^{T} y(t) y'(t) & \frac{1}{T} \sum_{t=1}^{T} y(t) \\
\frac{1}{T} \sum_{t=1}^{T} y'(t) & 1
\end{array} \right].
\]

(49)
Since the logarithm of the likelihood function is:

\[
L(\theta, \Omega) = -\frac{T}{2} \log 2\pi - \frac{1}{2} \log |\Omega| - \frac{1}{2} \sum_{t=1}^{T} \left[ y(t) - F(\theta) y(t-1) - g(\theta) \right] \Omega^{-1} 
\times \left[ y(t) - F(\theta) y(t-1) - g(\theta) \right],
\]

(50)

it follows from (49) that the covariance matrix of the limiting distribution of \(\sqrt{T}(\theta^{**} - \theta^0)\) is:

\[
- \left[ E\left( \frac{\partial^2 L}{\partial \theta_k \partial \theta_l} \right) \right]^{-1}.
\]

The estimator \(\theta^{**}\) is asymptotically efficient, therefore, in the sense defined by Cramér (1946, ch. 32).

For the purpose of predicting the future discrete observations of \(y(t)\), we are interested in the estimator

\[
[F^{**}, g^{**}] = [F(\theta^{**}), g(\theta^{**})]
\]

of the matrix of coefficients of the exact discrete model (31). By using (49) and the argument of Malinvaud (1970, p. 357) we can show that the concentration ellipsoid \(E^{**}\) [in \(n(n+1)\) dimensional space] of the limiting distribution of \([\sqrt{T}(F^{**} - F^0), \sqrt{T}(g^{**} - g^0)]\) is the set of \(n \times (n + 1)\) matrices \([F, g]\) that can be written in the form:

\[
[F, g] = \sum_{k=1}^{p} \theta_k \frac{\partial}{\partial \theta_k} [F(\theta^0), g(\theta^0)],
\]

(51)

for some vector \([\theta_1, \ldots, \theta_p]\) and satisfy the inequality

\[
\text{tr} \left[ F, g \right] \Omega^{-1} \left[ F, g \right] \leq 1.
\]

(52)

Since the concentration ellipsoid \(E^{*}\) of the limiting distribution of \([\sqrt{T}(F^{*} - F^0), \sqrt{T}(g^{*} - g^0)]\) is the set of all matrices \([F, g]\) satisfying (52) we have:

\[E^{**} \subset E^{*}\.

In geometrical terms, \(E^{**}\) is the intersection of \(E^{*}\) with the hyperplane of matrices defined by (51), i.e. it is the intersection of \(E^{*}\) with the hyperplane of
matrices satisfying the restrictions implied by the continuous time model in the
neighbourhood of \([F^0, g^0] = [F(\theta^0), g(\theta^0)].\)

This result implies that the asymptotic standard error of any element of
\([F**, g**]\) (or any linear combination of such elements) is at least as small as the
asymptotic standard error of the corresponding element of \([F^*, g^*]\) (or linear
combination of such elements). For it follows from the invariance property of the
concentration ellipsoid under a linear transformation [see Malinvaud (1970, ch. 5,
Lemma 1)] that the asymptotic standard errors of identical linear combinations of
elements of \([F**, g**]\) and \([F^*, g^*]\) can be compared by comparing the images of
\(E**\) and \(E^*\), respectively, under a linear transformation which transforms \(n(n + 1)
dimensional space into the appropriate one dimensional subspace. Provided,
therefore, that the sample size is sufficiently large we can obtain better predictions
of the future discrete observations by using the continuous time model than by
using the unrestricted least squares estimates of the coefficients of the exact
discrete model.

Finally we shall consider, very briefly, the maximum likelihood estimator. This
is obtained by maximising \(L(\theta, \Omega)\), as defined by (50), with respect to \(\theta\) and \(\Omega\).
We can do this in two stages. We first maximise \(L(\theta, \Omega)\) with respect to \(\Omega\) to
obtain \(\hat{\Omega}(\theta)\) and then substitute into \(L(\theta, \Omega)\) to obtain the concentrated likeli-
hood function:

\[
L(\theta) = L(\theta, \hat{\Omega}(\theta)) = c - \log|M(\theta)|,
\]

where \(c\) is a constant, \(|M|\) is the determinant of \(M\) and

\[
M(\theta) = \frac{1}{T} \sum_{t=1}^{T} [y(t) - F(\theta) y(t-1) - g(\theta)] [y(t) - F(\theta) y(t-1) - g(\theta)]'.
\]

Then the maximum likelihood estimator \(\hat{\theta}\) is the \(\theta \in \Theta\) that maximises \(L(\theta)\). The
estimation equations, obtained by equating to zero the partial derivatives of \(L(\theta)\)
with respect to \(\theta_1, \ldots, \theta_p\), are:

\[
\text{tr} M^{-1}(\theta) \frac{\partial}{\partial \theta_k} M(\theta) = 0, \quad k = 1, \ldots, p. \tag{53}
\]

The estimation equations for the minimum distance estimator \(\theta^{**}\) are:

\[
\text{tr} M^{-1}_{ee} \frac{\partial}{\partial \theta_k} M(\theta) = 0, \quad k = 1, \ldots, p. \tag{54}
\]

The system (54) is easier to solve than (53) since the matrix \(M^{-1}(\theta)\) in (53)
involves the unknown elements of \(\theta\), for which we are solving, whereas the matrix
\(M^{-1}_{ee}\) in (54) can be computed as an initial step without iteration. Under
Assumptions 1–6 \(\sqrt{T}(\hat{\theta} - \theta^0)\) has the same limiting distribution as \(\sqrt{T}(\theta^{**} - \theta^0)\).
This follows from the results of Dunsmuir and Hannan (1976) who consider a very general model which includes the exact discrete model (31) as a special case. We shall consider their results in more detail at a later stage.

It is of interest to compare the estimator \( \theta^{**} \) (or \( \hat{\theta} \)) obtained by using the exact discrete model with an estimator obtained by applying either the three-stage least squares or full information maximum likelihood method to an approximate simultaneous equations model. We know that if the sample size is sufficiently large the estimator obtained by using the exact discrete model will be better, since it is consistent, whereas the estimator obtained by using the approximate simultaneous equations model has an asymptotic bias of the order \( \delta^2 \), as we have seen. But, as the study of Bergstrom (1966) showed, for samples of the size available in practice the squared asymptotic bias in an estimator obtained by using the approximate simultaneous equations model can be small compared with the sampling variance.

A comparison of estimates obtained, from finite samples, by using the exact discrete model and an approximate simultaneous equations model was undertaken by Phillips (1972) who, for this purpose, wrote the first computer program for obtaining consistent and asymptotically efficient estimates of the parameters of a continuous time model, of the form (29), from discrete data. This program was for the computation of the minimum distance estimator \( \theta^{**} \), using the exact discrete model. The first program for the computation of the more complicated maximum likelihood estimator \( \hat{\theta} \), using the exact discrete model, was written by Wymer (1974) and applied to the model of Bergstrom and Wymer (1976) which will be discussed later. The main difficulty in computing either \( \theta^{**} \) or \( \hat{\theta} \) as compared with estimators obtained from an approximate simultaneous equations model is that \( F(\theta) \) must be expressed as a series of matrices and summed to a sufficient number of terms to give the desired degree of accuracy.

Phillips (1972) applied his program, in a Monte Carlo study, to a three equation trade cycle model based on the model of Phillips (1954) [see Bergstrom (1967, ch. 3)]. The model, in its deterministic form, is:

\[
\begin{align*}
DC(t) &= \alpha[(1 - s)Y(t) + a - C(t)], \\
DY(t) &= \lambda[C(t) + DK(t) - Y(t)], \\
DK(t) &= \gamma[uY(t) - K(t)].
\end{align*}
\]  

(55)  

(56)  

(57)

where \( C = \) consumption, \( Y = \) income and \( K = \) capital. By adding white noise disturbances and substituting for \( DK \) from (57) into (56) the model can be written:

\[
dy(t) = A(\theta) y(t) dt + b(\theta) + \xi(dt),
\]  

(58)

where

\[
\theta = (\alpha, \gamma, \lambda, s, v, a).
\]
A hundred synthetic samples each of 25 observations were generated by the exact discrete model derived from (58) and used in the estimation of $\theta$, both by applying the minimum distance estimation procedure to the exact discrete model and applying three-stage least squares to an approximate simultaneous equations model. The results are shown in Table 2.1. As can be seen, the estimates obtained by using the exact discrete model are, generally, superior to those obtained by using the approximate simultaneous equations model. Moreover, considering the smallness of the sample, the number of times that the 5\% confidence interval, computed from the estimated asymptotic standard errors, does not include the true value of the parameter is, for the estimates obtained by applying the minimum distance procedure to the exact discrete model, remarkably close to 5 (i.e. 7 for the parameters $\alpha$, $\gamma$, $\lambda$ and $s$, and 9 for $\nu$).

In the above example Phillips assumed the existence of point observations of all three variables. But the variables $C(t)$ and $Y(t)$ are flow variables and, in practice, could be observed only as the integrals $\int_{t-1}^{t} C(r) \, dr$ and $\int_{t-1}^{t} Y(r) \, dr$, $t = 1, 2, \ldots$. This does not, of course detract from the value of his study for the general purpose of comparing estimates derived from point observations using the exact discrete model and approximate simultaneous equations model. Moreover, at the time when the study was undertaken, the theoretical problems of obtaining consistent and asymptotically efficient estimates of the parameters of a continuous time stochastic model, of the form (29), from flow data had not been seriously studied. This is the problem to which we now turn. The essential difficulty is that, even when the continuous time model is a first-order system with white noise disturbances, the disturbances in the exact discrete model satisfied by the integral observations will be autocorrelated. The precise form of the autocorrelation is given in the following theorem.

