smfd10.tex
Day 10. 8.6.2012

5. CONDITION FOR STATIONARITY

We return to the general case. Just as in the $AR(2)$ example above, if the $AR(p)$ process has a moving-average representation

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i},$$

then if $\sigma^2 = var \epsilon_t$,

$$var X_t = \sigma^2 \sum_{i=0}^{\infty} \phi_i^2.$$

The condition

$$\sum_{i=0}^{\infty} \phi_i^2 < \infty$$

((ϕ_i) is *square-summable*, or is *in* L_2) is necessary and sufficient for
(i) $var X_t < \infty$;
(ii) the series $\sum \phi_i \epsilon_{t-i}$ in the moving-average representation to be convergent in mean square – or, in L_2. So if we interpret convergence in the mean-square sense, $\sum \phi_i^2 < \infty$ is the necessary and sufficient condition (NASC) for the moving-average representation of X_t to *exist*. Since $\sum \phi_i \epsilon_{t-i}$ is (when convergent) stationary (because (ϵ_t) is stationary):
if $\sum \phi_i^2 < \infty$, then (X_t) is stationary. The converse is also true, giving:

THEOREM (Condition for Stationarity). The following are equivalent:
(i) The parameters $\phi_1, \cdots, \phi_p$ in the $AR(p)$ model

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t, \qquad (\epsilon_t) \quad WN(\sigma^2) \qquad (*)$$

define a stationary process (X_t);
(ii) The roots of the polynomial

$$\phi(\lambda) := \phi_p \lambda^p + \cdots + \phi_1 \lambda - 1 = 0$$

lie outside the unit disc in the complex λ-plane;
(iii) X_t has the moving-average representation

$$X_t = \sum_{i=0}^{\infty} \phi_i \epsilon_{t-i}$$

with

$$\sum_{i=0}^{\infty} \phi_i^2 < \infty.$$

Proof. Substituting the moving-average representation into $(*)$,

$$\begin{aligned} \textstyle\sum_{i=0}^{\infty}\phi_i\epsilon_{t-i} &= \textstyle\sum_{k=1}^{p}\phi_k\sum_{i=0}^{\infty}\psi_i\epsilon_{t-k-i}+\epsilon_t \\ &= \textstyle\sum_{k=1}^{p}\phi_k\sum_{i=k}^{\infty}\phi_{i-k}\epsilon_{t-i}+\epsilon_t \\ &= \textstyle\sum_{i=1}^{\infty}(\sum_{k=1}^{\min(i,p)}\phi_k\phi_{i-k})\epsilon_{t-i}+\epsilon_t. \end{aligned}$$

Equating coefficients of ϵ_{t-i}, we obtain the difference equation

$$\phi_i = \textstyle\sum_{k=1}^{p}\phi_k\phi_{i-k} \qquad (i \geq p)$$

(with similar equations for $i = 0, 1, \cdots, p-1$, which provide starting-values for the difference equation above). The difference equation, of order p, has general solution

$$\phi_i = \textstyle\sum_{k=1}^{p} c_k\lambda_k^i,$$

where $\lambda_1, \cdots, \lambda_p$ are the roots of the characteristic polynomial

$$\lambda^p - \phi_1\lambda^{p-1} - \cdots - \phi_{p-1}\lambda - \phi_p = 0$$

(with appropriate modifications in the case of repeated roots, as before). [Check: if $\phi_i = \lambda^i$ is a trial solution of the difference equation, $\lambda^i = \sum_1^p\phi_k\lambda^{i-k}$. Multiply through by λ^{p-i}: $\lambda^p = \sum_1^p\phi_k\lambda^{p-k}$.] Now as $\phi_i = \sum_1^p c_k\lambda_k^i$ and $|\lambda_k^i| \to \infty$, $= 1$ or $\to 0$ as $i \to \infty$ according as $|\lambda_k| > 1$, $= 1$ or < 1, $\sum\phi_i^2 < \infty$ iff each $|\lambda_k| < 1$, i.e. each root of $\lambda^p - \phi_1\lambda^{p-1} - \cdots - \phi_p = 0$ is inside the unit disk, i.e. each root of

