smfd15.tex
Day 15. 5.11.2014.

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

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

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.

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*[2].

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'.

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

3. Non-parametric likelihood

At first glance, 'non-parametric likelihood' seems a contradiction in terms (an oxymoron – 'square circle', etc.) But it turns out that maximum-likelihood estimation (MLE) can indeed be usefully combined with non-parametrics. First, we interpret the empirical F_n as a non-parametric MLE (NPMLE) for the unknown true distribution F. For, if the data is $\{x_1, \ldots, x_n\}$, the *likelihood* of F is $L(F) := \prod_1^n \Delta F(x_i)$ (where $\Delta F(x) := F(x) - F(x-)$ is the probability mass on x), $F(\{x\})$). It makes sense to restrict attention to distributions F with support in $\{x_1, \ldots, x_n\}$, that is, absolutely continuous wrt the empirical F_n: $F << F_n$, and F_n does indeed maximise the likelihood over these F (Kiefer & Wolfowitz, 1956). Then it makes sense to call $T(F_n)$ a NPMLE for $T(F)$, where T is some functional – the mean, for example.

Let $X, X_1, \ldots, X_n \ldots$ be iid random p-vectors, with mean $EX = \mu$ and covariance matrix Σ of rank q. In higher dimensions, the distribution function, $P(. \leq .)$, which leads to *confidence intervals*, is replaced by $P(. \in .)$, which leads to *confidence regions* (which covers the unknown parameter with some probability); convexity is a desirable property of such confidence regions. For $r \in (0, 1)$, let

$$C_{r,n} := \{ \int X dF : F << F_n, L(F)/L(F_n) \geq r \}.$$

Then $C_{r,n}$ is a convex set, and

$$P(\mu \in C_{r,n}) \to P(\chi^2(q) \leq -2 \log r) \qquad (n \to \infty)$$

(the rate is $O(1/\sqrt{n})$ if $E[\|X\|^4] < \infty$). This is a non-parametric analogue of Wilks' Theorem (II.3 above) (A. Owen 1990; P. Hall 1990): "$-2 \log LR \sim \chi^2(q)$". For a monograph account, see Owen [O].

In view of results of this type, it is common practice, when we want the distribution of $T(F)$ when F is unknown, to use $T(F_n)$ as an approximation for it. This is commonly known as a *plug-in estimator* (just plug it in as an approximation when we need the exact answer but do not know it); 'empirical estimator', or 'NPMLE', would also be reasonable names.

Suppose we want to estimate an unknown density f, which is known to be *decreasing* on $[0, \infty)$(example: the exponential). A density is the derivative of a distribution; a concave function has a decreasing derivative (when differentiable). The NPMLE f_n of such a density is the (left-hand) derivative of the *least concave majorant* of F_n (Grenander, 1956). This example is interesting in that a CLT is known for it, but with an unusual rate of convergence

– *cube-root asymptotics* (Kim and Pollard 1990):

$$n^{1/3}(f_n(t) - f(t)) \to |4f'(t)f(t)|^{1/3}\text{argmax}_h(B(h) - h^2),$$

with B BM and argmax the argument where the maximum is attained .
Semi-parametrics.

Consider the *elliptical model*, with multidimensional density

$$f(\mathbf{x}) = const.g(Q(\mathbf{x})), \qquad Q(\mathbf{x}) = (\mathbf{x} - \mu)^T\Sigma^{-1}(\mathbf{x} - \mu).$$

Here $g : \mathbb{R}_+ \to \mathbb{R}_+$ is a function, the *density generator*, to be estimated. This is the *non-parametric* part of the model; (μ, Σ) is as above, the *parametric* part of the model. The model as a whole is then called *semi-parametric*.

Such models are very suited to financial applications. Notice how they generalise the multivariate normal or Gaussian (recall Edgeworth's theorem of IV.3). The parametric part (μ, Σ) is clearly needed in financial modelling, because of Markowitz's work on risk (Σ) and return (μ), and diversification (Σ again) (I.5, Day 2). The non-parametric part g allows us to choose a g that reflects the tail-behaviour observed in the data. For instance, for financial return data, it turns out that the *return interval*, Δ is crucial. For Δ long (monthly returns, say – though the rule of thumb is that 16 trading days suffice), the Gaussian ($g(x) = e^{-\frac{1}{2}x}$) suffices. This is an instance of *aggregational Gaussianity* – in other words, the Central Limit Theorem (CLT – see e.g. SP). For intermediate Δ – daily returns, say – the *generalised hyperbolic (GH)* distributions have been found to fit well. For short Δ – high-frequency data (tick data), g decreasing like a power (*Pareto tails*, or *heavy tails* – e.g. Student t) is both observed and predicted theoretically (the renormalisation group in Physics). These models have been extensively studied; see e.g. [BKRW], and [BFK] for some applications. In some cases, ignorance of one part of the model imposes no loss of efficiency when estimating the other part. This is the case for the elliptic model above, essentially for reasons to do with invariance under the action of the affine group. See [BKRW], 4.2.3, 6.3.9, 7.2.4, 7.8.3 for the theory, [BFK] for some applications.
Note. For *Gaussian* returns (say, monthly data), the density decreases extremely rapidly (far more so than is observed in practice!); the log-density decreases quadratically. In the generalised hyperbolic case (say, daily data), the log-density decreases only linearly (recall that a hyperbola approaches linear asymptotes). In the high-frequency case (say, tick data), the density decays like a power (say, like Student t).