Table 2.1

<table>
<thead>
<tr>
<th>Parameter:</th>
<th>$\alpha$</th>
<th>$\gamma$</th>
<th>$\lambda$</th>
<th>$s$</th>
<th>$\nu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True value:</td>
<td>0.6</td>
<td>0.4</td>
<td>4.0</td>
<td>0.25</td>
<td>2.0</td>
</tr>
</tbody>
</table>

**Minimum distance**

<table>
<thead>
<tr>
<th>Mean of the estimates</th>
<th>0.5734</th>
<th>0.4016</th>
<th>4.0709</th>
<th>0.2537</th>
<th>2.0021</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation of the estimates</td>
<td>0.1410</td>
<td>0.0153</td>
<td>0.7077</td>
<td>0.0259</td>
<td>0.0149</td>
</tr>
<tr>
<td>Root mean square error</td>
<td>0.1435</td>
<td>0.0154</td>
<td>0.7112</td>
<td>0.0262</td>
<td>0.0150</td>
</tr>
<tr>
<td>Number of wrong intervals</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

**Three-stage least squares**

<table>
<thead>
<tr>
<th>Mean of the estimates</th>
<th>0.6652</th>
<th>0.4182</th>
<th>2.7444</th>
<th>0.2767</th>
<th>1.995</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation of the estimates</td>
<td>0.1800</td>
<td>0.0241</td>
<td>0.8015</td>
<td>0.0937</td>
<td>0.0311</td>
</tr>
<tr>
<td>Root mean square error</td>
<td>0.1914</td>
<td>0.0302</td>
<td>1.4896</td>
<td>0.0974</td>
<td>0.0311</td>
</tr>
<tr>
<td>Number of wrong intervals</td>
<td>10</td>
<td>3</td>
<td>62</td>
<td>17</td>
<td>3</td>
</tr>
</tbody>
</table>

*a* Intervals not containing the true parameter value.
Theorem 8

If $\xi(dt)$ satisfies Assumption 1, $y(t)$ is the solution to (29) and $y_t$ is defined by:

$$y_t = \int_{t-1}^t y(r) \, dr,$$

then $y_t$ satisfies the system:

$$\begin{equation}
F = e^A, \quad g = (e^A - I)A^{-1}b,
\end{equation}$$

$$\begin{equation}
\Omega_0 = \int_0^1 A^{-1}(I - e^{\tau A}) \Sigma(I - e^{\tau A'}) A^{r-1} \, d\tau
+ \int_0^1 A^{-1}e^A(I - e^{(r-1)A}) \Sigma(I - e^{(r-1)A'}) e^{A' - 1} \, d\tau,
\end{equation}$$

$$\begin{equation}
\Omega_1 = \int_0^1 A^{-1}e^A(I - e^{(r-1)A}) \Sigma(e^{rA'} - I) A^{r-1} \, d\tau.
\end{equation}$$

Proof

From (29) we obtain:

$$y(t) - y(t-1) = A \int_{t-1}^t y(r) \, dr + b + \int_{t-1}^t \xi(dr),$$

and hence:

$$\int_{t-1}^t y(r) \, dr = A^{-1}[y(t) - y(t-1)] - A^{-1}b - A^{-1} \int_{t-1}^t \xi(dr).$$

Then, substituting from (31) into (63) and using (36), we obtain:

$$\begin{equation}
\int_{t-1}^t y(r) \, dr = A^{-1}F[y(t-1) - y(t-2)]
+ A^{-1} \left[ \int_{t-1}^t e^{(r-r')A} \xi(dr) - \int_{t-2}^{t-1} e^{(r-1-r')A} \xi(dr) \right]
- A^{-1}b - A^{-1} \int_{t-1}^t \xi(dr),
\end{equation}$$
and then, from (62) and (64):

\[
\int_{t-1}^{t} y(r) \, dr = A^{-1}FA \int_{t-2}^{t-1} y(r) \, dr + A^{-1}Fb + A^{-1} F \int_{t-2}^{t-1} \xi (dr) \\
+ A^{-1} \left[ \int_{t-1}^{t} e^{(t-r)A} \xi (dr) - \int_{t-2}^{t-1} e^{(t-1-r)A} \xi (dr) \right] \\
- A^{-1} b - A^{-1} \int_{t-1}^{t} \xi (dr),
\]

which, since

\[A^{-1} FA = A^{-1} \left[ I + A + \frac{1}{2} A^2 + \cdots \right] A = F\]

and

\[A^{-1} [F - I] b = [F - I] A^{-1} b = g,\]

reduces to (59) with \( \eta_t \) defined by:

\[
\eta_t = \int_{t-1}^{t} A^{-1} (e^{(t-r)A} - I) \xi (dr) \\
+ \int_{t-2}^{t-1} A^{-1} (e^{A} - e^{(t-1-r)A}) \xi (dr).
\]  \hspace{1cm} (65)

Finally, using the generalizations of (15) and (16) for a vector random measure we obtain:

\[
E(\eta_t, \eta_t') = \int_{t-1}^{t} A^{-1} (e^{(t-r)A} - I) \Sigma (e^{(t-r)A'} - I) A'^{-1} dr \\
+ \int_{t-2}^{t-1} A^{-1} (e^{A} - e^{(t-1-r)A}) \Sigma (e^{A'} - e^{(t-1-r)A'}) A'^{-1} dr \\
= \int_{0}^{1} A^{-1} (I - e^{A}A') \Sigma (I - e^{A'}) A'^{-1} dr \\
+ \int_{0}^{1} A^{-1} e^{A} (I - e^{(r-1)A}) \Sigma (I - e^{(r-1)A'}) e^{A'} A'^{-1} dr,
\]

\[
E(\eta_t, \eta_{t-1}) = \int_{t-2}^{t-1} A^{-1} (e^{A} - e^{(t-1-r)A}) \Sigma (e^{(t-1-r)A'} - I) A'^{-1} dr \\
= \int_{0}^{1} A^{-1} e^{A} (I - e^{(r-1)A}) \Sigma (e^{A'} - I) A'^{-1} dr. \quad \blacksquare
\]
It is clear from these results that \( \eta_t \) is a vector moving average process of the form:

\[
\eta_t = \epsilon_t + C \epsilon_{t-1},
\]

where

\[
E(\epsilon_t) = 0, \quad E(\epsilon_t \epsilon'_t) = K, \quad E(\epsilon_s \epsilon'_t) = 0, \quad s \neq t, \quad s, t = 1, 2, \ldots,
\]

and \( C \) and \( K \) satisfy the equations:

\[
K + CKC' = \Omega_0
\]

and

\[
CK = \Omega_1.
\]

Equations (67) and (68) imply that the elements of \( C \) and \( K \) are functions of the elements of \( \Omega_0 \) and \( \Omega_1 \) and, hence, of the elements of \( A \) and \( \Sigma \). The expressions (60) and (61) can be written as infinite series in ascending powers of \( A \) and \( A' \) by expressing the matrix exponential functions, in the integrals, in series form and integrating, term by term, with respect to \( r \). Evaluating the terms up to the first power of \( A \) in \( A' \) we obtain:

\[
\begin{align*}
\Omega_0 &= \frac{2}{3} \Sigma + \frac{1}{3} (A \Sigma + \Sigma A') + \cdots, \\
\Omega_1 &= \frac{1}{6} \Sigma + \frac{1}{8} A \Sigma + \frac{1}{24} \Sigma A' + \cdots.
\end{align*}
\]

Phillips (1978) shows that, if the observation period is \( \delta \), then:

\[
\begin{align*}
\Omega_0 &= \frac{2}{3} \delta^3 + \frac{\delta^4}{3} (A \Sigma + \Sigma A') + O(\delta^5), \\
\Omega_1 &= \frac{\delta^3}{6} + \frac{\delta^4}{8} A \Sigma + \frac{\delta^4}{24} \Sigma A' + O(\delta^5),
\end{align*}
\]

these equations being identical with (69) and (70) when \( \delta = 1 \). He also shows that:

\[
\begin{align*}
C &= \alpha I + \frac{\alpha \delta}{4} (A - \Sigma A' \Sigma^{-1}) + O(\delta^2), \\
K &= \frac{\delta^3}{6\alpha} \left[ \Sigma + \frac{\delta}{2} (A \Sigma + \Sigma A') \right] + O(\delta^2).
\end{align*}
\]
where \( \alpha = 0.268 \) is a root of the equation:

\[ z^2 - 4z + 1 = 0. \]

The first term on the right-hand side of (73) and (74) is easily obtained by substituting the first term on the right-hand side of (69) and (70) into (67) and (68) respectively and solving for \( C \) and \( K \).

