I. Review of time series models and their forecasts

1.1 Introduction:

A time series is a sequence of observations ordered over time. It is represented by a set of observations $\{z(t), t\varepsilon T\}$ where T is an index set.

The use of observations of a time series, which are available up to time t to forecast its value at future time points provides an important basis for many economic, business and technical decisions.

Forecasting means extrapolating historical data to predict the value of some variable at a future time. Various methods of extrapolating historical data have been developed, thus resulting in a number of different forecasting techniques. After giving a short discussion of time series models, various methods of forecasting are investigated.

1.2. Probabilistic Foundations of time series analysis:

Stochastic process

Let $(\Omega, \mathcal{S}, P)$ be a probability space. An indexed family of random variables $\{Z(t); t\varepsilon T\}$ defined on the probability space is called a stochastic process. If the index set T consists of equal spaced integers, the process is called a discrete time process. If the index set is the set

of real numbers $\{t; -\infty < t < \infty\}$, it is called a continuous time process. To put it into different words: A stochastic process can be thought as a statistical phenomenon that evolves in time according to probability laws.

It is useful to think of a random experiment modelled by a stochastic process in the following way: Underlying "nature" is thought to perform a trial of the experiment governed by the probability structure P and obtains a value $\omega\varepsilon\Omega$. One observes the function $Z(t,\omega)$ $t\varepsilon T, \omega\varepsilon\Omega$, where $Z(t,\omega)$ is the value of the random variable $Z(t)$ at ω. The values of $Z(t)$ which are generated by varying ω over Ω are called the sample function of the stochastic process. The time series is such a realization of the stochastic process.

The stochastic process determines the set of finite dimensional probability distributions $F(z_1, z_2, \ldots, z_n; t_1, t_2, \ldots, t_n) = P\{Z(t_1) \leq z_1; Z(t_2) \leq z_2, \ldots, Z(t_n) \leq z_n\}$ for any arbitrary set of t-values $(t_1, t_2, \ldots, t_n)$. Kolmogorov (1933) shows that the reverse is also true. Given a family of finite dimensional distributions, which satisfy the symmetry condition i.e. $F(z_{j_1}, z_{j_2}, \ldots, z_{j_n}, t_{j_1}, \ldots, t_{j_n}) = F(z_1, \ldots, z_n, t_1, \ldots, t_n)$ where $(j_1, j_2, \ldots, j_n)$ is any permutation of $(1,2,\ldots,n)$ and the compatibility condition i.e. $F(z_1, \ldots, z_m, \infty \ldots, \infty; t_1, t_2, \ldots, t_m, \ldots, t_n) = F(z_1, \ldots, z_m, t_1, \ldots, t_m)$, such a process exists.

Kolmogorov's theorem implies that the finite probability distributions completely determine the probability structure of the stochastic process.

Stationary stochastic process

A special case of the class of stochastic processes consists of stationary stochastic processes. A stochastic process is said to be strictly stationary if the joint probability distributions are invariant under translations in time, i.e.:

$$F(z_1,z_2,\ldots,z_n;t_1,t_2,\ldots,t_n) = F(z_1,\ldots,z_n;t_1+k,t_2+k,\ldots,t_n+k) \qquad (1.2.1)$$

for any set of times $(t_1,t_2,\ldots,t_n)$ and any k.

The condition of strict stationarity can be weakened by requiring that the multivariate moments $E\left(z_{t_1}^{\ell_1}\ldots z_{t_n}^{\ell_n}\right)$ up to order $L = \ell_1+\ell_2+\ldots+\ell_n$ depend only on the time differences. A process with this property is called weakly stationary of order L.

For the case $L = 2$ the condition of weak stationarity asserts that the mean is constant and the autocovariance function is a function of the time difference only.

i.e.: $E(z(t)) = \int z dF(z;t) = \mu$ for all t

$$E(z(s)z(t)) = \int z_1 z_2 dF(z_1,z_2;s,t) = \gamma_{|s-t|} \quad \text{for all } t,s \qquad (1.2.2)$$

If the probability distribution associated with any set of times is a multivariate Normal distribution, the process is called a Gaussian process. Since a Normal distribution is fully characterized by the moments of first and second order, in this case second order stationarity is sufficient to produce strict stationarity.

1.3: Linear prediction theory of stationary stochastic processes:

In the early part of the 20th century time series studies were based on models consisting of a nonstochastic and almost periodic trend term with added independent shocks (Schuster (1906)). Early criticism, however, prompted the search for models which could better describe the data and this led to stochastic models. These are based on Yule's idea (1927), that a time series in which successive values are highly dependent can be regarded as a linear aggregate of independent schocks a_t. Schocks or innovations are independent random drawings from some fixed distribution, with mean 0 and variance σ_a^2. In this thesis we call a sequence of in-

dependent random drawings from such a distribution a white noise sequence.

Wold's orthogonal decomposition

In his study of discrete models, Wold (1938) proved the fact that any weakly stationary process $z(t)$, $t = 0, \pm1, \pm2, \ldots$ can be uniquely represented as the sum of two processes:

$$z(t) = v(t) + u(t) \qquad (1.3.1)$$

where: i.) the process $u(t)$ is uncorrelated with the process $v(t)$

ii.) $u(t)$ has a one-sided moving average representation

i.e.: $u(t) = \sum_{k=0}^{\infty} \psi_k a_{t-k}$ with $\psi_o = 1$ and $\sum_{k=0}^{\infty} \psi_k^2 < \infty$

iii.) the process $v(t)$ is deterministic.

Such a decomposition of a stationary process in terms of uncorrelated or independent random variables is very important (orthogonal decomposition). This key decomposition of the weakly stationary process led to the formulation and solution of the linear prediction problem by Kolmogorov (1939, 1941a, 1941b). Kolmogorov also provides a geometric interpretation of weakly stationary processes. This geometric setting provides a natural framework for predicting future values of the

process from values obtained in the past.

