LECTURE 1

Origins

In a set of lectures on curve estimates in the case of independent and dependent observations, it is perhaps appropriate to have initial comments that are in part reminiscences of earlier days and work in the area before proceeding to a discussion of current problems and research. This can provide a motivation for the later development. It is rather doubtful whether comments of this sort can be taken seriously as scientific history. Perhaps attempts at such reconstruction can only convince one of the difficulties involved in writing history. Nonetheless they can give a personal perspective of the time.

During and after World War II there was a good deal of interest in dependent processes as models. One of the models examined probabilistically even before the war was that of a weakly stationary sequence of random variables x_t, $t = \dots, -1, 0, 1, \dots,$ that is, a sequence with constant mean and covariance function

$$r_{n-m} = \operatorname{cov}(x_n, x_m)$$

depending only on the time difference $n - m$. Such a sequence r_m is positive definite so that for any complex constants c_j,

$$\sum_{j,k=1}^{n} c_j r_{j-k} \bar{c}_k \geq 0$$

for each positive integer n. An old result due to Herglotz (1911) states that the covariances are Fourier–Stieltjes coefficients of a bounded nondecreasing function G,

$$r_n = \int_{-\pi}^{\pi} e^{in\lambda} \, dG(\lambda).$$

In current terminology, $G(\lambda)$ is called the spectral distribution function of the random sequence x_n. There is a continuous time analogue of this result in which one considers a random process $x(t)$ continuous in mean square,

$$\lim_{t \to s} E|x(t) - x(s)|^2 = 0.$$

The continuity in mean square implies that the weakly stationary process $x(t)$ has a continuous covariance function

$$r(t - \tau) = \operatorname{cov}(x(t), x(\tau)).$$

The covariance function $r(t)$ is positive definite as in the discrete case and so has a Fourier–Stieltjes representation in terms of a bounded nondecreasing function G which is now

$$r(t) = \int_{-\infty}^{\infty} e^{it\lambda}\, dG(\lambda)$$

(Bochner's theorem).

In common models, the spectral distribution function has a number of jumps (corresponding to discrete harmonics) and elsewhere a density

$$g(x) = \frac{dG(\lambda)}{d\lambda}.$$

If there are no discrete harmonics, $g(\lambda)$ is called the spectral density of the process. Suppose one has a sequence of observations $x_1, \dots, x_n$ of the process. Schuster (1898) suggested using the periodogram

$$I_n(\lambda) = \frac{1}{2\pi n} \left| \sum_{j=1}^{n} x_j e^{-ij\lambda} \right|^2$$

to detect harmonics in data. Of course, in those days, very simple models were used. The periodogram is basically the modulus squared of a finite Fourier transform of the data. Though the periodogram is useful in isolating harmonics, it has no direct value in estimating the continuous part of the spectrum or spectral density. If x_n is a Gaussian process with mean zero and a continuous spectral density, then

$$EI_n(\lambda) \to g(\lambda)$$

as $n \to \infty$ but

$$\lim_{n \to \infty} \operatorname{cov}[I_n(\lambda), I_n(\mu)] = \begin{cases} 0 & \text{if } \lambda \neq \mu,\ 0 \leq \lambda, \mu \leq \pi, \\ g^2(\lambda) & \text{if } \lambda = \mu,\ 0 < \lambda, \mu < \pi, \\ 2g^2(\lambda) & \text{if } \lambda = \mu,\ \lambda = 0, \pi. \end{cases}$$

It is clear from this that the periodogram is hopeless as an estimate of the spectral density since it is not consistent. However, Daniell (1946) noted the asymptotic orthogonality of periodogram values at different frequencies and suggested smoothing periodogram values in the neighborhood of the frequency of interest. Bartlett (1948) and Tukey (1949) initially considered estimates (local) of the spectral density function. This led to the consideration of estimates of the spectral density function,

$$g_n^*(\lambda) = \int_{-\pi}^{\pi} \omega_n(u-\lambda) I_N(u)\, du$$
$$= \frac{1}{2\pi} \sum_{\nu=-n}^{n} r_\nu^* \omega_\nu^{(n)} e^{-i\nu\lambda}$$

with

$$\omega_\nu^{(n)} = \int_{-\pi}^{\pi} \omega_n(u) e^{i\nu u}\, du$$

and

$$r_\nu^* = \frac{1}{n} \sum_{t=1}^{n-\nu} x_t x_{t+\nu}, \qquad r_\nu^* = r_{-\nu}^*, \qquad \nu \geq 0.$$

The smoothing weight functions $w_n(u)$ have total net mass one,

$$\int_{-\pi}^{\pi} w_n(u)\, du = 1,$$

and concentrate their mass more and more tightly in the neighborhood of zero as $n \to \infty$ [see Rosenblatt (1985)]. For each $\varepsilon > 0$,

$$\omega_n(u) \to 0$$

uniformly for $|u| \geq \varepsilon$ as $n \to \infty$. Under rather broad conditions such sequences $g_n^*(\lambda)$ are consistent in estimation of $g(\lambda)$ as $n \to \infty$, in particular if

$$\int_{-\pi}^{\pi} w_n^2(u)\, du = o(n)$$

as $n \to \infty$. Such estimates are intrinsically biased since on the basis of observations $x_1, \dots, x_n$, one cannot obtain nontrivial estimates of r_ν for $|\nu| > n$ and

$$g(\lambda) = \frac{1}{2\pi} \sum_\nu r_\nu e^{-i\nu\lambda}.$$

If x_t is a Gaussian process with mean zero and a continuous spectral density, the variance of the estimate

$$\sigma^2(g_n^*(\lambda)) \simeq \frac{2\pi}{n} g^2(\lambda) \int_{-\pi}^{\pi} w_n^2(u)\, du,$$

$0 < \lambda < \pi$, as $n \to \infty$. There is a doubling effect at $\lambda = 0$ and $\lambda = \pi$ and asymptotic uncorrelated behavior for the estimates corresponding to distinct nonnegative frequencies λ, μ.

