smfd9(13).tex

Day 9. 5.6.2013

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

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

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. Harvey 8.3, [BF] 9.4, [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 Ecomomics 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].

11. 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-) 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*[1]. We follow Hannan [H] III.8 (cf. [BD1] Ch. 12, [BD2] Ch. 8).

The Kalman filter has been extensively applied in Time Series, financial and otherwise. We cited Harvey's *Time series models* in D0; see also
A. C. HARVEY, *Forecasting, structural time series models and the Kalman filter*, CUP, 1991;
C. WELLS, *The Kalman filter in finance*, Springer, 1996;
J. DURBIN & S. KOOPMAN, *Time series analysis by state-space methods*, OUP, 2001.

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, or $y(n)$ at time n, a q-vector. The dynamics are represented by the following two equations, the *state equation* (SE) and the *observation equation* (OE):

$$x(n) = \Phi(n-1)x(n-1) + \epsilon(n-1), \qquad (SE)$$

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

Here $\Phi(.)$ is a $p \times p$ matrix, $H(.)$ a $q \times q$ matrix, both known. 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

$$cov(\epsilon(n)) = Q(n), \qquad cov(\eta(n)) = R(n), \qquad cov(\epsilon(n), \eta(n)) = S(n),$$

In the motivating trajectory example, $\Phi(.)$ comes from the dynamics of the vehicle being tracked, $H(.)$ from the properties of the tracking equipment.

We use the terminology of Hilbert space, which turns out to be the natural one for Time Series; for background, see e.g. Hannan [H], Brockwell & Davis [BD1], [BD2] or my survey
NHB, Szegö's theorem and its probabilistic descendants, *Probability Surveys* **9** (2012), 287-324 (and its multi-dimensional sequel, ibid. 325-339).
Hilbert space can be thought of as "Euclidean space of infinitely many dimensions" (though here we have finitely many finite-dimensional vectors, so things are in fact Euclidean). As in Euclidean space, one has a length (or norm), an inner product (generalising the ordinary dot product of vectors),

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

Pythagoras' theorem holds, and one can use projections (including drawing pictures of them), and think geometrically. We call subspaces A, B *orthogonal* if any $a \in A$ and $b \in B$ are orthogonal vectors, i.e. $(a, b) = 0$.

In this Hilbert-space setting, the best linear predictor given some information is the *projection* onto the (vector) subspace spanned by the vectors given. We need two facts here; see e.g. Rao [R], 1.c4(vi), (vii), p.47-8. The first is as in IV.6 D6; the second is the *Bayes linear estimator* of VII.7.7 D12.
(P1). The projection onto the space spanned by two orthogonal subspaces is the sum of the projections onto each of them separately.
(P2). The projection of x onto the space spanned by z is

$$\hat{x} = E[xz^*](E[zz^*])^{-1}z,$$

where $(.)^{-1}$ is the matrix inverse when this exists (which it will in our case), or a generalised inverse if it does not, and A^* denotes the transposed complex conjugate of a matrix A (complex Hilbert spaces are mathematically preferable to real ones and no harder, hence A^* for A^T even for real A).

We write $\mathcal{H}_n$ for the Hilbert space spanned by the errors $\eta(j)$ $(j \leq n)$, $\epsilon(k)$ $(k \leq n-1)$, together with $x(0)$, the initial state, $\mathcal{K}_n$ for the Hilbert space spanned by the signals $y(0), \ldots, y(n)$. From $(SE), (OE)$, $\mathcal{M}_n \subset \mathcal{H}_n$. We write $a \perp B$ to mean that vector a is orthogonal to (uncorrelated with, so independent of) every vector $b \in B$, $\hat{x}(n+m|n)$ for the best linear predictor (projection) of $x(n+m)$ on the space $\mathcal{M}_n$ spanned by $y(k)$, $k \leq n$.