Prediction

Suppose realizations of the process for times $s \leq n$ are available and the value for $z(n+\ell)$ for some $\ell \geq 1$ has to be predicted from this information. Since one never knows the value of the future observations, it is a reasonable procedure to select a linear function g of the past observations $z(s)$ $s \leq n$ which is good "on the average"; that is to select the function g such that the squared distance $E(z(n+\ell) - g(z(n), z(n-1), \ldots))^2$ is minimized.

Consider the random variables of the process $Z(t)$ $t = 0, \pm1, \pm2, \ldots$ as elements of the Hilbert space of real valued random variables X for which $EX = 0$; $EX^2 < \infty$.

A Hilbert space $\mathcal{H}$ is a complete inner product space. Inner product space means that there is an inner product $\langle x,y \rangle$ defined, which assigns scalar values to pairs of vector $x,y \in \mathcal{H}$ according to the following axioms: i.) $\langle x,y \rangle = \overline{\langle y,x \rangle}$

ii.) $\langle \alpha x+\beta y, z \rangle = \alpha \langle x,z \rangle + \beta \langle y,z \rangle$ where $x,y,z \in \mathcal{H}$; α,β scalars

iii.) $\langle x,x \rangle \geq 0$ with equality if and only if $x = 0$

Completeness means that every Cauchy sequence converges.

For the present case we define the inner product by $\langle x,y\rangle = Exy$ (the Hilbert space with this norm is called $L_2(P)$; see Koopmans (1974)). The linear subspace $\mathcal{M}^z$ generated by the process is the collection of all finite linear combinations of elements of the process and all limits of Cauchy sequences of these finite linear combinations.

$\mathcal{M}_n^z$ is the linear subspace generated by the linear past of the process up to time n; the prediction problem consists in finding an element $\hat{z}_n(\ell) \in \mathcal{M}_n^z$ such that $\langle z(n+\ell), \hat{z}_n(\ell)\rangle$ is minimized. The solution to this problem follows from an important property of the Hilbert space:

If $\mathcal{M}_n^z$ is a linear subspace and $z(n+\ell)$ is an element of the Hilbert space not in $\mathcal{M}_n^z$, then there exits a unique element $\hat{z}_n(\ell) \in \mathcal{M}_n^z$, for which

i.) $z_{n+\ell} - \hat{z}_n(\ell)$ is orthogonal to $\mathcal{M}_n^z$ and

ii.) $\langle z(n+\ell), \hat{z}_n(\ell)\rangle$ is minimized.

Spectral decomposition of a stationary process:

One extensively used decomposition of a stationary process is the spectral decomposition. This is an integral expansion with complex coefficients (compared to the sums and real coefficients in Wold's decomposition).

In the simplest form we can state this decomposition in the following way:

Consider the stationary stochastic process $z(t)$ with mean $Ez(t) = 0$ and autocovariance funtion $E(z(t)z(t-\tau)) = \gamma(\tau)$. Using Bochner's theorem (e.g. Koopmans (1974)) we can represent the nonnegative definite autocovariance function $\gamma(\tau)$ as

$$\gamma(\tau) = \int_{-\infty}^{+\infty} e^{i\tau\lambda} dF(\lambda) \text{ where} \qquad (1.3.2)$$

$F(\lambda)$ is real, nondecreasing and bounded. The function F is called the spectral distribution function defined on the frequency interval $-\infty<\lambda<\infty$ and its derivative (if it exists) is called the spectral density function $f(\lambda)$.

The spectral representation of the stochastic process is given by

$$z(t) = \int e^{i\lambda t} dS(\lambda) \qquad (1.3.3)$$

where for the case where t is discrete the limits of integration are $\pm\pi$ and for the continuous case the limits are $\pm\infty$. The stochastic process $S(\lambda)$ is such that:

$$E(S(\lambda_2) - S(\lambda_1))(S(\lambda_4) - S(\lambda_3)) = 0$$

for $\lambda_1 < \lambda_2 < \lambda_3 < \lambda_4$ (uncorrelated increments)

and

$$E|S(\lambda_2) - S(\lambda_1)|^2 = F(\lambda_2) - F(\lambda_1) \quad \lambda_2 < \lambda_1$$

$F(\lambda)$ is the spectral distribution function as defined in (1.3.2).

One recognizes that the spectral density of a stochastic process is merely a function of the autocovariance function of the process. The relation in (1.3.2) can be inverted to get

$$f(\lambda) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \gamma(k) e^{-i\lambda k} dk \quad \text{for continuous time processes}$$

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{+\infty} \gamma(k) e^{-i\lambda k} \quad \text{for discrete time processes} \qquad (1.3.4)$$

Equation (1.3.4) is called the Wiener-Khintchine relation.

In contrast to the parametric difference equations prefered by statisticians, Wiener (1949)characterizes the stationary stochastic disturbances by autocorrelation, cross-correlation and spectral density function, This approach is called non parametric in the literature. Wiener shows that the problem of specifying a linear filter for the minimum mean squared error prediction of a stochastic signal (or equivalently the separation of a stochastic signal from noise) leads to the Wiener-Hopf integral equation. For the case of stationary linear time series with rational spectra the

solution of this equation can be derived by spectral factorization.