It is convenient, at this stage, to write \( \Sigma \) as \( \Sigma(\mu) \) where \( \mu \) is a vector of parameters. If, as we have assumed so far, \( \Sigma \) is unrestricted then \( \mu \) will have \( n(n + 1)/2 \) elements. But we could, for example, require \( \Sigma \) to be a diagonal matrix in which case \( \mu \) would have \( n \) elements. We can now obtain the exact discrete model corresponding to the continuous time model (29) for the case in which the observations are in integral form. Combining (59) and (66) we obtain:

\[ y_t - F(\theta)y_{t-1} - g(\theta) = \varepsilon_t + C(\theta, \mu)e_{t-1}, \tag{75} \]

\[ E(\varepsilon_t) = 0, \quad E(\varepsilon_t \varepsilon_s') = K(\theta, \mu), \quad E(\varepsilon_t \varepsilon_s') = 0, \quad s \neq t, \quad s, t = 1, 2, \ldots. \]

An important point to notice is that, even though the covariance matrix \( \Sigma \) of the disturbance vector in the continuous time model does not depend on \( \theta \), the covariance matrix \( K \) of the random vector \( \varepsilon_t \) in the exact discrete model depends on \( \theta \) as well as \( \mu \), as is clear from (67) and (68).

If \( g \) is a zero vector (i.e., if \( b \), in the continuous time model, is a zero vector), then (75) is a special case of the model:

\[ y_t - \sum_{j=1}^{q} F_j(\theta, \mu)y_{t-j} = \varepsilon_t + \sum_{j=1}^{r} C_j(\theta, \mu)e_{t-j}, \tag{76} \]

\[ E(\varepsilon_t) = 0, \quad E(\varepsilon_t \varepsilon_s') = K(\theta, \mu), \quad E(\varepsilon_t \varepsilon_s') = 0, \quad s \neq t, \quad s, t = 1, 2, \ldots, \]

which was studied by Dunsmuir and Hannan (1976). They show, under certain assumptions, that the maximum likelihood estimator of \( \theta \), in the model (76), is strongly consistent (i.e., converges almost surely to \( \theta \)). And, for the case in which the matrices \( F_1, \ldots, F_q, C_1, \ldots, C_r \) do not depend on \( \mu \) and \( K \) does not depend on \( \theta \), so that these matrices can be written \( F_j(\theta), \ldots, F_q(\theta), C_j(\theta), \ldots, C_r(\theta) \) and \( K(\mu) \), they prove a central limit theorem. But, as we have shown, the matrices \( C \) and \( K \) in the model (75) depend on both \( \theta \) and \( \mu \). The case in which the matrices \( F_1, \ldots, F_q, C_1, \ldots, C_r \) and \( K \), in the model (76), all depend on both \( \theta \) and \( \mu \), was further considered by Dunsmuir (1979). Here he proves a central limit theorem for an estimator obtained by maximising an approximate likelihood function expressed in terms of the discrete Fourier transform of the data. [This estimator was proved to be strongly consistent by Dunsmuir and Hannan (1976).] His results imply that when \( \{ \varepsilon_t \} \) is Gaussian this estimator is asymptotically efficient.
(in the Cr˘amer sense). Kohn (1979) extends the model (76) by including exogenous variables (which cover the case in which \( g \), in (75), is not a zero vector); but he assumes that \( \Sigma \) is unrestricted. Although the properties of the exact maximum likelihood estimator \((\hat{\theta}, \hat{\mu})\) of the parameter vector \((\theta, \mu)\), in the model (75), cannot be definitely inferred from the results obtained in the above studies, these results leave little doubt that, under Assumptions 1–6 and the assumption that \( \Sigma \) is unrestricted, \((\hat{\theta}, \hat{\mu})\) is strongly consistent, asymptotically normal and asymptotically efficient.

In order to obtain the exact likelihood function for the parameters of (75) it is not necessary to find explicit formulae for \( C(\theta, \mu) \) and \( K(\theta, \mu) \). Instead we can work directly from the formulae for \( \Omega_0 \) and \( \Omega_1 \) which we have obtained from the continuous time model, thus avoiding the problem of solving (67) and (68). For this purpose we shall assume that \{ \( y(t) \) \} is stationary [i.e. that \( y(0) \), in the model (29), is not a fixed vector but a random vector generated by the application of this model over \(( - \infty, 0)\)]. Then defining

\[
m = E y(t),
\]

we obtain, from (59):

\[
m = Fm + g,
\]  

(77)

and hence

\[
m = m(\theta) = [I - F(\theta)]^{-1} g(\theta).
\]  

(78)

Then, from (59) and (77), we obtain:

\[
y_t - m(\theta) = F(\theta)[y_{t-1} - m(\theta)] + \eta_t,
\]  

(79)

whose solution is:

\[
y_t - m(\theta) = \sum_{j=0}^{\infty} [F(\theta)]^j \eta_{t-j}.
\]  

(80)

We note, incidentally, that (79) cannot be treated as a special case of the model (76), considered by Dunsmuir and Hannan, since \( y_t - m(\theta) \) is not observable.
Since $E(\eta_s \eta_t') = 0$, $|s-t| > 1$, we obtain, from (80):

$$E(y_s - m(\theta))(y_t - m(\theta))' = \sum_{j=0}^{\infty} [F(\theta)]^j \Omega_0(\theta, \mu)[F'(\theta)]^{t-s+j}$$

$$+ \sum_{j=0}^{\infty} [F(\theta)]^j \Omega_1(\theta, \mu)[F'(\theta)]^{t-s+j+1}$$

$$+ \sum_{j=0}^{\infty} [F(\theta)]^{j+1} \Omega_1'(\theta, \mu)[F'(\theta)]^{t-s+j}$$

$$= V_{st}(\theta, \mu), \quad t \geq s. \quad (81)$$

By using (60), (61) and the series expansion of $F(\theta) = e^{A(\theta)}$, the matrix $V_{st}(\theta, \mu)$ can be expressed as a power series in $A(\theta)$ and $A'(\theta)$ with each term involving $\Sigma(\mu)$. Then the maximum likelihood estimator $(\hat{\theta}, \hat{\mu})$ is obtained by maximising:

$$L(\theta, \mu) = \log|V(\theta, \mu)| + \left[ (y_1 - m(\theta))', \ldots, (y_T - m(\theta))' \right] [V(\theta, \mu)]^{-1}$$

$$\times \left[ \begin{array}{c} y_1 - m(\theta) \\ \vdots \\ y_T - m(\theta) \end{array} \right], \quad (82)$$

where $V(\theta, \mu)$ is the $nT \times nT$ matrix whose $(st)$th $n \times n$ block is $V_{st}(\theta, \mu)$.

Because of the computational difficulty of maximising $L(\theta, \mu)$ it is useful to consider estimates based on approximate models, even if these are not consistent. A simple approximate model is obtained by replacing $C$ by $\alpha I$, which is the limit, as $\delta \to 0$, of the right-hand side of (73). In place of (75) we then have:

$$y_t - F(\theta)y_{t-1} - g(\theta) = \epsilon_t + \alpha I \epsilon_{t-1}. \quad (83)$$

Then we can define the vectors $y_1^{(1)}, \ldots, y_T^{(1)}$ by the transformation:

$$\begin{bmatrix} y_1^{(1)} \\ y_2^{(1)} \\ \vdots \\ y_T^{(1)} \end{bmatrix} = \begin{bmatrix} I & 0 & \cdots & 0 \\ -\alpha I & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ (-\alpha)^{T-1} I & (-\alpha)^{T-2} I & \cdots & I \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} \quad (84)$$

and, assuming that $\epsilon(0) = 0$, we obtain:

$$y_t^{(1)} = F(\theta)y_{t-1}^{(1)} + g(\theta) + \epsilon_t. \quad (85)$$
which can be treated like (31). A procedure which is approximately equivalent to this was used by Bergstrom and Wymer (1976) who applied the transformation:

$$y_t^{(1)} = y_t - \alpha y_{t-1} + \alpha^2 y_{t-2} - \alpha^3 y_{t-3}. \quad (86)$$

We could also use the transformation (84) or (86) and then the approximate simultaneous equations model (38), which would be even simpler.

A method which can be expected to yield better estimates was studied by Phillips (1974b, 1978). His method is to obtain a preliminary estimate of $A$, ignoring the a priori restrictions on this matrix, from data transformed by (84) and then apply a second transformation:

$$\begin{bmatrix} y_1^{(2)} \\ y_2^{(2)} \\ \vdots \\ y_T^{(2)} \end{bmatrix} = \begin{bmatrix} I & 0 & \cdots & 0 \\ -C & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ (-C)^{T-1} & (-C)^{T-2} & \cdots & I \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \quad (87)$$

where $C$ is computed from the first two terms on the right-hand side of (73), using the preliminary estimates of $A$. He shows that the proportional asymptotic bias in the estimates obtained in this way tends to zero as the observation period tends to zero.

Having dealt with the cases in which the data are all point observations or all integrals we can easily deal with the case in which the data are mixed, with some variables being observed at points of time and others as integrals. Suppose, for example, that the first $m$ variables are stock variables and the remaining $n - m$ variables are flow variables, so that the observations are: $y_1(t), \ldots, y_m(t), \int_{t-1}^t y_{m+1}(r) \, dr, \ldots, \int_{t-1}^t y_n(r) \, dr, \ t = 1, \ldots, T.$ Then we can solve the first $m$ equations of the system (62) to express $\int_{t-1}^t y_1(r) \, dr, \ldots, \int_{t-1}^t y_m(r) \, dr$ in terms of $y_1(t) - y_1(t-1), \ldots, y_m(t) - y_m(t-1), \int_{t-1}^t y_{m+1}(r) \, dr, \ldots, \int_{t-1}^t y_n(r) \, dr, \int_{t-1}^t \xi_1(\, dr), \ldots, \int_{t-1}^t \xi_n(\, dr).$ Then substituting into (59), we obtain a system in which all variables are in the form in which they are observed and the disturbance vector is a vector moving average process whose autocorrelation properties can easily be obtained as in the proof of Theorem 8. The exact likelihood function can then be obtained in the same way as (82).

