V: TIME SERIES

The core of Time Series (TS) analysis is $ARMA(p,q)$ (V.1 below), a combination of $AR(p)$ (*autoregressive* with p parameters) and $MA(q)$ (*moving average* with q parameters). For a thorough textbook treatment, see (in addition to the references under V in W0)
[BJ] G. E. P. BOX and G. M. JENKINS, *Time series analysis: forecasting and control*, Holden-Day, 1970, 553 p (4th ed., with G. C. REINSEL, Wiley, 2008, 746p).
Box and Jenkins did not invent ARMA, but the Box-Jenkins approach did provide econometricians and other applied workers with a toolkit – a procedure – that they could use; this made TS much more accessible than it had been previously, and so the book was very influential in its time.

We will have to take $AR(p)$, $MA(q)$ and other introductory material for granted here. For background and details, see SMF1415 and QRM (Mikko Pakkanen's Quantitative Risk Management course last semester).

1. Autoregressive moving average processes, ARMA(p,q).

As above: because the theory of autoregressive models $AR(p)$ and moving-average models $MA(q)$ is covered in QRM, we can be brief here.

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 B is the *lag operator*, $B : X_t \mapsto X_{t-1}$ and

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

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 for $AR(p)$ and $MA(q)$, 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

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

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

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

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:

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

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

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

2. 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*, *state-space models and the Kalman filter*; see V.5 below.

Interpretation of parameters.

Recall the $ARMA(p,q)$ model

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

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

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

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*.

3. Wold decomposition; spectral methods; time domain and frequency domain

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. [D] 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 to the Wold decomposition and related results is the *Cramér representation* (Harald CRAMÉR (1893-1985) in 1942)

$$X_t = \int_0^{2\pi} e^{it\theta} dY(\theta) \qquad (t \in \mathbb{Z})$$

for a process $Y = (Y(\theta) : \theta \in [0, 2\pi))$ on the unit circle $\mathbb{T}$ ('T for torus'), parametrised by $\theta \in [0, 2\pi)$, or $\mathbb{R}/2\pi\mathbb{Z}$. Thus X is the *Fourier transform* (sequence of Fourier coefficients) of the random signed measure Y. Taking $E[X_t] = 0$, $var(X_t) = 1$ for simplicity, the autocorrelation function $r = (r_t)$ is given by

$$r_t = \int e^{-it\theta} d\mu(\theta),$$

where the *spectral measure* μ satisfies

$$E[|dY(\theta)|^2] = d\mu(\theta).$$

One can pass between the *time domain* – where one looks at the process $X = (X_t)$ in time, working on the L^2-space $L^2(\Omega)$ on the underlying probability space $(\Omega, \mathcal{F}, \mathcal{P})$ as in SP – and the *frequency domain* – where one looks at frequencies θ, and works on the L^2-space $L^2(\mu)$ of the spectral measure μ – via the *Kolmogorov isomorphism*

$$X_t \leftrightarrow e^{it\cdot} = (t \mapsto e^{it\theta}, \theta \in \mathbb{T}) \qquad (t \in \mathbb{Z})$$

(A. N. KOLMOGOROV (1903-1987) in 1941). These L^2-spaces are both *Hilbert spaces*, and Hilbert-space methods play the crucial role here[1].

[1]Hilbert space can be thought of as 'Euclidean space of infinitely many dimensions'. Its study belongs to Functional Analysis – roughly, analysis in infinitely many dimensions. The familiar dot product, or inner product, of vectors in Euclidean space $\mathbb{R}^d$ extends to the dot product of a Hilbert space. The two main examples are the function space L^2, with dot product $(f,g) := \int f\bar{g}$, and the sequence space ℓ^2, with dot product $(a,b) := \sum a_n \bar{b}_n$.

Szegö's theorem.

By *Szegö's theorem* (Gabor SZEGÖ (1895-1985) in 1915), 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).

Hidden frequencies.