Suppose that z_t is a weakly stationary process over discrete time with mean zero and continuous spectral density $f(\lambda)$ such that

$$\int_{-\pi}^{\pi} \log f(\lambda)\, d\lambda > -\infty.$$

Then the ℓ- step ahead linear predictor of $z_{n+\ell}$ is a linear combination of past observations $z_n, z_{n-1}, \ldots$.

$$\hat{z}_n(\ell) = \sum_{r \geq 1} b_r z_{n+1-r}$$

When the range of summation is infinite this is regarded as a suitable mean square limit. For $\ell = 1$, the mean squared error for such a predictor is given by

$$E\{[z_{n+1} - \hat{z}_n(1)]^2\} = E\{[z_{n+1} - \sum_{r \geq 1} b_r z_{n+1-r}]^2\}$$

$$= E\{[\sum_{r \geq 0} b_r z_{n+1-r}]^2\} \quad \text{defining } b_o = -1$$

$$= \sum_{t,s=0}^{\infty} b_t b_s \gamma(t-s)$$

$$= \sum_{t,s=0}^{\infty} b_t b_s \int_{-\pi}^{\pi} e^{-i\lambda(t-s)} f(\lambda) d\lambda$$

$$= \int_{-\pi}^{\pi} f(\lambda) \sum_{t,s=0}^{\infty} e^{-i\lambda(t-s)} b_t b_s d\lambda$$

$$= \int_{-\pi}^{\pi} |b(\lambda)|^2 f(\lambda) d\lambda \qquad (1.3.5)$$

where

$$b(\lambda) = 1 - \sum_{r \geq 1} b_r e^{ir\lambda}$$

Kolmogorov (1941) shows that for given $f(\lambda)$ the minimum value of (1.3.5) is given by

$$\sigma^2 = 2\pi \exp\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \log f(\lambda) d\lambda\right] \qquad (1.3.6)$$

and therefore the one step ahead predictor, which satisfies (1.3.6) is given by

$$|b(\lambda)|^2 = \frac{\sigma^2}{2\pi f(\lambda)} \qquad (1.3.7)$$

Relation (1.3.7) identifies the optimal predictor weights.

1.4. The general linear process, its parsimonious versions and their prediction.

Wold's decomposition theorem states that any weakly stationary stochastic process can be represented as a linear aggregate of random shocks. For practical problems (estimation however, it is important to employ parametric models which use parameters in a parsimonious fashion. This can be achieved by considering autoregressive, moving average and mixed autoregressive moving average processes.

Prediction theory, as discussed in the previous section, applies to weakly stationary processes. Frequently, however, economic, business and industrial series exhibit non-stationary behaviour in terms of changing mean and/or slope and/or periodicity. It will be necessary to relax the condition of stationarity to allow for particular kinds of homogeneous non-stationary behaviour.

The general linear process is a linear transformation of white noise.

$$z_t = \psi(B) a_t \qquad (1.4.1)$$

where: i.) z_t is the difference between the original observations and some deterministic component $f(t)$ which can be explained physically (e.g.deviations

from the mean)

ii.) $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$; $\psi_o = 1$ and B is the backshift operator

$B^m z_t = z_{t-m}$

iii.) $\{a_t\}$ is a white noise series with

$Ea_t = 0$ for all t

$Ea_t a_{t-k} = \begin{cases} 0 \text{ for } k \neq 0 \\ \sigma_a^2 \text{ for } k=0 \end{cases}$

Stationarity and invertibility of a general linear processes:

The linear process is said to be stationary if $\psi(B)$, the generating function of the ψ-weights, converges for $|B| \leq 1$. This condition ensures that the ψ-weights form a convergent series and that the variance of the process is finite and the matrix of autocovariances is positive definite.

An equivalent representation of the stochastic process in (1.4.1) is given by

$$\pi(B) z_t = a_t \qquad (1.4.2)$$

where

$$\pi(B) = \psi(B)^{-1} = 1 - \sum_{j=1}^{\infty} \pi_j B^j$$

The linear process is said to be invertible if $\pi(B)$, the

generating function of the π-weights, converges for $|B| \leq 1$. The invertibility condition is independent of the stationarity condition and is only necessary if one is interested in forecasting into a particular direction of time. As far as defining a stationary stochastic process is concerned, it is easily shown that the representation (1.4.1) is not unique. For example, the two representations

$$z_t = (1-\theta B)a_t \qquad (1.4.3)$$

$$z_t = (1-\theta F)a_t = (1-\theta^{-1}B)e_t \qquad (1.4.4)$$

where F is the foreward shift operator $F^m z_t = z_{t+m}$ and $e_t = -\theta a_{t+1}$, define the same stochastic process. Representation (1.4.3) is directed into the future and for $|\theta| < 1$ future values are derived as a convergent weighted sum of past observations. Representation (1.4.4) is directed into the past, and for $|\theta| < 1$ future values of the series will be a linear combination of past values, but with divergent weights.

Autoregressive moving average processes:

In general, representations (1.4.1) and (1.4.2) could contain an infinite number of parameters ψ_j and π_j and they would not be useful for estimation purposes. Parsimonious versions which are representational useful have to be found.

The consideration of a ratio of two polynomials in B, $\psi(B) = \frac{\theta_q(B)}{\phi_p(B)}$ where

$$\theta_q(B) = 1-\theta_1 B \ldots -\theta_q B^q$$

and

$$\phi_p(B) = 1-\phi_1 B \ldots -\phi_p B^p,$$

allows great flexibility with few parameters.

i.) $\phi_p(B)z_t = \theta_q(B)a_t$ (1.4.5)

is called autoregressive moving average process; ARMA (p,q)

ii.) $\phi_p(B)z_t = a_t$ (1.4.6)

is called autoregressive process of oder p; AR(p)

iii.) $z_t = \theta_q(B)a_t$ (1.4.7)

It is called moving average process of order q; MA(q)

It can be shown that a ARMA (p,q) process is stationary if the characteristic equation $\phi_p(B) = 0$ has all its roots outside the unit circle. It is invertible if the roots of $\theta_q(B) = 0$ lie outside the unit circle.

Nonstationary processes

Many time series encountered in industry and business exhibit non stationary behavior and in particular do not vary about a fixed mean. Nevertheless they exhibit homogeneity in the sense that apart from local level and/or trend and/or periodicity, one part of the series behaves very much like the other.

Stochastic processes which exhibit these characteristics have some roots of $\phi(B) = 0$ on the unit circle. This implies that the autoregressive operator will contain factors of the form $(1-B)$, $(1-B^2)$, $(1-\sqrt{3}B+B^2)$, $(1-B^s)$ etc. Box and Jenkins refer to such as simplifying operators.

