VI: NON-PARAMETRIC STATISTICS

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 – 2013)): *all 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. So it remains to prove that this holds *uniformly* in x.

Now fix a finite partition $-\infty = x_1 < x_2 < \ldots < x_m = +\infty$. By monotonicity of F and F_n,

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

(check). 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 – the *Kolmogorov-Smirnov 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*

$$B_0(t) := B(t) - tB(1), \qquad t \in [0,1],$$

one has

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

(B_0 is called the Brownian bridge, as $B_0(0) = 0$ and $B_0(1) = 0$, so B_0 gives a (Brownian) bridge betwen the points $(0,0)$ and $(0,1)$.) This is *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 (with distance between two probability distributions measured in any reasonable sense, e.g. by the Lévy metric), 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 – free, like (La)TeX, and the proprietary package S-Plus, are derived), see e.g. [VR], 5.6.

Note. By far the best choice of a general programming language for use with

statistical data is R nowadays.

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 (the second term).

Although the minimisation above is over an infinite-dimensional space, there is a unique finite-dimensional minimiser, the cubic spline with knots at the data points. The minimising value and the fitted values are derived in, e.g., [BF, §9.2]; cf. [LX, §7.2.4].

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.

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

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

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.5. 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. 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

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

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, W2). 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.

a. *Long (macro).*

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).

b. *Intermediate (meso).*

For intermediate Δ – daily returns, say – the *generalised hyperbolic (GH)* distributions have been found to fit well.

c. *Short (micro).*

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).

Note that it is sensible to think in terms of *returns*, rather than prices or log-prices – but that this commits us to think also in terms of the *return interval*. But this is a sensible discipline, in the financial world.

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 if

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \text{for all } i,\, j. \tag{DB}$$

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'.