$$\phi(\lambda) = \phi_p\lambda^p + \phi_{p-1}\lambda^{p-1} + \cdots + \phi_1\lambda - 1 = 0$$

is outside the unit disk. This is all that remained to be proved. //

In the stationary case, we thus have

$$\gamma_t = cov(X_t, X_{t+\tau}) = \sigma^2\textstyle\sum_{i=0}^{\infty}\phi_i\phi_{i+\tau},$$

with $\sum\phi_i^2 < \infty$ and $\phi_i = \sum_{k=1}^p c_k\lambda_k^i$, $|\lambda_i| < 1$. If λ_1 (say) is the root of largest modulus, $\phi_i \sim c_1\lambda_1^i$ for large i, and $\phi_i\phi_{i+\tau} \sim c_1^2\lambda_1^{\tau+2i}$. So for large τ, we can expect

$$\gamma_\tau \sim \sigma^2\sum c_1^2\lambda_1^{\tau+2i} \sim const.\lambda_1^\tau, \qquad \rho_\tau \sim \gamma_\tau/\gamma_0 \sim \lambda_1^\tau.$$

Thus for a stationary $AR(p)$ model, we expect that the autocorrelation decreases geometrically to zero for large lag τ (the decay rate being the characteristic root of largest modulus).

Note. For $AR(1)$, the autocorrelation function *is* geometrically decreasing: $\rho_\tau = \rho^\tau$. This holds exactly, even for small τ. Since the sample autocorrelation (correlogram) r_τ approximates the population autocorrelation $\rho_\tau = \rho^\tau$: for $AR(1)$,

$$r_\tau \sim \rho^\tau :$$

the sample ACF is *approximately* geometrically decreasing (i.e., geometrically decreasing plus sampling error), even for small lags τ. We can look for this pattern at the beginning of a plot of the ACF, and this is the signature of an $AR(1)$ process. For $AR(p)$, $p > 1$, matters are not so simple. The approximation above only holds for large τ, by which time r_τ will be small (it approximates ρ_τ, which tends to zero as τ increases), and the pattern of geometric decrease will tend to be swamped by sampling error. Consequently, it is *much harder* to interpret the correlogram of an $AR(p)$ for $p > 1$ than for an $AR(1)$.

By contrast, the moving average – $MA(q)$ – models considered below have autocorrelations that *cut off* - they are zero beyond lag q, apart from sampling error. This is the signature of the ACF of an $MA(q)$, and is easy to interpret; an $AR(1)$ signature is easy to interpret; that of an $AR(p)$ for $p > 1$ is (usually) not.

6. MOVING AVERAGE PROCESSES, MA(q).

Suppose we have a system in which new information arrives at regular intervals, and new information affects the system's response for a limited period. The new information might be economic, financial etc., and the system might involve the price of some commodity, for example.

The simplest possible model for the new information process, or *innovation process*, is white noise, $WN(\sigma^2)$, so we assume this. The simplest possible model for a response with such a limited time-influence is

$$X_t = \epsilon_t + \sum\nolimits_{j=1}^{q} \theta_j \epsilon_{t-j}, \qquad (\epsilon_t) \quad WN(\sigma^2).$$

This is called a *moving average process* or *order* q, $MA(q)$.

In terms of the lag operator B, $\epsilon_{t-j} = B^j \epsilon_t$, so if

$$\theta(B) := 1 + \sum\nolimits_{j=1}^{q} \theta_j B^j,$$

we can write

$$X_t = \theta(B)\epsilon_t.$$

Autocovariance. Since $E\epsilon_t = 0$, $EX_t = 0$ also. So writing $\theta_0 = 1$,

$$\begin{aligned}\gamma_k = cov(X_t, X_{t+k}) = E[X_t X_{t+k}] &= E[\textstyle\sum_{i=0}^q \theta_i \epsilon_{t-i} \sum_{j=0}^q \theta_j \epsilon_{t-k-j}] \\ &= \displaystyle\sum_{i,j=0}^q \theta_i \theta_j E[\epsilon_{t-i}\epsilon_{t-k-j}].\end{aligned}$$