Thus, a nonstationary series which exhibits homogeneous properties except in its level might be rendered stationary by considering its first difference $(1-B)z_t$.

Similarly, for a series for monthly data which shows a strong sinusoidal trend (e.g. ambient temperature data) an appropriate simplifying operator might be $(1-\sqrt{3}B+B^2)$.

For general 12-monthly seasonal pattern a complete set of sinusoids would be generated by the operator $(1-B^{12})$. This operator can be factored into operators corresponding to different components of sines and cosines. The factorization is shown in Table 1.1.

factor	root	period	frequency in cycles per year
$1-B$	1	constant	
$1-\sqrt{3}B+B^2$	$\frac{1}{2}(\sqrt{3}\pm i)$	12	1
$1-B+B^2$	$\frac{1}{2}(1\pm i)$	6	2
$1+B^2$	$\pm i$	4	3
$1+B+B^2$	$\frac{1}{2}(-1\pm i)$	3	4
$1+\sqrt{3}B+B^2$	$\frac{1}{2}(-\sqrt{3}\pm i)$	12/5	5
$1+B$	-1	2	6

Table 1.1 Factorization of $(1-B^{12})$.

Particular examples of the simplifying operators are the ordinary differences. The introduction of the operator $(1-B)^d$ allows for nonstationarity of the original series and up to the (d-1)st difference, while all higher differences are stationary but not necessarily invertible. For d = 1, the model allows for nonstationarity in terms of having no fixed level. For d = 2 it allows for nonstationarity in both level and slope. A theoretical account of such processes was first given by Yaglom (1955). An earlier procedure for time series analysis which used differencing of the data is the variate difference method developed by Tintner (1940), which however is different in its motivation,

Box and Jenkins call processes of the form

$$\phi_{p+d}(B)z_t = (1-B)^d\phi_p(B)z_t = \theta_q(B)a_t \qquad (1.4.8)$$

autoregressive integrated moving average processes; ARIMA (p,d,q). The roots of $\phi_p(B) = 0$ and $\theta_q(B) = 0$ are assumed to lie outside the unit circle, thus assuring stationarity and invertibility of the d.th difference.

Equivalent representations of ARIMA (p,d,q) processes:

The ARIMA (p,d,q) process can be represented in three equivalent forms:

i.) in terms of the difference equation of the model

$$z_t - \phi_1 a_{t-1} - \ldots - \phi_{p+d} z_{t-p-d} = a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \qquad (1.4.9)$$

ii.) in terms of current and previous shocks a_t

$$z_t = a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j} \qquad (1.4.10)$$

or in truncated form

$$z_t = C_k(t-k) + I_k(t-k) \qquad (1.4.11)$$

where the complementary function $C_k(t-k)$ is the general solution of $\phi_{p+d}(B)\ C_k(t-k) = 0$ and the particular integral $I_k(t-k)$ is any function which satisfies $\phi_{p+d}(B)I_k(t-k) = \theta(B)a_t$

iii.) in terms of a weighted sum of previous values z_{t-j} of the process and the current shock a_t

$$z_t = \sum_{j=1}^{\infty} \psi_j z_{t-j} + a_t \qquad (1.4.12)$$

Minimum mean squared error forecasts for ARIMA (p,d,q) models:

Minimum mean squared error forecasts for this class of models are easily derived using the general results by Wold, (1938), Kolmogorov (1941), Wiener (1949) and Whittle (1963). The minimum mean squared error forecast of $z_{n+\ell}$, denoted by $\hat{z}_n(\ell)$, is given by

$$\hat{z}_n(\ell) = \sum_{j=0}^{\infty} \psi_{n+j} a_{n-j} = E_n[z_{n+\ell}]$$

where

$$E_n[z_{n+\ell}]$$

is the conditional expectation of $z_{n+\ell}$ given knowledge of all the z's up to time n.

Forecasts can be represented in three basic forms corresponding to the three representations of the ARIMA (p,d,q) model.

i.) Forecasts from the difference equation:

Taking the conditional expectation of $z_{n+\ell}$ in (1.4,9) results in

$$\hat{z}_n(\ell) = [z_{n+\ell}] = \phi_1[z_{n+\ell-1}] + \ldots + \phi_{p+d}[z_{n+\ell-p-d}]$$

$$+ [a_{n+\ell}] - \theta_1[a_{n+\ell-1}] - \ldots - \theta_q[a_{n+\ell-q}] \qquad (1.4.13)$$

where:

$$[z_{n-j}] = E_n[z_{n-j}] \quad \text{for } j < 0 \qquad [a_{n-j}] = 0 \quad \text{for } j < 0$$

$$= z_{n-j} \quad \text{for } j \geq 0 \qquad = a_{n-j} \quad \text{for } j \geq 0$$

Forecasts can be easily updated using the relation

$$\hat{z}_{n+1}(\ell) = \hat{z}_n(\ell+1) + \psi_\ell a_{n+1} \qquad (1.4.14)$$

which for $\ell=0$ reduces to

$$z_{n+1} = \hat{z}_n(1) + a_{n+1} \quad \text{or} \quad a_{n+1} = z_{n+1} - \hat{z}_n(1)$$

The recursive relationship in (1.4.13) provides an easy method for obtaining forecasts and is conveniently implemented in practice.

ii.) Another way to represent minimum mean squared error (MMSE) forecasts is through the eventual forecast function. This representation gives more insight as far as the nature of the forecasts is concerned. The eventual forecast function is the solution of the difference equation

$$\phi_{p+d}(B)\hat{z}_n(\ell) = 0 \quad \text{for } \ell > q$$

which is given by:

$$\hat{z}_n(\ell) = b_1^*(n)f_1^*(\ell)+\ldots+b_{p+d}^*(n)f_{p+d}^*(\ell) \quad \ell > q-p-d \quad (1.4.15)$$