Spectral (or Fourier) methods are specially well adapted for searching for hidden frequencies. They can be traced back to work of Lord Kelvin on tides, and to work of Sir Arthur SCHUSTER (1851-1934) of 1897 and 1906 on sunspots (which show a periodicity of around 11 years). They are widely used for detecting the chemical composition of stars from analysing the frequencies found in starlight. An obviously relevant area for Math. Finance is analysing the *business cycle*. Under normal economic conditions (pre-Crash of 2007-8), economic life showed a natural rhythm, in which business activity tended to increase, leading to expansion (and eventually overheating) of the economy (employment increasing, and wages increasing as employers competed for labour), followed by contraction (and eventually depression) of the economy (employment and wages falling). The authorities would try to control this by increasing interest rates to slow the economy down (making it more expensive for firms to borrow to invest), and decreasing interest rates to stimulate the economy. Note that this has not applied since the Crash: we have had long periods of near-zero interest rates, combined with economic depression. The Japanese had even worse experiences, in the 1990s and later (the 'lost decade', or decades).

Wavelets.

One can combine time domain and frequency domain ('time-frequency analysis') by using *wavelets* (1980s on)[2]; we must omit details.

4. ARCH and GARCH; Econometrics ([BF, 9.4.1, 220-222)

There are a number of *stylised facts* in mathematical finance. E.g.:

(i). Financial data show *skewness*. This is a result of the asymmetry between profit and loss (large losses are lethal!)

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

(ii). Financial data have much *fatter tails* than the normal/Gaussian (I.5).
(iii) Financial data show *volatility clustering*. This is a result of the economic and financial environment, which is extremely complex, and which moves between good times/booms/upswings and bad times/slumps/downswings. Typically, the market 'gets stuck', staying in its current state for longer than is objectively justified, and then over-correcting. As investors are highly sensitive to losses (see (i) above), downturns cause widespread nervousness, which is reflected in higher volatility. The upshot is that good times are associated with periods of growth but low volatility; downturns spark extended periods of high volatility (and stagnation, or shrinkage, of the economy).
ARCH and GARCH.

We turn to models that can incorporate such features (volatility clustering, etc.).

The model equations are (with Z_t ind. $N(0,1)$)

$$X_t = \sigma_t Z_t, \qquad \sigma_t^2 = \alpha_0 + \sum_1^p \alpha_i X_{i-1}^2, \qquad (ARCH(p))$$

while in $GARCH(p,q)$ the σ_t^2 term becomes

$$\sigma_t^2 = \alpha_0 + \sum_1^p \alpha_i X_{i-1}^2 + \sum_1^q \beta_j \sigma_{t-j}^2. \qquad (GARCH(p,q))$$

The names stand for (generalised) autoregressive conditionally heteroscedastic (= variable variance). These are widely used in Econometrics, to model *volatility clustering* – the common tendency for periods of high volatility, or variability, to cluster together in time. See e.g. [BFK].
Integrated processes.

One standard technique used to reduce non-stationary processes to the stationary case is to *difference* them repeatedly (one differencing operation replaces X_t by $X_t - X_{t-1}$). If the series of dth differences in stationary but that of $(d-1)$th differences is not, the original series is said to be *integrated of order* d; one writes $(X_t) \sim I(d)$.
Co-integration.

If $(X_t) \sim I(d)$, we say that (X_t) is *cointegrated* with *cointegration vector* α if $\alpha^T X_t)$ is (integrated of) order less than d.

A simple example arises in random walks. If $X_n = \sum_{i=1}^n \xi_i$ with ξ_i iid random variables, $Y_n = X_n + \epsilon_n$ is a noisy observation of X_n, then

$(X, Y) = (X_n, Y_n)$ is cointegrated of order 1, with coint. vector $(-1, 1)^T$.

Cointegrated series are series that move together, and commonly occur in economics. These concepts arose in econometrics, in the work of R. F. ENGLE (1942-) and C. W. J. (Sir Clive) GRANGER (1934-2009) in 1987. Engle and Granger gave (in 1991) an illustrative example – the price of tomatoes in North Carolina and South Carolina. These states are close enough for a significant price differential between the two to encourage sellers to transfer tomatoes to the state with currently higher prices to cash in; this movement would increase supply there and reduce it in the other state, so supply and demand would move the prices towards each other.

Engle and Granger received the Nobel Prize in Economics in 2003. The citation included the following: "Most macroecomomic time series follow a stochastic trend, so that a temporary disturbance in, say, GDP has a long-lasting effect. These time-series are called non-stationary; they differ from stationary series which do not grow over time, but fluctuate around a given value. Clive Granger demonstrated that the statistical methods used for stationary time series could yield wholly misleading results when applied to the analysis of nonstationary data. His significant discovery was that specific combinations of nonstationary time series may exhibit stationarity, thereby allowing for correct statistical inference. Granger called this phenomenon cointegration. He developed methods that have become invaluable in systems where short-run dynamics are affected by large random disturbances and long-run dynamics are restricted to economic equilibrium relationships. Examples include the relations between wealth and consumption, exchange rates and price levels, and short- and long-term interest rates."