Theorem (Kalman Filter). (i) For $m > 1$,

$$\hat{x}(n+m|n) = \Phi(n+m-1)\hat{x}(n+m-1|n).$$

(ii)

$$\hat{x}(n+1|n) = \Psi(n)\hat{x}(n|n-1) + K(n)y(n),$$

where

$$\Psi(n) := \Phi(n) - K(n)H(n),$$

$$K(n) := \{\Phi(n)\Sigma(n)H(n)^T + S(n)\}\{H(n)\Sigma(n)H(n)^T + R(n)\}^{-1},$$

and $\Sigma(n)$ is defined recursively by

$$\Sigma(n+1) := \Phi(n)\Sigma(n)\Phi(n)^T + Q(n) - K(n)\{H(n)\Sigma(n)H(n)^T + R(n)\}K(n)^*.$$

Proof. $\mathcal{M}_{n-1} \subset \mathcal{M}_n$, and $\mathcal{M}_n = \mathcal{M}_{n-1} \oplus \mathcal{V}_n$, where $\mathcal{V}_n$ is the *orthogonal complement* of $\mathcal{M}_{n-1}$ in $\mathcal{M}_n$. We have $\eta(n) \perp \mathcal{M}_{n-1}$ (since the errors are

all orthogonal). So the part of $y(n)$ that depends on times $k \leq n-1$ is thus $H(n)\hat{x}(n|n-1)$. So writing $\Pi(.|V)$ for projection onto a subspace V,

$$\Pi(y(n)|\mathcal{M}_{n-1}) = H(n)\hat{x}(n|n-1).$$

Thus $\mathcal{V}_n$ is spanned by

$$I(n) := y(n) - H(n)\hat{x}(n|n-1),$$

the *innovation* at time n (Hannan uses $z(n)$ for this ('z after x and y'), but the suggestive name innovation and notation $I(n)$ is more usual now). So

$$I(n) = \eta(n) + H(n)\{x(n) - \hat{x}(n|n-1)\}. \qquad (I)$$

Projecting $x(n+1)$ onto $\mathcal{M}_n = \mathcal{M}_{n-1} \oplus \mathcal{V}_n$ and using (P1),

$$\hat{x}(n+1|n) = \Phi(n)\hat{x}(n|n-1) + u(n), \quad u(n) := \Pi(x(n+1)|\mathcal{V}_n). \qquad (*)$$

By (P2),

$$u(n) = E[x(n+1)I(n)^*][E(I(n)I(n)^*)]^{-1}I(n).$$

Write

$$\Sigma(n) := E[(x(n) - \hat{x}(n|n-1))(x(n) - \hat{x}(n|n-1))^T]$$

for the covariance matrix of $x(n) - \hat{x}(n|n-1)$. Then as $R(n)$ is the covariance matrix of $\eta(n)$ and $x(n) - x(n|n-1)$ is orthogonal to (independent of) the error $\eta(n)$ in $y(n)$, (I) gives

$$E[I(n)I(n)^*] = H(n)\Sigma(n)H(n)^T + R(n).$$

Similarly,

$$E[x(n+1)I(n)^*] = \Phi(n)\Sigma(n)H(n)^T.$$

Combining, and using the definition of $K(n)$ in the statement of the Theorem,

$$u(n) = K(n)I(n).$$

So

$$\begin{aligned} \hat{x}(n+1|n) &= \Phi(n)\hat{x}(n|n-1) + K(n)I(n) \\ &= \{\Phi(n) - K(n)H(n)\}\hat{x}(n|n-1) + K(n)\{I(n) + H(n)\hat{x}(n|n-1)\}. \end{aligned}$$

From the definitions of $\Psi(n)$ in the Theorem and $I(n)$ above, this says

$$\hat{x}(n+1|n) = \Psi(n)\hat{x}(n|n-1) + K(n)y(n),$$

giving (ii); (i) is similar. For the *Kalman recursion* for $\Sigma(n)$, see the handout on the course website, or the books cited.

VI. NON-PARAMETRICS

1. Empiricals; the Glivenko-Cantelli theorem

The first thing to note about Parametric Statistics is that the parametric model we choose will only ever be approximately right at best. We recall *Box's Dictum* (the English statistician George E. P. BOX (1919 –)): *al models are wrong – some models are useful.* For example: much of Statistics uses a normal model in one form or other. But no real population will ever be exactly normal. And even if it were, when we sampled from it, we would destroy normality, e.g. by the need to *round* data to record it; rounded data is necessarily rational, but a normal distribution takes irrational values a.s.