$\hat{z}_n(\ell)$ has the same form as the complementary function in (1.4.11). $f_1^*(\ell),\ldots,f_{p+d}^*(\ell)$ are functions of the lead time ℓ and depend only on the autoregressive part of the model. In general, these functions can be polynomials, exponentials, sines, cosines and products of these functions. For a given forecast origin n the coefficients $b_1^*(n),\ldots,b_{p+d}^*(n)$ are constants and are the same for all lead times ℓ. However, they change from one forecast origin to the other and can be updated conveniently (Box and Jenkins (1970)).

iii.) Forecasts as weighted average of previous observations:

$$\hat{z}_n(\ell) = [z_{n+\ell}] = \sum_{j=1}^{\infty} \pi_j[z_{n+\ell-j}]+[a_{n+\ell}] = \sum_{j=1}^{\infty} \pi_j^{(\ell)} z_{n+1-j} \quad (1.4.16)$$

where

$$\pi_j^{(\ell)} = \pi_{j+\ell-1} + \sum_{k=1}^{\ell-1} \pi_k \pi_j^{(\ell-k)}$$

An important property of minimum mean squared error forecasts is that the one step ahead forecast errors are uncorrelated. Any sensible forecast procedure should have this property since if one step ahead forecast errors were correlated one could improve the forecasts by using this correlation among the forecast errors.

The minimum mean squared error forecast in (1.4.13) is a point estimate; making assumptions about the distribution of the shocks a_t specifies the complete predictive distribution of future values.

1.5. The philosophy of iterative model building:

In the previous sections we discussed the minimum mean squared error forecasts for a given stochastic process. We assumed knowledge of the form and of the values of the parameters of the model. If the structure of the underlying process is known, the predictor is easily determined. In practice, however, the structure of the process is rarely, if ever, known, and one has to use past observations to derive an adequate model and estimate its parameters.

Box and Jenkins (1970) develop a three stage iterative

procedure Identification-Fitting-Diagnostic Checking to find adequate model(s) from historic data. This iterative procedure is shown diagramatically in Figure 1.1.

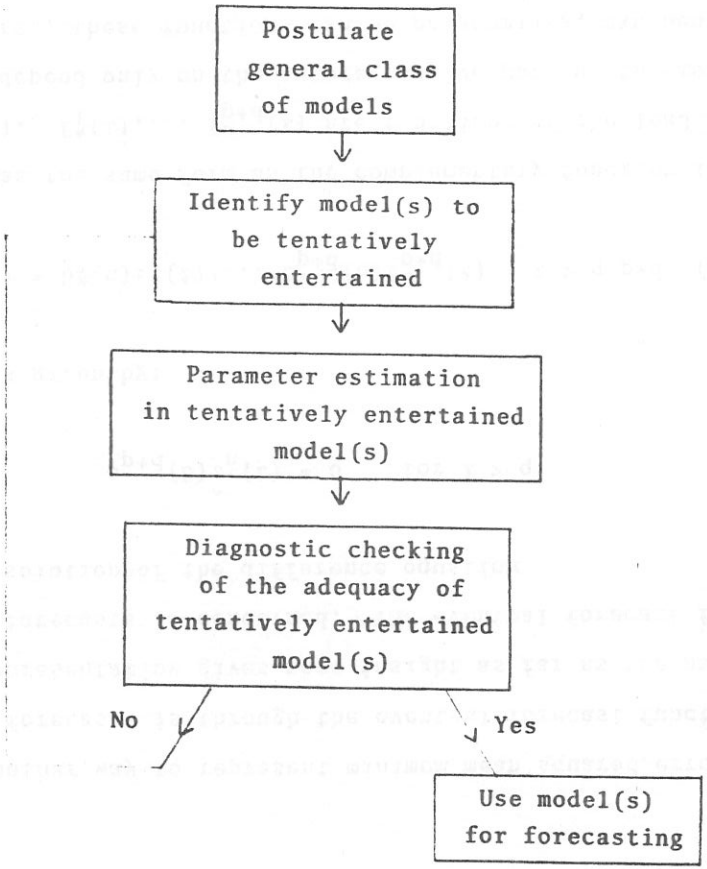

Figure 1.1.: Three stage iterative approach to model building

Model Identification;

At the first stage of the iterative approach to model building, the identification (specification) stage, one uses the data and the knowledge of the system to suggest an appropriate parsimonious subclass of models which may be tentatively entertained. At this stage the basic tools are plots of the data and the sample autocorrelation and sample partial autocorrelation function of the observations. The properties of the theoretical autocorrelations and partial autocorrelatio of ARIMA models provide a reference table. It is important to point out that at this stage statistically inefficient methods must necessarily be used, since no precise formulation of the problem is yet available.

Model Estimation:

At this stage efficient use of the data is made by making inferences about the parameters conditionally on the adequacy of the entertained model(s). Assuming Normally distributed shocks, Box and Jenkins (1970), Sredni (1970) derive the exact likelihood function for ARIMA (p.d.q) models. Maximum likelihood estimates can be obtained and as it was shown by Whittle (1953), the limiting properties of these estimates, which are usually established for the case of independent observations, also cover the case of stationary stochastic processes.

A Bayesian version of the estimation of parameters is also derived in the literature (Box and Jenkins (1970)). Combining Jeffrey's non informative prior distribution with the exact likelihood function results in the posterior distribution of the parameters.

For practical applications nonlinear least squares procedures can be employed to derive estimates of the parameters and their covariance matrix.

Diagnostic Checking:

After fitting the tentatively entertained model to the observed data one has to check the fitted model in its relation to the observations with intent to reveal model inadequacies and to achieve improvement. The residuals from the fitted model contain all the information about the adequacy of the fitted model. Inspection of the sample autocorrelation function and of the cumulative periodigram of the residuals indicate wether the entertained model can be considered adequate for the data set under study, or if and how the model should be revised.