Spurious regression.

Standard least-squares method work perfectly well if they are applied to *stationary* time series. But if they are applied to *non-stationary* time series, they can lead to spurious or nonsensical results. One can give examples of two time series that clearly have nothing to do with each other, because they come from quite unrelated contexts, but nevertheless have a high value of R^2. This would normally suggest that a correspondingly high propertion of the variability in one is accounted for by variability in the other – while in fact *none* of the variability is accounted for. This is the phenomenon of *spurious regression*, first identified by G. U. YULE (1871-1851) in 1927, and later studied by Granger and Newbold in 1974. We can largely avoid such pitfalls by restricting attention to stationary time series, as above.

From Granger's obituary (*The Times*, 1.6.2009): "Following Granger's

arrival at UCSD in La Jolla, he began the work with his colleague Robert F. Engle for which he is most famous, and for which they received the Bank of Sweden Nobel Memorial Prize in Economic Sciences in 2003. They developed in 1987 the concept of cointegration. Cointegrated series are series that tend to move together, and commonly occur in economics. Engle and Granger gave the example of the price of tomatoes in North and South Carolina .... Cointegration may be used to reduce non-stationary situations to stationary ones, which are much easier to handle statistically and so to make predictions for. This is a matter of great economic importance, as most macroeconomic time series are non-stationary, so temporary disturbances in, say, GDP may have a long-lasting effect, and so a permanent economic cost. The Engle-Granger approach helps to separate out short-term effects, which are random and unpredictable, from long-term effects, which reflect the underlying economics. This is invaluable for macroeconomic policy formulation, on matters such as interest rates, exchange rates, and the relationship between incomes and consumption."

Endogenous and exogenous variables.

The term 'endogenous' means 'generated within'. The $ARCH$ and $GARCH$ models above show how variable variance (or volatility) can arise in such a way. By contrast, 'exogenous' means 'generated outside'. Exogenous variables might be the effect in a national economy of international factors, or of the national economy on a specific firm or industrial sector, for example. Often, one has a vector autoregressive (VAR) model, where the vector of variables is partitioned into two components, representing the endogenous and exogenous variables. For monograph treatments in the econometric setting, see e.g. [G], [GM].

Discrete and continuous time.

While econometric data arrives discretely (monthly trade figures, daily closing prices for stocks, etc.), continuous time is more convenient for dynamic models of the economy. See e.g.

A. R. BERGSTROM: *Continuous-time econometric modelling*, OUP, 1990.

5. State-space models and the Kalman filter

State-space models originate in Control Engineering. This field goes back to the governor on a steam engine (James WATT (1736-1819) in 1788): to prevent a locomotive going too fast, the governor (a rotating device mounted on top of the engine) rose under centrifugal force as the speed increased, thus operating a valve to reduce the steam entering the cylinders. This was an

early form of *feedback control*.

The Kalman filter (Rudolf KALMAN (1930-2016) in 1960) was a device for online (or real-time) control, suitable for use with *linear* systems, *quadratic* loss and *Gaussian* errors (LQG) (the term *filter* is used because one 'filters out' the noise from the signal to reveal the best estimate of the state). This appeared just when it was needed, for online control of manned spacecraft during the 60s. We shall not develop the control aspects here; see e.g.

M. H. A. DAVIS, *Linear estimation and stochastic control*, Chapman & Hall, 1977,

M. H. A. DAVIS & R. B. VINTER, *Stochastic modelling and control*, Chapman & Hall, 1985.

But the power of the method even without control can be seen in applications such as to *mortar-locating radar*[3]. We follow Whittle ([W], Th. 12.5.2); cf. [BD1] Ch. 12, [BD2] Ch. 8.

The Kalman filter has been extensively applied in Time Series, financial and otherwise; see e.g.

[H] A. C. HARVEY, *Forecasting, structural time series models and the Kalman filter*, CUP, 1991.