The feasibility of the methods discussed in this section has been demonstrated by Bergstrom and Wymer (1976) who applied them in the construction of a continuous time model of the United Kingdom. This model is a closed first order system of 13 non-linear differential equations with 35 parameters including three trend parameters to represent technical progress, the growth of the labour supply and growth in the demand for exports. For the purpose of estimation the model
was represented by a system similar to (28), with the addition of trend terms, by taking a linear approximation about the sample means and adding white noise disturbances. The resulting matrix \([ A(\theta), b(\theta)]\) implies quite complicated non-linear cross equation restrictions, derived from economic theory. The estimate of \(\theta\) was obtained from quarterly data, for the years 1955–1966, by applying the method of full information maximum likelihood to a system similar to (85) (including derived trend terms) with the vector \(y_i(t)\) defined by (86).

An intensive mathematical study of the steady state and asymptotic stability properties of the original model (i.e. not the linear approximation used for estimation) shows that it generates plausible long-run behaviour for the estimated values of the parameters. Moreover, post sample predictions for the period 1969–1970 are remarkably accurate in view of the fact that the model contains no exogenous variables and the predictions are for a period up to eight quarters ahead of the latest data used in making them. But it should be possible to improve on this predictive performance by using a second or higher order system of differential equations. Such a system could represent more accurately the dynamics of the partial adjustment relations and allow a more satisfactory treatment of expectations.

3. Higher order systems

We shall consider the system:

\[
d[D^{k-1}y(t)] = A_1(\theta)D^{k-1}y(t) + \cdots + A_{k-1}(\theta)Dy(t) + A_k(\theta)y(t) + b(\theta) + \xi(dt), \tag{88}
\]

where \(\{y(t)\}\) is a vector random process, \(A_1(\theta), \ldots, A_k(\theta)\) are \(n \times n\) matrices whose elements are functions of the parameter vector \(\theta\), \(b(\theta)\) is an \(n \times 1\) vector whose elements are functions of \(\theta\) and \(\xi(dt)\) is a vector of white noise disturbances, i.e. a vector satisfying Assumption 1. The system (88) will be interpreted as meaning that \(y(t)\) satisfies:

\[
D^{k-1}y(t) - D^{k-1}y(0) = \int_0^t \left[ A_1(\theta)D^{k-1}y(r) + \cdots + A_{k-1}(\theta)Dy(r) + A_k(\theta)y(r) + b(\theta) \right] dr + \int_0^t \xi(dr), \tag{89}
\]

for all \(t\).

\(^2\)For a more general and comprehensive treatment of higher order systems, see Bergstrom (1982).
The assumption that the disturbances are white noise is more easily justified in a higher order system than in a first order system. An econometric model comprising a system of stochastic differential equations is usually obtained by, heuristically, adding disturbance functions to a non-stochastic system of differential equations which may be derived from certain optimisation assumptions and would hold exactly under certain ideal conditions. These conditions might include, for example, the conditions that each agent’s objective function remains constant and contains no variables other than those in the model, that his assumptions about the random processes generating these variables are constant and that the physical constraints on his behaviour are constant. The disturbances are added to take account of the fact that none of these things is really constant. Although it is difficult to justify the assumption that they are white noise disturbances, it is not unrealistic to assume that they are random processes generated by an unknown system of stochastic differential equations with white noise disturbances. Indeed the physical processes generating many of the non-economic variables that affect economic behaviour will be approximately of this form. We can, in this case, transform our original model into a higher order system of stochastic differential equations with white noise disturbances.

Suppose, for example, that the original model is a proper first order stochastic differential equation system:

\[ D_y(t) = A(\theta)y(t) + \xi(t), \]  

with a mean square continuous disturbance vector \( \xi(t) \) generated by this system:

\[ d\xi(t) = Q\xi(t) + \xi(dt), \]  

where \( Q \) is a \( n \times n \) matrix of unknown constants and \( \xi(dt) \) satisfies Assumption 1. The system (91) can be interpreted as meaning that

\[ \xi(t) - \xi(0) - Q\int_0^t \xi(r) \, dr = \int_0^t \xi(dr) \]  

holds for all \( t \). From (90) and (92) we obtain:

\[ D_y(t) - D_y(0) - Q\int_0^t D_y(r) \, dr = A(\theta)y(t) - A(\theta)y(0) - QA(\theta)\int_0^t y(r) \, dr + \int_0^t \xi(dr), \]
and hence
\[
\text{DY}(t) - \text{DY}(0) = \left[ Q + A(\theta) \right] \int_0^t \text{DY}(r) \, dr - QA(\theta) \int_0^t y(r) \, dr + \int_0^t \xi(dr).
\]  
\tag{93}

The system (93) is of the same form as (89) with \( k = 2 \). It can be written as:
\[
d[\text{DY}(t)] = \left[ Q + A(\theta) \right] \text{DY}(t) - QA(\theta) y(t) + \xi(dt),
\]  
\tag{94}

which is of the same form as (88) and is interpreted as meaning that (93) holds for all \( t \). Obviously \( Q \) and \( A \) will not be identifiable if \( A \) is unrestricted since, in this case, interchanging \( Q \) and \( A \) will not affect (93) and (94). But, in practice, \( A \) will be severely restricted by the requirement that its elements be known functions of \( \theta \).

Systems of stochastic differential equations of order \( k > 1 \) are discussed by Wymer (1972) who, following Sargan (1974, 1976) (the main results of which were available in a preliminary mimeographed paper in 1970), considers the properties of an approximate simultaneous equations model when the observation period tends to zero. Here, for simplicity, we shall start by considering the second order system, which is likely to be of considerable practical importance. We shall consider estimates based on both the exact discrete model and the approximate simultaneous equations model and then indicate, briefly, how the results can be extended to systems of order greater than two.

The second-order system to be considered is:
\[
d[\text{DY}(t)] = A_1(\theta) \text{DY}(t) + A_2(\theta) y(t) + b(\theta) + \xi(dt),
\]  
\tag{95}

which is interpreted as meaning that \( y(t) \) satisfies:
\[
\text{DY}(t) - \text{DY}(0) = \int_0^t \left[ A_1(\theta) \text{DY}(r) + A_2(\theta) y(r) + b(\theta) \right] \, dr + \int_0^t \xi(dr),
\]  
\tag{96}

for all \( t \). We know, from Theorem 3, that the first-order system (29) has a solution which is unique (with probability 1) in the class of mean square continuous vector processes. It is natural, therefore, to seek a solution to (96) which is unique in the class of vector processes whose first derivatives \( \{ \text{DY}(t) \} \) are mean square continuous. It is easy to see (from the definition of differentiation, integration and mean square continuity given in Section 2.1) that if \( \{ \text{DY}(t) \} \) is mean square continuous, then
\[
y(t) - y(0) = \int_0^t \text{DY}(r) \, dr.
\]  
\tag{97}
Combining (96) and (97) we have:

\[
\begin{bmatrix}
Dy(t) \\
y(t)
\end{bmatrix} - \begin{bmatrix}
Dy(0) \\
y(0)
\end{bmatrix} = \int_0^t \left( \begin{bmatrix}
A_1 & A_2 \\
I & 0
\end{bmatrix} \begin{bmatrix}
Dy(r) \\
y(r)
\end{bmatrix} + \begin{bmatrix}
b
0
\end{bmatrix} \right) dr \\
+ \int_0^t \xi(\sigma) d\sigma,
\]

which is a first-order system of the form (29) in the \(2n \times 1\) vector:

\[
\begin{bmatrix}
Dy(t) \\
y(t)
\end{bmatrix}.
\]

By Theorem 3 the system (98) has a solution which, for any given pair of \(n \times 1\) vectors \(y(0)\) and \(Dy(0)\), is unique in the class of random processes \(\{y(t)\}\) such that \(\{Dy(t)\}\) is mean square continuous (since if the process \(\{Dy(t)\}\) exists the process \(\{y(t)\}\) is, obviously, mean square continuous). And this solution satisfies the stochastic difference equation system:

\[
\begin{bmatrix}
Dy(t) \\
y(t)
\end{bmatrix} = \begin{bmatrix}
[e^A]_{11} & [e^A]_{12} \\
[e^A]_{21} & [e^A]_{22}
\end{bmatrix} \begin{bmatrix}
Dy(t-1) \\
y(t-1)
\end{bmatrix} + \begin{bmatrix}
g_1 \\
g_2
\end{bmatrix} \\
+ \int_{t-1}^{t} \begin{bmatrix}
[e^{(t-r)A}]_{11} & [e^{(t-r)A}]_{12} \\
[e^{(t-r)A}]_{21} & [e^{(t-r)A}]_{22}
\end{bmatrix} \xi(\sigma) d\sigma,
\]

where

\[
\begin{bmatrix}
[e^A]_{11} & [e^A]_{12} \\
[e^A]_{21} & [e^A]_{22}
\end{bmatrix} = \begin{bmatrix}
I & 0 \\
0 & I
\end{bmatrix} + \begin{bmatrix}
A_1 & A_2 \\
I & 0
\end{bmatrix} + \frac{1}{2} \begin{bmatrix}
A_1 & A_2 \\
I & 0
\end{bmatrix}^2 + \frac{1}{3!} \begin{bmatrix}
A_1 & A_2 \\
I & 0
\end{bmatrix}^3 + \ldots,
\]

and

\[
\begin{bmatrix}
g_1 \\
g_2
\end{bmatrix} = \begin{bmatrix}
[e^A]_{11} - I & [e^A]_{12} \\
[e^A]_{21} & [e^A]_{22} - I
\end{bmatrix} \begin{bmatrix}
A_1 & A_2 \\
I & 0
\end{bmatrix}^{-1} \begin{bmatrix}
b \\
0
\end{bmatrix}.
\]

The exact discrete model (99) cannot be used as a basis for estimation since \(Dy(t)\) is not observable, even at discrete intervals. This is the reason why the
second order system (or any higher order system) cannot be treated as a trivial
extension of the first order system as in the theory of ordinary linear differential
equations. An exact discrete model which can be used for estimation purposes can
be obtained by eliminating \(Dy(t)\) from the system (99) to obtain a second order
difference equation system in \(y(t)\). The precise form and properties of this system
are given in the following theorem.