4. Limit theorems; Markov chains; MCMC

We quote (see e.g. SP, PfS):

1. Strong Law of Large Numbers (SLLN): if $X_1, X_2, \ldots$ are independent and identically distributed (iid), with each $X_n, X \sim F$, then

$$\frac{1}{n}\sum_1^n X_i \to E[X] = \mu := \int x dF(x) \qquad (n \to \infty) \qquad a.s.$$

This includes as a special case the Weak Law of Large Numbers (WLLN), with convergence in probability in place of convergence a.s.

2. Central Limit Theorem (CLT). If also the X_n have variance $\sigma^2 < \infty$, then

$$\frac{1}{\sigma\sqrt{n}}\sum_1^n (X_i - \mu) \to N(0,1) \qquad (n \to \infty) \qquad \text{in distribution.}$$

So if f is such that $f(X_n)$ also has (finite) mean and variance, then

$$\frac{1}{n}\sum_1^n f(X_i) \to E[f(X)] \quad a.s.; \qquad \frac{1}{\sqrt{n\ var\ X}}\sum_1^n (f(X_i) - E[f(X)]) \to N(0,1).$$

The mode of convergence here is convergence in distribution, also known as weak convergence. This is weaker than convergence in probability, but when the limit is a constant (as in WLLN), the two are equivalent.

The convergence in the Glivenko-Cantelli theorem is uniform a.s., which is very strong. Similarly for weak convergence: for bounded continuous f,

$$\int f dF_n \to \int f dF : \qquad \frac{1}{n}\sum_1^n f(X_i) \to E[f(X)] \quad a.s.,$$

as above. The CLT above follows similarly from Donsker's theorem.

All this can be generalised far beyond the setting above of the iid case. We can work with *Markov chains* (see e.g. PfS VII) (discrete time will suffice for us, but the theory can be developed in continuous time). In PfS VII Markov chains are developed for discrete state spaces (finite or countably infinite). The definition of the Markov property is that, for predicting the future, knowing where one is at the present is all that matters – if we know where we are, how we got there is irrelevant. This irrelevance of the past suggests that as time passes the past 'becomes forgotten', and the chain settles down to some sort of steady state or equilibrium distribution, π – even

to a limit distribution π in favourable cases. Some Markov chains have no limit distribution (e.g., the trivial chain on the integers, which just moves 1 to the right at each step). But many Markov chains do have an equilibrium distribution, and even (if periodicity complications are absent) a limit distribution. See e.g. PfS VII for details. In particular, we need the idea of *detailed balance* (DB). A Markov chain with transition probability matrix $P = (p_{ij})$ and limiting distribution $\pi = \pi_i$ satisfies the *detailed balance* condition (DB) if $\pi_i p_{ij} = \pi_j p_{ji}$ for all i, j. We quote (Kolmogorov's theorem) that this is the same as *time-reversibility*.

When the Markov chain has suitably good properties (which ensure a limit distribution) – typically, appropriate *recurrence* properties, of returning repeatedly to its starting point – then the Markov chain satisfies a SLLN and a CLT as above. We shall not give details (see e.g. [MeyT] Ch. 17).

It turns out that all this carries over to continuous-state Markov chains (the case relevant to Statistics), subject to suitable restrictions on the chain, of which *Harris recurrence* is the best known.

Markov Chain Monte Carlo (MCMC); Hastings-Metropolis algorithm (HM)

We briefly sketch this; see VII.6 below for statistical applications.