Before proceeding to technicalities, we stress one great advantage of the Kalman filter and state-space methods: *they do not depend on stationarity*. Most processes encountered in economics and finance – and indeed, in life generally – are *not* stationary. One can induce stationarity by two main methods: *differencing* (as in the Box-Jenkins ARMA/ARIMA approach – rather brutal), or *discounting* (as in the Black-Scholes approach to option pricing, to get the EMM – but interest rates vary!). State-space models are more direct.

With the engineering example in mind for definiteness, suppose that the *state* of the system at time n is represented by some p-vector $x(n)$. Although the state x is what we are interested in, we cannot observe it directly; what we can observe is a *signal*, or *output* y, $y(n)$ at time n say, a q-vector. We apply a *control* $u(n-1)$, based on information $\mathcal{F}_{n-1}$ available at time $n-1$. The dynamics are represented by the following two equations, the *state equation* (SE) and the *observation equation* (OE):

$$x(n) = A(n-1)x(n-1) + B(n-1)u(n-1)\epsilon(n-1), \qquad (SE)$$

[3]Used in, e.g., the Siege of Sarajevo, 1992-96.

$$y(n) = C(n)x(n) + \eta(n). \qquad (OE)$$

Here $A(.), B(.), C(.)$ are known matrices. The errors $\epsilon(.)$, $\eta(.)$ are p- and q-vectors respectively, with means 0; the errors at different times are all uncorrelated (= independent, if the errors are Gaussian, as we may assume here); the covariance matrices are known matrices

$$cov(\epsilon(n)) = N(n), \qquad cov(\eta(n)) = M(n), \qquad cov(\epsilon(n), \eta(n)) = L(n),$$

In the motivating trajectory example, $A(.)$ comes from the dynamics of the vehicle being tracked, $C(.)$ from the properties of the tracking equipment, $B(.)$ from the control mechanism.

For simplicity, we restrict to the case where $A(n) = A$ for all n, and similarly for B and C; there is no difficulty in extending to the general case.

We write $\hat{x}(n)$ for the best linear predictor (in the sense of minimising expected squared error) of $x(n)$ given the information $\mathcal{F}(n)$ available at time n.

Theorem (Kalman filter).
(i) The conditional distribution of $x(n)$ given $\mathcal{F}(n)$ is $N(\hat{x}(n), V(n))$, where the estimate $\hat{x}(n)$ is given by Kalman filter in (ii) below and the covariance matrix $V(n)$ is given by the Kalman recursion in (iii) below.
(ii) (*Kalman filter*). The estimate $\hat{x}(n)$ is given by the recursion (updating relation)

$$\hat{x}(n) = A\hat{x}(n-1) + Bu(n-1) + H(n)(y(n) - C\hat{x}(n-1)),$$

where

$$H(n) := (L + AV(n-1)C^T)(M + CV(n-1)C^T)^{-1}.$$

(iii) (*Kalman recursion*). The covariance matrix $V(n)$ is given by the recursion (updating relation)

$$V(n) = N + AV(n-1)A^T - (L + AV(n-1)C^T)(M + CV(n-1)C^T)^{-1}(L^T + CV(n-1)A^T).$$

Proof. (i) We start with $x(0) \sim N(\hat{x}(0), V(0))$. That $x(n)|\mathcal{F}(n) \sim N(\hat{x}(n), V(n))$ is clear from the Gaussian Regression Formula of IV.6 on conditioning and regression for the multinormal, and is also proved by induction from the recursions (ii), (iii) below.

Write the estimation error as $\Delta(n) := x(n) - \hat{x}(n)$; then $V(n) = cov(\Delta(n))$.

Now

$$\begin{aligned}\Delta(n-1) &= x(n-1)-\hat{x}(n-1)\\ \epsilon(n-1) &= x(n)-Ax(n-1)-Bu(n-1)\\ \eta(n) &= y(n)-Cx(n-1)\end{aligned}$$

are jointly normal with mean 0 and covariance matrix

$$\begin{pmatrix} V & 0 & 0\\ 0 & N & L\\ 0 & L^T & M\end{pmatrix},$$

where for convenience we write V for $V(n-1)$. We now replace $x(n-1)$ (unobservable) by $\hat{x}(n-1)+\Delta(n-1)$ (we know the first, and know the covariance V of the second), and define

$$\begin{aligned}\zeta^*(n) &:= x(n)-A\hat{x}(n-1)-Bu(n-1)\\ &= x(n)-Ax(n-1)-Bu(n-1)+A(x(n-1)-\hat{x}(n-1))\\ &= \epsilon(n)+A\Delta(n-1),\end{aligned}$$

$$\begin{aligned}\zeta(n) &:= y(n)-C\hat{x}(n-1)\\ &= y(n)-Cx(n-1)+C(x(n-1)-\hat{x}(n-1))\\ &= \eta(n)+C\Delta(n-1).\end{aligned}$$