Now $E[.] = 0$ unless $i = j + k$, when it is σ^2. It suffices to take $k \geq 0$ (as $\gamma(-k) = \gamma(k)$). If also $k \leq q$, we can take $j = i - k$, and then the limits on j are $0 \leq j \leq q - k$, as $0 \leq i \leq q$. If however $k > q$, there are no non-zero terms as there are no $i = k + j$ with $0 \leq i, j \leq q$. So

$$\gamma(k) = \begin{cases} \sigma^2 \sum_{j=0}^{q-k} \theta_j \theta_{j+k}, & \text{if } k = 0, 1, \cdots, q, \\ 0 & \text{if } k > q, \end{cases}$$

$$\gamma_0 = \sigma^2 \textstyle\sum_{j=0}^q \theta_j^2,$$

so the autocorrelation is

$$\rho_k = \begin{cases} \sum_{i=0}^{q-k} \theta_i \theta_{i+k} / \sum_{i=0}^q \theta_i^2 & \text{if } k = 0, 1, \cdots, q, \\ 0 & \text{if } k > q. \end{cases}$$

This sudden cut-off of the autocorrelation after lag $k = q$ is the signature of an $MA(q)$ process.

First-order case: $MA(1)$.

The model equation is

$$X_t = \epsilon_t + \theta \epsilon_{t-1}.$$

By above,

$$\rho_0 = 1, \qquad \rho_1 = \theta/(1 + \theta^2), \qquad \rho_k = 0 \quad (k \geq 2).$$

In terms of the lag (backward shift) operator B:

$$X_t = (1 + \theta B)\epsilon_t.$$

Hence formally

$$\epsilon_t = (1 + \theta B)^{-1} X_t = \textstyle\sum_0^\infty (-\theta)^k B^k X_t = X_t + \sum_1^\infty (-\theta)^k X_{t-k} :$$

$$X_t = \epsilon_t - \sum_1^\infty (-\theta)^k X_{t-k}.$$

This is an *infinite-order autoregressive* representation of (X_t). For (mean-square) convergence on RHS, as in the AR theory above, we need

$$|\theta| < 1.$$

The $MA(1)$ model is then said to be *invertible*: the passage from the $MA(1)$ representation using $(1+\theta B)$ to the $AR(\infty)$ representation using $(1+\theta B)^{-1}$ is called *inversion.*
Note. If we replace θ by $1/\theta$, ρ_1 goes from $\theta/(1+\theta^2)$ to

$$(1/\theta)/[1+(1/\theta)^2] = \theta/(1+\theta^2)$$

– the same as before. So for $\theta \neq 1$, two *different* $MA(1)$ processes have the *same* ACF:we cannot hope to identify the process from the ACF, or its sample version, the correlogram. However, for $|\theta| \neq 1$, exactly *one* of these processes is invertible. So if we restrict attention to invertible MA processes, *identifiability* is restored in general ($|\theta| \neq 1$), but not in the exceptional case $|\theta| = 1$, $\theta \neq 1$.
General case: $MA(q)$. As above,

$$X_t = \epsilon_t + \sum_{j=1}^q \theta_j \epsilon_{t-j} = \theta(B)\epsilon_t, \quad \text{where} \quad \theta(\lambda) = 1 + \sum_{j=1}^q \theta_j \lambda^j.$$

So formally, if we can invert this to obtain

$$\epsilon_t = \theta(B)^{-1} X_t,$$

and as $\theta(\lambda) = 1 + \theta_1\lambda + \cdots$, $1/\theta(\lambda) = 1 + c.\lambda + \cdots$. So

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_i X_{t-i} + \cdots + \epsilon_t,$$

for some constants ϕ_i. This expresses the new value X_t at the current time t as a sum of two components:
(i) an (infinite) linear combination of previous values X_{t-i}, and
(ii) the new white-noise term ϵ_t, thought of as the *innovation* at time t. It is thus plausible that it should be possible to forecast future values of such a process given knowledge of its history.

