pfsl11.tex

Lecture 11. 1.11.2012 (half-hour: Problems)

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* (Kolmogorov in 1933, N. V. SMIRNOV (1900-1966) in 1944)

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

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 *Donsker 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* (below) in this context – the *von Mises calculus* (Richard von MISES (1883-1953), or *infinite-dimensional delta method.*

Variance-Stabilising Transformations

In exploratory data analysis (EDA), the scatter plot may suggest that the variance is not constant throughout the range of values of the predictor variable(s). But, the theory of the Linear Model *assumes* constant variance. Where this standing assumption seems to be violated, we may seek a systematic way to *stabilise* the variance – to make it constant (or roughly so), as the theory requires.

If the response variable is y, we do this by seeking a suitable function g (sufficiently smooth – say, twice continuously differentiable), and then *transforming* our data by

$$y \mapsto g(y).$$

Suppose y has mean μ:

$$Ey = \mu.$$

Taylor expand $g(y)$ about $y = \mu$:

$$g(y) = g(\mu) + (y - \mu)g'(\mu) + \frac{1}{2}(y - \mu)^2 g''(\mu) + \dots$$

Suppose the bulk of the response values y are fairly closely bunched around the mean μ. Then, approximately, we can treat $y - \mu$ as small; then $(y - \mu)^2$ is negligible (at least to a first approximation, which is all we are attempting here). Then

$$g(y) \sim g(\mu) + (y - \mu)g'(\mu).$$