smfd12(13a).tex
Day 12. 21.11.2013.

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 invert this we 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,

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

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:

$$X_t = \sum\nolimits_1^p \phi_i X_{t-i} + \epsilon_t + \sum\nolimits_1^q \theta_i \epsilon_{t-i}, \qquad (\epsilon_t) \quad WN(\sigma^2)$$

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 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).

$$X_t = (1-\phi B)^{-1}(1+\theta B)\epsilon_t = (1+\theta B)(\sum\nolimits_0^\infty \phi^i B^i)\epsilon_t$$

$$= \epsilon_t + \sum\nolimits_1^\infty \phi^i B^i \epsilon_t + \theta \sum\nolimits_0^\infty \phi^i B^{i+1}\epsilon_t = \epsilon_t + (\theta+\phi)\sum\nolimits_1^\infty \phi^{i-1}B^i\epsilon_t :$$

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

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

$$\gamma_0 = var X_t = E[X_t^2] = \sigma^2 + (\phi+\theta)^2 \sum\nolimits_1^\infty \phi^{2(i-1)}\sigma^2$$

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

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

Covariance: lag $\tau \geq 1$.

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

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

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

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

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

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

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, try an $AR(1,1)$.

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); see V.11 D9 below.

Interpretation of parameters.

Recall the $ARMA(p,q)$ model

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

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

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

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.

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

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

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

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

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

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

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

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).
Spectral methods; frequency domain.

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).

In contrast to the *time-domain* methods above, spectral methods belong to the *frequency domain.*
Wavelets.

One can combine time domain and frequency domain ('time-frequency analysis') by using *wavelets* (1980s on)[1]; we must omit details.

[1]Wavelets are a speciality of the Statistics Section here at Imperial College