Proceeding as in the proof of the Condition for Stationarity in Section 4, we find that ϕ_i is of the form

$$\phi_i = \sum_{k=1}^q c_k \lambda_k^i,$$

where the λ_k are the roots of the polynomial

$$\lambda^p + \theta_1 \lambda^{p-1} + \cdots + \theta_p = 0.$$

For $\phi_i \to 0$ as $i \to \infty$ – that is, for the influence of the remote past of the process to damp out to zero – we need all $|\lambda_i| < 1$. That is, all roots of the above polynomial (which is $\theta(1/\lambda)$) should lie *inside* the unit disc in the complex λ-plane. Equivalently, all roots of $\theta(\lambda) = 0$ lie *outside* the unit disc. Then as before, $\sum \phi_i^2 < \infty$ and the series $\sum \phi_i X_{t-i}$ converges in mean square. To summarise, we have:

THEOREM (Condition for Invertibility). For the $MA(q)$ model

$$X_t = \theta(B)\epsilon_t, \qquad (\epsilon_t) \quad WN$$

to be invertible as

$$\epsilon_t = \theta(B)^{-1} X_t,$$

it is necessary and sufficient that all roots λ_i of the polynomial equation

$$\lambda^p + \theta_1 \lambda^{p-1} + \cdots + \theta_p = 0$$

should lie outside the unit disc. Then

$$\epsilon_t = \sum\nolimits_1^\infty \phi_i X_{t-i}$$

with $\sum \phi_i^2 < \infty$ and the series convergent in mean square.

Note. 1. The Condition for Stationarity for $AR(p)$ processes and the Condition for Invertibility for $MA(q)$ processes exhibit a *duality*, in which the roles of X_t and ϵ_t are interchanged.
2. We shall confine ourselves in what follows to the invertible case. Then the parameters θ_j are uniquely determined by the autocorrelation function ρ_τ.
3. In the $MA(1)$ case, the above characteristic equation is

$$\lambda + \theta_1 = 0,$$

with root $\lambda = -\theta_1$. For invertibility, we need $|\theta_1| < 1$, as before. Invertibility avoids the ambiguity of both θ_1 and $1/\theta_1$ giving the same ACF

$$\rho_0 = 1, \qquad \rho_1 = \theta_1/(1+\theta_1^2), \qquad \rho_k = 0 \qquad (k \geq 2).$$

7. AUTOREGRESSIVE MOVING AVERAGE PROCESSES ARMA(p,q).

We can combine the $AR(p)$ and $MA(q)$ models as follows:

$$X_t = \sum_1^p \phi_i X_{t-i} + \epsilon_t + \sum_1^q \theta_i \epsilon_{t-i}, \qquad (\epsilon_t) \quad WN(\sigma^2)$$

or

$$\phi(B)X_t = \theta(B)\epsilon_t,$$

where

$$\phi(\lambda) = 1 - \phi_1\lambda - \cdots - \phi_p\lambda^p, \qquad \theta(\lambda) = 1 + \theta_1\lambda + \cdots + \theta_q\lambda^q.$$

We shall assume that the roots of $\phi(\lambda$ and $\theta(\lambda)$ all lie *outside the unit disc.* Then, as in the Conditions for Stationarity and Invertibility, the process (X_t) is both stationary and invertible, and

$$X_t = (\phi(B))^{-1}\theta(B)\epsilon_t.$$

Now $\theta(\lambda)/\phi(\lambda)$ is a rational function (ratio of polynomials). We shall assume that $\theta(\lambda)$, $\phi(\lambda)$ *have no common factors.* For if they do:
(i) the common factors can be cancelled from $(\phi(B))^{-1}\theta(B)$, leaving an equivalent model but with fewer parameters - so better;
(ii) we have no hope of *identifying* parameters in the factors thus cancelled. Thus the model is non-identifiable. So to get an *identifiable* model, we need to perform all possible cancellations. We assume this done in what follows.
Note. Generally in statistics, we try to work with *identifiable* models. These are the ones in which the task of estimating parameters from the data is possible in principle. Non-identifiable models are degenerate, or at least problematic.

Of course: $ARMA(p,0) \equiv AR(p)$, $ARMA(0,q) \equiv MA(q)$.