Model identification, estimation and diagnostic checking are important steps in the iterative model building procedure. After diagnostic checks satisfy the critic as to the adequacy of the model he originally sponsored, the derived model is used for forecasting purposes. Instead of

having population parameters available to derive forecasts one has to use estimated values of the parameters. Box and Jenkins (1970) investigated the effect of parameter estimation errors on the probability limits of the forecasts. Wichern (1969), Zellner (1971) used a Bayesian analysis and derived the predictive distribution of future observations treating the parameters in the ARIMA models as random variables.

Another approach to forecasting involves direct estimation of the spectral density. This approach is called nonparametric although strictly speaking it does involve parameters. The spectral density is a transformation of the autocorrelations which together with the mean are a sufficient statistic for any stationary Gaussian process. They are nonparametric in the sense that they don't use a specified class of models such as the family of ARIMA models. In a way they are rather nonparsimonious than nonparametric, since in the case of an ARIMA process the autocorrelations (spectral density) would contain the information, but not in a parsimonious way.

Such "nonparametric" estimates of spectral densities are widely discussed in the literature. For example in the books by Jenkins and Watts (1968), Hannan (1970), Koopmans (1974) a good introduction is given.

Substituting an estimate of the spectral density into

(1.3.7) one can derive the optimal predictor weights. Bloomfield (1972) shows that errors in the estimation of the spectral density function can have a larger effect on the variance of the forecast error than the parameter estimation errors of the parametric models as discussed above.

1.6. Exponential smoothing techniques and related forecasting procedures.

Exponential smoothing techniques have received broad attention in the existing literature, especially in the area of management science. These procedures are fully automatic which means that once a computer program has been written, forecasts for any given time series can be derived without manual intervention. The fact that they are fully automatic has been put forward as an advantage of the scheme. It can equally well be argued that this is a great disadvantage, since it discourages the use of the human mind in circumstances where this instrument could be used with profit.

The basic exponential smoothing equation replaces an observed series z by a smoothed series $\bar{z}$, an exponentially weighted average of current and past values of z.

$$\bar{z}_n = \alpha z_n + \alpha(1-\alpha) z_{n-1} + \alpha(1-\alpha)^2 z_{n-2} + \ldots = \alpha \sum_{j \geq 0} (1-\alpha)^j z_{n-j} \qquad (1.6.1)$$

The latest available smoothed value is used to forecast all future observations

$$\hat{z}_n(\ell) = \bar{z}_n \qquad (1.6.2)$$

A convenient updating algorithm for the forecasts is given by

$$\hat{z}_n(1) = \alpha z_n + (1-\alpha)\hat{z}_{n-1}(1) \qquad (1.6.3)$$

This basic exponential smoothing procedure which can be attributed to Holt (1957), Winters (1960), and Brown (1962) was, and still is, used frequently to derive forecasts of economic and business data. It was Muth (1960), who first asked the question: "What is the underlying process for which the exponential smoothing technique in (1.6.2) provides minimum mean squared error forecasts?"
He showed that the underlying process is given by

$$z_t = \alpha \sum_{j=0}^{\infty} a_{t-1-j} + a_t \qquad (1.6.4)$$

or in equivalent form as an ARIMA (0,1,1) process

$$(1-B) z_t = (1-(1-\alpha)B) a_t \qquad (1.6.4)$$

In order to take local trends and seasonality into account various modifications of the simple procedure outlined above have been considered in the literature. Holt and Winters assume additive trend and multiplicative seasonality. Denoting the trend factor at time t by T_t, the seasonal factor by S_t and the smoothed value by $\bar{z}_t$, their model can be written as

$$\begin{cases} T_t = \alpha_1(\bar{z}_t - \bar{z}_{t-1}) + (1-\alpha_1)T_{t-1} \\ S_t = \alpha_2 \dfrac{z_t}{\bar{z}_t} + (1-\alpha_2)S_{t-L} \text{ if the seasonal cycle has period } L \\ \bar{z}_t = \alpha_3 \dfrac{z_t}{S_{t-L}} + (1-\alpha_3)(\bar{z}_{t-1} + T_{t-1}) \end{cases} \tag{1.6.5}$$

The forecasts at time origin n are given by

$$\hat{z}_n(\ell) = (\bar{z}_n + \ell T_n) S_{n-L+j} \quad \text{where } \ell = kL+j \quad \text{for } k = 0,1,\ldots \quad j = 1,2,\ldots,L \tag{1.6.6}$$

The smoothing constants $\alpha_i (0<\alpha_i<1 \; i=1,2,3)$ are assumed to be known a priori.

The model by Harrison (1965) is a modification of the model by Holt and Winters obtained by smoothing the seasonal factors by Fourier methods. Instead of updating every

seasonal factor only once a season, as done in the Holt Winters method, this procedure updates all factors as each new observation becomes available. Other generalizations of exponential smoothing procedures have been considered by Brown and Meyer (1961), Brown (1962). They select forecast functions (fitting functions) $f_1(\ell),\ldots,f_m(\ell)$ such that the vector of values of these functions at lead time $\ell + 1$ is a linear combination of the values of the same functions at the previous lead time ℓ.

$$\underset{\sim}{f}(\ell+1) = L\underset{\sim}{f}(\ell) \qquad L = \begin{bmatrix} L_m & \cdots & L_{1m} \\ & & \\ L_{m1} & \cdots & L_{mm} \end{bmatrix} \qquad \underset{\sim}{f}(\ell) = \begin{bmatrix} f_1(\ell) \\ \vdots \\ f_m(\ell) \end{bmatrix} \tag{1.6.7}$$

where L is assumed nonsingular and $\underset{\sim}{f}(0)$ specified. Polynomials, sinusoids, exponentials and linear combinations and products of these fall into this class of fitting functions The fitting functions are fitted to the data $z_1, z_2, \ldots, z_n$ by discounted least squares. The coefficients $\underset{\sim}{b}'(n) = [b_1(n),\ldots,b_m(n)]$ are estimated by minimizing:

$$\sum_{j=0}^{n-1} \beta^j [z_{n-j} - \hat{p}(n,-j)]^2 \tag{1.6.8}$$

where

$$\hat{z}_n(j) = \hat{p}(n,j) = \sum_{i=1}^{m} b_i(n) f_i(j) = \underset{\sim}{b}'(n)\underset{\sim}{f}(j)$$