**Theorem 9**

If \(\xi(dt)\) satisfies Assumption 1, then for any given pair of \(n \times 1\) vectors \(y(0)\) and
\(Dy(0)\) the system (96) has a solution which is unique in the class of random
vector processes \(\{y(t)\}\) such that the process \(\{Dy(t)\}\) is mean square continuous,
i.e. if \(\hat{y}(t)\) is any other such solution then (21) holds for any interval \([0, T]\). This
solution satisfies the stochastic difference equation:

\[
y(t) = F_1(\theta) y(t - 1) + F_2(\theta) y(t - 2) + g(\theta) + \eta_i,
\]

\[
E(\eta_i) = 0, \quad E(\eta_i, \eta'_i) = \Omega_0, \quad E(\eta_i, \eta'_{i-1}) = \Omega_1,
\]

\[
E(\eta_s, \eta'_s) = 0, \quad |s - t| > 1, \quad s, t = 1, 2, \ldots,
\]

where

\[
F_1 = [e^A]_{21}[e^A]_{11}[e^A]_{21}^{-1} + [e^A]_{22},
\]

\[
F_2 = [e^A]_{21}[e^A]_{12} - [e^A]_{21}[e^A]_{11}[e^A]_{21}^{-1}[e^A]_{22},
\]

\[
g = [e^A]_{21}g_1 + \{I - [e^A]_{21}[e^A]_{11}[e^A]_{21}^{-1}\} g_2,
\]

\[
\Omega_0 = \int_0^1 \{[e^A]_{21}[e^{rA}]_{11} - [e^A]_{21}[e^A]_{11}[e^A]_{21}^{-1}[e^{rA}]_{21}\}
\]

\[
\times \Sigma\{[e^A]_{21}[e^{rA}]_{11} - [e^A]_{21}[e^A]_{11}[e^A]_{21}^{-1}[e^{rA}]_{21}\}' dr
\]

\[
+ \int_0^1 [e^{rA}]_{21}\Sigma[e^{rA}]_{21} dr,
\]

\[
\Omega_1 = \int_0^1 \{[e^A]_{21}[e^{rA}]_{11} - [e^A]_{21}[e^A]_{11}[e^A]_{21}^{-1}[e^{rA}]_{21}\}
\]

\[
\times \Sigma[e^{rA}]_{21} dr.
\]

**Proof**

The system (99) can be written:

\[
Dy(t) = [e^A]_{11} Dy(t - 1) + [e^A]_{12} y(t - 1) + g_1 + \int_{t-1}^t [e^{(t-r)A}]_{11} \xi(dr),
\]

\[
y(t) = [e^A]_{21} Dy(t - 1) + [e^A]_{22} y(t - 1) + g_2 + \int_{t-1}^t [e^{(t-r)A}]_{21} \xi(dr).
\]
From (103) we obtain:

\[
Dy(t-1) = [e^A]_{11}Dy(t-2) + [e^A]_{12}y(t-2) + g_1 \\
+ \int_{t-2}^{t-1} [e^{(t-1-r)A}]_{11}\xi(dr), \tag{105}
\]

and from (104):

\[
Dy(t-1) = [e^A]_{21}^{-1}y(t) - [e^A]_{21}^{-1}[e^A]_{22}y(t-1) - [e^A]_{21}^{-1}g_2 \\
- [e^A]_{21}^{-1}\int_{t-1}^{t} [e^{(t-r)A}]_{21}\xi(dr), \tag{106}
\]

\[
Dy(t-2) = [e^A]_{21}^{-1}y(t-1) - [e^A]_{21}^{-1}[e^A]_{22}y(t-2) - [e^A]_{21}^{-1}g_2 \\
- [e^A]_{21}^{-1}\int_{t-2}^{t-1} [e^{(t-1-r)A}]_{21}\xi(dr). \tag{107}
\]

Substituting from (106) and (107), for \(Dy(t-1)\) and \(Dy(t-2)\), respectively, into (105) and premultiplying by \([e^A]_{21}\), we obtain the system (100) with:

\[
\eta_t = \int_{t-2}^{t-1} \left\{ [e^A]_{21} [e^{(t-r)A}]_{11} - [e^A]_{21} [e^A]_{11} [e^A]_{21}^{-1} [e^{(t-r)A}]_{21} \right\} \xi(dr) \\
+ \int_{t-1}^{t} [e^{(t-r)A}]_{21}\xi(dr). \tag{108}
\]

Then, by using the generalizations of (15) and (16) for a vector random process, we obtain:

\[
E(\eta_t, \eta_t') = \int_{t-2}^{t-1} \left\{ [e^A]_{21} [e^{(t-r)A}]_{11} - [e^A]_{21} [e^A]_{11} [e^A]_{21}^{-1} [e^{(t-r)A}]_{21} \right\} \times \Sigma \left\{ [e^A]_{21} [e^{(t-r)A}]_{11} - [e^A]_{21} [e^A]_{11} [e^A]_{21}^{-1} [e^{(t-r)A}]_{21} \right\}' dr \\
+ \int_{t-1}^{t} [e^{(t-r)A}]_{21} \Sigma [e^{(t-r)A}]_{21}' dr \\
= \int_{0}^{1} \left\{ [e^A]_{21} [e^{rA}]_{11} - [e^A]_{21} [e^A]_{11} [e^A]_{21}^{-1} [e^{rA}]_{21} \right\} \times \Sigma \left\{ [e^A]_{21} [e^{rA}]_{11} - [e^A]_{21} [e^A]_{11} [e^A]_{21}^{-1} [e^{rA}]_{21} \right\}' dr \\
+ \int_{0}^{1} [e^{rA}]_{21} \Sigma [e^{rA}]_{21}' dr,
\]
\[
E(\eta, \eta_{t-1}) = \int_{t-2}^{t-1} \left\{ [e^{A}]_{21} [e^{(t-1-r)A}]_{11} - [e^{A}]_{21} [e^{A}]_{11} [e^{(t-1-r)A}]_{21} \right\} \times \Sigma [e^{(t+1-r)A}] dr \\
= \int_{0}^{1} \left\{ [e^{A}]_{21} [e^{rA}]_{11} - [e^{A}]_{21} [e^{A}]_{11} [e^{rA}]_{21} \right\} \times \Sigma [e^{rA}]_{21} dr,
\]
\[
E(\eta, \eta_{t}) = 0, \quad |s - t| > 1, \quad s, t = 1, 2, \ldots. \quad \blacksquare
\]

The expressions (101) and (102) can be written as infinite series by expressing the matrix exponential functions in the integrals in series form and integrating, with respect to \( r \), term by term. Evaluating the terms not involving \( A_1 \) and \( A_2 \) we obtain:

\[
\Omega_0 = \frac{2}{3} \Sigma + \cdots, \quad (109)
\]
\[
\Omega_1 = \frac{1}{6} \Sigma + \cdots. \quad (110)
\]

It is interesting to note that the first terms on the right-hand sides of (109) and (110) are identical with the first terms on the right hand sides of (69) and (70), respectively, which were obtained for the first order system with flow data.

It is clear from Theorem 9 that, if \( \Sigma = \Sigma(\mu) \), the exact discrete model corresponding to (96) is:

\[
y(t) - F_1(\theta) y(t-1) - F_2(\theta) y(t-2) - g(\theta) = \varepsilon_t + C(\theta, \mu) \varepsilon_{t-1}, \quad (111)
\]
\[
E(\varepsilon_t) = 0, \quad E(\varepsilon_t \varepsilon_{t}') = K(\theta, \mu), \quad E(\varepsilon_s \varepsilon_{t}') = 0, \quad s \neq t, \quad s, t = 1, 2, \ldots,
\]

where \( C(\theta, \mu) \) and \( K(\theta, \mu) \) satisfy (67) and (68) with \( \Omega_0(\theta, \mu) \) and \( \Omega_1(\theta, \mu) \) given by (101) and (102), respectively. The exact likelihood function can be obtained in the same way as it was for the model (75). And, in view of the results of Dunsmuir and Hannan (1976) and Dunsmuir (1979), we can expect the maximum likelihood estimates to be strongly consistent, asymptotically normal and asymptotically efficient under fairly general assumptions.