ARMA(1,1).

$$X_t = \phi X_{t-1} + \epsilon_t + \theta\epsilon_{t-1} : \qquad (1 - \phi B)X_t = (1 + \theta B)\epsilon_t.$$

Condition for Stationarity: $|\phi| < 1$ (assumed).
Condition for Invertibility: $|\theta| < 1$ (assumed).

$$X_t = (1 - \phi B)^{-1}(1 + \theta B)\epsilon_t = (1 + \theta B)(\sum_0^\infty \phi^i B^i)\epsilon_t$$

$$= \epsilon_t + \sum_1^\infty \phi^i B^i \epsilon_t + \theta \sum_0^\infty \phi^i B^{i+1}\epsilon_t = \epsilon_t + (\theta + \phi)\sum_1^\infty \phi^{i-1} B^i \epsilon_t :$$

$$X_t = \epsilon_t + (\phi + \theta)\sum_{i=1}^{\infty} \phi^{i-1}\epsilon_{t-i}.$$

Variance: lag $\tau = 0$. Square and take expectations. The ϵs are uncorrelated with variance σ^2, so

$$\gamma_0 = varX_t = E[X_t^2] = \sigma^2 + (\phi + \theta)^2 \sum_1^{\infty} \phi^{2(i-1)}\sigma^2$$

$$= \sigma^2 + \frac{(\phi + \theta)^2\sigma^2}{(1 - \phi^2)} = \sigma^2(1 - \phi^2 + \phi^2 + 2\phi\theta + \theta^2)/(1 - \phi^2) :$$

$$\gamma_0 = \sigma^2(1 + 2\phi\theta + \theta^2)/(1 - \phi^2).$$

Covariance: lag $\tau \geq 1$.

$$X_{t-\tau} = \epsilon_{t-\tau} + (\phi + \theta)\sum_{j=1}^{\infty} \phi^{j-1}\epsilon_{t-\tau-j}.$$

Multiply the series for X_t and $X_{t-\tau}$ and take expectations:

$$\gamma_\tau = cov(X_t, X_{t-\tau}) = E[X_t X_{t-\tau}],$$

which is

$$E\{[\epsilon_t + (\phi + \theta)\sum_{i=1}^{\infty} \phi^{i-1}\epsilon_{t-i}].[\epsilon_{t-\tau} + (\phi + \theta)\sum_{j=1}^{\infty} \phi^{j-1}\epsilon_{t-\tau-j}]\}.$$

The ϵ_t-term in the first [.] gives no contribution. The i-term in the first [.] for $i = \tau$ and the $\epsilon_{t-\tau}$ in the second [.] give $(\phi + \theta)\phi^{\tau-1}\sigma^2$. The product of the i term in the first sum and the j term in the second contributes for $i = \tau + j$; for $j \geq 1$ it gives $(\phi + \theta)^2\phi^{\tau+j-1}.\phi^{j-1}.\sigma^2$. So

$$\gamma_\tau = (\phi + \theta)\phi^{\tau-1}\sigma^2 + (\phi + \theta)^2\phi^\tau\sigma^2 \sum_{j=1}^{\infty} \phi^{2(j-1)}.$$

The geometric series is $1/(1 - \phi^2)$ as before, so for $\tau \geq 1$

$$\gamma_\tau = \frac{(\phi + \theta)\phi^{\tau-1}\sigma^2}{(1 - \phi^2)}.[1-\phi^2+\phi(\phi+\theta)] : \qquad \gamma_\tau = \sigma^2(\phi+\theta)(1+\phi\theta)\phi^{\tau-1}/(1-\phi^2).$$

Autocorrelation. The autocorrelation $\rho_\tau := \gamma_\tau/\gamma_0$ is thus

$$\rho_0 = 1, \qquad \rho_\tau = \frac{(\phi + \theta)(1 + \phi\theta)}{(1 + 2\phi\theta + \theta^2)}.\phi^{\tau-1} \qquad (\tau \geq 1).$$

Note that

$$\rho_1 = (\phi + \theta)(1 + \phi\theta)/(1 + 2\phi\theta + \theta^2), \qquad \rho_\tau/\rho_{\tau-1} = \phi \qquad (\tau \geq 1) :$$

$\rho_0 = 1$ always, ρ_1 is as above, and *then* ρ_τ decreases geometrically with common ratio ϕ. This is the signature of an $AR(1, 1)$ process: if the correlogram looks geometric after the r_1 term, to within sampling error, then an $AR(1, 1)$ model is suggested.