It is remarkable that the suggestions of Daniell were anticipated a long time ago in a recently rediscovered paper of Einstein (1914). Heuristically something like the representation of the covariance function as a Fourier integral is anticipated. Then a local smoothing of a periodogram as suggested by Daniell is discussed. An extended discussion of Einstein's paper is given in Yaglom (1987).

Perhaps the main point to note is that the research mentioned here focused on local estimation [estimation of the spectral density rather than that of a more global character] of the spectral distribution function [investigated in Grenander and Rosenblatt (1956)]. Also it is curious that the context was that of dependent stationary sequences. My own exposure to these questions in spectral estimation arose in joint work with Grenander [Grenander and Rosenblatt (1953)] who was then visiting the University of Chicago statistics group of which I was then a member.

Global estimation in the case of independent observations, estimation of the sample distribution function, had been of interest for some time. Of course the sample distribution function is an unbiased estimate of the distribution function and its covariance properties were well known. More elaborate questions relating to the Kolmogorov and von Mises metrics had been investigated. On seeing a paper of Fix and Hodges (1951) on discrimination using a simple probability density estimate, it was a natural question for me [Rosenblatt (1956a)] to reflect on the mean and covariance of probability density estimates given by a kernel function with bandwidth and examine their asymptotic properties as $n \to \infty$ and the bandwidth tends to zero. In particular, this seemed to be especially persuasive in terms of the parallel with spectral density estimates in the case of stationary sequences.

Given any method of smoothing or representing a function, one can adapt the method to the estimation of a probability density or a regression function. The kernel estimate of a density function can be associated with the method of approximating a function by convoluting with a narrowly concentrated kernel. For it is obtained by convoluting the sample distribution function with a narrowly focused kernel,

$$f_n(x) = \frac{1}{nb_n} \sum_{j=1}^{n} \omega\left(\frac{x - X_j}{b_n}\right) = \frac{1}{b_n} \int \omega\left(\frac{x-u}{b_n}\right) dF_n(u).$$

Here, of course, $F_n(u)$ is the sample distribution function of the independent, identically distributed random variables X_j, $j = 1, \dots, n$.

Another method of approximation or interpolation has been based on spline functions. Splines already appear in the work of Euler. But much of the interest in splines as a method of approximation or interpolation has been due to the work of Schoenberg (1964). It is clear that such methods based on splines have a greater global stability than methods using, for example, polynomial approximation. Boneva, Kendall and Stefanov (1971) proposed using histosplines in density estimation. This seems to have been the first paper on methods involving splines proposed in density estimation. Associated with a spline there are knots. A polynomial spline of degree m (for example, a cubic spline) is an mth degree polynomial between neighboring knots and is globally continuously differentiable $m - 1$ times. A family of splines called B-splines is sometimes introduced for convenience in some representations [see de Boor (1978)]. B-splines of order k are $(k-1)$st degree polynomial splines that have a relative local character. Given knots $t = (t_i)$, the ith B-spline for this series of knots is

$$B_{i,k,t}(x) = (t_{i+k} - t_i)[t_i, \dots, t_{i+k}](\cdot - x)_+^{k-1}$$

for all x with $[t_i, \dots, t_{i+k}]$ the kth difference operator at the points $t_i, \dots, t_{i+k}$. One can show only k B-splines might have a specific interval $[t_j, t_{j+1}]$ in their support, $B_{j-k+1}, \dots, B_j$.

What has been termed a smoothing spline was proposed by Schoenberg. Given approximate values

$$y_i = g(x_i) + \varepsilon_i \tag{1.1}$$

of a smooth function g at the points $x_1, \dots, x_n \in [0,1]$, the object is to approximate g by the function f that minimizes

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_0^1 [f^{(m)}(t)]^2 \, dt. \tag{1.2}$$

The solution f is a polynomial spline of degree $2m-1$ with knots at $x_1, \dots, x_n$ that satisfies the natural boundary conditions

$$f^{(j)}(0) = f^{(j)}(1) \quad \text{for } j = m, \dots, 2m-1.$$

The interval $[0,1]$ is taken here for convenience. Obviously there is the natural modification for any finite interval or even the whole real line. Splines satisfying these boundary conditions are sometimes called natural splines. The original idea qualitatively is due to Whittaker (1923) who considered uniformly spaced data points and suggested using mth order differences rather than $\int_0^1 [f^{(m)}(t)]^2 \, dt$. If one looks at the specification (1.1), it is plausible to think of formalizing the model in a statistical context by thinking of the errors ε_i as independent identically distributed errors. Much of the interest in smoothing splines and various extensions of this idea are due to Wahba (1973) and her collaborators.

It is apparent that one could also think of orthogonal series expansions of an appropriate character as providing reasonable estimators. Čencov (1962) introduced this idea.

Various ideas that come up naturally in approximation theory [see Shapiro (1969)] crop up again in a number of the questions that arise here in a stochastic context.

Given an initial motivation in curve estimation (spectral) for stationary dependent sequences, it is interesting that research in probability density estimation for many years was pursued only in the domain of independent observations.