The model (111) can, of course, be used only if we have point observations. An exact discrete model satisfied by the integral observations \( y_t = \int_{t-1}^{t} y(r) dr, \quad t = 1, 2, \ldots, \) can be obtained by combining the arguments used in the proofs of Theorems 8 and 9. Since the derivation is straight forward, but somewhat tedious, we shall not set it out in detail. The first step is to derive the system which is related to (98) in the same way as (59) is related to (29) with \( \eta_t \) in (59) replaced by
the expression on the right hand side of (65). This system is:

\[
\int_{t-1}^{t} \left[ Dy(r) \right] y(r) \, dr = \left[ \begin{bmatrix} e^t_{11} & e^t_{12} \\ e^t_{21} & e^t_{22} \end{bmatrix} \right] \int_{t-2}^{t-1} \left[ Dy(r) \right] y(r) \, dr + \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}
\]

\[
+ \int_{t-1}^{t} \begin{bmatrix} A_1 & A_2 \\ I & 0 \end{bmatrix}^{-1} \begin{bmatrix} e^{(t-r)A}_{11} - I \\ e^{(t-r)A}_{21} \end{bmatrix} \int_{t-2}^{t-1} \begin{bmatrix} e^{(t-r)A}_{11} & e^{(t-r)A}_{12} \\ e^{(t-r)A}_{21} & e^{(t-r)A}_{22} - I \end{bmatrix} y(r) \, dr
\]

\[
\times \left[ \begin{bmatrix} \xi(dr) \\ 0 \end{bmatrix} \right] \, dr
\]

\[
+ \int_{t-2}^{t-1} \begin{bmatrix} A_1 & A_2 \\ I & 0 \end{bmatrix}^{-1} \begin{bmatrix} e^{(t-r-1)A}_{11} & e^{(t-r-1)A}_{12} \\ e^{(t-r-1)A}_{21} & e^{(t-r-1)A}_{22} \end{bmatrix}
\]

\[
- \begin{bmatrix} A_1 & A_2 \\ I & 0 \end{bmatrix}^{-1} \begin{bmatrix} e^{(t-r)A}_{11} & e^{(t-r)A}_{12} \\ e^{(t-r)A}_{21} & e^{(t-r)A}_{22} \end{bmatrix}
\]

\[
\times \left[ \begin{bmatrix} \xi(dr) \\ 0 \end{bmatrix} \right] \, dr
\]

(112)

The exact discrete model, satisfied by the observations, is then obtained from (112) in the same way as (100) was obtained from (99).

Clearly, the disturbances in the exact discrete model, satisfied by the observations, will involve integrals with respect to \( \xi(dr) \) over the intervals \((t-3, t-2), (t-2, t-1)\) and \((t-1, t)\), so that, in place of (100), we have:

\[
y_t = F_1(\theta) y_{t-1} + F_2(\theta) y_{t-2} + g(\theta) + \eta_t,
\]

\[
E(\eta_t) = 0, \quad E(\eta_t y_t') = \Omega_0, \quad E(\eta_t y_{t-1}') = \Omega_1, \quad E(\eta_t y_{t-2}') = \Omega_2,
\]

\[
E(\eta_s y_t') = 0, \quad |s-t| > 2, \quad s, t = 1, 2, 
\]

where \(F_1(\theta), F_2(\theta)\) and \(g(\theta)\) are defined in the same way as for (100) and \(\Omega_0, \Omega_1\) and \(\Omega_2\) are derived, from the rather complicated expression which we obtain for \(\eta_t\) in the same way as \(\Omega_0\) and \(\Omega_1\) were derived for (100). In place of (111) we have an exact discrete model of the form:

\[
y_t - F_1(\theta) y_{t-1} - F_2(\theta) y_{t-2} - g(\theta) = \epsilon_t + C_1(\theta, \mu) \epsilon_{t-1} + C_2(\theta, \mu) \epsilon_{t-2},
\]

(114)

\[
E(\epsilon_t) = 0, \quad E(\epsilon_t \epsilon_t') = K(\theta, \mu), \quad E(\epsilon_s \epsilon_{t}') = 0, \quad s \neq t, \quad s, t = 1, 2, 
\]

Again the exact likelihood function can be derived in the same way as it was for (75).
An approximate discrete model can be obtained by first approximating the system (98) by the system:

\[
\begin{bmatrix}
\Delta D y(t) \\
\Delta y(t)
\end{bmatrix} = \begin{bmatrix} A_1 & A_2 \\
I & 0
\end{bmatrix} \begin{bmatrix} M D y(t) \\
M y(t)
\end{bmatrix} + \begin{bmatrix} b \\
0
\end{bmatrix} + \begin{bmatrix} u_{1t} \\
0
\end{bmatrix},
\] (115)

where

\[\Delta y(t) = y(t) - y(t-1), \quad M y(t) = \frac{1}{2} [y(t) + y(t-1)].\]

This is of the same form as (38), but cannot be used as a basis for estimation since \(D y(t)\) is not observable. Eliminating \(D y(t)\) we obtain the approximate simultaneous equations model:

\[\Delta^2 y(t) = A_1 M \Delta y(t) + A_2 M^2 y(t) + b + u_t.\] (116)

Wymer (1972) shows that the disturbance vector \(u_t\) is approximately a moving average process with coefficient matrix \(\alpha I\) where \(\alpha\) is a root of \(z^2 - 4z + 1 = 0\). We could, therefore, obtain a simultaneous equations model with an approximately serially uncorrelated disturbance by applying the transformation (84).

All of the above results can be extended to systems of order greater than two. We start by considering the system:

\[
\begin{bmatrix}
D^{k-1}y(t) \\
D^{k-2}y(t) \\
\vdots \\
y(t)
\end{bmatrix} = \begin{bmatrix} A_1 & A_2 & \cdots & A_{k-1} & A_k \\
I & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & I & 0 & 0
\end{bmatrix} \begin{bmatrix} D^{k-1}y(r) \\
D^{k-2}y(r) \\
\vdots \\
y(r)
\end{bmatrix} + \begin{bmatrix} b \\
0 \\
\vdots \\
0
\end{bmatrix} + \int_0^t \xi (dr),
\] (117)

which is obtained from (89). From (117) we obtain, for point observations, the exact discrete model:

\[y(t) - F_1 y(t-1) - \cdots - F_k y(t-k) - g = \epsilon_t + C_1 \epsilon_{t-1} + \cdots + C_{k-1} \epsilon_{t-k+1},\] (118)
which is a generalization of (111), and, for integral observations, the exact discrete model:

\[ y_t - F_1 y_{t-1} - \cdots - F_k y_{t-k} - g = C_1 \varepsilon_{t-1} + \cdots + C_k \varepsilon_{t-k}, \quad (119) \]

which is a generalization of (114). It is understood that the elements of the matrices \( F_1, \ldots, F_k \) in (118) and (119) are functions of the parameter vector \( \theta \) while the elements of \( C_1, \ldots, C_{k-1} \) in (118) and \( C_1, \ldots, C_k \) in (119) are functions of the extended parameter vector \( (\theta, \mu) \). The corresponding matrices \( F_i, i = 1, \ldots, k \), in (118) and (119) are identical, whereas the corresponding matrices \( C_i, i = 1, \ldots, k - 1 \), in (118) and (119) are different.

We can also obtain the approximate simultaneous equations model:

\[ A_k y(t) = A_1 M A_k^{-1} y(t) + A_2 M^2 A_k^{-2} y(t) + \cdots + A_k M^k y(t) + b + u, \quad (120) \]

in which \( u \) is approximately a vector moving average process of order \( k - 1 \) [see Wymer (1972)], and a similar model for integral observations, with a disturbance vector which is approximately a vector moving average process of order \( k \).

4. The treatment of exogenous variables and more general models

We shall now extend the model (28) by including a vector \( x(t) = [x_1(t), \ldots, x_m(t)]' \) of exogenous variables. The \( x_i(t), i = 1, \ldots, m \), can be either integrable non-random functions or integrable random processes satisfying the condition that \( \mathbb{E}[x_i(t)x_i(\tau)] = 0 \) for all \( t \) and all intervals \( \tau \) on the real line. In place of (28) we have:

\[ dy(t) = [A(\theta)y(t) + B(\theta)x(t)] dt + \xi(dt), \quad (121) \]

which is interpreted as meaning that:

\[ y(t) - y(0) = \int_0^t [A(\theta)y(r) + B(\theta)x(r)] dr + \int_0^t \xi(dr) \quad (122) \]

holds for all \( t \). The elements of the \( n \times m \) matrix \( B(\theta) \) are functions of the basic parameter vector \( \theta \). There is no need to include a vector of constants since this can be allowed for by letting \( x_m(t) = 1 \).

We are interested in the problem of estimating \( \theta \) from a sample of discrete observations of the \( n + m \) variables \( y_1(t), \ldots, y_n(t), x_1(t), \ldots, x_m(t) \) when, in general, the observations of some of these variables are point observations while the remainder are integral observations. The simplest case, in principle, is where the
vector \( x(t) \) is itself generated by a stochastic differential equation system:

\[
dx(t) = Rx(t) \, dt + \xi_x(dT), \tag{123}
\]

where \( \xi_x(dT) \) is a white noise disturbance vector. We can then treat the system:

\[
\begin{bmatrix}
y(t)
\end{bmatrix} - \begin{bmatrix}
y(0)
\end{bmatrix} = \int_0^T \begin{bmatrix}
A(\theta) & B(\theta)
\end{bmatrix} \begin{bmatrix}
y(r)
\end{bmatrix} \, dr + \int_0^T \xi(r) \, dr \tag{124}
\]

as a case of (29) by replacing \( \theta \) by an extended parameter vector comprising \( \theta \) and the \( m^2 \) elements of \( R \). The exact likelihood function for the parameters of this extended system can be obtained as in Section 2.3, the simplest case being where the observations of all the variables are point observations and the most complicated where the observations of some variables are point observations while the remainder are integral observations. The assumption that \( x(t) \) is generated by a system such as (123) will often be a good approximation even if \( x(t) \) is not, in fact, generated by such a system. An even better approximation might be obtained by assuming that \( x(t) \) is generated by a higher order system, which can then be combined with (121) and treated by the methods of Section 3. But, clearly, this would involve heavy computational costs.