So we avoid choosing a parametric model, and ask what can be done without it. We sample from an unknown population distribution F. One important tool is the *empirical* (distribution function) F_n of the sample $X_1, \ldots, X_n$. This is the (random!) probability distribution with mass $1/n$ at each of the data points X_i. Writing δ_c for the *Dirac* distribution at c – the probability measure with mass 1 at c, or distribution function of the constant c –

$$F_n := \frac{1}{n} \sum_1^n \delta_{X_i}.$$

The next result is sometimes called the *Fundamental Theorem of Statistics.* It says that, in the limit, we can recover the population distribution from the sample: *the sample determines the population in the limit.* It is due to V. I. GLIVENKO (1897-1940) and F. P. CANTELLI (1906-1985), both in 1933, and is a uniform version of Kolmogorov's Strong Law of Large Numbers (SLLN, or just LLN), also of 1933.

Theorem (Glivenko-Cantelli Theorem, 1933).

$$\sup_x |F_n(x) - F(x)| \to 0 \qquad (n \to \infty) \qquad a.s.$$

Proof. Think of obtaining a value $\leq x$ as Bernoulli trials, with parameter (= success probability) $p := P(X \leq x) = F(x)$. So by SLLN, for each fixed x,

$$F_n(x) \to F(x) \qquad a.s.,$$

as $F_n(x)$ is the proportion of successes. Now fix a finite partition $-\infty = x_1 < x_2 < \ldots < x_m = +\infty$. By monotonicity of F and F_n,

$$\sup_x |F_n(x) - F(x)| \leq \max_k |F_n(x_k) - F(x_k)| + \max_k |F(x_{k+1} - F(x_k)|.$$

Letting $n \to \infty$ and refining the partition indefinitely, we get

$$\limsup_n \sup_x |F_n(x) - F(x)| \leq \sup_x \Delta F(x) \qquad a.s.,$$

where $\Delta F(x)$ denotes the jump of F (if any – there are at most countably many jumps!) at x. This proves the result when F is continuous.

In the general case, we use the Probability Integral Transformation (PIT, IS, I). Let $U_1, \ldots, U_n \ldots$ be iid uniforms, $U_n \sim U(0,1)$. Let $Y_n := g(U_n)$, where $g(t) := \sup\{x : F(x) < t\}$. By PIT, $Y_n \leq x$ iff $U_n \leq F(x)$, so the Y_n are iid with law F, like the X_n, so wlog take $Y_n = X_n$. Writing G_n for the empiricals of the U_n,

$$F_n = G_n(F).$$

Writing A for the range (set of values) of F,

$$\sup_x |F_n(x) - F(x)| = \sup_{t \in A} |G_n(t) - t| \leq \sup_{[0,1]} |G_n(t) - t|, \to 0 \qquad a.s.,$$

by the result (proved above) for the continuous case. //

If F is continuous, then the argument above shows that

$$\Delta_n := \sup_x |F_n(x) - F(x)|$$

is *independent* of F, in which case we may take $F = U(0,1)$, and then

$$\Delta_n = \sup_{t \in (0,1)} |F_n(t) - t|.$$

Here Δ_n is the *Kolmogorov-Smirnov (KS) statistic*, which by above is *distribution-free* if F is continuous. It turns out that there is a uniform CLT corresponding to the uniform LLN given by the Glivenko-Cantelli Theorem: $\Delta_n \to 0$ at rate $\sqrt{n}$. The limit distribution is known – it is the *Kolmogorov-Smirnov (KS) distribution*

$$1 - 2\sum_1^\infty (-)^{k+1} e^{-2k^2x^2} \qquad (x \geq 0).$$

It turns out also that, although this result is a limit theorem for *random variables*, it follows as a special case of a limit theorem for *stochastic processes*. Writing B for Brownian motion, B_0 for the Brownian bridge ($B_0(t) := B(t) - t$, $t \in [0,1]$),

$$Z_n := \sqrt{n}(G_n(t) - t) \to B_0(t), \qquad t \in [0,1]$$

(*Donsker's Theorem*: Monroe D. DONSKER (1925-1991) in 1951 – originally, the *Erdös-Kac-Donsker Invariance Principle*). The relevant mathematics here is *weak convergence of probability measures* (under an appropriate topology). Thus, the KS distribution is that of the supremum of Brownian bridge. For background, see e.g. Kallenberg Ch. 14.
Higher dimensions.