Then

$$\begin{pmatrix}\zeta^*(n)\\ \zeta(n)\end{pmatrix}=\begin{pmatrix}A\Delta(n-1)+\epsilon(n)\\ C\Delta(n-1)+\eta(n)\end{pmatrix}=\begin{pmatrix}A & 1 & 0\\ C & 0 & 1\end{pmatrix}\begin{pmatrix}\Delta(n-1)\\ \epsilon(n)\\ \eta(n)\end{pmatrix}\sim N(0,\Sigma),$$

where the covariance matrix Σ is given by

$$\Sigma=\begin{pmatrix}A & 1 & 0\\ C & 0 & 1\end{pmatrix}\begin{pmatrix}V & 0 & 0\\ 0 & N & L\\ 0 & L^T & M\end{pmatrix}\begin{pmatrix}A^T & C^T\\ 1 & 0\\ 0 & 1\end{pmatrix}=\begin{pmatrix}A & 1 & 0\\ C & 0 & 1\end{pmatrix}\begin{pmatrix}VA^T & VC^T\\ N & L\\ L^T & M\end{pmatrix}:$$

$$\Sigma=\begin{pmatrix}N+AVA^T & L+AVC^T\\ L^T+CVA^T & M+CVC^T\end{pmatrix}.$$

Both (ii) (conditional means) and (iii) (conditional variances) now follow from the Gaussian Regression Formula of IV.6, W4. //

It is difficult to overestimate the practical importance of this key result. It has proved invaluable in many areas since its introduction in 1960.

Extensions.

1. *Non-Gaussian errors.*

The result extends beyond the context of Gaussian errors (multivariate normal distribution) above. One does not obtain the full distribution, but works instead with means and variances. See e.g. Whittle [W], 12.8 and Th. 12.9.4; see also the Bayes linear estimate (SMF1415, VII.7.8 D19).

2. *Prediction further into the future.*

The method above can be readily adapted to prediction k time-steps into the future. This is done in detail in [BD2], 12.3.

3. *Smoothing.*

Instead of predicting the future, one can instead seek to get the best fit we can to the data. The mathematics is very similar; see e.g. [BD2], Prop. 12.2,3, 4.

4. *Riccati equation.*

The non-linear recursion (iii) is a matrix *Riccati equation*, and this name is often used instead of Kalman recursion.

5. *Off-line calibration.*

To use a Kalman filter, one needs the relevant matrices, A, B, C, L, M, N. In practice, these will have to be estimated numerically. This can be done off-line, 'at leisure'. Once accurate (enough) numerical estimates of these matrices are known, and the recursions (ii) and (iii) programmed, the filter can be used online (in real time).

6. *Innovations.*

The *innovations* are $I(n) := y(n) - C\hat{x}(n-1)$. These are the differences between an observation $y(n)$ and the prediction $C\hat{x}(n-1)$ we would have made for it at time $n-1$ (from the observation equation (OE)). This is the new information at time n – beyond what we could have predicted. They are mutually uncorrelated (independent, in the Gaussian case, as here). One can base the theory on them ([W], 12.7).

7. *Continuous time.*

One can work instead in continuous time, where the recurrence (or difference) equations above are replaced by differential equations. See e.g. [W],

Ch. 20. Hence the name Riccati – Riccati's differential equation.

8. *Hilbert-space methods.*

The prediction above is done in the least-squares sense – to minimise the expected squared errors. This has a nice geometrical interpretation in terms of projections (see e.g. [BF], Ch. 4). In our finite-dimensional setting, this just involves Euclidean geometry, but the method works just as well in infinitely many dimensions – *Hilbert space* ('Euclidean space of infinitely many dimensions' – see V.9).