This can be thought of as a generalized least squares problem. For large n (steady state) the solution is given by

$$\hat{\underset{\sim}{b}}(n) = F^{-1}\underset{\sim}{r}(n) \qquad (1.6.9)$$

where:

$$F = \sum_{j=0}^{\infty} \beta^j \underset{\sim}{f}(-j)\underset{\sim}{f}'(-j) \text{ and } \underset{\sim}{r}(n) = \sum_{j=0}^{\infty} \beta^j z_{n-j}\underset{\sim}{f}(-j)$$

The smoothing constant $\beta(0<\beta<1)$ is assumed to be known and Brown suggests that β^m should be picked between .7 and .9, where m is the number of coefficients to be estimated. Although Brown (1974) mentions some unpublished work which supposedly developed theoretical reasons why this should be so, actual study of time series gives no empirical support to this assertion and no theoretical reasons seem to be available for discussion.

In Chapter 2 of this thesis we will show that the smoothing constant has to be decided by the underlying stochastic model and cannot be picked arbitrarily and set equal to .9.

It will be shown that the choice of fitting functions and the choice of some specific β implies an ARIMA model with a certain relation among its parameters.

Brown and Meyer (1961) prove what they call the fundamental theorem of exponential smoothing, which is summarized below:

If the behavior of the series is represented by a polynomial of degree k, the coefficients of the polynomial, estimated by minimizing (1.6.8) can equivalently be estimated by linear combinations of the smoothed values $s_t^{[j]}(z)$ $j = 1,2,\ldots,k$.

The j^{th} order smoothing is defined as:

$$s_t^{[j]}(z) = \alpha s_t^{[j-1]}(z) + (1-\alpha)s_{t-1}^{[j]}(z) \qquad (1.6.10)$$

As example consider

$$\hat{p}(n,j) = \underset{\sim}{b}'(n)\underset{\sim}{f}(j)$$

where

$$\underset{\sim}{f}(j) = \begin{bmatrix} 1 \\ j \end{bmatrix} \quad \text{and} \quad \underset{\sim}{b}(n) = \begin{bmatrix} b_1(n) \\ b_2(n) \end{bmatrix}$$

Then

$$\hat{\underset{\sim}{b}}(n) = F^{-1}\underset{\sim}{r}(n) = \begin{bmatrix} 1-\beta^2 & (1-\beta)^2 \\ (1-\beta)^2 & \frac{(1-\beta)^3}{\beta} \end{bmatrix} \begin{bmatrix} \sum_{j=0}^{\infty} \beta^j z_{n-j} \\ -\sum_{j=0}^{\infty} j\beta^j z_{n-j} \end{bmatrix}$$

$$= \begin{bmatrix} (1-\beta^2) \sum_{j=0}^{\infty} \beta^j z_{n-j} - (1-\beta)^2 \sum_{j=0}^{\infty} j\beta^j z_{n-j} \\ (1-\beta)^2 \sum_{j=0}^{\infty} \beta^j z_{n-j} - \frac{(1-\beta)^3}{\beta} \sum_{j=0}^{\infty} j\beta^j z_{n-j} \end{bmatrix}$$

$$= \begin{bmatrix} 2s_n^{[1]}(z) - s_n^{[2]}(z) \\ \frac{1-\beta}{\beta}\{s_n^{[1]}(z) - s_n^{[2]}(z)\} \end{bmatrix}$$

where

$$s_n^{[1]}(z) = (1-\beta)z_n + \beta s_{n-1}^{[1]}(z) = (1-\beta)\sum_{j=0}^{\infty} \beta^j z_{n-j}$$

$$s_n^{[2]}(z) = (1-\beta)s_n^{[1]}(z) + \beta s_{n-1}^{[2]}(z) = (1-\beta)^2 \sum_{j=0}^{\infty} (j+1)\beta^j z_{n-j} \qquad (1.6.11)$$

This procedure is called multiple (second order) exponential smoothing.

Reid (1971) did extensive empirical work on over 100 economic series comparing several forecasting techniques such as Box-Jenkins iterative model building procedures, Brown's exponential smoothing methods, Holt-Winters procedures and the forecasting technique proposed by Harrison. He shows that Box-Jenkins procedures outperformes its competitors in terms of having smaller mean squared error for one step ahead forecasts as well as for forecasts over longer lead times. A similar study with similar conclusions on different data sets is reported by Newbold and Granger (1974).

Groff (1973) came to different conclusions. However he does not seem to have understood the iterative approach to model building described by Box-Jenkins and his results and conclusions appear incorrect.

Harrison and Stevens (1971) approach the problem of forecasting from a Bayesian point of view. They consider a model, in which the observations follow

$$z_t = \mu_t + \varepsilon_t$$

$$\text{with mean} \quad \mu_t = \mu_{t-1} + \beta_t + \gamma_t \qquad (1.6.12)$$

$$\text{and trend} \quad \beta_t = \beta_{t-1} + \delta_t \,.$$

The observations can be in a number of different states ($j=1,2,\ldots,N$) according to

$$\varepsilon_t \sim N(0,V_\varepsilon^j); \quad \gamma_t \sim N(0,V_\gamma^j); \quad \delta_t \sim N(0,V_\delta^j)$$

(e.g. no change, transient change, step change, etc.).

Posterior distributions for mean and trend and their updating formulae are derived. However, one has to specify a large number of parameters before using this scheme, and hence its practical application is very doubtful.