8. ARMA MODELLING; THE GENERAL LINEAR PROCESS

The model equation $\phi(B)X_t = \theta(B)\epsilon_t$ for an $ARMA(p,q)$ process may sometimes have a direct interpretation in terms of the mechanism generating the model. Usually, however, $ARMA$ models are tried and fitted to the data empirically. Their principal use is that $ARMA(p,q)$ models are so flexible: a wide range of different examples may be satisfactorily fitted by an $ARMA$ model with small values of p and q, so with a small number $p+q$ of parameters. This ability to use a small number of parameters is an advantage, by the Principle of Parsimony. The drawback is that the $ARMA$ model may not correspond well with the actual data-generating mechanism, and so the $p+q$ parameters ϕ_i, θ_j may lack any direct interpretation - or indeed, any basis in reality. An alternative approach is to try to build a model whose structure reflects the actual data-generating mechanism. This leads to *structural time-series models* (Harvey [H], 5.3), *state-space models and the Kalman filter* ([H], Ch. 4), but these are too advanced for a first course on TS such as this.

Interpretation of parameters. Recall the $ARMA(p,q)$ model

$$X_t = \textstyle\sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}, \qquad (\epsilon_t) \quad WN(\sigma^2). \qquad (*)$$

Think, for example, of X_t as representing the value at time t of some *particular* economic/financial/business variable - the current price of a particular company's stock, or of some particular commodity, say. Think of ϵ_t as representing the current value of some *general* indicator of the overall state of the economy. We are trying to predict the value of the particular variable X_t, given information of two kinds:

(i) on the past values of the X-process (*particular* information),
(ii) on the past and present values of the ϵ-process (*general* information).
Then (relatively) large values of a coefficient ϕ_i, or θ_j, indicate that this variable - particular information at lag i, or general information at lag j - is important in determining the variable X_t of interest. By contrast, a (relatively) low value suggests that we may be able to discard this variable.

Another illustration, from geographical or climatic data rather than an economic/financial setting, is in modelling of river flow, or depth. Here X_t might be the depth of a particular river at time t; ϵ_t might be some general indicator of recent rainfall in the area - e.g., precipitation at some weather station in the river's watershed.

The General Linear Process. An infinite-order MA process

$$X_t - \mu = \sum\nolimits_{i=0}^{\infty} \phi_i \epsilon_{t-i}, \qquad \sum \phi_i^2 < \infty, \qquad (\epsilon_t) \quad WN$$

is called a *general linear process.* Both AR and MA processes are special cases, as we have seen. But since there are infinitely many parameters ϕ_i in the above, the model is only useful in practice if it reduces to a finite-dimensional model such as an $AR(p), MA(q)$ or $ARMA(p,q)$.

However, the general linear process is important theoretically, as we now explain. Consider a stationary process (X_t) (the general linear process is stationary), and write σ^2 for the variance of X_t (rather than ϵ_t, as before). Then σ^2 measures the *variability* in X_t. Suppose now that we are *given* the values of X_s up to X_{t-q}. This knowledge makes X_t *less variable*, so

$$\sigma_q^2 := var(X_t | \cdots, X_{t-q-2}, X_{t-q-1}, X_{t-q}) \leq \sigma^2.$$

As we increase q, the information given decreases (recedes further into the past), so X_t given this information becomes more variable: σ_q^2 increases with q. So

$$0 \leq \sigma_q^2 \uparrow \sigma_\infty^2 \leq \sigma^2 \qquad (q \to \infty).$$

One possibility is that $\sigma_q = 0$ for all q, and then $\sigma_\infty = 0$ also. Now if a random variable has *zero variance*, it is *constant* (with probability one) – i.e., non-random or deterministic. The case $\sigma_q \equiv 0$ does occur, in cases such as

$$X_t = a \cos(\omega t + b),$$

where a, b, ω may be random variables, but do not depend on time. Then three values of X_t are enough to find the three values a, b, ω, and then *all* future values of X_t are completely determined. In this case, each X_t is a random variable, but (X_t) as a stochastic process is clearly degenerate: there is no 'new randomness', and the dependence of randomness on time – the essence of a stochastic process (and even more, of a time series!) – is trivial. Such a process is called *singular* or *purely deterministic.*