We turn now to less costly approximate methods. An obvious approximation is the simultaneous equations model:

\[
y(t) - y(t-1) = \frac{1}{2} A(\theta) [y(t) + y(t-1)] + \frac{1}{2} B(\theta) [x(t) + x(t-1)] + u_t, \tag{125}
\]

\[
E(u_t) = 0, \quad E(u_t u_s^*) = \Sigma, \quad E(u_s u_t^*) = 0, \quad s \neq t,
\]

\[
E(x_s u_t^*) = 0, \quad s, t = 1, 2, \ldots,
\]

which is a natural extension of (38). This model is approximate in the sense that, if \( u_t \) is defined in such a way that (125) holds exactly, then the conditions \( E(u_s u_t^*) = 0, s \neq t, E(x_s u_t^*) = 0, s, t = 1, 2, \ldots \), will be only approximately satisfied. The use of the model (125) is particularly convenient when the only restrictions on \( A \) and \( B \) are that certain elements of these matrices are zero (or some other specified numbers) so that \( \theta \) is a vector of the unknown elements of \( A \) and \( B \). For this case, Sargan (1976) has made a thorough study of the behaviour of the two-stage least squares, three-stage least squares and full information maximum likelihood estimators as the observation period \( \delta \) tends to zero [for which purpose the model can be reformulated like (47)]. He introduces three alternative assump-
tions about the exogenous variables. Assumption 7 is the simpler of his two alternative assumptions for the case of non-stochastic exogenous variables while Assumption 8 is his assumption for the case of stochastic exogenous variables.

**Assumption 7**

(i) $\frac{d^2x}{dt^2}$ exists and is bounded and continuous for all $t$, except a countable set of points $S$. There exists a time period $p$ such that the number of points of $S$ lying in the time interval $(s, s + p)$ is less than or equal to $d$ for any $s$.

(ii) $\frac{dx}{dt}$ exists and is bounded for all $t$ except points of $S$.

(iii) $x(t)$ is bounded for all $t$. The size of the discontinuity of $x(t)$ (which can occur only at points of $S$) is bounded for all points of $S$.

**Assumption 8**

(i) $x(t)$ is generated by a strictly stationary ergodic process with $E[x(t)x'(t + r)] = \Omega_x(r)$, all $t$.

(ii) $\Omega_x(r)$ has one-sided derivatives at the origin up to the fourth order, so that a one-sided Taylor series expansion of $\Omega_x(r)$ at the origin, up to the fourth power of $r$, exists.

(iii) $\frac{d^2}{dx^2} \Omega_{xy}(r) = E[(x(t)y'(t + r))]$ has positive one-sided first and second derivatives at the origin, so that a one-sided Taylor series expansion of $\Omega_{xy}(r)$ at the origin, up to the second power of $r$, exists.

Sargan's results imply that, under Assumptions 1, 2, 3 and either 7 or 8, the asymptotic bias of the two-stage least squares, three-stage least squares and full information maximum likelihood estimates are $O(\delta^2)$ as $\delta \rightarrow 0$. Moreover, under these assumptions, the difference between the limits in probability of the three-stage least squares and full information maximum likelihood estimates are $O(\delta^5)$ as $\delta \rightarrow 0$.

Assumption 8 will be satisfied if $x(t)$ is generated by the system (123). But it is easy to prove directly, by an extension of the argument used in the proof of Theorem 5, that, in this case, the asymptotic bias of the two-stage least squares estimator is $O(\delta^2)$ as $\delta \rightarrow 0$. We can show that if $A$ and $\epsilon_r$, as used in Theorem 5, are redefined for the extended model (124), then the elements of the first $n$ rows of the matrix:

$$
\lim_{T \rightarrow \infty} \sum_{t=1}^{T} \left( \frac{1}{\delta} I - \frac{1}{2} A \right) \epsilon_{it} x_i'
$$

are $O(\delta^2)$ as $\delta \rightarrow 0$, where $x_i = x(it\delta)$. It is obvious from this that Theorem 5 holds for the $n$ equations of the model (125) when $z_i^{(t)}$, $i = 1, \ldots, n$, are the vectors of instruments that yield two-stage least squares estimates for these equations. Moreover it would not be difficult to extend this argument to three-stage least squares estimators.
The discussion of the closed model, in Section 2, suggests that we should next consider estimators based on the exact discrete model:

\[ y(t) = e^{A(t)}y(t-1) + \int_{t-1}^{t} e^{(t-r)A(t)}x(t-r)dr + \int_{t-1}^{t} e^{(t-r)A(t)}\xi(dr), \quad (126) \]

which can be obtained by a fairly obvious revision of Theorems 2 and 3 to include exogenous variables. Since we do not have a continuous record of \(x(t)\), the model (126) cannot be used directly as a basis for estimation. But it was used by Phillips (1974a, 1976) in order to obtain a more complicated approximate discrete model than that studied by Sargan. This is obtained by replacing \(x(t-r)\) in the first integral on the right hand side of (126) by:

\[ \hat{x}(t-r) = x(t) + \frac{r}{2}[3x(t) - 4x(t-1) + x(t-2)] + \frac{r^2}{2}[x(t) - 2x(t-1) + x(t-2)], \]

which is the quadratic function of \(r\) chosen so that \(\hat{x}(t) = x(t)\), \(\hat{x}(t-1) = x(t-1)\) and \(\hat{x}(t-2) = x(t-2)\). Evaluating the integral we obtain the approximate discrete model:

\[ y(t) = F(\theta)y(t-1) + G_1(\theta)x(t) + G_2(\theta)x(t-1) + G_3(\theta)x(t-2) + v_t, \quad (127) \]

\[ \mathbb{E}(v_t) = 0, \quad \mathbb{E}(v_t v_s') = \Omega, \quad \mathbb{E}(v_t v_s') = 0, \quad s \neq t, \]

\[ \mathbb{E}(x_t v_s') = 0, \quad s,t = 1,2,\ldots, \]

where

\[ F = e^A, \]

\[ G_1 = \left[ (A^{-2} + A^{-3})e^A - A^{-1} - \frac{3}{2}A^{-2} - A^{-3} \right]B, \]

\[ G_2 = \left[ (A^{-1} - 2A^{-3})e^A + 2A^{-2} + 2A^{-3} \right]B, \]

\[ G_3 = \left[ (-\frac{1}{2}A^{-2} + A^{-3})e^A - \frac{1}{2}A^{-2} - A^{-3} \right]B. \]

In the above derivation we have identified the time unit with the observation period as we would in practical applications. But, for the purpose of considering the behaviour of the estimators as the observation period tends to zero, Phillips follows Sargan's procedure of introducing a parameter \(\delta\) to represent the observation period. If \(\delta \neq 1\), we must replace \(A\) and \(B\) by \(\delta A\) and \(\delta B\), respectively, in the expressions for \(F, G_1, G_2\) and \(G_3\).
Phillips considers, first, the estimator of $\theta$ obtained by applying full information maximum likelihood to the model (127) as if this were the true model, i.e. the vector $\theta$ that minimises:

$$\det\left[ \frac{1}{T} \sum_{t=1}^{T} \left\{ y(t) - F(\theta) y(t-1) - G_1(\theta) x(t) \\
- G_2(\theta) x(t-1) - G_3(\theta) x(t-2) \right\} \right].$$

He shows, under certain assumptions, that this estimator has a limiting normal distribution as $T \to \infty$ and that the asymptotic bias is $O(\delta^2)$ as $\delta \to 0$. But the assumptions made for this purpose are stronger than Assumptions 7 or 8 and require the exogenous variables to follow a smoother time path, whether random or non-random, than the latter assumptions. In particular, they rule out the case of exogenous variables generated by the first order stochastic differential equation system (123), with white noise disturbances.

Phillips then considers the properties of an instrumental variables estimator in which the vector $[y'(t-2), x'(t), x'(t-2), x'(t-3)]$ is used as a vector of instruments. He shows that the asymptotic bias of this estimator is $O(\delta^3)$ under much weaker assumptions which do not exclude exogenous variables generated by the system (123). This method can be expected to give better estimates, therefore, than the use of the approximate simultaneous equations model (125).

We turn, finally, to a powerful method due to Robinson (1976a). This makes use of a discrete Fourier transformation of the data and is applicable to a very general model which includes, as special cases, systems of stochastic differential equations and mixed systems of stochastic differential and difference equations. Moreover, it does not assume that the disturbances are white noise. They are assumed to be strictly stationary ergodic processes with unknown correlation of functions. But the method is not applicable to a closed model.

The model considered by Robinson can be written:

$$y(t) = B \int_{-\infty}^{\infty} \Gamma(r, \theta) x(t-r) \, dr + \xi(t),$$

(128)

where $y(t)$ is an $n \times 1$ vector of endogenous variables, $x(t)$ is an $m \times 1$ vector of exogenous variables, $\xi(t)$ is a disturbance vector, $B$ is an $n \times l$ matrix of parameters which are subject to specified linear restrictions (e.g. certain elements of $B$ could be specified as zero), $\Gamma(r, \theta)$ is an $l \times m$ matrix of generalized functions and $\theta$ is a $1 \times p$ vector of parameters. An even more general model in
which \( \Gamma \) is not required to belong to a finite dimensional space was considered by Sims (1971). (His investigation is confined to the single equation case, but the results could be extended to a system of equations.) In this case \( \Gamma \) is not identifiable from discrete data. Moreover, the results obtained by Sims show that it can be very misleading to approximate \( \Gamma(r, \theta) \) by smoothing the lag distribution in the equivalent discrete model.