Granger and Newbold (1974) investigate the question wether the combination of forecasts (derived by using different forecast procedures) can produce a forecast which is better than the individual forecasts. The philosophy behind this study is to combine several inexpensive, fully automatic, computerized forecasting procedures (such as Holt-Winters, exponential smoothing methods, stepwise autoregressive methods) and to do almost as well as the more time consuming and more experience demanding iterative Box-Jenkins procedure. Various methods for combining forecasts are proposed. However, as Stigler (1974) points out, the methodology used by the authors in deriving optimal weights for the different forecasts could have led to a "well-disguised, but very real instance of the selection fallacy".

REFERENCES

Bloomfield, P. (1972), "On the error of prediction of a time series," Biometrika, 59, 501-507.

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

Brown, R.G., (1962), Smoothing, Forecasting and Prediction of Discrete Time Series, Prentice-Hall; New Jersey.

Brown, R.G., (1974), "Forecasting," in: Handbook of Operations Research, forthcoming.

Brown, R.G., and Meyer, R. F. (1961), "The fundamental theorem of exponential smoothing," Operations Res., 9, 673-685.

Groff, G. K. (1973), "Empirical comparison of models for short range forecasting," Management Science, 20, 22-31.

Hannan, E. J. (1970), Multiple Time Series, John Wiley; New York.

Harrison, P. J. (1965), "Short-term sales forecasting," Applied Stat., 14, 102-139.

Harrison, P. J. and Stevens, C. F. (1971), "A Bayesian approach to short term forecasting," Operations Research Quarterly, 22, 341-362.

Holt, C. C. (1957), "Forecasting trends and seasonals by exponentially weighted moving averages," O.N.R. Memorandum, Carnegie Institute of Technology, No. 52.

Jenkins, G. M. and Watts, D. G. (1968), Spectral Analysis, Holden-Day; San Francisco.

Kolmogorov, A. N. (1933), "Grundbegriffe der Wahrscheinlichkeitsrechnung," Erg. Mat., 2, No. 3.

Kolmogorov, A. N. (1939), "Sur l'interpolation et l' extrapolation des suites stationaires," C. R. Acad. Sci., Paris, 208, 2043.

Kolmogorov, A. N. (1941), "Stationary sequences in Hilbert space," Bull. Math., Univ. Moscow, 2, No. 6.

Kolmogorov, A. N. (1941), "Interpolation und Extrapolation von stationären sufälligen Folgen," Bull. Acad. Sci. (Nauk) USSR Ser. Math., 5, 3-14.

Koopmans, L. H. (1974), The Spectral Analysis of Time Series, Academic Press; New York and London.

Muth, G. F. (1960), "Optimal properties of exponentially weighted forecasts of time series with permanent and transitory components," JASA, 55, 299-306.

Newbold, P. and Granger, C. W. J. (1974), Experience with forecasting univariate time series and the combination of forecasts," Jour. of Royal Statis, Soc., Series A, 137, 131-164.

Reid, D. J. (1971), "A comparison of forecasting techniques on economic time series," Joint conference organized by Society for Long Range Planning and Forecasting Group of the Operational Research Society, London School of Economics.

Schuster, A. (1906), "On the periodicities of sunspots," Phil. Trans. Royal Soc., A206, 69-100.

Stigler, S. (1974), "Contribution to the discussion of the paper "Experience with forecasting univariate time series and the combination of forecasts" by Newbold, P. and Granger, C. W. J. (1974), Jour. of Royal Statis. Soc., Series A, 137, 131-164.

Sredni, J. (1970), "Problems of design, estimation, and lack of fit in model building," Ph.D. thesis, University of Wisconsin-Madison.

Tintner, G. (1940), The Variate Difference Method, Principia Press; Bloomington, Indiana.

Whittle, P. (1953), "Estimation and information in stationary time series," Arkiv für Mathematik, 2, 423.

Whittle, P. (1963), Prediction and Regulation, D. Van Nostrand; Princeton, New Jersey.

Wichern, D. W. (1969), "Studies in the identification and forecasting of non stationary time series," Ph.D. thesis, University of Wisconsin-Madison.

Wiener, N. (1949), Extrapolation, Interpolation and Smoothing of Stationary Time Series, John Wiley; New York.

Winters, P. R. (1960), "Forecasting sales by exponentially weighted moving averages," Management Science, 6, 324-342.

Wold, H. O. (1938), A Study in the Analysis of Stationary Time Series, Almquist and Wicksell; Uppsala (2nd ed. 1954).

Yaglom, A. M. (1955), "The correlation theory of processes whose n'th difference constitutes a stationary process," Matem. Sb. 37, 141.

Yaglom, A. M. (1962), An Introduction to the Theory of Random Functions, English translation by A. Silverman, Prentice-Hall; Englewood Cliffs, N. J.

Yule, G. U. (1927), "On a method of investigating periodicities in disturbed series, with special reference to Wölfer's sunspot numbers," Phil. Trans. 226, 267.

Zellner, A. (1974), An Introduction to Bayesian Inference in Econometrics, John Wiley; New York.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report No. 446	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) TOPICS IN TIME SERIES ANALYSIS I. Review of time series models and their forecasts		5. TYPE OF REPORT & PERIOD COVERED Scientific Interim 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) JOHANNES LEDOLTER GEORGE E. P. BOX		8. CONTRACT OR GRANT NUMBER(s) AFOSR AF-72-2363-C
9. PERFORMING ORGANIZATION NAME AND ADDRESS Dept. of Statistics 1210 W. Dayton Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research Building 410 Bolling AFB, D. C. 20332		12. REPORT DATE Dec., 1975 13. NUMBER OF PAGES 38
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Distribution of this document is unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES None		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Stochastic process stationarity spectral representation general linear process ARIMA models linear prediction iterative model building exponential smoothing techniques		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This paper provides a summary of the literature concerned with time series analysis and forecasting. We discuss the probabilistic foundations of time series analysis, the linear prediction theory, the general linear process, its parsimonious versions and their prediction. Furthermore, the philosophy of iterative model building and exponential smoothing techniques are discussed.		

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)