9. WOLD DECOMPOSITION

At the other extreme to the deterministic case, we may have

$$\sigma_q \uparrow \sigma_\infty = \sigma \qquad (q \to \infty).$$

Then as information given recedes into the past, its influence dies away to nothing – as it should. Such a process is called *purely nondeterministic.*

We quote the

THEOREM (Wold Decomposition Theorem: Wold (1938)). A (strictly) stationary stochastic process (X_t) possesses a unique decomposition

$$X_t = Y_t + Z_t,$$

where
(i) Y_t is purely deterministic,
(ii) Z_t is purely nondeterministic,
(iii) Y_t, Z_t are uncorrelated,
(iv) Z_t is a general linear process,

$$Z_t = \sum \phi_i \epsilon_{t-i},$$

with the ϵ_t uncorrelated.

This result is due to the Swedish statistician Hermann Wold (1908-1992) in 1938. It shows that infinite moving-average representations $\sum \phi_i \epsilon_{t-i}$, far from being special, are general enough to handle the stationary case apart from degeneracies such as purely deterministic processes. For proof, see e.g. J. L. DOOB (1953): *Stochastic processes*, Wiley (XII.4, Th. 4.2).

COROLLARY. If (X_t) has no purely deterministic component – so

$$X_t = \textstyle\sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}, \qquad \sum \psi_i^2 < \infty, \qquad (\epsilon_t) \quad WN(\sigma^2) \;\; - -$$

then
(i) $\gamma_k := cov(X_t, X_{t+k}) = \sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k}$,
(ii) $\gamma_k \to 0$, $\rho_k := corr(X_t, X_{t+k}) \to 0$ $(k \to \infty)$: the autocovariance and autocorrelation tend to zero as the lag k increases.

Proof.

$$\gamma_k = cov(X_t, X_{t+k}) = E(X_t, X_{t+k}) = E[(\textstyle\sum_{i=0}^{\infty} \psi_i \epsilon_{t-i})(\sum_{j=0}^{\infty} \psi_j \epsilon_{t-k-j})]$$

$$= \textstyle\sum\sum_{i,j} \psi_i \psi_j E(\epsilon_{t-i}\epsilon_{t-k-j}).$$

Here $E(.) = 0$ unless $i = j + k$, when it is σ^2, so

$$\gamma_k = \sigma^2 \textstyle\sum_{j=0} \psi_j \psi_{j+k},$$

proving (i). For (ii), use the Cauchy-Schwarz inequality:

$$|\gamma_k| = \sigma^2 |\textstyle\sum_{i=0}^{\infty} \psi_i \psi_{i+k}| \leq (\sum_{i=0}^{\infty} \psi_i^2)^{1/2} \sum_{i=0}^{\infty} \psi_{i+k}^2)^{1/2} \to 0 \quad (k \to \infty),$$

as $\sum \psi_i^2 < \infty$, so $\sum_{i=k}^{\infty} \psi_i^2$ is the tail of a convergent series. //

Szegö's theorem.

The key theoretical result in the prediction theory of stationary stochastic processes with discrete time is *Szegö's theorem* (Gabor SZEGÖ (1895-1985) in 1915), according to which the deterministic component in the Wold decomposition is absent (the 'nice case') iff

$$\int_0^{2\pi} \log w(\theta) d\theta > -\infty,$$

where w is the density of the *spectral measure* μ of the process (the logarithm of the density enters here in connection with the concept of *entropy*, which arises in Statistical Mechanics and Thermodynamics). More precisely, by *Kolmogorov's formula* (A. N. KOLMOGOROV (1903-1987) in 1941), the *one-step prediction error* is

$$\sigma^2 := E[(X_0 - E(X_0|X_t, t < 0))^2] = \exp(\frac{1}{2\pi} \int \log w(\theta) d\theta) =: G(\mu) > 0. \tag{K}$$

More general models. We mention a few generalisations here.
1. $ARIMA(p, d, q)$. The 'I' here stands for 'integrated'; the d for how many times. Differencing d times (e.g. to give stationarity) gives $ARMA(p, q)$.
2. $SARIMA$. Here 'S' is for 'seasonal': many economic time series have a seasonal effect (e.g., agriculture, building, tourism).