At this stage we shall give a few results, relating to the spectral representation of a stationary process, which are essential for an understanding of Robinson's method. It should be remarked, however, that Robinson (1976) is concerned with the use of Fourier methods in the estimation of the parameters of a particular model formulated in the time domain and not with the more general spectral analysis of time series. The latter is discussed in Chapter 17 of this Handbook.

A wide sense stationary random vector process \( \{x(t)\} \) has [Rozanov (1967, Theorem 4.2] the Cramér representation:

\[
x(t) = \int_{-\infty}^{\infty} e^{i\lambda t} \Phi_x(d\lambda),
\]

where \( \Phi_x \) is a complex valued random measure of the general type discussed in Section 2.2 and integration with respect to a random measure is defined as in Section 2.2. The random measure \( \Phi_x \) is called the random spectral measure of the process \( \{x(t)\} \), and if \( F_{xx}(d\lambda) \) is defined by:

\[
F_{xx}(d\lambda) = E[\Phi_x(d\lambda)\Phi^*_x(d\lambda)],
\]

where \( \Phi^*_x(d\lambda) \) denotes the complex conjugate of the transpose of \( \Phi_x(d\lambda) \), we call \( F_{xx}(d\lambda) \) the spectral measure of \( \{x(t)\} \). If \( F_{xx}(d\lambda) \) is a matrix of absolutely continuous set functions with the derivative matrix:

\[
f_{xx}(\lambda) = \lim_{d\lambda \to 0} \frac{1}{d\lambda} F_{xx}(d\lambda),
\]

then \( f_{xx}(\lambda) \) is called the spectral density of \( \{x(t)\} \). The random spectral measure \( \Phi_x(d\lambda) \) can be obtained from \( \{x(t)\} \) by [Rozanov (1967, p. 27)] the inverse Fourier transformation:

\[
\Phi_x(\Delta) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \left\{ \frac{e^{i\lambda_2 t} - e^{-i\lambda_1 t}}{-it} \right\} x(t) dt,
\]

which holds for any interval \( \Delta = (\lambda_1, \lambda_2) \) such that:

\[
\Phi_x(\lambda_1) = \Phi_x(\lambda_2) = 0.
\]
Returning now to Robinson’s model, let $\Gamma(\lambda, \theta)$ be defined by:

$$\Gamma(\lambda, \theta) = \int_{-\infty}^{\infty} e^{i\lambda t} \Gamma(t, \theta) \, dt.$$ 

Then it can be shown [Rozanov (1967, p. 38)] that the random spectral measure of $\int_{-\infty}^{\infty} \Gamma(r, \theta) x(t - r) \, dr$ is $\overline{\Gamma}(-\lambda, \theta) \Phi_{x}(d\lambda)$. By replacing each of the terms in (128) by its Cramér representation and applying the inverse Fourier transformation (130), we obtain, therefore:

$$B \overline{\Gamma}(-\lambda, \theta) \lim_{N \to \infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \frac{e^{-i\lambda t} - e^{-i\lambda_{1} t}}{-i t} \right\} y(t) \, dt = B \overline{\Gamma}(-\lambda, \theta) \lim_{N \to \infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \frac{e^{-i\lambda_{2} t} - e^{-i\lambda_{1} t}}{-i t} \right\} x(t) \, dt + \Phi_{x}(\Delta). \quad (131)$$

The equation system (131) holds exactly. Moreover, if $\Delta_{1}, \ldots, \Delta_{n}$ are disjoint intervals the disturbance terms $\Phi_{x}(\Delta_{1}), \ldots, \Phi_{x}(\Delta_{n})$ are uncorrelated, although they are not homoscedastic when the intervals $\Delta_{1}, \ldots, \Delta_{n}$ are of equal length. But we cannot estimate (131) directly since we cannot observe the integrals.

In order to derive an approximate model we first note that:

$$\frac{e^{-i\lambda_{2} t} - e^{-i\lambda_{1} t}}{-i t} = e^{-i\lambda_{1} t}(\lambda_{2} - \lambda_{1}) + O(\lambda_{2} - \lambda_{1})^{2}. \quad (132)$$

If we now divide the interval $(-\pi, \pi)$ into $N$ subintervals $\Delta_{1}, \ldots, \Delta_{n}$, each of length $2\pi/N$, and use (132), we obtain from (131) the approximate system:

$$\lim_{N \to \infty} \frac{1}{N} \int_{-\pi}^{\pi} \left\{ \frac{e^{-i\lambda_{2} t} - e^{-i\lambda_{1} t}}{-i t} \right\} y(t) \, dt = B \overline{\Gamma}(-\lambda, \theta) \lim_{N \to \infty} \frac{1}{N} \int_{-\pi}^{\pi} \left\{ \frac{e^{-i\lambda_{2} t} - e^{-i\lambda_{1} t}}{-i t} \right\} x(t) \, dt + \Phi_{x}(\Delta_{s}), \quad (133)$$

If we normalize (133) by dividing by $(2\pi/N)^{1/2}$ we obtain a system with a disturbance vector $(N/2\pi)^{1/2} \Phi_{x}(\Delta_{s})$ whose covariance matrix is approximately the spectral density of $\xi(t)$ at the frequency $\lambda_{s}$. If we then conjugate and replace the integrals by discrete Fourier transforms of the observations we obtain:

$$w_{y}(s) = B \overline{\Gamma}(\lambda_{s}, \theta) w_{x}(s) + w_{x}(s), \quad (134)$$

where

$$w_{x}(s) = (2\pi N)^{-1/2} \sum_{n=1}^{N} x(n) e^{in\lambda_{s}}.$$
The model (134) is the approximation used by Robinson for estimation purposes. Estimation is carried out in two stages. We first minimise the sums of squares of the errors which are then used to compute estimates of the spectral density of $\xi(t)$ in the frequency bands corresponding to various values of $s$. These estimates are then used in the construction of a Hermitian form in the errors in (134) which is minimised with respect to $B$ and $\theta$, subject to the restrictions, in order to obtain estimates of the parameters. In another article Robinson (1976b) considers the application of this general method specifically to a system of stochastic differential equations. The differential equations model is also treated by an instrumental variables method in Robinson (1976c). These two articles contain interesting Monte Carlo studies of the results of the application of the two methods.

Robinson (1976a) shows, under certain assumptions, that the estimation procedure described above, for the model (128) using the approximate discrete model (134), yields estimates which are strongly consistent, asymptotically normal and asymptotically efficient. The most restrictive of his assumptions is that the spectral density of $x(t)$ is zero outside the frequency range $(-\pi, \pi)$. This assumption is necessary when estimating the parameters of such a general model from equispaced discrete observations because of aliasing, to which we have already referred in Section 2.3. The assumption would not be satisfied if, for example, $x(t)$ were generated by the stochastic differential equation system (123), with white noise disturbances. But in this case, we can always extend the system, as we have seen, so that it can be treated as a closed model. And, if necessary, we can transform the model into a higher order system so that the assumption that the disturbances are white noise is approximately satisfied.

5. Conclusion

In this chapter we have described statistical methods which are applicable to a class of continuous time stochastic models and discussed the theoretical foundations of these methods. An important feature of the class of models considered is that such models allow for the incorporation of a priori restrictions, such as those derived from economic theory, through the structural parameters of the continuous time system. They can be used, therefore, to represent a dynamic system of causal relations in which each variable is adjusting continuously in response to the stimulus provided by other variables and the adjustment relations involve the basic structural parameters in some optimization theory. These structural parameters can be estimated from a sample comprising a sequence of discrete observations of the variables which will, generally, be a mixture of stock variables (observable at points of time) and flow variables (observable as integrals). In this way it is possible to take advantage of the a priori restrictions derived from
economic theory (which are very important in econometric work, because of the smallness of the samples) without making the unrealistic assumption that the economy moves in discrete jumps between successive positions of temporary equilibrium.

The feasibility of constructing a continuous adjustment model of an economy, using the methods described in this chapter, was demonstrated by Bergstrom and Wymer (1976) to whose work we have referred in Section 2.3. The methods are now being widely used, and the Bergstrom–Wymer model has been used as a prototype for a larger econometric model of the United Kingdom [see Knight and Wymer (1978)] as well as for models of various other countries [see, for example, Jonson, Moses and Wymer (1977)]. There have also been some applications of the models, not only for forecasting, but also for the investigation of the effects of various types of policy feed-back [see, for example, Bergstrom (1978, 1984)]. And, in addition to these macroeconomic applications there have been applications to commodity and financial markets [see, for example, Richard (1978) and Wymer (1973)]. The results of these various studies, which are concerned with models formulated, mainly, as first order systems of stochastic differential equations, are very encouraging. They suggest that further empirical work with higher order systems of differential equations or more general continuous time models is a promising field of econometric research.

On the theoretical side an important and relatively unexplored field of research is in the development of methods of estimation for systems of non-linear stochastic differential equations. So far these have been treated by replacing the original model by an approximate system of linear stochastic differential equations, which is treated as if it were the true model for the purpose of deriving the "exact discrete model" or, alternatively, making a direct approximation to the non-linear system of differential equations with a non-linear simultaneous equations model. In some cases it may be possible to derive the exact likelihood function in terms of the discrete observations generated by a system of non-linear stochastic differential equations. But, more generally, we shall have to rely on approximate methods, possibly involving the use of numerical solutions to the non-linear differential equations system.

References