9. *Nonlinear systems.*

The Kalman filter is linear, and (as linearity and Gaussianity are so closely linked) works very well in the Gaussian case. However, in practice one encounters non-linear systems (and non-Gaussian errors). The *extended Kalman filter* reduces to the linear case by linearisation. This works well in some applications (such as GPS – geographic positioning systems). But it does not always give good results – for example, it may not be numerically stable. Also, to implement it one needs computer-intensive methods such as MCMC (Markov chain Monte Carlo), particle filters etc.[4]

10. *Financial applications.*

The Kalman filter has been extensively applied in finance (e.g., for calibration of interest-rate models). For background, see e.g.

C. WELLS, *The Kalman filter in finance*, Springer, 1996;

11. *State-space models for time series.*

The Kalman filter, and state-space models generally, have also been extensively used in Time Series; see e.g. Harvey [H] above, and

J. DURBIN & S. KOOPMAN, *Time series analysis by state-space methods*, OUP, 2001.

12. *Change-point detection.*

One important application is in automatic control of industrial production. If a machine in use begins to deteriorate, or deviate from its required performance level (for lack of maintenance, etc.), it is important to be able to *detect* this *as quickly as possible*. Such quick-detection problems are an important area of application of the Kalman filter and its relatives.

13. *Control.*

For further background on Control Theory, see e.g.

M. H. A. DAVIS, *Linear estimation and stochastic control*, Chapman & Hall,

[4]MCMC and particle filters are specialities of the Imperial College Mathematics Department, and Professor Dan Crisan.

1977,
M. H. A. DAVIS & R. B. VINTER, *Stochastic modelling and control*, Chapman & Hall, 1985.
Professor M. H. A. (Mark) Davis (1945-) was the founding Professor of Mathematical Finance at Imperial College, and set up this MSc in 2000.

§6. The Yule-Walker equations.

Recall the model for $AR(p)$:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t, \qquad (*)$$

with (ϵ_t) WN as before.

Multiply $(*)$ through by X_{t-k} and take expectations. Since $E[X_{t-k}X_{t-i}] = \rho(|k-i|) = \rho(k-i)$, this gives

$$\rho(k) = \phi_1\rho(k-1) + \cdots + \phi_p\rho(k-p) \qquad (k>0). \qquad (YW)$$

These are the *Yule-Walker equations*, due to G. Udny YULE (1871-1951) in 1926 and Sir Gilbert WALKER (1868-1958) in 1931.

The Yule-Walker equations (YW) have the form of a *difference equation* of *order p*. The *characteristic polynomial* of this difference equation is

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

which by above is

$$\phi(1/\lambda) = 0.$$

If the roots are $\lambda_1, \cdots, \lambda_p$, the trial solution $\rho(k) = \lambda^k$ is a solution iff λ is one of the roots λ_i. Since the equation is linear,

$$\rho(k) = c_1\lambda_1^k + \cdots + c_p\lambda_p^k$$

(for $k \geq 0$, and use $\rho(-k) = \rho(k)$ for $k < 0$) is a solution for all choices of constants $c_1, \cdots, c_p$. This is the *general solution* of (YW) if all the roots λ_i are distinct, with appropriate modifications for repeated roots (if $\lambda_1 = \lambda_2$, use $c_1\lambda_1^k + c_2 k\lambda_1^k$, etc.).

Now $|\rho(k)| \leq 1$ for all k (as $\rho(.)$ is a correlation coefficient), and this is *only possible* if

$$|\lambda_i| \leq 1 \qquad (i = 1, \cdots, p)$$

– that is, all the roots lie inside (or on) the unit circle. This happens (as our polynomial is $\phi(1/\lambda)$) if and only if *all the roots of the polynomial* $\phi(\lambda)$ *lie outside (or on) the unit circle.* Then $|\rho(k)| \leq 1$ for all k, and when there are no roots of unit modulus, also $\rho(k) \to 0$ as $k \to \infty$ – that is, the influence of the remote past tends to zero, as it should. This is also the condition for the $AR(p)$ process above to be *stationary.*

The remote past.

In some physical systems, the influence of the remote past does indeed become negligible. Example: bathwater – when we run a bath, the detailed thermal history is forgotten, as the hot and cold water *thermalise* when they mix. Sometimes it does not. Example: tempered steel – here the detailed thermal history is locked in by the tempering process, and is what gives tempered steel its special qualities.