The aim here is to sample from a distribution π. This may be straightforward (see IS); if not, we may proceed as follows. We construct a Markov chain $X = (X_n)$ for which π is the limit distribution (we assume this has a density, also written π). HM selects a transition density $q(x, .)$ (see below for choice of q), and then at each step, conditional on $X_{k-1} = x$, HM *proposes* a new value Y_k drawn from this transition density $q(x, .)$. This value Y_k is *accepted* as the new value X_k with probability

$$p(x, y) := \min\Big(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\Big);$$

otherwise, X_k is taken as the previous value X_{k-1}. One can check that this does indeed define a Markov chain, which satisfies (the continuous form of) (DB) and has invariant (= equilibrium) distribution π. Here $q(x, y) := p(|x - y|)$, for some transition density p of a symmetric random walk (the choice is usually not critical). What is critical in applying MCMC in practice is the rate of convergence. We have to run the chain for a long enough 'burn-in' period for it to be 'approximately in equilibrium'.

VII. BAYESIAN STATISTICS

1. Classical statistics and its limitations.

Broadly speaking, statistics splits into two main streams: (i) classical, or frequentist, and (ii) Bayesian. Much of classical statistics is devoted to the following general areas: Estimation of parameters (I), Hypothesis testing (II). Again, this is not exhaustive: the main remaining area is Non-parametric statistics (VI). Estimation of parameters itself splits, into
(ia). Point estimation [e.g., maximum-likelihood estimates],
(ib). Interval estimation [e.g., confidence intervals].
Both these are open to interpretational objections. A point estimate is a single number, which will almost certainly be wrong [i.e., will differ from the value of the parameter it estimates]. How wrong? And how to proceed?

A confidence interval is more informative, because it includes an error estimate. For instance, its mid-point can be regarded as a point estimate, and half its length as an error estimate – leading to conclusions of the form

$$\theta = 3.76 \pm 0.003 \qquad (*)$$

– with *confidence* 95% [or 99 %, or whatever]. What does this mean? It is not a probability statement:
either θ lies between 3.73 and 3.79 [(*) is true, so holds with pr. 100 %]
or it doesn't [(*) is false, so holds with pr. 0 %].
Problem: We don't know which!
Interpretation. If a large number of statisticians independently replicated the analysis leading to (*), then about 95 % of them would succeed in producing confidence intervals covering the unknown parameter θ. But
(a) We wouldn't know *which* 95 %,
(b) This is of doubtful relevance anyway. The large number of independent replications will usually never take place in practice. So confidence statements like (*) lack, in practice, a direct interpretation. [They are 'what happens to probability statements in classical statistics when we put the numbers in'.]

A further problem is that small changes in our data can lead to abrupt discontinuities in our conclusions. In borderline situations, θ 'just within' the confidence interval and 'just outside' represent diametrically opposite outcomes, but the data may be very close. Small changes in input *should* only lead to small changes in output, rather than abrupt changes.

Hypothesis testing is open to similar objections. It is usual to have a null

hypothesis, H_0, representing our present theory (the 'default option'), and an alternative hypothesis, H_1, representing some proposed alternative theory. At the end of the investigation, we have to choose between two alternatives. We may be wrong: we may
reject H_0 when it is true, and choose H_1 [Type I error, probability α, the significance level], or
reject H_1 when it is true, and choose H_0 [Type II error, probability β].
We then have a trade-off between α and β. It is not always clear how to do this sensibly, still less optimally [it is customary to choose $\alpha = 0.05$ or 0.01, and then try to minimise β, but this is merely conventional]. Again, problems present themselves:
(i) We won't know whether our choice between H_0 and H_1 was correct;
(ii) Small changes in the data can lead to abrupt changes between choosing H_0 and choosing H_1.

Thus both the main branches of classical parametric statistics lead to abruptly discontinuous conclusions and present interpretational difficulties. One justification for Bayesian statistics is that it avoids these. There are many others: we shall argue for Bayesian statistics below on its merits.

2. Prior knowledge and how to update it.

The difficulties identified above arise because in classical statistics we rely *entirely* on the data, that is, on the sample we obtained. The mathematics involved in classical statistics amounts to comparing the sample we actually obtained with the large (usually, infinite) class of hypothetical samples we might have obtained but didn't. These include the samples that we would obtain if we repeated our sampling independently – or that other statisticians would obtain if they independently replicated our work. This is where the term 'frequentist' for classical statistics originates: e.g., in 95 % confidence intervals, independently replicated confidence intervals would cover the parameter θ with frequency 0.95.

The other aspect of classical statistics crucial for our purposes is that it ignores everything before sampling. This is often unreasonable. For instance, we may know a good deal about the situation under study, based on prior experience. Such situations are typical in, e.g., industrial quality control: suppose we are employed by a rope manufacturer, and are testing the breaking strain of ropes in a current batch. We may have to hand large amounts of data obtained from tests on previous batches from the same production line. Similarly for a scientist, testing a scientific hypothesis.