In one dimension, the half-lines $(-\infty, x]$ form the obvious class of sets to use – e.g., by differencing they give us the half-open intervals $(a, b]$, and we know from Measure Theory that these suffice. In higher dimensions, obvious analogues are the half-spaces, orthants (sets of the form $\prod_{k=1}^{n}(-\infty, x_k]$), etc. – the geometry of Euclidean space is much richer in higher dimensions. We call a class of sets a *Glivenko-Cantelli class* if a uniform LLN holds for it, a *Donsder class* if a uniform CLT holds for it. For background, see e.g.
[vdVW] A. W. van der VAART & J. A. WELLNER, *Weak convergence and empirical processes, with applications to statistics*, Springer, 1996, Ch. 2.
This book also contains a good treatment of the *delta method* in this context – the *von Mises calculus* (Richard von MISES (1883-1953), or *infinite-dimensional delta method.*

Variants on the problem above include:
1. *The two-sample Kolmogorov-Smirnov test.*

Given two populations, with unknown distributions F, G, we wish to test whether they are the same, on the basis of empiricals F_n, G_m.
2. *Kolmogorov-Smirnov tests with parameters estimated from the data.*

A common case here is *testing for normality.* In one dimension, our hypothesis of interest is whether or not $F \in \{N(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma > 0\}$. Here (μ, σ) are *nuisance parameters*: they occur in the formulation of the problem, but not in the hypothesis of interest.

Although the Glivenko-Cantelli Theorem is useful, it does not tell us, say, whether or not the law F is absolutely continuous, discrete etc. For, there are discrete G arbitrarily close to an absolutely continuous F (discretise), and absolutely continuous F arbitrarily close to a discrete F (by smooth approximation to F at its jump points). So sampling alone cannot tell us what *type* of law F is – absolutely continuous (with density f, say), discrete, continuous singular, or some mixture of these. So it makes sense for the statistician to *choose* what kind of population distribution he is going to assume. Often (usually), this will be absolutely continuous; again, it makes sense to *assume* what smoothness properties of the density f we will assume. This leads on to the important subject of density estimation, to which we now turn.

2. Curve and surface fitting.

We begin with some background. Suppose we have n points (x_i, y_i), with the x_i distinct, and we wish to *interpolate* them – find a function f with $f(x_i) = y_i$, $i = 1, \ldots, n$. One can of course do this by linear interpolation between each adjacent pair of points, obtaining a continuous piecewise-linear function – but this is not smooth enough for many purposes. One might guess that as a polynomial of degree $n-1$ contains n degrees of freedom (its n coefficients), it might be possible to interpolate by such a polynomial, and this is indeed so (Lagrangian interpolation, or Newtonian divided-difference interpolation). There is a whole subject here – the Calculus of Finite Differences (the discrete analogue of the ordinary ('infinitesimal') calculus).

The degree n may be large (should be large – the more data, the better). But, polynomials of large degree are very oscillatory and numerically unstable. We should and do avoid them. One way to do this is to use *splines*. These are continuous functions, which are polynomials of some chosen low degree (*cubic splines* are the usual choice in Statistics) *between* certain special points, called *knots* (or *nodes*), across which the function and as many derivatives as possible are continuous. So a cubic spline is piecewise cubic; it and its first two derivatives continuous are across the knots.

Another relevant piece of background is the *histogram*, familiar from elementary Statistics courses. One represents discrete data diagrammatically, with vertical bars showing how many data points fall in each subinterval.

Computer implementation is necessary to use methods of this kind in practice. For a general account using the computer language S (from which R, and the proprietary package S-Plus, are derived), see e.g. [VR], 5.6.

Roughness penalty.

Using polynomials of high degree, we can fit the data exactly. But we don't, because the resulting function would be too rough ('too wiggly'). It is better to fit the data approximately rather than exactly, but obtain a nice smooth function at the end. One way to formalise this (due to I. J. GOOD (1916-2009) and his pupil R. A. Gaskins in 1971) is to use a *roughness penalty* – to measure the roughness of the function by some integrated measure – $\int (f'')^2$ is the usual one for use with cubic splines – and minimise a combination of this and the relevant sum of squares (see IV, [BF] 9.2):

$$\min \quad \sum_1^n (y_i - f(x_i))^2 + \lambda^2 \int (f'')^2.$$

Here λ^2 is the *smoothing parameter*. It is under the control of the statistician, who can choose how much weight to give to goodness of fit (the first term) and how much to smoothness (roughness being measured by the second term).
1. Density estimation. Suppose we want to find as good a fit to the data as possible using a density function with smoothness properties that we have chosen (see above). One way to do this is to make two key choices:
(a) the *kernel* $K(.)$. This is a density with the required smoothness properties;
(b) the *bandwidth* $h > 0$ (also called the *window width*).
One then defines the *kernel density estimator*

