smfd14(13a).tex

Day 14. 28.11.2013.

12. Complements.

1. Akaike Information Criterion (AIC).

In choosing a model, we are torn between two conflicting objectives. One is *goodness of fit*, and here we can achieve a better fit by using a more complex model, with more parameters. The other is simplicity – according to the *Principle of Parsimony*, one should use the simplest model that will do the job. In order to achieve a sensible balance between these two, we use a *penalised likelihood* method – likelihood, penalised by the number of parameters. The simplest and commonest is the *Akaike information criterion (AIC)*,: if there are p parameters in the model,

$$AIC := -2\text{log-likelihood} + 2(p+1)$$

($p+1$ parameters, counting the variance σ^2, usually unknown) (H. Akaike (1927-) in 1974). For computer implementation, see e.g.

[VR] W. N. VENABLES & B. D. RIPLEY, *Modern applied statistics with S*, 4th ed., Springer, 2002, p. 174.

There is a variant on AIC, the *Bayesian information criterion (BIC)* (Gideon E. Schwarz, 1978); see e.g. [BD], p.291, and Ch. VII below.

Over-interpretation/over-fitting.

It is important to note that, in general terms, one should resist the temptation to achieve a better fit with a more complex model. Our data is to be modelled; given a model, what we see is data = model prediction + noise. The first term is the signal, or trend; the second term is the error. Our task is to try to iron out the noise or error so as to reveal the signal or trend more clearly. Going for a better fit treats the error with "too much respect": the errors we actually observed (accidental, and of no interest in themselves) are left in our model too much. This gives a better fit to the data already observed, but a misleadingly complicated model which will give a worse fit to future data. The phenomenon is called *over-interpretation*, or *over-fitting*. It is quite general, and not specific to the present context of Time Series. For further discussion, see a Statistics textbook, e.g. [BF].

2. Unit roots and the (augmented) Dickey-Fuller test.

For simplicity, recall [V.3, Day 10] the $AR(1)$ model:

$$X_t = \phi X_{t-1} + \epsilon_t; \qquad (*)$$

$$E[X_t|\mathcal{F}_{t-1}] = \phi X_{t-1}.$$

In probability language, this says that X is a martingale for $\phi = 1$, a submg for $\phi > 1$ and a supermg for $\phi < 1$. In econometric language, the impact of X_t will die out with time if $\phi < 1$, but not if $\phi = 1$. Thus the influence of the current state dies away with time for $\phi < 1$ (which we need to have a stationary model), but not for $\phi = 1$. The distinction is vital for econometrics (compare the discussion in V.10, Day 13, about the work of Engle and Granger): if an economy is (say) depressed, policy makers want it to recover, not get stuck in its current state.

For such reasons, *unit roots* are dangerous. One needs to avoid them, and to be able to test for their presence. One way to avoid unit roots is to *difference* the data (as many times as necessary; cf. the above discussion of $ARIMA(p, d, q)$ extensions to $ARMA(p, q)$. Various tests for unit roots have been developed, most notably the *Dickey-Fuller (DF)* and *augmented Dickey-Fuller (ADF)* tests. The theory would take us too far afield here; for background and details, we must refer to an econometrics text, e.g.
T. C. MILLS, *The econometric modelling of financial time series*, 2nd ed., CUP, 1999 (1st ed. 1993), §§3.1.2, 3.1.3,
J. Y. CAMPBELL, A. W. LO & A. C. MacKINLAY, *The econometrics of financial markets*, Princeton UP, 1997, §2.7.
For computer implementation, see e.g.
R. A. CARMONA, *Statistical analysis of financial data in S-Plus*, Springer, 2004, §5.4.7.

3. Residuals and the Ljung-Box test.

When fitting a model to data, one has

data = trend [or signal] + error [or noise] = fitted value + residual.

We are using *white noise* as our error – no structure or pattern.

The first thing to check is whether there *is* a signal (if not, there is no point in trying to estimate it!). Exploratory data analysis (EDA) will suggest absence of a signal if the data seems patternless; we then test for this (below).

If there is a signal, our fitted model should reveal as much as possible of the signal, leaving (ideally) merely noise – patternless. So a good model produces *patternless residuals*. One can (and should!) inspect for this by EDA. One can also test for it.

One revealing aspect of pattern here is *serial correlation* (recall the correlogram $r = (r_n)$ of §1). In the Gaussian case (which we can restrict to here), uncorrelatedness is the same as independence. Recall that the sum of squares of n iid $N(0, \sigma)$s is distributed as $\sigma^2 \chi^2(n)$. From this, Ljung and Box (1978) developed their test: for white noise (patternless),

$$Q_m := n(n+2) \sum_{k=1}^{m} r_k^2/(n-k) \sim \chi^2(m) \qquad (n \text{ large, } m << n).$$

One rejects the null hypothesis (Ch. II!) of no serial correlation if the Q_m statistic is too big. Our current model then fails to fit well enough, and we must look further for one that does. For details, see e.g.
P. J. DIGGLE, *Time series: A biostatistical introduction*, OUP, 1990, §2.5.

4. Markov chain Monte Carlo (MCMC); sequential MCMC (particle filters).

We mentioned [V.11, Day 13] the extended Kalman filter, and its application to *non-linear* situations (recall the Kalman filter applies to the LQG case – linear, quadratic, Gaussian), and that its implementation involves computer-intensive methods such as MCMC and particle filters.

We discuss MCMC briefly in VI.4 [Day 15] below, and its statistical application in VII.6. In brief: to simulate from a distribution we cannot tackle directly, it may be possible instead to construct a Markov chain whose limiting distribution is the desired distribution. Then if we run the chain for long enough, its distribution will approximate the desired limit.

Particle filters involve various ideas:
Importance sampling in Simulation. Here, we focus our simulation effort on the parts of the region which are of most interest to us.
Non-parametrics (Ch. VI below), in particular *empiricals*. The idea here is to replace an unknown distribution by a (random) empirical distribution – a weighted average of points drawn from the unknown distribution.
Each of these random points is regarded as a *particle*. Particles in regions of less importance are weeded out, and replaced by copies drawn from regions of greater importance. The method is powerful; for a full treatment, see e.g.
A. BAIN & D. CRISAN, *Fundamentals of stochastic filtering*, Springer, 2009, Ch. 9, 10.

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 $X_1, \dots, X_n$. 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 $-\infty = x_1 < x_2 < \dots < x_m = +\infty$. 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

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

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, $Y_n \le x$ iff $U_n \le F(x)$, 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,

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

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 – the *Kolmogorov-Smirnov distribution*

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

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]. 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 the law F is absolutely continuous, discrete etc. For, there are discrete G arbitrarily close to an abs. cts F (discretise), and abs. cts 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. So we have to *choose* what kind of population distribution to assume. Often this will have a density f; we have to *assume* how smooth to take f. This leads on to *density estimation*, below.

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 e.g. [Sil], which contains 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*[1] causing this heterogeneity, disaggregate accordingly, and 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. One can extend this to a non-parametric setting, using roughness penalties, cubic splines etc.; see e.g. [BF], 9.2.

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