$$\hat{f}(x) := \frac{1}{nh} \sum_1^n K\Big(\frac{x - X_i)}{h}\Big).$$

This is again a density, with the same smoothness properties as K. It turns out that the properties of $\hat{f}$ are mainly determined by h, and the choice of K is less important. We must refer for detail here to a specialised text, e.g. Silverman [Sil], Tapia and Thompson [TapT]. Such books contain graphics, comparing kernel density estimates with histograms of the data.

Silverman's book (4.2.3 Scatter plots, p. 81-83, Figs 4.6 – 4.8) contains a contour plot of the two-dimensional density of a clinical measurement in the treatment of a disease. Fig. 4.7 reveals that the contour plot is *bimodal* – has two peaks (this will be familiar to those of you with map-reading experience in hilly country, and is visually clear anyway). This suggested – correctly – that there were in fact two different sub-populations present. Two different types of this disease were identified, and different treatments developed for them – a good example of an unexpected benefit from density estimation.

One can see similar effects more easily, in one dimension. If a histogram of adult heights were plotted, it would again be bimodal. The reason is obvious: males are statistically taller than females. So here *sex*, or gender, is a relevant *factor* (recall that we met factor analysis briefly in III.3, III.5).

A less obvious example arises in teaching UK undergraduate mathematics students. Again, exam scores after one year are bimodal. This reflects the still-visible effects of having some students with single maths at A Level and some with double maths. This difference is much less marked in later years.

The statistical moral here is clear. Bi- or multi-modality of a population suggests that the population is heterogeneous. We should seek to identify relevant *factors*[2] causing this heterogeneity, disaggregate accordingly, and

[2]There is a whole subject, Factor Analysis – see [MKB], [K].

analyse the sub-populations separately. Otherwise the aspect we wish to study becomes entangled with (*confounded with*) these factors.
2. Non-parametric regression. This extends and complements the parametric regression in Ch. IV. The ideas above can be used to extend these ideas to a non-parametric setting, using roughness penalties, cubic splines etc. We refer for detail to, e.g., [BF], 9.2.
3. Semi-parametric regression. This combines Ch. IV and VI: see e.g.
D. RUPPERT, M. P. WAND & R. J. CARROLL: Semi-parametric regression during 2003-07. *Electronic J. Statistics* **3** (2009), 1193-1256 [free, online], + refs there, and book *Semi-parametric regression* (same authors, CUP, 2003).
4. Volatility surfaces. The volatility σ in the Black-Scholes formula is unknown, and has to be estimated – either as *historic volatility* from time-series data (Ch. V), or as *implied volatility* – the Black-Scholes price is (continuous and) increasing in σ ('options like volatility'), so one can infer 'what the market thinks σ is' from the prices at which options currently trade. Closer examination reveals that the volatility is not constant, but varies – e.g., with the strike price ('volatility smiles'). Volatility is observed to vary so unpredictably that it makes sense to model is as a stochastic process (*stochastic volatility, SV*). Market data is discrete, but for visual effect it is better to use computer graphics and a continuous representation of such *volatility surfaces*. For a monograph treatment, see Gatheral [Gat].
Note. Because of the asymmetry between profit and loss, one often encounters skewness in financial data. In the context of the volatility smile, one obtains a skew smile, known as the *volatility smirk*[3].

The VIX – volatility index (colloquially called the 'fear index') is widely used, and is the underlying for volatility derivatives. It has even affected literature (see e.g. John Harris' novel *The fear index*, Hutchinson, 2011).
5. Stochastic volatility and state-space models. Compare with V.11. In each, one has a coupled set of equations (difference equations in discrete time, differential equations in continuous time). The state variable plays the role of the volatility – both unobserved.
6. Image enhancement. Images (of faces, moonscapes etc.) are typically corrupted by 'noise'. When these are digitised, into pixels, techniques such as the *Gibbs sampler* (VI.4, VII.6) can improve quality, by iterations in which a pixel is changed to improve agreement with 'a consensus of neighbours'.

[3] A smirk is a smile one is ashamed of, and this negative feeling is often betrayed by a visible asymmetry.